Maahanmuutot ja Suomen julkinen talous

Osa II

Elinkaarivaikutukset: Suomi, Irak ja Somalia

Samuli Salminen

SUOMEN PERUSTA

Kiitokset

Haluan kiittää tämän tutkimuksen tekijänä erityisesti Suomen Perusta -säätiötä kärsivällisyydestä, jolla tätä tutkimusprojektin II osan valmistumista on odotettu. Tämä mahdollisti aiheeseen perusteellisen syventymisen. Suomen Perustasta haluan vielä erikseen kiittää toiminnanjohtaja Simo Grönroosia, sillä ilman hänen myötävaikutustaan tätä tutkimustekstiä ei olisi koskaan syntynyt.

Lisäksi haluan kiittää Tilastokeskusta, Terveyden ja hyvinvoinnin laitosta (THL) sekä Maahanmuuttovirastoa hyvästä yhteistyöstä tutkimusaineiston kokoamisessa ja muokkaamisessa sekä aineistoa koskevassa konsultoinnissa.

Henkilöitä, joita haluan kiittää heidän antamastaan avustaan tämän tutkimusprojektin eri vaiheissa, ovat: Kai Järvikare, Matti Virén, Pauli Vahtera, Matti Sarvimäki (VATT), Marja Riihelä (VATT), Timo Matala (Tilastokeskus), Kaarlo Reipas (ETK), Tuuli Halonen, Joonas Reunamo, Teemu Vauhkonen, Harri Turunen sekä Aleksi Karhula.

Tilastotieteen ja ekonometrian puolelta kiitän seuraavia henkilöitä, joiden aikoinaan erinomaista opetusta tässä tutkimuksessa käytettäviin menetelmiin olen nyt parhaani mukaan pyrkinyt soveltamaan:

- Yrjö Vartiaa, joka innosti perehtymään tarkemmin todennäköisyyslaskentaan ja tilastolliseen päättelyyn.
- Petri Koistista, jonka erinomainen opetus johdatti laskennallisen tilastotieteen (*engl. computational statistics*) pariin.
- Kari Aurasta johdannosta selviytymisaikamallinnukseen ja pitkittäisaineistojen käyttöön tilastollisessa mallinnuksessa.

Ohjelmistopuolelta oman kiitoksensa ansaitsevat:

- Pitkälti automatisoitu posteriorijakaumien simulointiohjelma *Stan* (mc-stan.org).
- R-ohjelman parketti *Brms* (Buerkner 2015), joka antoi hyvän johdannon Stan-ohjelman käyttöön.

Helsingissä 18.3.2019, Samuli Salminen

© Suomen Perusta 2019

ISBN 978-952-7145-42-5 (PDF)

Etukannen ulkoasu: Janne Turunen.

Suomen Perusta -ajatuspaja on saanut toimintaansa tukea Opetus- ja kulttuuriministeriöltä.

Abstrakti

Suomen Perusta -ajatuspaja käynnisti vuonna 2013 tutkimushankkeen, jossa tutkimme maahanmuuttojen vaikutuksia Suomen julkiselle taloudelle. Tutkimuksen ensimmäinen osa *Maahanmuutot ja Suomen julkinen talous. Osa 1: Toteutuneet julkisen talouden tulot ja menot* julkaistiin 2.4.2015 ja sitä päivitettiin 17.4. lisäämällä kymmenen suurimman maahanmuuttomaan lisäksi maahanmuuttajaryhmien aluekohtainen jaottelu.

Julkisen talouden nettovaikutukset ovat henkilötasolla julkisen talouden tulojen ja menojen välinen erotus. Kun tutkimuksen I osassa selvitettiin jo *toteutuneita* julkisen talouden nettovaikutuksia, niin tässä tutkimuksen II osassa selvitetään julkisen talouden nettovaikutuksia kokonaisten *elinkaarien yli*. Suomessa syntyvälle kantaväestön henkilöille elinkaarivaikutukset tarkoittavat julkisen talouden nettovaikutuksia syntymästä kuolemaan, maahanmuuttajille puolestaan Suomeen muutosta kuolemaan tai lopulliseen poismuuttoon.

Tässä tutkimuksen 2. osassa elinkaarivaikutukset selvitetään sekä Suomessa syntyville kantaväestön henkilöille että Irakissa ja Somaliassa syntyneille maahanmuuttajille.

Tutkimustekstin rakenteesta. Tutkimusteksti on jaettu 3 osaan.

Ensimmäisessä osassa esitetään yleisiä asioita tutkimuksesta. Näitä ovat muun muassa tutkimuskysymykset sekä käsitteidenmäärittely (luku 1). Sen jälkeen ensimmäisessä osassa määritetään ehdolliset todennäköisyysjakaumat uusien maahanmuuttajien (ja Suomessa syntyvien kantaväestön henkilöiden) julkisen talouden nettovaikutusten elinkaarivaikutusten ennusteille (luku 2).

Toisessa osassa mallinnetaan johdannossa esitetyt uusien maahanmuuttajien elinkaarivaikutusten ennusteet sekä esitetään näiden (diskonttaamattomat) tulokset (luku 9). Elinkaarivaikutukset ja niiden ennusteet koostuvat 7 komponentista, joiden mallintaminen on käsitelty luvuittain (luvut 3–8). Elinkaarivaikutusten 7 komponentin mallintamiselle asetetaan tavoite, jonka mukaan niiden mallinnusten pohjalta on kyettävä luomaan uskottavat ja realistiset elinkaarivaikutusten ennusteet.

Tämän tavoitteen saavuttamiseksi elinkaarivaikutusten komponenttien mallien täytyy täyttää kolme kriteeriä:

1. Mallien pohjalta tehdyt ennusteet toistavat alkuperäisen havaintoaineiston.
2. Malleissa huomioidaan otanta-asetelman havaintoaineistoihin (paneeliaineistoja) tuottamat riippuvuudet. Paneeliaineistot mahdollistavat myös nk. havaitsemattoman heterogeenisuuden huomioon ottamisen ”uusia” tilastoyksiköitä (henkilöt, maassaolovuodet, maahanmuuttovuodet ja kalenterivuodet) koskevissa ennusteissa.
3. Malleissa huomioidaan erityisesti tärkeimmät aikaan liittyvät vaikutukset. Aikaan liittyviä vaikutuksia ovat ikään, maahanmuuttoikään, maassaoloaikaan, maassaolovuoteen ja kalenterivuoteen liittyvät vaikutukset.

Kolmannessa osassa esitetään, millaisia johtopäätöksiä elinkaarivaikutusennusteiden pohjalta voidaan tehdä niin kutsuttua *päätöksentekoteoriaa* soveltamalla koskien uusien maahanmuuttajien valintaa.

Kolmannessa osassa käydään aluksi läpi tutkimuksesta puuttuvat nettovaikutukset (luku 10). Koska päätöksentekoteoriaa varten tarvitaan nykyarvoon diskontatut elinkaariennusteet, niiden mallinnus ja ennustaminen esitetään luvussa 11. Luvussa 12 esitetään päätöksentekoteorian perusteet, ja miten sitä voidaan soveltaa uusien maahanmuuttajien valintaan.

Luvussa 13 esitetään nykarvoon diskontatut elinkaariennusteet ja mitä johtopäätöksiä niiden pohjalta voidaan tehdä uusien maahanmuuttajien valinnasta. Tulos on, että Suomen julkisen talouden kannalta Somaliassa ja Irakissa syntyneitä uusia maahanmuuttajia ei kannata valita muuttavaksi Suomeen. Lisäksi luvussa viitataan muihin pohjoismaisiin tutkimuksiin, joissa on tutkittu pakolaistaustaisten maahanmuuttajien julkisen talouden vaikutuksia/mittareita (esimerkiksi työllisyys ja työtulot) verrattuna kantaväestöihin.

Liitteissä on esitetty mallinnuksissa käytettyjä R-ohjelman koodeja, joiden käyttöä myös demonstroidaan.

Sisältö

I Johdanto **14**

1 Yleistä tutkimuksen rakenteesta **15**

1.1 Tutkimuksen motivaatio 15

1.2 Keskeisimmät määritelmät 16

1.3 Tutkimuskysymykset 18

1.4 Tutkimusaineisto 19

1.5 Miksi elinkaariennusteet? 20

1.5.1 Elinkaariennusteiden kaksi käyttötarkoitusta 21

1.5.2 Elinkaariennusteet vain uusille maahanmuuttajille ja vastasyntyneille 21

1.5.3 Elinkaariennusteet vain Irakissa ja Somaliassa syntyneille sekä Suomessa syntyville kantaväestön henkilöille 22

1.6 Huomautus englanninkielisten termien käytöstä 22

2 Elinkaarivaikutusten rakenne ja ennustaminen uusille maahanmuuttajille **23**

2.1 Elinkaarivaikutusten rakenne 23

2.1.1 Elinkaarivaikutusten komponentit 23

2.1.2 Elinkaarivaikutusten funktio 25

2.2 Elinkaarivaikutusten ennustaminen uusille maahanmuuttajille 26

2.2.1 Elinkaariennusteiden komponentit 26

2.2.2 Elinkaariennusteiden komponenttien parametrit 27

2.2.3 Elinkaariennusteissa käytetty havaintoaineisto 27

2.2.4 Elinkaariennusteiden komponenttien, komponenttien parametrien, muiden ennusteparametrien sekä havaintoaineiston yhteistodennäköisyysjakauma 28

2.2.5 Uusien maahanmuuttajien elinkaarivaikutusten ennustejakaumat 30

2.3 Elinkaarivaikutusten ennusteiden ehdolliset todennäköisyysjakaumat kuvaavat ennusteisiin sisältyviä epävarmuuksia 31

2.3.1 Elinkaariennusteiden ehdollinen todennäköisyysjakauma ja ennusteiden kokonaisepävarmuus 31

2.3.2 Elinkaarivaikutusten ennusteet ovat jatkuvia ehdollisia todennäköisyysjakaumia . 37

2.3.3 Epävarmuustekijät elinkaarivaikutusten ennustejakaumissa 42

2.3.4 Havaintoaineiston ulkopuolisten tekijöiden vaikutukset elinkaarivaikutusten ennustejakaumissa 48

II Elinkaarivaikutusten mallinnus ja ennustaminen uusille maahanmuuttajille 56

3 Olotilamallinnus 57

3.1 Olotilamallinnuksen tavoitteet ja rakenne 57

3.1.1 Olotilamallinnuksen tavoite 57

3.1.2 Olotilojen aineisto ja olotilasiirtymät 57

3.2 Suomessaoloaikojen mallinnus 61

3.2.1 Selviytymisaika- ja tapahtumahistoria-analyysi 62

3.2.2 Paranemisosuusmallin luokan valinta 69

3.2.3 Suomessaoloaikojen mallinnus De Castron et al. (2014) menetelmällä 75

3.2.4 Estimoitavien parametrien priorijakaumat ja posteriorijakauma 83

3.2.5 Posteriorijakauman simulointi ja konvergoituminen tavoitejakaumaan 85

3.2.6 Mallin sopivuuden tarkistaminen: posterioriennustetarkistukset 89

3.2.7 Havainnot, estimointitulokset ja posterioriennustetarkistukset 92

3.3 Ulkomaillaoloaikojen mallinnus 109

3.3.1 Ulkomaillaoloaikojen mallinnus De Castron et al. (2014) menetelmällä 109

3.3.2 Havainnot, estimointitulokset ja posterioriennusteet 111

3.4 Kuolemien mallinnus 126

4 Työikäisten ja nuorten nettovaikutusmallinnus 127

4.1 Mallinnustavoitteet 127

4.1.1 Alkuperäisen havaintoaineiston toistaminen 127

4.1.2 Otanta-asetelman havaintoaineistoon tuottamien riippuvuuksien huomiointi . . . 128

4.1.3 Tärkeimpien aikaan liittyvien vaikutusten huomiointi 129

4.2 Havaintoaineisto 129

4.2.1 Havaintoaineisto on kuvattu kattavasti toteutuneita nettovaikutuksia käsitteleväs-sä tutkimusprojektin I osassa 129

4.2.2 Erot havaintoaineistossa verrattuna tutkimusprojektin I osassa esitettyihin julki-sen talouden nettovaikutuksiin 130

4.3 Nettovaikutusten mallinnus 132

4.3.1 Paneeliaineistot, havaitsematon heterogeenisuus ja hierarkkiset mallit 133

4.3.2 Nettovaikutusten aineistomallit 138

4.3.3 Aineistomallin mallinnusaskeleet 142

4.3.4 Estimoitavien parametrien priorijakaumat 158

4.3.5 Estimoitavien parametrien yhteisposteriorijakauma 160

4.3.6 Posteriorijakauman simulointi ja konvergoituminen tavoitejakaumaan 160

4.3.7 Mallin sopivuuden tarkistaminen: posterioriennustetarkistukset 165

4.4 Havaintoaineistot, estimointitulokset ja posterioriennustetarkistukset 169

4.4.1 Irak 173

4.4.2 Somalia 198

4.4.3 Suomi 221

5 Palkkasummien mallinnus 241

5.1 Mallinnustavoitteet 241

5.2 Palkkasummien havaintoaineisto 243

5.2.1 Palkkasummien laskeminen eläkemaksuja koskevista rekisteritiedoista 243

5.2.2 Palkkasumma-aineistojen kuvaus 245

5.3 Palkkasummien mallinnus 247

5.3.1 Palkkasummien aineistomallit 248

5.3.2 Estimoitavien parametrien priorijakaumat 257

5.3.3 Estimoitavien parametrien yhteisposteriorijakauma 259

5.3.4 Posteriorijakauman simulointi ja konvergoituminen tavoitejakaumaan 259

5.3.5 Mallin sopivuuden tarkistaminen: posterioriennustetarkistukset 260

5.4 Havainnot, estimointitulokset ja posterioriennustetarkistukset 260

5.4.1 Somalia 263

5.4.2 Irak 288

5.4.3 Suomi 312

6 Muut ennusteparametrit ja eläkkeisiin liittyvien nettovaikutusten ennustaminen 330

6.1 Muut ennusteparametrit 330

6.1.1 Diskonttauskertoimet ja niillä ehdolliset parametrit 332

6.1.2 Työnantajien maksujen osuudet palkoista 336

6.1.3 Eläkkeiden karttumisen osuus palkoista ja palkattomilta ajoilta 340

6.1.4 Maksettavien kansaneläkkeiden määrä 341

6.1.5 Ennustetuista eläkeajan tuloista maksettavat verot (veroprosentit) 342

6.2 Eläkkeisiin liittyvien vaikutusten ennustaminen palkkasummaennusteista ja muista ennusteparametreista 344

6.2.1 Ennustetut työnantajien maksamat lakisääteiset työeläkevakuutusmaksut ja muut sosiaalivakuutusmaksut 345

6.2.2 Ennustetut saatavat kokonaiseläkkeet (työeläke + kansaneläke + takuueläke) . . . 345

6.2.3 Ennustetut maksettavat verot eläkeikäisenä 347

6.2.4 Ennustetut maksettavat eläkeläisten asumistuet 349

7 Alle 15-vuotiaiden ja yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutukset 353

7.1 Alle 15-vuotiaiden nettovaikutukset 353

7.1.1 7–14-vuotiaiden nettovaikutukset 353

7.1.2 0–6-vuotiaiden nettovaikutukset 354

7.2 Yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutukset 355

7.2.1 Yli 62-vuotiaiden terveys- ja sosiaalipalveluiden menot 356

7.2.2 Yli 62-vuotiaiden muiden julkisten palveluiden menot 359

8 Maahanmuuttajien lasten elinkaarivaikutukset 360

8.1 Yleistä: maahanmuuttajien lasten yhteenlasketut elinkaarivaikutukset 360

8.2 Maahanmuuttajien lasten nettovaikutusten ennustaminen 361

8.2.1 Alle 8-vuotiaina muuttaneet ja Suomessa syntyneet lapset 361

8.2.2 Tiedot maahanmuuttajien lapsista 22-vuotiaina 362

8.2.3 Julkisen talouden ennustetut nettovaikutukset 32–35-vuotiaana 362

8.2.4 32–62-vuotiaiden nettovaikutusennusteet 32–35-vuotiaiden ennustetuista nettovaikutuksista 363

8.2.5 Maahanmuuttajien lasten 32–62-vuotiaiden nettovaikutusten ennusteiden todennäköisyysjakauma 364

8.2.6 32–62-vuotiaiden palkkasummaennusteiden ero verrattuna kantaväestöön 35–39-vuotiaiden ennustetuista palkkasummista 366

8.3 Mallinnus- ja ennusteaineistot 367

8.3.1 Suomessa syntyneiden 22-vuotiaiden mallinnusaineisto 367

8.3.2 Maahanmuuttajien lasten määritelmästä ja ennusteissa käytettävästä osajoukosta . 367

8.3.3 Havaitut tiedot 22-vuotiaista maahanmuuttajien lapsista 368

8.3.4 Maahanmuuttajaryhmäkohtaiset tiedot 22-vuotiaista maahanmuuttajien lapsista vuosina 2005–2014 371

8.4 Mallinnus: julkisen talouden nettovaikutukset 32–35-vuotiaana 22-vuotiaita koskevien tietojen perusteella 377

8.4.1 Estimoitavien parametrien posteriorijakauman määritys ja simulointi 377

8.4.2 Havainnot, estimointitulokset ja posterioriennustetarkistus 378

8.5 Mallinnus: palkkasummat 35–39-vuotiaana 22-vuotiaita koskevien tietojen perusteella . 387

8.5.1 Estimoitavien parametrien posteriorijakauman määritys ja simulointi 387

8.5.2 Havainnot, estimointitulokset ja posterioriennustetarkistus 388

8.6 Lasten määrien ennustaminen 395

8.6.1 Äitien mukana Suomeen muuttaneiden lasten lukumäärät 396

8.6.2 Naisten Suomessa saamien lasten lukumäärien ennustaminen 396

8.7 Yhteenveto: maahanmuuttajien lasten yhteenlaskettujen elinkaarivaikutusten ennusteet . 400

9 Elinkaariennusteet 401

9.1 Yleistä elinkaariennusteista 401

9.2 Ennustepopulaatioiden poiminta 402

9.3 Elinkaariennusteiden simulointi ennustepopulaatioille 404

9.4 Elinkaarivaikutusten komponenttien ennusteet mallinnuksen pohjalta 405

9.4.1 Olotila-aikojen ennusteet 405

9.4.2 Nettovaikutusennusteet työikäisille ja nuorille 407

9.4.3 Palkkasummaennusteet ja eläkkeisiin liittyvät nettovaikutusennusteet 410

9.4.4 Alle 15-vuotiaiden ja yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutusennusteet 414

9.4.5 Maahanmuuttajien lasten yhteenlasketut nettovaikutusennusteet 414

9.4.6 Suomessa syntyneiden elinkaariennusteiden odotusarvoksi nolla euroa 419

9.4.7 Parannettavaa elinkaariennusteissa 419

9.4.8 Elinkaariennusteiden komponenttien keskimääräiset kontribuutiot elinkaariennusteisiin: esimerkki 420

9.5 Uusien maahanmuuttajien (diskonttaamattomat) elinkaariennusteet 422

9.5.1 Somalia 422

9.5.2 Irak 443

9.6 Suomessa syntyvien (diskonttaamattomat) elinkaariennusteet 459

9.6.1 Elinkaariennusteet koko ennustepopulaatiolle yhteensä 459

9.6.2 Elinkaariennusteet sukupuolen mukaan 460

III Elinkaariennusteet ja uusien maahanmuuttajien valinta **461**

10 Elinkaariennusteista puuttuvat julkisen talouden vaikutukset **462**

10.1 Elinkaariennusteista puuttuvat julkisen talouden vaikutukset 462

10.2 Puuttuvat julkisen talouden menot 463

10.2.1 Puutteet maahanmuuttajien palveluiden menojen osalta 463

10.2.2 Muut puutteet julkisten palveluiden menojen osalta 466

10.3 Puuttuvat julkisen talouden tulot 467

10.4 Tutkimukseen sisältyvät ja siitä puuttuvat yhteisöjen ja yhtymien vaikutukset 468

10.4.1 Henkilöverotuksen kautta tutkimukseen sisältyvät julkisen talouden vaikutukset . 469

10.4.2 Yhteisöverot eivät sisälly julkisen talouden vaikutuksiin tässä tutkimuksessa . . . 469

10.5 Yhteenveto puuttuvista julkisen talouden vaikutuksista 472

11 Elinkaariennusteiden diskonttaus nykyarvoon **476**

11.1 Ennustettujen julkisen talouden nettovaikutusten nettonykyarvo 476

11.2 Diskonttauskertoimien havaintoaineisto 478

11.2.1 Rahanarvonkertoimen vuosimuutokset 1922–2015 478

11.2.2 Suomen Pankin peruskorko vuosina 1867–2015 480

11.2.3 Diskonttauskertoimet vuosina 1950–2015 481

11.3 Diskonttauskertoimien mallinnus ja ennustaminen 485

11.3.1 Diskonttauskertoimien mallinnus 485

11.3.2 Diskonttauskertoimien ennustaminen 487

12 Päätöksentekoteoria ja Suomeen muuttavien uusien maahanmuuttajien valinta **494**

12.1 Yleistä päätöksentekoteoriasta 494

12.1.1 Ennustetun hyödyn odotusarvo valinnasta 494

12.1.2 Ennustetun hyödyn odotusarvon maksimointi valinnasta 495

12.1.3 Tuloksen hyötyfunktio ja sen posteriorijakauman päivitys lisäinformaatiolla . . . 496

12.2 Päätöksentekoteorian sovellus: julkinen talous ja potentiaalisen uuden maahanmuuttajan valinta . . . 498

12.2.1 Potentiaalisen uuden maahanmuuttajan valinta . . . 498

12.2.2 Julkisen talouden hyöty maahanmuuttajaa koskevan valintapäätöksen tuloksesta . 500

12.2.3 Tulos maahanmuuttajan valintapäätöksestä: uuden maahanmuuttajan elinkaarivaikutuksen posterioriennustejakauma . . . 502

12.2.4 Julkisen talouden ennustetun hyödyn odotusarvon maksimointi valintapäätöksestä 504

12.3 Aineiston ulkopuolisten tekijöiden vaikutukset ja potentiaalisen uuden maahanmuuttajan valinta . . . 505

12.3.1 Aineiston ulkopuolisten tekijöiden vaikutukset ja julkisen talouden hyödyn odotusarvon maksimointi . . . 507

12.3.2 Aineiston ulkopuolisten tekijöiden vaikutus yhteen malliparametriin . . . 509

12.3.3 Aineiston ulkopuolisten tekijöiden vaikutusten vektori ja elinkaarivaikutusten ennusteiden todennäköisyysjakauma . . . 515

12.3.4 Aineiston ulkopuolisten tekijöiden vaikutukset ja elinkaarivaikutusten ennusteet tässä tutkimuksessa . . . 520

12.4 Potentiaalisten uusien maahanmuuttajien valinta: ennustetun hyödyn odotusarvon maksimointi yli ajan . . . 521

12.4.1 Malliparametrien vektorin posteriorijakauman päivitys uudella havaintoaineistolla 522

12.4.2 Ennustemallien laajennus uusilla parametreilla ja ennustemalleissa käytettävien todennäköisyysjakaumien luokkien vaihtaminen . . . 523

12.4.3 Julkisen talouden ennustetun hyödyn odotusarvo yli ajan . . . 524

12.4.4 Julkisen talouden ennustetun hyödyn odotusarvon maksimointi yli ajan . . . 525

12.5 Tiivistelmä . . . 527

12.5.1 Yleistä päätöksentekoteoriasta . . . 527

12.5.2 Päätöksentekoteorian soveltaminen potentiaalisen uuden maahanmuuttajan valintaan . . . 528

13 Johtopäätökset: Irakissa ja Somaliassa syntyneiden potentiaalisten uusien maahanmuuttajien valinta 533

13.1 Yhteenveto . . . 533

13.1.1 Elinkaariennusteet ovat todennäköisyysjakaumia (I osa) 533

13.1.2 Elinkaariennusteiden komponenttien uskottava mallinnus ja ennustaminen uusille maahanmuuttajille (II osa) 534

13.1.3 Elinkaariennusteet ja päätöksentekoteoria: Potentiaalisten uusien maahanmuuttajien valinta (III osa) 536

13.2 Nykyarvoon diskontatut elinkaariennusteet 541

13.2.1 Somalia 542

13.2.2 Irak 558

13.2.3 Suomi 574

13.2.4 Lisäepävarmuutta sisältävät nykyarvoon diskontatut elinkaariennusteet 576

13.3 Johtopäätökset: Irakissa ja Somaliassa syntyneiden potentiaalisten uusien maahanmuuttajien valinta 580

13.3.1 Johtopäätökset diskontattujen elinkaariennusteiden odotusarvojen perusteella: Irakissa ja Somaliassa syntyneitä ei kannata valita muuttavaksi Suomeen 580

13.3.2 Johtopäätökset diskontattujen, lisäepävarmuutta sisältävien elinkaariennusteiden odotusarvojen perusteella: Irakissa ja Somaliassa syntyneitä ei kannata valita muuttavaksi Suomeen 581

13.3.3 Havaintoaineiston ulkopuolinen informaatio, mitä elinkaariennusteen tekijällä pitäisi olla, jotta valintapäätökset muuttuisivat. 581

13.4 Katsaus pohjoismaisiin tutkimuksiin, joissa tulokset on eritelty pakolaistaustaisuuden mukaan 583

13.4.1 Maassaoloaika 584

13.4.2 Työllisyys, työtulot ja saadut tulonsiirrot 587

13.4.3 Humanitaaristen maahanmuuttajien lasten peruskoulutodistukset Ruotsissa . . . 602

13.4.4 Julkisen talouden nettovaikutukset 605

A Olotilamallinnus: De Castron et al. (2014) kehittämän posteriorijakauman yksiulotteisen tapauksen simulointi Gibbsin otannalla. 620

B Vinon Studentin t-jakauman simulointi Stan-ohjelmalla. Nettovaikutusmallinnuksessa käytetyt Stan- ja BRMS-koodit. 637

C Hurdle-gamma-sekoitusjakauman simulointi Stan-ohjelmalla. Palkkasummamallinnuksessa käytetyt Stan- ja BRMS-koodit. 655

Osa I

Johdanto

Luku 1

Yleistä tutkimuksen rakenteesta

Maahanmuutot ja Suomen julkinen talous on kaksiosainen tutkimus, jossa tarkastellaan eri maahanmuuttajaryhmien julkisen talouden vaikutuksia. Maahanmuuttajaryhmät on määritetty tutkimuksessa syntymävaltion tai syntymävaltion maa-alueen mukaan. Aluejako on esitetty tutkimuksen I osassa: *Maahanmuutot ja Suomen julkinen talous. Toteutuneet julkisen talouden tulot ja menot* (Salminen 2015a). Syntymävaltiotasoinen tarkastelu keskittyy 10 suurimpaan maahanmuuttajaryhmään. Maahanmuuttajaryhmäkohtaisia vaikutuksia verrataan tutkimuksessa Suomessa syntyneen kantaväestön vaikutuksiin.

Tutkimuksen I osassa on esitetty tutkimuksen motivaatio (luku 1, Johdanto), yleisesti käytetyt menetelmät tutkimuskirjallisuudessa (luku 2), käytetyt käsitteet ja määritelmät (luku 3) sekä tutkimuskysymykset (luku 4). Käyn nämä lyhyesti ja valikoiduin osin läpi seuraavaksi, mutta kehotan samalla lukijaa tutustumaan I osan mainittuihin lukuihin.

1.1 Tutkimuksen motivaatio

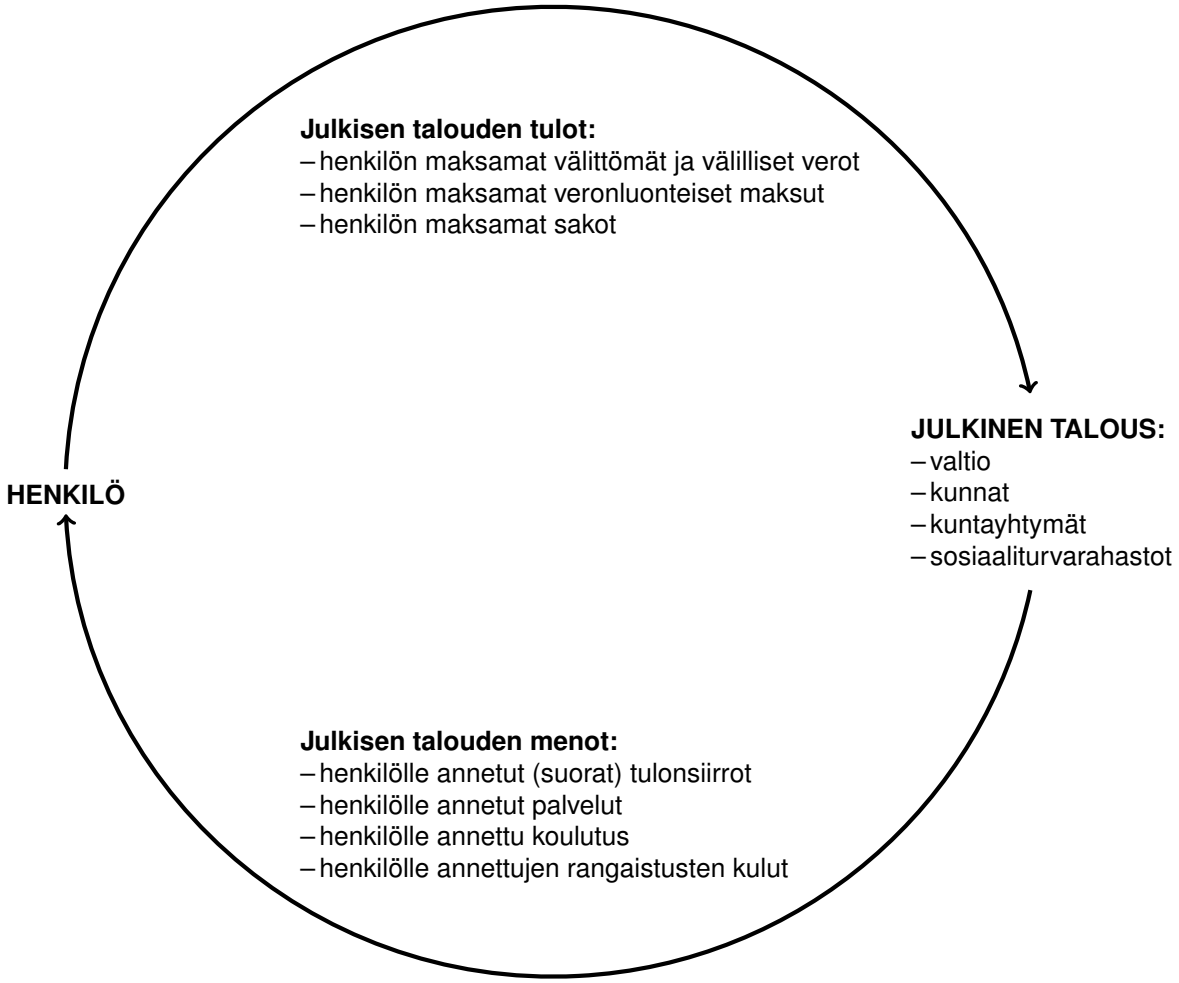
Tutkimuksen motivaationa on tutkia maahanmuuttajien julkisen talouden vaikutuksia siten, että syntymävaltion mukaan määriteltyjä maahanmuuttajaryhmiä käsitellään *eritellysti*. Korostaakseni tämän erittelyn tärkeyttä käytän tutkimuksen nimessä monikkoa ”maahanmuutot”. Kuten kirjoitan tutkimuksen I osan johdannossa, suurin osa (viittaamistani) alan tutkimuksista käsittelee maahanmuuton julkisen talouden vaikutuksia yhtenä kokonaisuutena. Tutkimuksen I osan tulokset (I osa, luku 6.9) osoittavat kuitenkin, että keskimääräiset julkisen talouden nettovaikutukset vaihtelevat ”paljon” eri maahanmuuttajaryhmien ja Suomessa syntyneen kantaväestön välillä.

Kymmenen suurimman maahanmuuttajaryhmän joukossa negatiivisimmat toteutuneet keskimääräiset julkisen talouden nettovaikutukset ovat Somaliassa ja Irakissa syntyneillä. Positiivisimmat keskimääräiset julkisen talouden nettovaikutukset ovat puolestaan Saksassa syntyneillä sekä Suomessa syntyneellä kantaväestöllä. Syntymävaltioiden maa-alueista negatiivisimmat keskimääräiset julkisen talouden nettovaikutukset ovat Lähi-Idässä, Pohjois-Afrikassa tai Keski-Aasiassa syntyneillä, positiivisimmat puolestaan

länsimaissa syntyneillä.

1.2 Keskeisimmät määritelmät

Julkisen talouden tulot ja menot on määritetty tutkimuksessa niin, että ne kohdistuvat tutkimuksessa henkilöihin, kuten kuviossa 1.1 on esitetty. Julkisen talouden tulot tulevat aina tietyltä henkilöltä. Julkisen talouden menot ovat joko yksilöllisiä, jolloin ne kohdistuvat tiettyyn henkilöön, tai kollektiivisia, jolloin menolla kustannettavan palvelun saajia ei ole mahdollista yksilöidä. Määritelmä ei pidä sisällään henkilöiden epäsuoria vaikutuksia julkiseen talouteen eli sellaisia mahdollisia vaikutuksia, jotka eivät sisälly alla esitettäviin määritelmiin julkisen talouden tuloista ja menoista, jotka kohdistuvat henkilöihin.



Kuva 1.1: Julkisen talouden tulot ja menot tutkimuksessa.

Julkisen talouden tulot

Julkinen talous saa tuloja henkilöiltä maksettujen verojen ja veronluonteisten maksujen sekä sakkojen muodossa. Verot voivat olla joko välittömiä tai välillisiä. Tähän tutkimukseen sisältyvät, rekisteritietoihin perustuvat, välittömät verot ja veronluonteiset maksut on lueteltu tutkimuksen I osan liitteessä 2. Tulokset maksetuista veroista ja veronluonteisista maksuista on esitetty tutkimuksen I osan luvussa 6.2.2. Estimaatit henkilöiden maksamista välillisistä veroista sisältyvät tutkimukseen (ks. tutkimuksen I osa, luku 6.2.4). Välillisillä veroilla tarkoitan tässä tutkimuksessa sekä arvonlisäveroja että valmisteveroja.

Tutkimukseen eivät sisälly yhteisöverot, sillä niissä veronmaksajana on yhteisö. Sen sijaan ne yritystoiminnasta maksetut verot, jotka henkilöt maksavat ansiotulo- ja pääomatuloveroina, sisältyvät tutkimukseen (ks. tarkemmin yritystoimintaan liittyvistä veroista luku 10.4). Henkilöiden maksamat kiinteistöverot eivät kuulu mukaan tutkimukseen, sillä en saanut niitä käyttöön tutkimusaineistooni kuin vuoden 2011 osalta. Lisäksi myös joitain muita, kertymältään pieniä veroja puuttuu tutkimuksesta (ks. edelleen luku 10.3 sekä tutkimuksen I osan luku 5.5 tutkimuksen ulkopuolelle jäävistä julkisten talouden tuloista ja menoista).

Työnantajien maksamat pakolliset eläkevakuutusmaksut ovat tutkimuksen määritelmien kannalta hieman ongelmallisia, sillä niissä maksajana ei ole henkilö, niin kuin julkisen talouden tulot on määritetty kuviossa 1.1 . Sisällytän ne kuitenkin mukaan elinkaarivaikutusten ennusteisiin, sillä henkilöiden saamia eläkkeitä maksetaan suureksi osaksi työnantajien maksamista pakollisista eläkevakuutusmaksuista. Havaintoaineistoon sisältyviä lakisääteisiä eläkevakuutusmaksuja sekä niiden ennustamista palkkasummista käsittelen luvuissa 5 ja 6. Ennustettujen lakisääteisten eläkevakuutusmaksujen vaikutuksia elinkaarivaikutusten ennusteisiin – erityisesti eläkeiän osalta – käsittelen edelleen luvussa 6.

Julkisen talouden menot

Julkisen talouden menot koostuvat tässä tutkimuksessa suorista tulonsiirroista, julkisen talouden maksamista sosiaali-, terveys- ja koulutuspalveluista sekä oikeusmenoista, kuten kuviossa 1.1 on esitetty. Tähän tutkimukseen sisältyvät, rekisteritietoihin perustuvat, henkilöiden saamat tulonsiirrot on lueteltu tutkimuksen I osan liitteessä 2. Tulokset saaduista tulonsiirroista on esitetty tutkimuksen I osan luvussa 6.2.1. Tutkimuksen I osan luvussa 6.3 esitetään tulokset toteutuneista sosiaali- ja terveyspalveluiden menoista, luvussa 6.4 opiskelumenoista, luvussa 6.5 oikeusmenoista, luvussa 6.6 työvoimakoulutusten menoista, luvussa 6.7 turvapaikanhakijoina ja kiintiöpakolaisina maahan tulleiden vastaanotto- ja kotoutumismenoista sekä luvussa 6.8 kollektiivisesti kulutetuista julkisen talouden palveluista. Julkisen talouden menojen osalta tutkimusaineistossa on puutteita erityisesti turvapaikanhakijoina ja kiintiöpakolaisina maahan tulleiden vastaanotto- ja kotoutumismenojen osalta (ks. tarkemmin I osan luvut 5.5 ja 6.7 sekä tämän II osan luku 10.2.1).

Kollektiiviset menot

Koska laadin elinkaariennusteet tässä tutkimuksen II osassa vain Suomessa *syntyville* henkilöille että uusille maahanmuuttajille, eli ulkomailta Suomeen *juuri muuttamassa* oleville, näiden henkilöiden osalta kollektiivisten palveluiden rajakustannusten oletan olevan nolla. Myös VATT (2014a) olettaa näin yh-

den Suomeen muuttavan tai syntyvän kohdalla (Liite VATT-analyysin lukuun 5 (VATT 2014b))). Vaikka lukumäärältään "suuren"henkilöjoukon kohdalla näiden henkilöiden yhteenlaskettu kollektiivisten palveluiden rajakustannus olisi nollaa suurempi – käsiteltäessä joko Suomessa syntyvien joukkoa tai jotain Suomeen juuri muuttamassa olevien joukkoa – oletan, että rajakustannus on näillä eri henkilöjoukoilla yhtä suuri. Kun tällöin tarkastellaan julkisten talouden nettovaikutusten elinkaariennusteiden summien erotuksia näiden joukkojen *välillä*, rajakustannusten välinen erotus on nolla, joten ne eivät vaikuta näiden joukkojen välisiin ennustetuloksiin.

Julkisen talouden nettovaikutus

Julkisen talouden nettovaikutus on julkisen talouden tulojen ja menojen välinen erotus.

1.3 Tutkimuskysymykset

Tutkimuksen I osan 4 luvussa esitettiin tutkimuksen I osan sekä tutkimuksen II osan tutkimuskysymykset. Esitän ne toistamiseen tässä.

Tutkimuksen ensimmäisen osan tutkimuskysymys on:

- Kuinka suuria ovat olleet toteutuneet julkisen talouden vuotuiset tulot, menot ja nettovaikutukset, kun ne eritellään syntymävaltioiden ja syntymäalueiden mukaan? Erittely tapahtuu syntymävaltioiden ja syntymäalueiden lisäksi sukupuolen, ikäluokan, maassaolovuosien sekä pääasiallisen toiminnan mukaan.

Tutkimuksen toisen osan tutkimuskysymys on:

- Kuinka suuria ovat toteutuvat julkisen talouden nettovaikutukset koko elinkaaren yli summattuina, kun ne eritellään syntymävaltioiden ja syntymäalueiden mukaan? Erittely tapahtuu syntymävaltioiden ja syntymäalueiden lisäksi sukupuolen, ikäluokan sekä maahanmuuttovuoteen liittyvien tekijöiden mukaan. Maahanmuuttovuoteen liittyviä tekijöitä voivat olla ikä maahan muuttaessa ja maahanmuuton kalenterivuosi.

  Tutkimuksen toiseen kysymykseen voidaan vastata joko Suomessa jo olevien maahanmuuttajien osalta tai Suomeen vasta muuttavien maahanmuuttajien osalta.

Ensimmäinen tutkimuskysymys määritti sen, että tutkimuksen I osassa tavoitteena oli *kuvata toteutuneita* julkisen talouden vaikutuksia. Toinen tutkimuskysymys puolestaan määrittää sen, että tässä tutkimuksen II osassa tavoitteena on *ennustaa tulevia* julkisen talouden nettovaikutuksia kokonaisten elinkaarien yli. Elinkaaren alku on maahanmuuttajilla sinä kalenterivuonna, kun he muuttavat ensimmäisen kerran Suomeen. Suomessa syntyneillä elinkaaren alku on syntymän kalenterivuosi. Kaikkien henkilöiden kohdalla, joille elinkaariennusteet tehdään, elinkaaren loppu on se kalenterivuosi, jolloin henkilö kuolee.

Tässä tutkimuksen elinkaarivaikutusten ennusteita käsittelevässä osassa esitän myös, kuinka julkisen talouden elinkaarivaikutusten ennusteita voidaan käyttää uusien maahanmuuttajien valitsemisessa niin, että odotettavissa olevat julkisen talouden vaikutukset Suomeen muuttavaksi – tai ei muuttavaksi – valittavista uusista maahanmuuttajista ovat mahdollisimman positiiviset.

1.4 Tutkimusaineisto

Tutkimuksen I osan luvussa 6 ja 7 esitettiin toteutuneita julkisen talouden tuloja ja menoja koskevat tulokset eli vastaus tutkimuksen I osan tutkimuskysymykseen. Vaikutukset on eritelty maahanmuuttajaryhmittäin, sukupuolittain, ikäluokittain sekä maassaoloajan mukaan. Osa tuloksista on eritelty myös pääasiallisen toiminnan mukaan.

Tutkimuksen I osan tulokset perustuvat tutkimusaineistoon, joka koostuu lähestulkoon pelkästään laajoista henkilövuositasoisista rekisteriaineistoista. Tutkimusaineistoon kuuluvat seuraavat osat:

1. Henkilön saamat rahamääräiset tulonsiirrot.
2. Henkilön maksamat välittömät ja välilliset verot sekä veronluonteiset maksut.
3. Henkilön yksilöllisesti kuluttamat julkisen talouden menot: esimerkiksi sosiaali- ja terveyspalvelut, koulutus.
4. Henkilöiden kollektiivisesti kuluttamat julkisen talouden menot: esimerkiksi maanpuolustus.

Yksilölliset julkisen talouden tulot ja menot (kohdat 1.–3.) selvitettiin tutkimuksessa pääasiassa käyttäen Tilastokeskuksen, THL:n ja Maahanmuuttoviraston hallinnoimia henkilövuositasoisia rekisteriaineistoja. Tarvittavat yksikkökustannustiedot on poimittu joko suomalaisista tutkimuksista, valtion talousarvioesityksistä tai Suomen virallisesta tilastosta. Tutkimuksen aineistollisena pohjana toimiva Tilastokeskuksen FLEED-aineisto sisältää kaikki Suomessa vuosina 1988–2011 asuneet 15–70-vuotiaat maahanmuuttajat. Lisäksi aineistossa oli 50 000 henkilön satunnaisotos samat kriteerit täyttävistä Suomessa syntyneistä. Tässä tutkimuksen II osassa on lisäksi käytössä henkilövuositasoiset rekisteriaineistot koskien lakisääteisiä eläkevakuutusmaksuja vuosina 2012–2014. Nämä tiedot löytyvät myös henkilöille, jotka ovat olleet ensimmäisen kerran FLEED-aineistossa vuosina 2012–2014.

Lähteenä toimivat rekisteriaineistot on luetteloitu tutkimuksen I osan lähdeluettelon rekistereitä koskevassa ensimmäisessä osassa. Tutkimuksen I osan luvussa 5 on esitetty tarkemmin käytetyt aineistot, ja liitteessä 2 on lueteltu ne suorat tulonsiirrot, jotka sisältyvät käytettyihin rekisteriaineistoihin.

Koska rekisteriaineistot eivät sisällä tietoja ainakaan osasta vastaanotto- ja kotouttamispalveluita, joitakin julkisen talouden tuloja ja menoja jää tutkimuksen ulkopuolelle. Näitä on käsitelty tutkimuksen I osan luvussa 5.5 ja tämän toisen osan luvussa 10, jossa käsitellään puuttuvien tietojen vaikutusta elinkaariennusteisiin.

Tutkimuksen II osan aineisto on varsinkin 20–62-vuotiaiden työikäisten nettovaikutusten osalta sama kuin tutkimuksen I osassa. Tältä osin tutkimuksen I osaa voi pitääkin osittaisena tutkimuksen II osan aineistonkuvauksena. Myös olotila-aikojen mallinnuksessa (luku 3) ja palkkasummien mallinnuksessa (luku 5) käytetyt aineistot pohjautuvat samaan, tässä lyhyesti esiteltyyn, rekisteriaineistokokonaisuuteen. Jokaisen elinkaarivaikutuksen komponentin kohdalla niiden mallinnuksessa käytetty aineisto esitellään vielä tarkemmin niitä koskevissa tämän tutkimuksen luvuissa.

Sekä tutkimuksen I osassa että tässä II osassa kaikki tutkimusaineistoja ja tuloksia koskevat euromääräiset luvut on ilmaistu elinkaarivaikutuksia lukuun ottamatta vuoden 2011 euroissa kuten tutkimuksen I osassa. Elinkaarivaikutusten tulokset on esitetty vuoden 2015 euroissa. Muunnokset vuoden 2011 ja vuoden 2015 euroiksi on tehty Tilastokeskuksen rahanarvonkertoimella (SVT: *Kuluttajahintaindeksi*).

1.5 Miksi elinkaariennusteet?

Ennen varsinaista elinkaarivaikutusten ennusteiden mallinnusta on vielä syytä paneutua siihen, miksi päätelmiä julkisen talouden nettovaikutuksista ei voi tehdä pelkästään toteutuneiden nettovaikutusten perusteella. Vastaus tähän kysymykseen perustelee tarpeen tutkimuksen II osalle.

Olen jo oikeastaan vastannut tähän kysymykseen tutkimuksen I osan luvussa 2, jossa käsitellään menetelmiä tutkimuskirjallisuudessa. Tutkimuksen I osan tulokset julkisen talouden nettovaikutuksista on esitetty aikasarjoina vuosille 1995–2011 (ks. I osan luku 6.9). Näitä tuloksia ei voi kuitenkaan suoraan käyttää uskottavasti elinkaarivaikutusten mallina, sillä:

”Mikäli julkisen talouden tulo- ja menolaskelmia tehdään vain yhdelle vuodelle, sen perusteella tehtävät arviot maahanmuuton vaikutuksista julkiselle taloudelle ovat hyvin suhdanneriippuvaisia. Julkisen talouden tulo- ja menolaskelmat voidaan tällöin laatia useammalle vuodelle aikasarjana. Tällainen perättäisiin poikkileikkauksiin perustuva malli ei kuitenkaan ota huomioon esimerkiksi maahanmuuttajien iän, sukupuolen, koulutuksen, perhetaustan, maahantulovuoden (tai maassaoloajan) vaikutuksia julkisen talouden tuloihin ja menoihin.”. (Salminen 2015a, luku2).

Yllä lyhyesti esitetty tutkimusaineisto on itsessään kuitenkin paneeliaineisto (eikä perättäisiin poikkileikkauksiin perustuva aikasarja), sillä se koostuu aineistoon kuuluvia henkilöitä koskevista aikasarjoista. Paneeliaineisto mahdollistaa luoda sellaiset elinkaariennusteiden mallit, jotka ottavat huomioon edellisessä lainauksessa mainitut tekijät:

”Tutkimuskirjallisuudessa dynaamisiksi malleiksi kutsutaan nettonykyarvolaskelmia, joissa yksittäisen maahanmuuttajan yhteenlasketut (diskontatut) tulevat julkisen talouden tulot ja menot estimoidaan esimerkiksi maahanmuuttajan iän, sukupuolen, lähtömaan, maahantulovuoden ja maassaoloajan sekä muiden merkitsevien selittävien tekijöiden perusteella.[...]”. (Salminen 2015a, luku2).

Tällaisten elinkaariennusteiden mallien luominen, jossa huomioidaan esimerkiksi aikaan liittyvien muuttujien vaikutus julkisen talouden nettovaikutuksiin, on tavoitteena tässä tutkimuksen II osassa. Aikaan liittyviä muuttujia ovat ikä, maassaoloaika, maahanmuuttovuosi ja kalenterivuosi sekä näiden muuttujien

muunnokset.

1.5.1 Elinkaariennusteiden kaksi käyttötarkoitusta

Valmiilla julkisen talouden nettovaikutusten elinkaariennusteilla on kaksi käyttötarkoitusta. Ensiksi niiden avulla julkisen talouden nettovaikutukset voidaan esittää niin, että maahanmuuttajien maassaoloaika, sukupuoli, ikä, maahanmuuttovuosi ja muut tekijät otetaan samaan aikaan huomioon. Tämä mahdollistaa myös yksittäisten tekijöiden vaikutusten tarkastelun siten, että muiden tekijöiden vaikutus ”vakioidaan” pois. Voidaan esimerkiksi tarkastella, mikä on maassaoloajan vaikutus nettovaikutuksiin, kun muut tekijät on huomioitu.

Toinen elinkaariennusteiden käyttötarkoitus on soveltaa elinkaarivaikutusten ennustetuloksia uusien maahanmuuttajien optimaaliseen valintaan julkisen talouden kannalta. Optimaalisella valinnalla tarkoitetaan tässä yhteydessä uusien maahanmuuttajien valitsemista Suomeen muuttavaksi niin, että julkisen talouden hyödyn odotusarvo maksimoituu. Odotusarvo maksimoituu tässä tutkimuksessa silloin, kun uudet maahanmuuttajat valitaan niin, että uusien maahanmuuttajien yhteenlaskettu odotettavissa oleva julkisen talouden euromääräinen elinkaarivaikutusten summa maksimoituu. Elinkaariennusteiden soveltamista tällaisessa päätöksenteossa käsitellään tämän tutkimustekstin osassa III ja siinä erityisesti luvussa 12.

1.5.2 Elinkaariennusteet vain uusille maahanmuuttajille ja vastasyntyneille

Yllä esitetyssä tutkimuksen toisessa tutkimuskysymyksessä todetaan, että elinkaarivaikutukset on tarkoitus selvittää joko Suomessa jo olevien maahanmuuttajien osalta tai Suomeen vasta muuttavassa olevien uusien maahanmuuttajien osalta. Laadin kuitenkin julkisen talouden elinkaariennusteet vain uusille, Suomeen ensimmäistä kertaa muuttavissa oleville henkilöille. Tähän on kaksi syytä.

Ensiksi, elinkaariennusteiden mallit estimoidaan aineistosta, joka koskee jo maassa olevia maahanmuuttajia. Tästä syystä elinkaarivaikutusten ennusteissa uusien ja jo maassa olevien maahanmuuttajien välillä on vain se ero, että uusien maahanmuuttajien osalta ennusteissa huomioidaan se lisäepävarmuus (satunnaisvaihtelu), joka aiheutuu siitä, että ennusteet tehdään sellaisille henkilöille, jotka eivät sisälly mallien estimointiaineistoon. Jo maassa olevien maahanmuuttajien elinkaariennusteissa tätä henkilöiden välisistä havaitsemattomista eroista johtuvaa satunnaisvaihtelua ei tarvitse erikseen lisätä ennusteisiin, sillä se sisältyy jo käytettyjen nk. *hierarkkisten* mallien estimaatteihin. Toisin sanoen, mikäli ennusteet voidaan tehdä uusille maahanmuuttajille, ne voidaan tehdä samalla (hieman yksinkertaisemmin) jo maassa oleville maahanmuuttajille.

Toiseksi, tutkimuksen II osan motivaationa on soveltaa päätöksentekoteoriaa uusien maahanmuuttajien optimaaliseen valintaan Suomen julkisen talouden kannalta. Jotta tällainen sovellus voitaisiin tehdä, tätä varten on oltava käytössä uskottavat uusia maahanmuuttajia koskevat elinkaariennusteet.

Koska elinkaariennusteet laaditaan Suomeen vasta muuttavassa oleville uusille maahanmuuttajille, en-

nusteet on laadittava koskemaan uusia maahanmuuttajia siltä pohjalta, mitä uusista maahanmuuttajista tiedetään, kun he muuttavat Suomeen. Tästä syystä elinkaariennusteiden komponenttien malleissa selittävinä muuttujina ovat tiedot maahanmuuttajista siltä vuodelta, kun he muuttavat ensimmäisen kerran Suomeen. Aikaan liittyvien determinististen muuttujien, kuten iän, kalenterivuoden ja maassaoloajan suhteen, ei tarvitse kuitenkaan rajoittua pelkästään muuttovuoden tietoihin, koska näiden muuttujien osalta maassaolovuoden tiedoista saadaan laskettua niiden arvot deterministisesti kaikille tuleville vuosille.

1.5.3 Elinkaariennusteet vain Irakissa ja Somaliassa syntyneille sekä Suomessa syntyville kantaväestön henkilöille

Yllä esitetyssä tutkimuksen toisessa tutkimuskysymyksessä sanotaan, että elinkaarivaikutuksia tarkastellaan eritellysti sekä syntymävaltioiden ja syntymäalueiden mukaan. Laadin tässä tutkimuksen II osassa julkisen talouden elinkaariennusteet kuitenkin vain Irakissa ja Somaliassa syntyneille maahanmuuttajien ryhmälle sekä Suomessa syntyneelle kantaväestölle. Tämä tarkoittaa sitä, että en laadi elinkaariennusteita muulle kahdeksalle suurimmalle maahanmuuttajaryhmälle enkä myöskään syntymäalueiden mukaan eritellyille ryhmille. Tähän on kaksi syytä.

Ensiksi, uskottavien elinkaariennusteiden laatiminen perusteellisesti Irakissa ja Somaliassa syntyneille sekä Suomessa syntyneelle kantaväestölle on jo itsessään sekä työ- että sivumäärältään niin suuri tutkimus, että nopeuttaakseni tutkimuksen II osan valmistumista pitäydyn vain mainittujen syntymävaltiokohtaisten ennusteiden laatimisessa. Irakissa ja Somaliassa syntyneiden maahanmuuttajaryhmät on esitetty tutkimuksen I osan luvussa 6.1 sekä sen liitteessä 1.

Toiseksi, aluekohtaisissa ennustemalleissa olisi oman ennakkokäsitykseni mukaan huomioitava jotenkin se heterogeenisyys, jonka jokaiseen alueeseen sisältyvät erilaiset syntymävaltiot uskoakseni tuovat aineistoon. Esimerkiksi regressiomallien vakiokertoimien arvot saattavat vaihdella tällaisissa aluekohtaisissa malleissa syntymävaltioittain, ja tämä olisi jotenkin huomioitava mallinnuksessa. Tämä syntymävaltiokohtainen vaihtelu aluekohtaisissa ennusteissa toisi siis nähdäkseni yhden mallinnettava vaihtelun ”tason” lisää verrattuna pelkkiin syntymävaltiokohtaisiin ennusteisiin. Tämä lisäisi erityisesti mallikohtaista työmäärää.

1.6 Huomautus englanninkielisten termien käytöstä

Käytän tutkimuksen II osassa paljon sellaista erityisesti tilastotieteen ja kansantaloustieteen piiriin kuuluvaa terminologiaa, joka esitetään englanninkielisinä viittaamisssani lähteissä. Olen yleisesti ottaen kursivoinut termien suomenkieliset nimet erityisesti silloin, kun käytän niitä ensimmäisen kerran tekstissä. Koska haluan lukijan tietävän myös näiden termien kirjoitusasun alkuperäisisissä englanninkielisissä lähteissä, olen useimmiten lisännyt käyttämieni termien perään niiden englanninkielisen nimen suluissa. Näin lukija voi helpommin löytää terminologiaa koskevaa kirjallisuutta tai muita lähteitä. Lisäksi erityisesti monet käyttämäni *bayesilaisen tilastollisen päättelyn* (*engl. Bayesian inference*) piiriin kuuluvat termit ovat sellaisia, että niille ei ole ole tietääkseni vielä olemassa omaa vakiintunutta suomennosta.

Luku 2

Elinkaarivaikutusten rakenne ja ennustaminen uusille maahanmuuttajille

2.1 Elinkaarivaikutusten rakenne

2.1.1 Elinkaarivaikutusten komponentit

Elinkaarivaikutukset muodostuvat tässä tutkimuksessa seitsemästä eri komponentista. Yhden henkilön elinkaarivaikutusten kohdalla komponentit ovat:

1. Olotila ikävuosittain 0–100-vuotiaana: joko Suomessa, ulkomailla vai kuollut.
2. Julkisen talouden nettovaikutukset alle 15-vuotiaana.
3. Julkisen talouden nettovaikutukset nuorena ja työikäisenä (15–62-vuotiaana).
4. Palkkasummat 18–62-vuotiaana.
5. Muut kuin eläkkeisiin liittyvät julkisen talouden nettovaikutukset eläkeikäisenä (yli 62-vuotiaana).
6. ”Muut parametrit” (π), joita käytetään muodostamaan eläkkeisiin liittyvät vaikutukset muista komponenteista. Lisäksi käytetään elinkaarivaikutusten diskonttauksessa nykyarvoon.
7. Henkilön lasten elinkaarivaikutukset yhteensä.

1. **Olotila-ajat**. Henkilö on koko elinkaarensa yli yhdessä kolmesta olotilasta (ikävuosittain): joko Suomessa, ulkomailla tai kuollut. Maksimiseuranta-aika elinkaariennusteissa ulottuu siihen vuoteen, jolloin henkilö täyttää/täyttäisi 101 vuotta. Maahanmuuttajilla elinkaari alkaa siitä (ikä)vuodesta, jolloin he muuttavat ensimmäisen kerran Suomeen. Suomessa syntyneillä syntymästä. Olotilojen mallinnusta ja ennustamista käsittelen luvussa 3.

2. **Nettovaikutukset alle 15-vuotiaana**. Henkilön julkisen talouden nettovaikutukset alle 15-vuotiaana muodostuvat pääasiassa peruskoulusta ja päivähoidosta aiheutuvista menoista, sillä lapsiin liittyvät tulonsiirrot eli käytännössä perhe-etuudet menevät lapsen huoltajan julkisen talouden menojen alle, sillä huoltaja saa nämä tulonsiirrot. Alle 15-vuotiaiden osalta menot ovat melko vakioarvoisia, ja käsittelen niitä lyhyesti luvussa 7.1.

3. **Nuorten ja työikäisten nettovaikutukset**. Julkisen talouden nettovaikutukset nuorena ja työikäisenä eli 15–62-vuotiaana muodostavat tutkimuksessa suurimmat erot henkilöiden välille elinkaarivaikutuksissa. Tämä johtuu siitä, että tutkimuksessa sekä lapsuuden (alle 15-vuotiaat) että eläkeiän (yli 62-vuotiaat) nettovaikutukset eläkevaikutuksia lukuun ottamatta ovat tutkimuksessa odotusarvoltaan suunnilleen yhtä suuria kaikkien henkilöiden osalta. Tutkimuksen I osa keskittyi suurimmaksi osaksi käsittelemään työikäisten eli 20–62-vuotiaiden toteutuneita julkisen talouden nettovaikutuksia. Tutkimuksen I osa osoitti, että erot keskimääräisissä vuosittaisissa työiän nettovaikutuksissa ovat ”suuria” eri maahanmuuttajaryhmien ja Suomessa syntyneen kantaväestön välillä.

   Työikäisistä 20–62-vuotiaista poiketen nettovaikutusten odotusarvot ovat melko samansuuruisia 15–19-vuotiailla nuorilla eri ryhmissä. Nuorten ja työikäisten nettovaikutusten mallintamista ja ennustamista käsittelen luvussa 4.

4. **Palkkasummat**. Henkilön vuotuinen palkkasumma koostuu henkilön kaikkien niiden palkkojen summasta, joista maksetaan lakisääteisiä eläkevakuutusmaksuja. Palkkasummien mallintamisen ja ennustamisen varsinainen tarkoitus on ennustaa maksettavien lakisääteisten eläkevakuutusmaksujen määrää. Eläkevakuutusmaksujen ennusteet saadaan palkkasummaennusteista ”muiden malliparametrien” avulla. Muut malliparametrit tarkoittavat tässä tapauksessa palkkojen reaalista kehitystä sekä eläkevakuutusmaksuprosentteja. Lakisääteisten eläkevakuutusmaksujen ennusteita käytetään elinkaarivaikutuksissa *eläkevaikutusten* ennustamiseen.

   Eläkevaikutukset koostuvat kaikista lakisääteisiin työeläkkeisiin liittyvistä julkisen talouden vaikutuksista lukuun ottamatta henkilöiden (palkansaajien) itse työikäisinä maksamia lakisääteisiä työeläkemaksuja, jotka sisältyvät jo yllä esitettyyn nuorten ja työikäisten nettovaikutuksiin. Eläkevaikutuksiin kuuluvat näin 18–62-vuotiaiden osalta työnantajien maksamat lakisääteiset työeläkemaksut sekä kaikki eläkkeistä suoraan johtuvat julkisen talouden nettovaikutukset eläkeiässä eli yli 62-vuotiaana. Näitä vaikutuksia ovat esimerkiksi maksetut työeläkkeet, maksetut kansaneläkkeet ja eläkkeistä maksetut tuloverot. Oletan VATT:n tutkimuksen tavoin, että kaikki jäävät eläkkeelle 63-vuotiaina (VATT 2014b).

   Käsittelen palkkasummien havaintoaineiston muodostamista sekä näiden palkkasummien mallinnusta ja ennustamista luvussa 5. Luvussa 6 puolestaan esitän, kuinka palkkasummaennusteista lasketaan eläkkeisiin liittyvät vaikutukset alla lyhyesti esitettävien ”muiden malliparametrien” avulla.

5. **Muut kuin eläkkeisiin liittyvät nettovaikutukset eläkeikäisenä**. Muut nettovaikutukset eläkeikäisenä (yli 62-vuotiaana) koostuvat niistä julkisen talouden nettovaikutuksista, jotka eivät sisälly eläkevaikutuksiin. Näitä ovat erityisesti sosiaali- ja terveyshuollon menot. Käsittelen näitä julkisen talouden menoja luvussa 7.2.

6. **Muut parametrit (π)**. Kaikki elinkaarivaikutusten yllä esitetyt komponentit riippuvat tietyistä (satunnaisista) parametreista, joihin viittaan tässä luvussa alempana lyhyesti sekä tarkemmin jokaisen

komponentin mallinnuksen yhteydessä. Näiden komponenttien parametrien lisäksi elinkaarivaikutukset muodostuvat ”muista parametreista”, joiden lista muodostaa vektorin $\boldsymbol{\pi}$.

Muihin malliparametreihin sisältyvät parametrit, joiden avulla eläkemaksuista lasketaan eläkevaikutukset. Näiden lisäksi muihin malliparametreihin sisältyvät ennustettavien nettovaikutusten diskonttauskertoimet silloin, kun ennustettavat nettovaikutukset halutaan ilmaista nykyarvossa (ks. luku 11).

7. **Lasten elinkaarivaikutukset yhteensä**. Lasten elinkaarivaikutukset ovat erillinen aihe edellisiin komponentteihin verrattuna, sillä ne käsittelivät henkilön omia elinkaarivaikutuksia. Lasten elinkaarivaikutukset puolestaan koostuvat henkilön kaikkien lasten yhteenlasketuista elinkaarivaikutuksista. Koska lasten elinkaarivaikutukset muodostavat erillisen osan elinkaarivaikutuksista ja niiden ennusteista, ilmaisen elinkaarivaikutusten ennusteita koskevat tulokset henkilöille sekä ilman henkilön lasten elinkaarivaikutuksia sekä niiden kanssa.

   Lasten elinkaarivaikutusten osalta en mallinna ja ennusta erikseen eläkevaikutuksia, joten kaikki erot lasten elinkaarivaikutuksissa tulevat käytännössä työikää koskevista ennusteista. Koska usean maahanmuuttajaryhmän kohdalla havaintoaineistoon sisältyvät vanhimmat Suomeen alle kouluikäisinä muuttaneet lapset ovat vasta vähän yli 20-vuotiaita, lasten elinkaarivaikutukset perustuvat niihin tietoihin, joita maahanmuuttajien lapsista on 22-vuotiaina tutkimusaineistossa. Käsittelen lasten elinkaarivaikutusten mallinnusta ja ennustamista luvussa 8.

2.1.2 Elinkaarivaikutusten funktio

Määritetään nyt sellainen funktio (surjektio) f, joka on kuvaus yllä esitetyistä komponenteista elinkaarivaikutuksiin:

$$\begin{aligned} f : &\{\textbf{olotila-ajat}\} \times \{\text{nettovaikutukset: alle 15-v.}\} \times \{\text{nettovaikutukset: nuoret ja työikäiset}\} \times \\ &\{\text{palkkasummat: 18–62-v.}\} \times \{\text{muut nettovaikutukset: yli 62-v.}\} \times \{\boldsymbol{\pi}\} \times \\ &\{\text{elinkaarivaikutukset: lapset}\} \to \{\text{elinkaarivaikutus}\}. \end{aligned} \tag{2.1}$$

Funktion määrittelyjoukko muodostuu elinkaarivaikutuksen komponenttien *mahdollisten arvojen* joukkojen karteesisesta tulosta. Esimerkiksi joukko $\{\textbf{olotila-ajat}\}$ koostuu kaikista mahdollisista olotila-ajoista eli kaikista mahdollisista suomessaoloajoista, ulkomaillaoloajoista ja kuolleena olemisen ajoista. Funktion arvojoukko $\{\text{elinkaarivaikutus}\}$ muodostuu puolestaan kaikista elinkaarivaikutusten mahdollisista arvoista.

Elinkaarivaikutusten komponenttien mahdollisten arvojen joukot ovat yksi- tai moniulotteisia reaalilukujen joukkoja tai niiden osajoukkoja. Elinkaarivaikutusten mahdollisten arvojen joukko on reaalilukujen joukko. Näiden joukkojen avulla mainittu funktio f voidaan lausua myös muodossa:

$$f : \mathbb{R}^3_{\geq 0} \times \mathbb{R}_{<0} \times \mathbb{R} \times \mathbb{R}_{\geq 0} \times \mathbb{R}_{\leq 0} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}.$$

Ensimmäinen joukko on 3-ulotteinen, sillä olotiloja on kolme mahdollista. Käytännössä olotila-aikojen joukot ovat suljettuja ja rajoitettuja tässä tutkimuksessa. Toinen joukko koostuu vain negatiivisista reaaliluvuista, sillä tässä tutkimuksessa alle 15-vuotiaiden nettovaikutukset ovat aina negatiivisia. Työikäisten nettovaikutukset ovat reaalilukujen joukko. Neljäs joukko koostuu ei-negatiivisista reaaliluvuista, sillä se on mahdollisten palkkasummien joukko, ja ainakin tässä tutkimuksessa palkkasummat eivät voi saada negatiivisia arvoja. Viides joukko koostuu ei-positiivista reaaliluvuista, sillä tässä tutkimuksessa yli 62-vuotiaiden muut kuin eläkevaikutuksista johtuvat nettovaikutukset ovat aina joko nollia tai negatiivisia, sillä ne muodostuvat vain julkisen talouden menoista. ”Muiden malliparametrien” vektorin dimensio on d eli se on näiden parametrien lukumäärä. Nuorten ja työikäisten mahdollisten nettovaikutusten, lasten kaikkien mahdollisten nettovaikutusten sekä kaikkien mahdollisten elinkaarivaikutusten joukot koostuvat reaalilukujen joukoista.

Käytännössä funktio määrittää sen, miten elinkaarivaikutusten komponenttien arvoista lasketaan elinkaarivaikutus. Käsittelen asiaa tarkemmin luvussa 9.

2.2 Elinkaarivaikutusten ennustaminen uusille maahanmuuttajille

2.2.1 Elinkaariennusteiden komponentit

Kutsun ennusteita elinkaarivaikutusten tulevista arvoista elinkaariennusteiksi. Kuten olen edellä esittänyt, tämän tutkimuksen II osan tavoitteena on laatia elinkaariennusteet sekä uusille maahanmuuttajille että Suomessa syntyville (kantaväestöön kuuluville). Elinkaariennusteet ovat tässä tutkimuksessa *satunnaismuuttujia*, joilla on omat (ehdolliset) *todennäköisyysjakaumansa*. Tutkimuksen II osan tavoitteena on selvittää nämä todennäköisyysjakaumat. ”Ehdollisuus” todennäköisyysjakaumiin tulee siitä, että jokaisen uuden maahanmuuttajan elinkaariennusteissa otetaan huomioon, mitä jo Suomessa olleista maahanmuuttajista tiedetään sekä, mitä uudesta maahanmuuttajasta tiedetään, kun hän muuttaa ensimmäisen kerran Suomeen.

Ennustetut elinkaarivaikutukset eli elinkaariennusteet muodostuvat samoista komponenteista kuin elinkaarivaikutukset edellä. Jotta elinkaariennuste voidaan muodostaa, näiden komponenttien arvot on ennustettava. Kuten elinkaariennusteet, myös nämä komponentit ovat satunnaismuuttujia, joilla on omat (ehdolliset) todennäköisyysjakaumansa. Koska elinkaariennusteet muodostetaan yllä esitetyn funktion mukaisesti komponenteista ja muista malliparametreista, elinkaariennusteen todennäköisyysjakauma on verrannollinen näiden ennustekomponenttien ja muiden malliparametrien yhteistodennäköisyysjakaumaan. Käytetään uuden maahanmuuttajan tai Suomessa syntyvän henkilön elinkaariennusteen ennustekomponenttien vektorista merkintää: $\boldsymbol{y}_{uusi}$. Tämä vektori koostuu kuudesta ennustekomponentista, joten siihen ei sisälly muita ennusteparametreja ($\boldsymbol{\pi}$), jonka esittelin edellä myös elinkaarivaikutusten yhtenä komponenttina. Kun viittaan tästä lähtien elinkaariennusteiden komponentteihin $\boldsymbol{y}_{uusi}$, tarkoitan muita ennustekomponentteja muita ennusteparametreja ($\boldsymbol{\pi}$) lukuun ottamatta. Komponenttien ja muiden malliparametrien ennusteiden muodostamista käsittelen tämän tutkimuksen luvuissa, kuten olen edellä esittänyt. Esimerkiksi olotila-aikojen ennustamista käsittelen luvussa 3.

2.2.2 Elinkaariennusteiden komponenttien parametrit

Jokainen ennustekomponentti $\boldsymbol{y}_{uusi}$ riippuu niitä koskevista omista malliparametreista. Esimerkiksi työikäisten nettovaikutusennusteita varten laadin regressiomallin, joka sisältää useita malliparametreja. Käytetään nyt kaikkien ennustekomponenttien kaikkien malliparametrien yhteisestä vektorista merkintää: $\boldsymbol{\theta}$.

Tämän vektorin seitsemän komponenttia ovat ennustekomponenttikohtaisia parametrivektoreita: $\boldsymbol{\theta} = \begin{bmatrix}\boldsymbol{\theta}_{\text{olotila-ajat}} & \cdots & \boldsymbol{\theta}_{\text{muut nettovaikutukset: yli 62-v.}}\end{bmatrix}$. Tässä esimerkiksi parametrivektori $\boldsymbol{\theta}_{\text{olotila-ajat}}$ koostuu kaikista niistä parametreista, joita käytetään olotila-aikojen mallintamiseen ja ennustamiseen. Parametrivektorin $\boldsymbol{\theta}$ parametrien arvoihin sisältyy epävarmuuksia. Kuten elinkaariennusteiden komponentit, ne ovat myös satunnaismuuttujia, ja niillä on omat todennäköisyysjakaumansa. Nämä parametrien todennäköisyysjakaumat estimoidaan pääasiassa Suomessa jo olevia ”vanhoja” maahanmuuttajia koskevaan havaintoaineistoon perustuen.

Huomautan, että käytän tässä tekstissä useimmiten samoja merkintöjä sekä *satunnaismuuttujista* että niiden *toteutuneista arvoista*. Mikäli käyttäisin formaalimpaa merkintätapaa, missä nämä on erotettu toisistaan, elinkaariennusteiden komponenttien vektorin satunnaismuuttaja kirjoitettaisiin isolla kirjaimella, $\boldsymbol{Y}_{uusi}$, ja vektorin toteutuneita arvoja puolestaan – nyt kummassakin tapauksessa käytetyllä – pienellä kirjaimella: $\boldsymbol{y}_{uusi}$. Malliparametrien vektorin kohdalla vastaavat merkinnät satunnaismuuttujan ja toteutuneiden arvojen osalta olisivat: $\boldsymbol{\Theta}$ ja $\boldsymbol{\theta}$. Tässäkin tapauksessa käytän jälkimmäistä merkintää sekä satunnaismuuttujasta että sen toteutuneista arvoista. (ks. tarkemmin merkintätavoista esimerkiksi Koistinen (2012), luku 1).

2.2.3 Elinkaariennusteissa käytetty havaintoaineisto

Olen edellä esittänyt (luku 1.4), mistä (rekisteri)lähteistä tutkimusaineisto koostuu. Havaintoaineisto sisältää tietoja koskien sekä jo Suomessa olleita ”vanhoja” maahanmuuttajia että Suomeen vasta ensimmäisen kerran muuttamassa olevia ”uusia” maahanmuuttajia (sama koskien myös Suomessa aiemmin syntyneitä henkilöitä sekä vastasyntyneitä). Elinkaariennusteet laaditaan uusille maahanmuuttajille.

Käytetään nyt niistä elinkaarivaikutusten komponenttien vektorista, jotka on havaittu ”vanhoilta” maahanmuuttajilta, merkintää: $\boldsymbol{y}_{vanha}$. Tämä noudattaa edellä esitettyä merkintätapaa, jossa ”uutta” maahanmuuttajaa koskevista ennustettavista elinkaarivaikutusten komponenttien vektorista käytettiin merkintää: $\boldsymbol{y}_{uusi}$. Käytännössä $\boldsymbol{y}_{vanha}$:n komponentit toimivat selitettävinä muuttujina elinkaarivaikutusten komponenttien regressiomalleissa tässä tutkimuksessa.

Käytän elinkaarivaikutusten komponenttien regressiomalleissa selittävinä tekijöinä sellaisia havaintoja ”vanhoilta” maahanmuuttajilta, jotka mahdollisesti ennustavat elinkaarivaikutusten komponenttien arvoja. Käytetään näiden havaittujen selittävien muuttujien vektorista merkintää: $\boldsymbol{x}_{vanha}$. Käytetään samalla tavoin ”uusia” maahanmuuttajia koskevien selittävien tekijöiden vektorista merkintää: $\boldsymbol{x}_{uusi}$. Nämä ovat sellaisia tietoja, joita uusista maahanmuuttajista tiedetään, kun he muuttavat ensimmäisen kerran Suomeen ja joita voidaan käyttää uuden maahanmuuttajan elinkaarivaikutusten komponenttien $\boldsymbol{y}_{uusi}$

ennustamiseen. Elinkaariennusteiden todennäköisyysjakaumat ovat toisin sanoen ehdollisia todennäköisyysjakaumia, koska ne ehdollistetaan $\boldsymbol{x}_{uusi}$:lla (ja muulla havaintoaineistolla).

Kirjoitetaan nyt kaiken havaintoaineiston sisältävä vektori muodossa: $\boldsymbol{h} = \begin{bmatrix} \boldsymbol{x}_{uusi} & \boldsymbol{y}_{vanha} & \boldsymbol{x}_{vanha} \end{bmatrix}$. Kaikki havaintoaineisto koostuu siis ”vanhoja” maahanmuuttajia koskevista selittävistä ja selitettävistä muuttujista sekä ”uusia” maahanmuuttajia koskevista selittävistä muuttujista. ”Vanhoja” maahanmuuttajia koskevat selitettävät muuttujat, $\boldsymbol{y}_{vanha}$, ovat elinkaarivaikutusten komponenttien arvoja, jotka on havaittu Suomessa jo asuneilta maahanmuuttajilta.

Kaikkeen havaintoaineistoon liittyy käytännössä aina jonkin verran mittausvirheen aiheuttamaa epävarmuutta. Koska tässä tutkimuksessa kyse on edellä esitetyistä suomalaisista rekisteriaineistoista, niin uskallan olettaa, että mittausvirhe on rekisteriaineistojen osalta tässä yhteydessä sen verran pientä, että sen huomiotta jättäminen ei vaikuta juurikaan elinkaarivaikutusten ennusteiden jakaumiin. Toisaalta, kuten luvussa 10 esitän, tutkimuksen havaintoaineistoon ei sisälly kaikkia tutkimuksen määritelmän mukaisia julkisen talouden nettovaikutuksia. Koska havaintoaineiston on tarkoitus mitata julkisen talouden nettovaikutuksia, havaintoaineistosta puuttuvat julkisen talouden nettovaikutukset voisi esittää systemaattisena virheenä havaintoaineistossa.

2.2.4 Elinkaariennusteiden komponenttien, komponenttien parametrien, muiden ennusteparametrien sekä havaintoaineiston yhteistodennäköisyysjakauma

Olen yllä esittänyt uusien maahanmuuttajien julkisen talouden elinkaarivaikutusten ennusteiden komponentit, $\boldsymbol{y}_{uusi}$, joista elinkaariennusteet muodostetaan yhdessä ”muiden malliparametrien” vektorin $\boldsymbol{\pi}$ kanssa edellä esitetyn funktion mukaisesti. Olen myös esittänyt niiden malliparametrien vektorin $\boldsymbol{\theta}$, joista elinkaariennusteiden komponenttien ennusteet, $\boldsymbol{y}_{uusi}$, riippuvat. Samoin olen esittänyt havaintoaineiston vektorin $\boldsymbol{h}$, joka koostuu ”vanhoja” ja ”uusia” maahanmuuttajia koskevista havaituista tiedoista, kuten edellä esitin. Sekä elinkaariennusteiden komponentit, muut malliparametrit, elinkaariennusteiden komponenttien malliparametrit että havaintoaineisto ovat satunnaismuuttujia. Havaintoaineisto on satunnaismuuttuja silloin, kun sillä viitataan kaikkiin mahdollisiin havaintoaineistoihin, eikä jo havaittuihin eli toteutuneisiin havaintoaineistoihin. Toteutuneet havaintoaineistot ovat siis poimintoja kaikkien mahdollisten havaintoaineistojen joukosta.

Määritetään nyt näille satunnaismuuttujille yhteistodennäköisyysjakauma, josta käytetään tässä tutkimuksessa merkintätapaa: $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h})$. Tämä käytetty merkintätapa yhteistodennäköisyysjakaumasta on puutteellinen, mutta se on kuitenkin yleisesti käytössä oleva, sillä se tiivistää todennäköisyysjakaumista käytettyjä merkintöjä. Täydellisempi merkintätapa mainittujen neljän satunnaismuuttujan yhteisestä todennäköisyystiheysfunktiosta olisi: $f_{\boldsymbol{\Pi},\boldsymbol{Y}_{uusi},\boldsymbol{\Theta},\boldsymbol{H}}(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h})$. Tässä jälkimmäisessä merkintätavassa todennäköisyystiheysfunktioon on merkitty erikseen ne satunnaismuuttujat, joita yhteistodennäköisyystiheysfunktio koskee ja funktion argumenteiksi puolestaan näiden satunnaismuuttujien (mahdolliset) toteutuvat arvot. Käytän tutkimuksessa todennäköisyysjakaumista suurimmaksi osaksi lyhyempää mutta puutteellisempaa merkintätapaa: $p(\cdot)$. (Koistinen (2012), s. 3).

Sovelletaan nyt mainittujen satunnaismuuttujien yhteistodennäköisyysjakaumaan $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h})$ *Baye-*

sin kaavaa (teoreemaa):

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h}) = \frac{p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h})}{p(\boldsymbol{h})}. \tag{2.2}$$

Yhtälön vasemmalla on $\boldsymbol{\pi}$:n, $\boldsymbol{y}_{uusi}$:n ja $\boldsymbol{\theta}$:n ehdollinen yhteistodennäköisyysjakauma, missä ehdollistajana on havaintoaineisto $\boldsymbol{h}$. Yhtälön oikealla puolella olevan osamäärän osoittajassa on yllä esitettyjen satunnaismuuttujien yhteistodennäköisyysjakauma, ja osamäärän nimittäjässä on puolestaan havaintoaineiston ehdollistamaton todennäköisyysjakauma eli reunajakauma.

Koska havaitun aineiston reunajakauma $p(\boldsymbol{h})$ saa aina arvokseen vain jonkin vakioarvoisen luvun, $\boldsymbol{\pi}$:n, $\boldsymbol{y}_{uusi}$:n ja $\boldsymbol{\theta}$:n ehdollinen yhteistodennäköisyysjakauma on näin ollen verrannollinen kaikkien satunnaismuuttujien yhteistodennäköisyysjakaumaan:

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h}) \propto p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h}).$$

Nyt satunnaismuuttujien havaintoaineistolla ehdollista yhteistodennäköisyysjakaumaa voidaan kehittää edelleen todennäköisyyslaskennan perussääntöjen avulla:

$$\begin{aligned}
&p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h}) \\
&= p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}|\boldsymbol{h}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{h}) \\
&= p(\boldsymbol{\pi})p(\boldsymbol{y}_{uusi}|\boldsymbol{h}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{h}) \\
&\propto p(\boldsymbol{\pi})p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta}).
\end{aligned} \tag{2.3}$$

Toinen rivi seuraa tässä ensimmäisestä rivistä Bayesin kaavaa soveltamalla. Kolmas rivi seuraa toisesta rivistä siitä syystä, että muut malliparametrit ($\boldsymbol{\pi}$) ovat tilastollisesti riippumattomia elinkaarivaikutusten komponenttien ennusteista $\boldsymbol{y}_{uusi}$.

Viimeinen eli neljäs rivi yllä on verrannollinen kolmanteen riviin kahdesta syystä:

Ensiksi, ”uusien” maahanmuuttajien elinkaarivaikutusten ennustekomponentit ovat ehdollisesti riippumattomia ”vanhoja” maahanmuuttajia koskevasta havaintoaineistosta ($\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}$). Ehdollistajana tässä ovat elinkaariennusteiden komponenttien malliparametrit $\boldsymbol{\theta}$. Käytännössä tämä tarkoittaa sitä, että ”vanhat” maahanmuuttajat vaikuttavat ”uusien” maahanmuuttajien elinkaariennusteiden todennäköisyysjakaumiin vain näiden malliparametrien $\boldsymbol{\theta}$ kautta. Uusien maahanmuuttajien elinkaariennusteiden komponentit riippuvat lisäksi havaintoaineistojen siitä osasta, joka koskee näitä ”uusia” maahanmuuttajia ($\boldsymbol{x}_{uusi}$).

Toiseksi, malliparametrien ehdollinen jakauma voidaan lausua Bayesin kaavan avulla:
$p(\boldsymbol{\theta}|\boldsymbol{h}) = p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \propto p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

Tässä malliparametrien ehdollinen jakauma $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$ on malliparametrien nk. *posteriorijakauma*. Havaintoaineiston ehdollista jakaumaa, $p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})$, nimitetään puolestaan *otosjakaumaksi* (*engl. sampling distribution*). Koska selitettävät muuttujat $\boldsymbol{y}_{vanha}$ ovat ehdollisia myös selittävillä muuttujilla $\boldsymbol{x}_{vanha}$ – ja koska selittävien muuttujien arvot oletetaan aina tunnetuiksi – otosjakauman voi kirjoittaa selvyyden vuoksi myös muodossa: $p(\boldsymbol{y}_{vanha}|\boldsymbol{\theta}, \boldsymbol{x}_{vanha})$. Malliparametrien nk. *priorijakauma* $p(\boldsymbol{\theta})$ puolestaan kertoo, mitä malliparametreista tiedetään ennen, kuin niiden jakauma ”päivitetään”

havaintoaineiston tarjoamalla informaatiolla. Yleisesti ottaen käytän tässä tutkimuksessa ”epäinformatiivisia” priorijakaumia, mikä tarkoittaa sitä, että malliparametrien jakaumat määräytyvät (lähes) täysin havaintoaineiston perusteella.

Elinkaariennusteiden komponenttien mallinnuksessa on tässä tutkimuksessa kyse siitä, että pyritään komponenteittain selvittämään malliparametrien posteriorijakaumat. Esimerkiksi olotila-aikamallinnuksen yhteydessä halutaan selvittää olotila-aikoja koskevien parametrien posteriorijakauma: $p(\boldsymbol{\theta}_{\text{olotila-ajat}}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$. Havaintoaineistona tässä olotila-aikoja koskevien parametrien jakaumien selvityksessä käytetään Suomessa jo olleita ”vanhoja” maahanmuuttajia koskevia olotila-aikatietoja ($\boldsymbol{y}_{vanha}$) sekä tietoja, joiden avulla olotila-aikoja voisi ennustaa ($\boldsymbol{x}_{vanha}$).

Elinkaariennusteiden komponenttien (ehdollinen) riippumattomuus

Elinkaarikaariennusteissa oletetaan, että elinkaariennusteiden komponenttien todennäköisyysjakaumat ovat toisistaan (ehdollisesti) riippumattomia, kun niiden malliparametrit huomioidaan. Tämä oletus on epärealistinen ainakin palkkasummien ehdollisen todennäköisyysjakauman ja nuorten ja työikäisten nettovaikutusten todennäköisyysjakauman osalta. Tämä johtuu siitä, että palkkasummat ja työiän nettovaikutukset ovat oletettavasti tilastollistesti riippuvaisia toisistaan, vaikka niitä koskevat malliparametrit huomioitaisiin. Elinkaariennusteissa käytetään kuitenkin riippumattomuusoletusta, sillä tämä tekee elinkaariennusteiden komponenttien ehdollisten todennäköisyysjakaumien mallinnuksen ja ennustamisen helpommaksi.

2.2.5 Uusien maahanmuuttajien elinkaarivaikutusten ennustejakaumat

Yllä saatiin parametrien yhteistä posteriorijakaumaa koskevaksi tulokseksi: $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h})$ $\propto p(\boldsymbol{\pi})p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Tämän tuloksen avulla ”uudelle” maahanmuuttajalle voidaan määrittää ehdollinen elinkaarivaikutusten komponenttien ennusteiden ja muiden malliparametrien todennäköisyysjakauma, joka riippuu vain havaintoaineistosta: $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}|\boldsymbol{h})$. Elinkaarivaikutusten ennusteet saadaan laskettua elinkaarivaikutusten komponenttien ennusteiden ja muiden malliparametrien ennustejakaumasta edellä esitetyn elinkaarivaikutusten funktion avulla.

Tämä ennustejakauma saadaan, kun parametrien havaintoaineistolla ehdollisesta yhteistodennäköisyysjakaumasta, $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h})$, ”integroidaan pois” elinkaarivaikutusten komponenttien malliparametrit $\boldsymbol{\theta}$.

Uuden maahanmuuttajan elinkaarivaikutuksen komponenttien ja muiden malliparametrien posterioriennustejakauma:

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}|\boldsymbol{h}) \propto p(\boldsymbol{\pi}) \int_{\boldsymbol{\theta}} p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.4}$$

Tätä jakaumaa voidaan kutsua elinkaarivaikutusten komponenttien ja muiden malliparametrien *posterioriennustejakaumaksi* (*engl. posterior predictive distribution*) (Gelman et al. 2014, 7). Nimitys seuraa siitä,

että jakauma on sekä posteriorijakama että ennustejakauma. Jakauma on posteriorijakauma, koska se on ehdollinen toteutuneella havaintoaineistolla $\boldsymbol{h}$. Ennustejakauma se on puolestaan, koska siinä ennustetaan tulevia elinkaarivaikutusten komponenttien arvoja. Tarvittaessa myös muiden malliparametrien $\boldsymbol{\pi}$ arvot ennustetaan. Tässä tutkimuksessa on tavoitteena selvittää uusien maahanmuuttajien elinkaarivaikutusten ehdolliset posterioriennustejakaumat.

Käytännössä tämä tapahtuu tämän tutkimusraportin II osassa niin, että ensiksi elinkaarivaikutisten komponenttien malliparametrien posteriorijakaumat, $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, estimoidaan havaintoaineistosta. Estimointi koskee myös muiden malliparametrien jakaumien selvittämistä silloin, kun ne käsitetään satunnaisiksi. Tämän jälkeen malliparametrien $\boldsymbol{\theta}$ ”pois” integrointi tapahtuu käytännössä poimimalla simuloituja arvoja niiden posteriorijakaumasta, $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, ”tarpeeksi” monta kertaa, sillä lausekkeen 2.4 integraalilla $\boldsymbol{\theta}$:n suhteen ei käytännössä ole nk. ”analyyttista” ratkaisua.

Koska elinkaarivaikutusten komponenttien ennusteiden ja muiden malliparametrien posteriorijakaumalla, $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}|\boldsymbol{h})$, ei ole myöskään analyyttista ratkaisua, tämä jakauma myös simuloidaan poimimalla siitä otoksia ”tarpeeksi” monta kertaa. Käytännössä tämä tapahtuu jokaisen elinkaarivaikutusten komponentin kohdalla niin, että komponenttiin liittyvien malliparametrien $\boldsymbol{\theta}$ jokaisen simulaatiokerran jälkeen simuloidut malliparametrien $\boldsymbol{\theta}$ arvot sijoitetaan kyseessä olevan komponentin ennustejakaumaan, $p(y_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})$. Tästä jakaumasta poimitaan puolestaan simuloitu y_{uusi}:n arvo. Eli aluksi poimitaan $\boldsymbol{\theta}$:n simuloitu arvo ja sen avulla saadaan poimittua y_{uusi}:lle simuloitu arvo sen ehdollisesta jakaumasta. Mikäli muut malliparametrit $p(\boldsymbol{\pi})$ ovat satunnaisia, myös niiden ennustejakaumat simuloidaan poimimalla niille riittävä määrä arvoja niiden ennustejakaumista.

Elinkaarivaikutusten ennustejakauma saadaan näin karttuneista elinkaarivaikutusten komponenttien ja muiden malliparametrien ennusteista niin, että niiden simuloidut arvot sijoitetaan elinkaarivaikutusten edellä esitettyyn funktioon.

2.3 Elinkaarivaikutusten ennusteiden ehdolliset todennäköisyysjakaumat kuvaavat ennusteisiin sisältyviä epävarmuuksia

2.3.1 Elinkaariennusteiden ehdollinen todennäköisyysjakauma ja ennusteiden kokonaisepävarmuus

Edellisessä alaluvussa määritettiin uuden muuttajan elinkaariennusteen ehdollinen todennäköisyysjakauma, joka on nk. posterioriennustejakauma. Tässä alaluvussa käsittelen sitä, miten uuden maahanmuuttajan elinkaarivaikutusten posterioriennustejakaumaan sisältyy ennusteiden kokonaisepävarmuus.

Esitän aluksi lyhyen määritelmän siitä, miten todennäköisyys käsitetään (tässä tutkimuksessa) kuvauksena johonkin asiaan liittyvästä epävarmuudesta. Koska elinkaariennusteiden laadinta tehdään nk. bayesilaisen tilastollisen päättelyn piirissä, käyn läpi lyhyesti ne syyt, miksi tämä ei tee elinkaariennusteista erityisen ”subjektiivisia” – ainakaan verrattuna ei-bayesilaiseen päättelyyn perustuvissa ennusteissa. Tämän jälkeen esitän, miten posterioriennustejakaumaan sisältyy elinkaariennusteiden kokonaisepävarmuus. Koska

informaatioteoriassa satunnaismuuttujaan liittyvä epävarmuus ilmaistaan *entropian* avulla, esitän myös elinkaariennusteiden kokonaisentropian (Cover and Thomas 2006, 13).

2.3.1.1 Todennäköisyyden määritelmästä

Todennäköisyyden – tai *todennäköisyysavaruuden* – matemaattinen malli on Kolmogorovin (1933) esittämä. Kolmogorovin esittämässä mallissa esitetään todennäköisyyden kolme aksioomaa, jotka toimivat yleisenä perustana sille, että erilaisia satunnaisilmiöitä voidaan tutkia matemaattisesti (Tuominen 2004, 15). En käsittele näitä aksioomia tässä (ks. esittely esim. Tuominen 2004, 15–17).

Vaikka Kolmogorovin kehittämä todennäköisyysavaruus mahdollistaa satunnaisilmiöiden käsittelyn matemaattisesti, se ei kuitenkaan vastaa filosofisempaan kysymykseen siitä, mitä todennäköisyys oikein on (Tuominen 2004, 3). Hájek (2012) esittää kuusi keskeisintä tai yleisintä tulkintaa todennäköisyydelle, jotka on edelleen jaettavissa kolmeen eri luokkaan. En käsittele näitä tulkintoja tässä yhteydessä, mutta niiden olemassaolo osoittaa, että todennäköisyydelle käsitteenä ei ole olemassa objektiivista tulkintaa. Noudatan tässä tutkimuksessa Gelmanin et al. (2014) esittämää tulkintaa todennäköisyydelle, joka voidaan esittää sekä jotain asiaa koskevan tiedon tai epävarmuuden kautta:

"The guiding principle is that the state of knowledge about anything unknown is described by a probability distribution." (Gelman et al. 2014, 12).

"[...] probability a reasonable way of quantifying uncertainty[...]" (Gelman et al. 2014, 12).

Todennäköisyysjakaumat kuvaavat ensimmäisen lainatun virkkeen mukaan tiedon määrää jonakin ajanhetkenä (*"state of knowledge"*) jostakin tuntemattomasta asiasta. Toisen lainauksen mukaan todennäköisyys on järkevä tapa lukumäärällistää epävarmuus. Gelman et al. 2014 vaikuttavat käyttävän tietoa ja epävarmuutta toisilleen käänteisinä suureina: eli mitä enemmän jostain tuntemattomasta asiasta on tietoa, sitä vähemmän siihen liittyy epävarmuutta. Käytän tätä määritystä tiedolle ja epävarmuudelle käänteisinä suureina tässä tutkimuksessa. Lisäksi käytän tässä tutkimuksessa tietoa ja informaatiota synonyymeina toisilleen.

2.3.1.2 Huomautus bayesilaisen päättelyn "subjektiivisuudesta"

Elinkaariennusteiden laatiminen tehdään tässä tutkimuksessa bayesilaisen tilastollisten päättelyn piirissä. Päädyin tähän ratkaisuun ainakin 3 syystä: 1. Monet elinkaariennusteiden komponenttien mallinnuksessa käytetyt menetelmät on kehitetty bayesilaisen päättelyn piirissä, 2. Posterioriennustejakaumaan on helppoa lisätä "ketjuttamalla" lisää (ehdollisia) todennäköisyysjakaumia, joten esimerkiksi erilaisia lisäepävarmuuksia on suhteellisen helppoa elinkaariennusteisiin, 3. elinkaariennusteen komponenttien mallien sopivuuden tarkistukset on suhteellisen helppoa suorittaa nk. posterioriennustetarkistuksilla, joista kerrotaan tarkemmin mallinnuksen yhteydessä.

Bayesilaisen päättelyn piiriin kuuluvista menetelmistä sanotaan toisinaan, että ne ovat erityisen *subjektiivisia* (verrattuna ei-bayesilaisiin menetelmiin) (Gelman et al. 2014, 13). Esimerkiksi edellä viitatussa

artikkelissa (Hájek 2012), jossa luokitellaan todennäköisyyskäsitykset kuuteen luokkaan, bayeselainen päättely mainitaan vain luokkaan, jossa käsitellään subjektiivista todennäköisyyttä tai todennäköisyyttä uskon asteena. Gelman et al. (2014) kuitenkin argumentoivat, että bayesilaiset menetelmät olisivat erityisen subjektiivisia verrattuna muihin menetelmiin:

"Bayesian methods are sometimes said to be especially subjective because of their reliance on a prior distribution, but in most problems, scientific judgment is necessary to specify both the 'likelihood' and 'prior' parts of the model." (Gelman et al. 2014, 13).

Ja edelleen:

"In any case, however, certain elements requiring scientific judgment will remain, notably the choice of data included in the analysis, the parametric forms assumed for the distributions, and the ways in which the model is checked." (Gelman et al. 2014, 13).

Gelmanin et al. (2014) argumentti on, että vaikka bayesilaisissa menetelmissä tilastolliseen malliin voidaan tuoda vaihteleva määrä havaintoaineiston ulkopuolista informaatiota nk. priorijakaumien muodossa, bayesilaiset menetelmät eivät ole silti erityisen subjektiivisia verrattuna ei-bayesilaisiin menetelmiin. Tämä johtuu siitä, että kaikissa mallinnustilanteissa – riippumatta siitä, käytetäänko bayesilaisia vai ei-bayesilaisia menetelmiä – mallin laatijan on vähintäänkin valittava analyysissa käytettävä havaintoaineisto, valittava käytetty tilastollinen malli ja siihen kuuluvat parametrit sekä valittava, miten malli tarkistetaan.

Se, että käytän tässä tutkimuksessa bayesilaisia menetelmiä, ei lisää elinkaariennusteiden subjektiivisuutta verrattuna siihen, että käyttäisin ei-bayesilaisia menetelmiä. Tähän on kaksi syytä:

1. **Havaintoaineiston ulkopuolinen informaatio ei vaikuta lähes ollenkaan mallinnustuloksiin.**

   Tämä johtuu ensiksi siitä, että havaintoaineistot ovat malleissa yleisesti hyvin suuria. Toiseksi tämä johtuu siitä, että käytettävät priorijakaumat – eli havaintoaineiston ulkopuolisen informaation summaavat jakaumat – ovat lähes aina "melko epäinformatiivisia". Havaintoaineiston ulkopuolisen informaation hyvin pientä "painoa" käsitellään enemmän alla.

2. **Kaikkien mallien sopivuus suhteessa havaintoaineistoihin tarkistetaan nk. posterioriennustetarkistuksilla.**

   Posterioriennustetarkistuksilla tarkistetaan, että elinkaariennusteissa käytettävät ennustemallit toistavat alkuperäisen havaintoaineiston. Se, että mallit toistavat alkuperäiset havaintoaineistot, tekee ennusteista uskottavia ja vähemmän subjektiivisia. Tämän vuoksi elinkaariennusteiden komponenttien ensimmäinen mallinnustavoite on, että mallit toistavat alkuperäiset havaintoaineistot. Pyrin noudattamaan mm. tältä osin Gelman and Shalizi (2013):n esittämää bayesilaisen mallinnuksen käytäntöjä ja filosofiaa.

2.3.1.3 Elinkaariennusteiden kokonaisepävarmuus todennäköisyysjakaumin ilmaistuna

Määritin edellisessä alaluvussa uuden maahanmuuttajan elinkaarivaikutuksen komponenttien ja muiden malliparametrien posterioriennustejakauman (ks. edellä lauseke 2.4). Se on havaintoaineistolla ehdollinen todennäköisyysjakauma, josta saadaan elinkaarivaikutusten funktiolla muodostettua elinkaarivaikutusten ennustejakaumat uusille maahanmuuttajille.

Käsitellään nyt yhden elinkaarivaikutusten komponentin ennusteen (esim. palkkasummaennuste), y_{uusi}, ja sitä koskevien malliparametrien, $\boldsymbol{\theta}$, yhteistä posteriorijakaumaa: $p(y_{uusi}, \boldsymbol{\theta}|\boldsymbol{x}_{uusi})$. Tällä posteriorijakaumalle verrannollinen todennäköisyysjakauma saadaan lausekkeesta 2.4, kun otetaan lausekkeesta kaikki, mikä on lausekkeessa integraalin sisällä:

$$p(y_{uusi}, \boldsymbol{\theta}|\boldsymbol{x}_{uusi}) \propto p(y_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.5}$$

Otetaan tämän verrannollisuuden avulla käsittelyyn se, mistä elinkaarivaikutusten komponentin ja komponenttia koskevien malliparametrien yhteinen kokonaisepävarmuus koostuu. Yllä, epävarmuutta ja todennäköisyysjakaumia käsittelevässä alaluvussa, sanottiin, että todennäköisyysjakauma kuvaa epävarmuutta. Tämän perusteella jakauma $p(y_{uusi}, \boldsymbol{\theta}|\boldsymbol{x}_{uusi})$ kuvaa elinkaarivaikutuksen komponentin ja sitä koskevien malliparametrien epävarmuutta, kun huomioidaan uutta maahanmuuttajaa koskevat tiedot $\boldsymbol{x}_{uusi}$. Verrannollisuuden oikealla puolella on yllä kolme todennäköisyysjakaumaa, joten elinkaarivaikutuksen komponentin ennusteen ja malliparametrin yhteisposteriorijakauman kokonaisepävarmuus muodostuu kolmesta epävarmuuden osasta eli kolmesta todennäköisyysjakaumasta:

Elinkaariennusteen komponentin ennusteen ja malliparametrien θ kokonaisepävarmuus:

- Havaintoaineiston ulkopuolinen epävarmuus koskien malliparametreja: $p(\boldsymbol{\theta})$.
- Malliparametrien epävarmuuden päivitys havaintoaineistolla: $p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})$.
- Komponentin ennustejakauman epävarmuus: $p(y_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})$.

Ensimmäinen jakauma, nk. *priorijakauma*, $p(\boldsymbol{\theta})$, summaa havaintoaineiston ulkopuolisen epävarmuuden (tai informaation) koskien malliparametreja $\boldsymbol{\theta}$. Toinen jakauma, nk. *otosjakauma*, ”päivittää” malliparametreja koskevan epävarmuuden (tai informaation) vanhoja maahanmuuttajia koskevalla havaintoaineistolla. Viimeinen todennäköisyysjakauma, $p(y_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})$, kuvaa sitä, kuinka paljon elinkaarivaikutuksen komponentin ennusteeseen sisältyy epävarmuutta (tai informaatiota sen mahdollisista arvoista). Koska komponentin ennusteen jakauma on ehdollinen malliparametreilla, komponentin ennusteen epävarmuuden tasoon vaikuttaa malliparametrien epävarmuuden taso. Jäljempänä, alaluvussa 2.3.3, käsittelen vielä eritellymmin elinkaariennusteisiin sisältyvän kokonaisepävarmuuden koostumista.

2.3.1.4 Malliparametrien posteriorijakauma summaa niitä koskevan kokonaisinformaation

Käsittelen vielä hieman tarkemmin pelkkiin malliparametreihin, $\boldsymbol{\theta}$, liittyvää kokonaisinformaatiota (kokonaisepävarmuutta). Malliparametreilla on posteriorijakauma, joka summaa kaiken informaation liit-

tyen malliparametreihin. Tämä posteriorijakauma kirjoitetaan verrannollisuutena: $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \propto p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Tämän verrannollisuuden oikea puoli sisältyy uuden maahanmuuttajan elinkaarivaikutusten posterioriennustejakaumaan sekä yllä lausekkeeseen 2.5. Kuten jo yllä mainitaan, priorijakauma $p(\boldsymbol{\theta})$ summaa malliparametreja koskevan informaation aineiston ulkopuolelta ja otosjakauma, $p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})$, päivittää malliparametreja koskevan informaation havaintoaineiston tarjoamalla informaatiolla.

Malliparametrien kokonaisinformaation (posteriorijakauma) muodostuminen havaintoaineiston informaatiosta (otosjakauma) ja havaintoaineiston ulkopuolisesta informaatiosta (priorijakauma) voidaan kirjoittaa skemaattisesti auki seuraavana summana:

Malliparametrien kokonaisinformaatio
=
(Havaintoaineiston informaatio $\times$ havaintoaineiston paino)
+
(Havaintoaineiston ulkopuolinen informaatio $\times$ prioripaino)

Malliparametrin kokonaisinformaatio muodostuu toisin sanoen havaintoaineiston informaation ja havaintoaineiston ulkopuolisen informaation (priori-informaatio) painotettuna summana. Tämä summa voidaan käsittää myös havaintoaineiston informaation ja priori-informaation painotettuna keskiarvona, jos yllä laatikossa esitetty summa jaetaan painojen summalla.

Havaintoaineiston paino ja prioripaino kertovat sen, missä suhteessa painotetaan havaintoaineiston tarjoamaa informaatiota suhteessa havaintoaineiston ulkopuoliseen informaatioon, kun määritetään malliparametria koskevaa kokonaisinformaatiota. Karkeasti ajatellen painot ovat suhteessa niitä koskeviin havaintomääriin: havaintoaineiston paino vastaa havaintojen lukumäärää ja prioripaino havaintoaineiston ulkopuolisten ”havaintojen” lukumäärää (ks. erityinen esimerkki, jossa painot ovat suoraan havaintomääriä: Koop 2003, 19–23).

Tässä tutkimuksessa havaintoaineiston paino on suuri verrattuna prioripainoon. Mikäli havaintoaineiston paino on suuri verrattuna prioripainoon, malliparametrin kokonaisinformaatio muodostuu suurimmaksi osaksi havaintoaineiston informaatiosta. Tämä on tilanne tässä tutkimuksessa, eli havaintoaineiston paino on hyvin suuri verrattuna prioripainoon. Tämä johtuu kahdesta syystä: 1. Useimmissa malleissa havaintojen lukumäärä on todella suuri (eli useita kymmeniä tuhansia) ja 2. Priorijakaumat, $p(\theta)$, ovat todella leveitä.

Prioripainon kasvattaminen. Luvussa 12 esitetään, kuinka elinkaariennusteiden tuloksia voidaan soveltaa uusien maahanmuuttajien valintaan. Tässä yhteydessä myös selvennetään esimerkein, mitä elinkaariennusteiden tulosten kannalta tarkoittaa, jos elinkaariennusteisiin lisätään aineiston ulkopuolisten tekijöiden vaikutusten aiheuttamaa epävarmuutta. Malliparametrien kokonaisinformaation kannalta tämä tarkoittaa käytännössä sitä, että prioripainoa kasvatetaan suhteessa havaintoaineiston painoon. Äärimmäisessä esimerkissä (ks. esimerkki luvussa 12.3.2) tämä tarkoittaa sitä, että havaintoaineiston paino lähenee nollaa suhteessa prioripainoon, jolloin havaintoaineiston ulkopuolinen informaatio määrittää malliparametria koskevan kokonaisinformaation.

Parametrin kokonaisinformaation päivitys uudella havaintoaineistolla. Malliparametria koskeva informaatio ei ole staattista. Kuten luvussa 12.4.1 esitetään, malliparametria koskeva informaatio voidaan päivittää uuden havaintoaineiston tuomalla informaatiolla. Tällöin parametrin ”vanha” posteriorijakauman jakauman sisältämä informaatio päivitetään uuden havaintoaineiston tarjoamalla informaatiolla, jolloin saadaan parametrin ”uusi” posteriorijakauma, joka summaa jälleen kaiken informaation koksien parametria.

2.3.1.5 (Elinkaariennusteiden kokonaisentropia)

Informaatioteoriassa *entropia* on satunnaismuuttujan epävarmuuden mitta: *”[...] entropy, which is a measure of the uncertainty of a random variable.”* (Cover and Thomas 2006, 13). Koska entropian avulla voidaan mitata epävarmuutta, esitän elinkaariennusteisiin sisältyvän kokonaisentropian. En kuitenkaan palaa entropian käsitteeseen jäljempänä, koska todennäköisyys ennusteisiin sisältyvän epävarmuuden kuvaajana riittää. Joka tapauksessa satunnaismuuttujan entropia on ”lähellä” satunnaismuuttujaan liittyvää todennäköisyysjakaumaa. Jos X on satunnaismuuttuja, ja siihen liittyvä todennäköisyysjakauma on $p(X)$, niin tuon satunnaismuuttujan entropia on $-E[log(p(X))]$ (Cover and Thomas 2006, 243). Toisin sanoen entropia on satunnaismuuttujan todennäköisyysjakauman logaritmin odotusarvo kerrottuna -1:llä.

Elinkaariennusteen yhden komponentin, y_{uusi}, ehdollisen ennusteen kokonaisentropia eli ehdollinen kokonaisepävarmuus saadaan laskettua ehdollisen entropian avulla (Cover and Thomas 2006, 249). Tässä komponentin ennustejakauma on ehdollinen malliparametreilla, $\boldsymbol{\theta}$, ja havaintoaineistolla, $\boldsymbol{h}$. Logaritmin laskusääntöjen ja integrointisääntöjen avulla tämän ehdollisen ennustejakauman kokonaisentropia voidaan ilmaista seuraavasti:

$$H(y_{uusi}|\boldsymbol{\theta},\boldsymbol{h}) = \tag{2.6}$$

$$-\int_{Y_{uusi}}\int_{\boldsymbol{\Theta}}\int_{\boldsymbol{H}} p(y_{uusi},\boldsymbol{\theta},\boldsymbol{h}) \times log(p(y_{uusi}|\boldsymbol{\theta}))\, d\boldsymbol{h}d\boldsymbol{\theta}dy_{uusi} \tag{2.7}$$

$$-\int_{Y_{uusi}}\int_{\boldsymbol{\Theta}}\int_{\boldsymbol{H}} p(y_{uusi},\boldsymbol{\theta},\boldsymbol{h}) \times [log(p(\boldsymbol{h}|\boldsymbol{\theta})) - log(p(\boldsymbol{h}))]\, d\boldsymbol{h}d\boldsymbol{\theta}dy_{uusi} \tag{2.8}$$

$$-\int_{Y_{uusi}}\int_{\boldsymbol{\Theta}}\int_{\boldsymbol{H}} p(y_{uusi},\boldsymbol{\theta},\boldsymbol{h}) \times log(p(\boldsymbol{\theta}))\, d\boldsymbol{h}d\boldsymbol{\theta}dy_{uusi}. \tag{2.9}$$

Tässä kokonaisentropian määritelmässä Y_{uusi}, $\boldsymbol{\Theta}$ ja $\boldsymbol{H}$ ovat satunnaismuuttujia ja y_{uusi}, $\boldsymbol{\theta}$ sekä $\boldsymbol{h}$ ovat satunnaismuuttujien toteutuneita arvoja. Käsitän tässä yhteydessä havaintoaineiston $\boldsymbol{H}$ satunnaiseksi. Lisäksi komponentin ennuste Y_{uusi} on ehdollisesti riippumaton havaintoaineistosta $\boldsymbol{H}$, kun ehdollistajana ovat malliparametrit $\boldsymbol{\Theta}$.

Yllä oleva elinkaariennusteen yhden komponentin ehdollisen ennusteen kokonaisentropian lauseke voidaan lausua suomeksi (yllä olevien lausekkeiden numerot suluissa):

Komponentin ennusteen kokonaisentropia (2.6)
=
Komponentin ennustemallin osuus kokonaisentropiasta (2.7)
+
Komponentin aineistomallin ja havaintoaineiston osuus kokonaisentropiasta (2.8)
+
Komponenttia koskevan aineiston ulkopuolisen informaation osuus kokonaisentropiasta (2.9)

Komponentin ennusteen kokonaisentropia on ehdollinen malliparametreilla ja havaintoaineistolla, kuten lauseke 2.6 yllä kertoo. Ehdollinen entropia muodostuu sitä pienemmäksi, mitä enemmän komponentin ennusteella Y_{uusi} on keskinäistä informaatiota (*engl. mutual information*) malliparametrien $\boldsymbol{\Theta}$ ja havaintoaineiston $\boldsymbol{H}$ kanssa (Cover and Thomas 2006, 251). Tämä tarkoittaa sitä, että mitä enemmän malliparametrit ja havaintoaineisto ”selittävät” komponentin ennusteen arvoja, sitä pienemmäksi ehdollisen ennusteen entropia (ja epävarmuus) muodostuvat. Äärimmäisessä tapauksessa, jos malliparametrit ja havaintoaineisto määrittäisivät deterministisesti ennusteiden arvot, ennusteiden ehdollinen entropia saisi arvoksi nolla, eli ennusteisiin ei liittyisi mitään epävarmuuksia.

Suhteellinen entropia. Entropia on käsitteenä epäsuorasti mukana myös useiden elinkaarivaikutusten komponenttien mallinvalinnassa. Tämä johtuu siitä, että mallinvalintojen yhteydessä eri mallien ennustetarkkuutta vertaillaan ennustetarkkuuden mittareilla (mallinvalinta tässä tutkimuksessa perustuu kuitenkin ensi sijassa muihin kriteereihin). Nämä ennustetarkkuuden mittarit perustuvat siihen, että sellaiset mallit ovat parhaimpia, joiden tuottamat ennusteet poikkeavat vähiten todellisista toteutuvista arvoista. Tätä poikkeamaa ennusteen ja todellisten arvojen välillä – tai ennusteiden sopivuutta – mitataan nk. *Kullback-Leibler-divergenssillä*. Kullback-Leibler-divergenssi on puolestaan toinen nimitys *suhteelliselle entropialle* kahden satunnaismuuttujan välillä (Cover and Thomas 2006, 19, 251). Suhteellinen entropia on hyödyllinen väline ennustetarkkuuden mittaamisessa, koska havaintomäärän kasvaessa rajatta sillä mallilla, jolla on pienin suhteellinen entropia, on odotusarvoisesti suurin ennustetodennäköisyystiheyden logaritmi ja suurin posterioriennustejakauman todennäköisyys. (Gelman et al. 2014, 167).

2.3.2 Elinkaarivaikutusten ennusteet ovat jatkuvia ehdollisia todennäköisyysjakaumia

Elinkaarivaikutusten ennusteet ovat todennäköisyysjakaumia. Todennäköisyysjakaumat summaavat sen saatavilla olevan informaation, joka koskee tulevien elinkaariennusteiden mahdollisia arvoja, sekä epävarmuuden, joka ennusteisiin väistämättä sisältyy. Kuten edellä olen esittänyt, mahdollisten toteutuvien elinkaariennusteiden joukko muodostuu reaalilukujen joukosta, koska elinkaariennusteet ovat euromääräisiä. Tästä syystä elinkaariennusteiden todennäköisyysjakaumat ovat *jatkuvia* jakaumia, joilla on todennäköisyystiheysfunktiot. Tämä tarkoittaa sitä, että näiden jakaumien *kertymäfunktiot* ovat (absoluuttisesti) jatkuvia.

Käytännössä tämä tarkoittaa sitä, että elinkaariennusteiden todennäköisyysjakaumat määrittävät todennä-

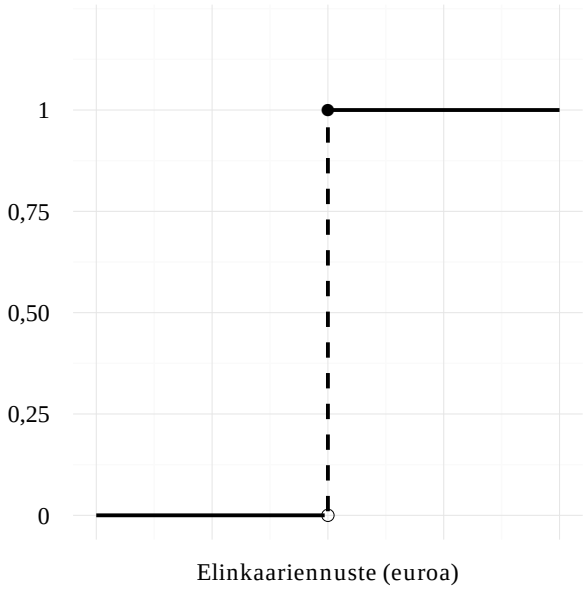
köisyyden sille, että uuden maahanmuuttajan tuleva toteutuva elinkaarivaikutus sijoittuu jollekin tietylle euromääräiselle *välille* tai *intervallille*. Elinkaariennusteiden jakaumat määrittävät esimerkiksi todennäköisyyden sille, että uuden maahanmuuttajan tuleva toteutuva elinkaarivaikutus on nollaa (euroa) suurempi. Samoin jakaumat määrittävät todennäköisyyden sille, että tuleva toteutuva elinkaarivaikutus on kahden euromääräisen luvun välissä, esimerkiksi $+100\,000$:n euron ja $+200\,000$:n euron välisellä intervallilla. Koska elinkaarivaikutusten ennustejakaumat ovat jatkuvia jakaumia, kaikkien yksittäisten euromääräisten (reaali)lukujen toteutumisen todennäköisyys on nolla. Esimerkiksi todennäköisyys sille, että toteutuva elinkaarivaikutus olisi esimerkiksi tasan $-200\,000$ euroa, on nolla.

Elinkaarivaikutuksen ennuste, kun uutta potentiaalista maahanmuuttajaa ei valita muuttavaksi Suomeen. Poikkeus tähän esiintyy päätöksentekoteoriaa ja uusien potentiaalisten maahanmuuttajien valintaa koskevassa luvussa 12. Luvussa esitetään uusia potentiaalisia maahanmuuttajia koskeva päätöksentekotilanne, jossa uuden maahanmuuttajan elinkaarivaikutus on *täysin varmasti* tasan nolla euroa, mikäli uutta potentiaalista maahanmuuttajaa ei valita muuttavaksi Suomeen (ks. tarkemmin 12.2.3). Tässsä tapauksessa uuden maahanmuuttajan elinkaarivaikutusten ennusteen kertymäfunktio, $KF(elinkaariennuste)$, on *degeneroitunut* niin, että kertymäfunktio saa avoimella välillä ennen nollaa euroa arvoksi nollan. Nollan euron ja siitä suurempien arvojen kohdalla kertymäfunktion arvo on puolestaan yksi.

Elinkaarivaikutusten ennusteen kertymäfunktio on degeneroitunut seuraavasti, kun potentiaalinen uusi maahanmuuttaja ei muuta Suomeen lainkaan:

$$KF(elinkaariennuste) = \begin{cases} 0, \text{ kun } elinkaariennuste < 0 \text{ (euroa)} \\ 1, \text{ kun } elinkaariennuste \geq 0 \text{ (euroa).} \end{cases} \tag{2.10}$$

Erityistapaus: elinkaariennusteeseen ei sisälly lainkaan epävarmuuksia. Uusien maahanmuuttajien elinkaariennusteiden jakaumien kertymäfunktiot olisivat myös degeneroituneita siinä hypoteettisessa tapauksessa, että tulevat toteutuvat elinkaarivaikutukset pystyttäisiin ennustemaan lainkaan ilman mitään epävarmuuksia, kuten kuvassa 2.1 esitetään. Tällöin kaikki ”todennäköisyysmassa” olisi kasautunut yhden euromääräisen arvon kohdalle: sen arvon, jonka tiedettäisiin tässä hypoteettisessa tapauksessa täysin varmasti toteutuvan. Seuraten VATT:n (2014) terminologiaa tässä tapauksessa kyse olisi uuden maahanmuuttajan ”hintalapusta”– mikäli hintalapulla tarkoitetaan yhtä varmasti toteutuvaa euromääräistä elinkaarivaikutusta.

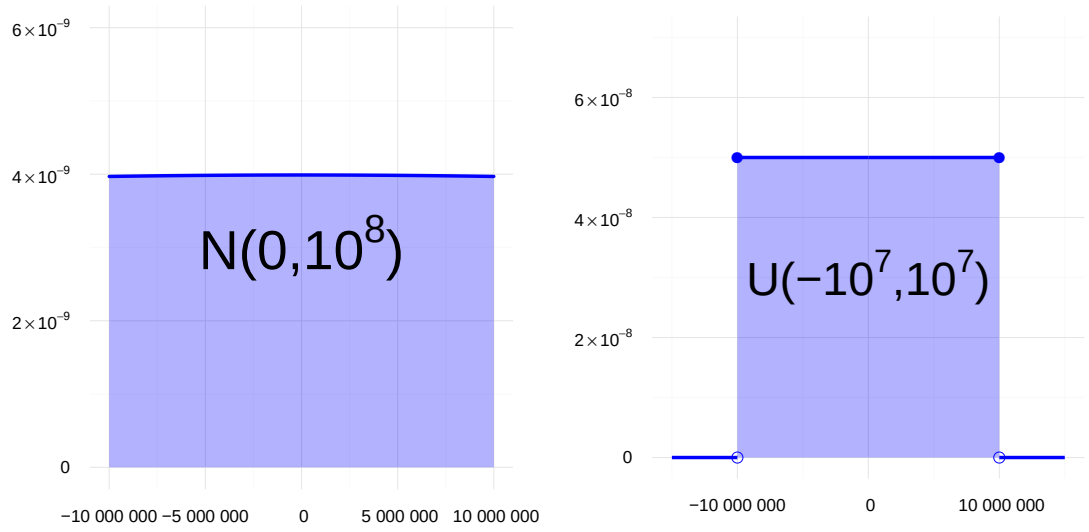


Kuva 2.1: *Hypoteettinen esimerkki degeneroituneesta elinkaariennusteen kertymäfunktiosta. Elinkaariennusteeseen ei sisälly tällöin lainkaan epävarmuuksia. Pystyakselilla on todennäköisyys ja vaaka-akselilla euromääräinen elinkaarivaikutus.*

Erityistapaus: elinkaariennusteista ei lähes minkäänlaista informaatiota. Vastakkaisessa hypoteettisessa esimerkissä uuden maahanmuuttajan tulevasta toteutuvasta elinkaarivaikutuksesta ei olisi (lähes) minkäänlaista informaatiota. Tätä (lähes) täydellistä epävarmuutta voisi kuvata esimerkiksi todennäköisyystiheysfunktiolla, joka saisi saman – tai lähes saman – arvon eli todennäköisyystiheyden (suljetulla) välillä miinus 10 miljoonasta eurosta plus 10 miljoonaan euroon.

Kuviossa 2.2 on esitetty kaksi esimerkkiä mainitunlaisista epäinformatiivisista jakaumista: keskistetty normaalijakauma, jonka keskihajonta on 100 miljoonaa euroa, sekä tasajakauma miinus 10 miljoonasta plus 10 miljoonaan euroon. Mikäli elinkaariennusteet saisivat tällaisia todennäköisyysjakaumia, tämä tarkoittaisi sitä, että epävarmuus yhden uuden maahanmuuttajan tulevien toteutuvien elinkaarivaikutusten osalta olisi niin täydellistä, että olisi esimerkiksi (lähes) yhtä todennäköistä, että tämän maahanmuuttajan tuleva elinkaarivaikutus olisi -9 miljoonan ja -10 miljoonan euron välillä kuin, että se olisi nollan euron ja $+1$ miljoonan euron välillä.

Elinkaariennusteiden todennäköisyysjakaumat ja niihin sisältyvä epävarmuus käytännössä. Käytännössä nämä esitetyt ääripäät eivät esiinny uusien maahanmuuttajien elinkaariennusteiden jakaumissa. Eli epävarmuus koskien tulevia toteutuvia elinkaarivaikutuksia ei ole niin pientä, että kaikki todennäköisyysmassa keskittyisi jonkin yksittäisen euromääräisen luvun ympäristöön. Toisaalta epävarmuudet eivät ole niin suuria, että jakaumien todennäköisyysmassa olisi levinnyt tasaisesti esimerkiksi mainittujen miinus 10 miljoonan ja plus 10 miljoonan euron välille. Nämä esimerkit auttanevat kuitenkin hahmottamaan



(a) *Esimerkki epäinformatiivisesta normaalijakaumasta. Pystyakselilla todennäköisyystiheys ja vaaka-akselila elinkaarivaikutus euroina.*

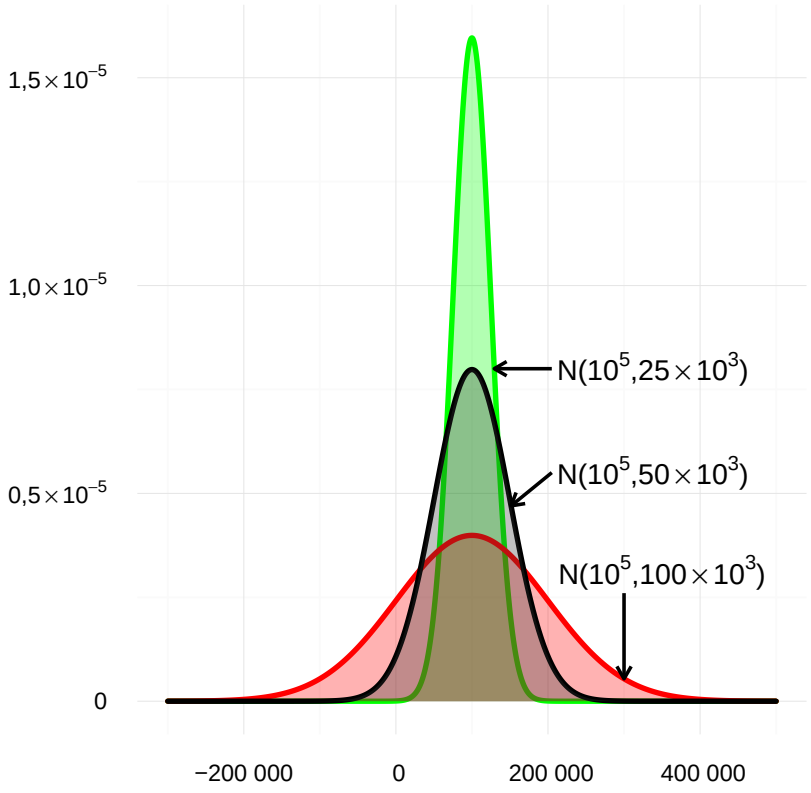
(b) *Esimerkki epäinformatiivisesta tasajakaumasta. Pystyakselilla todennäköisyystiheys ja vaaka-akselila elinkaarivaikutus euroina.*

Kuva 2.2: *Esimerkkejä elinkaarivaikutusten ennusteiden epäinformatiivisista todennäköisyysjakaumista.*

sitä, että elinkaarivaikutusten todennäköisyysjakaumat kuvaavat sitä, kuinka suuri epävarmuus tuleviin toteutuviin elinkaarivaikutuksiin sisältyy saatavilla olevan informaation perusteella.

Elinkaariennusteiden jakaumien leveydet kuvaavat sitä, kuinka suuria ovat tuleviin elinkaariennusteisiin liittyvät epävarmuudet. Epävarmuudet ovat sitä pienempiä, mitä kapeampia jakaumia ovat, ja puolestaan sitä suurempia, mitä leveämpiä elinkaariennusteiden jakaumat ovat. Toisin sanoen, mitä pienempiä epävarmuudet ovat, sitä kapeammalle (euromääräiselle) alueelle elinkaariennusteiden jakaumien todennäköisyysmassa keskittyy. Tämä tarkoittaa kääntäen sitä, että mitä suurempia epävarmuudet ovat, sitä leveämmälle (euromääräiselle) alueelle elinkaariennusteiden jakaumien todennäköisyysmassa levittäytyy.

Olen havainnollistanut kuvassa 2.3 epävarmuuksien kuvaamista elinkaariennusteissa niiden todennäköisyysjakaumien kautta. Kuviossa on kolme havainnollistamista varten keksittyä elinkaarivaikutusten ennustejakaumaa. Tässä keksityssä esimerkissä ennustejakaumat ovat normaalijakaumia, joilla on sama odotusarvo, +100 000 euroa, mutta niiden keskihajonnat poikkeavat toisistaan.



Kuva 2.3: *Havainnollistus: elinkaariennusteisiin sisältyviä epävarmuuksia kuvaavat erilaiset todennäköisyysjakaumat. Jakauman leveys kuvaa elinkaariennusteeseen sisältyvää epävarmuutta. Pystyakselilla todennäköisyystiheys ja vaaka-akselilla elinkaarivaikutus euroina.*

Kaikkein kapeimmalla ennustejakaumalla keskihajonta on kaikkein pienin, 25 000 euroa. Tämän ennustejakauman todennäköisyysmassa on keskittynyt lähes täysin nollan ja +200 000 euron välille. Uuden maahanmuuttajan, jolla on tällainen elinkaarivaikutusten ennustejakauma, toteutuva elinkaarivaikutus sijoittuu siis mainitulle välille lähestulkoon todennäköisyydellä yksi. Kaikkein leveimmän ennustejakauman keskihajonta on kaikkein suurin, 100 000 euroa. Mikäli uuden maahanmuuttajan elinkaarivaikutusten ennustejakauma noudattaa tätä jakaumaa, toteutuvaan elinkaarivaikutukseen sisältyy suurempia epävarmuuksia, kuin jos ennustejakauma noudattaisi kaikkein kapeinta jakaumaa. Kun kapeimman jakauman kohdalla elinkaarivaikutuksen todennäköisyys sijoittua nollan ja +200 000 euron välille on suurempi kuin 99,99 prosenttia, niin leveimmän jakauman kohdalla tämä todennäköisyys on noin 68 prosenttia. Todennäköisyys, että elinkaarivaikutus on nollaa pienempi, on jälkimmäisen jakauman kohdalla noin 16 prosenttia. Kaiken kaikkiaan, kuten edellä jo kirjoitin, mitä leveämpi elinkaarivaikutusten ennustejakauma on, sitä enemmän epävarmuuksia sisältyy siihen, mille euromääräiselle välillä toteutuva elinkaarivaikutus sijoittuu.

2.3.3 Epävarmuustekijät elinkaarivaikutusten ennustejakaumissa

Erittelen tässä luvussa ne epävarmuudet, joista elinkaarivaikutusten ennusteiden ehdolliset todennäköisyysjakaumat muodostuvat. Käsittelen aluksi epävarmuuksia, jotka koskevat edellä esitettyä uusien maahanmuuttajien posterioriennustejakaumaa (ks.lauseke 2.4, s. 30). Tämän jälkeen käsittelen tutkimusaineiston ulkopuolisten epävarmuuksien lisäämistä elinkaariennusteiden todennäköisyysjakaumiin. Tällä lisäepävarmuudella viittaan siihen ”ylimääräiseen” epävarmuuteen ennusteissa, mikä aiheutuu siitä, että elinkaariennusteet sijoittuvat tulevaisuuteen eli tämän tutkimuksen havaintoaineiston jälkeiseen aikaan.

Yhteistodennäköisyysjakauma. Toimikoon lähtökohtana elinkaarivaikutusten ennusteiden epävarmuuksien tarkastelemiseen edellä (luvussa 2.2.4) määritetty yhteistodennäköisyysjakauma elinkaariennusteiden komponenteille ($\boldsymbol{Y}_{uusi}$), näiden komponenttien malliparametreille ($\boldsymbol{\Theta}$), muille ennusteparametreille ($\boldsymbol{\Pi}$) sekä havaintoaineistoille ($\boldsymbol{H}$): $f_{\boldsymbol{\Pi},\boldsymbol{Y}_{uusi},\boldsymbol{\Theta},\boldsymbol{H}}(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h})$.

Tässä yhteydessä on syytä huomata, että $\boldsymbol{H}$ on satunnaismuuttuja, joka kuvaa siitä epävarmuutta, joka liittyy erilaisten mahdollisten havaintoaineistojen poimintaan. Mahdollisia havaintoaineistoja koskevaan satunnaismuuttujaan liittyvä todennäköisyysjakauma (todennäköisyystiheysfunktio) $f_{\boldsymbol{H}}(\boldsymbol{h})$ puolestaan kertoo, millä todennäköisyydellä erilaiset toteutuvat havaintoaineistot ($\boldsymbol{h}$) sijoittuvat jollekin mahdollisten havaintoaineistojen välille. Käytössä oleva tutkimusaineisto on tällainen eräs toteutunut havaintoaineisto, joka on poimittu kaikkien mahdollisten havaintoaineistojen muodostamasta (perus)joukosta.

Edellä (luvussa 2.2.5) esitettiin uusien maahanmuuttajien elinkaarivaikutusten ennustekomponenttien, näiden komponenttien malliparametrien ja muiden ennusteparametrien posteriorijakauma. Tämä tarkoittaa sitä, että näiden satunnaismuuttujien ehdollinen yhteistodennäköisyysjakauma on ehdollistettu toteutuneella havaintoaineistolla $\boldsymbol{h}$. Kuten edellä on esitetty, tämä posteriorijakauma voidaan lausua verrannollisuutena:

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h}) \propto p(\boldsymbol{\pi})p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta})p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

Tarkastellaan nyt posteriorijakaumaan $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h})$ sisältyviä epävarmuuksia komponentti kerrallaan.

2.3.3.1 Epävarmuudet tutkimusaineistossa $\boldsymbol{h}$

Yllä esitetty posteriorijakauman $p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}|\boldsymbol{h})$ verrannollisuus perustuu sille oletukselle, että käytössä olevaan tutkimusaineistoon (havaintoaineistoon) ei liity enää epävarmuuksia sen jälkeen, kun se on havaittu. Esitin kuitenkin edellä (ks. 2.3.3.1), että erityisesti havaittujen työikäisten nettovaikutusten osalta havaintoaineistossa on puutteita, jotka voidaan käsittää systemaattiseksi mittausvirheeksi.

Käytössä oleva tutkimusaineisto eli havaintoaineisto $\boldsymbol{h}$ koostuu kolmesta vektorikomponentista, niin kuin edellä on esitetty: $\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}, \boldsymbol{x}_{uusi}$. Mainittu systemaattinen (mittaus)virhe kohdistuu näistä komponenteista jo Suomessa olevia maahanmuuttajia koskevaan selittävien muuttujien havaintoaineistoon:

$\boldsymbol{x}_{vanha}$. Systemaattinen virhe selittävissä muuttujissa koskee lähinnä työikäisten julkisen talouden nettovaikutuksia. Tarkemmin näitä puutteita aineistossa käsittelen luvussa 10. Julkisen talouden menojen puolella puutteita on erityisesti maahanmuuttajien erityispalveluiden menoja koskien, sillä näitä ei ole tutkimuskäytössä olevissa rekisteriaineistoissa ainakaan kattavasti. Erityispalvelut koskevat erityisesti Suomeen turvapaikanhakijoina, kiintiöpakolaisina tai näiden perheenjäseninä muuttaneita. Julkisen talouden tulojen puolella on joitakin veroja ja veronluonteisia maksuja, joita aineistosta puuttuu, esimerkiksi henkilöiden maksamat kiinteistöverot.

Systemaattiset virheet tutkimusaineistossa. Olkoon työikäisten nettovaikutusten systemaattiset puutteet satunnaismuuttuja, josta käytetään merkintää ξ. Se saa havainnoittain euromääräisiä arvoja, ja sen arvo riippuu käytännössä eritoten lähtömaasta ja maassaoloajasta Suomessa, mutta yksinkertaisemman esitystavan vuoksi ei kiinnitetä tässä yhteydessä systemaattisten puutteiden satunnaismuuttujan ehdollisuuteen huomiota. Olkoon nyt $p(\xi)$ systemaattisiin puutteisiin liittyvä todennäköisyysjakauma, joka kuvaa puutteiden euromääräiseen suuruuteen liittyvää epävarmuutta, ja $p(\boldsymbol{y}_{vanha}|\xi)$ selitettävien muuttujien arvoja koskeva ehdollinen todennäköisyysjakauma – käytännössä otosjakauma – joka riippuu systemaattisten puutteiden ξ arvosta.

Edellä elinkaarivaikutusten malliparametrien posteriorijakaumaksi määritettiin $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$. Koska havaintoaineistoon ei oletettu sisältyvän enää mitään epävarmuuksia sen jälkeen, kun se oli havaittu, tämä posteriorijakauma oli verrannollinen otos- ja priorijakaumiin seuraavasti: $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \propto p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

Mikäli työikäisten nettovaikutusten systemaattiset puutteet ξ ja niihin sisältyvä epävarmuus huomioitaisiin nyt malliparametrien posteriorijakaumassa, niin tässä tapauksessa $\boldsymbol{\theta}$:n ja systemaattisten puutteiden ξ yhteisposteriorijakauma olisi: $p(\boldsymbol{\theta}, \xi|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \propto p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta}, \xi)p(\boldsymbol{\theta})p(\xi)$. Otosjakauma $p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta}, \xi)$ – tai $p(\boldsymbol{y}_{vanha}|\boldsymbol{\theta}, \xi, \boldsymbol{x}_{vanha})$ – on nyt ehdollinen sekä malliparametreilla $\boldsymbol{\theta}$, systemaattisia puutteita kuvaavalla satunnaismuuttujalla ξ sekä tunnetuiksi oletetuilla selittävillä muuttujilla $\boldsymbol{x}_{vanha}$. Systemaattisten puutteiden reunajakauma (priorijakauma) on yllä esitetty $p(\xi)$. Tässä systemaattisten puutteet ja malliparametrit on oletettu toisistaan tilastollisesti riippumattomiksi satunnaismuuttujiksi.

En kuitenkaan huomioi tätä systemaattista virhettä elinkaarivaikutusten ennusteiden posteriorijakaumissa uusille maahanmuuttajille, sillä systemaattisen virheen jakauman $p(\xi)$ ja ehdollisen jakauman $p(\boldsymbol{y}_{vanha}|\xi)$ määrittäminen perustuisi mielestäni liiaksi omiin käsityksiini. Koska systemaattisen virheen huomiointi muuttaisi erityisesti niiden maahanmuuttajaryhmien, joissa on paljon turvapaikanhakijoina, kiintiöpakolaisina tai heidän perheenjäseninään Suomeen muuttaneita, työikäisten julkisen talouden nettovaikutuksia negatiivisemmiksi (ks. luku 10), jätän systemaattisen virheen huomioinnin pois. Mikäli kuitenkin tulevaisuudessa erityisesti maahanmuuttajien erityispalveluiden menoja koskevat tiedot muodostuvat kokonaisvaltaisemmiksi ja luotettavammiksi, saatan jatkossa sisällyttää ne täysimääräisinä elinkaarivaikutusten ennusteisiin.

Muut mittausvirheet havaintoaineistossa. Työikäisten nettovaikutuksia koskevan systemaattisen mittausvirheen lisäksi havaintoaineistoon $\boldsymbol{h}$ sisältyy luultavasti muutakin mittausvirhettä. Olen kuitenkin

edellä perustellut, että koska havaintoaineisto perustuu luotettaviin suomalaisiin rekisteritietoihin, uskon, että tämä mittausvirhe on luonteeltaan satunnaista ja pientä – eli lähinnä valkoista kohinaa pienellä varianssilla. Poikkeuksen tästä muodostavat lähinnä maahanmuuttajia koskevat koulutustiedot, jotka ovat puutteellisia.

2.3.3.2 Epävarmuudet elinkaariennusteiden komponenttien havaintoaineistolla ehdollisissa malliparametreissa $\boldsymbol{\theta}|\boldsymbol{h}$

Elinkaariennusteiden komponentit $\boldsymbol{y}_{uusi}$ riippuvat niitä koskevista malliparametreista $\boldsymbol{\theta}$, kuten olen edellä esittänyt elinkaariennusteiden posteriorijakauman johtamisen yhteydessä sekä luvussa 2.2.2. Esimerkiksi olotilojen ennustamista varten laadin regressiomallin, joka sisältää useita malliparametreja, $\boldsymbol{\theta}_{\text{olotila-ajat}}$, joka on yksi $\boldsymbol{\theta}$:n seitsemästä vektorikomponentista.

Elinkaariennusteiden komponenttien malliparametrien posteriorijakauma, $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, summaa malliparametreja koskevan informaation – ja myös epävarmuudet – niiden mahdollisia arvoja koskien. Koska posteriorijakauma on verrannollinen otosjakauman, $p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta})$, ja malliparametrien priorijakauman, $p(\boldsymbol{\theta})$, tuloon nähden, posteriorijakauma on eräänlainen painotettu keskiarvo θ:n todennäköisyysjakaumalle havaintoaineiston ja havaintoaineiston ulkopuolisen informaation välillä. Toisin sanoen elinkaariennusteissa malliparametreihin sisältyvän epävarmuuden suuruus riippuu siitä, kuinka paljon epävarmuuksia malliparametreihin jää sen jälkeen, kun sekä havaintoaineiston tarjoama informaatio (otosjakauma) että muu havaintoaineiston ulkopuolinen informaatio (priorijakauma) on otettu huomioon.

Käytetyt parametrien priorijakaumat ovat melko epäinformatiivisia. Käytän kaikkien malliparametrien osalta nk. melko epäinformatiivisia priorijakaumia. Tämä tarkoittaa sitä, että tutkimusaineiston ulkopuolinen epävarmuus malliparametreissa on melko suurta. Oletan siis, ettei malliparametrien arvoista tiedetä tutkimusaineiston ulkopuolisen informaation perusteella kovinkaan paljon.

Havaintomäärät ovat suuria, joten otosepävarmuus jää pieneksi. Kun malliparametrien arvot ”päivitetään” käytössä olevan havaintoaineiston eli tutkimusaineiston tarjoaman informaation avulla, malliparametrien arvoihin liittyvä epävarmuus muuttuu ”pieneksi”. Tämä johtuu siitä, että tutkimusaineisto on maahanmuuttajaryhmäkohtaisesti havaintomääräisesti suuri. Jokaisen maahanmuuttajaryhmän kohdalla esimerkiksi työikäisten nettovaikutusmallinnus sisältää vähintään kymmeniä tuhansia havaintoja useilta tuhansilta eri henkilöiltä.

Koska malliparametrien estimointi perustuu otosaineistoon, malliparametreja koskevaa epävarmuutta jää aina jonkin verran jäljelle, vaikka havaintoaineisto olisikin suuri, kuten tässä tutkimuksessa on. Parametrien arvoihin sisältyvä epävarmuus lähestyy nollaa vasta, kun havaintomäärä lähestyy ääretöntä. On syytä huomioida, mistä joukosta tutkimuksessa käytettävissä oleva havaintoaineisto on poimittu. Jos tavoitteena olisi selvittää parametrien arvot – kuten esimerkiksi sukupuolen vaikutus työiän nettovaikutuksiin – vain Suomessa jo asuvia maahanmuuttajia koskien, tämä estimointi voitaisiin tehdä ilman jäljelle

jäävää epävarmuutta (olettaen, että havaintoaineistossa ei ole mittausvirhettä). Tämä johtuu siitä, että tutkimusaineisto sisältää kaikki Suomessa asuneet maahanmuuttajat, joten tutkimusaineistosta voidaan suoraan estimoida parametrien ”todelliset” arvot kiinnostuksen kohteena olevassa perusjoukossa.

Jo Suomessa olevat maahanmuuttajat otos kaikkien mahdollisten maahanmuuttajien perusjoukosta. Koska tutkimuksen tavoitteena on kuitenkin laatia elinkaariennusteet uusille maahanmuuttajille, eli Suomeen vasta ensimmäisen kerran muuttamassa oleville henkilöille, käsitän tutkimusaineistoon sisältyvät jo Suomessa asuneet ”vanhat” maahanmuuttajat otokseksi (perus)joukosta, joka sisältää kaikki mahdolliset nykyiset ja tulevat (”vanhat” ja ”uudet”) maahanmuuttajat Suomeen. Tällöin yksittäiset henkilöitä koskevat vuosittaiset havainnot ovat puolestaan otos kaikkien mahdollisten havaintojen (perus)joukosta, jotka koskevat kaikkia mahdollisia nykyisiä ja tulevia maahanmuuttajia.

Malliparametrien posteriorijakaumien simulointi. Koska malliparametrien posteriorijakaumalla $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$ ei ole analyyttista ratkaisua, se on simuloitava käyttämällä nk. *MCMC*-menetelmiä. Malliparametrien posteriorijakaumien simulointia näillä MCMC-menetelmillä käsitellään yksityiskohtaisesti jokaisen elinkaariennusteen komponentin mallinnuksen yhteydessä. Malliparametrien posteriorijakaumien simulointi aiheuttaa pienen lisäepävarmuuden malliparametrien arvoihin. Sitä virhettä, joka aiheutuu simuloinnista parametrien posteriorijakaumista tehtäviin (bayesilaisiin) tilastollisiin päätelmiin, kutsutaan *simulaatiovirheeksi* (*engl. simulation error*). Posterioriestimaattien kohdalla tätä simulaatiovirhettä kutsutaan *Monte Carlo -keskivirheeksi* (*engl. Monte Carlo Standard Error*). Koska malliparametrien posteriorijakaumien odotusarvojen Monte Carlo -keskivirheet ovat hyvin pieniä kaikissa elinkaarivaikutusten komponenttien malleissa, tämän perusteella oletan, että simulaatiovirheiden tuoma lisäepävarmuus malliparametrien jakaumiin on hyvin pientä verrattuna epävarmuuteen, joka niihin jää otosaineiston perusteella tehdyn estimoinnin jälkeen.

2.3.3.3 Epävarmuudet elinkaariennusteiden ehdollisissa komponenteissa $\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta}$

Uusien maahanmuuttajien elinkaariennusteiden komponenttien $\boldsymbol{y}_{uusi}$ jakaumat ovat ehdollisia malliparametreilla $\boldsymbol{\theta}$ ja uusia maahanmuuttajia koskevilla tiedoilla $\boldsymbol{x}_{uusi}$. Tämän vuoksi edellä esitetyt malliparametreja koskevat epävarmuudet sisältyvät varsinaisiin elinkaariennusteiden komponenttien jakaumien epävarmuuksiin. Oletan tässäkin yhteydessä, että havaintoaineistoon ei sisälly epävarmuuksia (käytännössä mittausvirhettä) sen jälkeen, kun se on havaittu. Tämä oletus perustuu siihen, että Suomessa rekisteriaineistojen tiedot ovat pääasiassa luotettavia. Tästä syystä uusia maahanmuuttajia koskevat tiedot $\boldsymbol{x}_{uusi}$ eivät sisällä epävarmuuksia, jotka sisältyisivät myös elinkaariennusteiden komponenttien jakaumiin.

Elinkaarivaikutusten komponenttien ehdollisiin jakaumiin sisältyvä epävarmuus aiheutuu siitä, että malliparametrien $\boldsymbol{\theta}$ ja uusia maahanmuuttajia koskevien selittävien muuttujien $\boldsymbol{x}_{uusi}$ muodostamat ennustesovitteet elinkaarivaikutusten komponenteille eivät tee näistä ennusteista deterministisiä. Tämä tarkoittaa sitä, että vaikka uusista maahanmuuttajia koskien onkin olemassa tiedot $\boldsymbol{x}_{uusi}$, niin näiden tietojen avulla ei voi ennustaa ilman epävarmuuksia, kuinka kauan uusi maahanmuuttaja on Suomessa, mitkä ovat hänen

julkisen talouden nettovaikutuksena työiässä ja kuinka paljon hän maksaa lakisääteisiä eläkevakuutusmaksuja ennen eläkkeelle jäämistään.

Mikäli malliparametrien ja uusia maahanmuuttajia koskevien tietojen avulla voitaisiin luoda täysin deterministisiä elinkaarivaikutusten komponenttien ennusteita, niin tämä näkyisi malliparametrien mallinnuksessa siten, että selittävien muuttujien $\boldsymbol{x}_{vanha}$ ja malliparametrien $\boldsymbol{\theta}$ muodostamien mallisovitteiden avulla voitaisiin täysin selittää/ennustaa selitettävien muuttujien $\boldsymbol{y}_{vanha}$ arvot. Esimerkiksi olotila-aikojen mallinnuksen yhteydessä tämä tarkoittaisi sitä, että uutta maahanmuutttajaa koskevien selittävien muuttujien ja malliparametrien avulla jokaisen uuden maahanmuuttajan suomessaoloaika voitaisiin ennustaa täysin varmasti. Samoin työikäisten nettovaikutuksia koskevissa ennusteissa voitaisiin jokainen yksittäinen nettovaikutus ennustaa täysin ilman epävarmuuksia malliparametrien sekä niiden tietojen avulla, joita uusista maahanmuuttajista tiedetään, kun he muuttavat ensimmäisen kerran Suomeen. Näin ei kuitenkaan ole, ja esimerkiksi työikäisten nettovaikutusennusteisiin jää ennustevirhettä, jonka hajonta on nollaa suurempaa.

Epävarmuudet ovat rakentuneet hierarkkisesti. Tutkimusaineisto h, joka on nk. paneeliaineisto tai pitkittäisaineisto, on rakenteeltaan ja otoksensa puolesta luonteeltaan hierarkkinen. Tämä tarkoittaa sitä, että yksittäiset havainnot ovat ryvästyneet henkilöihin, eli yksittäisiltä henkilöiltä on yksi tai useampi havainto yli ajan karttuneessa havaintoaineistossa. Henkilöt puolestaan voidaan jakaa edelleen omiin ryppäisiinsä sen perusteella, minä vuonna he ovat muuttaneet ensimmäisen kerran Suomeen. Yksittäiset havainnot voidaan ryvästää myös kalenterivuosien tai maassaoloajan vuosien mukaan. Koska aineisto on tällä tavalla hierarkkinen, myös satunnaisvaihtelu (tai epävarmuus) elinkaarivaikutusten komponenttien parametrien $\boldsymbol{\theta}$ malleissa on luonteeltaan hierarkkista. Epävarmuuksien hierarkkisuutta voidaan kutsua myös satunnaisvaihtelun monitasoisuudeksi.

Epävarmuuksien hierarkkisuus tai monitasoisuus tulee kaikkein selvemmin ilmi työikäisten nettovaikutusmalleissa. Niissä malleihin sisältyvä epävarmuus eli satunnaisvaihtelu on jaettu hierarkkisesti tai ”tasoittain” seuraavien termien kesken:

- Henkilövuositason eli yksittäisten havaintojen virhetermi.
- Henkilötason satunnaisvaikutus.
- Maahanmuuttovuositason tai kalenterivuositason satunnaisvaikutus.
- Maassaoloajan tason satunnaisvaikutus.

Satunnaisvaikutusten käsitteeseen palataan tarkemmin elinkaarivaikutusten komponenttien mallinnuksen yhteydessä. Joka tapauksessa tässä yhteydessä tärkeintä on ymmärtää, että työikäisten nettovaikutusmalleissa selittämätön vaihtelu on jaettu hierarkkisesti tai tasoittain mainittujen termien kesken. Myös palkkasummia koskeva mallinnus on luonteeltaan hierarkkinen, koska siinäkin havaintoaineisto on paneeli- tai pitkittäisaineisto.

Työikäisten nettovaikutusten ja eläkevakuutusmaksujen hierarkkisella mallinnuksella on se seuraus, että niitä koskevien elinkaarivaikutusten komponenttien ennusteet ovat myös rakenteeltaan hierarkkisia.

Tämä tarkoittaa sitä, että näiden komponenttien osalta niihin sisältyvä epävarmuus on luonteeltaan hierarkkista. Työikäisten nettovaikutusten ennusteiden osalta niihin sisältyvä epävarmuus voidaan jaotella hierarkkisesti tai tasoittain samalla tavoin, kuin yllä olevassa listassa on esitetty:

- Yksittäiseen nettovaikutusennusteeseen liittyvä epävarmuus.
- Henkilön vaikutukseen liittyvä epävarmuus.
- Maahanmuuttovuoden tai kalenterivuoden vaikutukseen liittyvä epävarmuus.
- Maassaoloajan vaikutukseen liittyvä epävarmuus.

Henkilöiden vaikutukseen liittyvä epävarmuus syntyy siitä, että uusia maahanmuuttajia koskeva informaatio, $\boldsymbol{x}_{uusi}$, ei sisällä kaikkia niitä henkilöihin liittyviä/henkilötason tekijöitä, jotka vaikuttavat tuleviin nettovaikutusiin. Tällaisia tekijöitä voivat olla esimerkiksi henkilön maahanmuuton syy tai henkilön tarkka koulutustieto muuttohetkellä, sillä nämä henkilöön liittyvät tiedot eivät sisälly havaintoaineistoon, mutta ne voivat hyvin vaikuttaa tulevien nettovaikutusten arvoihin. Samoin esimerkiksi yksittäisiin (tuleviin) kalenterivuosiin saattaa liittyä sellaisia tekijöitä, jotka vaikuttavat tulevien nettovaikutusten arvoihin, mutta joista ei ole tietoja tutkimusaineistossa.

Simuloin elinkaariennusteiden komponenttien ehdolliset todennäköisyysjakaumat, p($\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta}$). Tämän vuoksi näihin ennusteisiin sisältyy simulaation tuottama virhe tai epävarmuus samalla tavoin kuin malliparametrien simuloituihin jakaumiin edellä. Tässäkin tapauksessa simulaatiovirhe on kuitenkin pieni sen perusteella, että elinkaarivaikutusten komponenttien jakaumien odotusarvojen Monte Carlo - keskivirheet ovat pieniä.

2.3.3.4 Epävarmuudet muissa ennusteparametreissa π

Mikäli muut ennusteparametrit $\boldsymbol{\pi}$ käsitetään satunnaisiksi ja niiden estimointia varten on saatavilla havaintoaineistoa, muille ennusteparametreille voidaan muodostaa posteriorijakauma: $p(\boldsymbol{\pi}|\boldsymbol{d})$, missä $\boldsymbol{d}$ on muita ennusteparametreja koskeva havaintoaineisto. Tällöin muita ennusteparametreja koskevat epävarmuudet ovat samanlaisia kuin elinkaariennusteiden komponenttien malliparametreja koskevat epävarmuudet, joita käsittelin edellä. Koska muita ennusteparametreja koskevat havaintoaineistot ovat kuitenkin pieniä – niitä koskevat aikasarjat käsittävät tyypillisesti joitakin kymmeniä havaintoja – epävarmuudet koskien niiden arvoja ovat suuria vielä sen jälkeen, kun ne on estimoitu pieniin otoksiin perustuvista havaintoaineistoista. Malliparametrien $\boldsymbol{\theta}$ osalta epävarmuudet olivat estimoinnin jälkeen suurimmaksi osaksi pieniä, sillä niitä koskevat havaintoaineistot olivat suurimmaksi osaksi suuria.

2.3.4 Havaintoaineiston ulkopuolisten tekijöiden vaikutukset elinkaarivaikutusten ennustejakaumissa

2.3.4.1 Havaintoaineiston ulkopuoliset tekijät

Esitin edellä elinkaarivaikutusten ennusteisiin liittyvät epävarmuudet. Erittelin epävarmuudet sen mukaan, minkä satunnaismuuttujan (ehdolliseen) todennäköisyysjakaumaan ne sisältyvät ennusteissa. Epävarmuuden lähteet ennusteissa voidaan näin eritellä tutkimusaineistoon, malliparametreihin, elinkaarivaikutusten komponentteihin ja muihin ennusteparametreihin liittyviin epävarmuuksiin. Tässä malliparametrien todennäköisyysjakauma on ehdollinen tutkimusaineistolla ja elinkaarivaikutusten komponenttien ennusteiden todennäköisyysjakauma puolestaan malliparametreilla sekä uusia maahanmuuttajia koskevilla tiedoilla.

Parametrien posteriorijakaumat perustuvat käytännössä havaintoaineiston tarjoamaan informaatioon. Tässä yhteydessä voidaan palata siihen, että malliparametrien $\boldsymbol{\theta}$ ehdollinen todennäköisyysjakauma (posteriorijakauma) $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$ riippuu käytännössä havaintoaineistosta. Tämä johtuu siitä, kuten edellä esitin, että tutkimusaineisto on havaintomäärältään hyvin suuri ja siitä, että käytän tässä tutkimuksessa yleisesti ottaen vain melko epäinformatiivisia priorijakaumia. Koska priorijakauma tarjoaa (tutkimusaineistoon verrattuna) hyvin vähän informaatiota malliparametrien $\boldsymbol{\theta}$ arvoja koskien, mutta havaintomäärältään suuri tutkimusaineisto tarjoaa tätä informaatiota ”paljon”, malliparametrien todennäköisyysjakaumat määräytyvät käytännössä täysin käytettävissä olevan havaintoaineiston eli tutkimusaineiston perusteella.

Havaintoaineiston ulkopuoliset kalenterivuodet ja niiden mahdolliset vaikutukset. Laadin elinkaarivaikutusten ennusteita sellaisille uusille maahanmuuttajille, jotka muuttavat Suomeen ensimmäisen kerran vuoden 2011 jälkeen. Tutkimuksessa käytössä oleva havaintoaineisto ulottuu kuitenkin vain vuoden 2011 loppuun asti lukuun ottamatta palkkasumma-aineistoa, joka ulottuu vuoteen 2014. Koska elinkaariennusteiden komponenttien malliparametrien $\boldsymbol{\theta}$ ehdollinen todennäköisyysjakauma (posteriorijakauma) $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$ riippuu käytännössä vain tutkimusaineiston tarjoamasta informaatiosta, voidaan kysyä, että poikkeaako malliparametrien $\boldsymbol{\theta}$ todennäköisyysjakauma vuoden 2011 jälkeen jollain tavalla siitä todennäköisyysjakaumasta ($\boldsymbol{\theta}$:n posteriorijakaumasta), joka perustuu vuoteen 2011 ulottuvaan tutkimusaineistoon. Kenties vuoden 2011 jälkeiseen aikaan liittyy sellaisten tekijöiden vaikutuksia, jotka muuttavat malliparametrien $\boldsymbol{\theta}$ todennäköisyysjakaumaa siitä, mitä se on tutkimusaineistoon perustuvan estimoinnin jälkeen. Muutokset elinkaariennusteiden komponenttien malliparametrien todennäköisyysjakaumassa aiheuttaa myös muutoksia elinkaariennusteiden todennäköisyysjakaumassa. Koska tällaiset elinkaariennusteiden jakaumiin vaikuttavat tulevaisuuteen liittyvät tekijät tulevat tutkimusaineiston ulkopuolelta, kutsun näitä tekijöitä ”aineiston ulkopuolisiksi tekijöiksi”.

Havaintoaineiston ulkopuoliset maassaolovuodet ja niiden mahdolliset vaikutukset. Aineiston ulkopuoliset tekijöiden vaikutukset eivät ole kaikkien yksittäisten malliparametrien θ kohdalla vain vuoden

2011 jälkeiseen tulevaisuuteen liittyviä vaikutuksia. Työikäisten nettovaikutusennusteissa ja palkkasummaennusteissa maassaoloajan regressiokerroin on sellainen parametri, jonka jakaumaan vaikuttaa sekä vuoden 2011 jälkeiseen tulevaisuuteen liittyvät tekijät että maassaoloajan vaikutuksen ekstrapolointiin liittyvät tekijät. Maassaoloajan vaikutuksen ekstrapoloinnilla viittaan siihen, että tutkimusaineistossa maassaoloaikoja on – vuosissa mitattuna – 0:sta 23 maassaolovuoteen. Koska nuorten ja työikäisten nettovaikutusennusteissa maassaoloajan vaikutuksen todennäköisyysjakauma on kuitenkin määritettävä maksimissaan 46 maassaolovuodelle (62-vuotiaalle, joka muuttaa ensimmäisen kerran Suomeen 16-vuotiaana), voi olla, että maassaoloajan regressiokertoimen todennäköisyysjakauma maassaolovuosille 24–46 poikkeaa siitä jakaumasta, joka on estimoitu käytettävissä olevan havaintoaineiston sisältävien maassaolovuosien 0–23 perusteella. Koska tutkimusaineiston ulkopuolella jääviin maassaolovuosiin 24–46 voi liittyä sellaisia tekijöitä, jotka muuttavat maassaoloajan regressiokertoimen jakaumaa – verrattuna tutkimusaineistosta estimoituun jakaumaan – kutsun myös tällaisia ekstrapolointiin liittyviä tekijöitä aineiston ulkopuolisiksi tekijöiksi. Käsittelen tarkemmin eri aineiston ulkopuolisen tekijöiden vaikutusta maassaoloajan regressiokertoimen jakaumaan luvussa 12.3.2.

Satunnaisvaikutuksista havaintoaineistosta estimoituna epävarmuuden lähteenä. Aineiston ulkopuolisten tekijöiden osalta on huomattava, etteivät ne tarkoita hierarkkisiin (työikäisten nettovaikutusten) ennustemalleihin jo sisältyviä satunnaisvaikutuksia. Nämä satunnaisvaikutukset ottavat huomioon sen, että esimerkiksi maassaoloajan vaikutus ei ole kaikkien maassaoloaikojen kohdalla vakioarvoinen, vaan se saattaa vaihdella ”satunnaisesti” maassaolovuodesta toiseen. Samoin esimerkiksi sukupuolen vaikutus saattaa vaihdella ”satunnaisesti” maassaoloajan tai maahanmuuttovuoden mukaan. Toki aineiston ulkopuoliset tekijät voivat vaikuttaa myös näiden satunnaisvaikutusten suuruuteen siten, että ne vaikuttavat satunnaisvaikutusten suuruuden määräävien parametrien todennäköisyysjakaumiin.

Havaintoaineiston ulkopuoliset tekijät voivat myös pienentää elinkaariennusteisiin liittyviä epävarmuuksia. En kutsu aineiston ulkopuolisia tekijöitä myöskään ”epävarmuustekijöiksi”, sillä aineiston ulkopuoliset tekijät voivat – ainakin teoriassa – pienentää elinkaarivaikutusten komponenttien parametreihin liittyviä epävarmuuksia ennusteissa. Tällainen tilanne syntyy, jos aineiston ulkopuolisten tekijöiden avulla jonkin malliparametrin arvo tiedetään tarkemmin tulevaisuudessa kuin aineiston perusteella laaditun estimoinnin perusteella.

Esimerkkejä havaintoaineiston ulkopuolisten tekijöiden vaikutuksista tutkimuskirjallisuudessa

Ennen kuin esitän formaalimmin, kuinka aineiston ulkopuolisten tekijöiden vaikutukset sijoittuvat elinkaarivaikutusten ennustejakaumiin, käyn tutkimuskirjallisuuteen viitaten lyhyesti läpi, mitä aineiston ulkopuoliset tekijöiden vaikutukset voivat käytännössä olla.

Aineiston ulkopuolisten tekijöiden vaikutusten konkretisoimiseksi voidaan esittää kysymys siitä, millaisia ovat mahdollisesti sellaiset aineiston ulkopuoliset tekijät, joiden vaikutuksesta malliparametrien jakaumat muuttuvat tutkimusaineiston ulkopuolella. Toisin sanoen, millaisia voivat käytännössä olla sellaiset tekijät, jotka vaikuttavat siihen, että maassaoloajan vaikutus on erilainen 23 vuoden jälkeen kuin 0–23 maassaolovuoden aikana? Tai millaisia voivat käytännössä olla tekijät, jotka vaikuttavat siihen, että ka-

lenterivuoteen liittyvät vaikutukset ovat vuoden 2014 jälkeen erilaisia kuin vuosina 1988–2014? Samaa voidaan kysyä myös niitä maahanmuuttovuosia koskien, jotka tulevat vuoden 2014 jälkeen.

Tutkimusartikkeleissa Sarvimäki (2017b) ja VATT (2014a) sekä VATT (2014b) esitetään sellaisia tiettyihin malliparametreihin liittyviä vaikutuksia, joita kutsun edellä esitetyn perusteella aineiston ulkopuolisiksi tekijöiden vaikutuksiksi.

Kalenterivuosiin liittyvät aineiston ulkopuolisten tekijöiden vaikutukset. Ensin mainitun artikkelin yhteenvedossa viitataan kahteen mahdolliseen aineiston ulkopuolisten tekijöiden vaikutukseen (s. 21). Ensiksi, samasta lähtömaasta kotoisin olevien maahanmuuttajien (kohorttien) julkisen talouden vaikutukset voivat vaihdella maahanmuuttovuoden mukaan siitä syystä, että muuttajat voivat olla erilaisia eri vuosina. Aineiston pohjalta tehdyissä ennusteissa on huomioitu maahanmuuttovuoden – tai ainakin maahanmuuttovuosiluokan – vaikutukset, mutta aineisto ei sisällä tietoja aineiston viimeisimmän kalenterivuoden (vuosi 2014) jälkeisistä maahanmuuttajista. Se *epävarmuuden osa*, joka kohdistuu vuosien 2014 jälkeisten maahanmuuttovuosien kohorttivaikutusten suuruuteen – mikä johtuu havaintoaineiston jälkeiseen *tulevaisuuteen liittyvästä epävarmuudesta* – on tässä tapauksessa aineiston ulkopuolisten tekijöiden vaikutus. Kohorttivaikutuksen kohdalla vaikutus ennusteisiin välittyy erityisesti maahanmuuttovuosivaikutuksen kautta.

Maahanmuuttovuosiin liittyvät aineiston ulkopuolisten tekijöiden vaikutukset. Toinen mainitun artikkelin, Sarvimäki (2017b), esittämä, ja myös lähinnä maahanmuuttovuoden vaikutukseen liittyvä, aineiston ulkopuolisten tekijöiden vaikutus muodostuu uusista kotouttamispolitiikan lähestymistavoista, jotka voivat olla tehokkaampia kuin aiemmat: *"[...] the new approaches in integration policies that Finland is currently adopting may be more efficient than previous policies."* (s. 21). Koska kotouttaminen sijoittuu maassaoloajan alkupäähän, nämä uudet kotouttamiskäytännöt koskevat tällöin lähinnä sellaisia maahanmuuttajia, jotka ovat vasta muuttaneet ensimmäisen kerran Suomeen tai muuttavat niin tulevaisuudessa. Näin ollen uudet kotouttamispolitiikan lähestymistavat ovat sellainen aineiston ulkopuolisten tekijöiden vaikutus, joka voi vaikuttaa elinkaariennusteiden jakaumiin lähinnä maahanmuuttovuoden vaikutusten kautta.

VATT (2014a) esittää yhdeksi ennusteita koskevaksi epävarmuustekijäksi "'tulevaisuuden politiikkamuutokset", missä politiikkamuutoksiin sisältyvät esimerkiksi verojärjestelmän muutokset (s. 34). Tässä tapauksessa vaikutukset elinkaariennusteisiin välittyvät eniten kalenterivuoden vaikutusten kautta, sillä politiikkamuutokset voidaan liittää tiettyihin kalenterivuosiin. Politiikkamuutoksiin liittyvä aineiston ulkopuolisten tekijöiden vaikutus on jälleen se ylimääräinen, vuoden 2011 jälkeisestä tulevaisuudesta johtuva epävarmuus koskien politiikkamuutosten suuruutta.

2.3.4.2 Havaintoaineiston ulkopuolisten tekijöiden vaikutukset elinkaarivaikutusten ennusteiden todennäköisyysjakaumissa

Olkoon nyt kaikkien aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektori $\boldsymbol{\epsilon}$. Sen pituus on yhtä suuri kuin kaikkien malliparametrien vektoreiden $\boldsymbol{\theta}$ pituus. Jokaista yksittäistä malliparametria kohden on olemassa yksi aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektorin komponentti. Jokaista yksittäistä malliparametria θ koskeva aineiston ulkopuolisten tekijöiden komponentti ϵ määrää, kuinka aineiston ulkopuoliset tekijät vaikuttavat malliparametrin todennäköisyysjakaumaan. Tätä käsitellään tarkemmin alla.

Yhteistodennäköisyysjakauma, joka sisältää aineiston ulkopuolisten tekijöiden vaikutukset. Kirjoitetaan nyt uudestaan aiemmin tässä luvussa esitetty kaikkien käsiteltävien satunnaismuuttujien yhteistodennäköisyysjakauma, sekä posteriorijakauma, joka on verrannollinen tälle yhteisjakaumalle. Aiemmasta poiketen, nyt aineiston ulkopuolisten tekijöiden vaikutusten vektori $\boldsymbol{\epsilon}$ sisältyy sekä posteriorijakaumaan että yhteistodennäköisyysjakaumaan:

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{h}) \propto p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{\epsilon}). \tag{2.11}$$

Edelleen, aiemmin johdettuun satunnaismuuttujien posteriorijakaumaan viitaten, nyt posteriorijakauma voidaan kirjoittaa aineiston ulkopuolisten tekijöiden vaikutusten kanssa:

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}, \boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{h}) \propto p(\boldsymbol{\pi})p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta}, \boldsymbol{\epsilon})p(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}). \tag{2.12}$$

Aineiston ulkopuolisten tekijöiden vaikutuksen malliparametrien posteriorijakaumassa. Edellä, kun aineiston ulkopuolisten tekijöiden vaikutuksia ei huomioitu, riitti määrittää malliparametrien ”vanhoja” maahanmuuttajia koskevasta havaintoaineistosta riippuva posteriorijakauma: $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$. Nyt, kun aineiston ulkopuolisten tekijöiden vaikutukset huomioidaan, on määritettävä malliparametrien ja aineiston ulkopuolisten tekijöiden yhteisposteriorijakauma: $p(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$. Tämä jälkimmäinen posteriorijakauma voidaan lausua seuraavasti, kun huomioidaan, että aineiston ulkopuoliset tekijät ja havaintoaineisto ovat toisistaan riippumattomia:

$$p(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) = p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})p(\boldsymbol{\epsilon}). \tag{2.13}$$

Jotta aineiston ulkopuolisten tekijöiden vaikutus huomioitaisiin elinkaarivaikutusten ennustejakaumissa, on määritettävä todennäköisyysjakaumat:

1. $p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$.
2. $p(\boldsymbol{\epsilon})$.

Ensimmäinen näistä jakaumista on malliparametrien havaintoaineistolla sekä aineiston ulkopuolisilla tekijöillä ehdollinen jakauma. Toinen jakauma on puolestaan aineiston ulkopuolisten tekijöiden vaikutusten

reunajakauma.

1. Havaintoaineiston ulkopuolisten tekijöiden vaikutukset malliparametrien ehdollisissa jakaumissa

Ehkä helpoin tapa, jolla voidaan määrittää malliparametrien havaintoaineistolla sekä aineiston ulkopuolisten tekijöiden vaikutuksilla ehdollinen jakauma, on mieltää aineiston ulkopuoliset tekijät malliparametrien todennäköisyysjakaumaa siirtävänä vaikutuksena:

$$p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \doteq p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) + \boldsymbol{\epsilon}. \tag{2.14}$$

Nyt tässä malliparametrien posteriorijakauma, $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, on sama jakauma kuin edellä määritetyssä elinkaarivaikutusten ennustejakaumassa, jossa ei huomioitu aineiston ulkopuolisten tekijöiden vaikutuksia. Tähän jakaumaan lisätään aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektori $\boldsymbol{\epsilon}$, mikä siirtää malliparametrien jakaumaa, mikäli aineiston ulkopuolisten tekijöiden satunnaisvektorin arvo poikkeaa nollavektorista.

On muitakin mahdollisuuksia määrittää aineiston ulkopuolisten tekijöiden vaikutus malliparametreihin. Esimerkiksi malliparametreihin liittyvä epävarmuus voi joko kasvaa tai pienentyä, jos aineiston ulkopuolisten tekijöiden vaikutuksesta ne skaalautuvat uudelleen:

$$p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \doteq p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) * \boldsymbol{\epsilon}. \tag{2.15}$$

Esimerkki havaintoaineiston ulkopuolisten tekijöiden vaikutuksesta malliparametrin posteriorijakaumaan. Tarkastellaan nyt vielä esimerkkinä aineiston ulkopuolisten tekijöiden vaikutusta yhden malliparametrin (θ) posteriorijakaumaan $p(\theta|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$. Olkoon tämä malliparametri selvyyden vuoksi tässä riippumaton muista malliparametreista sekä annetaan sen olla jonkin elinkaarivaikutuksen komponentin mallin regressiokerroin. Koska regressiokertoimet ovat malleissa jatkuvia ja rajoittamattomia ja havaintoaineistomäärät ovat malleittain vähintään tuhansia, eli havaintomäärät ovat ”suuria”, regressiokertoimien havaintoaineistoihin perustuvat posteriorijakaumat ovat likimain normaalijakautuneita (ks. Gelman et al. 2014, 88). Toisin sanoen: $p(\theta|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim N(\mu_h, \sigma_h^2)$, missä μ_h ja σ_h^2 ovat tutkimusaineistosta $\boldsymbol{h}$ saadut estimaatit normaalijakauman odotusarvolle ja varianssille.

Tämän yksittäisen regressiokertoimen havaintoaineistolla ja aineiston ulkopuolisten tekijöiden vaikutuksilla ehdollinen todennäköisyysjakauma voidaan nyt lausua yllä esitettyä määritelmää seuraten:

$$p(\theta|\epsilon, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim N(\mu_h, \sigma_h^2) + \epsilon = N(\mu_h + \epsilon, \sigma_h^2). \tag{2.16}$$

Tässä ϵ on satunnaismuuttuja, joka on se aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektorin komponentti, joka vaikuttaa kyseessä olevan yksittäisen regressiokertoimen ehdolliseen todennäköisyysjakaumaan.

Mikäli ϵ saa nollaa suurempia arvoja, regressiokertoimen normaalijakauma siirtyy oikealle: regressiokerroin saa siis tällöin todennäköisemmin suurempia arvoja, kuin jos aineiston ulkopuolisten tekijöiden

vaikutusta ei huomioitaisi. Mikäli ϵ saa puolestaan nollaa pienempiä arvoja, regressiokertoimen normaalijakauma siirtyy puolestaan vasemmalle: regressiokerroin saa siis tällöin todennäköisemmin pienempiä arvoja, kuin jos aineiston ulkopuolisten tekijöiden vaikutusta ei huomioitaisi. Yleisesti ottaen, koska ϵ on satunnaismuuttuja, sen huomiointi tuo regressiokertoimen jakaumaan lisäepävarmuutta verrattuna pelkästään havaintoaineistoon perustuvaan jakaumaan, jossa aineiston ulkopuolisia tekijöitä ei huomioida.

2. Havaintoaineiston ulkopuolisten tekijöiden vaikutusten reunajakauma

Yllä määritettiin, että aineiston ulkopuoliset tekijät vaikuttavat malliparametrien ehdolliseen jakaumaan siirtämällä jakaumaa. Näiden siirtymien eli aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektori on $\boldsymbol{\epsilon}$. Kuten yllä määritettiin, tämän vektorin pituus on malliparametrien vektorin pituus, sillä jokaiselle malliparametrille määritetään aineiston ulkopuolisten tekijöiden vaikutus malliparametrin ehdolliseen todennäköisyysjakaumaan.

Satunnaisvektorille $\boldsymbol{\epsilon}$ on määritettävä sen todennäköisyysjakauma eli reunajakauma $p(\boldsymbol{\epsilon})$. Tämä todennäköisyysjakauma kuvastaa elinkaariennusteen laatijan epävarmuutta sitä koskien, kuinka aineiston ulkopuoliset tekijät vaikuttavat malliparametrien $\boldsymbol{\theta}$ ehdollisiin todennäköisyysjakaumiin, eli kuinka ne – yllä olevaa määritelmää noudattaen – siirtävät näitä jakaumia. Yleisesti ottaen, mitä leveämmäksi elinkaariennusteen laatija määrittää tämän todennäköisyysjakauman, sitä vähemmän hänellä on informaatiota aineiston ulkopuolisten tekijöiden vaikutuksesta. Todennäköisyysjakauma on puolestaan sitä kapeampi, mitä enemmän elinkaariennusteen laatijalla on informaatiota aineiston ulkopuolisten tekijöiden vaikutuksesta malliparametrien jakaumiin. Tämä informaatio tulee määritelmän mukaisesti tutkimusaineiston ulkopuolelta, sillä kyseessä on aineiston *ulkopuolisten* tekijöiden vaikutus.

Elinkaariennusteen laatija voi määrittää aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektorin $\boldsymbol{\epsilon}$ noudattavan esimerkiksi d-ulotteista ”leveää” Studentin t-jakaumaa, missä d on satunnaisvektorin ja malliparametrien vektorin pituus. Studentin t-jakauma on ”leveä” silloin, kun sen vapausasteiden määrä on ”pieni”, esimerkiksi 4. Tällaisella jakaumalla elinkaariennusteen laatija mahdollistaa isojenkin poikkeamien suhteellisen suuren mahdollisuuden aineiston ulkopuolisten tekijöiden vaikutuksesta. Tällöin ulkopuolisten tekijöiden vaikutusten satunnaisvektorin reunajakauma voidaan kirjoittaa:

$$\boldsymbol{\epsilon} \sim T_d(4, \boldsymbol{\mu}_\epsilon, \boldsymbol{\Sigma}_\epsilon), \tag{2.17}$$

missä 4 on d-ulotteisen Studentin t-jakauman vapausasteiden määrä, $\boldsymbol{\mu}_\epsilon$ on jakauman d-ulotteinen odotusarvovektori ja $\boldsymbol{\Sigma}_\epsilon$ on jakauman hajontaparametrien d-ulotteinen matriisi. Yksinkertaisuuden vuoksi määritän myös moniulotteisen t-jakauman reunajakaumien eli yksittäisten aineiston ulkopuolisten tekijöiden vaikutusten jakaumien $p(\epsilon)$ noudattavan yksiulotteisia t-jakaumia, joiden vapausaste on 4.

Erityistapaus: aineiston ulkopuolisten tekijöiden vaikutusten reunajakauman odotusarvovektori on nollavektori. Mainitsen tästä jakaumasta erityistapauksena jakauman, jossa jakauman odotusarvovektori on d-ulotteinen 0-vektori eli:

$$E_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}] = \boldsymbol{\mu}_\epsilon = \mathbf{0}_{\mathrm{d}}. \tag{2.18}$$

Tämä jakauma sopii mielestäni kuvastamaan elinkaariennusteen laatijan epävarmuutta aineiston ulkopuolisten tekijöiden vaikutuksesta sellaisessa tilanteessa, jossa elinkaariennusteen laatijalla ei ole sellaista informaatiota näitä vaikutuksia koskien, joka kertoisi ovatko vaikutukset todennäköisemmin nollaa suurempia vai nollaa pienempiä. Eli jos elinkaariennusteen laatija ei tiedä, siirtyvätkö malliparametrien tutkimusaineistosta estimoidut jakaumat todennäköisemmin vasemmalle vai oikealle – eli pienempiä vai suurempia arvoja saaviksi – aineiston ulkopuolisten tekijöiden vaikutuksesta, tällaista epävarmuutta kannattaa kuvata todennäköisyysjakaumilla, jotka on keskistetty nollaan ja ovat symmetrisiä sen suhteen.

Elinkaariennusteen laatija voi tietenkin määrittää aineiston ulkopuolisten tekijöiden vaikutusten reunajakauman muulla tavalla, kuin tässä on esitetty. Tässä esitettyä realistisemman jakauman voisi määrittää siten, että aineiston ulkopuolisten tekijöiden vaikutuksiin liittyvä epävarmuus kasvaa sitä suuremmaksi, mitä pidemmälle tulevaisuuteen ennusteet sijoittuvat. Studentin t-jakaumaa käyttäessä tämä onnistuisi niin, että hajontaparametrien matriisin $\boldsymbol{\Sigma}_\epsilon$ komponentit eivät olisi kiinteitä, vaan ne saisivat sitä suurempia arvoja, mitä pitemmälle tulevaisuuteen mennään.

3. Malliparametrien ja havaintoaineiston ulkopuolisten tekijöiden vaikutusten yhteisposteriorijakauma

Olen yllä määrittänyt malliparametrien ehdollisen jakauman, $p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, sekä aineiston ulkopuolisten tekijöiden vaikutusten reunajakauman $\boldsymbol{\epsilon}$. Malliparametrien sekä aineiston ulkopuolisten tekijöiden vaikutusten yhteisposteriorijakauma saadaan näiden tulona niin, kuin yllä on esitetty: $p(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) = p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})p(\boldsymbol{\epsilon})$.

Esitän tämän yhteisposteriorijakauman esimerkinomaisesti vielä yhdelle malliparametrille sekä sitä koskevalle aineiston ulkopuolisten tekijöiden vaikutukselle. Malliparametri on tässä regressiokerroin, sillä sen pelkästä tutkimusaineistosta estimoitu posteriorijakauma on likimain normaalijakautunut, kuten yllä on esitetty. Lisäksi tässä aineiston ulkopuoliset tekijät vaikuttavat regressiokertoimen jakaumaan siirtämällä sitä. Muutkin vaihtoehdot vaikutuksesta – kuten regressiokertoimen jakauman skaalaus – olisivat mahdollisia. Oletan tässä, että regressiokerroin sekä sitä koskeva aineiston ulkopuolisten tekijöiden vaikutus ovat riippumattomia muista malliparametreista. Tällöin yhteisposteriorijakauma voidaan esittää:

$$p(\theta, \epsilon|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim \begin{cases} N(\mu_h, \sigma_h^2), & \text{kun } 1995 \leq \textit{kalenterivuosi} \leq 2011 \\ N(\mu_h + \epsilon, \sigma_h^2)\, T(4, \mu_\epsilon, \sigma_\epsilon), & \text{kun } \textit{kalenterivuosi} > 2011. \end{cases} \tag{2.19}$$

Olen tässä määritelmässä korostanut sitä, että malliparametrit (tässä tapauksessa yksittäinen regressiokerroin) eivät käytännössä riipu aineiston ulkopuolisten tekijöiden vaikutuksesta, kun kyse on kalenterivuosista, joihin tutkimusaineisto sijoittuu. Nämä kalenterivuodet ovat tutkimusaineistoa koskevat vuodet 1995–2011.

Tämä johtuu tutkimusaineiston laajuudesta: on vaikeaa kuvitella, että olisi olemassa sellaista aineiston ulkopuolista informaatiota, joka muuttaisi ”käytännön kannalta merkitsevästi” malliparametrien posteriorijakaumia tutkimusaineiston tarjoaman informaation perusteella estimoiduista jakaumista, kun on kyse vuosista 1995–2011. Tästä syystä esimerkkinä olevan regressiokertoimen ja aineiston ulkopuolisten te-

kijöiden vaikutuksen yhteisposteriorijakauma vuosina 1995–2011 vastaa pelkästään tutkimusaineistosta estimoitua posteriorijakaumaa regressiokertoimelle: $p(\theta|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim N(\mu_h, \sigma_h^2)$. Tämä voidaan myös mieltää erityistapauksena θ:n ja ϵ:n yhteisposteriorijakaumasta: nyt ϵ:n arvo on nolla, ja se on vakio.

Tutkimusaineiston ulkopuolella tulevaisuudessa (eli vuodesta 2012 eteenpäin) yhteisposteriorijakauma on regressiokertoimen havaintoaineistolla ja aineiston ulkopuolisten tekijöiden vaikutuksilla ehdollisen jakauman, $N(\mu_h + \epsilon, \sigma_h^2)$, sekä aineiston ulkopuolisten tekijöiden vaikutusten reunajakauman, $T(4, \mu_\epsilon, \sigma_\epsilon)$, välinen tulo. Erityistapauksessa, jossa elinkaariennusteen laatijalla ei ole aineiston ulkopuolista informaatiota siitä, onko regressiokertoimen jakauma tutkimusaineiston jälkeisessä tulevaisuudessa, eli vuoden 2011 jälkeen, todennäköisemmin vasemmalla vai oikealla (pienempi vai suurempi) suhteessa tutkimusaineistosta estimoituun jakaumaan, aineiston ulkopuolisten tekijöiden vaikutusten jakauman odotusarvoksi voidaan määrittää nolla. Tällöin edellä esitetyn perusteella $\epsilon \sim T(4, 0, \sigma_\epsilon)$.

2.3.4.3 Havaintoaineiston ulkopuolisten tekijöiden vaikutus ja uusien maahanmuuttajien optimaalinen valinta

Käsittelen aineiston ulkopuolisten tekijöiden vaikutuksia ja niihin liittyviä epävarmuuksia vielä luvussa 12.3. Määritän luvussa ehdon sille, että uusien maahanmuuttajien elinkaarivaikutusten ennusteiden jakaumien odotusarvot muuttuvat, kun elinkaariennusteissa otetaan aineiston ulkopuolisten tekijöiden vaikutus huomioon. Määritän lisäksi ehdon sille, että julkisen talouden kannalta optimaalinen päätös uuden potentiaalisen maahanmuuttajan valitsemisesta Suomeen muuttavaksi *muuttuu*, kun aineiston ulkopuolisten tekijöiden vaikutus otetaan huomioon.

Osoitan myös, että mikäli elinkaariennusteen laatijalla on sellaista informaatiota aineiston ulkopuolisten tekijöiden vaikutuksesta, että ne lisäävät malliparametreihin $\boldsymbol{\theta}$ liittyviä epävarmuuksia – mutta ei sellaista informaatiota, joka kertoisi, mihin suuntaan malliparametrien jakaumat siirtyvät aineiston ulkopuolisten tekijöiden vaikutuksesta – aineiston ulkopuolisten tekijöiden vaikutuksilla ja niihin liittyvillä epävarmuuksilla ei ole vaikutusta uuden maahanmuuttajan elinkaarivaikutuksen jakauman odotusarvoon ja siten uusien maahanmuuttajien optimaaliseen valintaan.

Osa II

Elinkaarivaikutusten mallinnus ja ennustaminen uusille maahanmuuttajille

Luku 3

Olotilamallinnus

3.1 Olotilamallinnuksen tavoitteet ja rakenne

3.1.1 Olotilamallinnuksen tavoite

Henkilöllä voi olla tutkimuksen viitekehyksessä kolme vaihtoehtoista tilaa: hän voi olla joko Suomessa, ulkomailla tai kuollut. Olotilamallinnuksen tavoitteena on mallintaa Suomeen muuttavien henkilöiden suomessaoloaika heidän koko elinikänsä aikana. Mallinnustuloksia käytetään uskottavien olotilaennusteiden laatimiseen elinkaariennusteissa. Ennusteiden uskottavuuden varmennan sillä, että olotilamallien tuottamilla ennusteilla voidaan toistaa alkuperäinen olotila-aineisto.

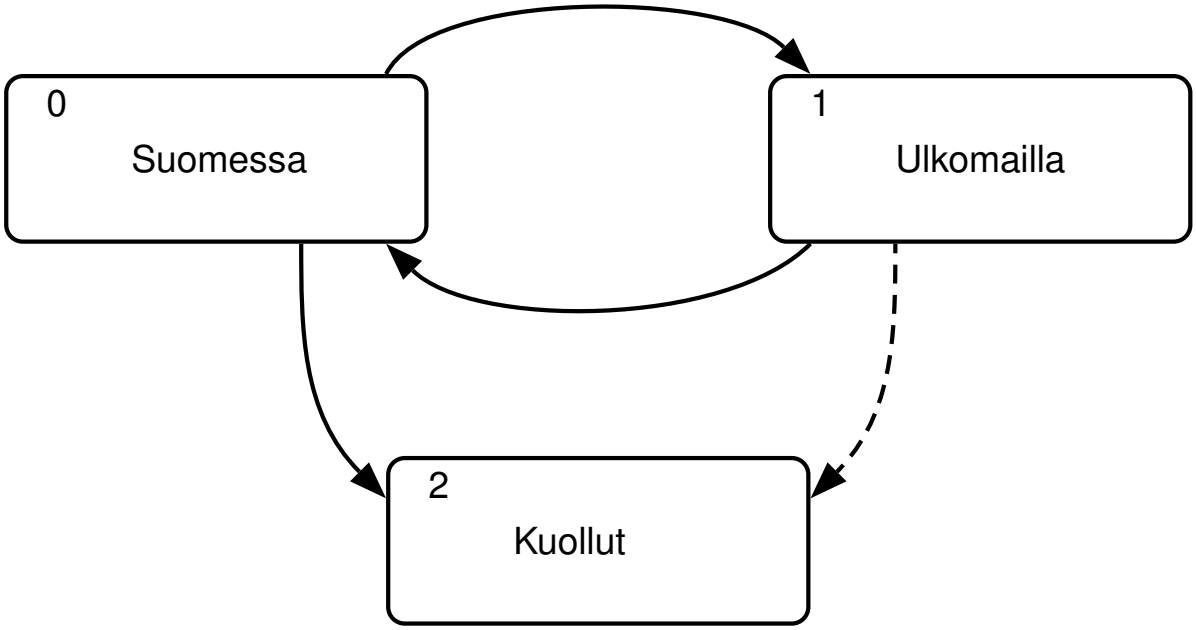
3.1.2 Olotilojen aineisto ja olotilasiirtymät

Maassaoloaika ja suomessaoloaika tarkoittavat tässä tutkimuksessa samaa asiaa. Olen määritellyt suomessaoloajan aiemmin tutkimuksessa (ks. tutkimuksen I osa: *3 Käsitteet ja määritelmät: Maassaolovuodet.*). Lyhyesti esitettynä henkilö on Suomessa, mikäli häneltä tai hänen mahdolliselta puolisoltaan löytyy merkintöjä tutkimuskäytössä olevista rekistereistä (tulonsiirrot, terveyspalvelut, opiskelu, ym.). Henkilö on puolestaan muuttanut pois Suomesta vuonna t, mikäli hän ei täytä yllä olevaa suomessaolon kriteeriä vuosina $t + 1$ ja $t + 2$. Likimain samaa maastamuuton määritelmää käyttää myös Sarvimäki (2011) (s. 668).

Laskin 15–70-vuotiaana Suomessa vietetyn maassaoloajan jokaiselle tutkimusaineiston henkilölle vuosien 1988–2009 aikana. Tutkimuskäytössäni oleva rekisteriaineisto sisältää tiedot myös vuosien 2010 ja 2011 osalta, mutta johtuen maastamuuton määritelmästä näiden vuosien osalta ei ole mahdollista tietää varmasti, ovatko henkilöt olleet Suomessa kyseisinä vuosina.

Suomessaolon tila on voinut aineistossa loppua kahteen tapahtumaan: joko henkilön muuttoon pois Suomesta tai henkilön kuolemaan. Mahdolliset tilat on esitetty alla kuviossa 3.1 Suomessa olon tila

on kuviossa 0-tila eli alkutila, koska kiinnostuksen kohteena oleva aika eli maassaoloaika alkaa siitä, kun henkilö on määritelmän mukaisesti Suomessa ensimmäisen kerran vähintään yhden kalenterivuoden aikana. Ulkomaillaolon tilassa henkilö ei ole Suomessa edellä lyhyesti esitetyn määritelmän mukaan. Hän ei ole tällöin myöskään kuollut. Mikäli henkilö muuttaa Suomesta ulkomaille ja taas takaisin Suomeen, maassaoloaika alkaa jälleen nollasta. Kuoleman olotila on nk. *absorboiva* tila. Yleisen markovilaisiin prosesseihin liittyvän määritelmän mukaan todennäköisyys pysyä absorboivassa tilassa kaikkina tulevina ajanhetkinä on yksi sen jälkeen, kun absorboivaan tilaan on siirrytty.



Kuva 3.1: *Olotilat ja olotilasiirtymät. Suomessa olon tila on alkutila. (Olotilasiirtymä 1 → 2 on merkitty katkoviivalla, koska havaintoja siirtymästä ei ole aineistossa.).*

Sekä tutkimusaineistossa että jäljempänä laadittavissa olotilamalleissa ja -ennusteissa mahdollisia siirtymiä olotilasta toiseen on neljä kappaletta: suomessaolon tilasta siirtyminen ulkomaillaolon tilaan (siirtymä $0 \rightarrow 1$), ulkomaillaolon tilasta takaisin suomessaolon tilaan (siirtymä $1 \rightarrow 0$), kuoleminen, kun on Suomessa (siirtymä $0 \rightarrow 2$) sekä kuoleminen, kun on ulkomailla (siirtymä $1 \rightarrow 2$). Viimeisen siirtymän olen merkinnyt kuvioon 3.1 katkoviivalla, sillä aineisto ei sisällä tietoja kyseisestä siirtymästä. Tärkeimpänä mielenkiinnon kohteena on Suomesta ulkomaille muuton (siirtymä $0 \rightarrow 1$) mallinnus, sillä suurin osa Suomesta poismuuttaneista ei enää palaa takaisin Suomeen, kuten jäljempänä esitettävät tulokset osoittavat. Kuolemat puolestaan tapahtuvat pääasiassa iässä, jossa ei ole suurta poismuuttaneisuutta yli 60-vuotiaiden poismuuttaneisuuden perusteella. Koska tutkimusaineiston seuranta-aika ulottuu vain 70 ikävuoteen asti, oletan, ettei poismuuttaneisuus kasva yli 70-vuotiailla suhteessa yli 60-vuotiaisiin.

Suomessa syntyneiden osalta en mallintanut lainkaan Suomesta ulkomaille muuttamista, sillä poismuuttaneita oli aineistossa hyvin vähän. Suomesta poismuuttaneita oli aineistossa 22 vuoden, eli maksimiseurantaajan, jälkeen vain noin 3 %:ia. Suomessa syntyneiden osalta poismuuttaneisuus oli suurinta niiden henkilöiden kohdalla, joita oli alettu seuraamaan heidän ollessaan 20–30-vuotiaita. Suomessa syntyneiden

osalta poismuuttaneisuus oli koko seuranta-ajan suunnilleen yhtä suurta, noin 0,1–0,2 %:ia väestöstä vuodessa, mikä vastaa likimäärin Suomen virallisen tilaston lukuja koko Suomen väestön 15–70-vuotiaiden poismuuttaneisuudesta vuosina 1990–2011 (SVT: *muuttoliike*).

Alle 15-vuotiaiden henkilöiden olotiloja en käsitellyt olotilamallinnuksessa lainkaan.

Oikeanpuoleinen sensorointi olotilojen aineistossa

Niin kutsuttu *oikeanpuoleinen sensorointi* esiintyy aineistossa, kun henkilö joko oli jo täyttänyt 71 vuotta tai ei ollut vielä kuollut vuoden 2009 loppuun mennessä. Ulkomaille muuttaneiden osalta jälkimmäisestä eli kuolemisesta ei ole tietoa, ja oletan tämän vuoksi, että kuolleisuus on yhtä suurta ulkomailla kuin Suomessa. Tällä ei ole suurta vaikutusta laadittaviin olotilaennusteisiin, sillä suurin osa ulkomaille muuttaneista ei palaa enää takaisin ja kuolleisuus aikuisiässä alle 71-vuotiailla ei ole normaalioloissa suurta ainakaan suurimmassa osassa maailman maita.

Mikäli henkilö on aineistossa sensoroitu vuoden 2009 lopussa, merkitsin henkilön sensoroiduksi vuonna 2010, jotta nämä tapaukset erottuisivat vuonna 2009 tapahtuneista mahdollisista olotilojen siirtymistä. Yhden kalenterivuoden aikana henkilö voi joko pysyä samassa olotilassa, missä oli vuoden alussa, tai sitten hänen olotilansa voi muuttua kerran. Kaikki aineistossa tapahtuvat kuolemat ovat Suomessa olon aikana tapahtuneita, sillä ulkomailla tapahtuneista kuolemista ei ole tietoa.

Oletan jäljempänä olotilakohtaisissa mallinnuksissa sensoroinnin olevan *riippumatonta*, mikä tarkoittaa sitä, että sensorointi ei ole yhteydessä olotilasiirtymien todennäköisyyksiin. Erityisen tärkeä tämä oletus on pääasiallisen mielenkiinnon kohteen eli Suomesta poismuuton osalta. Pidän tätä oletusta hyvin realistisena, sillä se tarkoittaa, ettei vuodella 2010 eikä 71-vuotiaaksi tulemisella ole erityistä yhteyttä Suomesta muuttamisen todennäköisyyden kanssa.

Olotilahavaintojen jakaumien epäjatkuvuus aineistossa

Suomessa olon ajan minimi toteutuu aineistossa silloin, kun henkilö on sekä muuttanut Suomeen että muuttanut pois Suomesta yhden kalenterivuoden aikana. Koska aineisto ei sisällä tarkkoja muuttopäivämääriä, olen tässä tapauksessa määrittänyt henkilön maassaoloajaksi 0,5 vuotta. Kun henkilö on ollut Suomessa kahtena kalenterivuotena ilman poismuuttoa, henkilön maassaoloaika on tuolta ajalta yhteensä 1,5 vuotta. Tällä tavoin laskettuna maksimiaika Suomessa vietettyinä vuosina on aineistossa 22,5 vuotta. Tämä koskee niitä henkilöitä, jotka ovat joko jo asuneet Suomessa tai muuttaneet Suomeen vuonna 1988 ja ovat siitä lähtien asuneet ainoastaan Suomessa vuoden 2009 loppuun asti siten, etteivät he ole tuona aikana kuolleet tai täyttäneet 71 vuotta.

Sekä suomessaoloaikojen että ulkomaillaoloaikojen jakaumat olivat aineistossa näin ollen diskreettejä eli epäjatkuvia. Käsittelen kuitenkin mallinnuksissa näitä jakaumia ajassa jatkuvina, mitä ne myös todellisuudessa ovat. Tämä ei aiheuta lopullisiin suomessaoloajan ja ulkomaillaoloajan ennusteisiin käytännössä muutoksia, sillä ennustetulokset suomessaoloajan mallinnuksesta olivat käytännössä muuttumattomat, kun kokeilin mallintaa ”todellista” maassaoloaikojen jakaumaa, jonka loin edellä esitellystä havaintoaineistosta lisäämällä tasajakaumasta $\mathcal{U}(-0{,}5; 0{,}5)$ diskreetteihin maassaoloaikoihin sopivan satunnaisen osan. Tasajakauman parametriarvot seuraavat siitä oletuksesta, jota käytän epäjatkuvan aineiston kans-

sa, että muutot tapahtuvat keskellä kalenterivuotta. Lisäämällä määritelty satunnainen osa ne tapahtuvat pitkin koko kalenterivuotta. Tässä yksinkertaisessa imputoinnissa oletuksena oli, että maastamuuton todennäköisyys kalenterivuoden sisällä on samansuuruinen läpi koko vuoden ja että tämä todennäköisyys on riippumaton mallin selittävistä muuttujista.

Vasemmanpuoleinen katkaisu (*engl. truncation*) olotilojen aineistossa

Niin kutsuttu vasemmanpuoleinen katkaisu esiintyy aineistossa niiden henkilöiden kohdalla, jotka ovat muuttaneet Suomeen ennen vuotta 1988. Ennen vuotta 1988 muuttaneiden osalta ei ole tietoa, milloin he olivat muuttaneet Suomeen, ja siten heidän osaltaan ei ole tietoa, kuinka kauan he todella ovat olleet Suomessa. Tällaisten henkilöiden muuttovuodeksi määritin aineistossa (virheellisesti) vuoden 1988. Eniten tällä ratkaisulla on vaikutusta Ruotsissa ja Saksassa syntyneiden olotilamallinnukseen ja -ennustamiseen, sillä näissä maahanmuuttajaryhmissä ennen vuotta 1988 Suomeen muuttaneita henkilöitä oli eniten. Koska yleisesti ottaen koko tutkimusaineistossa poismuuton todennäköisyys laskee, kun maassaoloaika kasvaa – ainakin populaatiotasolla – vuonna 1988 muuttaneiksi määritettyjen osalta poismuuttaneisuus on maassaoloajan suhteen havaintoaineistossa pienempää, kuin mitä se todellisuudessa oli. Tämän seurauksena erityisesti Ruotsin ja Saksan maassaolomallien kohdalla vuonna 1988 muuttaneita koskeviin tuloksiin ei voi luottaa.

Toisaalta esimerkiksi Irakissa ja Somaliassa syntyneiden kohdalla ratkaisulla ei ole vaikutusta olotilamallinnukseen ja niistä tehtäviin ennusteisiin, sillä lähes kaikki näissä maissa syntyneet ovat muuttaneet Suomeen vuonna 1988 tai sen jälkeen.

Suomessaoloajan, ulkomaillaoloajan sekä kuolemisen todennäköisyyden mallinnus

Tavoitteena olotilamallinnuksessa on mallintaa edellä kuviossa 3.1 esitettyjen neljän siirtymän ehdolliset todennäköisyydet. Ehdollistajina ovat tiedot henkilöistä silloin, kun he muuttavat ensimmäisen kerran Suomeen. Palaan näihin olotilojen muutoksia ”selittäviin” tekijöihin jäljempänä, kun käsittelen suomessaolo- ja ulkomaillaolomallien selittäviä muuttujia.

Edellä kuvion 3.1 yhteydessä määritettiin olotilat $\{0, 1, 2\}$, missä 0 on Suomessa oleminen, 1 ulkomailla oleminen ja 2 kuolleena oleminen. Siirtymiä merkittiin nuolella siten, että esimerkiksi siirtymä $0 \rightarrow 1$ tarkoittaa siirtymää Suomesta ulkomaille ja $0 \rightarrow 0$ puolestaan ”siirtymänä”’ tarkoittaa Suomessa olemisen jatkumista. Yleisesti ottaen nuolen vasemmalla oleva numero kertoo henkilön olotilan jokaisen kalenterivuoden alussa seuranta-aikana ja nuolen oikealla puolella oleva numero olotilan jokaisen vuoden lopussa.

Merkitään nyt todennäköisyyttä siirtyä olotilasta toiseen yhden kalenterivuoden aikana: $P(\text{olotila(alku)} \rightarrow \text{olotila(loppu)})$, missä nuolen vasemmalla puolella oleva olotila on olotila vuoden alussa ja olotila nuolen oikealla puolella puolestaan olotila vuoden lopussa. Tällöin esimerkiksi todennäköisyys siirtyä Suomesta ulkomaille yhden kalenterivuoden aikana on $P(0 \rightarrow 1)$ ja todennäköisyys pysyä Suomessa on puolestaan $P(0 \rightarrow 0)$. Todennäköisyys pysyä Suomessa voidaan ilmaista myös: $1 - (P(0 \rightarrow 1) + P(0 \rightarrow 2))$, sillä jollei henkilö muuta ulkomaille ($0 \rightarrow 1$) tai kuole Suomessa ollessaan ($0 \rightarrow 2$), hän pysyy edelleen Suomessa.

Kaikkien siirtymien todennäköisyydet voidaan esittää näin matriisimuodossa, missä jokaisen matriisin $\boldsymbol{P}$ rivin todennäköisyyksien summa on 1:

$$\boldsymbol{P} = \begin{bmatrix} P(0 \to 0) & P(0 \to 1) & P(0 \to 2) \\ P(1 \to 0) & P(1 \to 1) & P(1 \to 2) \\ 0 & 0 & 1 \end{bmatrix}$$

Koska kuolema on absorboiva olotila, siitä poissiirtymisen todennäköisyydet muihin olotiloihin ovat nollia. Käytännössä mallinnan siirtymätodennäköisyydet erikseen koskien suomessaoloa, ulkomaillaoloa ja kuolemista. Laadin aluksi mallin suomessaoloajalle. Esitän yksityiskohtaisesti, kuinka päädyin käyttämääni malliin, kuinka estimoin sen ja miten sen tuloksia tulkitaan. Tämän jälkeen sovellan käytännössä samoja periaatteita ulkomaillaoloajan mallinnukseen.

Oletan, että suomessaoloaika on riippumaton mahdollisesta ulkomaillaoloajasta. Voi olla, ettei oletus pidä paikkaansa, mutta sillä ei ole suurta merkitystä, koska vain harvalla henkilöllä aineistossa on useampi kuin yksi muutto pois Suomesta. Tutkimusaineisto sisälsi noin 155 000 henkilöä 10 suurimman syntymävaltion perusteella määritetyn maahanmuuttajaryhmän osalta. Heistä noin 23 000:lla oli yksi muutto pois Suomesta, mutta henkilöitä, joilla oli 2 muuttoa pois Suomesta, oli vain 1 375, eli alle yksi prosentti koko aineiston henkilömäärästä. Edelleen henkilöitä, joilla oli 3 muuttoa pois Suomesta, oli 108 henkilöä, ja yli 3 poismuuton henkilöitä oli 10 henkilöä. Ulkomaillaoloajan annan riippua Suomessa kumulatiivisesti vietetystä ajasta.

Kuolemistodennäköisyyksien osalta oletan, että ne ovat riippumattomia muuttotodennäköisyyksistä. Tämä oletus ei juurikaan vaikuta lopputuloksiin – vaikkei se pitäisikään paikkaansa – koska kuolleisuus on alle 70-vuotiailla suhteellisen pientä.

3.2 Suomessaoloaikojen mallinnus

Maassaoloaikojen mallinnuksen kuvaus alkaa siitä, että esittelen lyhyesti, mihin tilastotieteen osa-alueeseen mallinnus luontevasti pohjautuu, kun otetaan huomioon edellä esitetty aineiston rakenne ja mallinnustavoitteen luonne. Tämä sisältää toistoa yllä esitetystä. Kun olen esittänyt sopivan tilastotieteen osa-alueen, perustelen millaiseen mallinnusluokkaan perustan maassaoloajan (ja siten ulkomaillaoloajan) mallinnukseen. Esitän sitten tarkemmin, mitä estimoitavia parametreja lopullisiin maahanmuuttajaryhmäkohtaisiin maassaolomalleihin kuuluu ja mikä niiden tulkinta on. Lopuksi esitän menetelmän sille, kuinka valitun mallinnusluokan mallit estimoidaan sekä kuinka tarkistaa ja arvioida estimoitujen mallien hyvyyttä vertaamalla niiden tuottamia ennusteita alkuperäisiin maahanmuuttajaryhmäkohtaisiin havaintoaineistoihin.

3.2.1 Selviytymisaika- ja tapahtumahistoria-analyysi

Edellä kuviossa 3.1 on esitetty olotilat ja mahdolliset siirtymät olotilojen välillä. Tietyssä tilassa kulunutta aikaa sekä siirtymätodennäköisyyksiä eri tilojen välillä mallinnetaan muun muassa lääketieteessä ja yhteiskuntatieteissä *selviytymisaika- ja tapahtumahistoria-analyysiin* (*engl. failure time and event history analysis*) perustuvilla tilastollisilla malleilla. Selviytymisaika-analyysista on karkeasti ottaen kyse silloin, kun mallinnetaan vain yhdessä tilassa olemisen aikaa sekä siitä poissiirtymisen todennäköisyyttä. Lääketieteessä tämä voi tarkoittaa esimerkiksi jäljellä olevaa elinaikaa siitä hetkestä, kun on saanut kuolemaan johtavan sairauden diagnoosin, siis kirjaimellisesti selviytymisaikaa. Tapahtumahistoria-analyysissa puolestaan mahdollisia mallinnettavia siirtymiä eri tilojen välillä on enemmän kuin yksi: esimerkiksi sairauden diagnoosin jälkeen tauti voi johtaa joko kuolemaan tai sairaudesta paranemiseen.

Tässä tutkimuksessa ”selviytymisaika” tietyssä olotilassa tarkoittaa joko:

1. Suomessa vietettyä aikaa vuosina (tila 0) ennen poismuuttoa (tila 1) tai kuolemaa (tila 2).

tai

2. Ulkomailla vietettyä aikaa vuosina (tila 1) ennen takaisin Suomeen muuttamista (tila 0) tai kuolemaa (tila 2).

Kuolema on niin kutsutta *absorboiva* tila, sillä siitä ei enää siirrytä muihin tiloihin. Tämän tutkimuksen tilamalli vastaa skemaattisesti lääketieteen *sairaus–kuolema-mallia parantumisella* (*engl. illness-death with recovery*, ks. esim. Andersen and Keiding (2002), 95). Tällöin suomessaolo vastaa skemaattisesti terveenä olemista ja sairastaminen ulkomaillaoloa. Olotilasiirtymissä sairastuminen vastaa ulkomaille muuttoa ja paraneminen puolestaan vastaa ulkomailta takaisin Suomeen muuttoa. Kaikista tiloista on aina mahdollista kuolla.

Käytettävissä olevassa ja edellä esitellyssä havaintoaineistossa suurin osa maassaoloajoista ei pääty Suomesta poismuuttoon tai kuolemaan ennen kuin seuranta-aika aineistossa loppuu. Tämä niin kutsuttu *oikea sensorointi* (*engl. right censoring*) tapahtuu havaintoaineistossa, kun kuolemaa ei ole tapahtunut vuoden 2009 loppuun mennessä tai henkilö on täyttänyt 71 vuotta seuranta-aikana. Käytännössä käsitän myös kuolemat sensorointitapauksina, kun laadin maassaoloajan ja ulkomaillaoloajan mallit. Tällöin sensorointivuosi on kuolinvuosi. *Vasen katkaisu* (*engl. left truncation*) aineistossa koskee ennen vuotta 1988 Suomeen muuttaneita: heidän todellisesta maassaoloajastaan ei ole varmaa tietoa, kuten edellä esitin.

Selviytymisajan mallinnnukseen ei käytetä yhteiskuntatieteissä yleisesti käytössä olevia lineaarisia ja logistisia regressiomalleja, sillä ne eivät ota huomioon sensorointia. Lisäksi logistisella mallilla voidaan mallintaa vain sitä, onko henkilö jossakin (olo)tilassa – tai ei ole – mutta ei kiinnostuksen kohteena olevaa selviytymisaikaa. Lineaarisessa mallissa puolestaan oletetaan selviytymisajan jakaumaksi normaalijakauma, ja tämä oletus on väärä: mahdollisten selviytymisaikojen joukko sisältää vain positiiviset reaaliluvut. Lisäksi selviytymisaikojen funktio ajan suhteen on monotonisesti vähenevä, joten normaalijakauma kuvaa sitä huonosti. Tästä syystä selviytymisaikojen mallinnus perustuu menetelmiin, joissa niiden yhteistodennäköisyysjakaumassa otetaan huomioon sensorointi (ja katkaisu) ja jakaumaoletuksena on esimerkiksi eksponentti-, weibull- tai gammajakauma, sillä ne sopivat (sopivin parametrivalinnoin) muodoltaan kuvaamaan selviytymisaikojen jakaumaa.

Yllä lyhyesti esitellyssä tavanomaisessa selviytymisaikojen mallinnuksessa oletuksena on, että todennäköisyys sille, että tutkimuksen kohteena oleva tilasiirtyminen on tapahtunut ajankohtaan T mennessä, lähestyy yhtä, kun seuranta-aikaa t kasvatetaan rajatta (ks. esim.Kalbfleisch 2002, 6). Toisin sanoen selviytymistodennäköisyys lähestyy nollaa seuranta-ajan kasvaessa rajatta. Tämän tutkimuksen viitekehyksessä tällainen mallinnusoletus tarkoittaisi sitä, että kaikkien henkilöiden oletettaisiin muuttavan pois Suomesta, kun maassaoloaika kasvaa tarpeeksi. Yksikään henkilö ei jäisi Suomeen pysyvästi sellaisessa skenaariossa, jossa kuolemia ei tapahtuisi.

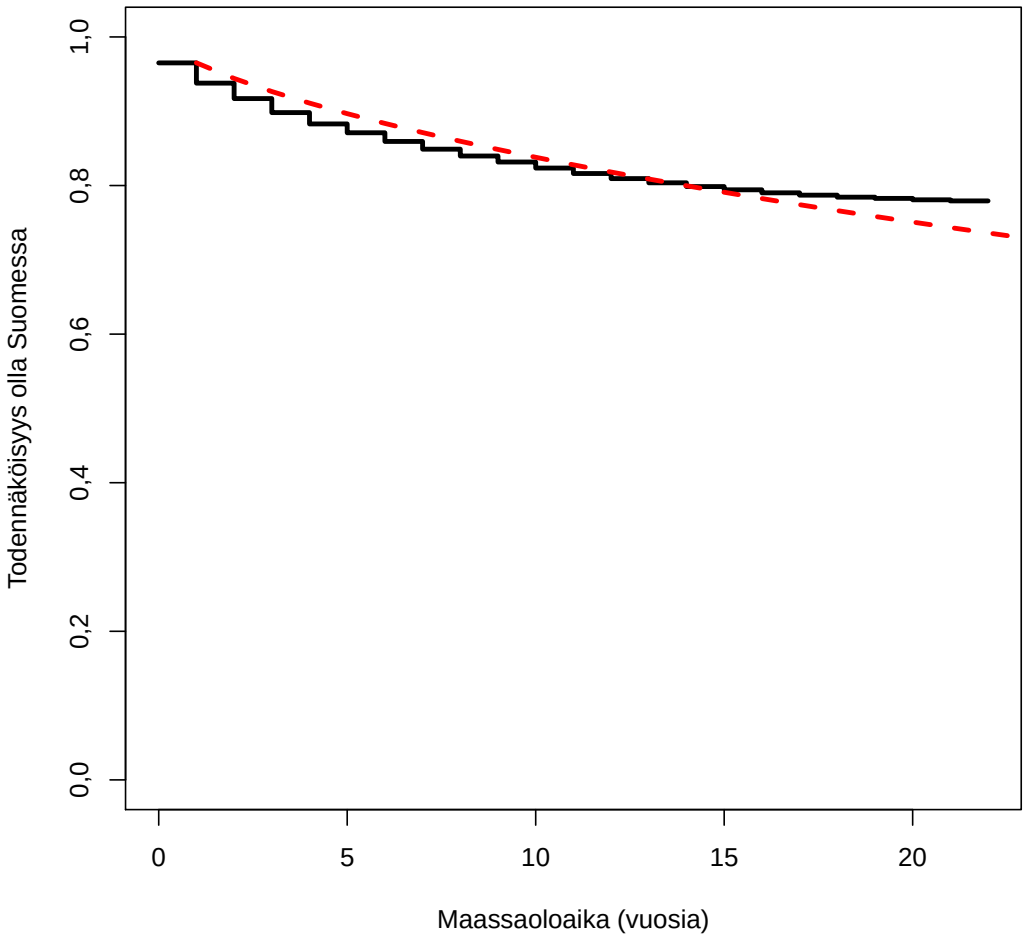
Kaplan-Meier-estimaatit havaintoaineistolle

Niin kutsuttu *Kaplan-Meier-estimaattori* (*lyh. K-M-estimaattori*) estimoi populaation selviytymistodennäköisyyden funktiota seuranta-ajan suhteen aineiston havaintojen perusteella (ks. esim. Aalen et al. (2008), 90–94). Havaintoja ovat toteutuneet siirtymät sekä sensoroinnit, joiden tapahtuma-ajat tiedetään. K-M-estimaattori jakaa seuranta-ajan pieniin aika-intervalleihin. Intervallikohtainen K-M-estimaattorin arvo riippuu siitä, kuinka moni aineiston havainnoista, jotka intervallin alussa kuuluivat tilasiirtymää koskevaan riskiväestöön, eli eivät olleet vielä siirtyneet eivätkä olleet tulleet sensoroiduiksi, eivät ”selviytyneet” intervallin yli siirtymättä toiseen tilaan. K-M-estimaattorin arvo seuranta-ajan hetkellä t on kaikkien niiden intervallikohtaisten arvojen tulo, jossa viimeinen intervalli päättyy seuranta-ajan hetkellä t.

Tässä viitekehyksessä seuranta-aika on maassaoloaika vuosina, ja selviytymisfunktio kertoo, kuinka todennäköistä on olla vielä Suomessa, kun Suomeen muutosta on kulunut maassaoloajan verran vuosia. Alla kuviossa 3.2 on esitetty K-M-estimaatit havaintopopulaatiolle (musta viiva), joka koostuu 10 suurimman maahanmuuttajaryhmän 15–70-vuotiaista väestöistä yhteensä. Kuviossa 3.2 on myös usein käytettyyn weibull-jakaumaoletukseen perustuva estimoitu sovite selviytymisfunktiolle (punainen katkoviiva).

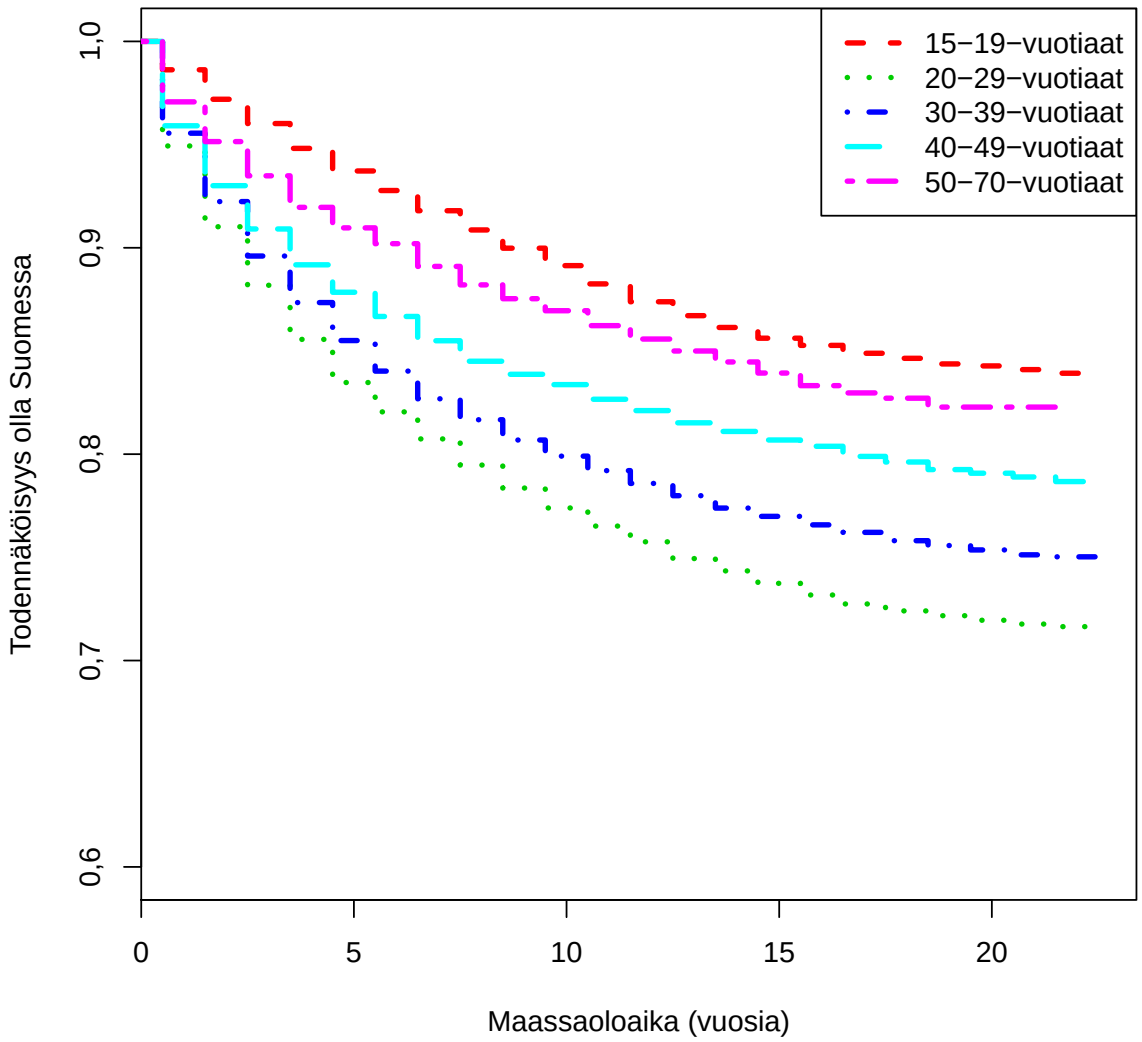
Kuviosta 3.2 selviää, että todennäköisyys olla Suomessa pienenee nopeiten ensimmäisten maassaolovuoisen aikana, mikä ilmenee siinä, että K-M-estimaatin ”askelmat” ovat aluksi suurempia. Ne kuitenkin pienenevät siten, että 20 vuoden maassaolovuoden kohdalla todennäköisyys olla Suomessa ei vaikuta juuri enää pienenevän. Todennäköisyys olla Suomessa vaikuttaa ”tasoittuvan” noin 80 %:n kohdalle tai hieman sen alle.

Kokeilin mallintaa selviytymisaikaa – eli todennäköisyyttä olla vielä Suomessa tietyn maassaolovuosien jälkeen – varsin yleisesti käytössä olevalla weibull-jakaumaan perustuvalla estimoinnilla. Estimoinnin tulos on piirretty kuvioon 3.2 punaisella katkoviivalla. Tavanomaisen selviytymisaikamallin oletuksiin perustuva estimaatti poikkeaa K-M-estimaateista erityisesti 15 maassaolovuoden jälkeen: sovitteen mukaan todennäköisyys olla Suomessa on pienempi kuin, mitä K-M-estimaatit osoittavat. Lisäksi weibull-jakaumaan perustuva estimaatin kuvaajan kulmakerroin ei pienene maassaoloajan kasvaessa toisin kuin K-M-estimaattien kuvaaja. Tämä johtuu siitä, että weibull-jakaumaan perustuva estimaatti perustuu oletukseen, että selviytymisfunktio lähestyy nollaa, kun maassaoloaikaa kasvatetaan rajatta. K-M-estimaattien perusteella vaikuttaa kuitenkin siltä, että selviytymisfunktio lähestyy arvoa 0,80. Toisin sanoen kuvion perusteella näyttää siltä, että noin 80 %:ia kymmeneen suurimpaan maahanmuuttajaryhmään kuuluvista henkilöistä jää pysyvästi Suomeen.



Kuva 3.2: *Kaplan-Meier-estimaatit maassaoloajan mukaan 10 suurimman maahanmuuttajaryhmän muodostamalle väestölle (musta viiva). Weibull-jakaumaan perustuva estimaatti selviytymisfunktiolle (punainen katkoviiva).*

Alla kuviossa 3.3 K-M-estimaatit on esitetty maahanmuuttoikäluokittain havaintopopulaatiolle, joka koostuu 10 suurimman maahanmuuttajaryhmän 15–70-vuotiaista väestöistä yhteensä. Esimerkiksi 30–39-vuotiaiden ryhmä koostuu niistä henkilöistä, jotka ovat muuttaneet Suomeen, kun he ovat olleet 30–39-vuotiaita. Nuorimpaan 15–19-vuotiaiden ikäluokkaan kuuluvat ne henkilöt, jotka ovat muuttaneet alun perin alle 15-vuotiaina Suomeen.



Kuva 3.3: *Kaplan-Meier-estimaatit maassaoloajan ja maahanmuuttoikäluokan mukaan 10 suurimman maahanmuuttajaryhmän muodostamalle väestölle. Pystyakseli on katkaistu kuviossa, jotta kuvion tarkastelu olisi helpompaa.*

Kuviosta 3.3 ilmenee, että kaikissa maahanmuuttoiän luokissa selviytymisfunktion kulmakerroin lähenee kohti nollaa noin 20 maassaolovuoden jälkeen. Kuvion perusteella kaikissa maahanmuuttajaikäryhmissä henkilöt eivät enää muuta (melkein lainkaan) pois Suomesta, kun he ovat olleet Suomessa yhtäjaksoisesti vähintään 20 vuoden ajan. Kuvion perusteella nuorimmasta (15–19-vuotiaat) ja vanhimmasta (50–70-vuotiaat) maahanmuuttoikäluokasta jää suurempi osuus henkilöistä pysyvästi Suomeen kuin näiden ikä-

luokkien väliin jäävistä muista ikäluokista. Pienin osuus Suomeen pysyvästi jääviä on 20–29-vuotiaana Suomeen muuttaneiden ikäluokassa: noin 70 %:ia kaikista Suomeen muuttaneista.

En selvitä tässä tutkimuksessa sitä, miksi osa maahanmuuttajapopulaatiosta jää pysyvästi Suomeen, enkä sitä, miksi jotkut poismuuttavista muuttavat pois nopeammin kuin toiset poismuuttajat. Mainitsen kuitenkin lyhyesti Bodvarsson and Van den Berg (2013) listan *maahanjääntitekijöistä* (*engl. stay factors*) (s. 6). Maahanjäännin tekijöitä ovat pääasiassa sosiaaliset ja taloudelliset siteet maahan kuten ystävyyssuhteet, perhesiteet ja työllisyys. Kenties nämä lisääntyvät ainakin populaatiotasolla maassaoloajan kasvaessa ja siten selittävät sitä, että poismuuttaneisuus laskee lähelle nollaa maassaoloajan kasvaessa.

Yksilötasolla maahan jääminen ja poismuuttaneisuus voivat vaihdella populaatiotasosta poiketen myös silloin, kun mallin selittävät muuttujat otetaan huomioon. Tätä havaitsemattomista tekijöistä johtuvaa vaihtelua (*engl. frailty*) käsittelen erikseen mallinvalinnassa.

Suomeen pysyvästi jääminen

Lähden rakentamaan maassaolomallia siitä oletuksesta käsin, että kaikista maahanmuuttajaryhmistä tietty osuus jää pysyvästi Suomeen. Perustelen tätä esitetyillä selviytymisfunktion K-M-estimaattien kulmakertoimilla, jotka kaikkien luokkien (esimerkiksi maahanmuuttajaryhmä) mukaan jaoteltuina lähestyvät nollaa noin 20 maassaolovuoteen mennessä. Lisäksi pidän uskottavana sitä, että edellä esitetyt maahanjääntitekijät, eli sosiaaliset ja taloudelliset siteet Suomeen, lisääntyvät ainakin populaatiotasolla, kun maassaoloaika kasvaa.

Oletuksena on myös, että Suomesta poismuuttaneisuus ei enää kasva populaatiotasolla nollan läheisyydestä sen jälkeen, kun maassaolovuosia on kulunut yli aineiston maksimiseuranta-ajan eli 22,5 maassaolovuotta. Tämä tarkoittaa sitä, että esimerkiksi 30 vuotta Suomessa olleilla maastamuutto ei enää ”ala” uudestaan, kun maastamuuttaneisuus on ehtinyt hidastua (aidosti vähenevästi) nollaan tai lähelle nollaa seuranta-ajan puitteissa.

Tämän oletuksen paikkansapitävyyteen viittaa vahvasti se, että K-M-estimaattien perusteella selviytymisfunktioiden kulmakertoimet lähestyvät kohti nollaa kaikissa maahanmuuttoikäryhmissä (kuten yllä kuviossa 3.3) kaikkien maahanmuuttajaryhmien osalta. Koska eri ikäisinä Suomeen muuttaneet vanhenevat samaa tahtia maassaoloajan kanssa, mainittujen K-M-estimaattien tulosten perusteella yhdessäkään maahanmuuttajaryhmässä ei ole olemassa mitään tiettyä ikää, jolloin maahanmuuttajat alkaisivat palata syntymävaltioihinsa. Mikäli oletus maahanjääntitekijöiden kasaantumisesta maassaolovuosien mukaan pitää myös paikkansa, yli tutkimuksen seuranta-ajan meneviä maassaolovuosilla maahanmuuttajapopulaatioissa on vain entistä enemmän maahanjääntitekijöitä.

Paranemisosuusmallit

Selviytymisaika- ja tapahtumahistoria-analyysin yhden osa-alueen muodostavat *paranemisosuusmallit* (*engl. cure (rate) models*) eli mallit, joissa on mukana *pitkän ajan selviytyjiä* (*engl. long-term survivors*). Nimitys tulee lääketieteestä, jossa seuranta-aineistojen avulla tutkitaan esimerkiksi selviytymisaikaa syöpäsairauksien yhteydessä. Tällaiset aineistot sisältävät merkittäviä osuuksia tutkittavista väestöistä, jotka parantuvat sairauksista, eivätkä siten kuole niihin, vaikka seuranta-aikaa kasvatettaisiin rajatta (Ibrahim

et al. 2001, 155).

Käytännössä tämä havaitaan esimerkiksi Kaplan-Meier-estimaattorista, jonka estimaattien havaitaan ”tasoittuvan” johonkin nollaa suurempaan tasoon (Aalen et al. 2008, 245). Näin tapahtui edellä maassaoloajan todennäköisyyksien K-M-estimaattien kuvioissa.

Kuten jäljempänä esitän alan menetelmäkirjallisuuteen edelleen viitaten, tilastotieteellisessä mallinnuksessa ”paraneminen” – tai tässä tapauksessa Suomeen pysyvästi jääminen – tarkoittaa sitä, että osalla tutkittavasta väestöstä ei seuranta-ajan alkaessa ole lainkaan riskiä kokea tutkittavaa tapahtumaa. Esimerkiksi maassaoloajan mallinnuksen kohdalla tämä tarkoittaa sitä, että osalla Suomeen juuri muuttaneesta väestöstä riski muuttaa pois Suomesta heidän elinikänsä aikana on nolla, kun riski poismuutttoon käsitetään todennäköisyytenä. Tämä ominaisuus, eli täysi riskittömyys siirtyä (alkuperäisestä) olotilasta pois, on useimmissa tutkimussovelluksissa – niin kuin tässäkin tutkimuksessa – latentti eli havaitsematon.

Keskeinen menetelmäkirjallisuus, jota käytän paranemisosuusmallien soveltamisessa maassaoloaikojen mallinnukseen, koostuu kirjoista *Bayesian Survival Analysis* (Ibrahim et al. 2001), *Survival and Event History Analysis* (Aalen et al. 2008), *Frailty Models in Survival Analysis* (Wienke 2011) sekä artikkelista *Bayesian Transformation Models for Multivariate Survival Data* (De Castro et al. 2014). Sovellan viimeksi mainitun artikkelin menetelmää maassaoloaikojen mallinnukseen. Mainittu kirjallisuus antaa tarkemmat matemaattiset määritelmät esimerkiksi selviytymisfunktiolle ja sen ominaisuuksille sekä K-M-estimaattorille.

Esitän seuraavaksi lyhyesti muutaman keskeisimmän yleisesti selvitymisaikamallinnukseen liittyvän funktion, joihin viitataan myös jäljempänä tekstissä. Nämä funktiot on esittänyt (hieman eri muodossa) esim. Aalen et al. 2008, 5–6. En paneudu esitettäviin funktioihin sen tarkemmin tässä tekstissä ja kehotankin näistä funktioista tai laajemmin aihepiiristä kiinnostunutta lukijaa tutustumaan tarvittaessa esimerkiksi tässä luvussa esitettyyn keskeisimpään kirjallisuuteen.

Selviytymisaikamallien keskeisimpien funktioiden määritelmiä

Olkoon ei-negatiivinen T selviytymisaikaa – tässä tapauksessa maassaoloaikaa – kuvaava satunnaismuuttuja. Olkoon tähän satunnaismuuttujaan liittyvä todennäköisyystiheysfunktio $f_T(t)$ ja sen kertymäfunktio $F_T(t)$.

Selviytymisajan (maassaoloajan) kertymäfunktio $F_T(t)$ esittää todennäköisyyttä, että tutkittava tapahtuma (maastamuutto) on tapahtunut aikaan t mennessä:

$$F_T(t) = P(T \leq t) = \int_0^t f_T(x)dx. \tag{3.1}$$

Tällöin *selviytymisfunktio* $S_T(t)$ on:

$$S_T(t) = P(T > t) = 1 - F_T(t) = \int_t^\infty f_T(x)dx. \tag{3.2}$$

Selviytymisajan T *hasardin funktio* (*engl. hazard function*) $\lambda_T(t)$ on:

$$\lambda_T(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \tag{3.3}$$

Hasardi kertoo, mikä on todennäköisyys sille, että tutkittava tapahtuma tapahtuu ajanhetkien t ja $t + dt$ muodostamalla (puoliavoimella) välillä $[t, t + dt)$, mikäli tapahtumaa ei ole vielä sattunut aikaan t mennessä, eli tiedetään, että $T \geq t$. Tässä dt on "pieni" lisäys aikaan t. Hasardi on tällöin tapahtuman todennäköisyys, kun annetaan pienen aikalisäyksen dt lähestyä kohti nollaa.

Hasardin funktio voidaan esittää myös tapahtuman todennäköisyystiheysfunktion ja selviytymisfunktion osamääränä:

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)}. \tag{3.4}$$

Tästä osamäärästä voidaan johtaa hasardin funktiolle myös pelkästään selviytymisfunktion sisältämä yhtälö, kun huomioidaan, että lausekkeesta 3.2 voidaan päätellä, että $-f_T(t)$ on selviytymisfunktion $S_T(t)$ derivaatta:

$$\lambda_T(t) = -\frac{d}{dt} log(S_T(t)). \tag{3.5}$$

Selviytymisfunktio voidaan nyt puolestaan esittää hasardin avulla:

$$S_T(t) = exp\left(-\int_0^t \lambda_T(x)dx\right). \tag{3.6}$$

Tässä sulkeiden sisällä oleva integraali on selviytymisajan (maassaoloajan) *kertyneen hasardin funktio* (*engl. cumulative hazard function*) $H_T(t)$:

$$H_T(t) = \int_0^t \lambda_T(x)dx. \tag{3.7}$$

Noudatan näiden keskeisempien selviytymisaikamallien funktioiden osalta jäljempänä tekstissä viittaamieni artikkeleiden ja muun menetelmäkirjallisuuden käyttämiä merkintätapoja. Tämä helpottanee ainakin niiden löytämistä ja tulkitsemista lähteenä olevasta kirjallisuudesta. On myös syytä huomioida, että paranemisosuusmallien osalta keskeisimmät määritelmät poikkeavat selviytymisaikamallien määritelmistä. Esimerkiksi selviytymisfunktion $S_T(t)$ arvo lähenee selviytymisaikamallinnuksessa aina nollaa, kun selviytymisajan T annetaan kasvaa rajatta. Paranemisosuusmalleissa tämä ei yleisesti ottaen pidä paikkaansa ainakaan koko tutkittavan populaation osalta, sillä osalle populaatiota tutkittava tapahtuma ei tapahdu koskaan, vaikka selviytymisajan annettaisiin kasvaa rajatta.

3.2.2 Paranemisosuusmallin luokan valinta

Esitän tässä luvussa perustelut sille, miksi mallinsin maassaoloajat käyttäen De Castro et al. (2014) kehittämää paranemisosuusmallien luokkaa. Tätä varten käyn lyhyesti läpi paranemisosuusmallien kehittämisen historiaa, sillä samalla voin tuoda esille niitä sopivia ominaisuuksia, jotka löytyvät mainitusta De Castro et al. (2014) artikkelin menetelmästä.

3.2.2.1 Bayesilainen viitekehys

Paranemisosuusmalleja on kehitetty sekä bayesilaisen että frekventistisen tilastollisen päättelyn piirissä. Maassaoloaikojen mallinnuksessa käytän bayesilaisen päättelyn piiriin kuuluvia tilastollisia malleja.

Ibrahim et al. (2001) mainitsevat joitakin yleisempiä, ei vain selviytymisaikamallinnukseen liittyviä, syitä käyttää bayesilaisia menetelmiä. Näitä ovat priori-informaation eli havaintoaineiston ulkopuolisen informaation mahdollinen käyttäminen, pienempi riippuvuus asymptoottisista tuloksista, ei-hierarkkisten mallien vertailu sekä puuttuvien havaintojen mallinnuksen käsitteellinen helppous. Erityisesti selviytymisaikamallinnukseen liittyviä syitä käyttää bayesilaisia menetelmiä ovat yleisesti vaikeasti estimoitavien selviytymysaikamallien estimointi MCMC-menetelmiä käyttäen sekä ennusteiden laatimisen helppous. Jälkimmäinen ominaisuus on hyödyllinen tässä tutkimuksessa, sillä käytän bayesilaisista malleista tehtäviä posterioriennusteita sekä mallinnustulosten diagnostiikkaan että varsinaisten uusien maahanmuuttajien elinkaariennusteiden simulointiin. (Ibrahim et al. 2001, 26–27).

Bayesilainen viitekehys mahdollistaa mallien estimointiin liittyvän epävarmuuden ottamisen huomioon tässä tutkimuksissa tehtävissä elinkaariennusteissa. Pelkkiin piste-estimaatteihin nojaavat ennusteet eivät sisällä tätä estimointiin liittyvää epävarmuutta, mikä tekee näistä ennusteista epärealistisempia, sillä mallin parametrien tarkkoja arvoja ei käytännössä tunneta varmasti mallin estimoinnin jälkeen. Bayesilaisen päättelyn käytön yleisiä perusteluita esittävät esimerkiksi Gelman et al. (2014) (s. 24–25) ja Robert (2007) (s. 507–518).

Paranemisosuusmalleja käsittelevästä frekventistisestä kirjallisuudesta mainitsen esimerkkeinä artikkelit Cook et al. (2002), Ma and Yin (2008) sekä Cai et al. (2013). Cook et al. (2002) on hyödyllistä luettavaa, koska siinä määritetään olotilojen välisen muutosintensiteetti- ja muutostodennäköisyysmatriisin (standardin) moniulotteisen paranemisosuusmallin tapauksessa. Cai et al. (2013) käsittelee sen kirjoittajien kehittämää R-ohjelman pakettia *smcure* (Cai et al. 2012), jolla voi sovittaa nk. semiparametrisia suhteellisen hasardin paranemisosuusmalleja. Käytän tämän tutkimuksen liitteessä A smcure-paketin mukana tulevaa aineistoa, jonka avulla havainnollistan käyttämäni De Castro et al. (2014) mallinnusluokan toimivuutta.

3.2.2.2 Standardi paranemisosuusmalli

Sumathi and Rao K. (2008) käyvät menetelmäkatsauksessaan läpi paranemisosuusmallien kehittämisen

keskeisimmät kehitysvaiheet heidän artikkelinsa ilmestymiseen asti. Perustan seuraavan katsauksen paranemisosuusmalleihin pitkälti heidän artikkeliinsa. Sekä Ibrahim et al. (2001) että artikkelit, joihin viittaan jäljempänä, esittävät myös lyhyesti paranemisosuusmallien keskeisimmät kehitysvaiheet.

Paranemisosuusmallit syntyivät Ibrahim et al. (2001):n mukaan artikkelien Boag (1949) sekä Berkson and Cage (1952) myötä. Niissä tutkittiin syöpäpotilaiden selviytymisaikaa siten, että osa syöpäpotilaista mallinnettiin parantuvaksi sairaudesta. Bergin ja Gagen malli on sittemmin tunnettu *standardina paranemisosuusmallina* (*engl. standard cure rate model*) tai *paranemissekoitusmallina* (*engl. cure mixture model*). Jälkimmäinen nimitys viittaa mallin rakenteeseen:

$$S(t) = \pi + (1 - \pi) \times S^*(t), \tag{3.8}$$

missä t on seuranta-aika, $S(t)$ on selviytymisfunktio koko populaatiossa, π osuus populaatiosta, joka parantuu, ja $S^*(t)$ selviytymisfunktio sille osuudelle populaatiosta, joka ei parannu. $S^*(t)$ on tavanomainen selviytymisfunktio, eli sen määrittämä selviytymistodennäköisyys lähenee nollaa, kun seuranta-ajan annetaan kasvaa rajatta. (Ibrahim et al. 2001, 155).

Standardin paranemisosuusmallin estimoinnissa ollaan kiinnostuneita selvittämään paranemisosuus populaatiossa, π, sekä selviytymisfunktion $S^*(t)$ muoto, joka voi noudattaa esimerkiksi eksponentti-, weibull- tai gammajakaumaa. Käytännön sovellukset voivat olla regressiomalleja, joissa sekä π:n että $S^*(t)$:n arvot voivat riippua erilaisista selittävistä muuttujista.

Maassaolomallinnuksen kohdalla standardin paranemisosuusmallin soveltaminen tarkoittaisi sitä, että π olisi se osuus Suomeen muuttaneista, jotka jäävät Suomeen pysyvästi. Todennäköisyys $1 - \pi$ on puolestaan se osuus Suomeen muuttaneista, jotka muuttavat pois Suomesta jossain vaiheessa, jolleivät he ehdi kuolla ennen poismuuttoa. Selviytymisfunktio $S^*(t)$ puolestaan kuvaa sitä, kuinka suuri todennäköisyys niillä, jotka jossain vaiheessa muuttavat pois Suomesta, on olla vielä Suomessa maassaoloajan t kuluttua Suomeen muutosta. Sekä Suomeen jääneiden osuus että Suomesta poismuuttavien selviytymisfunktio voivat riippua regressiomalleissa esimerkiksi sukupuolesta, maahanmuuttoiästä ja maahanmuuttovuodesta.

3.2.2.3 Rajoitettu kertyneen hasardin malli (BCH)

Edellä esitelty standardi paranemisosuusmalli on varteenotettava vaihtoehto mallinnusluokaksi, sillä se on yksinkertainen ja siitä on olemassa paljon käytännön sovelluksia. Standardilla paranemisosuusmallilla on kuitenkin kaksi puutetta, jotka tulevat esille monissa sovellustilanteissa.

Ensimmäinen puute on, että standardissa paranemisosuusmallissa *suhteellisten hasardien* (*engl. proportional hazards*) rakenne ei säily, kun paranemisosuutta (π) mallinnetaan selittävillä muuttujilla binomijakaumaan perustuvissa regressiomalleissa. *Hasardi* (*engl. hazard rate*) on todennäköisyyden raja-arvo jollain seuranta-ajan hetkellä t siirtyä tilasta toiseen, eli ei ”selviytyä”. Maassaolomallinnuksessa se on maastamuuton todennäköisyyden raja-arvo maassaoloaikana t.

Suhteellisten hasardien ominaisuus tarkoittaa sitä, että mallissa kaikilla selittävien muuttujien mukaan kahteen eri ryhmään (esimerkiksi miehiin ja naisiin) kuuluvilla on sama kiinteä hasardi suhteessa toiseen ryhmään kuuluviin (Aalen et al. 2008, 258). Tämä on perusoletuksena esimerkiksi paljon käytetyssä Coxin suhteellisten hasardien mallissa, mutta tämä oletus ei siis pidä paikkaansa standardin paranemisosuusmallin kohdalla, kun mallissa on mukana selittäviä muuttujia.

Toiseksi, jos selittävien muuttujien regressiokertoimilla on bayesilaisissa binomijakaumaan perustuvissa regressiomalleissa *epäsopivia todennäköisyysjakaumia* (*engl. improper probability distributions*) priorijakaumina, monilla tällaisilla malleilla on myös epäsopivat posteriorijakaumat. Epäsopivat posteriorijakaumat eivät täytä todennäköisyysjakauman määritelmää, joten niitä ei voi käyttää bayesilaisissa malleissa. (Ibrahim et al. 2001, s. 156 ja 161).

Yakovlev et al. (1993) kehittivät vaihtoehtoisen muodon standardille paranemisosuusmallille. Hieman myöhemmin Chen et al. kehittivät tästä mallista bayesilaisen yksiulotteisen (1999) että moniulotteisen (2002) version. Nämä mallit on esitetty myös heidän kirjansa luvussa 5 (Ibrahim et al. 2001).

Yakovlev et al. (1993) kehittämää mallia kutsutaan joko *rajoitetuksi kertyneen hasardin* malliksi (*engl. bounded cumulative hazard, BCH*), *puhkeamisajan malliksi* (*engl. promotion time model*) tai *parametriseksi paranemisosuuden* malliksi (*engl. parametric cure rate model*) (Sumathi and Rao K. 2008). Kaksi ensimmäistä nimitystä viittaavat BCH-mallin mallin muotoon:

$$\begin{aligned} S_{pop}(t) &= P(\textit{ei tutkittavaa tapahtumaa aikaan t mennessä}) \\ &= P(N=0) + P(Z_1 > t, \ldots, Z_N > t, N \geq 1), \end{aligned} \tag{3.9}$$

missä $S_{pop}(t)$ on selviytymisfunktion määrittämä selviytymistodennäköisyys koko populaatiossa, N on niiden latenttien eli ei-havaittavien tekijöiden lukumäärä, jotka voivat aikaansaada tutkittavan tapahtuman, ja vektori $\mathbf{Z} = (Z_1, \ldots, Z_N)$ sisältää näiden tekijöiden *puhkeamisajat* (*engl. promotion time*). (Ibrahim et al. 2001, 156).

Yakovlev et al. (1993) artikkelissa tutkimuksen kohteena oli syövän etäpesäkkeen havaitsemiseen kulunut aika alkuperäisestä syöpädiagnoosista. Selviytymisaika tarkoittaa tässä yhteydessä aikaa ilman havaittua etäpesäkettä. N puolestaan on niiden syöpäsolujen määrä, jotka aiheuttavat etäpesäkkeitä ja Z_n kunkin tällaisen solun kohdalla aika sen aiheuttaman etäpesäkkeen havaitsemiseen. Tässä yhteydessä paranemisen, eli sen ettei etäpesäkkeitä havaita milloinkaan, todennäköisyys on $P(N = 0)$.

Kun etäpesäkkeitä aiheuttavien syöpäsolujen havaitsemattoman määrän N oletetaan noudattavan poissonjakaumaa parametrilla θ, $S_{pop}(t)$ voidaan kirjoittaa muodossa (Ibrahim et al. 2001, 157):

$$S_{pop}(t) = exp(-\theta) + \sum_{N=1}^{\infty} S^*(t)^N \times \frac{\theta^N}{N!} exp(-\theta), \tag{3.10}$$

missä parametri θ on latentin lukumäärämuuttujan odotusarvo, ensimmäinen termi on $exp(-\theta) = P(N = 0)$, eli se on paranemisosuus populaatiossa, ja $S^*(t)$ on *sopiva* (*engl. proper*) selviytymisajan

funktio sille osuudelle populaatiosta, joilla on soluja, jotka aiheuttavat syövän etäpesäkkeitä. BCH-malli voidaan lausua myös standardin paranemisosuusmallin muodossa (Ibrahim et al. 2001, 158):

$$S_{pop}(t) = exp(-\theta) + (1 - exp(-\theta))S^*(t). \tag{3.11}$$

Näin määritellyllä BCH-mallilla suhteellisten hasardien ominaisuus säilyy, kun paranemisosuutta mallinnetaan selittävien muuttujien kautta – toisin kuin aiemmin määritellyllä standardilla paranemisosuusmallilla. Aalen et al. (2008) esittävät yleiset ehdot sille, milloin nk. *frailty*-mallit säilyttävät suhteellisten hasardien ominaisuuden selviytymismalleissa (s. 257–260, frailty-malleista enemmän jäljempänä). Lisäksi, kun tietyt ehdot täyttyvät, BCH-mallin posteriorijakauma noudattaa jotain sopivaa todennäköisyysjakauma bayesilaisessa viitekehyksessä (Ibrahim et al. 2001, 159–161). Sumathi and Rao K. (2008) kirjoittavat, että suurin osa BCH-mallia koskevasta kirjallisuudesta sijoittuu bayesilaiseen viitekehykseen, koska BCH-mallin posteriorijakaumat voivat olla sopivia todennäköisyysjakaumia, vaikka ei-selviytyvien selviytymisfunktio $S^*(t)$ edellä ei noudatakaan sopivaa todennäköisyysjakauma (Ibrahim et al. 2001 osoittavat tämän, s. 158).

Maassaoloaikojen mallinnuksen kannalta BCH-mallien luokka tarjoaa hyvin käyttökelpoisen vaihtoehdon, sillä suhteellisten hasardien ominaisuus tarkoittaa sitä, että erilaisilla paranemisosuutta – eli Suomeen pysyvästi jäämistä – selittävillä muuttujilla on selkeä tulkinta, kuten niillä on suhteellisten hasardien mallissa. Esimerkiksi, jos selittävänä muuttujana on Suomeen muuttaminen opiskelun vuoksi, BCH-mallilla voidaan tutkia, millainen on Suomeen pysyvästi jäämisen todennäköisyys suhteessa Suomeen muista syistä kuin opiskelun vuoksi muuttaneilla.

BCH-malli tarjoaa mielestäni myös sopivan käsitteellisen kehikon maassaoloaikojen mallinnusta varten. BCH-mallin havaitsemattomat latentit tekijät, jotka jokainen voivat aiheuttaa tutkittavan tapahtuman, tässä tapauksessa Suomesta poismuuttamisen, vastaavat sellaisia Suomesta poismuuton aiheuttavia tekijöitä, mitä havaintoaineistossa ei ole saatavilla.

Tällaisia tekijöitä voivat olla (em. artikkeliin Bodvarsson and Van den Berg (2013) viitaten) esimerkiksi sosiaaliset suhteet Suomessa (tai niiden puute), vahvat suhteet alkuperäiseen maahan ja aikomukset muuttaa Suomeen vain väliaikaisesti. BCH-mallin teoreettisen kehyksen mukaan ne henkilöt, joilla havaitsemattomia Suomesta muuton syitä ei ole, jäävät pysyvästi maahan. Niillä, jotka muuttavat Suomesta pois, Suomesta poismuutto tapahtuu sitä nopeammin, mitä enemmän näitä havaitsemattomia Suomesta poismuuton syitä heillä on. Näiden latenttien Suomesta poismuuton aiheuttavien tekijöiden määrä voi riippua aineistossa olevista selittävistä muuttujista.

BCH-malli olettaa latenttien tekijöiden olevan riippumattomia ja samoin jakautuneita. Tämä oletus ei välttämättä pidä paikkaansa maassaoloaikojen mallinnuksen kohdalla, mikäli latentteja maassaolosyitä konkretisoidaan yllä mainitulla tavalla. BCH-mallissa riittää kuitenkin, että aineiston on voitu ajatella syntyneen BCH-mallin olettamalla tavalla latenteista tekijöistä ilman, että latenttien tekijöiden ominaisuuksia pitäisi määritellä konkreettisesti (Ibrahim et al. 2001, 157).

BCH-mallissa, niin kuin se on määritetty edellä mainitussa kirjallisuudessa, latenttien Suomesta poismuuton aiheuttavien tekijöiden lukumäärä on jokaisella henkilöllä sama koko heidän Suomessa olonsa

ajan. Vielä realistisemman mallista saisi, jos nämä tekijät voisivat kertyä Suomessaoloajan kuluessa. Aalen et al. (2008) esittelevät tällaisen mallin, s. 434–435.

3.2.2.4 Yleinen paranemisosuusmalli

Guosheng and Ibrahim (2005) huomasivat, että on olemassa yleinen paranemisosuusmallien luokka, jonka erikoistapauksia edellä esitellyt standardi paranemisosuusmalli ja BCH-malli ovat. He johtivat tämän yleisen mallin tekemällä selviytymisfunktiolle Box-Cox-tyyppisen muunnoksen, joka sisältää muunnosparametrin, jota merkitään kirjaimella r.

Tällaisen mallin populaation selviytymisfunktio on muotoa (Guosheng and Ibrahim 2005, 562):

$$S_{pop}(t) = \{1 - r\theta \times F(t)\}^{(1/r)}, \tag{3.12}$$

missä r on muunnosparametri ja $F(t)$ on kiinnostuksen kohteena olevan tapahtuman todennäköisyyden kertymäfunktio. Yleinen malli lähestyy BCH-mallia, kun muunnosparametri r lähestyy nollaa, ja se on standardi paranemisosuusmalli, kun $r = 1$. Parametrin θ tulkinta riippuu muunnosparametrin r arvosta. Mikäli $r = 1$, eli kyseessä on standardi paranemisosuusmalli, niin $1 - \theta$ on paranemisosuus. Jos r lähestyy nollaa, niin tällöin θ on latentin lukumäärämuuttujan odotusarvo, kuten edellä määritettiin. Paranemisosuus yleisessä mallissa on $\{1 - r\theta\}^{(1/r)}$.

Yleinen paranemisosuusmalli tarjoaa laajemman valikoiman mahdollisia malleja, sillä muunnosparametri r voi saada periaatteessa mitä tahansa arvoja positiivisten reaalilukujen joukossa. Selittävien muuttujien regressioparametreilla ei ole kuitenkaan yleisessä paranemisosuusmallissa selkeää tulkintaa lukuun ottamatta tapausta, missä r lähestyy nollaa oikealta, eli kun kyse on BCH-mallista.

Guosheng (2008) esitti yleiselle paranemisosuusmallille moniulotteisen version. Sekä Yinin ja Ibrahimin (2005) että Yinin (2008) esittämissä malleissa muunnosparametrin arvoa r ei estimoida, vaan se kiinnitetään johonkin tiettyyn arvoon. Käyttämässäni De Castron et al. (2014) esittämässä mallinnusluokassa r on satunnainen, eli se myös estimoidaan mallissa.

3.2.2.5 Frailty eli havaitsematon heterogeenisuus

Frailty (vapaasti suomennettuna *alttius*) viittaa selviytymisaika-analyysissa aineiston sisältämään havaitsemattomaan heterogeenisuuteen. Havaitsematon heterogeenisuus aiheutuu siitä, että käytännössä selviytymisaikamalleissa ei ole koskaan mukana kaikkia selittäviä muuttujia niin, että kaikki selittämätön vaihtelu johtuisi vain puhtaasta satunnaisvaihtelusta. Esimerkiksi maassaoloaikojen mallinnuksen kohdalla tutkimusaineisto ei sisällä tarkkoja tietoja Suomeen muuton syistä, jotka saattavat vaikuttaa suomessaoloajan pituuteen. Tämä havaitsematon vaihtelu saattaa olla suurta ja merkittävää. (Aalen et al. 2008, 231–234).

Nimitys frailty tulee selviytymisaikamallien historiasta, jossa tutkittiin alun perin kuolemisen todennäköisyyttä selviytymisajan suhteen. Jotkut tutkittavat henkilöt olivat alttiimpia kuolemalle, eli he kuolivat nopeammin, kuin toiset henkilöt, joita koskien oli saatavilla samat taustatiedot: korkeampi alttius kuolemalle johtui ensin mainittujen henkilöiden kohdalla havaitsemattomista kuolemistodennäköisyyteen vaikututtavista syistä (Wienke 2011, 64). Tämä havaitsemattomien tekijöiden vaikutus tutkittavaan ilmiöön – eli tässä tapauksessa siirtymätodennäköisyyteen – on havaitsematonta heterogeenisuutta.

Edellä esitettyihin paranemismalleihin sisältyy jo yleinen oletus frailtyn rakenteesta. Standardissa paranemisosuusmallissa frailty on havaitsematon paranemisosuus. BCH-mallissa frailty on puolestaan havaitsemattomien tutkittavan tapahtuman laukaisevien tekijöiden määrä. Frailty voi noudattaa erilaisia todennäköisyysjakaumia, kuten esimerkiksi gamma-jakaumaa joillain parametriarvoilla. BCH-mallissa frailty noudattaa *compound Poisson* -jakaumaa (Aalen et al. 2008, 239–240).

Havaintokohtaiset frailtyn arvot voivat olla korreloituneita toisten havaintojen kanssa esimerkiksi havaintojen ryvästymisestä johtuen. Frailty-mallit ovat erityisen käytettyjä tällaisissa nk. moniulotteisissa (ks. alla) selviytymisaikamalleissa, sillä niiden avulla (Aalen et al. 2008, luku7; Ibrahim et al. 2001, luku 4) mallin selittämätön vaihtelu voidaan eritellä ryppäiden *sisäiseksi* (*engl. within*) ja ryppäiden *väliseksi* (*engl. between*) vaihteluksi. Ryppäät voivat muodostua esimerkiksi henkilöistä. Idea on sama kuin satunnaisvaikutuksia sisältävissä tilastollisissa malleissa.

Yksiulotteisissa selviytymisaikamalleissa frailty on hankalasti identifioituva, koska niissä on vaikeaa löytää tarvittavaa informaatiota koskien havaitsematonta heterogeenisuutta (Aalen et al. 2008, 275). Käsittelen frailtya De Castron et al. (2014) kehittämässä mallinnusluokassa tarkemmin jäljempänä.

3.2.2.6 Moniulotteisuus

Moniulotteisuus selviytymisaikamallinnuksen yhteydessä tarkoittaa sitä, että aineisto sisältää joko useita havaintoja yhdeltä henkilöltä, tai havaintoja, jotka ovat ryvästyneet esimerkiksi otanta-asetelman johdosta. Edellinen on erityistapaus jälkimmäisestä.

Maassaolomallinnuksen kohdalla moniulotteisuus tarkoittaa joko sitä, että yhdeltä henkilöltä voi löytyä useita tietoja Suomesta poismuuttamisesta, tai yleisesti ottaen sitä, että poismuuttohavainnot eivät ole toisistaan riippumattomia. Käsittelen tutkimusaineiston moniulotteisuutta ja mallinnusluokan valintaa tarkemmin jäljempänä.

3.2.2.7 Semiparametrisuus

Selviytymisaikamallin *hasardin vertailutaso* (*engl. baseline hazard*) on – ainakin yksinkertaisimpien mallien kohdalla – mallin hasardi silloin, kun malli ei sisällä lainkaan selittäviä muuttujia eikä frailtya. *Parametrisissa* (*engl. parametric*) malleissa hasardin vertailutason määräävät parametrit koskevat koko selviytymisaikaa nollasta äärettömyyteen. Esimerkiksi jos selviytymisaika noudattaa eksponenttijakau-

maa, jonka ainoa parametri (olkoon tässä θ) pysyy samansuuruisena kaikilla selviytymisajan arvoilla, tällöin hasardin vertailutaso riippuu vain θ:sta.

Semiparametrisissa (*engl. semiparametric* tai *piecewise*) malleissa seuranta-aika jaetaan intervalleihin, joita on J kappaletta (Ibrahimin et al. 2001 merkintöjä seuraten). Jokainen intervalli on avoin vasemmalta ja suljettu oikealta: $(s_k - 1, s_k], k \in \{1, 2, \ldots, J\}$, missä $0 = s_0 < s_1 < \ldots < s_J < \infty$ ja s_J on käytettävissä olevan havaintoaineiston suurin mahdollinen selviytymis- tai sensorointiaika. Intervallien ei tarvitse olla samansuuruisia. Semiparametrisessa mallissa jokaisella intervallilla on oma hasardin vertailutasonsa, eli niitä on J kappaletta. Kun $J = 1$, saadaan parametrinen malli.

Semiparametrisilla eksponenttijakautuneilla malleilla voidaan approksimoida mitä tahansa ei-parametrista hasardin vertailutasoa, joten ne tuovat lisää joustavuutta hasardin mallinnukseen. Kun intervallien määrää J kasvatetaan, mallin sopivuus yleisesti ottaen paranee suhteessa aineistoon. Kääntöpuolena kuitenkin on, että intervallikohtaisten hasardin vertailutasojen parametrien estimaattien arvoihin sisältyy suurempi epävarmuus (Ibrahim et al. 2001, 106–107 ja 171). Lisäksi tilastollisessa mallinnuksessa suositaan yleensä säästäväisyysperiaatetta parametrien määrän suhteen. Täysparametrisilla malleilla – semiparametristen mallien sijaan – saattaa olla myös helpompaa selittää niitä ilmiöitä, jotka ovat selviytymisfunktion muodon taustalla.

Maassaoloaikojen mallinnuksen kohdalla semiparametrisuus tarkoittaa sitä, että suomessaoloaika jaetaan intervalleihin niin, että viimeisen intervallin oikea reuna on 22,5 vuodessa, joka on seuranta-ajan maksimi aineistossa. De Castron et al. (2014) esittämä mallinnusluokka on semiparametrinen. Käsittelen tarkemmin sopivien intervallien valintaa mallinnuksen yhteydessä jäljempänä.

3.2.3 Suomessaoloaikojen mallinnus De Castron et al. (2014) menetelmällä

Käytin maassaoloaikojen maahanmuuttajaryhmäkohtaiseen mallinnukseen De Castron et al. (2014) kehittämää paranemisosuusmallien luokkaa. Esitän seuraavaksi tämän mallien luokan keskeisemmät kohdat huomioiden erityisesti edellisessä luvussa esitetyt malleilta toivottavat ominaisuudet. Esitän samalla yleisesti mallinnusluokan estimoitavat parametrit sekä miten valitsin käyttämäni mallit tarkemmin ottaen. Käytän parametrien kohdalla samoja merkintätapoja (mm. kreikkalaisten kirjainten suhteen) kuin De Castro et al. (2014) käyttävät artikkelissaan.

3.2.3.1 Yleistä mallinnusluokasta: muunnosparametri (r) ja paranemisosuus (ρ)

De Castro et al. (2014) johtavat selviytymisaikamallien luokan, jonka pohjalta he puolestaan johtavat yleisen paranemisosuusmallin luokan. Selviytymisaikamallin johtaminen alkaa selviytymisaikamallin kumulatiivisen hasardin määrittämisestä (ks. artikkelin lauseke (1)). He johtavat sen yleisesti käytetyn Cox-mallin kumulatiivisesta hasardista tekemällä sille muunnos, jonka parametrina on estimoitava muunnosparametri r. Lisäksi näin muunnettu kumulatiivinen hasardi kerrotaan frailtylla. Frailty saa arvoja henkilöittäin.

Muunnosparametri r on mallissa satunnainen toisin kuin joissain edellä mainituissa vanhemmissa malleissa. Tämä tuo malliin sen epävarmuuden, joka sisältyy r:n arvoon myös mallin estimoinnin jälkeen. Lisäksi r:n käsitteleminen satunnaisena parametrina tarkoittaa sitä, että sen estimointi ei edellytä useiden mallien ajamista ja niiden vertailuja useilla eri kiinnitetyillä r:n arvoilla. Muunnosparametrin arvo $r = 0$ toteuttaa paranemisosuusmallien osalta erikoistapauksena suhteellisten hasardien paranemisosuusmallin ja arvolla $r = 1$ puolestaan erikoistapauksena standardin paranemisosuusmallin. Varsinaisessa malliestimoinnissa r:stä käytetään logaritmista muunnosta, jota merkitään alkuperäisen artikkelin mukaan kreikkalaisella kirjaimella ksii, ξ. Muunnetun muunnosparametrin arvoja voidaan tulkita niin, että mikäli sen jakauma sijoittuu suureksi osaksi ”pienille” arvoille, eli ainakin jos $\xi < -5$ on suurella todennäköisyydellä, tällöin ollaan hyvin lähellä suhteellisten hasardien mallia (BCH-malli). Jos puolestaan sen jakauma sijoittuu suurimmaksi osaksi nollan ympäristöön, kyse on standardista paranemisosuusmallista.

De Castro et al. (2014) saavat paranemisosuusmallin muokkaamalla alkuperäisestä, muokkaamattomasta kumulatiivisesta hasardista kertymäfunktion ominaisuudet omaava funktio (ks. s. 190 loppu artikkelissa). Viitatussa artikkelissa ja tässä tekstissä paranemisosuudesta käytetään kreikkalaista kirjainta ρ. Suomessaoloaikamallien kohdalla ”paranemisosuus” tarkoittaa populaatiotasolla Suomeen muuttavan henkilön todennäköisyyttä jäädä Suomeen loppuelämäksi ilman, että muuttaa pois edes osaksi aikaa. Esitän mallinnustulosten yhteydessä nämä todennäköisyydet maahanmuuttajaryhmittäin.

Sensorointi

Kuten yleisesti selviytymisaikamalleissa, myös tässä mallinnusluokassa sensorointi otetaan huomioon. Sensoroinnin oletetaan olevan oikeanpuoleista ja sen oletetaan olevan myös ehdollisesti riippumatonta sekä kiinnostuksen kohteena olevan tapahtuman tapahtuma-ajoista että frailtysta. Kuten olen edellä esittänyt (ks. luku 3.1.2), oikeanpuoleisen sensoroinnin riippumattomuus toteutuu käytettävissä olevalla aineistolla maassaoloaikojen mallinnuksessa. Mallinnusluokka ei kuitenkaan ota huomioon vasemmanpuoleista katkaisua (*engl. left truncation*), joka sisältyy aineistoon. Olen edellä käsitellyt myös vasemmanpuoleisen vaikutuksia erityisesti Ruotsissa ja Saksassa syntyneiden osalta (ks. luku 3.1.2).

3.2.3.2 Semiparametrisuus ja intervallikohtaiset hasardit (λ)

Mallinnusluokka on semiparametrinen eli hasardin vertailutasot vaihtelevat seuranta-ajan intervallien mukaan. Viimeistä intervallia lukuun ottamatta hasardin vertailutasot ovat vakioarvoisia kussakin intervallissa, eli intervallikohtaiset hasardit noudattavat käytännössä eksponenttimallia, jossa hasardin arvo ei riipu ajasta. Viimeisessä intervallissa, joka on avoin väli viimeisestä tapahtumasta äärettömyyteen, hasardin vertailutaso pienenee eksponentiaalisesti seuranta-ajan t suhteen: $\lambda_J \times e^{-(t-s_{(J-1)})}$. Intervallin j hasardin vertailutason parametrista käytän De Castron et al. (2014) tavoin merkintätapaa: λ_j.

Käytännössä tämä muotoilu tarkoittaa sitä, että mallin mukaan maahanmuuttajapopulaatiosta ei ole enää sen jälkeen käytännössä poismuuttoja, kun maahanmuuttajat ovat olleet Suomessa 25 vuotta. Perustan tämän oletuksen pääasiassa havaittuihin maahanmuuttajaryhmäkohtaisiin selviytymisfunktioihin. Lisäksi, vaikka poismuuttaneisuus ei olisikaan aivan tasan nolla, oletan, ettei se poikkea Suomessa syntyneen

kantaväestön poismuuttaneisuudesta eli noin 0,1–0,2 %:sta vuosittain. Koska en mallinna lainkaan Suomessa syntyneiden osalta poismuuttoa – vaikka sitä todellisuudessa on mainitun osuuden verran – se, että pieni poismuuttaneisuus jää huomioimatta 25 maassaolovuoden jälkeen myös maahanmuuttajien osalta, ei vaikuta elinkaariennusteiden tuloksiin, kun niitä verrataan Suomessa syntyneiden ja ulkomailla syntyneiden välillä.

Intervallien valinta

Semiparametrisilla malleilla voidaan tavoittaa muodoltaan seuranta-ajan suhteen hyvin vaihtelevia hasardifunktioita. Mitä enemmän malli sisältää intervalleja, sitä monimutkaisempi malli ja sen hasardifunktio voivat olla. Intervallien, ja siten hasardien vertailuluokkien parametrien, määrää rajoittaa ennen kaikkea tilastollisen mallinnuksen säästäväisyyden periaate. Lisäksi hasardien vertailuluokkien estimaatit sisältävät sitä enemmän epävarmuutta niiden arvoista, mitä enemmän niitä on, koska estimaatit perustuvat pienempään määrään havaintoja.

De Castro et al. (2014) suosittelevat intervallien määrän valinnassa kokeilemaan eri intervallivalintoja ja tarkastelemaan näiden valintojen vaikutusta mallin posterioriestimaatteihin (s. 190). Lisäksi valinnassa voi käyttää mallinvalintakriteerejä, joita he esittelevät artikkelissaan. Intervallien pituudet De Castro et al. (2014) valitsivat niin, että jokaisessa intervallissa olisi suunnilleen saman verran havaintoja (s. 189).

Valitsin intervallien määrän ja niiden leveydet kokeilemalla eri vaihtoehtoja ja vertaamalla näiden vaihtoehtoisten mallien estimaateista saatuja selviytymisfunktion posterioriennusteita alkuperäisen aineiston selviytymisfunktioon. Esitän jäljempänä, kuinka posterioriennusteet saadaan tässä mallinnusluokassa. Jäljempänä esitettävät mallinnustulokset sisältävät kuviot, joissa posterioriennusteiden tuottamia selviytymisfunktioita verrataan alkuperäisen aineiston vastaaviin, joten nämä kuviot auttavat hahmottamaan tämän mallinnusvaiheen periaatteita.

Alustaviin maassaoloajan malleihin, joiden avulla valitsin intervallien määrän ja ajalliset pituudet, otin mukaan selittävinä muuttujina vain maahanmuuttoiän ja sukupuolen. Koska tässä mallinnusvaiheessa valitaan käytännössä vain hasardin vertailutasot, niin olisin voinut käyttää myös nk. *tyhjää mallia*, eli mallia, joka ei sisällä lainkaan selittäjiä.

Aloitin jokaisen maahanmuuttajaryhmän kohdalla intervallivalinnan kokeilemalla aluksi kolmea intervallia, sillä kokemus osoitti nopeasti, että pienemmillä määrillä näiden mallien pohjalta laadittujen posterioriennusteiden sopivuus alkuperäiseen havaintoaineistoon oli heikkoa. Ensimmäisen ja toisen intervallin välisen katkaisupisteen asetin noin 3–4 maassaolovuoden kohdalle, sillä useimpien maahanmuuttajaryhmien kohdalla aineiston selviytymisfunktio näytti laskevan ensimmäisten maassaolovuosien aikana nopeammin kuin, mitä esimerkiksi vakioarvoiseen poismuuttaneisuuteen perustuva täysparametrinen eksponenttifunktiomalli olettaisi. Viimeinen intervalli alkoi De Castron et al. (2014) menetelmän mukaisesti siitä maastamuuton havainnosta, jonka kohdalla maassaoloaika oli kaikkein pisin aineistossa.

Estimoin nämä alustavat mallit käyttäen De Castron et al. (2014) esittämää *Gibbsin otantaa* (estimoinnista tarkemmin jäljempänä). Jokainen tällainen Gibbsin otantaan perustuva posteriorijakauman simulointi sisälsi noin 5000–7000 simulaatiota kahdessa simulaatioketjussa niin, että mallit (eli parametrien yhteisposteriorijakaumat) suunnilleen konvergoituvat. Käytin pienempiä simulaatiomääriä intervallien va-

lintamalleissa, jotta estimointiaika yhden kokeilumallin osalta pysyisi ”järkevänä” (noin tunnissa), sillä jokaisen maahanmuuttajaryhmän kohdalla intervallien selvittäminen lopullista mallia varten vaati noin viiden tällaisen kokeilumallin ajon.

Kun yksittäinen kokeilumalli oli estimoitu, laadin sen parametrien posterioriestimaateista selviytymisfunktion posterioriennusteen. Vertasin tätä sitten kuvion avulla alkuperäisen aineiston vastaavaan. Mikäli posterioriennusteen selviytymisfunktio poikkesi kuvion perusteella aineiston vastaavasta, lisäsin seuraavaksi estimoitavaan kokeilumalliin joko uuden intervallin tai siirsin jo olemassa olevien intervallien katkaisukohtia niihin kohtiin, joissa posterioriennusteen selviytymisfunktio poikkesi eniten aineiston vastaavasta.

Jokaisen maahanmuuttajaryhmän kohdalla käytin ensimmäisiä kokeilumalleja lähinnä siihen, että lisäsin intervalleja. Jälkimmäisiä kokeilumalleja käytin siihen, että siirsin intervallien katkaisukohdat sopiviksi. Jatkoin tällä tavalla intervallien ”iterointia”, kunnes viimeisimmän kokeilumallin posterioriennusteen selviytymisfunktion odotusarvo vastasi likimain aineiston selviytymisfunktiota. ”Likimain” tarkoittaa tässä yhteydessä sitä, että ennusteen odotusarvon ja havaintoaineiston selviytymisfunktioiden ero kaikilla massaoloajoilla oli korkeintaan pari prosenttiyksikköä.

3.2.3.3 Frailty ja moniulotteisuus (w ja varianssiparametri θ)

Mallinnusluokassa jokaisella henkilöllä $i = 1, \ldots, n$ on oma frailtynsa w_i. Frailtyt w_i ovat riippumattomia ja samoin jakautuneita. Ne noudattavat $gamma(1/\theta, 1/\theta)$-jakaumaa, missä $\theta > 0$ on w_i:n varianssi. Gamma-jakaumat ovat kaikkein käytetyimpiä jakaumia frailtyn mallinnuksessa (Aalen et al. 2008, 239). Koska frailty on määritelmänsä mukaan havaitsematonta, De Castro et al. (2014) integroivat maassaoloajan uskottavuusfunktion kaikkien mahdollisten frailtyn arvojen yli, jotta mallin parametrien yhteisposteriorijakauma ehdollistettaisiin vain havaitulla aineistolla (s. 190). Integroinnin seurauksena uskottavuusfunktioon ja siten posteriorijakaumaan ilmaantuu θ-parametri. Jokaiselle henkilölle saadaan posterioriennusteissa poimittua oma frailtyn arvo $gamma(1/\theta, 1/\theta)$-jakaumasta, missä käytetään θ:n posteriorijakauman simuloituja arvoja.

Moniulotteisuus ja frailty

Tutkimusaineisto sisältää henkilöitä, joilla on enemmän kuin yksi poismuuttotieto Suomesta, eli aineisto on moniulotteista. De Castron et al. (2014) kehittämä mallinnusluokka on moniulotteinen, ja frailtyn oletetaan olevan samansuuruinen kaikilla yhtä henkilöä koskevilla havainnoilla. Koska nämä havainnot jakavat saman frailtyn, tätä kutsutaan *jaetuksi frailtyksi* (*engl. shared frailty*). Kuten jo edellä esitin, aineisto sisältää kuitenkin niin vähän henkilöitä, joilla on enemmän kuin yksi poismuutto Suomesta (alle 1 %:lla henkilöistä), että en ottanut valmiissa mallissa huomioon tätä moniulotteisuutta. Mikäli samankaltainen tutkimus toteutettaisiin tulevaisuudessa, aineistossa saattaisi olla tarpeeksi paljon henkilöitä, joilla on useita maastamuuttotietoja, jotta moniulotteisuus kannattaisi mallintaa.

De Castron et al. (2014) mallinnusluokkaa voisi kenties jatkokehittää maassaolomallinnuksen kohdalla kokeilemalla erilaisia frailty-jakaumia, joita esittelee esimerkiksi Wienke (2011). Koska ulkomailla asu-

minen ja ajan kuluminen saattavat vaikuttaa frailtyyn henkilön kohdalla, mainitun jaetun frailtyn sijaan voisi käyttää *korreloitunutta frailtya* (*engl. correlated frailty*), jossa saman henkilön eri havainnoilla ei olisi täsmälleen sama frailty, mutta ne olisivat silti korreloituneita keskenään.

Tutkimusaineisto sisältää (maahanmuuttajia), jotka ovat aviopuolisoita toisten (maahanmuuttajien) kanssa. Näiden henkilöiden kohdalla voidaan miettiä, että pitäisikö aviopuolisoilla olla sama jaettu frailty tai ainakin korreloitunut frailty, sillä aviopuolisoilla saattaa olla sellaisia samankaltaisia havaitsemattomia tekijöitä, jotka vaikuttavat maassaoloaikaan. En kuitenkaan soveltanut aviopareille tällaisia frailty-jakaumia. Mallinnuksen helpottamisen lisäksi toinen syy, ettei aviopuolisoiden jaettuja frailty-vaikutuksia sisällytetty malliin, on kuitenkin se, että maassaolomallit – kuten muutkin tutkimuksen elinkaariennusteiden komponenttien mallit – tehdään sen pohjalta, mitä maahanmuuttajista tiedetään, kun he muuttavat ensimmäisen kerran Suomeen. Koska suuri osa aviopareista muuttaa Suomeen eri vuosina, pelkistä muuttovuoden tiedoista ei selviä tietoa mahdollisesti myöhemmin Suomeen muuttavasta aviopuolisosta. Mallinnusasetelma on valittu muuttovuoden tietoihin perustuvaksi sen vuoksi, että mallinnustuloksia voitaisiin käyttää Suomeen saapuvien *uusien maahanmuuttajien* elinkaarivaikutusten ennustamiseksi.

3.2.3.4 Selittävien muuttujien valinta (regressiokertoimet β)

Valittu mallinnusluokka voi sisältää erilaisia maassaoloaikaa selittäviä muuttujia. Näiden selittävien muuttujien regressiokertoimilla ei ole kuitenkaan kuin erikoistapauksissa selkeää tulkintaa siitä, mitä ne tarkoittavat hasardien välisille suhteille. Ensimmäinen erikoistapaus on, kun muunnosparametri r lähestyy arvoa nolla. Tällöin regressiokertoimien tulkinta on se, että ne kuvaavat hasardien suhteita selittävien muuttujien eri arvoilla. Kun muunnosparametri $r = 1$, regressiokertoimet kuvaavat hasardien välisiä *vedonlyöntisuhteita* (*engl. odds*) selittävien muuttujien eri arvoilla.

Koska bayesilaisessa viitekehyksessä muunnosparametri r:n arvo on satunnainen, se ei käytännössä koskaan saa estimoitaessa arvoksi pelkkiä pisteitä 0 tai 1. Vaikka regressiokertoimilla ei olekaan selvää tulkintaa useimmissa tapauksissa, ne kertovat kuitenkin selittävien muuttujien mahdollisten vaikutusten merkistä, eli siitä pidentävätkö vai vähentävätkö muutokset selittävien muuttujien arvoissa odotettavissa olevaa maassaoloaikaa. Lisäksi ne mahdollistavat selittävillä muuttujilla ehdollisten (posteriori)ennusteiden laatimisen. (De Castro et al. 2014, 187 ja 198).

Selittävät muuttujat olisin voinut valita jokaisen maahanmuuttajaryhmän maassaolomalliin erikseen. Tämä olisi voinut tapahtua lisäämällä malliin selittäviä muuttuja yksi kerrallaan ja vertaamalla pienempää ja suurempaa mallia jollakin tunnetuilla mallinvalintakriteereillä, esimerkiksi myöhemmin tässä tutkimuksessa käytettävillä *WAIC*:lla (*Widely applicable information criterion*) ja *LOO-CV*:lla (*Leave-one-out cross-validation*).

Päädyin kuitenkin valitsemaan jokaisen maahanmuuttajaryhmän malliin samat selittävät muuttujat. Tärkein syy valintaani oli käytettävissä oleva aika. Vertailtavia malleja olisi ollut jokaisen maahanmuuttajaryhmän mallin kohdalla 20–30, mikäli olisin halunnut vertailla malleja edes kohtuullisen kattavasti. Koska yhden mallin konvergoitumiseen koko tutkimusaineistoa käyttäen olisi mennyt vähintäänkin tunteja, kaikkien mallien ajaminen olisi vienyt helposti tuhansia tunteja aikaa eli useita kuukausia. Lisäksi

”turhien” muuttujien tuoma epätarkkuus ennusteisiin voi olla pieni hinta siitä, että kaikki potentiaaliset selittäjät tulevat malliin, sillä niiden jättäminen pois mallista kokonaan tarkoittaa käytännössä niiden arvojen kiinnittämistä tasan nollaan, mikä on bayesilaisessa viitekehyksessä varsin kummallista (Gelman et al. 2014, 367).

Lisäksi maahanmuuttajaryhmäkohtaisia malleja on helppo verrata keskenään, kun ne sisältävät samat selittävät muuttujat. Selittävien muuttujien valinnassa pyrin siihen, että kaikki keskeisimmät aineistosta löytyvät maassaoloaikoja selittävät – ja siten ennustavat – muuttujat olisivat mukana malleissa. Perustelen seuraavaksi näiden muuttujien valinnan. Samat muuttujat esiintyvät suurimmaksi osaksi jäljempänä nuorten ja työikäisten nettovaikutusmalleissa ja palkkasummamalleissa.

Selittävien muuttujien valinnassa en voinut nojata muihin tutkimuksiin, koska yhdessäkään julkisen talouden tutkimuksessa, joihin viittaan, maassaoloaikojen mallinnusta ei ole tehty ainakaan yhtä kattavalla tavalla kuin tässä tutkimuksessa. Selittävien muuttujien valinta perustuu kolmeen valintakriteeriin: ensiksi maahanmuuttajaryhmäkohtaisiin aineiston eri selittävien muuttujien arvojen mukaan jaoteltuihin Kaplan-Meier-estimaattien kuvioihin, toiseksi muuttujien yleiseen esiintyvyyteen yhteiskuntatieteellisessä tutkimuksessa ilmiöitä selittävinä tekijöinä sekä kolmanneksi siihen, miten selittäjät kuvaavat sosiaalisia ja taloudellisia siteitä Suomeen.

Kaplan-Meier-estimaattien kuvioiden avulla tarkastelin, mitkä olivat niitä selittäviä muuttujia, joiden eri arvoilla selviytymisfunktiot vaikuttivat poikkeavan toisistaan. Yhteiskuntatieteellisessä tutkimuksessa ainakin sukupuoli, ikä ja koulutus ovat sellaisia tekijöitä, joiden mahdollisia yhteyksiä ilmiöihin tutkitaan yleensä. Kuten olen edellä maininnut Bodvarssoniin ja Van den Bergiin (2013) viitaten, sosiaaliset ja taloudelliset siteet voivat vaikuttaa maassaoloaikaan.

Selittävät muuttujat valitsin yllä mainittujen perusteiden nojalla *a priori* eli ennen mallien estimointia ja tulosten näkemistä. Jäljempänä, kun esitän tulokset, en juurikaan esitä tulosten suhteen *a posteriori* -päättelyitä mahdollisista syy–seuraus-suhteista selittävien muuttujien ja maassaoloaikojen välillä. Poikkeuksen tein tässä suhteessa maahanmuuttovuosiluokan vaikutusten kohdalla siinä, että pohdin tulosten yhteydessä, mistä syystä poismuuttaneisuus vaihtelee suuresti joidenkin maahanmuuttajaryhmien kohdalla maahanmuuttovuosiluokan mukaan.

Valitut selittävät muuttujat

Selittävät muuttujat sisältävät jokaisen henkilön kohdalla tietoja henkilöstä siltä vuodelta, kun hän muuttaa ensimmäisen kerran Suomeen. Alle 15-vuotiaana muuttaneiden kohdalla tiedot koskevat sitä vuotta, jolloin henkilö täytti 15 vuotta. Selittävät muuttujat, jotka valitsin malleihin, ovat:

– *Nainen* (referenssiluokka: mies).

– *Maahanmuuttoikä*.

– *Nainen* $\times$ *maahanmuuttoikä* (referenssiluokka: mies).

– *Maahanmuuttoikä*^{2} (jaettuna 100:lla).

– *Korkein koulutus: vähintään 2. asteen koulutus* (referenssiluokka: ei 2. asteen koulutusta tai ei tietoa).

– *Puoliso* (referenssiluokka: ei puolisoa).

– *Suomessa syntynyt puoliso* (referenssiluokka: ei Suomessa syntynyttä puolisoa).

– *Alle 7-vuotiaiden lasten lukumäärä* (perheessä).

– *Julkisen talouden menot* (tuhansia euroja) (vain ne menot, joista tiedot vuosille 1988–2011).

– *Ei julkisen talouden tuloja* (referenssiluokka: julkisen talouden tulot yli nollan. Vain ne tulot, joista tiedot vuosille 1988–2011).

– *Pääasiallinen toiminta: työllinen* (referenssiluokka: muu kuin työllinen tai opiskelija).

– *Pääasiallinen toiminta: opiskelija* (referenssiluokka: muu kuin työllinen tai opiskelija).

– *Maahanmuuttovuosi 1988–1993* (referenssiluokka: maahanmuuttovuosi 2000–2009).

– *Maahanmuuttovuosi 1994–1999* (referenssiluokka: maahanmuuttovuosi 2000–2009).

Valintaperusteet selittäville muuttujille

– *Nainen* (referenssiluokka: mies): yleisesti mukana yhteiskuntatieteellisissä tutkimuksissa. Lisäksi monen maahanmuuttajaryhmän kohdalla erot sukupuolten väillä esimerkiksi työllistymisessä ovat suuret, joten kenties eroja on myös maassaoloajoissa.

– *Maahanmuuttoikä*: ikä yleisesti mukana yhteiskuntatieteellisissä tutkimuksissa.

– *Nainen* $\times$ *maahanmuuttoikä* (referenssiluokka: Mies): interaktio kahden potentiaalisesti tärkeän muuttujan välillä.

– *Maahanmuuttoikä*^{2} (jaettuna 100:lla): Kaplan-Meier-estimaattien kuvioiden perusteella suurempi osa sekä nuorena että vanhana muuttaneista jää pysyvästi Suomeen kuin noin 20–40-vuotiaina muuttaneet. Jaoin muuttujan arvot sadalla, jotta muuttujan regressiokertoimien estimaatit olisivat enemmän samassa mittakaavassa muiden regressiokertomien estimaattien kanssa.

– *Korkein koulutus: vähintään 2. asteen koulutus* (referenssiluokka: ei 2. asteen koulutusta tai ei tietoa): koulutus yleisesti mukana yhteiskuntatieteellisissä tutkimuksissa. Tilastokeskuksen rekisterit ovat koulutustiedon suhteen puutteelliset, ja rekisterissä tiedon puuttuminen kuuluu samaan kategoriaan kuin ei 2. asteen koulutusta.

– *Puoliso* (referenssiluokka: ei puolisoa): yleisesti mukana yhteiskuntatieteellisissä tutkimuksissa.

– *Suomessa syntynyt puoliso* (referenssiluokka: ei Suomessa syntynyttä puolisoa): Suomessa syntynyt puoliso voi lisätä siteitä Suomeen ja siten vähentää muuttoalttiutta.

Kaikkien maahanmuuttajaryhmien kohdalla selittävinä muuttujina ei ole sekä puolisoindikaattoria että Suomessa syntynyt puoliso -indikaattoria, mikäli havaintoja, joilla indikaattorimuuttuja saa arvon yksi, on vähän, tai, mikäli indikaattorit ovat lähes päällekkäisiä. Jälkimmäinen tarkoittaa sitä, että lähes kaikki puolisot ovat Suomessa syntyneitä.

– *Alle 7-vuotiaiden lasten lukumäärä* (perheessä): lasten määrä ja lapsiperheys (lasten määrä > 0) ovat yleisesti mukana yhteiskuntatieteellisissä tutkimuksissa.

– *Julkisen talouden menot* (tuhansia euroja) (vain ne menot, joista tiedot vuosille 1988–2011): menot ovat nettovaikutusten miinusmerkkinen puoli, joten niiden mahdollinen vaikutus maastamuuttoon on mielenkiintoista tutkia. Tietoon sisältyvät vain ne julkisen talouden menot, joista on tiedot vuodesta 1988 saakka.

– *Ei julkisen talouden tuloja* (referenssiluokka: julkisen talouden tulot yli nollan. Vain ne tulot, joista tiedot vuosille 1988–2011): tulot ovat nettovaikutusten plusmerkkinen puoli, joten niiden mahdollinen vaikutus on mielenkiintoista tutkia.

Julkisen talouden tulojen puute viittaa taloudellisten siteiden puuttumiseen maahanmuuton alussa, ja tällä saattaa olla vaikutuksia koko maassaoloaikaan. Koska monen maahanmuuttajaryhmän kohdalla nollatuloisia on yli puolet väestöstä, päätin jakaa muuttujan mainittuun kahteen luokkaan. Tietoon sisältyvät vain ne julkisen talouden tulot, joista on tiedot vuodesta 1988 saakka.

– *Pääasiallinen toiminta: työllinen* (referenssiluokka: muu kuin työllinen tai opiskelija): pääasiallinen toiminta ensimmäisenä vuotena Suomessa kertoo epäsuorasti maahanmuuton syystä. Maahanmuuton syy saattaa puolestaan kertoa siitä, millaisia siteitä henkilöllä syntyy Suomeen.

– *Pääasiallinen toiminta: opiskelija* (referenssiluokka: muu kuin työllinen tai opiskelija): pääasiallinen toiminta ensimmäisenä vuotena Suomessa kertoo epäsuorasti maahanmuuton syystä. Maahanmuuton syy saattaa puolestaan kertoa siitä, millaisia siteitä henkilöllä syntyy Suomeen.

– *Pääasiallinen toiminta: muu kuin työllinen tai opiskelija*: luokkaan kuuluvat työttömät, eläkeläiset, varusmiehet ja siviilipalvelumiehet sekä muut (työvoiman ulkopuolella olevat).

– *Maahanmuuttovuosi 1988–1993* (referenssiluokka: maahanmuuttovuosi 2000–2009): maahanmuuttoluokka sisältää myös ennen vuotta 1988 Suomeen muuttaneet. Maahanmuuttovuoden vaikutukset ovat koko tämän tutkimuksen läpi kiinnostuksen kohteena, sillä se on aikaan liittyvä muuttuja. Haluan ottaa aikaan liittyvät muuttujat huomioon kaikissa malleissa. Muita aikaan liittyviä muuttujia ovat ikä, maassaoloaika ja kalenterivuosi sekä niiden muunnokset ja interaktiot muiden muuttujien kanssa.

– *Maahanmuuttovuosi 1994–1999* (referenssiluokka: Maahanmuuttovuosi 2000–2009).

Suurin osa mallinnuksen ulkopuolelle jättämistäni mahdollisista selittävistä muuttujista löytyy Tilastokeskuksen FLEED-rekisteriaineiston aineistokuvauksesta (Tilastokeskus 2016). Kokeilin joidenkin maahanmuuttajaryhmien mallien kohdalla ottaa mukaan esimerkiksi tiedon siitä, onko henkilö asunut muuttovuonna Uudellamaalla (referenssiluokka: asunut muualla kuin Uudellamaalla), tai siitä, oliko asuntokunnalla, johon henkilö kuului, omistusasunto muuttovuonna (referenssiluokka: ei omistusasuntoa). Näiden muuttujien posteriorijakaumat olivat kuitenkin keskittyneet nollan ympärille, joten päättelin tästä, etteivät ne ennusta juurikaan maassaoloaikaa.

Selittäviksi muuttujiksi olisi voinut valita lisäksi erilaisia interaktioita muuttujien välillä. Näitä olisi voinut tutkia erityisesti maahanmuuttajaryhmäkohtaisesti, mutta tämä olisi vaatinut liikaa aikaa.

3.2.4 Estimoitavien parametrien priorijakaumat ja posteriorijakauma

Olen edellä esittänyt De Castron et al. (2014) kehittämän mallinnusluokan keskeisimmät periaatteet, niihin liittyvät malliparametrit sekä millä perusteella olen valinnut maassaolomalleihin niiden selittävät muuttujat. Esitän seuraavaksi De Castron et al. (2014) artikkeliin ja sen liitteisiin viitaten maassaolomallien estimoitavien parametrien yhteisen posteriorijakauman sekä siihen liittyvät priorijakaumat hyperprioriparametrivalintoineen. Tämän jälkeen esitän, kuinka estimoin mallien posteriorijakaumat ja kuinka tarkistin mallien estimaattien konvergoitumisen. Lisäksi, ennen varsinaisten estimointitulosten esittämistä, käyn läpi, kuinka tarkistin posterioriennusteiden avulla estimoitujen mallien sopivuuden suhteessa alkuperäiseen aineistoon.

Priorijakaumat

Mallin estimoitavat parametrit, joille on määritettävä yhteinen priorijakauma, ovat $\boldsymbol{\beta}$-vektori, $\boldsymbol{\lambda}$-vektori, θ ja r. Tämä estimoitavien malliparametrien yhteispriorijakauma koostuu $\boldsymbol{\beta}$-vektorin, $\boldsymbol{\lambda}$-vektorin, θ:n ja r:n toisistaan riippumattomista jakaumista.

$\boldsymbol{\beta}$-vektorilla on tutkimusalan kirjallisuudessa paljon käytetty epäsopiva (engl. improper) jakauma:

$$\boldsymbol{\beta} \propto 1. \tag{3.13}$$

θ- ja r-parametri ovat *käänteisgammajakautuneita* (*engl. inverse-gamma, IG*), koska frailtylla on gamma-jakauma, ja valinta helpottaa posteriorijakauman estimointia. $\boldsymbol{\lambda}$-vektorin alkioilla on gamma-jakaumat. Mikäli näin määritelty malli täyttää hyperprioriparametrien valinnan suhteen sekä aineiston suhteet tietyt väljät kriteerit, posteriorijakauma on *sopiva* (*engl. proper*) todennäköisyysjakauma.

Nämä hyperprioriparametriehdot ovat De Castron et al. (2014) merkintöjä seuraten:

$\boldsymbol{\lambda}$-vektorin hyperprioriparametreja koskien:

$\gamma_{05_j} = \gamma_{06_j} = 0$, jolloin:

$$\lambda_j \sim gamma(0,0), \textit{kun } j = 1, \ldots, J-1 \tag{3.14}$$

ja $\gamma_{05_J} > 0$ ja $\gamma_{06_J} > 0$, jolloin:

$$\lambda_J \sim gamma(\gamma_{05_J}, \gamma_{05_J}). \tag{3.15}$$

θ-parametrin hyperprioriparametreja koskien:

$\gamma_{01} > 0$ ja $\gamma_{02} > 0$, jolloin:

$$\theta \sim IG(\gamma_{01}, \gamma_{02}). \tag{3.16}$$

Sekä muunnosparametrin r hyperprioriparametreja koskien:

$\gamma_{03} > 0$ ja $\gamma_{04} > 0$, jolloin:

$$r \sim IG(\gamma_{03}, \gamma_{04}). \tag{3.17}$$

Yllä mainitut sopivan todennäköisyysjakauman kriteerit täyttyvät käyttämäni aineiston ja mallien suhteen, joten näiden mallien posteriorijakaumat ovat sopivia todennäköisyysjakaumia. (De Castro et al. 2014, 191–192 ja Liite A).

Käytin malleissa yleisesti ottaen sellaisia hyperprioriparametreja, että priorijakaumat ovat epäinformatiivisia mallien estimoitavien parametrien suhteen. Poikkeuksena tästä oli θ-parametrin priorijakauma, jonka kohdalla määritin hyperprioriparametrit niin, että jakauma oli *melko* informatiivinen. Syynä tähän oli, että huomasin malleja estimoitaessa varsin nopeasti, että jos θ-parametrin priorijakauma oli täysin epäinformatiivinen, θ-parametrin posteriorisimulaatioissa esiintyi lyhyitä ”piikkejä” suunnilleen 10 000 simulaation välein, mikä esti erityisesti tämän parametrin posteriorijakaumaa konvergoitumasta kunnolla jäljempänä esitettävien konvergoitumiskriteerien mukaan.

Määritin tästä syystä θ-parametrille priorijakauman $IG(1, 1)$. Tämä on melko informatiivinen jakauma, sillä θ-parametrin näin määritellyn kertymäfunktion mukaan sen arvo oli noin 50 %:n todennäköisyydellä alle 1,5; 80 %:n todennäköisyydellä alle 4,5 ja noin 90 %:n todennäköisyydellä alle 10:n. Kymmenen jälkeen suurempiin arvoihin mentäessä kertymäfunktion kulmakerroin kuitenkin pienentyy lähelle nollaa, joten myös θ-parametrin suuret arvot (yli 10) ovat mahdollisia. Koska θ-parametri on frailtyn w_i:n varianssi, tällä priorijakauman valinnalla suuret frailty-varianssin – ja frailtyn toteumien – arvot tulevat epätodennäköisemmiksi.

Muunnosparametrin r:n priorijakaumana käytin myös $IG(1, 1)$-jakaumaa. Yksittäisten λ_j-parametrien kohdalla käytin epäsopivaa $gamma(0, 0)$-priorijakaumaa, kun $j < J$, ja $gamma(1, 1)$-priorijakaumaa, kun $j = J$.

Posteriorijakauma

Kun mallin havaintoaineiston uskottavuusfunktiolle (De Castron et al. (2014) artikkelin lauseke 5) ja edellä määritetylle estimoitavien malliparametrien yhteispriorijakaumalle (artikkelin lauseke 7) määritetään niiden välinen tulo, saadaan vakioa vaille mallin estimoitavien parametrien yhteisposteriorijakauma. Posteriorijakauma on tässä yksiuloitteinen toisin kuin viitatussa artikkelissa. Yksiulotteisuus tarkoittaa sitä tässä yhteydessä, että mallinnuksessa ei oteta erikseen huomioon sitä oletettua havaitsematonta heterogeenisuutta havaintoaineistossa, mikä aiheutuu siitä havaintojen osittaisesta ryvästymisestä, että yksi henkilö on voinut muuttaa enemmän kuin yhden kerran Suomeen.

Kun parametreille tehdään seuraavat muunnokset: $\lambda_j^* = r\lambda_j$ ja $r^* = r/\theta$, posteriorijakauma $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}^*, \theta, \boldsymbol{r}^* | \boldsymbol{D}_{\mathrm{o}})$

voidaan esittää seuraten De Castron et al. 2014 merkintätapoja muodossa:

$$
\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\lambda^*}, \theta, \boldsymbol{r^*}|\boldsymbol{D}_{\mathrm{o}}) \propto \prod_{i=1}^{n} \Bigg\{ \Bigg[\prod_{j=1}^{J} \{h_i^*(\lambda_j^*) e^{\boldsymbol{x_i'}\boldsymbol{\beta}}/r^*\}^{\delta_{ij}\nu_i} \{1 + \delta_{ij} H_i^*(\boldsymbol{\lambda^*}) \\
\times e^{\boldsymbol{x_i'}\boldsymbol{\beta}}\}^{-\delta_{ij}\nu_i} \Bigg] \frac{\Gamma(\theta^{-1} + \nu_i)}{\Gamma(\theta^{-1})} \Bigg[1 + \sum_{j=1}^{J} \log\{1 + \delta_{ij} H_i^*(\boldsymbol{\lambda^*}) e^{\boldsymbol{x_i'}\boldsymbol{\beta}}\}/r^* \Bigg]^{-\theta^{-1}-\nu_i} \Bigg\} \\
\times \left\{ (r^*)^{-(\gamma_{03}+1)} \, e^{-\gamma_{04}/(\theta r^*)} \prod_{j=1}^{J} (\lambda_j^*/\theta r^*)^{\gamma_{05_j}-1} \, e^{-\gamma_{06_j}\lambda_j^*/(\theta r^*)} \right\} \\
\times (r^*)^{-J} \, \theta^{-\gamma_{03}-\sum_{j=1}^{J}\gamma_{05_j}} \, \pi(\theta),
\end{aligned}
\tag{3.18}
$$

missä estimoitavien parametrien posteriorijakauma on ehdollistettu havaitulla aineistolla: $\boldsymbol{D}_{\mathrm{o}} = (n, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{v})$. Henkilön i havainto $y_i = min(T_i, C_i)$, missä T_i on maastamuuton aika ja C_i on sensorointiaika. Lisäksi henkilöltä i on havaittu sensorointi-indikaattori $v_i = I(T_i <= C_i)$, missä $v_i = 0$, jos y_i on oikeanpuoleisesti sensoroitu. Havaintomäärä on n, ja selittävien muuttujien matriisi on $\boldsymbol{X}$. Yksittäisiä intervalleja merkitään $j = 1, \ldots, J$, ja intervallien reunat ovat, kuten edellä (s. 74) on esitetty. Merkintätavan mukaisesti $\pi(\theta)$ on θ-parametrin priorijakauma.

Lisäksi posteriorijakaumassa yllä:

$h_i^*(\lambda_j^*) = \lambda_j^* e^{-\delta_{iJ}(y_i - s_{J-1})}$

ja

$$
H_i^*(\boldsymbol{\lambda^*}) = \lambda_j^*(y_i - s_{j-1})^{1-\delta_{iJ}}(1 - e^{-(y_i - s_{J-1})})^{\delta_{iJ}} + \sum_{g=1}^{j-1} \lambda_g^*(s_g - s_{g-1}), \tag{3.19}
$$

missä $\boldsymbol{\lambda^*} = (\lambda_1^*, \ldots, \lambda_J^*)'$. (De Castro et al. 2014, Liite B).

3.2.5 Posteriorijakauman simulointi ja konvergoituminen tavoitejakaumaan

3.2.5.1 Posteriorijakauman simulointi Gibbsin otannalla

De Castro et al. (2014) esittävät artikkelinsa liitteessä B, kuinka edellä esitetyn posteriorijakauman simulointi toteutetaan. Koska posteriorijakaumalla ei ole nk. *analyyttista ratkaisua*, eikä sitä voi simuloida poimimalla posteriorijakaumasta suoraan otoksia (nk. *Monte Carlo* -simulointimenetelmällä), artikkelin kirjoittajat käyttävät sopivaa MCMC-simulointimenetelmää (*MCMC* eli *Markov Chain Monte Carlo*).

Tässä tapauksessa he käyttävät kenties yleisimmin käytettyä MCMC-menetelmää, *Gibbsin otantaa*. Gibbsin otannassa estimoitavien parametrien yhteisposteriorijakaumaa, josta käytän yleisesti pelkkää nimitystä posteriorijakauma, simuloidaan poimimalla jokaisen parametrin ehdollisesta jakaumasta vuorotellen aina

uusi simuloitu arvo tälle parametrille. Nämä jakaumat ehdollistetaan havaitulla aineistolla ja kaikkien muiden parametrien uusimmilla simuloiduilla arvoilla.

Liitteessä B artikkelin kirjoittajat johtavat tarvittavat parametrien ehdolliset jakaumat Gibbsin otantaa varten. Tätä varten he luovat kaksi uutta parametri-vektoria ($\boldsymbol{U}$ ja $\boldsymbol{Z}$), joiden ainoa tarkoitus on helpottaa estimointia.

Lisäksi he tekevät kaksi satunnaismuuttujan muunnosta. Ensimmäisessä satunnaismuuttujan muunnoksessa he muuntavat θ-parametrin ehdollisen jakauman η-parametrin ehdolliseksi jakaumaksi, missä muunnos parametrien välillä on $\eta = \theta^{-1}$. Muunnosparametri r uudelleenparametroidaan siten, että $r^* = r/\theta$. Toisessa satunnaismuuttujan muunnoksessa r^*-parametrin ehdollinen jakauma muutetaan ξ-parametrin ehdolliseksi jakaumaksi, missä muunnos parametrien välillä on $\xi = log(r^*)$. Käytän tuloksia esittäessä näitä muunnettuja satunnaismuuttujia, sillä simulaatiot suoritetaan niiden ehdollisilla jakaumilla.

Suoritin posteriorijakauman simuloinnin Gibbsin otannalla kuudesta yksi- tai moniulotteisesta ehdollisesta jakaumasta, jotka on esitetty viitatun artikkelin liitteessä B. Moniulotteiset ehdolliset jakaumat koostuvat toisistaan riippumattomista yksiulotteisista reunajakaumista.

Kolmella näistä jakaumista on tunnettu muoto: jokaisella $\boldsymbol{\lambda}^*$-vektorin alkiolla λ_j^*, jokaisella apuparametrin vektorin $\boldsymbol{Z}$ alkiolla z_i sekä jokaisella apuparametrin vektorin $\boldsymbol{U}$ alkiolla u_i on riippumaton yksiulotteinen ehdollinen jakauma, joka noudattaa gamma-jakaumaa annetuin parametrein. $\boldsymbol{\beta}$-parametrivektorin, η-parametrin ja ξ-parametrin ehdolliset jakaumat eivät vastaa mitään tunnettuja todennäköisyysjakaumia, joista simuloituja arvoja poimimalla ne voitaisiin simuloida suoraan. $\boldsymbol{\beta}$-vektorin komponenttien $\beta_1, \ldots, \beta_p$, ξ-parametrin ja η-parametrin ehdollisten jakaumien todennäköisyystiheysfunktiot ovat log-konkaaveja, joten jakaumista voidaan suorittaa poimintoja Gilksin ja Wildin (1992) kehittämällä *adaptive rejection* -otannalla. Käytin osana Gibbsin otantaa tätä Gilksin ja Wildin kehittämää simulointimenetelmää, jonka suoritin käytännössä R-ohjelman *ars* (Rodriguez 2009) avulla.

Tämän tutkimuksen liitteessä A suoritan esimerkinomaisesti yllä esitetyn posteriorijakauman simuloinnin Gibbsin otantaa käyttäen. Käytän tässä aineistona vapaasti saatavilla olevaa aineistoa R-paketista *smcure* (Cai et al. 2012). Esimerkkiaineisto on selviytymisaika-aineisto, joka sisältää ”parantuvien” osuuden, eli osuuden havainnoista, jotka eivät koskaan koe tutkittavaa olotilasiirtymää.

Keskistin selittävät muuttujat Gibbsin otantaa varten. Keskistäminen tarkoittaa sitä, että vähensin kunkin muuttujan havainnon arvosta kyseisen muuttujan keskiarvon, tässä tapauksessa havainnoista lasketun otoskeskiarvon.

Suoritin suurimmassa osassa mallien posteriorijakaumien simulointia kaksi ketjua eli perättäisten simulointien sarjaa. Kummassakin ketjussa käytin parametreille sellaisia alkuarvoja, jotka olivat reaalilukujen akselilla vastakkaisissa suunnissa ja kaukana niistä arvoista, joita simulaatiot saivat ketjujen alun jälkeen. Tällä halusin varmistaa, että simuloiduilla parametrien posteriorijakaumilla ei ole useita moodeja. Kummassakin ketjussa poimin jokaisesta kuudesta ehdollisesta jakaumasta noin 50 000 simuloitua arvoa.

Simulaatiomäärä oli melko suuri, koska perättäisten simulaatioiden välinen autokorrelaatio, josta tarkemmin alla *Monte Carlo -keskivirheen* yhteydessä, oli korkea ξ-, η- ja λ_j^*-parametrien osalta. Myös näiden parametrien simulaatioiden väliset keskinäiset korrelaatiot olivat korkeita. Tämä on tyypillistä Gibbsin

otannassa, ja osittain tästä syystä valitsin nettovaikutus- ja palkkasummamallien posteriorijakaumien simulointiin, joita käsittelen jäljempänä tässä tutkimuksessa, Gibbsin otantaa (monesti) tehokkaamman simulaatiomenetelmän.

Posterioripäättelyitä varten tiputin kummankin ketjun alusta pois 20 000 ensimmäistä simulaatiota. Tätä alkuvaihetta kutsutaan *lämmittelyksi* (*engl. warm-up* tai *burn-in*). Ideana lämmittelyssä on, että simulaatioissa, joista posterioripäättelyt tehdään, ei ole mukana simulaatioita, jotka halutun posteriorijakauman sijaan edustavat enemmän valittuja alkuarvoja. Kummastakin jäljelle jääneiden ketjujen 30 000 simulaatiosta valitsin tehtäviä posterioripäättelyitä ja posterioriennusteita varten joka 25:nnen tai 50:nnen simulaation. Tämä *harvennus* (*engl. thinning*) nopeutti huomattavasti tehtäviä posterioripäättelyitä ja erityisesti posterioriennusteita.

3.2.5.2 Simulaatioketjujen konvergoituminen tavoiteposteriorijakaumaan

Gibbsin otannalla posteriorijakaumaa simuloidaan iteratiivisesti. Tämän vuoksi on syytä tarkistaa, että simulaatiot kuvaavat tarpeeksi hyvin haluttua posteriorijakaumaa, eli että ne ovat konvergoituneet tavoiteposteriorijakaumaan. Tutkin simulaatioiden konvergenssia Gelmanin et al. (2014) suosittelemilla menetelmillä (s. 282–285, ks. myös Gelman and Shirley 2011).

Ensimmäinen konvergenssin tarkistusmenetelmä koskee simulaatioaikasarjakuvioiden visuaalista tarkastelua. Jotta simulaatioketjujen sarjat konvergoituvat tavoiteposteriorijakaumaan, jokaisen simulaatiosarjan tulee olla stationaarinen sekä kaikkien simulaatiosarjojen tulee *sekoittua* (*engl. to mix*) keskenään. Stationaarisuuden arvioiminen visuaalisesti tarkoittaa sitä, että kun simulaatioketjut esitetään aikasarjakuvioina, ketjujen pitää näyttää samanlaisilta riippumatta simulaation järjestysnumerosta. Eli simulaatioiden arvojen pitää esimerkiksi pysyä samassa tasossa sekä niissä esiintyvät vaihtelu pitää olla samansuuruista lämmittelyvaiheen jälkeen läpi koko simulaatioketjun. Simulaatioketjujen sekoittumisen arvioiminen visuaalisesti tarkoittaa, sitä kaikkien – tässä tutkimuksessa kahden – simulaatioketjujen simulaatioiden täytyy saada aikasarjakuvioissa ”samanlaisia” arvoja kaikilla simulaatioiden järjestysnumeroilla. Käytännössä tämä tarkoittaa sitä, että jos simulaatioketjujen aikasarjoja tarkastellaan samassa kuviossa, ne vaikuttavat ”sekoittuvan”.

Tarkastelen simulaatioketjujen konvergenssia – eli stationaarisuuta ja sekoittumista – myös formaalimmalla testillä. Testisuureen laskemiseksi yksittäiset simulaatioketjut jaetaan aluksi osaketjuihin esimerkiksi puolittamalla alkuperäiset ketjut. Simulaatioiden kokonaisvarianssi koostuu tällöin *osaketjujen välisestä varianssista* (*engl. between-sequence variance*) sekä näiden *ketjujen sisäisestä varianssista* (*engl. within-sequence variance*). Testi perustuu siihen tulokseen, että konvergensissa simulaatioiden varianssia dominoi osaketjujen sisäinen varianssi niiden välisen varianssin sijaan. Mitä pienemmäksi testisuure muuttuu kohti lukua yksi, sitä konvergoituneempia tavoiteposteriorijakaumaan simulaatioketjut ovat.

Testisuuretta kutsutaan *potentiaaliseksi skaalan pienennykseksi* (*engl. potential scale reduction*) tai sen kirjoitustavan, $\hat{R}$, mukaan yksinkertaisesti *R-hatuksi*. Gelman et al. (2014) kertovat tarkemmin, kuinka testisuure lasketaan ja mistä sen nimi tulee (s. 284–285). Hyväksyttävän konvergenssin arvoina pidän sitä, että $\hat{R} <= 1{,}10$ (mt, 287). R-hattu konvergenssin mittarina perustuu vahvoihin oletuksiin, jotka

eivät aina toteudu. Tämän vuoksi R-hattu voi pysyä suurena, vaikka simulaatiomäärää kasvatettaisiin rajatta (Stan Development Team 2015, 457). Gelman et al. (2014) suosittelevat enemmän kuin kahden simulaatioketjun suorittamista konvergenssitarkasteluita varten, mutta tämä olisi vaatinut liikaa aikaa.

Epäsuoraan konvergenssitarkasteluun käytin osaltaan myös jäljempänä esitettäviä posterioriennusteita. Posterioriennusteiden sopivuus suhteessa havaintoaineistoon kertoo myös osaltaan, että konvergoituminen on tapahtunut. Parametrien posterioripäättelyiden estimaattien järkevyys viittaa myös konvergenssiin.

3.2.5.3 Posterioriestimaattien luotettavuus (*Monte Carlo -keskivirhe*)

Simulaatioiden konvergenssin lisäksi on syytä tarkastella simulaatioista tehtävien posterioripäättelyiden luotettavuutta silloin, kun kyse on piste-estimaateita. Käytännössä tämä tarkoittaa *Monte Carlo -keskivirheiden, MC-kv* (*engl. Monte Carlo Standard Error, MCSE*) esittämistä parametrien posteriorijakaumien odotusarvoestimaateille. MC-kv:ta kutsutaan myös *numeeriseksi keskivirheeksi* (*engl. numerical standard error*) (Koop 2003, 9). MC-kv kertoo, millainen virhe piste-estimaattiin sisältyy sen vuoksi, että se on simuloinnin tulos. Jotta piste-estimaattiin voi luottaa, sitä koskevan MC-kv:n on syytä olla niin pieni, ettei simuloinnista aiheutuva virhe vaikuta piste-estimaatista tehtäviin päätelmiin. (Flegal and Jones 2011, erityisesti. s 194).

Seuraava MC-keskivirheen määritelmä on Flegal and Jones (2011) esittämä. Noudatan heidän merkintätapojaan määritelmässä:

Olkoon $\bar{g}_{\mathrm{s}}$ satunnaismuuttuja g:lle laskettu piste-estimaatti (keskiarvo) S:stä simulaatiosta MCMC-ketjusta, jonka invariantti jakauma on π. Olkoon lisäksi $E_\pi g$ satunnaismuuttujan g todellinen odotusarvo. Mikäli tällöin MCMC-menetelmän tuottaman Markov-ketjuaineiston *Monte Carlo -virheelle* (*engl. Monte Carlo error*), $\bar{g}_{\mathrm{s}} - E_\pi g$, pätee Markov-ketjujen *keskeinen raja-arvolause* (*engl. central limit theorem, CLT*), Monte Carlo -virhe on asymptoottisesti normaalijakautunut:

$$\sqrt{S}(\bar{g}_{\mathrm{s}} - E_\pi g) \xrightarrow{d} N(0, \sigma_g^2), \text{kun} S \to \infty \text{ja} \sigma_g^2 \in (0, \infty). \tag{3.20}$$

Käytännössä simulaatiovarianssin σ_g^2:n todellista arvoa ei koskaan tunneta, joten se estimoidaan simulaatioaineistosta:

$$\hat{\sigma}_{\mathrm{s}}/\sqrt{S}. \tag{3.21}$$

Tämä estimaattori on MC-keskivirhe. Koska simulaatiot ovat autokorreloituneita, $\hat{\sigma}_{\mathrm{s}}$:n laskeminen ei ole yksinkertaista, ja siihen on olemassa erilaisia menetelmiä. Tästä syystä eri tilasto-ohjelmat voivat tuottaa samalle simulaatioketjujen aineistolle erisuuruisia MC-keskivirheitä. (Flegal and Jones 2011, 176).

Markov-ketjun CLT:n olemassaolo edellyttää sitä, että Markov-ketju on *geometrisesti ergodinen* (*engl. geometrically ergodic*) ja tietyt *momenttiehdot* (*engl. moment conditions*) täyttyvät (Flegal and Jones 2011, 193). Näiden ehtojen täyttyminen maassaolomallinnuksen piste-estimaattien kohdalla vaatisi kenties oman tutkimuksensa, joten en paneudu niihin sen enempää. Koska edellä esitetty simulaatioketjujen

sekoittuminen implikoi ketjujen ergodisuuden, tyydyn olettamaan, että Markov-ketjujen CLT:t ovat olemassa maassaolomallinnuksen piste-estimaattien kohdalla, kun niiden simulaatioketjut sekoittuvat.

Esitän niin maassaolomallien tulosten kuin muiden tämän tutkimuksien mallien osalta MC-keskivirheet parametrien odotusarvoestimaateille. MC-keskivirheet koskien odotusarvoestimaatteja ovat tuloksissa hyvin pieniä verrattuina parametrien odotusarvo- ja keskihajontaestimaatteihin. Simulaatiovirhe ei siis käytännössä vaikuta odotusarvoestimaatteihin. MC-keskivirheet voisi laskea myös muille piste-estimaateille, mutta se ei olisi nähdäkseni sitä varten tarvittavan suuren työmäärän arvoista (Flegal and Jones 2011, s. 187). Lisäksi odotusarvoestimaattien MC-keskivirheiden pienuuden johdosta oletan myös muiden posterioriestimaattien MC-keskivirheiden olevan pieniä. Maassaoloaikamallien osalta odotusarvoestimaattien MC-keskivirheet laskin simulaatioaineistosta R-ohjelman *coda*-paketin (Plummer et al. 2015) avulla.

3.2.6 Mallin sopivuuden tarkistaminen: posterioriennustetarkistukset

3.2.6.1 Yleistä posterioriennusteista

Mallin sopivuuden tarkistaminen tarkoittaa sitä, että mallin posterioriestimaatteja ja niistä tehtäviä päätelmiä verrataan havaintoaineistoon sekä mahdolliseen muuhun saatavilla olevaan informaatioon tutkittavasta ilmiöstä. Ilman tarkistusta on vaikeaa sanoa, soveltuvatko mallin tuottamat tulokset todella niistä tehtäviin päätelmiin. Tästä syystä mallin sopivuus suhteessa havaintoaineistoon ja muuhun saatavilla olevaan informaatioon on syytä osoittaa.

Mallin sopivuuden tarkistamisen voi aloittaa tutkimalla, onko mallista tehtävissä (posteriori)päättelyissä järkeä mallin ulkopuolisen informaation valossa (Gelman et al. 2014, 142). Mallin ”sisäinen” johdonmukaisuus tarkoittaa puolestaan sitä, että havaintoaineiston tulisi näyttää mallin posterioriennustejakaumalta. Poikkeamat havaintoaineiston jakauman ja posterioriennustejakauman välillä voivat johtua joko sattumasta tai mallin epäsopivuudesta. Systemaattiset poikkeamat viittaavat mallin epäsopivuuteen (Gelman et al. 2014, 143). Mallin sopivuuden tarkistaminen tarkoittaa tässä tapauksessa poimintojen tekemistä mallin estimoitavien parametrien yhteisestä posterioriennustejakaumasta ja vertaamalla näitä alkuperäiseen aineistoon. (ks. posterioriennusteista mallin tarkistuksessa tarkemmin: Gelman et al. 2014, 143–162).

Bayesilaisessa tilastollisessa päättelyssä posterioriennusteet perustuvat seuraavaan tulokseen:

$$p(y^{enn}|y^{hav}) = \int p(y^{enn}|\theta)p(\theta|y^{hav})d\theta, \tag{3.22}$$

missä $p(y^{enn}|y^{hav})$ on posterioriennustejakauma ennusteiden vektorille y^{enn} ja y^{hav} on havaintoaineiston vektori sekä θ on malliparametrien vektori. Tulos pätee, kun y^{enn} ja y^{hav} ovat ehdollisesti (ehdollistajana θ) riippumattomia. Posterioriennustejakauman ideana on, että aineistosta y^{hav} estimoidut posterioriestimaatit θ toimivat priorijakaumana $p(\theta|y^{hav})$ tulevien havaintojen otosjakaumalle $p(y^{enn}|\theta)$, eli kaikki havaintojen tarjoama informaatio ennusteita varten saadaan $p(\theta|y^{hav})$:stä. Integrointi θ -vektorin mahdollisten arvojen yli tapahtuu käytännössä simuloimalla, kun integraalille ei tunneta analyyttista ratkaisua. Simuloinnissa käytetään θ -vektorin posterioriestimoinnissa saatuja simuloituja arvoja.

3.2.6.2 Posterioriennusteet uudelle aineistolle

Posterioriennusteet voivat koskea joko havaittua aineistoa tai uutta aineistoa (Gelman et al. 2014, 108 ja 118). Havaitulle aineistolle ennusteiden tekeminen tarkoittaa sitä, että ennusteet tehdään samalle aineistolle, jota käytettiin parametrien estimointiin. Uusi aineisto tulee puolestaan parametrien estimoinnin jälkeen ”ulkopuolelta”. Tässä tutkimuksessa laadin posterioriennusteet vanhalle aineistolle siten, kuin ne tehtäisiin uudelle aineistolle, sillä haluan varmistaa tulosten soveltuvan uusien maahanmuuttajien elinkaarivaikutusten ennustamiseen.

Tällä tarkoitan sitä, että kaikille havainnoille yhteisten *havaittujen vaikutusten* lisäksi ennustan myös havaintoja koskevat satunnaisvaikutukset, jotka koskevat mallien *havaitsematonta heterogeenisuutta* (vrt. frailty edellä). Jos tavoitteena olisi vain laatia ennusteita pelkälle havaitulle aineistolle, satunnaisvaikutuksille ei tarvitsisi erikseen poimia ennusteita, koska havaintojen muodostamat ryppäät/ryhmät (esimerkiksi henkilöt), olisivat samat sekä mallia estimoitaessa että ennusteita tehdessä.

Uusi aineisto sisältää kuitenkin uudet ryppäät/ryhmät, joiden väliset satunnaisvaikutukset on syytä myös ennustaa, mikäli uusien havaintojen ryppäiden aiheuttama epävarmuus ennusteissa halutaan ottaa huomioon. Koska tutkimuksen tavoitteena on laatia uskottavat elinkaariennusteet vasta Suomeen muuttaville – eli uudelle aineistolle – tarvittavat satunnaisvaikutukset on myös ennustettava. Tästä syystä posterioriennustetarkistuksissa käsittelen havaintoaineistoa ikään kuin se olisi uutta aineistoa, eli sisällytän myös posterioriennustetarkistuksiin tarvittavien satunnaisvaikutusten simuloinnin.

Uudelle aineistolle tehtävät posterioriennusteet voidaan ilmaista seuraavasti (seuraten Gelman et al. 2014, 108 merkintätapoja):

$$p(y^{enn}|y^{hav}) = \int\int p(y^{enn}|\theta,\phi)p(\theta,\phi|y^{hav})d\phi d\theta, \tag{3.23}$$

missä:

$$p(\theta,\phi|y^{hav}) \propto p(y^{hav}|\theta)p(\theta|\phi)p(\phi). \tag{3.24}$$

Posterioriennusteiden jakauma $p(y^{enn}|y^{hav})$ on muuten kuten yllä lausekkeessa 3.22, mutta nyt nk. *otosjakauman* $p(y^{hav}|\theta)$ parametreilla θ on ehdollinen jakauma $p(\theta|\phi)$, nk. *populaatiomalli*, missä ϕ edustaa populaatiomallin priorijakauman *hyperprioriparametreja*. Tässä tutkimuksessa populaatiomalli on satunnaisvaikutusten malli, joka olotilamallinnuksen yhteydessä tarkoittaa frailtyn mallia. Mikäli satunnaisvaikutus noudattaa esimerkiksi normaalijakaumaa, niin silloin mainitut hyperprioriparametrit ovat kyseisen normaalijakauman odotusarvo- ja varianssiparametri.

Uuden aineiston sisältämät ”uudet” henkilöt (ja nettovaikutusmallin yhteydessä myös uudet maahanmuuttovuodet ja uudet maassaoloajat) ovat toteutumia/poimintoja populaatiomallista. Tällä tavalla laadittaviin posterioriennusteisiin sisältyy siten se epävarmuus, joka aiheutuu siitä, että uusilla maahanmuuttajilla on sellaisia havaitsemattomia tekijöitä (havaitsematonta heterogeenisuutta), jotka vaikuttavat maassaoloaikaan, nettovaikutuksiin ja palkkasummiin.

Käytännössä posterioriennusteet uudelle aineistolle poimitaan kolmivaiheisesti niin, että aluksi populaatiomallin hyperprioriparametreille ϕ poimitaan aluksi toteuma niiden omasta posteriorijakaumasta.

Tämän jälkeen populaatiomallista $p(\theta|\phi)$ poimitaan toteuma poimitulla ϕ:n arvolla ehdollistettuna. Lopulta toteuma posterioriennusteen jakaumasta saadaan poimimalla arvo aineistomallista $p(y^{enn}|\theta, \phi)$, missä θ ja ϕ ovat niille mainitulla tavalla poimitut toteumat.

3.2.6.3 Suomessaoloajan posterioriennusteet

Maassaolomallien posterioriennustetarkistuksissa vertasin maahanmuuttajaryhmäkohtaisia posterioriennusteiden selviytymisfunktioita havaintoaineistojen selviytymisfunktioihin. Kuviot vertailuista löytyvät jäljempänä tuloksista. Selviytymisfunktiot on kuvioissa estimoitu Kaplan-Meier-estimaattorilla, jonka estimaatit on laskettu R-ohjelman paketin *survival* avulla (Therneau and Lumley 2015).

Posterioriennustetarkistuksissa vertaan maahanmuuttajaryhmien kohdalla kaikkia havaintoja koskevien posterioriennusteiden selviytymisfunktioita havaintoaineistojen selviytymisfunktioihin. Lisäksi tein vastaavia vertailuja joidenkin luokiteltujen selittävien muuttujien luokkien mukaan. Näitä selittäviä muuttujia olivat lähinnä sukupuoli ja maahanmuuttovuoden luokka, koska näiden muuttujien luokkien selviytymisfunktiot vaikuttivat havaintoaineistotarkastelussa poikkeavan paljon toisistaan usean maahanmuuttajaryhmän kohdalla. Vertailukuvioissa on vain selviytymisfunktioiden odotusarvojen piste-estimaatit, koska suurten havaintomäärien vuoksi selviytymisfunktion (posteriori)jakauma on keskittynyt lähelle odotusarvoa.

Selviytymisfunktion arvo maassaoloajan hetkellä t tarkoittaa sitä todennäköisyyttä, että maastamuuttoa ei ole vielä tapahtunut tuohon hetkeen mennessä. Maassaolomallien selviytymisfunktioiden posterioriennusteet laskin hyödyntämällä posteriorijakaumien simuloinneista saatuja parametrien arvoja sekä De Castron et al. (2014) mallinnusluokan selviytymisfunktiota (s. 190).

Henkilön i selviytymisfunktion posterioriennuste on:

$$S_i(t^{enn}|\boldsymbol{y}) = \int_{\mathbb{R}^J_{>0}} \int_{\mathbb{R}^p} \int_{\mathbb{R}} \int_{\mathbb{R}_{>0}} S_i(t^{enn}|w_i, \boldsymbol{x_i}, \xi, \boldsymbol{\beta}, \boldsymbol{\lambda})\, p(w_i|\eta)\, p(\eta, \xi, \boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y})\, d\eta\, d\xi\, d\boldsymbol{\beta}\, d\boldsymbol{\lambda}\,, \tag{3.25}$$

missä $\boldsymbol{y}$ sisältää havaitut maastamuutto- ja sensorointiajat (kuten lausekkeen 3.18 ohessa on edellä esitetty), t^{enn} on Suomesta poismuuton ajan ennuste, w_i on henkilön i frailty ja $\boldsymbol{x_i}$ selittävien muuttujien vektori henkilölle i, ja muut parametrit ovat, kuten edellä on esitetty.

Simuloin jokaiselle henkilölle i erikseen posterioriennustejakaumasta niin monta selviytymisaikaa – eli maassaoloaikaa vuosina – kuin varsinaisessa mallin posterioriennusteesta kertyi lämmittelyvaiheen jälkeisiä simulaatioita. Tämä simulaatioiden lukumäärä oli maahanmuuttajaryhmittäin 1 200–2 500 simulaatiota.

Selviytymisaikojen posterioriennustejakaumien simulointiin käytetty R-koodi löytyy tutkimuksen liitteestä A. Mainitsin edellä, että suoritan liitteessä A De Castron et al. (2014) esittelemän mallinnusluokan posteriorijakauman simuloinnin käyttäen tähän tarkoitukseen R-paketista smcure löytyvää sopivaa

esimerkkiaineistoa. Hyödyntäen posteriorin simuloinnista kertyneitä parametriarvoja laadin selviytymisfunktiolle posterioriennusteet ja vertaan niitä esimerkkihavaintoaineiston selviytymisfunktion arvoihin.

3.2.7 Havainnot, estimointitulokset ja posterioriennustetarkistukset

Esitän tässä luvussa maassaoloaikoja koskevat havainnot, syntymävaltiokohtaiset mallinnustulokset sekä posterioriennustetarkistukset. Aloitan esittämällä syntymävaltiokohtaiset mallinnusaineistojen havaintolukumäärät. Sen lisäksi esitän maassaolomallien parametrien selitetaulukon, joka toimii apuna syntymävaltiokohtaisten mallien tulosten tarkasteluun. Olen esittänyt malliparametrien tarkemman selostuksen sekä valintaperiatteet edellä.

Koska maassaolomallien estimointi tapahtui syntymävaltiokohtaisesti, käyn estimointitulokset ja posterioriennustetarkistukset myös syntymävaltio kerrallaan läpi. Esitän jokaisen syntymävaltion kohdalla aluksi niiden mallinnusaineistojen selittävien muuttujien jakaumien tärkeimmät tunnusluvut: keskiarvot ja kvartiilit. En esittele erikseen selitettävää muuttujaa eli maassaoloaikojen havaintoja – ja sensorointeja – sillä maassaoloaikojen havaintoaineisto esitetään käytännössä posterioriennustetarkistusten yhteydessä.

Posterioriennustetarkistuksissa vertailen havaintoaineiston selviytymisfunktiota estimoitujen mallien pohjalta laadittujen maassaoloajan posterioriennusteiden selviytymisfunktioihin. Kuten osoitan kuvioiden avulla, posterioriennusteiden selviytymisfunktiot vastaavat hyvin havaintoaineistojen selviytymisfunktioita. Tämän perusteella pidän estimoitujen mallien tuottamia maassaoloajan ennusteita luotettavina maassaoloajan ennusteina myös uusien maahanmuuttajien elinkaarivaikutusten ennusteissa.

Posterioriennustarkistuksissa vertaan posterioriennusteita havaintoihin joko kaikkien havaintojen osalta tai selittävien muuttujien luokkien mukaan jaoteltuina. Mikäli teen vertailun selittävien muuttujien luokkien perusteella, olen valinnut vertailuun sellaisen selittävän muuttujan, jolla on estimointitulosten perusteella ”paljon” vaikutusta maassaoloaikaan. Usean syntymävaltion kohdalla maahanmuuttovuoden luokka vaikuttaa paljon maassaoloaikaan.

3.2.7.1 Havaintolukumäärät ja parametrien selitteet

Taulukossa 3.1 on esitetty syntymävaltiokohtaiset maassaoloajan mallinnuksen havaintojen ja henkilöiden lukumäärät. Henkilöt ovat aineistossa 15–70-vuotiaita, kun he ovat muuttaneet Suomeen. Alle 15-vuotiaana todellisuudessa muuttaneilla muuttoikä on 15. Havainnot muodostuvat joko maastamuuttotiedoista tai sensoirointitiedoista. Havainto sensorointiin, mikäli henkilö täytti 71 vuotta, hän kuoli tai seuranta-aika jatkui vuoteen 2010 ilman maastamuuttoa tai sensorointia muusta syystä.

Havaintoja on jokaisen syntymävaltion kohdalla hieman enemmän kuin henkilöitä. Tämä johtuu siitä, että havaintoaineisto sisältää sellaisia henkilöitä, joita koskien on useampi kuin yksi havainto. Tällaisissa tapauksissa henkilö on muuttanut aluksi pois Suomesta, mutta sittemmin hän on muuttanut Suomeen takaisin. Jokainen ”uusi” kerta Suomessa luo aineistoon aina uuden havainnon.

Taulukko 3.1: *Maassaoloajan mallinnusaineistojen havaintojen ja henkilöiden lukumäärät.*

Syntymävaltio	**Havainnot**	**Henkilöt**
Irak	5 161	5 078
Somalia	6 269	6 060
Thaimaa	4 168	4 085
Turkki	4 201	4 049
Saksa	6 029	5 674
Kiina	6 244	6 021
Entinen Jugoslavia	6 927	6 799
Viro	7 973	7 711
Entinen Neuvostoliitto	10 348	10 060
Ruotsi	11 066	10 450

Taulukossa 3.2 on esitetty maassaoloajan syntymävaltiokohtaisten mallien estimoitavien parametrien nimet ja niiden selitteet. Parametrit on esitetty siinä järjestyksessä, missä ne esitetään parametrien estimointitulosten syntymävaltiokohtaisissa taulukoissa jäljempänä. Parametrit, joista ensimmäinen on ”nainen”, ja viimeinen on, ”maahanmuuttovuosi v. 1994–1999”, ovat selittävien muuttujien regressiokertoimia. Parametrien nimet ja selitteet kertovat, mihin selittävään muuttujaan kukin regressiokerroin liittyy. Indikaattorimuuttujien osalta selitetaulukossa ovat näiden vertailuluokat.

Intervallikohtaiset hasardin vertailutasot ovat parametreja λ. Estimointituloksissa jokaisen niiden mukana on kaksi numeroa, jotka määrittävät sen maassaoloajan intervallin, jossa kyseinen parametri on voimassa. Intervallikohtaisia hasardiparametreja on syntymävaltiokohtaisissa malleissa 4–5 kappaletta. Muut kaksi malliparametria ovat frailty-parametrin muunnos η sekä mallin muunosparametrin muunnos ξ. Sisällytän estimointituloksiin vielä ρ:n, joka ei ole varsinainen malliparametri, sillä se lasketaan muiden parametrien tuottamista tuloksista. Se on kuitenkin mukana tuloksissa, sillä se on paranemisosuusmallin ”paranemisosuus”. Maassaoloajan mallinnuksen yhteydessä se on todennäköisyys jäädä Suomeen pysyvästi, kun selittävien muuttujien vaikutusta ei huomioida.

Taulukko 3.2: *Maassaoloajan mallien estimoitavien parametrien nimet ja niiden selitteet.*

Parametrinimi	**Selite**
Nainen	Vertailuluokka: mies.
Maahanmuuttoikä	
Nainen × maahanmuuttoikä	Maahanmuuttoiän vaikutus: naisilla verrattuna miehiin.
Maahanmuuttoikä^2	Regressiokerroin. Maahanmuuttoiän neliö jaettuna 100:lla.
Korkein koulutus: vähintään 2. asteen koulutus	Vertailuluokka ei 2.asteen koulutusta/ei tietoa.
Puoliso	Vertailuluokka: ei puolisoa.
Suomessa syntynyt puoliso	Vertailuluokka: ei Suomessa syntynyttä puolisoa.
Alle 7-vuotiaiden lasten lukumäärä	Lasten lukumäärä henkilön perheessä.
Julkisen talouden menot	Yksikkö: tuhat euroa. Vain me menot, jotka kaikille vuosille 1988–2011.
Ei julkisen talouden tuloja	Vertailuluokka: julkisen talouden tulot yli nollan. Vain me tulot, jotka kaikille vuosille 1988–2011.
Pääasiallinen toiminta: työllinen	Vertailuluokka: pääasiallinen toiminta: muu..
Pääasiallinen toiminta: opiskelija	Vertailuluokka: pääasiallinen toiminta: muu..
Maahanmuuttovuosi v. 1988–1993	Regressiokerroin. Vertailuluokka: Maahanmuuttovuosi v. 2000–2009.
Maahanmuuttovuosi v. 1994–1999	Regressiokerroin. Vertailuluokka: Maahanmuuttovuosi v. 2000–2009.
$\lambda_{nro}(alku; loppu)$	Hasardin intervallikohtainen vertailutaso. *nro* on intervallin järjestysluku, *alku* on intervallin alkupiste ja *loppu* on intervallin loppupiste.
η	Parametrilla on muunnos: $\eta = 1/\theta$. Henkilötasoinen frailty on $gamma(1/\theta, 1/\theta)$-jakautunut.
ξ	Parametrilla on muunnos: $\xi = log(r^*)$, missä $r^* = r/\theta$ ja r on mallin muunnosparametri.
ρ	ρ on mallin "paranemisosuus"eli muuttovuoden (ehdollistamaton) todennäköisyys jäädä Suomeen pysyvästi.

3.2.7.2 Estimointitulokset ja posterioriennustetarkistukset: Irak

Taulukossa 3.3 on esitetty Irakissa syntyneiden maassaolomallin selittävien muuttujien jakaumien keskiarvot ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden mallinnusväestössä: esimerkiksi ”puolison” keskiarvo kertoo, että 38 %:lla oli puoliso, kun he muuttivat Suomeen. Maahanmuuttoiän neliö on jaettu sadalla.

Taulukko 3.3: *Irakissa syntyneet 15–70-vuotiaat. Maassaoloajan mallin selittävien muuttujien jakaumien tunnusluvut.*

	Keskiarvo	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,41	0,00	0,00	1,00
Maahanmuuttoikä	26,59	16,00	25,00	33,00
Nainen $\times$ maahanmuuttoikä	10,70	0,00	0,00	20,00
Maahanmuuttoikä^2	8,30	2,56	6,25	10,89
Vähintään 2. asteen koulutus	0,14	0,00	0,00	0,00
Puoliso	0,38	0,00	0,00	1,00
Alle 7-vuotiaiden lasten lukumäärä	0,47	0,00	0,00	1,00
Julkisen talouden menot	4,08	1,29	3,75	6,17
Ei julkisen talouden tuloja	0,67	0,00	1,00	1,00
Pääasiallinen toiminta: työllinen	0,04	0,00	0,00	0,00
Pääasiallinen toiminta: opiskelija	0,30	0,00	0,00	1,00
Pääasiallinen toiminta: muu	0,66	0,00	1,00	1,00
Maahanmuuttovuosi 1988–1993	0,12	0,00	0,00	0,00
Maahanmuuttovuosi 1994–1999	0,26	0,00	0,00	1,00
Maahanmuuttovuosi 2000–2009	0,62	0,00	1,00	1,00

Estimointitulokset

Taulukossa 3.4 on esitetty Irakissa syntyneiden estimoidun maassaolomallin parametrien posteriorieestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli yhtä, en käsittele niitä tarkemmin.

Taulukko 3.4: *Irakissa syntyneiden15–70-vuotiaiden maassaoloajan mallinnus. Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Nainen	0,30	0,12	0,08	0,22	0,30	0,38	0,53	0,0027	1,00
Maahanmuuttoikä	0,005	0,010	-0,014	-0,002	0,005	0,011	0,023	0,0002	1,00
Maahanmuuttoikä^2	-0,03	0,04	-0,12	-0,06	-0,03	0,00	0,06	0,0010	1,00
Nainen $\times$ maahanmuuttoikä	-0,037	0,011	-0,059	-0,044	-0,036	-0,029	-0,016	0,0002	1,01
Vähintään 2. asteen koulutus	0,16	0,15	-0,15	0,06	0,16	0,26	0,46	0,0035	1,00
Puoliso	-0,21	0,14	-0,47	-0,31	-0,21	-0,12	0,05	0,0031	1,00
Alle 7-vuotiaiden lasten lukumäärä	-0,23	0,07	-0,37	-0,28	-0,23	-0,18	-0,10	0,0016	1,00
Julkisen talouden menot	-0,26	0,19	-0,63	-0,39	-0,26	-0,12	0,11	0,0044	1,00
Ei julkisen talouden tuloja	0,07	0,14	-0,19	-0,02	0,07	0,16	0,34	0,0031	1,01
Pääasiallinen toiminta: työllinen	-0,29	0,31	-0,91	-0,50	-0,28	-0,08	0,29	0,0071	1,00
Pääasiallinen toiminta: opiskelija	-0,44	0,16	-0,76	-0,55	-0,43	-0,32	-0,13	0,0038	1,00
Maahanmuuttovuosi 1988–1993	0,64	0,16	0,33	0,53	0,64	0,74	0,97	0,0037	1,00
Maahanmuuttovuosi 1994–1999	0,44	0,14	0,17	0,34	0,44	0,54	0,72	0,0032	1,00
$\lambda_1[0;4{,}5)$	0,0064	0,0029	0,0026	0,0044	0,0057	0,0076	0,0138	0,00007	1,03
$\lambda_2[4{,}5;10{,}5)$	0,0042	0,0019	0,0018	0,0029	0,0038	0,0050	0,0094	0,00004	1,04
$\lambda_3[10{,}5;15{,}5)$	0,0048	0,0023	0,0020	0,0033	0,0043	0,0057	0,0105	0,00005	1,06
$\lambda_4[15{,}5;19{,}5]$	0,0056	0,0028	0,0023	0,0037	0,0049	0,0066	0,0126	0,00006	1,09
$\lambda_5(19{,}5;23)$	0,0041	0,0050	0,0001	0,0010	0,0024	0,0054	0,0180	0,00012	1,02
η	1,32	0,84	0,35	0,73	1,11	1,69	3,45	0,0194	1,02
ξ	-0,68	0,92	-2,44	-1,31	-0,68	-0,06	1,12	0,0210	1,00
ρ	0,84	0,02	0,79	0,82	0,84	0,85	0,87	0,0005	1,00

Regressiokertoimilla ei ole taulukossa 3.4 standardin paranemisosuusmallin tai BCH-mallin kaltaista selkeää tulkintaa. Tämä johtuu siitä, että ξ-parametrin posteriorijakauma ei ole keskittynyt ”pieniin” arvoihin – jolloin regressiokertoimilla olisi BCH-mallin suhteellisten hasardien tulkinta – eikä myöskään nollan ympärille – jolloin puolestaan regressiokertoimilla olisi standardin paranemisosuusmallin (log-)vedonlyöntisuhteiden tulkinta.

Regressiokertoimet kuitenkin kertovat, mihin suuntaan muutokset selittävät muuttujissa vaikuttavat maassapysymisen todennäköisyyteen. Kun regressiokertoimet saavat positiivisia arvoja, tämä tarkoittaa sitä, että positiivinen muutos niitä vastaavissa selittävissä muuttujissa *lisää* maastamuuton todennäköisyyttä. Samaten, mikäli regressiokertoimet saavat negatiivisia arvoja, tämä tarkoittaa sitä, että positiivinen muutos niitä vastaavissa selittävissä muuttujissa *vähentää* maastamuuton todennäköisyyttä. Koska indikaattorimuuttujat ovat 0–1-muuttujia, indikaattorimuuttuja saa arvoksi *vertailuluokassa* nollan ja muuten ykkösen. Indikaattorimuuttujien regressiokertoimet ilmaisevat täten maastamuuton todennäköisyyttä verrattuna vertailuluokkaan. Indikaattorimuuttujien vertailuluokat on esitetty edellä parametrien selitetaulukossa eli taulukossa 3.2.

Koska indikaattorimuuttujaa ”nainen” koskevan regressiokertoimen posteriorijakauma on keskittynyt positiivisiin (reaali)lukuihin, tämä tarkoittaa sitä, että naisten maastamuuttotodennäköisyys on suurempi kuin miesten, kun kaikki muut mallissa olevat selittävät muuttujat (ja frailty) otetaan huomioon. Negatiivisiin lukuihin keskittyneen interaktion ”nainen $\times$ maahanmuuttoiän” regressiokertoimen posteriorijakauman tulkinta puolestaan on, että iän vaikutus naisilla maastamuuton todennäköisyyteen on erilainen kuin miehillä: miehillä ikä lisää enemmän maastamuuton todennäköisyyttä kuin naisilla.

Sekä ”puoliso”-indikaattorin että alle 7-vuotiaiden lasten määrää koskevien regressiokertoimien jakaumat ovat kummatkin keskittyneet negatiivisiin lukuihin. Tämä viittaa siihen, että jos Irakissa syntyneellä henkilöllä on perhettä (puoliso ja/tai lapsia), kun hän muuttaa Suomeen, hänen todennäköisyytensä pysyä Suomessa kasvaa sellaiseen henkilöön verrattuna, jolla perhettä ei ole muuttaessa.

Julkisen talouden menojen osalta tulosten perusteella on 75 %:n ja 97,5 %:n välillä oleva todennäköisyys sille, että mitä enemmän Irakissa syntyneellä henkilöllä on julkisen talouden menoja maahanmuuttovuonna, sitä suurempi on hänen todennäköisyytensä pysyä Suomessa. Taulukon 3.4 tulosten perusteella vaikuttaa lähes varmalta, että jos henkilön pääasiallinen toiminta on ollut opiskelijuus maahanmuuttovuonna, hänen todennäköisyytensä pysyä Suomessa kasvaa verrattuna henkilöön, jonka pääasiallinen toiminta on ollut jokin muu kuin opiskelija tai työllinen muuttovuonna.

Taulukon 3.4 tulosten perusteella vaikuttaa myös selvältä, että maastamuuton todennäköisyys on suurempi niillä Irakissa syntyneillä, jotka ovat muuttaneet Suomeen alun perin joko vuosina 1988–1993 tai vuosina 1994–1999 – verrattuna niihin, jotka ovat muuttaneet Suomeen vuosina 2000–2009. Lisäksi 1988–1993 muuttaneilla poismuuttaneisuus vaikuttaa olevan suurempaa kuin 1994–1999 Suomeen muuttaneilla. Koska maahanmuuttovuoden luokka vaikuttaa näiden tulosten ja alkuperäisen havaintoaineiston perusteella vaikuttavan selvästi maassaolon todennäköisyyteen, tarkastelen Irakissa syntyneiden osalta posterioriennustetarkistuksissa juuri maahanmuuttovuoden luokkaa.

Taulukossa 3.4 ρ:n eli ”paranemisosuuden” jakauma on keskittynyt lähelle 84 %:ia. Tämä tarkoittaa sitä, että havaintoaineistossa Irakissa syntyneiden (selittävillä muuttujilla ehdollistamaton) todennäköisyys jäädä Suomeen pysyvästi – ilman poismuuttoja – on noin 84 %:ia.

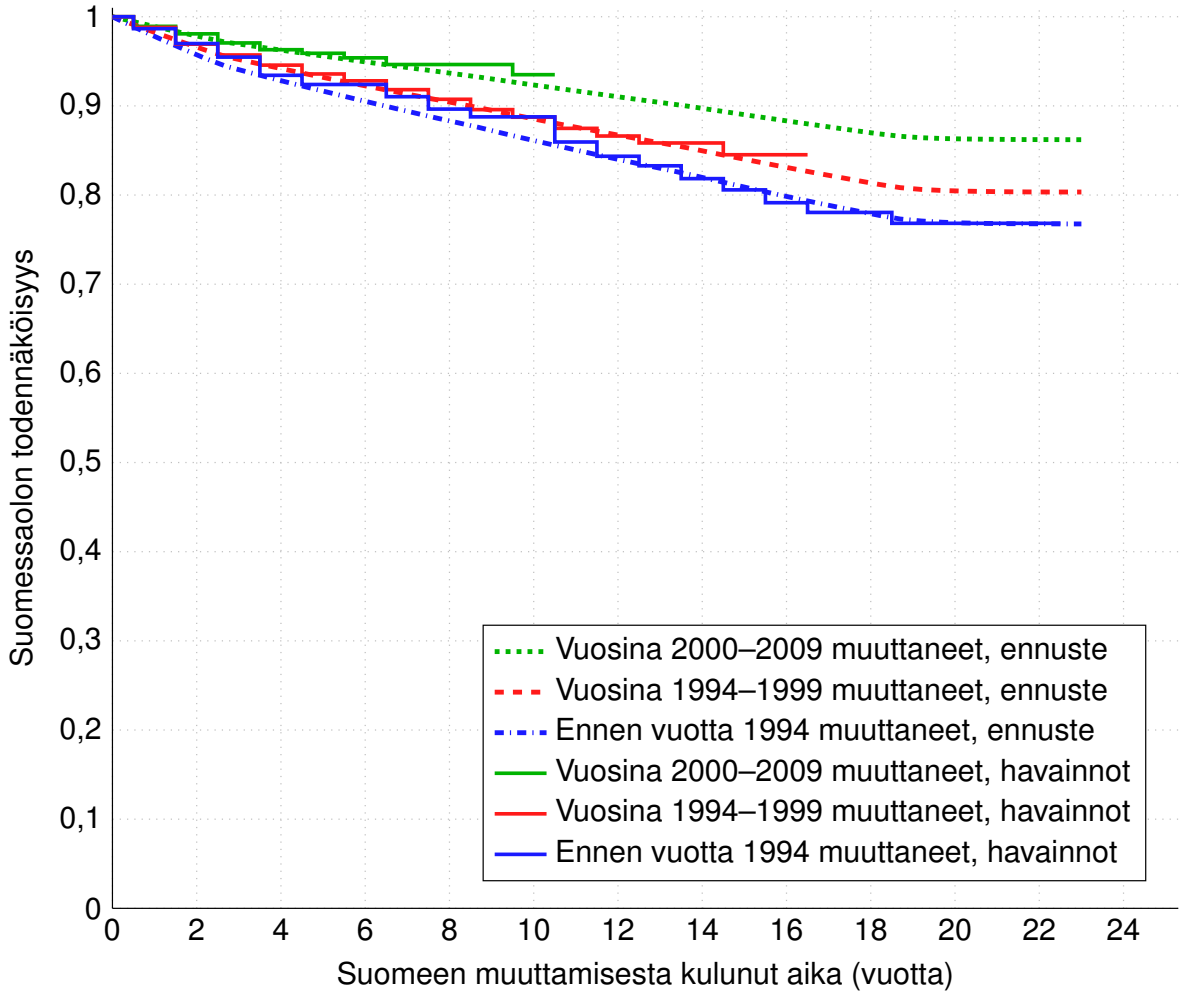
Posterioriennustetarkistus

Irakissa syntyneiden maassaoloajan posterioriennustetarkistuksissa vertaan havaintoaineiston selviytymisfunktioita maahanmuuttovuoden luokan mukaan posterioriennusteiden vastaaviin selviytymisfunktioihin. Valitsin maahanmuuttovuoden luokan tarkastelun kohteeksi, sillä sekä havaintoaineiston että taulukossa 3.4 esitettyjen estimointitulosten perusteella maahanmuuttovuoden luokka vaikuttaa maassaolon todennäköisyyteen.

Posterioriennustetarkistuksen tulokset on esitetty kuvassa 3.4. Kiinteät, portaittaiset viivat esittävät selviytymisfunktion – eli suomessaolon todennäköisyyttä maassaoloajan mukaan – arvoja eri maahanmuuttovuoden luokan mukaan. ”Hyppäykset” tai portaat havaintoaineiston kuvaajissa vastaavat henkilöiden poismuuttoja Suomesta tietyn maassaoloajan jälkeen.

Mikäli havaintoaineistossa yksi ”hyppäys” tai porras aiheutuu alle 5 henkilön poismuutosta, olen yhdistänyt yksityisyydensuojan vuoksi useamman perättäisen hyppäyksen yhdeksi hyppäykseksi siten, että poismuutot vastaavat vähintään 5 henkilöä. Tällöin olen sijoittanut hyppäyksen suuremman maassaoloajan arvon kohdalle. Esimerkiksi, jos 7 ja 8 maassaolovuoden välillä kaksi henkilöä on muuttanut pois ja 8 ja 9 maassaolovuoden välillä 3 henkilöä on muuttanut pois, olen tulkinnut tämän kuvioon niin, että viisi henkilöä on muuttanut pois 8 ja 9 maassaolovuoden välillä.

Jatkuvat, katkonaiset viivat esittävät kuviossa puolestaan estimoidusta maassaoloajan mallista tuotettujen posterioriennusteiden odotusarvoja maahanmuuttovuoden luokan mukaan. Kuten kuviosta 3.4 ilmenee, posterioriennusteiden selviytymisfunktioiden odotusarvot vastaavat hyvin havaintoaineiston selviytymisfunktioita.



Kuva 3.4: *Posterioriennustetarkistus. Irakissa syntyneet 15–70-vuotiaat. Suomessaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle maahanmuuttovuoden luokan mukaan.*

Kuviosta käy myös ilmi, että suomessaolon todennäköisyys, kun otetaan huomioon maassaoloaika, on suurempaa vuosina 2000–2009 muuttaneilla kuin tätä aiemmin muuttaneilla. Ennen vuotta 2000 muuttaneilla suomessaolon todennäköisyys pysyi samalla tasolla noin 10 maassaolovuoteen asti, mutta sen jälkeen suurempi osa ennen vuotta 1994 muuttaneista on muuttanut pois kuin vuosina 1994–1999 Suomeen muuttaneista. Suomeen pysyvästi jäävien osuus on ennen vuotta 1994 muuttaneilla noin 75 %:ia ja ennusteen mukaan vuosina 1994–1999 muuttaneilla noin 80 %:ia. Ennusteen mukaan vuosina 2000–2009 muuttaneista noin 85 %:ia jää pysyvästi Suomeen.

Tämä on mielestäni varsin realistinen ennuste, sillä kuten olen aiemmin esittänyt, kaikissa havaintoaineiston luokissa (ei vain Irakissa syntyneillä) selviytymisfunktioiden kulmakertoimet ovat karkeasti ottaen

väheneviä. Tosin tämä ei vaikuta juuri Irakin ja Somalian kohdalla niin selkeältä intervallikohtaisten hasardien vertailutasojen (vrt. λ-parametrien arvot estimointituloksissa) perusteella – joskin näidenkin maiden osalta poismuuttaneisuus *ei kasva* maassaoloajan kasvaessa. Lisäksi myös Irakin ja Somalian kohdalla selviytymisfunktio vaikuttaa lopulta tasoittuvan jonkin suomessaolon todennäköisyyden arvon kohdalle. Kuvion 3.4 perusteella vuosina 2000–2009 muuttaneilla tämä ”tasoittuminen” jonkin suomessaolon todennäköisyyden kohdalle vaikuttaa tapahtuvan jo 10 maassaolovuoden jälkeen, joten ennuste noin 85 %:n jäämisestä pysyvästi Suomeen voi olla jopa liian matala, sillä ennuste pohjautuu osittain aiemmin Suomeen muuttaneita koskevaan informaatioon mallissa.

3.2.7.3 Estimointitulokset ja posterioriennustetarkistukset: Somalia

Taulukossa 3.5 on esitetty Somaliassa syntyneiden maassaolomallin selittävien muuttujien jakaumien keskiarvot ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden mallinnusväestössä: esimerkiksi ”puolison” keskiarvo kertoo, että 27 %:lla oli puoliso, kun he muuttivat Suomeen. Maahanmuuttoiän neliö on jaettu sadalla.

Taulukko 3.5: *Somaliassa syntyneet 15–70-vuotiaat. Maassaoloajan mallin selittävien muuttujien jakaumien tunnusluvut.*

	Keskiarvo	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,47	0,00	0,00	1,00
Maahanmuuttoikä	24,31	15,00	21,00	28,00
Nainen $\times$ maahanmuuttoikä	11,51	0,00	0,00	21,00
Maahanmuuttoikä^2	7,04	2,25	4,41	7,84
Vähintään 2. asteen koulutus	0,05	0,00	0,00	0,00
Puoliso	0,27	0,00	0,00	1,00
Alle 7-vuotiaiden lasten lukumäärä	0,58	0,00	0,00	1,00
Julkisen talouden menot	3,77	0,38	3,30	6,14
Ei julkisen talouden tuloja	0,80	1,00	1,00	1,00
Pääasiallinen toiminta: työllinen	0,02	0,00	0,00	0,00
Pääasiallinen toiminta: opiskelija	0,33	0,00	0,00	1,00
Pääasiallinen toiminta: muu	0,65	0,00	1,00	1,00
Maahanmuuttovuosi 1988–1993	0,26	0,00	0,00	1,00
Maahanmuuttovuosi 1994–1999	0,25	0,00	0,00	1,00
Maahanmuuttovuosi 2000–2009	0,49	0,00	0,00	1,00

Estimointitulokset

Taulukossa 3.6 on esitetty Somaliassa syntyneiden estimoidun maassaolomallin parametrien posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulu-

kossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli yhtä, en käsittele niitä tarkemmin.

Regressiokertoimilla on taulukossa 3.6 *keskimäärin* standardin paranemisosuusmallin tulkinta, eli regressiokertoimet ovat log-vedonlyöntisuhteita (*engl. log-odds*). Tämä johtuu siitä, että ξ-parametrin posteriorijakauman odotusarvo on lähellä nollaa.

Esimerkiksi ”ei julkisen talouden tuloja”-indikaattorimuuttujan kohdalla tämä tarkoittaa sitä, että Suomesta poismuuton todennäköisyyden vedonlyöntisuhde on odotusarvoisesti noin 1,25-kertainen ($exp(0{,}22) \approx 1{,}25$) niillä, joilla ei ollut julkisen talouden tuloja (eli maksettuja veroja yms.) maahanmuuttovuonna, verrattuna sellaisiin henkilöihin, joilla tuloja oli. Mikäli Suomesta poismuuton todennäköisyyttä merkitään niillä, joilla julkisen talouden tuloja oli: $p(tuloja)$, ja puolestaan niiden osalta, joilla tuloja ei ollut: $p(\textit{ei tuloja})$, niin odotusarvoisesti:

$$\frac{p(\textit{ei tuloja})}{1 - p(\textit{ei tuloja})} \approx 1{,}25 \times \frac{p(tuloja)}{1 - p(tuloja)}.$$

Regressiokertoimien tulkintaan log-vedonlyöntisuhteina ei kuitenkaan kannata takertua liikaa, sillä ξ-parametrin posteriorijakauma on varsin leveä, eli suuri osa siitä ei sijaitse nollan ”välittömässä läheisyydessä”. Joka tapauksessa regressiokertoimien yleinen tulkinta tässä mallinnusluokassa on, että regressiokertoimien positiiviset arvot tarkoittavat sitä, että positiiviset muutokset niitä vastaavien selittävien muuttujien arvoissa merkitsevät suurempaa todennäköisyyttä muuttaa maasta.

Taulukon 3.6 regressiokertoimien posteriorijakaumia koskevien estimaattien perusteella Somaliassa syntyneiden maasta poismuuton todennäköisyyttä koskevat esimerkiksi seuraavat tulokset:

- Maahanmuuttoiän neliö: maastamuuton todennäköisyys on pienempää nuorimmissa ja vanhimmissa maahanmuuttoikäluokissa kuin niiden välillä olevissa ikäluokissa.
- Pääasiallinen toiminta: mikäli ollut maahanmuuttovuonna työllinen tai opiskelija, niin maastamuuton todennäköisyys on pienempää kuin niillä, joilla pääasiallinen toiminta oli jokin muu maahanmuuttovuonna.
- Maahanmuuttovuoden luokka: vuosina 1988–1993 ja 1994–1999 muuttaneilla suurempi poismuuton todennäköisyys kuin vuosina 2000–2009 muuttaneilla.
- ρ: Somaliassa syntyneiden (selittävillä muuttujilla ehdollistamaton) todennäköisyys jäädä Suomeen pysyvästi – eli ilman poismuuttoja – on noin 65 %:n luokkaa.

Taulukko 3.6: *Somaliassa syntyneiden15–70-vuotiaiden maassaoloajan mallinnus. Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

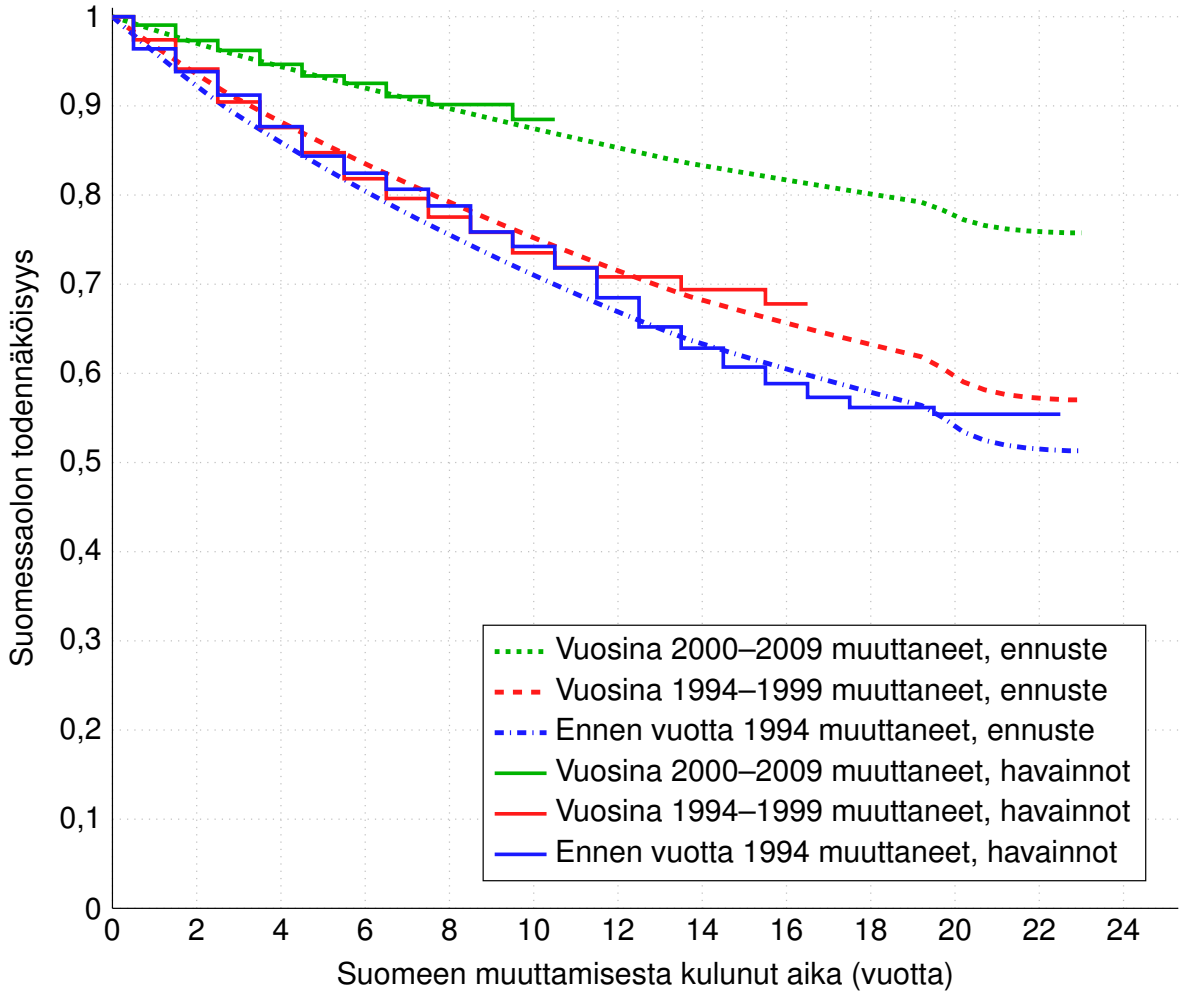
	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Nainen	0,02	0,07	-0,11	-0,02	0,02	0,07	0,16	0,0017	1,00
Maahanmuuttoikä	-0,020	0,007	-0,034	-0,024	-0,020	-0,015	-0,007	0,0002	1,00
Nainen $\times$ maahanmuuttoikä	0,01	0,03	-0,06	-0,01	0,01	0,04	0,08	0,0008	1,01
Maahanmuuttoikä^2	-0,028	0,007	-0,042	-0,033	-0,028	-0,023	-0,014	0,0002	1,00
Vähintään 2. asteen koulutus	-0,22	0,15	-0,51	-0,32	-0,22	-0,12	0,06	0,0034	1,00
Puoliso	-0,11	0,08	-0,28	-0,17	-0,11	-0,06	0,05	0,0020	1,00
Alle 7-vuotiaiden lasten lukumäärä	0,04	0,03	-0,01	0,02	0,05	0,07	0,11	0,0008	1,00
Julkisen talouden menot	-0,02	0,11	-0,24	-0,09	-0,02	0,06	0,21	0,0026	1,01
Ei julkisen talouden tuloja	0,22	0,09	0,04	0,15	0,22	0,28	0,40	0,0021	1,00
Pääasiallinen toiminta: työllinen	-0,30	0,31	-0,95	-0,51	-0,30	-0,09	0,26	0,0071	1,00
Pääasiallinen toiminta: opiskelija	-0,25	0,09	-0,43	-0,31	-0,25	-0,19	-0,07	0,0021	1,01
Maahanmuuttovuosi 1988–1993	0,99	0,10	0,80	0,92	0,99	1,05	1,19	0,0023	1,00
Maahanmuuttovuosi 1994–1999	0,79	0,10	0,59	0,72	0,79	0,85	0,99	0,0023	1,00
$\lambda_1[0;4{,}5)$	0,0053	0,0013	0,0031	0,0043	0,0052	0,0061	0,0082	0,00011	1,00
$\lambda_2[4{,}5;10{,}6)$	0,0046	0,0012	0,0028	0,0038	0,0046	0,0053	0,0072	0,00010	1,00
$\lambda_3[10{,}6;15{,}6)$	0,0047	0,0012	0,0028	0,0038	0,0046	0,0054	0,0074	0,00010	1,00
$\lambda_4[15{,}6;19{,}5]$	0,0038	0,0011	0,0022	0,0030	0,0037	0,0044	0,0063	0,00008	1,00
$\lambda_5[19{,}5;56]$	0,0180	0,0198	0,0005	0,0047	0,0114	0,0239	0,0753	0,00053	1,00
η	4,12	1,23	2,15	3,26	3,97	4,85	6,95	0,0826	1,07
ξ	-0,12	0,47	-1,08	-0,43	-0,11	0,19	0,77	0,0358	1,03
ρ	0,65	0,05	0,53	0,63	0,65	0,68	0,71	0,0010	1,00

Posterioriennustetarkistus

Somaliassa syntyneiden maassaoloajan posterioriennustetarkistuksissa vertaan havaintoaineiston selviytymisfunktioita maahanmuuttovuoden luokan mukaan posterioriennusteiden vastaaviin selviytymisfunktioihin. Valitsin maahanmuuttovuoden luokan tarkastelun kohteeksi, sillä sekä havaintoaineiston että taulukossa 3.6 esitettyjen estimointitulosten perusteella maahanmuuttovuoden luokka vaikuttaa maassaolon todennäköisyyteen.

Posterioriennustetarkistuksen tulokset on esitetty kuvassa 3.5. Olen esittänyt selviytymisfunktioiden tulkinnan ja mahdolliset havaintoaineistoon liittyvien ”hyppäysten” yhdistämiset, kuten selostin Irakissa syntyneiden posterioriennustetarkistusten kuvaa koskien (kuva 3.4, s. 99). Kuten kuviosta 3.5 ilmenee, posterioriennusteiden selviytymisfunktioiden odotusarvot vastaavat hyvin havaintoaineiston selviytymisfunktioita.

Kuviosta myös käy ilmi, että suomessaolon todennäköisyys, kun otetaan huomioon maassaoloaika, on suurempaa vuosina 2000–2009 muuttaneilla kuin tätä aiemmin muuttaneilla. Kaikilla ennen vuotta 2000 muuttaneilla suomessaolon todennäköisyys pysyi samalla tasolla noin 12 maassaolovuoteen asti, mutta sen jälkeen suurempi osa ennen vuotta 1994 muuttaneista on muuttanut pois kuin vuosina 1994–1999 Suomeen muuttaneista. Suomeen pysyvästi jäävien osuus on ennen vuotta 1994 muuttaneilla noin 50–55 %:ia ja ennusteen mukaan vuosina 1994–1999 muuttaneilla noin 60 %:ia. Ennusteen mukaan vuosina 2000–2009 muuttaneista noin 75 %:ia jää pysyvästi Suomeen. Suomeen jäävien osuudet ovat Somaliassa syntyneillä siten pienempiä kuin Irakissa syntyneillä (vrt. 3.4, s. 99).



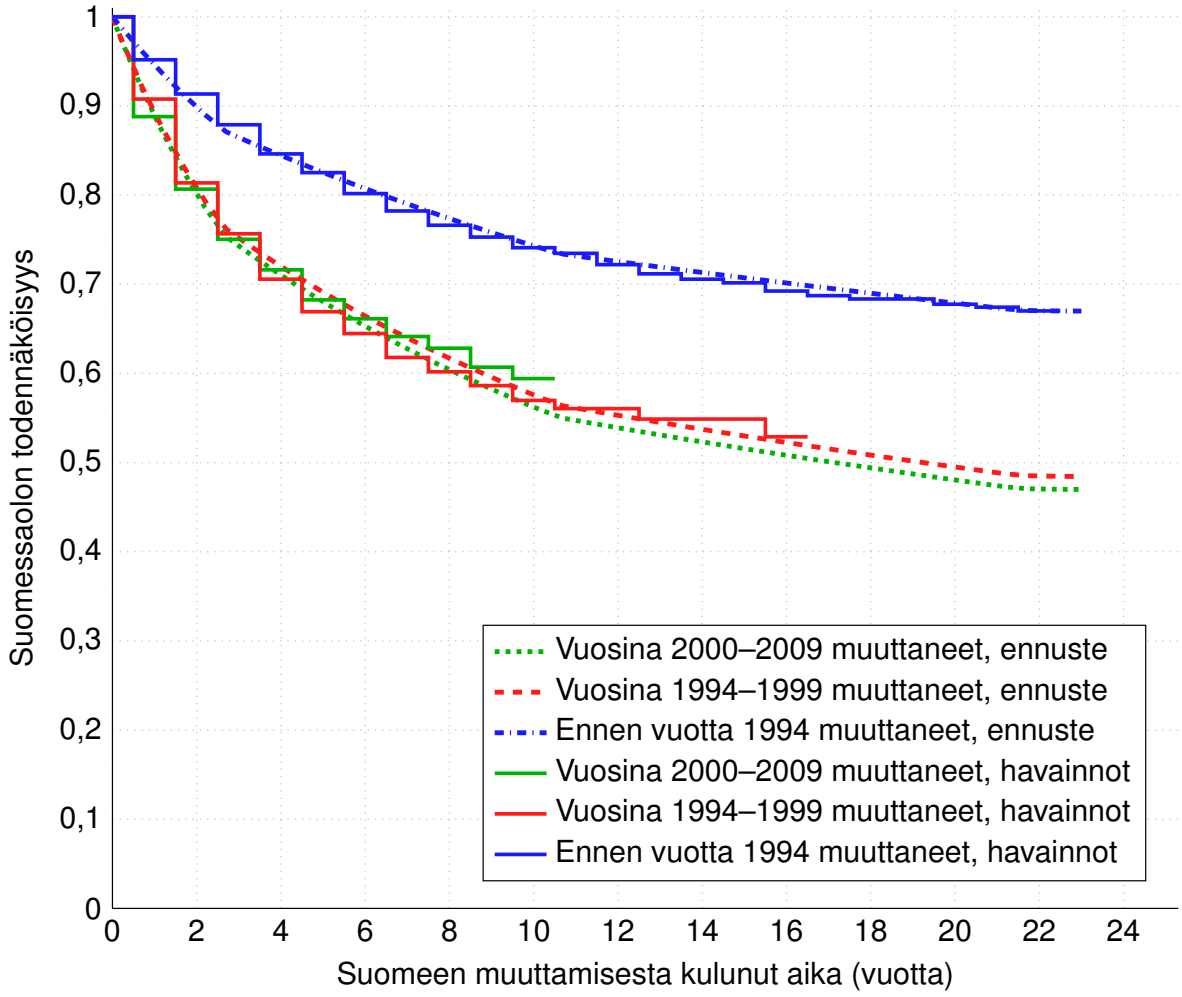
Kuva 3.5: *Posterioriennustetarkistus. Somaliassa syntyneet 15–70-vuotiaat. Suomessaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle maahanmuuttovuoden luokan mukaan.*

3.2.7.4 Posterioriennustetarkistukset: esimerkkejä muista syntymävaltioista

En julkaise tutkimuksen II osan tässä ensimmäisessä versiossa elinkaarivaikutusten tuloksia kuin Suomessa, Irakissa ja Somaliassa syntyneitä koskien. Julkaisen tässä kuitenkin muiden suurimpien maahanmuuttajaryhmien osalta joitakin kiinnostavimpia posterioriennustetarkistusten tuloksia kuvina. Kuvailen niissä olevia tuloksia hyvin lyhyesti. Niistä käy myös yleisesti ottaen ilmi, että posterioriennusteet vastaavat hyvin havaintoaineistoja.

Kuvassa 3.6 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston

selviytymisfunktiot Saksassa syntyneille maahanmuuttovuoden luokaan mukaan.

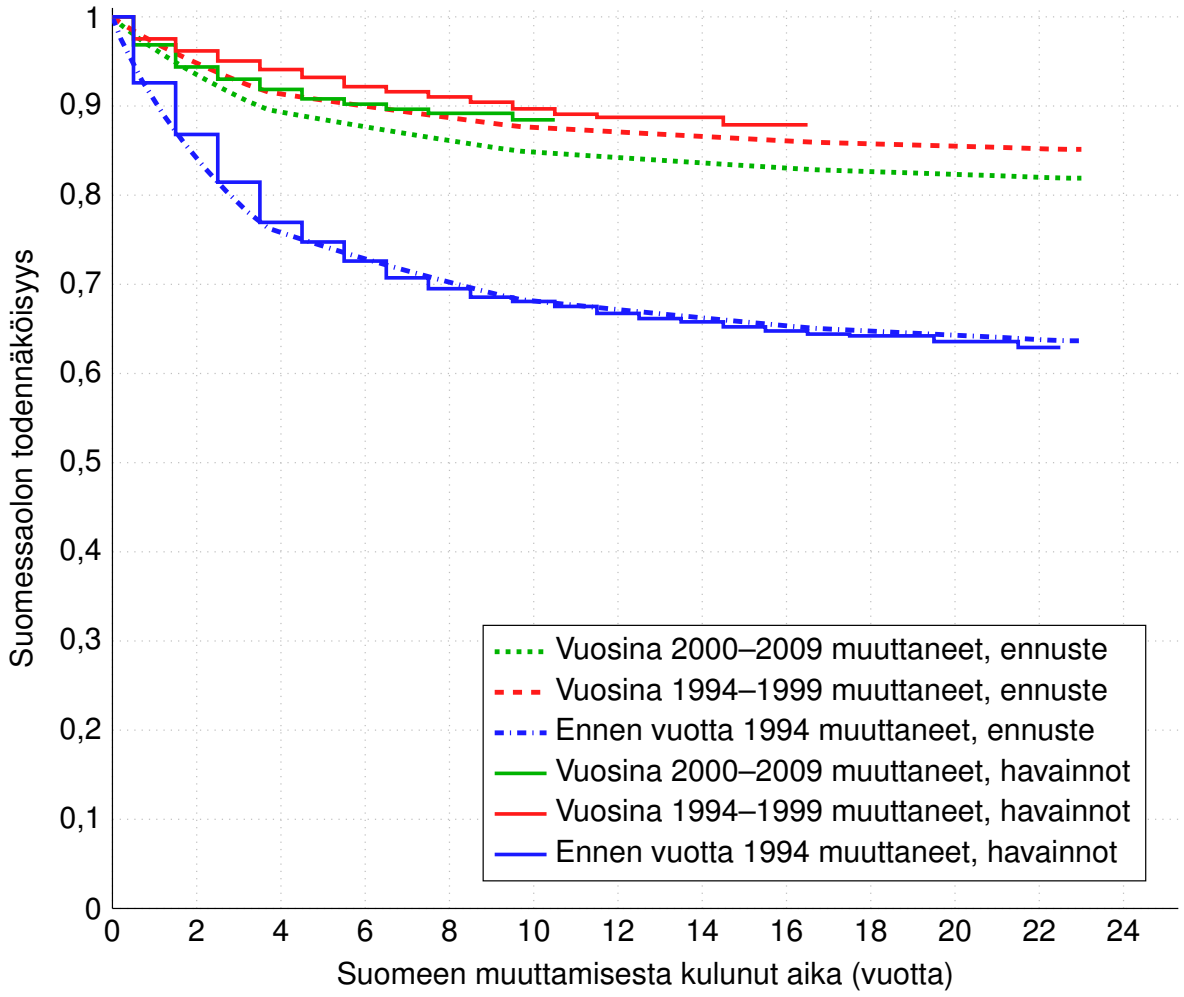


Kuva 3.6: *Posterioriennustetarkistus. Saksassa syntyneet 15–70-vuotiaat. Suomessaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle maahanmuuttovuoden luokan mukaan.*

Kuvan perusteella maastamuuton todennäköisyys olisi vastakkaista Somaliassa ja Irakissa syntyneisiin: aiemmin muuttaneilla (eli ennen vuotta 1994) poismuuttotodennäköisyys vaikuttaisi olevan pienempi kuin myöhemmin muuttaneilla. Todellisen eron sijaan kyse lienee kuitenkin erityisesti Saksassa syntyneiden aineistossa olevasta *vasemmasta katkaisusta* (*engl. left truncation*). Koska havaintoaineistossa maahanmuuttovuodet ovat tiedossa vain vuodesta 1988 alkaen – ja koska olen merkinnyt aineistoon ennen vuotta 1988 muuttaneet vuonna 1988 Suomeen muuttaneiksi – Saksassa syntyneiden aineistossa ennen vuotta 1994 muuttaneiden luokassa on henkilöitä, joiden maassaoloaika on todellista maassaoloaikaa pienempi. Koska ainakin populaatiotasolla maastamuuton todennäköisyys laskee maassaoloajan kasvaessa,

tämä saa aikaiseksi kuvan 3.6 ilmiön.

Kuvassa 3.7 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston selviytymisfunktiot entisessä Neuvostoliitossa syntyneille maahanmuuttovuoden luokaan mukaan.

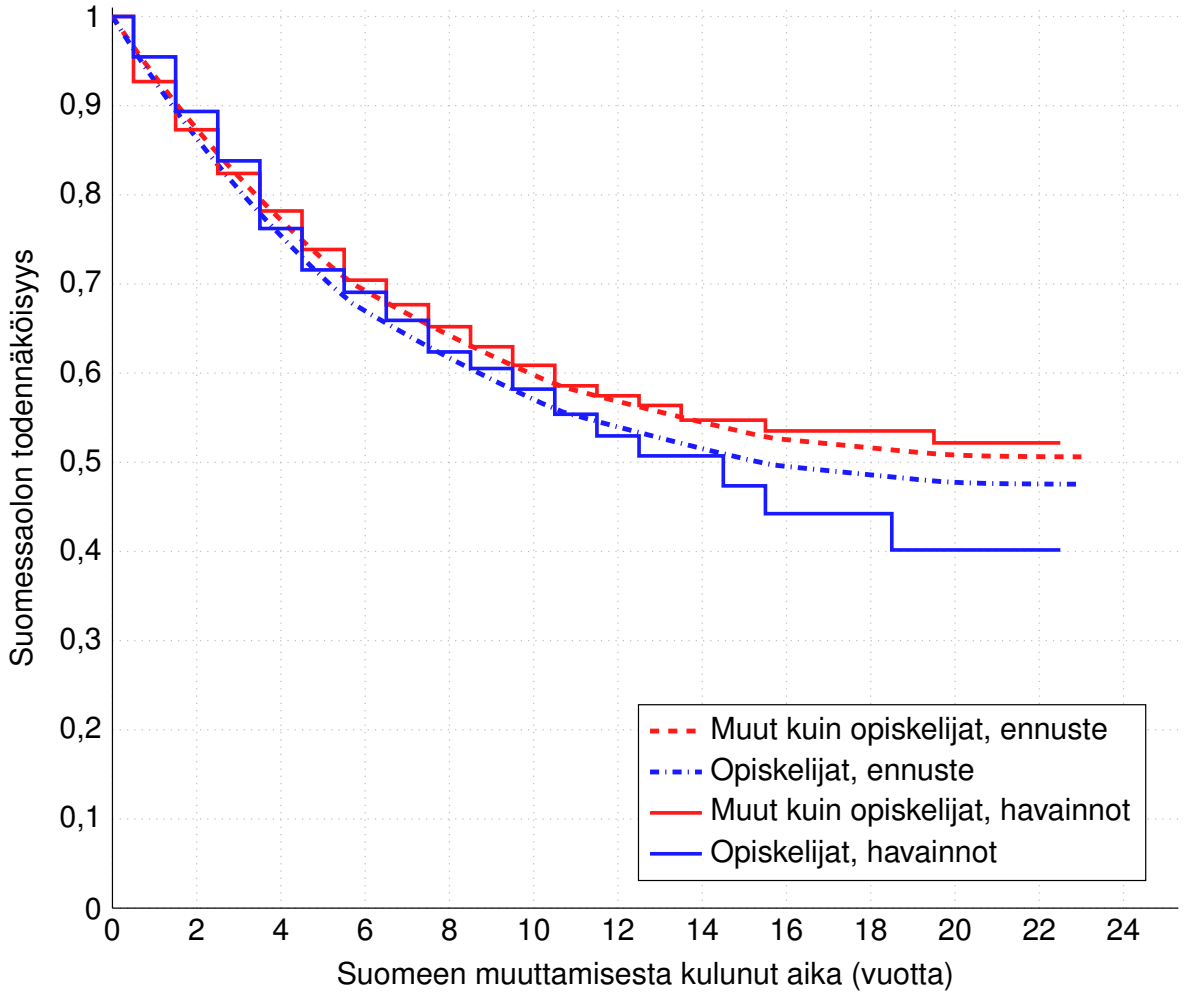


Kuva 3.7: *Posterioriennustetarkistus. Entisessä Neuvostoliitossa syntyneet 15–70-vuotiaat. Suomessaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle maahanmuuttovuoden luokan mukaan.*

Kuvan perusteella maastamuuton todennäköisyys on entisessä Neuvostoliitossa syntyneillä maahanmuuttovuosiluokan mukaan samanlaista kuin Somaliassa ja Irakissa syntyneillä: ennen vuotta 1994 muuttaneilla poismuuttotodennäköisyys on suurempi kuin myöhemmin muuttaneilla. Entisen Neuvostoliiton kohdalla ero on kuitenkin siinä, että sekä 1994–1999 muuttaneilla että 2000–2009 muuttaneilla pois-

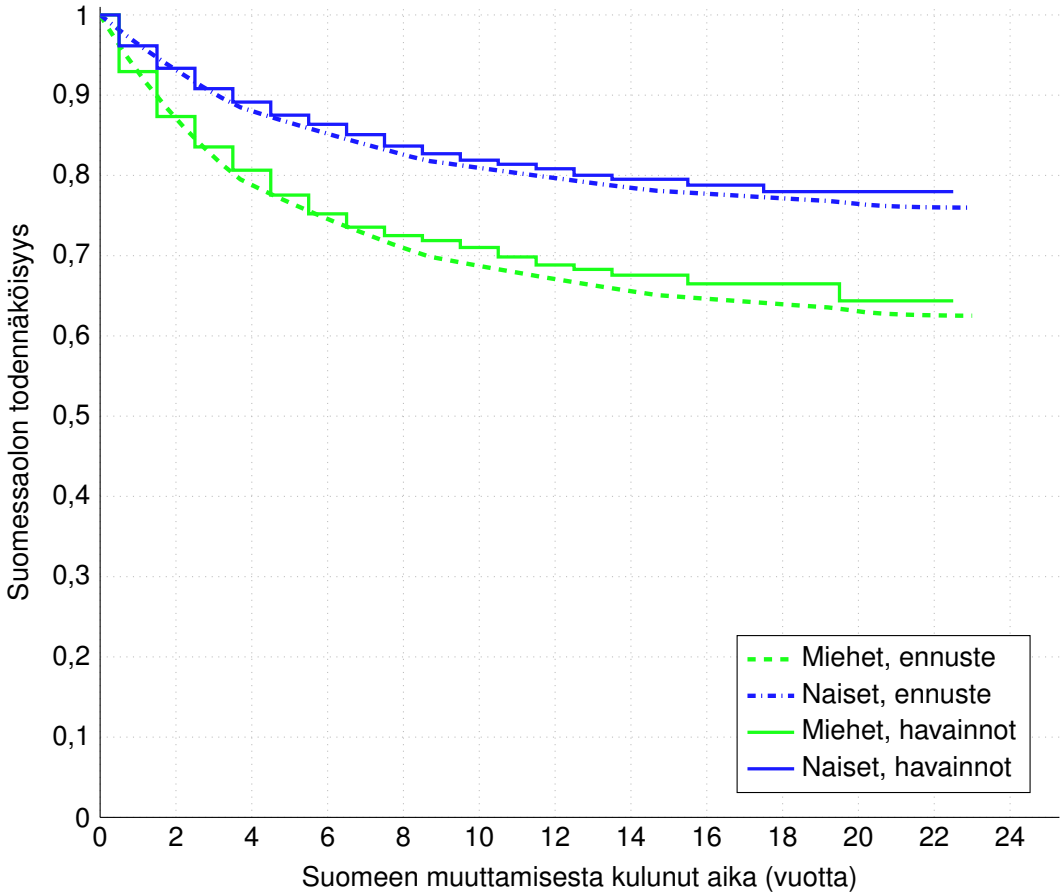
muuton todennäköisyys vaikuttaisi suunnilleen samansuuruiselta maassaoloajan suhteen. Kuvan 3.7 perusteella vaikuttaisikin siltä, että *ennen* Neuvostoliiton lakkaamista Suomeen muuttaneet ovat jostain syystä muuttaneet useammin pois kuin Neuvostoliiton lakkaamisen *jälkeen* Suomeen muuttaneet. Koska poismuuttaneisuus on suurta ennen vuotta 1994 muuttaneilla erityisesti muutaman ensimmäisen vuoden aikana – eli Neuvostoliiton lakkaamisen aikoihin – nämä ilmiöt voivat olla sidoksissa toisiinsa.

Kuvassa 3.8 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston selviytymisfunktiot Kiinassa syntyneille sen mukaan, onko henkilön pääasiallinen toiminta ollut maahanmuuttovuonna opiskelija vai muu kuin opiskelija.



Kuva 3.8: *Posterioriennustetarkistus. Kiinassa syntyneet 15–70-vuotiaat. Suomessaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle sen mukaan, oliko opiskelija maahanmuuttovuonna.*

Tutkimuksen I osan liitetaulukon 2 mukaan 10 suurimman maahanmuuttajaryhmän joukossa Kiinassa syntyneillä opiskelijoiden osuus on suurin työikäisestä väestöstä. Päätin siksi tarkastella Kiinassa syntyneiden osalta, että onko sillä vaikutusta poismuuton todennäköisyyteen, jos maahanmuuttovuonna henkilön pääasiallinen toiminta on ollut opiskelija. Kuvan 3.8 perusteella poismuuttaneisuudessa ei ole eroa ensimmäisen viiden maassaolovuoden aikana. Tämän jälkeen opiskelijoina Suomeen tulleista useampi vaikuttaa muuttaneen pois kuin niistä, jotka eivät olleet opiskelijoita maahanmuuttovuonna. Kuvan perusteella opiskelijoina Suomeen tulleista jää maahan pysyvästi noin 40 %:ia ja muista Kiinassa syntyneistä noin 50 %:ia.



Kuva 3.9: *Posterioriennustetarkistus. Virossa syntyneet 15–70-vuotiaat. Suomessaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle sukupuolen mukaan.*

Kuvassa 3.9 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston selviytymisfunktiot Virossa syntyneille sukupuolen mukaan. Kuvan perusteella Virossa syntyneet miehet muuttavat useammin pois Suomesta kuin Virossa syntyneet naiset.

3.3 Ulkomaillaoloaikojen mallinnus

3.3.1 Ulkomaillaoloaikojen mallinnus De Castron et al. (2014) menetelmällä

Koska osa Suomesta (takaisin) ulkomaille muuttaneista palaa takaisin Suomeen, myös ulkomaillaoloajat on syytä mallintaa ja mallinnustulosten pohjalta ennustaa. Ulkomaillaoloaika tarkoittaa aikaa ulkomailla Suomesta poismuuton jälkeen siihen asti, kunnes henkilö joko muuttaa takaisin Suomeen tai kuolee.

Ulkomaillaoloaikojen aineistoa koskee samanlainen oikeanpuoleinen sensorointi kuten suomessaoloaikoja: henkilö sensoroidaan ulkomaillaoloajan aineistosta, mikäli hän on ulkomailla sinä vuonna, kun hän täyttää 71 vuotta tai hän oli ulkomailla vuoden 2009 lopussa. Jälkimmäisessä tapauksessa tutkimusaineistossa ulkomaillaolon sensorointivuodeksi tuli vuosi 2010. Mallinnuksessa käsittelin sensorointina myös henkilön kuolemaa. Kuten maassaolomallinuksen kohdalla, myös ulkomaillaoloaikojen kohdalla oikeanpuoleisen sensoroinnin olettaminen riippumattomaksi takaisin Suomeen muuttamisesta, on mielestäni realistista.

Mallinsin ulkomaillaoloajat samaa De Castron et al. (2014) artikkelin tarjoamaa yleisen paranemisosuusmallien luokkkaa käyttäen, kuin mallinsin myös maassaoloajat Suomessa. Ulkomaillaoloaikamallien yhteydessä ”paranemisosuus” tarkoittaa sitä osuutta Suomesta poismuuttaneista, jotka eivät enää muuta takaisin Suomeen.

Ulkomaillaoloaikamalleihin valitsin osan samoista selittävistä muuttujista kuin, mitä valitsin suomessaoloaikojen malleihin. Näiden lisäksi valitsin selittäväksi muuttujaksi henkilön kumulatiivisen suomessaoloajan, koska oletin, että suomessaoloaika kertoo sidoksista Suomeen: mitä pitempään henkilö oli ollut Suomessa, sitä todennäköisempää olisi, että hän muuttaisi myös takaisin Suomeen. On syytä huomioida, että selittävien muuttujien arvot koskevat – Suomessa henkilön yhteensä viettämää aikaa lukuun ottamatta – ensimmäisen Suomeen muuton vuotta. Koska aineistossa ei ole paljon henkilöitä, jotka olisivat muuttaneet pois Suomesta 10 maassaolovuoden jälkeen, oletin varmuuden vuoksi, ettei kumulatiivisen suomessaoloajan vaikutus enää muutu 10 Suomessa vietetyn vuoden jälkeen.

Selittävät muuttujat sisältävät jokaisen henkilön kohdalla tietoja henkilöstä siltä vuodelta, kun hän muuttaa ensimmäisen kerran Suomeen. Alle 15-vuotiaana muuttaneiden kohdalla tiedot koskevat sitä vuotta, jolloin henkilö täytti 15 vuotta. Selittävät muuttujat, jotka valitsin maahanmuuttajaryhmäkohtaisiin ulkomaillaoloaikamalleihin, olivat:

– *Nainen* (referenssiluokka: mies).

– *Maahanmuuttoikä*.

– *Puoliso* (referenssiluokka: ei puolisoa).

– *Suomessa syntynyt puoliso* (referenssiluokka: ei Suomessa syntynyttä puolisoa).

– *Pääasiallinen toiminta: työllinen* (referenssiluokka: muu kuin työllinen tai opiskelija).

– *Pääasiallinen toiminta: opiskelija* (referenssiluokka: muu kuin työllinen tai opiskelija).

– *Maahanmuuttovuosi 1988–1993* (referenssiluokka: maahanmuuttovuosi 2000–2009).

– *Maahanmuuttovuosi 1994–1999* (referenssiluokka: maahanmuuttovuosi 2000–2009).

– *Suomessaoloaika yhteensä*.

Valintaperusteet selittäville muuttujille olivat samat kuin suomessaoloaikamallien yhteydessä: maahanmuuttajaryhmäkohtaiset aineiston eri selittävien muuttujien arvojen mukaan jaotellut Kaplan-Meier-estimaateien kuviot, muuttujien yleinen esiintyvyys yhteiskuntatieteellisessä tutkimuksessa ilmiöitä selittävinä tekijöinä sekä se, miten selittäjät kuvaavat sosiaalisia ja taloudellisia siteitä Suomeen.

Kaplan-Meier-estimaattien kuvioiden avulla tarkastelin, mitkä olivat niitä selittäviä muuttujia, joiden eri arvoilla selviytymisfunktiot vaikuttivat poikkeavan toisistaan. Yhteiskuntatieteellisessä tutkimuksessa ainakin sukupuoli, ikä ja koulutus ovat sellaisia tekijöitä, joiden mahdollisia yhteyksiä ilmiöihin ainakin yleensä tutkitaan. Kuten olen edellä maininnut Bodvarssoniin ja Van den Bergiin (2013) viitaten, sosiaaliset ja taloudelliset siteet voivat vaikuttaa maassaoloaikaan.

Selittävät muuttujat, jotka olivat mukana suomessaoloaikamalleissa, mutta jotka jätin pois ulkomaillaoloaikamalleista, olivat: interaktio *Nainen* $\times$ *maahanmuuttoikä*, *Maahanmuuttoikä*^{2}, *Korkein tiedossa oleva koulutus: vähintään 2. asteen koulutus*, *Alle 7-vuotiaiden lasten lukumäärä perheessä*, *Julkisen talouden nollatulot ensimmäisenä vuonna Suomessa* sekä *Julkisen talouden menot ensimmäisenä vuonna Suomessa*. Nämä muuttujat eivät vaikuttaneet joko Kaplan-Meier-estimaattien tai joidenkin maahanmuuttajaryhmien kohdalla tehdyillä kokeilumalleilla potentiaalisilta selittäjiltä. Edellisen kohdalla K-M-estimaatit eivät eronneet silmämääräisesti eri selittävien muuttujien arvoilla toisistaan. Jälkimmäisen kohdalla näiden selittäjien posteriorijakaumat olivat keskittyneet nollan ympärille.

3.3.2 Havainnot, estimointitulokset ja posterioriennusteet

Esitän tässä luvussa ulkomaillaoloaikoja koskevat havainnot, syntymävaltiokohtaiset mallinnustulokset sekä posterioriennustetarkistukset. Aloitan esittämällä syntymävaltiokohtaiset mallinnusaineistojen havaintolukumäärät. Sen lisäksi esitän ulkomaillaolomallien parametrien selitetaulukon, joka toimii apuna syntymävaltiokohtaisten mallien tulosten tarkasteluun. Olen esittänyt malliparametrien tarkemman selostuksen sekä valintaperiatteet edellä.

Koska ulkomaillaolomallien estimointi tapahtui syntymävaltiokohtaisesti, käyn estimointitulokset ja posterioriennustetarkistukset myös syntymävaltio kerrallaan läpi. Esitän jokaisen syntymävaltion kohdalla aluksi niiden mallinnusaineistojen selittävien muuttujien jakaumien tärkeimmät tunnusluvut: keskiarvot ja kvartiilit. En esittele erikseen selitettävää muuttujaa eli ulkomaillaoloaikojen havaintoja – ja sensorointeja – sillä ulkomaillaoloaikojen havaintoaineisto esitetään käytännössä posterioriennustetarkistusten yhteydessä.

Posterioriennustetarkistuksissa vertailen havaintoaineiston selviytymisfunktiota estimoitujen mallien pohjalta laadittujen ulkomaillaoloajan posterioriennusteiden selviytymisfunktioiden odotusarvoihin. Kuten osoitan kuvioiden avulla, posterioriennusteiden selviytymisfunktiot vastaavat hyvin havaintoaineistojen selviytymisfunktioita. Tämän perusteella pidän estimoitujen mallien tuottamia ulkomaillaoloajan ennusteita luotettavina ulkomaillaoloajan ennusteina myös uusien maahanmuuttajien elinkaarivaikutusten ennusteissa.

Posterioriennustarkistuksissa vertaan posterioriennusteita havaintoihin joko kaikkien havaintojen osalta tai selittävien muuttujien luokkien mukaan jaoteltuina. Mikäli teen vertailun selittävien muuttujien luokkien perusteella, olen valinnut vertailuun sellaisen selittävän muuttujan, jolla on estimointitulosten perusteella ”paljon” vaikutusta ulkomaillaoloaikaan.

3.3.2.1 Havaintolukumäärät ja parametrien selitteet

Taulukossa 3.7 on esitetty syntymävaltiokohtaiset ulkomaillaoloajan mallinnuksen havaintojen ja henkilöiden lukumäärät. Henkilöt ovat aineistossa 15–70-vuotiaita, kun he ovat muuttaneet Suomeen. Alle 15-vuotiaana todellisuudessa muuttaneilla muuttoikä on 15. Havainnot muodostuvat joko Suomeen suuntautuvien paluumuuttojen tiedoista tai sensoirointitiedoista. Havainto sensorointiin, mikäli henkilö täytti 71 vuotta tai seuranta-aika jatkui vuoteen 2010 ilman paluumuuttoa tai sensorointia muusta syystä.

Havaintoja on jokaisen syntymävaltion kohdalla hieman enemmän kuin henkilöitä. Tämä johtuu siitä, että havaintoaineisto sisältää sellaisia henkilöitä, joita koskien on useampi kuin yksi havainto. Tällaisissa tapauksissa henkilö on paluumuuttanut aluksi Suomeen, mutta sittemmin hän on muuttanut jälleen pois Suomesta. Jokainen ”uusi” kerta Suomesta poismuutolle luo aineistoon aina uuden havainnon.

Taulukossa 3.8 on esitetty ulkomaillaoloajan syntymävaltiokohtaisten mallien estimoitavien parametrien nimet ja niiden selitteet. Parametrit on esitetty siinä järjestyksessä, missä ne esitetään parametrien

Taulukko 3.7: *Ulkomaillaoloajan mallinnusaineistojen havaintojen ja henkilöiden lukumäärät.*

Syntymävaltio	**Havainnot**	**Henkilöt**
Irak	446	432
Somalia	1 328	1 285
Thaimaa	375	364
Turkki	577	551
Entinen Jugoslavia	651	626
Viro	1 474	1 433
Kiina	1 861	1 817
Entinen Neuvostoliitto	1 881	1 816
Saksa	1 971	1 906
Ruotsi	2 283	2 169

estimointitulosten syntymävaltiokohtaisissa taulukoissa jäljempänä. Parametrit, joista ensimmäinen on ”nainen”, ja viimeinen on, ”Suomessa syntynyt puoliso”, ovat selittävien muuttujien regressiokertoimia. Parametrien nimet ja selitteet kertovat, mihin selittävään muuttujaan kukin regressiokerroin liittyy. Indikaattorimuuttujien osalta selitetaulukossa on esitetty näiden vertailuluokat. Suomessaoloaikaa lukuun ottamatta selittävien muuttujien tiedot koskevat henkilöä, kun hän on muuttanut ensimäisen kerran *Suomeen*. Suomessaoloaika on henkilön yhteensä Suomessa viettämä aika, kun hän muuttaa Suomesta ulkomaille. Tämä aika on voinut karttua useamman kuin yhden Suomessa vietetyn ajanjakson aikana.

Intervallikohtaiset hasardin vertailutasot ovat parametreja λ. Estimointituloksissa jokaisen niiden mukana on kaksi numeroa, jotka määrittävät sen ulkomaillaoloajan intervallin, jossa kyseinen parametri on voimassa. Intervallikohtaisia hasardiparametreja on syntymävaltiokohtaisissa malleissa 3–4 kappaletta. Muut kaksi malliparametria ovat frailty-parametrin muunnos η sekä mallin muunosparametrin muunnos ξ. Sisällytän estimointituloksiin vielä ρ:n, joka ei ole varsinainen malliparametri, sillä se lasketaan muiden parametrien tuottamista tuloksista. Se on kuitenkin mukana tuloksissa, sillä se on paranemisosuusmallin ”paranemisosuus”. Ulkomaillaoloajan mallinnuksen yhteydessä se on todennäköisyys jäädä ulkomaille pysyvästi (eli ei paluumuuttoa Suomeen), kun selittävien muuttujien vaikutusta ei huomioida.

Taulukko 3.8: *Ulkomaillaoloajan mallien estimoitavien parametrien nimet ja niiden selitteet.*

Parametrinimi	**Selite**
Nainen	Vertailuluokka: mies.
Suomessaoloaika yhteensä	Suomessa yhteensä oltu aika 15–70-vuotiaana, vuosia.
Maahanmuuttoikä	
Pääasiallinen toiminta: työllinen	Vertailuluokka: pääasiallinen toiminta: muu.
Pääasiallinen toiminta: opiskelija	Vertailuluokka: pääasiallinen toiminta: muu.
Maahanmuuttovuosi v. 1988–1993	Regressiokerroin. Vertailuluokka: Maahanmuuttovuosi v. 2000–2009.
Maahanmuuttovuosi v. 1994–1999	Regressiokerroin. Vertailuluokka: Maahanmuuttovuosi v. 2000–2009.
Puoliso	Vertailuluokka: ei puolisoa.
Suomessa syntynyt puoliso	Vertailuluokka: ei Suomessa syntynyttä puolisoa.
$\lambda_{nro}(alku; loppu)$	Hasardin intervallikohtainen vertailutaso. *nro* on intervallin järjestysluku, *alku* on intervallin alkupiste ja *loppu* on intervallin loppupiste.
η	Parametreilla on muunnos: $\eta = 1/\theta$. Henkilötasoinen frailty on $gamma(1/\theta, 1/\theta)$-jakautunut.
ξ	Parametreilla on muunnos: $\xi = log(r^*)$, missä $r^* = r/\theta$ ja r on mallin muunnosparametri.
ρ	ρ on mallin "paranemisosuus"eli ulkomaille muuton vuoden (ehdollistamaton) todennäköisyys jäädä ulkomaille pysyvästi.

3.3.2.2 Estimointitulokset ja posterioriennustetarkistukset: Irak

Taulukossa 3.9 on esitetty Irakissa syntyneiden ulkomaillaolomallin selittävien muuttujien jakaumien keskiarvot ja kvartiilit. Tiedot koskevat ensimmäisen *Suomeen* muuton vuotta. Vain yhteensä karttunut suomessaoloaika on tieto siltä hetkeltä, kun henkilö on muuttanut Suomesta ulkomaille. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden mallinnusväestössä: esimerkiksi ”puolison” keskiarvo kertoo, että 41 %:lla oli puoliso, kun he muuttivat Suomeen.

Regressiokertoimilla ei ole taulukossa 3.10 standardin paranemisosuusmallin tai BCH-mallin kaltaista selkeää tulkintaa. Tämä johtuu siitä, että ξ-parametrin posteriorijakauma ei ole keskittynyt ”pieniin” arvoihin – jolloin regressiokertoimilla olisi BCH-mallin suhteellisten hasardien tulkinta – eikä myöskään nollan ”läheisyyteen” – jolloin puolestaan regressiokertoimilla olisi standardin paranemisosuusmallin (log-)vedonlyöntisuhteiden tulkinta.

Regressiokertoimet kuitenkin kertovat, mihin suuntaan muutokset selittävät muuttujissa vaikuttavat ulkomaillaolon pysymisen todennäköisyyteen. Kun regressiokertoimet saavat positiivisia arvoja, tämä tarkoit-

Taulukko 3.9: *Irakissa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan mallin selittävien muuttujien jakaumien tunnusluvut.*

	Keskiarvo	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,45	0,00	0,00	1,00
Suomessaoloaika (vuosia)	6,35	2,50	5,50	9,50
Maahanmuuttoikä	28,07	19,00	27,00	34,00
Pääasiallinen toiminta: työllinen	0,04	0,00	0,00	0,00
Pääasiallinen toiminta: opiskelija	0,20	0,00	0,00	0,00
Pääasiallinen toiminta: muu	0,76	0,00	1,00	1,00
Maahanmuuttovuosi 1988–1993	0,29	0,00	0,00	1,00
Maahanmuuttovuosi 1994–1999	0,42	0,00	0,00	1,00
Maahanmuuttovuosi 2000–2009	0,29	0,00	0,00	1,00
Puoliso	0,41	0,00	0,00	1,00

taa sitä, että positiivinen muutos niitä vastaavissa selittävissä muuttujissa *lisää* Suomeen paluumuuton todennäköisyyttä. Samaten, mikäli regressiokertoimet saavat negatiivisia arvoja, tämä tarkoittaa sitä, että positiivinen muutos niitä vastaavissa selittävissä muuttujissa *vähentää* Suomeen paluumuuton todennäköisyyttä.

Taulukko 3.10: *Irakissa syntyneiden15–70-vuotiaiden ulkomaillaoloajan mallinnus. Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Nainen	-0,32	0,35	-1,03	-0,54	-0,32	-0,10	0,38	0,0078	1,00
Suomessaoloaika (vuosia)	0,174	0,046	0,088	0,144	0,172	0,203	0,267	0,0011	1,00
Maahanmuuttoikä	-0,041	0,022	-0,088	-0,055	-0,040	-0,026	-0,002	0,0005	1,00
Pääasiallinen toiminta: työllinen	0,46	0,78	-1,25	0,00	0,49	0,98	1,89	0,0179	1,00
Pääasiallinen toiminta: opiskelija	-0,38	0,42	-1,25	-0,65	-0,36	-0,09	0,43	0,0097	1,00
Maahanmuuttovuosi 1988–1993	-0,30	0,55	-1,36	-0,68	-0,30	0,05	0,83	0,0126	1,00
Maahanmuuttovuosi 1994–1999	-0,35	0,48	-1,28	-0,67	-0,35	-0,03	0,59	0,0096	1,00
Puoliso	-0,20	0,37	-0,92	-0,45	-0,20	0,04	0,58	0,0085	1,01
$\lambda_1[0;4{,}5)$	0,0135	0,0067	0,0051	0,0089	0,0121	0,0162	0,0302	0,00019	1,00
$\lambda_2[4{,}5;10{,}5)$	0,0116	0,0066	0,0035	0,0069	0,0101	0,0141	0,0297	0,00018	1,00
$\lambda_3[10{,}5;15{,}5)$	0,0094	0,0070	0,0022	0,0049	0,0075	0,0116	0,0274	0,00017	1,00
$\lambda_3[15{,}5;23)$	0,0239	0,0317	0,0005	0,0048	0,0136	0,0293	0,1117	0,00076	1,02
η	1,63	0,96	0,44	0,94	1,39	2,03	4,11	0,0268	1,01
ξ	-0,39	0,84	-2,09	-0,97	-0,39	0,19	1,19	0,0246	1,00
ρ	0,76	0,07	0,60	0,72	0,77	0,81	0,87	0,0015	1,00

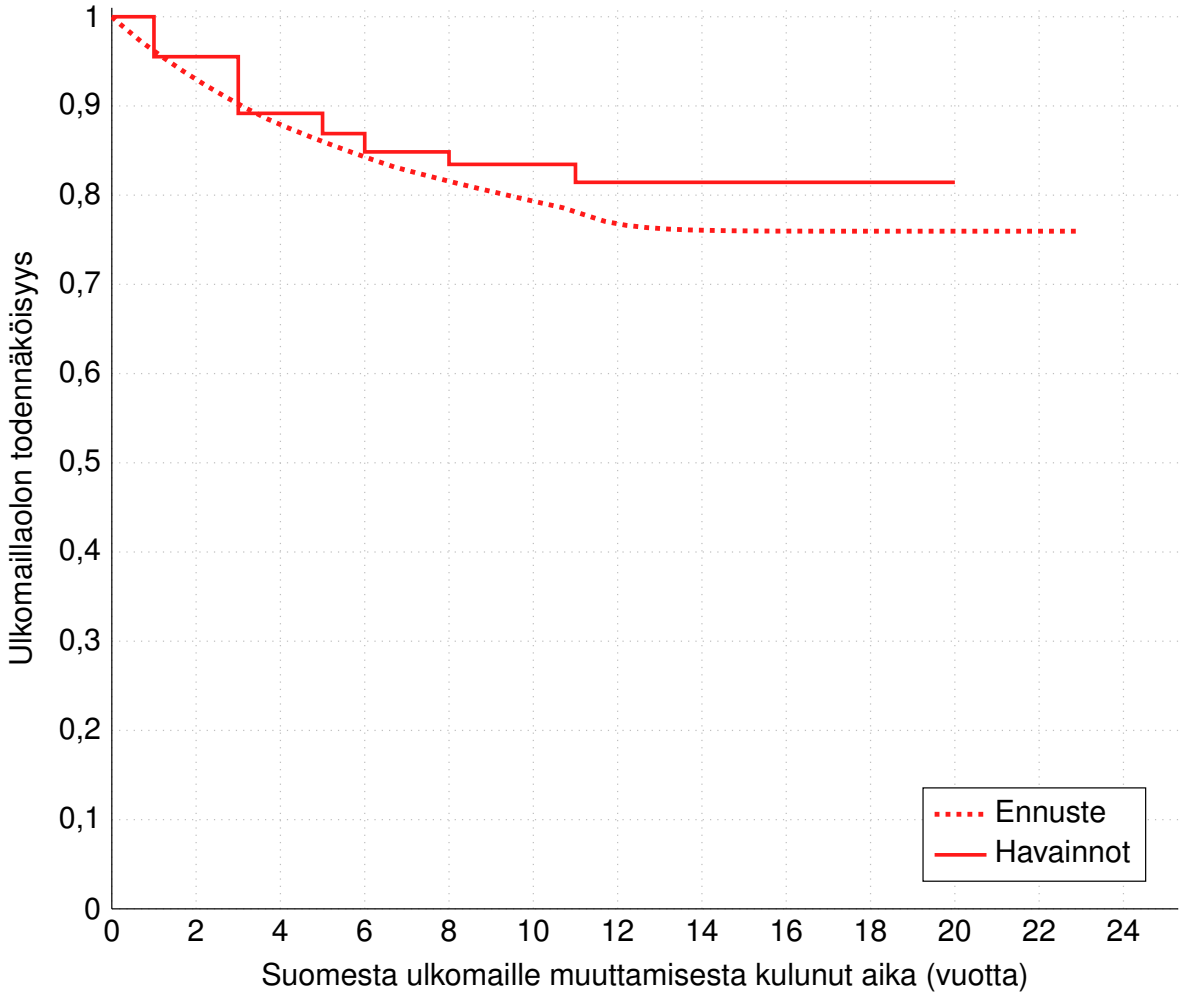
Taulukon 3.10 tulokset viittaavat siihen, että mikäli henkilö on nainen, hän muuttaa epätodennäköisemmin takaisin Suomeen kuin siinä tapauksessa, että hän on mies. Lisäksi vaikuttaa siltä, että mitä nuorempana henkilö on muuttanut alun perin Suomeen, sitä epätodennäköisempää on, että hän paluumuuttaa takaisin Suomeen. Selkein – ja odotettavissa oleva – vaikutus paluumuuton todennäköisyyteen on Suomessa vietetyllä ajalla: Mitä pidempään henkilö on ehtinyt yhteensä asua Suomessa, sitä todennäköisempää on, että hän paluumuuttaa takaisin Suomeen ulkomailta.

Taulukon 3.10 mukaan ρ:n eli ”paranemisosuuden” jakauma sijoittuu n. 95 %:sti suunnilleen 0,60:n ja 0,90:n välille. Odotusarvo on 0,76. Tämä tarkoittaa sitä, että havaintoaineistossa Irakissa syntyneiden (selittävillä muuttujilla ehdollistamaton) todennäköisyyden odotusarvo jäädä ulkomaille pysyvästi – eli ilman paluumuuttoa Suomeen – on noin 76 %:ia.

Posterioriennustetarkistus

Irakissa syntyneiden ulkomaillaoloajan posterioriennustetarkistuksissa vertaan havaintoaineiston selviytymisfunktiota posterioriennusteiden selviytymisfunktion odotusarvoon. Posterioriennustetarkistuksen tulokset on esitetty kuvassa 3.10. Kiinteät, portaittaiset viivat esittävät selviytymisfunktion – eli ulkomaillaolon todennäköisyyttä ulkomaillaoloajan mukaan. ”Hyppäykset” tai portaat havaintoaineiston kuvaajissa vastaavat henkilöiden paluumuuttoja Suomeen tietyn ulkomaillaoloajan jälkeen.

Mikäli havaintoaineistossa yksi ”hyppäys” tai porras aiheutuu alle 5 henkilön paluumuutosta, olen yhdistänyt yksityisyydensuojan vuoksi useamman perättäisen hyppäyksen yhdeksi hyppäykseksi siten, että paluumuutot vastaavat vähintään 5 henkilöä. Tällöin olen sijoittanut hyppäyksen suuremman ulkomaillaoloajan arvon kohdalle. Esimerkiksi, jos 7 ja 8 ulkomaillaolovuoden välillä kaksi henkilöä on paluumuuttanut Suomeen ja 8 ja 9 ulkomaillaolovuoden välillä 3 henkilöä on paluumuuttanut Suomeen, olen tulkinnut tämän kuvioon niin, että viisi henkilöä on paluumuuttanut 8 ja 9 ulkomaillaolovuoden välillä.



Kuva 3.10: *Posterioriennustetarkistus. Irakissa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle.*

Jatkuvat, katkonaiset viivat esittävät kuviossa puolestaan estimoidusta ulkomaillaoloajan mallista tuotettujen posterioriennusteiden selviytymisfunktioiden odotusarvoa. Kuten kuviosta 3.10 ilmenee, posterioriennusteiden selviytymisfunktioiden odotusarvot vastaavat kohtalaisen hyvin havaintoaineiston selviytymisfunktioita.

Irakin, kuten usean muun syntymävaltion kohdalla, posterioriennusteiden selviytymisfunktio vaikuttaa olevan hieman alempana kuin havaintoaineiston selviytymisfunktio. Tämä tarkoittaa sitä, että posterioriennusteen mukaan hieman useampi paluumuuttaa Suomeen kuin, mitä havaintoaineisto antaa ymmärtää. Tämä pätee erityisesti viiden ulkomaillaoloajan jälkeisille vuosille.

Oletan, että tämä ero johtuu siitä, että ulkomaillaoloajan havaintoaineistossa niillä henkilöillä, joilla on

paljon ulkomaillaoloajan vuosia, täytyy olla vain vähän suomessaoloajan vuosia. Esimerkiksi Irakissa syntyneiden kohdalla ne henkilöt, joilla on yli 10 ulkomaillaolovuotta, ovat ”ehtineet” olla Suomessa maksimissaan noin viisi vuotta. Koska suomessaoloaika vaikuttaa paluumuuton todennäköisyyteen, näillä ”aikaisin” Suomesta pois muuttaneilla todennäköisyys muuttaa Suomeen on pienempi kuin henkilöillä, joilla on enemmän suomessaoloajan vuosia. Ulkomaillaoloajan havaintoaineistossa sellaiset henkilöt, joilla voi olla paljonkin suomessaoloajan vuosia, ovat keskimäärin ulkomaillaoloajan alkupäässä. Posterioriennusteissa kuitenkin näillekin henkilöille laaditaan paluumuuton ennusteet koko mahdolliselle ulkomaillaoloajalle. Tämä ”laskee” posterioriennusteiden selviytymisfunktion kuvaajan suuremmilla ulkomaillaoloajan vuosilla havaintoaineiston kuvaajan alle.

3.3.2.3 Estimointitulokset ja posterioriennustetarkistukset: Somalia

Taulukossa 3.11 on esitetty Somaliassa syntyneiden ulkomaillaolomallin selittävien muuttujien jakaumien keskiarvot ja kvartiilit. Tiedot koskevat ensimmäisen *Suomeen* muuton vuotta. Vain yhteensä karttunut suomessaoloaika on tieto siltä hetkeltä, kun henkilö muuttaa Suomesta ulkomaille. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden mallinnusväestössä: esimerkiksi ”puolison” keskiarvo kertoo, että 28 %:lla oli puoliso, kun he muuttivat Suomeen.

Taulukko 3.11: *Somaliassa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan mallin selittävien muuttujien jakaumien tunnusluvut.*

	Keskiarvo	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,46	0,00	0,00	1,00
Suomessaoloaika (vuosia)	7,04	3,50	6,50	10,50
Maahanmuuttoikä	23,65	17,00	22,00	27,00
Maahanmuuttovuosi 1988–1993	0,50	0,00	1,00	1,00
Maahanmuuttovuosi 1994–1999	0,35	0,00	0,00	1,00
Maahanmuuttovuosi 2000–2009	0,15	0,00	0,00	0,00
Puoliso	0,28	0,00	0,00	1,00

Estimointitulokset

Taulukossa 3.12 on esitetty Somaliassa syntyneiden estimoidun ulkomaillaolomallin parametrien posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli yhtä, en käsittele niitä tarkemmin.

Taulukko 3.12: *Somaliassa syntyneiden15–70-vuotiaiden ulkomaillaoloajan mallinnus. Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Nainen	-0,65	0,23	-1,12	-0,79	-0,64	-0,48	-0,20	0,0047	1,00
Suomessaoloaika (vuosia)	0,160	0,026	0,110	0,142	0,160	0,178	0,215	0,0006	1,00
Maahanmuuttoikä	-0,015	0,013	-0,041	-0,023	-0,014	-0,006	0,009	0,0003	1,00
Maahanmuuttovuosi 1988–1993	-1,22	0,33	-1,86	-1,44	-1,23	-1,00	-0,56	0,0075	1,00
Maahanmuuttovuosi 1994–1999	-0,58	0,29	-1,14	-0,78	-0,58	-0,39	0,01	0,0067	1,00
Puoliso	-0,34	0,25	-0,84	-0,49	-0,33	-0,17	0,14	0,0060	1,00
$\lambda_1[0; 4{,}5)$	0,0119	0,0053	0,0048	0,0082	0,0109	0,0145	0,0250	0,00025	1,01
$\lambda_2[4{,}5; 10{,}6)$	0,0052	0,0028	0,0017	0,0033	0,0046	0,0064	0,0125	0,00011	1,00
$\lambda_3[10{,}6; 15{,}6)$	0,0051	0,0061	0,0001	0,0012	0,0032	0,0067	0,0220	0,00017	1,01
η	1,71	0,95	0,58	1,05	1,50	2,13	4,03	0,0387	1,01
ξ	-0,25	0,80	-1,82	-0,79	-0,26	0,30	1,30	0,0346	1,01
ρ	0,82	0,04	0,75	0,80	0,83	0,85	0,89	0,0009	1,00

Regressiokertoimilla ei ole taulukossa 3.12 standardin paranemisosuusmallin tai BCH-mallin kaltaista selkeää tulkintaa. Tämä johtuu siitä, että ξ-parametrin posteriorijakauma ei ole keskittynyt ”pieniin” arvoihin – jolloin regressiokertoimilla olisi BCH-mallin suhteellisten hasardien tulkinta – eikä myöskään nollan ”läheisyyteen” – jolloin puolestaan regressiokertoimilla olisi standardin paranemisosuusmallin (log-)vedonlyöntisuhteiden tulkinta.

Regressiokertoimet kuitenkin kertovat, mihin suuntaan muutokset selittävät muuttujissa vaikuttavat ulkomaillaolon pysymisen todennäköisyyteen. Kun regressiokertoimet saavat positiivisia arvoja, tämä tarkoittaa sitä, että positiivinen muutos niitä vastaavissa selittävissä muuttujissa *lisää* Suomeen paluumuuton todennäköisyyttä. Samaten, mikäli regressiokertoimet saavat negatiivisia arvoja, tämä tarkoittaa sitä, että positiivinen muutos niitä vastaavissa selittävissä muuttujissa *vähentää* Suomeen paluumuuton todennäköisyyttä.

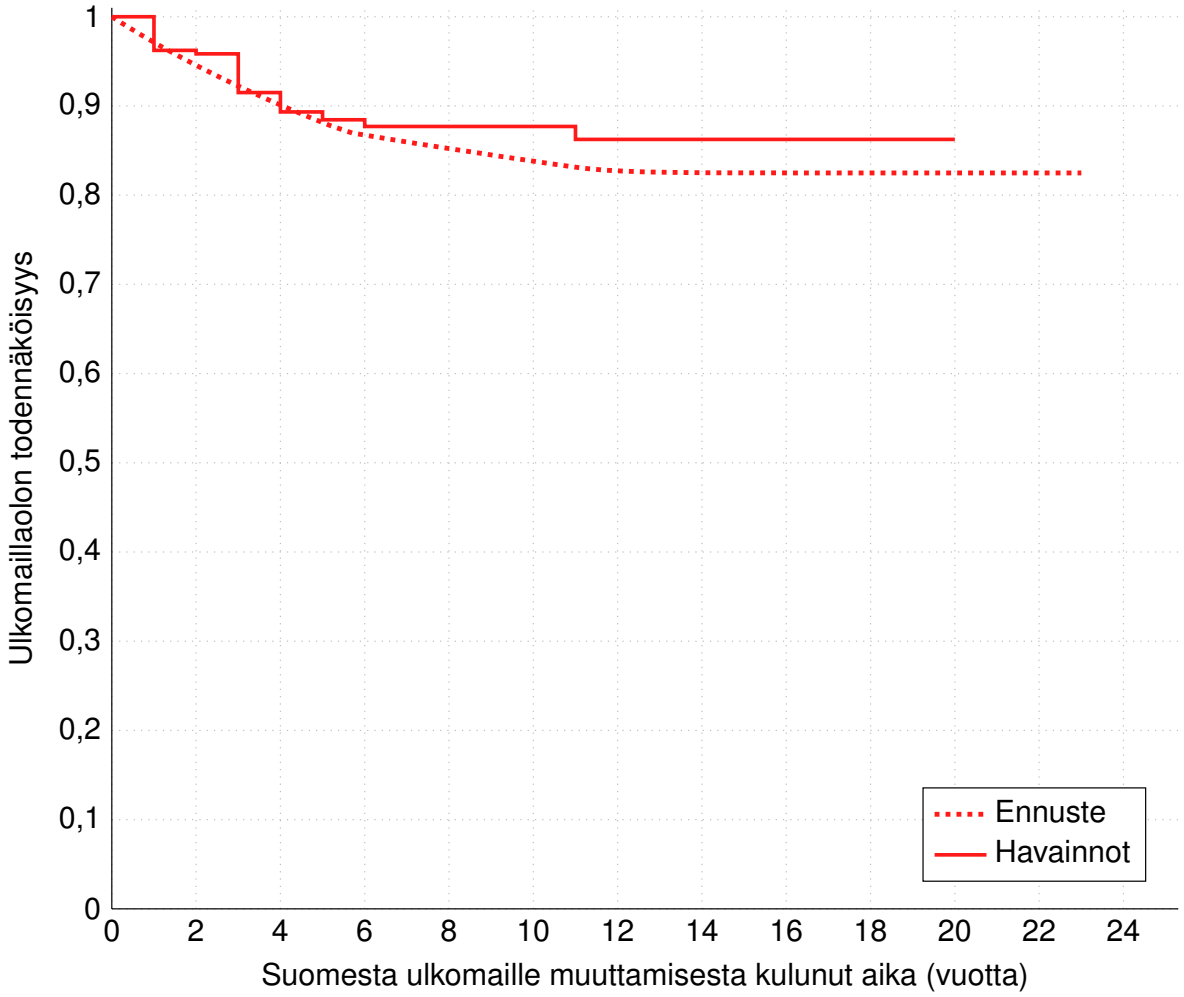
Taulukon 3.12 tulokset viittaavat siihen, että mikäli henkilö on nainen, hän muuttaa vähemmän todennäköisemmin takaisin Suomeen kuin siinä tapauksessa, että hän on mies. Lisäksi vaikuttaa siltä, että mitä nuorempana henkilö on muuttanut alun perin Suomeen, sitä vähemmän todennäköisempää on, että hän paluumuuttaa takaisin Suomeen. Myös alkuperäisen Suomeen muuton vuoden luokka näyttää vaikuttavan selkeästi paluumuuton todennäköisyyteen: aiempina vuosina Suomeen muuttaneet muuttavat vähemmän todennäköisesti takaisin Suomeen kuin 2000-luvulla Suomeen muuttaneet. Tässä yhteydessä voi muistaa, että aiemmin Suomeen muuttaneet myös muuttavat pois Suomesta todennäköisemmin. Kenties populaatiotasolla 1990-luvun alussa Suomeen muuttaneilla Somaliassa syntyneillä on heikommat siteet Suomeen kuin myöhemmin Suomeen muuttaneilla, mikä voisi selittää tämän.

Suomessa vietetyllä ajalla on odotettu vaikutus paluumuuton todennäköisyyteen: Mitä pidempään henkilö on ehtinyt yhteensä asua Suomessa, sitä todennäköisempää on, että hän paluumuuttaa takaisin Suomeen ulkomailta.

Taulukon 3.12 mukaanρ:n eli ”paranemisosuuden” jakauma sijoittuu n. 95 %:sti suunnilleen 0,75:n ja 0,90:n välille. Odotusarvo on 0,82. Tämä tarkoittaa sitä, että havaintoaineistossa Somaliassa syntyneiden (selittävillä muuttujilla ehdollistamaton) todennäköisyyden odotusarvo jäädä ulkomaille pysyvästi – eli ilman paluumuuttoa Suomeen – on noin 82 %:ia.

Posterioriennustetarkistus

Somaliassa syntyneiden ulkomaillaoloajan posterioriennustetarkistuksissa vertaan havaintoaineiston selviytymisfunktiota posterioriennusteiden selviytymisfunktion odotusarvoon. Posterioriennustetarkistuksen tulokset on esitetty kuvassa 3.11. Kiinteät, portaittaiset viivat esittävät selviytymisfunktion – eli ulkomaillaolon todennäköisyyttä ulkomaillaoloajan mukaan. ”Hyppäykset” tai portaat havaintoaineiston kuvaajissa vastaavat henkilöiden paluumuuttoja Suomeen tietyn ulkomaillaoloajan jälkeen.



Kuva 3.11: *Posterioriennustetarkistus. Somaliassa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle.*

Jatkuvat, katkonaiset viivat esittävät kuviossa puolestaan estimoidusta ulkomaillaoloajan mallista tuotettujen posterioriennusteiden selviytymisfunktioiden odotusarvoa. Kuten kuviosta 3.11 ilmenee, posterioriennusteiden selviytymisfunktioiden odotusarvot vastaavat kohtalaisen hyvin havaintoaineiston selviytymisfunktioita.

Somalian, kuten esimerkiksi Irakin kohdalla, posterioriennusteiden selviytymisfunktio vaikuttaa olevan hieman alempana kuin havaintoaineiston selviytymisfunktio. Oletan, että tämä johtuu siitä, että niillä, joilla on paljon ulkomaillaolovuosia aineistossa, on vähän suomessaoloajan vuosia. Lisäksi posterioriennusteet on laadittu kaikille henkilöille koko ulkomaillaoloajan yli, eli myös niille henkilöille, joilla on vähän ulkomaillaoloajan vuosia, mutta mahdollisesti paljon suomessaoloajan vuosia. Olen laatinut tästä pitemmän selvennyksen Irakissa syntyneiden ulkomaillaoloajan posterioriennustetarkistusten yhteydessä.

3.3.2.4 Posterioriennusteet: esimerkkejä muista syntymävaltioista

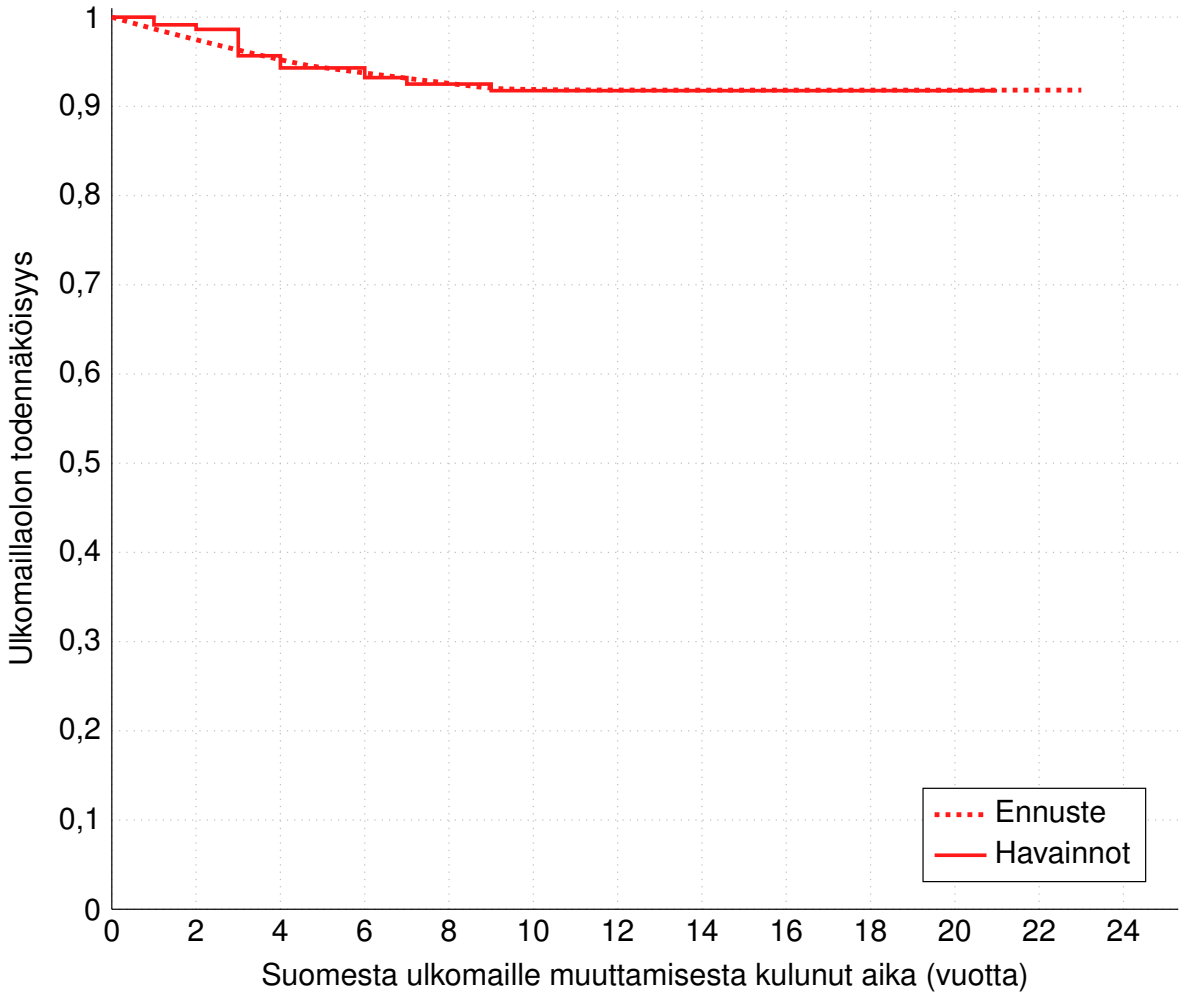
En julkaise tutkimuksen II osan tässä ensimmäisessä versiossa elinkaarivaikutusten tuloksia kuin Suomessa, Irakissa ja Somaliassa syntyneitä koskien. Julkaisen tässä kuitenkin muiden suurimpien maahanmuuttajaryhmien osalta joitakin posterioriennustetarkistusten tuloksia kuvina. Kuvailen niissä olevia tuloksia hyvin lyhyesti. Niistä käy myös yleisesti ottaen ilmi, että posterioriennusteet vastaavat hyvin havaintoaineistoja.

Kuvassa 3.12 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston selviytymisfunktiot Kiinassa syntyneille. Kiinassa syntyneet ovat tässä esimerkkinä maahanmuuttajaryhmästä, jonka osalta vain alle joka kymmenes paluumuuttaa takaisin Suomeen sen jälkeen, kun on muuttanut aluksi pois Suomesta.

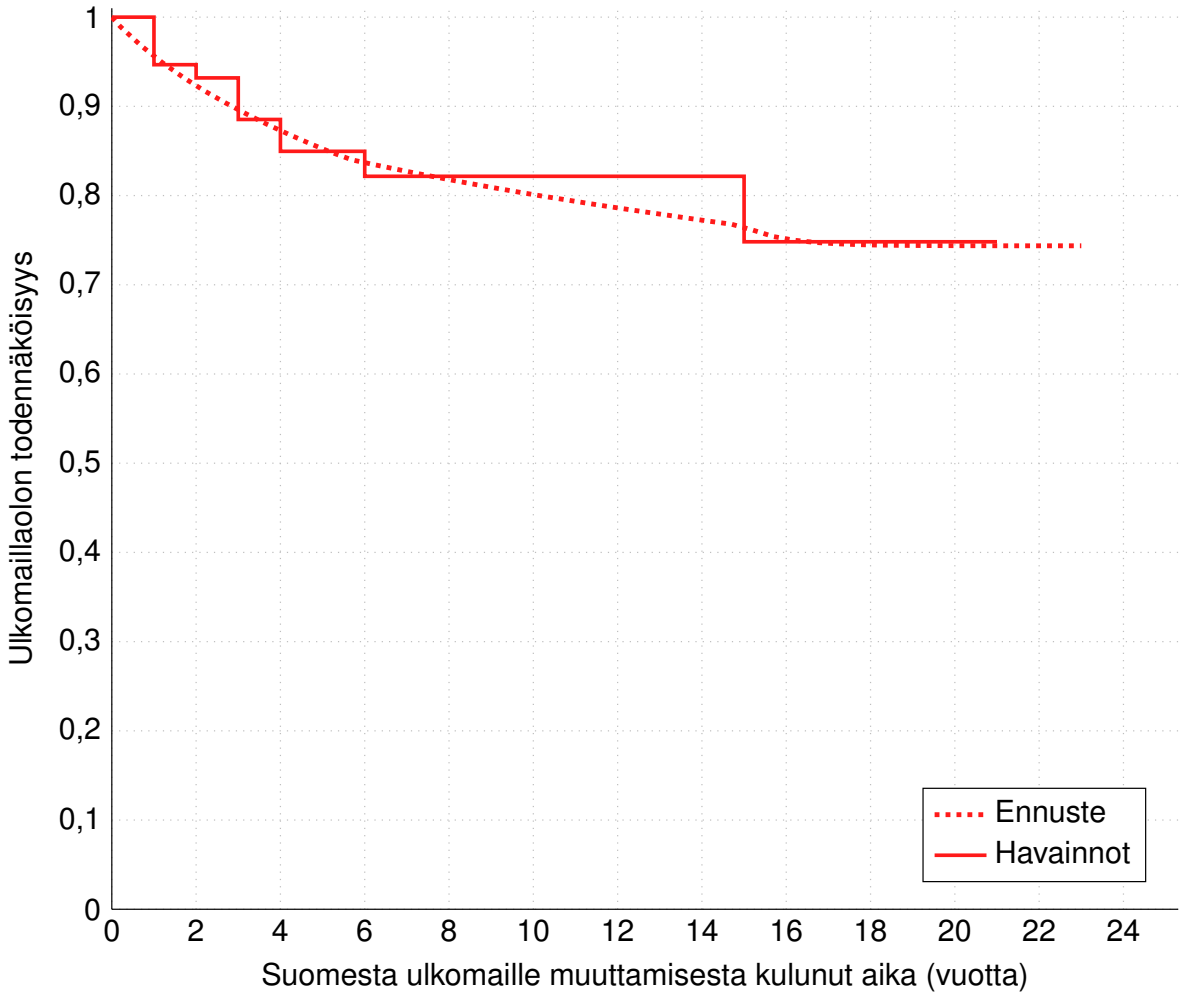
Kuvassa 3.13 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston selviytymisfunktiot Thaimaassa syntyneille. Thaimaassa syntyneistä ja Suomesta poismuuttaneista noin joka neljäs vaikuttaa paluumuuttavan takaisin Suomeen.

Kuvassa 3.14 on esitetty posterioriennusteiden selviytymisfunktioiden odotusarvot ja havaintoaineiston selviytymisfunktiot Turkissa syntyneille alkuperäisen Suomeen muuton maahanmuuttovuoden luokan mukaan.

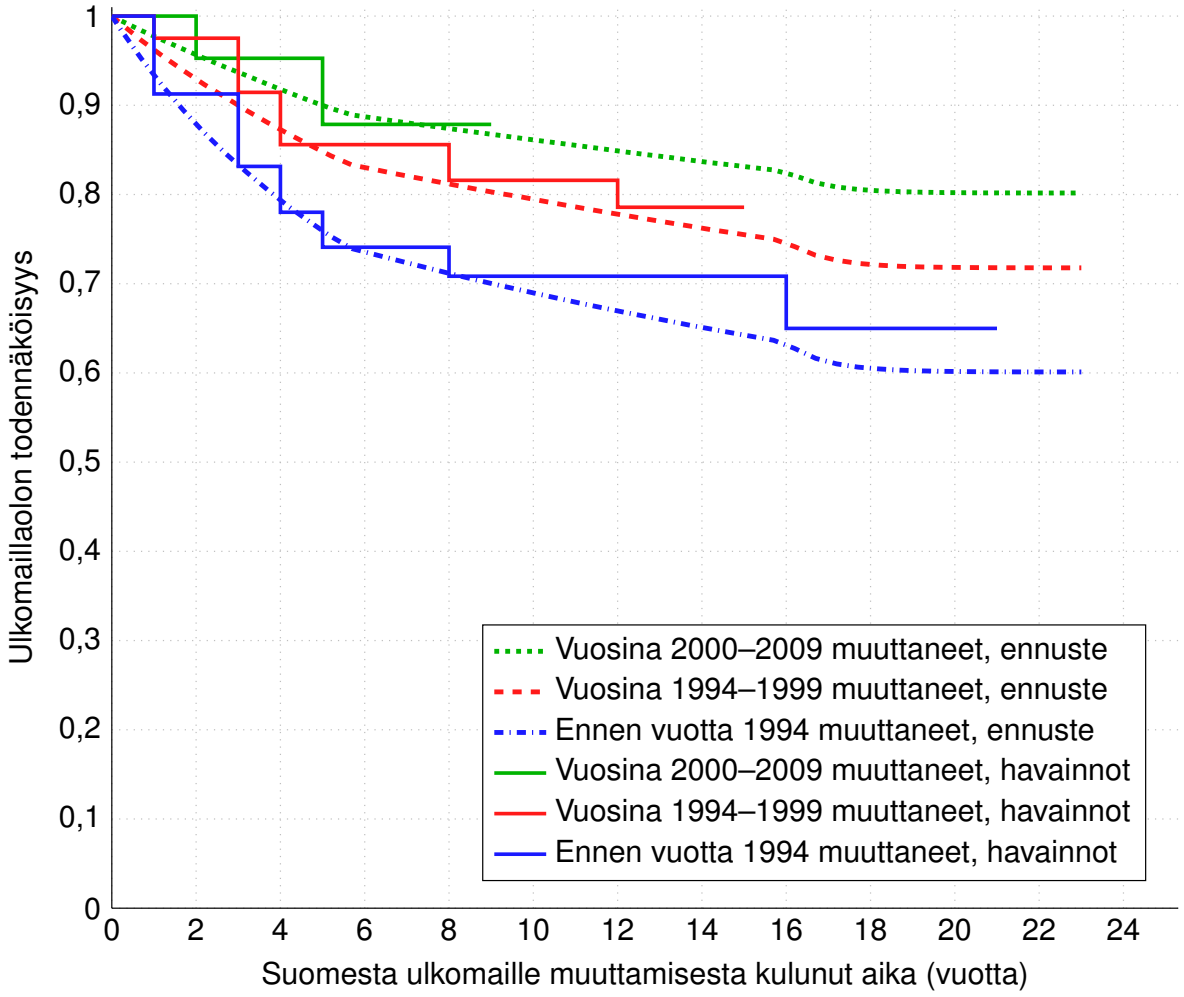
Turkissa syntyneiden osalta vaikuttaa siltä, että ne jotka ovat muuttaneet aiempina vuosina Suomeen, paluumuuttavat todennäköisemmin Suomeen kuin ne, jotka ovat muuttaneet Suomeen vasta 2000-luvulla. Ennen vuotta 1994 Suomeen muuttaneista on paluumuuttanut Suomeen, ulkomaille muuton jälkeen, noin 30–40 %:ia. Ennusteen mukaan puolestaan niistä, jotka ovat muuttaneet Suomeen 2000–2009, paluumuuttaa Suomeen, ulkomaille muuton jälkeen, noin 20 %:ia.



Kuva 3.12: *Posterioriennustetarkistus. Kiinassa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle.*



Kuva 3.13: *Posterioriennustetarkistus. Thaimaassa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posterioriennusteille ja havaintoaineistolle.*



Kuva 3.14: *Posteriorienustetarkistus. Turkissa syntyneet 15–70-vuotiaat. Ulkomaillaoloajan selviytymisfunktioiden Kaplan-Meier-estimaatit posteriorienusteille ja havaintoaineistolle maahanmuuttovuoden luokan mukaan.*

3.4 Kuolemien mallinnus

Kuollut on olotiloissa nk. absorboiva tila (ks. kuvio 3.1, s. 58), eli jos henkilö on kuollut tiettynä vuonna, hän on kuollut todennäköisyydellä 1 myös kaikkina tulevina vuosina.

Ikävuosittaiset kuolinennusteet ovat yksinkertaisia: mikäli henkilö ei ole vielä kuollut edeltävinä vuosina, niin hänen kuolintodennäköisyytensä tiettynä vuonna riippuu vain henkilön sukupuolesta sekä hänen iästään. Sukupuoli- ja ikävuosikohtaiset kuolintodennäköisyydet eli kuolleisuusluvut ovat Tilastokeskuksen kuolleisuusennusteista Helsinkiä koskien vuodelle 2015 (SVT: *väestöennuste*).

Ikävuodelle 100 lukua ei ennusteista löydy, joten käytin sitä koskien sukupuolittain kuolleisuuslukuja, jotka on saatu pyöristämällä 99:nnen ikävuoden sukupuolittaiset kuolleisuusluvut ylöspäin lähimpään satalukuun.

Vaikka kuolleisuuslukujen odotetaan laskevan tulevaisuudessa, käytin silti vuoden 2015 lukuja, sillä tutkimuksessa tehtävät elinkaariennusteet eivät ota huomioon eliniän pitenemiseen liittyviä vaikutuksia kuten esimerkiksi elinaikakerrointa karttuneiden eläkkeiden arvon määrittämisessä. Yksinkertaisuuden vuoksi oletan, että kuolleisuusluvut ovat samat niin ulkomailla olevilla henkilöillä kuin Suomessa olevilla henkilöillä.

Luku 4

Työikäisten ja nuorten nettovaikutusmallinnus

4.1 Mallinnustavoitteet

Työikäisten (20–62-vuotiaat) ja nuorten (15–19-vuotiaat) julkisen talouden nettovaikutusten mallintamisen tavoitteena on laatia maahanmuuttajaryhmäkohtaiset mallit, joita voidaan käyttää julkisen talouden elinkaarivaikutusten maahanmuuttajaryhmäkohtaisissa ennusteissa. Esitän seuraavaksi malleille kriteerit, jotka niiden tulee täyttää, jotta ne voivat toimia uskottavina malleina ennusteille.

Elinkaarivaikutusten ennusteet laadin siten, että ulkomailla syntyneiden osalta ennusteet perustuvat selittävien muuttujien osalta henkilöiden ensimmäisen Suomeen muuton vuoden tietoihin. Ulkomailla syntyneiden osalta työikäisten ja nuorten nettovaikutusmallit perustuvat selittävien muuttujien osalta siten sille, mitä heistä tiedetään, kun he muuttavat ensimmäisen kerran Suomeen. Suomessa syntyneiden osalta selittävät muuttujat perustuvat siihen, mitä henkilöistä tiedetään, kun he syntyvät.

Käytän jatkossa työikäisten ja nuorten nettovaikutusmallista joko nimitystä työikäisten nettovaikutusmalli tai pelkästään nettovaikutusmalli. Käsittelen mallinnuksessa käytettävien havaintoaineistojen esittelyn yhteydessä sitä, miten nuorten (15–19-vuotiaat) mallinnus eroaa maahanmuuttajaryhmittäin 20–62-vuotiaiden nettovaikutusmallinnuksesta.

4.1.1 Alkuperäisen havaintoaineiston toistaminen

Maahanmuuttajaryhmäkohtaisten nettovaikutusmallien on toistettava alkuperäisten aineistojen nettovaikutusten jakaumat. Tämä tarkoittaa sitä, että mallien on oltava sellaisia, että aineistot olisivat voineet olla generoituja näistä malleista. Arvioin tämän kriteerin täyttymistä posterioriennustetarkistusten yhteydessä.

Ensisijaisen tärkeää on, että maahanmuuttajaryhmäkohtaisten nettovaikutusten (ehdollistamattomien) ha-

vaintojen otoskeskiarvot ovat ”lähellä” mallien tuottamien posterioriennusteiden keskiarvoja. Läheisyyttä käsittelen tarkemmin posterioriennustetarkistusten yhteydessä. Tällä tavoin nettovaikutusten ennusteiden keskiarvoista tulee luotettavia.

Toiseksi tärkeintä on, että nettovaikutusten aineistojen jakaumien kvartiilit (alakvartiilit, mediaani, yläkvartiili) ovat ”lähellä” mallien tuottamien posterioriennusteiden samoja kvartiileja. Tällä tavoin nettovaikutusten ennusteiden tuottamista ennusteiden jakaumista tulee luotettavia muutenkin kuin vain keskiarvojen suhteen, erityisesti jakauman ”keskiosan” suhteen.

Mikäli myös mallien posterioriennustejakaumien alakvartiileja pienemmät prosenttipisteet sekä yläkvartiileja suuremmat prosenttipisteet ovat lähellä vastaavia nettovaikutusten aineistojen jakaumien pisteitä, ennusteet ovat luotettavia myös jakaumien häntien osalta. Ihanteellisessa tapauksessa mallien tuottamien nettovaikutusten posterioriennustejakaumien prosenttipisteet vastaavat aineistojen vastaavia 1 %:sta 99 %:iin. Tämän välin ulkopuolella sijaitsevat jakaumien ääriarvot jäävät tämän tutkimuksen kiinnostuksen kohteen ulkopuolelle jo siitä syystä, että niitä koskevien havaintojen julkaiseminen vaarantaisi yksityisyydensuojan.

Jakaumapistevertailujen lisäksi vertaan havaintojen jakaumien keskihajontoja posterioriennusteiden jakaumien keskihajontoihin. Tässäkin yhteydessä tavoitteena on näiden lukujen vastaavuus.

4.1.2 Otanta-asetelman havaintoaineistoon tuottamien riippuvuuksien huomiointi

Nettovaikutusmalleissa on otettava huomioon otanta-asetelman tuottamat riippuvuudet aineistossa. Kuten jäljempänä esitän, nettovaikutusaineistot ovat asetelmaltaan nk. paneeli- tai pitkittäisaineistoja. Tämä on käytännössä esitetty jo tutkimuksen I osassa, ja tarkoittaa sitä, että havaintoaineisto koostuu ajassa seurattavilta samoilta henkilöiltä poimituista vuosittaisista havainnoista.

Tämän seurauksena aineistossa on havaintojen välistä vaihtelua niin henkilöiden *välillä* (*engl. between*) kuin henkilöiden ”*sisällä*” (*engl. within*) eli tietyltä henkilöltä eri vuosina saatujen havaintojen välillä. Jälkimmäinen vaihtelu koostuu ajassa *perättäisten havaintojen välisestä riippuvuudesta* (*engl. serial dependence*) sekä muusta vaihtelusta, joka jää jäljelle.

Nettovaikutusmalleissa on otettava huomioon yllä mainitut kolme komponenttia: vaihtelu henkilöiden välillä, yhden henkilön ajassa toisiaan lähellä olevien havaintojen samankaltaisuus sekä jäljelle jäävä vaihtelu. Lisäksi henkilöiden voidaan katsoa olevan ryvästyneitä maahanmuuttovuoden mukaan, mikä tuo havaintoaineistoon yhden hierarkkisen tason lisää.

Aineistoon eivät sisälly kaikki nettovaikutuksia selittävät henkilöihin, maahanmuuttovuosiin ja maassaolovuosiin liittyvät tekijät. Näiden vaikutusten puuttumisen aiheuttama *havaitsematon heterogeenisuus* on myös huomioitava mallinnuksessa.

4.1.3 Tärkeimpien aikaan liittyvien vaikutusten huomiointi

Nettovaikutusmalleissa on oltava mukana selittävinä muuttujina aikaan liittyvät tärkeimmät muuttujat. Tämä johtuu siitä, että käytännössä kaikissa alan tutkimuksissa, joihin viittaan (ks. tutkimuksen I osa: *Tutkimuskirjallisuus*) nettovaikutusten kehitystä tarkastellaan ainakin iän mukaan. Ennusteiden luotettavuuden kannalta on myös tärkeää, että niissä on otettu huomioon iän lisäksi erityisesti maassaoloajan vaikutus.

Iän ja maassaoloajan lisäksi muita aikaan liittyviä selittäviä muuttujia ovat maahanmuuttovuosi sekä kalenterivuosi. Maassaoloaika, maahanmuuttovuosi ja kalenterivuosi eivät voi olla kuitenkaan samaan aikaan yhdessä mallissa, koska jokaisen niistä saa muodostettua kahdesta muusta muuttujasta lineaarisella muunnoksella, esimerkiksi: $kalenterivuosi = maahanmuuttovuosi + maassaoloaika$.

Aikaan liittyvien selittävien muuttujien päävaikutusten lisäksi on syytä tarkastella ainakin joitakin niiden tärkeimpiä interaktioita muiden selittävien muuttujien kanssa. Käsittelen mallien selittävien muuttujien valintaa tarkemmin jäljempänä.

Kuten mainitsin yllä (2. mallinnustavoite), aineistoon sisältyvä havaitsematon heterogeenisuus on huomioitava malleissa. Käytännössä tämä tarkoittaa havaitsemattoman heterogeenisuuden mallintamista erilaisten satunnaisvaikutusten avulla. Tämä mahdollistaa ennusteiden tekemisestä uskottavia aikamuuttujien osalta siten, että niihin sisältyy epävarmuus, joka johtuu ”uusista” maahanmuuttovuosista ja maassaoloajoista. Esimerkiksi vuosi 2015 on uusi maahanmuuttovuosi, sillä sitä ei ole mallinnusaineistossa, joka ulottuu vuoteen 2011.

4.2 Havaintoaineisto

4.2.1 Havaintoaineisto on kuvattu kattavasti toteutuneita nettovaikutuksia käsittelevässä tutkimusprojektin I osassa

Nettovaikutusmalleissa käytettävän aineiston muodostaminen on esitetty kattavasti tämän tutkimuksen I osassa *Toteutuneet julkisen talouden tulot ja menot*. Tutkimuksen ensimmäistä osaa voi pitää tässä yhteydessä aineistonkuvauksena työikäisten nettovaikutusten mallinnusta varten.

Esitän tässä lyhyesti, mistä nettovaikutusaineisto koostuu ja mitä eroja mallinnuksessa käytettävässä aineistossa on suhteessa tutkimuksen I osaan. Kehotan lukijaa tutustumaan tutkimuksen I osaan yksityiskohtaisempaa kuvausta varten.

Julkisen talouden nettovaikutusten määritelmä on esitetty tutkimuksen I osassa. Nettovaikutus on henkilötasolla henkilön julkisen talouden tulojen ja menojen välinen erotus. Julkisen talouden tulot koostuvat lähinnä henkilön maksamista välittömistä ja välillisistä veroista sekä veronluonteisista maksuista. Julkisen talouden menot puolestaan koostuvat lähinnä henkilölle annetuista tulonsiirroista ja palveluista (ml. koulutus). Maahanmuuttajaryhmäkohtaiset erot toteutuneissa julkisen talouden nettovaikutuksissa

syntyvät suurimmaksi osaksi eroista nettotulonsiirroissa, jotka ovat henkilöiden maksamien verojen ja veronluonteisten maksujen sekä saatujen tulonsiirtojen välisiä erotuksia. (Salminen 2015a, 9).

Julkisen talouden nettovaikutusten tietojen lähteinä ovat lähestulkoon pelkästään Tilastokeskuksen, THL:n ja Maahanmuuttoviraston henkilövuositasoiset rekisterit (Salminen 2015a, 155). Tutkimukseen sisältyy julkisen talouden nettovaikutuksia niin kuin niitä löytyy näistä rekistereistä henkilövuositasoisina tietoina. Tutkimuksen I osan liitteessä 2 on listattu ne suorat tulonsiirrot, jotka sisältyvät tutkimukseen (mt, 165–167). Suorista tulonsiirroista tutkimuksen ulkopuolelle jäävät esimerkiksi henkilöiden maksamat kiinteistöverot, sillä niitä ei ollut saatavilla koko tutkimuksen kattamalle ajanjaksolle 1995–2011. Lisäksi maahanmuuttajien erityispalveluista (esimerkiksi tulkkauksesta) koituvat menot puuttuvat suurimmaksi osaksi tutkimuksesta, sillä niistä ei ole olemassa valtakunnallisia rekisteritietoja. Käsittelen luvussa 10 puuttuvien tietojen vaikutusta tehtäviin julkisen talouden elinkaarivaikutusten ennusteisiin.

Otanta-asetelmasta. Otanta-asetelmaltaan nettovaikutusaineisto vastaa ryväsotannalla saatua paneeli- tai pitkittäisaineistoa. Tämä tarkoittaa sitä, että aineisto koostuu samoilta henkilöiltä poimituista vuosittaisista havainnoista. Ulkomailla syntyneiden osalta aineisto sisältää vuosittaiset nettovaikutustiedot ajanjaksolla 1995–2011 kaikkien niiden henkilöiden osalta, joilla on ollut kotikunta Suomessa vähintään yhtenä vuotena mainittujen vuosien aikana ja jotka ovat olleet samaan aikaan 15–70-vuotiaita. Otanta muodostaa ulkomailla syntyneiden osalta erityistapauksen ryväsotannasta, sillä sekä ryppäiden (henkilöiden) että ryväskohtaisten havaintojen (vuosittaiset nettovaikutukset) sisältymistodennäköisyys aineiston muodostavaan otokseen on 100 %.

Suomessa syntyneiden osalta aineiston muodostava otos sisältää 50 000 Suomessa syntynyttä henkilöä, jotka ovat olleet 15–70-vuotiaita ajanjaksolla 1988–2011. Näistä nettovaikutusmallinnuksessa käytetään henkilöitä, joille kriteeri täyttyy ajanjaksolla 1995–2011. Vuoden 2011 perusteella (otokseen sisältyvien henkilöiden ja saman ikäisten Suomessa syntyneiden henkilöiden osamäärä) ryppäiden sisältymistodennäköisyys otokseen on noin 1 %. Ryväskohtaisten havaintojen (vuosittaiset nettovaikutukset) sisältymistodennäköisyydet tutkimusaineiston muodostavaan otokseen ovat 100 %.

4.2.2 Erot havaintoaineistossa verrattuna tutkimusprojektin I osassa esitettyihin julkisen talouden nettovaikutuksiin

Korjaus välillisten verojen laskentaan. Tutkimusprojektin vuonna 2015 julkaistussa I osassa laskettiin henkilöiden maksamat välilliset verot osana julkisen talouden nettovaikutuksia. Välillisten verojen laskennassa tapahtui kuitenkin virhe, joka on korjattu tätä elinkaarivaikutuksia käsittelevää tutkimusprojektin II osaa varten. Virhe johtui siitä, että bruttorahatulot laskettiin (kotitalouden) käytettävistä tuloista lisäämällä niihin henkilöiden *saamat* tulonsiirrot, vaikka todellisuudessa käytettäviin tuloihin olisi pitänyt lisätä henkilöiden *maksamat* tulonsiirrot.

Jos tarkastellaan tämän korjauksen vaikutusta esimerkiksi vuoden 2011 toteutuneisiin nettovaikutuksiin, niin vaikutus kaikkien maahanmuuttajien keskiarvoon on hyvin pieni: ennen korjausta keskiarvo oli -3 620 euroa, ja korjauksen jälkeen se on -3 676 euroa. Eniten tämä korjaus vaikuttaa kuitenkin Somaliaa ja Irakia koskeviin keskiarvoihin, koska näissä maissa syntyneet saavat keskimäärin eniten tulonsiirtoja

suurimmista maahanmuuttajaryhmistä. Irakin osalta vuoden keskimääräinen 20–62-vuotiaiden nettovaikutus muuttui -13 348 eurosta korjauksen jälkeen -14 357 euroon, ja Somalian osalta muutos oli -13 845 eurosta -14 875 euroon. Suomessa syntyneiden osalta keskiarvo muuttui 81 eurosta 366 euroon.

Vain Suomessa olleiden havainnot sisältyvät havaintoaineistoon. Tutkimuksen I osan toteutuneita julkisen talouden nettovaikutuksia käsittelevä aineisto ei vastaa aivan täysin tässä tutkimuksen II osassa nettovaikutusmallinnukseen käytettävää aineistoa. Nettovaikutusmallinnukseen käytettävät aineistot sisältävät otokseen sisältyviltä henkilöiltä vain niiden vuosien havainnot, kun henkilö on ollut Suomessa tutkimuksen maassaolomääritelmän mukaan. Tutkimuksessa käytettävä havaintoaineisto sisältää havaintoja, joissa henkilö ei tutkimuksessa käytetyn määritelmän (ks. tutkimuksen I osa, s. 11) mukaan ole ollut kyseisenä vuonna Suomessa.

Nämä havainnot sisältyvät tutkimuksen I osan aineistoon, mutta eivät tässä II osassa tehtäviin nettovaikutusmalleihin. Syitä näiden havaintojen poistamiseen on kaksi. Ensiksi haluan, että elinkaarivaikutusten ennusteet perustuvat malleihin, jotka puolestaan perustuvat aineistoihin, jotka koskevat todella Suomessa olleita. Suomessaolo mallinnetaan ja ennustetaan erikseen. Toiseksi, poistettavien havaintojen nettovaikutus on käytännössä nolla, joten tämä vähentää nollaan kertyvien havaintojen määrää mallinnuksessa ja siten helpottaa mallinnusta, joka perustuu jatkuviin todennäköisyysjakaumiin.

15–19 nettovaikutukset mukaan havaintoaineistoon. Tutkimuksen I osassa käsiteltiin vain työikäisten 20–62-vuotiaiden toteutuneita julkisen talouden nettovaikutuksia. Tässä luvussa Irakin ja Somalian osalta myös 15–19-vuotiaiden nettovaikutukset sisällytetään mukaan samaan havaintoaineistoon 20–62-vuotiaiden kanssa. Suomessa syntyneiden 15–19-vuotiaiden nettovaikutukset mallinnetaan erikseen.

Maksetut eläkkeet poistettu nuorten ja työikäisten havaintoaineistosta, maksettavat vanhuuseläkkeet ennustetaan erikseen. Toisin kuin tutkimuksen I osassa, maksetut eläkkeet eivät sisälly nettovaikutusaineistoon tutkimuksen II osassa. Vanhuuseläkkeitä varten palkkasummasta karttuneet maksut mallinnetaan erikseen, ja maksetut eläkkeet ennustetaan erikseen. Mikäli maksetut eläkkeet olisi samalla tavalla poistettu myös tutkimuksen I osassa, tämä olisi lähinnä parantanut Suomessa syntyneiden työikäisten keskimääräisiä julkisen talouden nettovaikutuksia suhteessa eri maahanmuuttajaryhmiin, sillä Suomessa syntyneiden ikärakenteen johdosta, heidän joukossaan oli suhteellisesti enemmän eläkkeensaajia kuin muissa tutkittavissa ryhmissä.

Vastaanotto- ja kotouttamismenot on poistettu nuorten ja työikäisten nettovaikutuksia käsittelevästä havaintoaineistosta. Irakissa ja Somaliassa syntyneisiin liittyvät vastaanotto- ja kotouttamismenot on poistettu havaintoaineistosta. Syynä tähän on se, että tutkimuksen I osassa (luku 6.7) arvioidut (muutenkin puutteelliset) vastaanotto- ja kotouttamismenot löytyvät täysimääräisinä vain viimeisille aineiston kalenterivuosille. Tämä puolestaan vääristäisi esimerkiksi maahanmuuttovuoden estimoitua vaikutusta työikäisten nettovaikutuksiin. Elinkaariennusteissa vastaanotto- ja kotouttamismenojen oletetaan olevan yhteensä 50 000 euroa per henkilö ja sen jakaantuvan tasaisesti viidelle ensimmäiselle maassaolovuodelle (ks. tarkemmin luvut 9.4.2 ja 10.2.1).

Kollektiiviset menot poistettu nuorten ja työikäisten havaintoaineistosta. Toisin kuin tutkimuksen I osassa, menot nk. kollektiivisesti kulutetuista julkisen talouden palveluista (ks. tutkimuksen I osa, s. 10 ja 128) eivät sisälly nettovaikutusaineistoon tutkimuksen II osassa. Näiden palveluiden kuluttajia ei voida yksilöidä tarkalleen, ja niihin sisältävät esimerkiksi infrastruktuurimenot.

Syy siihen, että kollektiivisesti kulutettujen palveluiden menot eivät sisälly elinkaarivaikutusten ennusteisiin, on, että laadin elinkaariennusteet tässä tutkimuksen II osassa vain Suomessa *syntyville* että uusille maahanmuuttajille, eli ulkomailta Suomeen *juuri muuttamassa* oleville. Oletan näiden henkilöiden osalta kollektiivisten palveluiden rajakustannusten olevan nolla (euroa). Myös VATT (2014a) olettaa näin yhden Suomeen muuttavan tai syntyvän kohdalla (Liite VATT-analyysin lukuun 5 (VATT 2014b))).

Vaikka lukumäärältään "suuren"henkilöjoukon kohdalla näiden henkilöiden yhteenlaskettu kollektiivisten palveluiden rajakustannus olisikin nollaa suurempi – käsiteltäessä joko Suomessa syntyvien joukkoa tai jotain Suomeen juuri muuttamassa olevien joukkoa – oletan, että rajakustannus on näillä eri henkilöjoukoilla yhtä suuri. Kun tällöin tarkastellaan julkisten talouden nettovaikutusten elinkaariennusteiden summien erotuksia näiden joukkojen *välillä*, rajakustannusten välinen erotus on nolla, joten ne eivät vaikuta näiden joukkojen välisiin ennustetuloksiin.

Aineistoa koskevat ryhmäkohtaiset henkilömäärät, nettovaikutusten jakaumatiedot sekä mallien selittävien muuttujien jakaumatiedot esitän tulosten yhteydessä jäljempänä.

4.3 Nettovaikutusten mallinnus

Nettovaikutusten mallinnus tapahtuu bayesilaisessa viitekehyksessä. Perusteet tälle ovat käytännössä samat kuin olotilamallinnuksen kohdalla (ks. luku 3.2.2.1).

Bayesilaisessa viitekehyksessä tavoitteena on määrittää estimoitavien parametrien havaintoaineistolla ehdollinen yhteistodennäköisyysjakauma eli posteriorijakauma: $p(\boldsymbol{\theta}|\boldsymbol{y})$, missä $\boldsymbol{\theta}$ on estimoitavien parametrien vektori ja $\boldsymbol{y}$ on havaintoaineisto. Kun parametrivektori $\boldsymbol{\theta}$ koskee nuorten ja työikäisten nettovaikutusten mallinnusta, sen posteriorijakaumasta tehdään tässä tutkimuksen luvussa tarvittavat tilastolliset päättelyt, sekä sitä käytetään nuorten ja työikäisten nettovaikutusten ennustamiseen uusien maahanmuuttajien elinkaariennusteissa.

Olkoon nyt nuorten ja työikäisten nettovaikutusten yhteistodennäköisyysjakauma (tilastollinen malli), joka on ehdollinen havaintojen malliparametreilla: $p(\boldsymbol{y}|\boldsymbol{\theta})$. Kutsun tätä tilastollista mallia aineistomalliksi, sillä sen on tarkoitus kuvata se dataa generoiva prosessi, jonka toteutumana havaintoaineisto on syntynyt. Malliparametrien reunajakauma $p(\boldsymbol{\theta})$ on puolestaan nk. *priorijakauma*, sillä se summaa $p(\boldsymbol{\theta})$-vektoria koskevan havaintoaineiston ulkopuolisen informaation ”ennen” havaintojen tuomaa lisäinformaatiota. Bayesin kaavaa soveltamalla voidaan osoittaa, että (normeeraamaton) posteriorijakauma on verrannollinen aineistomallin ja priorijakauman tulolle: $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

Työikäisten nettovaikutusmallien parametrien posteriorijakauman selvittämiseksi on täten määritettävä aineistomalli sekä parametreja koskeva priorijakauma. Määritän aluksi nettovaikutuksille aineistomallin.

Tämä tapahtuu askeltaen siten, että käytännössä jokaisen maahanmuuttajaryhmän kohdalla estimoin useita perättäisiä aineistomalleja, joista valitsin sopivimman käytettäväksi uusien maahanmuuttajien nuorten ja työikäisten nettovaikutusennusteissa. Luvuissa 4.3.2 ja 4.3.3 perustelen ja määritän aineistomallit. Luvussa 4.3.4 määritän puolestaan aineistomallien estimoitaville parametreille niiden priorijakaumat. Määritän kaikille parametreille priorijakaumiksi *sopivat* (*engl. proper*) todennäköisyysjakaumat, mikä varmistaa sen, että posteriorijakauma on vakiokerrointa vaille sopiva todennäköisyysjakauma.

Ennen kuin määritän perättäiset kokeiltavat aineistomallit, esitän perustelut sille, miksi haluan aineistomallien olevan hierarkkisia tilastollisia malleja. Tämä pätee erityisesti sen aineistomallin kohdalla, jota sovellan uusien maahanmuuttajien elinkaarivaikutusten ennustamiseen.

4.3.1 Paneeliaineistot, havaitsematon heterogeenisuus ja hierarkkiset mallit

Perustelen tässä luvussa, miksi käytin satunnaisvaikutuksia sisältäviä hierarkkisia malleja nettovaikutusten aineistomalleina. Olennaista tässä on huomioida paneeliaineiston rakenne, edellä esitetyt mallinnustavoitteet sekä se, että käytän lopullisia malleja ennustamaan nettovaikutuksia *uudelle aineistolle* eli sellaisille henkilöille, joista olevia tietoja ei käytetty mallien estimoinnissa. Nämä henkilöt uusia maahanmuuttajia eli vasta Suomeen muuttavassa olevia.

Paneeliaineistoista ja niiden mallintamisesta satunnaisvaikutuksia sisältävien (hierarkkisten) mallien avulla löytyy huomattava määrä kirjallisuutta niin ”frekventistisen” kuin bayesilaisen päättelyn piirissä. Mainitsen kirjallisuudesta esimerkkeinä Digglen et al. (1994), Hsiaon (2003) sekä erityisesti bayesilaiselta puolelta Koopin (2003, luku 7) ja Chibin (1996) sekä tutkimuksessa yleisesti paljon käyttämäni Gelmanin et al. (2014). Seuraavaksi esitettävät paneeliaineiston ja satunnaisvaikutuksia sisältävien mallien valinnan perustelut perustuvat pitkälti näihin teoksiin.

4.3.1.1 Paneeliaineiston edut: ryhmä- ja aikavaikutusten erittely, havaitsemattoman heterogeenisuuden mallinnus

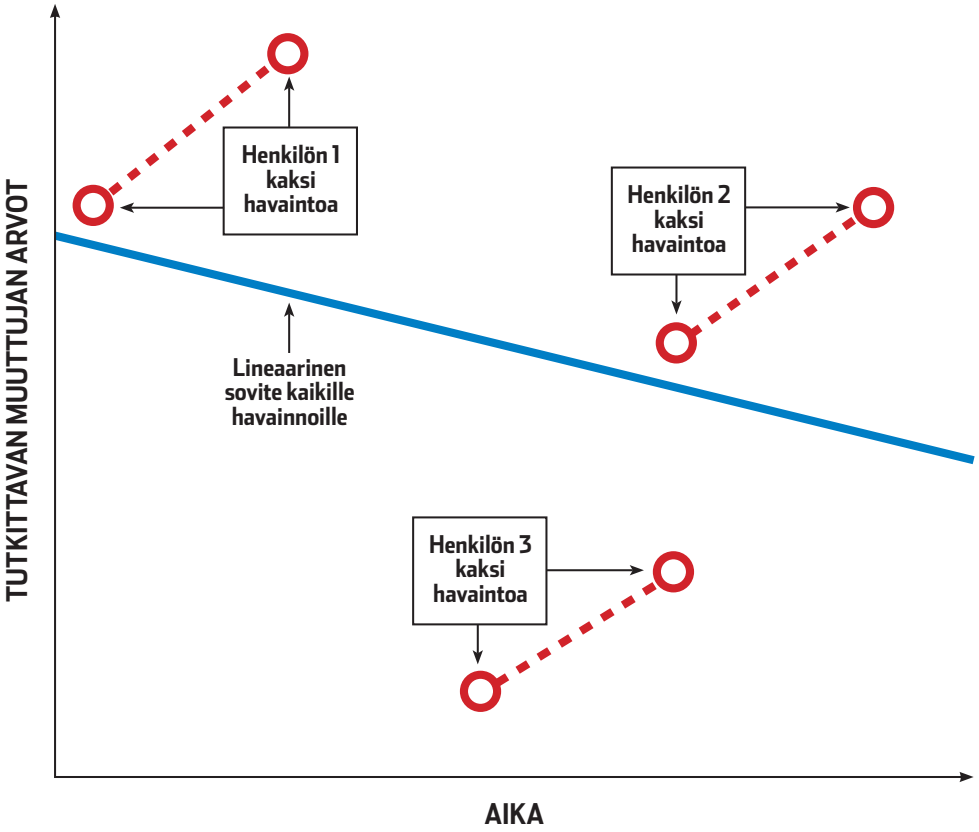
Paneeliaineisto tai pitkittäisaineisto tarjoaa mahdollisuuden mallintaa vaikutuksia, joita ei voisi selvittää joko pelkkiin poikkileikkausaineistoihin tai yksittäisiin (yksiulotteisiin) aikasarjoihin perustuvissa malleissa. Mikäli käytettävä aineisto on poikkileikkausaineisto, eli se ei sisällä samoista tutkittavista poikkileikkausyksiköistä (esimerkiksi henkilöistä) havaintoja eri ajankohtina, aineiston perusteella on käytännössä mahdotonta eritellä toisistaan niitä vaikutuksia, jotka johtuvat eroista poikkileikkausyksiköiden välillä, niistä vaikutuksista, jotka liittyvät muutoksiin ajassa (Hsiao 2003, 4; Diggle et al. 1994, 1–2 ja 17–18). Yksittäisistä aikasarjoista koostuvat aineistot eivät puolestaan sisällä tietoa yksikköjen (esimerkiksi henkilöiden) välisistä eroista, joka voi aiheuttaa muun muassa vaikeuksia estimoinnissa, vaikka yksiköiden väliset erot eivät olisi edes kiinnostuksen kohteena mallinnuksessa (Hsiao 2003, 5).

Ikä-, kohortti- ja periodivaikutusten selvittäminen ja eritteleminen on yksi esimerkki mallinnustilanteesta, johon tarvitaan paneeliaineistoa. Nettovaikutusmallien aineistossa kohortteja ovat eri kalenterivuosina Suomeen muuttaneet ja periodeita ovat puolestaan kalenterivuodet ja maassaolovuodet. Käytettävissä

oleva paneeliaineisto mahdollistaa siten iän, maahanmuuttovuoden, maassaoloajan ja kalenterivuoden vaikutusten erittelyn ja selvittämisen, mikä on eräs mallinnustavoitteista – joskaan aivan kaikki nämä vaikutukset eivät voi olla samaan aikaan yhdessä mallissa.

Käytännössä yksikään tilastollinen malli, jonka estimointiin käytettyä aineistoa ei ole tuotettu kontrolloidussa koeasetelmassa, ei sisällä kaikkia niitä tilastollisiin yksiköihin liittyviä tekijöitä, jotka vaikuttavat selitettävän muuttujan vaihteluun. Näitä tilastollisia yksiköitä ovat poikkileikkausyksiköt, jotka nettovaikutusmalleissa ovat henkilöitä, ja aikayksiköt, joita nettovaikutusmalleissa ovat maahanmuuttovuodet ja maassaoloajat. Näitä tilastoyksiköittäin vaihtelevia havaitsemattomia tekijöitä kutsutaan mallin *havaitsemattomaksi heterogeenisuudeksi* (*engl. unobserved heterogeneity*). Nettovaikutusmalleissa henkilöiden kohdalla esimerkiksi tarkka koulutustieto tai tarkka maahanmuuton syy voivat olla tällaisia havaitsemattomia tai latentteja tekijöitä, jotka vaikuttavat nettovaikutusten arvojen vaihteluun henkilöittäin. (Hsiao 2003, 8–9).

Alla kuvassa 4.1 on havainnollistettu, miksi paneeliaineistoja tarvitaan, jotta esimerkiksi henkilöiden väliset vaikutukset voitaisiin eritellä aikaan liittyvistä vaikutuksista.



Kuva 4.1: ***Havainnekuva paneeliaineistojen hyödyllisyydestä****, kun halutaan eritellä aikaan ja henkilöihin liittyviä vaikutuksia. Havainnekuvassa kolmelta eri henkilöltä on havaittu tutkittavan muuttujan arvot kahtena eri ajanhetkenä. Mikäli vaihtelua henkilöiden välillä ei huomioitaisi, kaikkien kuuden havainnon perusteella estimoidun yksinkertaisen lineaarisen sovitteen mukaan ajalla olisi negatiivinen vaikutus tutkittavan muuttujan arvoon (kuvan laskeva sininen suora). Jos taas paneeliaineiston mahdollistama henkilöiden välinen vaihtelu otetaan huomioon, kun tutkitaan ajan vaikutusta tutkittavaan muuttujaan, tuloksena on, että ajalla onkin positiivinen vaikutus tutkittavaan muuttujaan. Tämä johtuu siitä, että kaikilla henkilöillä tutkittavan muuttujan arvot nousevat ajassa (punaiset katkoviivat henkilöiden havaintojen välillä). Tätä jälkimmäistä analyysia, jossa henkilöiden välinen vaihtelu otetaan huomioon, ei voisi tehdä pelkällä poikkileikkausaineistolla, jossa jokaiselta henkilöltä on vain yksi havainto. Mikäli henkilöiden välinen vaihtelu ei "selity"malliin kuuluvilla (havaituilla) selittävillä muuttujilla, kyse on henkilöiden välisestä havaitsemattomasta heterogeenisuudesta. Havainnekuvassa henkilöiden väliset erot tutkittavan muuttujan suhteen voivat johtua esimerkiksi koulutuseroista, mutta jos analyysin tekijällä ei ole käytössä henkilöiden koulutustietoja, henkilöiden väliset erot jäävät havaitsemattomaksi heterogeenisuudeksi.*

4.3.1.2 Havaitsematon heterogeenisuus ja satunnaisvaikutukset

Havaitsemattoman heterogeenisuuden aiheuttama vaihtelu tilastoyksiköiden välillä voi olla suuruudeltaan kolmenlaista. Joko vaihtelua ei ole, eli vaihtelun suuruus on nolla, vaihtelu on ääretöntä, tai vaihtelu on jotain nollan ja äärettömän väliltä. Kaksi ensimmäistä tapausta näistä ovat erityistapauksia, ja ensimmäisessä tapauksessa havaitsematonta heterogeenisuutta ei ole ollenkaan. Kaikilla kolmella tapauksella on vaikutuksensa niiden tilastolliseen mallintamiseen.

Ei ryväsvaikutuksia. Mikäli havaitsematonta heterogeenisuutta ei ole lainkaan – eli mallinnusaineisto on täysin kontrolloidusta kokeesta, sillä ei ole merkitystä, minkä tilastoyksikön muodostamaan ”ryppääseen” kukin havainto kuuluu. Toisin sanoen ryväsvaikutuksia ei ole. Tällöin kaikki havainnot voidaan "sekoittaa"täysin mallinnuksessa (*engl. complete pooling*), ja mallin vakiokertoimen ja regressiokertoimien arvot eivät vaihtele tilastoyksiköittäin. Mikäli oletus havaitsemattoman heterogeenisuuden puuttumisesta on kuitenkin virheellinen, tämä johtaa esimerkiksi mallin regressiokertoimien mahdollisesti harhaisiin estimaatteihin sekä liian pieniin keskivirheisiin.

Kiinteät vaikutukset . Mikäli havaitsematon heterogeenisuus on ääretöntä (tai käytännössä ainakin hyvin suurta), tilastoyksiköiden välinen vaihtelu muodostaa mallin kaiken selittämättömän vaihtelun. Jokaiselle tilastoyksikölle mallinnetaan tällöin erikseen omat vakioterminsä ja regressiokertoimensa. Tässä *kiinteiden* (ryhmä/ryväs)*vaikutusten* (*engl. fixed effects*) mallissa havaintojen sekoittamista eri tilastoyksiköiden muodostamien havaintojen ryppäiden välillä ei ole ollenkaan (*engl. no pooling*). Mikäli tämä oletus on virheellinen, mallissa on liikaa parametreja, eikä siinä käytetä hyväksi tilastoyksiköiden välisen vaihtelun tuomaa tilastollista voimaa.

Satunnaisvaikutukset. Mikäli havaitsematon heterogeenisuus ei noudata kumpaakaan edellä mainituista ääri- tai erityistapauksista, kyse on *satunnaisvaikutuksista* (*engl. random effects*). Ääritapauksissa joko ryppäiden *välinen* vaihtelu (kiinteiden vaikutusten malli) tai ryppäiden *”’sisäinen”’* vaihtelu (complete pooling) dominoi mallin kokonaisvaihtelua. Satunnaisvaikutusten mallissa ryppäiden välisen ja sisäisen vaihtelun suhde on näiden ääripäiden välissä. Mikäli satunnaisvaikutuksia sisältävässä tilastollisessa mallissa vakiotermin annetaan vaihdella tilastoyksiköittäin, kyse on *satunnaisvakiosta* (*engl. random intercept*). Tilastoyksiköittäin vaihtelevien regressiokertoimien osalta kyse on puolestaan *satunnaisista kulmakertoimista/regressiokertoimista* (*engl. random slopes*).

Gelmania et al. (2014) lainaten huomautan, että bayesilaisessa yhteydessä puhuminen ”kiinteistä” ja ”satunnaisista” vaikutuksista on hieman harhaanjohtavaa, sillä bayesilaisissa malleissa kaikki estimoitavat parametrit ovat satunnaisia siinä mielessä, että niillä on oma todennäköisyysjakaumansa (s. 383). Nimitykset tulevatkin bayesilaisen päättelyn ulkopuolelta.

Usein yksi malli saattaa sisältää sekä satunnaisia että kiinteitä vaikutuksia. Tällaista mallia kutsutaan *sekavaikutusten malliksi* (*engl. mixed effects model*). Kuten edellä mainitsin, satunnaisvaikutuksia sisältävässä mallissa regressiokertoimien arvot voivat vaihdella joko poikkileikkausyksiköittäin (esimerkiksi

henkilöittäin) tai aikayksiköittäin (nettovaikutusmallissa maahanmuuttovuoden tai maassaoloajan mukaan).

Nettovaikutusmalleissa käytän saatavilla olevaa paneeliaineistoa yllä mainituista syistä johtuen. Mallit sisältävät myös satunnaisvaikutuksia, sillä ryppäiden välisen vaihtelun mallintaminen mahdollistaa ennusteiden laatimisen niille *uuden aineiston* yksiköille (esimerkiksi henkilöille eli uusille maahanmuuttajille), jotka eivät kuulu itse mallinnusaineistoon. Käytännössä tämä tapahtuu nk. *hierarkkisten mallien* (*engl. hierchical models*) avulla (Gelman et al. 2014, 381).

4.3.1.3 Hierarkkiset mallit ja uuden aineiston ennusteet

Hierarkkisia malleja voidaan soveltaa tilanteissa, joissa aineisto on rakentunut hierarkkisesti. Paneeliaineistot ovat tällaisia. Esimerkiksi nettovaikutusaineistossa yksittäiset havainnot nettovaikutuksista ovat ryvästyneet aineistossa henkilöiden alle. Henkilöt puolestaan ovat ryvästyneet aineistossa maahanmuuttovuosien mukaan.

Hierarkkiset mallit mahdollistavat satunnaisvaikutusten mallintamisen siten, että kukin ryväskohtainen parametri on toteuma ryvästen yhteisestä kyseisen parametrin *populaatiotason* jakaumasta. Nämä parametrit ovat edellä esiteltyjä satunnaisvaikutuksia.

Hierarkkiset mallit ovat hyödyllisiä, koska mallin kokonaisvaihtelu voidaan jakaa eri ”tasoille” noudattaen sen hierarkkista rakennetta. Tästä syystä hierarkkisia malleja kutsutaan myös monitasomalleiksi. Hierarkkisten mallien avulla voidaan näin tutkia, kuinka mallin vaihtelu jakaantuu näille eri tasoille.

Uuden aineiston ennusteet. Tämä on hyödyllinen ominaisuus, kun mallin pohjalta laaditaan *uudelle aineistolle* ennusteita. Uuden aineiston ennuste tarkoittaa, että ennusteet laaditaan sellaisille tilastoyksiköille (esimerkiksi henkilöille), joita ei ollut havaintoaineistossa, minkä pohjalta malli estimoitiin. Tällä tavoin eri vaihtelun tasoja koskeva havaitsematon heterogeenisuus voidaan ottaa mukaan ennusteisiin satunnaisvaikutusten kautta niin, ettei ennusteista tule epärealistisen tarkkoja. Esimerkiksi, kun ennusteita tehdään uusille maahanmuuttajille, ennusteisiin sisältyy siten myös se epävarmuus, joka aiheutuu siitä, ettei henkilöihin liittyvien havaitsemattomien tekijöiden arvoja ole voitu sisällyttää suoraan ennusteisiin, koska niitä ei tunneta. (Gelman et al. 2014, 381 ja 387).

Vaihdettavuus ryppäiden välillä. Keskeinen hierarkkisten mallien käsite ja hierarkkisilta malleilta edellytettävä vaatimus on *vaihdettavuuden* (*engl. exchangeabilty*) toteutuminen. Vaihdettavuuden toteutuminen tarkoittaa sitä, että hierarkkisen mallin jokaisella tasolla ryväsvaikutukset (esimerkiksi henkilövaikutukset) ovat ehdollisesti toisistaan riippumattomia, kun mallin selittävät muuttujat huomioidaan. Tällöin eri ryppäät (esimerkiksi henkilöt) identifioivat indeksit voidaan vaihtaa keskenään ilman, että tällä olisi vaikutusta niiden ehdolliseen yhteistodennäköisyysjakaumaan. Vaihdettavuuden toteutuminen tarkoittaa sitä, että ryppäitä koskevien parametrien priorijakaumina voidaan käyttää samaa todennäköisyysjakaumaa. (Gelman et al. 2014: s. 5, 104–107 ja 383).

Mallin kokonaisvaihtelun eri komponentit. Hierarkkisiin malleihin sisältyy ryppäiden *välisen* (*engl. between*) vaihtelun lisäksi ryppäiden *sisäinen* (*engl. within*) vaihtelu. Kuten olen jo edellä esittänyt, sisäinen vaihtelu puolestaan koostuu kahdesta osasta paneeliaineistossa: perättäisten havaintojen välisestä riippuvuudesta (*engl. serial correlation*) sekä jäljelle jäävästä havaintotason virhetermin vaihtelusta. Nämä kaikki vaihtelun komponentit on mallinnettava (ks. esim. Diggle et al. 1994).

4.3.2 Nettovaikutusten aineistomallit

4.3.2.1 Nettovaikutusaineiston hierarkkinen rakenne

Nettovaikutusaineiston rakenne on hierarkkinen, kuten olen edellä esittänyt. Käytän jäljempänä hierarkkisten mallien yhteydessä nettovaikutuksista seuraavaa merkintätapaa nettovaikutusten hierarkkisesta rakenteesta johtuen:

Y_{tim}.

Varsin yleiseen tapaan merkitsen nettovaikutusta eli mallin selitettävää muuttujaa Y-kirjaimella. Alaindeksit t,i ja m puolestaan viittaavat nettovaikutusaineistojen hierarkkiseen rakenteeseen. Jokainen niistä edustaa nettovaikutusaineistojen eri ”tasoja”. Näiden alaindeksien arvot ilmoittavat havaintojen sijainnin aineistojen hierarkkisessa rakenteessa. Alaindeksin kirjain t viittaa maassaolovuosien määrään ja se on se on siten käytännössä henkilöä i koskevan havainnon *järjestysnumero*, missä järjestysnumerot ovat aineistossa: $t = 0, 1, 2, \ldots, 22, 23$. Yksittäiset havainnot ovat ryvästyneet henkilöittäin, missä alaindeksi i on henkilön tunnus. Henkilöt ovat puolestaan ryvästyneet maahanmuuttovuosittain. Alaindeksi m viittaa maahanmuuttovuoteen.

Kaikkien selittävien muuttujien ”taso” voidaan ilmaista myös yllä mainittujen alaindeksien avulla. Henkilöittäin, muttei maahanmuuttovuoden tai maassaoloajan, mukaan vaihtelevien selittävien muuttujien alaindeksi on pelkkä i. Pelkästään maahanmuuttovuoden mukaan vaihtelevilla muuttujilla indeksi on m, ja vain maassaoloajan mukaan vaihtelevilla se on t. Ikä vaihtelee paitsi henkilöittäin myös maassaoloajan mukaan, joten sen alaindeksi on it.

Edellä esitin *vaihdettavuuden* käsitteen. Tavoitteena nettovaikutusmallinnuksessa on, että elinkaariennusteissa käytettävissä malleissa toteutuu vaihdettavuus. Tämä tarkoittaa sitä, että nettovaikutusten ehdolliset todennäköisyydet pysyvät samoina – kun selittävät tekijät huomioidaan – vaikka alaindeksien t,i,m saamat arvot vaihdettaisiin ennusteiden välillä.

4.3.2.2 Aineistomallien valinta noudattaen mallinnuksen tavoitteita

Maahanmuuttajaryhmäkohtaisten aineistomallien valinta perustui edellä esitettyihin mallinnustavoitteisiin sekä joidenkin selittävien muuttujien valinnan kohdalla myös erilaisten mallien kokeiluihin eri maahanmuuttajaryhmien kohdalla. Koska valitsin ennusteissa käytettäviksi nettovaikutusmalleiksi lähes sa-

manlaiset mallit jokaisen maahanmuuttajaryhmän kohdalla, käytän yksikkömuotoa, mallinvalinta, tässä luvussa.

Mallinvalinnan suoritin käytännössä askeltaen. Tämä tarkoittaa sitä, että aloitin mallinvalinnan yksinkertaisimmista mallista ja etenin mallinvalinnassa laajentaen aina edellistä mallia tarpeelliseksi katsomillani parametreilla. Esitän jäljempänä jokaisen mallinnusaskeleen kohdalla, kuinka mallinlaajennus paransi mallitavoitteiden täyttymistä. Mallinnusaskeleita kertyi jokaisen maahanmuuttajaryhmän (ja Suomessa syntyneiden) kohdalla 5 kappaletta.

Määritin näiden kaikkien mallinnusaskeleiden parametreille priorijakaumat ja simuloin näin syntyneet nettovaikutusten posteriorijakaumat. Mallinnustuloksissa esitän kaikkien mallinnusvaiheiden mallien posterioriestimaatit parametrien (posteriori)jakaumien odotusarvoille. Lopullisen mallin eli elinkaariennusteissa käytettävän mallin parametrien (posteriori)jakaumille esitän myös muita jakaumatunnuslukuja. Mallinvalinta askeltaen noudatti noudatti karkeasti Macnabin et al. (2004) esimerkkiä: Aluksi määritin nk. *tyhjän mallin*, sen jälkeen lisäsin kaikille tilastoyksiköille yhteiset vaikutukset. Lopuksi lisäsin muut kuin henkilötason vakiotermin satunnaisvaikutukset sekä paransin edelleen mallia nettovaikutusten otosjakauman vinouden lisäämisellä malliin.

4.3.2.3 Perättäisten aineistomallien uuden aineiston ennustetarkkuuden mittaaminen: WAIC ja LOO

Erityisesti ensimmäinen mallinnustavoite (ks. luku 4.1.1, s. 127) tähtää *uuden aineiston* (*engl. out-of-sample*) ennustamisen hyvään ennustetarkkuuteen. ”Uusi aineisto” viittaa tässä ennusteisiin, joita tehdään nettovaikutusten osalta muille kuin aineistoon jo sisältyville tilastoyksiköille (uudet henkilöt eli uudet maahanmuuttajat, maahanmuuttovuodet ja maassaoloajat). Mallinnustavoitteessa mainittu ennustetarkkuus viittaa lähinnä posterioriennustetarkistuksilla todennettavaan tarkkuuteen, mutta ennustetarkkuutta koskien on olemassa formaalimpiakin mittareita.

Käytin ennustetarkkuuden mittaamisen näistä formaalimmista mittareista *WAIC*:ia (*Watanabe-Akaike tai Widely Available Information Criterion*, Watanabe (2010)) ja *LOO-CV*:ta (*Leave-one-out cross-validation*). Esitän erityisesti Gelmania et al. (2014) ja Vehtaria et al. (2015) lainaten, miksi käytän juuri näitä mittareita enkä muita ehkä tunnetumpia mittareita kuten esimerkiksi *AIC*:tä.

Laskin jokaisen mallinnusvaiheen estimoidulle mallille nämä ennustarkkuudet R-paketin *loo* (Vehtari et al. 2016) avulla. Ilmoitan näiden uuden aineiston ennustetarkkuuden mittareiden saamat arvot jokaisen mallin kohdalla tulostiedoissa jäljempänä. En käyttänyt ennustetarkkuuden mittareita mallinvalintaan, mutta niiden vertailu eri mallinnusaskelten mallien välillä osoittaa kuitenkin, että parametrimäärältään suuremmat mallit ovat parempia ennustajia uudelle aineistolle kuin pienemmät mallit, mikä osaltaan oikeuttaa mallien laajennukset uusilla parametreilla.

Keskeisessä roolissa ennustetarkkuuden mittaamisessa on aineistomallin *log-ennustetiheysfunktio* (tai log-uskottavuusfunktio). Havaintomäärän kasvaessa rajatta mallilla, jolla on pienin *Kullback-Leibler-informaatio*, on suurin log-ennustetiheysfunktion odotusarvo (*engl. log-predictive density mean*) ja siten suurin posterioritodennäköisyys (Gelman et al. (2014), 167; Kullback-Leibler-informaatiosta ks.

esimerkiksi Tutz 2012, 500–503). On syytä huomata, että ennustetarkkuutta mitataan nimenomaan aineistomallille $p(\boldsymbol{y}|\boldsymbol{\theta})$ eikä posteriorijakaumalle $p(\boldsymbol{\theta}|\boldsymbol{y})$, joten tarkastelun kohteena on aineistomallin log-tiheysfunktio log-posteriorin sijaan.

Uuden aineiston (*engl. out-of-sample*) log-tiheysfunktio saadaan log-posteriorista integroimalla estimoitavien parametrien vektorin $\boldsymbol{\theta}$ mahdollisten arvojen yli. Koska tulevan eli ennustettavan aineiston arvoja ei tunneta, ne puolestaan ”integroidaan pois” uuden aineiston log-ennustetiheysfunktion lauseesta. Näin saadaan *log-ennustetiheysfunktion odotusarvo* jokaiselle uudelle datapisteelle (*engl. expected log predictive density for a new data point, elpd*). On olemassa erilaisia menetelmiä estimoida tämä ennusteodotusarvo uudelle aineistolle. (Gelman et al. 2014, 168–170 ja Vehtari et al. 2015).

Käytännössä tämä voi tapahtua estimoimalla olemassa olevasta aineistosta datapisteiden log-ennustefunktiot ja korjata sitten jollain funktiolla harhaa, joka johtuu siitä, että ennusteeseen käytetään samaa aineistoa, josta malli on estimoitu. Tällaiseen menetelmään perustuvat varsin yleisesti käytetyt *AIC* (*Akaike information criterion*), *DIC* (*Deviance information criterion*) sekä edellä mainittu WAIC. Toinen tapa välttää ennustetarkkuudesta tulemasta liian suuri on poistaa aineistosta aina yksi datapiste kerrallaan, estimoida malli ilman sitä ja laatia ennustetarkkuus poistetulle havainnolle. Kun malli simuloidaan (bayesilaisessa yhteydessä) S kertaa kaikille havainnoille erikseen, saadaan ennustetarkkuuden mittariksi edellä mainittu LOO-CV. (Gelman et al. 2014, 168–170 ja Vehtari et al. 2015).

Käytän nettovaikutusten aineistomallien ennustetarkkuuksien mittaamisessa WAIC:ia ja LOO:ta. Sekä Gelman et al. (2014) (s. 174) että Vehtari et al. (2015) (s. 20) suosivat bayesilaisissa malleissa näitä mittareita AIC:n ja DIC:n sijaan, sillä AIC ja DIC ovat olennaisesti ei-bayesilaisia mittareita, sillä ne perustuvat estimoitavien parametrien piste-estimaatteihin. WAIC:n ja LOO:n bayesilaisen version kohdalla hyödynnetään sen sijaan estimoitavien parametrien koko posteriorijakaumia (Gelman et al. 2014, 174).

Kuten jo edellä mainitsin, mallikohtaiset estimaatit ja keskivirheet WAIC:lle ja LOO:lle sain R-ohjelman loo-paketin (Vehtari et al. 2016) avulla. Paketin menetelmät on esitetty Vehtarin et al. artikkelissa (2015). Koska LOO:n estimointi suoraan on erittäin aikaavievää, sen simulointiin on kehitetty *importance-otantaan* (ks. esim. Geyer 2011, 295–297) perustuva menetelmä (Vehtari et al. 2015, 3). Koska tämän menetelmän nk. *importance-painoilla* on kuitenkin suuri tai ääretön varianssi, mikä aiheuttaa ongelmia (mt, 3), Vehtari et al. (2015) ovat kehittäneet tähän ratkaisun, jossa importance-painojen suurimmat arvot estimoidaan ja *sileytetään* (*engl. smoothing*) erikseen *yleistettyyn pareto-jakaumaan* perustuvalla menetelmällä (mt, s. 3–5). Mallien tulosten yhteydessä esitetyt LOO-estimaatit ovat siten todellisuudessa pareto-sileytettyjä importance-otantaan perustuvia estimaatteja (*engl. lyh. PSIS*) LOO:lle.

Kuten mallien tulostietojen yhteydessä käy ilmi, WAIC- ja LOO-estimaatit ovat joko täysin identtisiä tai hyvin lähellä toisiaan ensimmäisen kolmen mallinnusaskeleen kohdalla. Tämä oli odotettavissa, sillä WAIC:n on osoitettu saavan asymptoottisesti yhtä suuria arvoja kuin LOO: Irakissa, Somaliassa ja Suomessa syntyneiden mallinnusaineistoissa oli jokaisessa noin 40 000 havaintoa, eli havaintomäärät olivat melko suuria (Gelman et al. 2014, 176). Mallinnusaskeleiden 4 ja 5 osalta, joissa otosjakaumana on *vino* Studentin t-jakauma, mallikohtaiset WAIC- ja LOO-estimaatit kuitenkin eroavat toisistaan. En osaa sanoa tälle poikkeamalle tarkkaa syytä.

4.3.2.4 Nettovaikutusten havaintoaineistojen jakaumat

Nettovaikutusten havaintoaineistojen jakaumien lyhyt kuvaus. Jokaisen 10 suurimman maahanmuuttajaryhmän kohdalla kaikkien nettovaikutushavaintojen yhteisjakaumat ovat yksihuippuisia – ainakin siinä tapauksessa, että nuorten eli 16–19-vuotiaiden nettovaikutukset mallinnetaan erikseen, eli tätä ikäluokkaa koskevat havainnot poistetaan työikäisten nettovaikutusmallinnuksesta. Mikäli 16–19-vuotiaat olisivat mukana esimerkiksi Suomessa syntyneiden osalta, tämä aiheuttaisi jakaumaan pienen, paikallisen toisen huipun, sillä suurin osa ikäluokan henkilöistä on opiskelijoita ja 16–19-vuotiaiden opiskelijoiden julkisen talouden nettovaikutukset ovat keskittyneet hyvin pienelle alueelle. Koska Irakissa ja Somaliassa syntyneiden osalta tätä toista 16–19-vuotiaista johtuvaa huippua ei ollut – tai se oli hyvin pieni – sisällytin Irakissa ja Somaliassa syntyneiden osalta myös 16–19-vuotiaat samaan mallinnusaineistoon kuin työikäiset 20–62-vuotiaat.

Nettovaikutusten kaikkien havaintojen jakaumat ovat kaikkein syntymävaltioiden osalta melko symmetrisiä. Ne vaikuttavat kuitenkin histogrammien että jakaumien tunnuslukujen perusteella olevan vasemmalle vinoja (ks. tulokset, luku 4.4).

Nuorten (16–19-vuotiaat) nettovaikutusten mallinnus. Koska Suomessa syntyneiden osalta 16–19-vuotiaiden sisällyttäminen samaan mallinnusaineistoon 20–62-vuotiaiden kanssa olisi aiheuttanut nettovaikutusten mallinnusaineistoon toisen, paikallisen huipun, mallinsin 16–19-vuotiaiden nettovaikutukset erikseen hyvin yksinkertaisella tavalla. Suomessa syntyneiden 16–19-vuotiaiden nettovaikutus muodostuu opiskelijoiden nettovaikutusten ja ei-opiskelijoiden nettovaikutusten painotetuista keskiarvoista. Painoina toimivat opiskelijoiden ja ei-opiskelijoiden ikävuosittaiset lukumäärät. Tämä yksinkertainen menettelytapa on varsin hyvin perusteltu 16–19-vuotiaiden kohdalla, sillä suurin osa 16–19-vuotiaiden nettovaikutusten kaikista havainnoista on keskittynyt kahden pisteen ympäristöön: opiskelijoiden kohdalla vuosittaisten opiskelumenojen kohdalle ja ei-opiskelijoiden kohdalla lähelle nollaa.

Ulkomailla syntyneiden osalta aineistossa vain yli 15-vuotiaana Suomeen muuttaneet. Koska haluan laatia elinkaariennusteet yli 15-vuotiaana Suomeen muuttaneille, sisällytin nettovaikutusten mallinnusaineistoon vain ne henkilöt, jotka ovat muuttaneet Suomeen yli 15-vuotiaana.

Havaintojen keskittyminen nollan ympäristöön. Henkilö on tiettynä kalenterivuonna Suomessa, kun hän täyttää tutkimuksen määritelmän asiaa koskien (ks. tutkimuksen I osa: *3 Käsitteet ja määritelmät: Maassaolovuodet*). Sisällytin henkilöitä koskevat nettovaikutushavainnot mukaan nettovaikutusten mallinnusaineistoon vain niiden kalenterivuosien osalta, jolloin henkilö oli tutkimuksen määritelmän mukaan Suomessa. Tämä osaltaan vähensi aineistossa niiden havaintojen lukumäärää, joiden nettovaikutus on tasan nolla tai hyvin lähellä sitä. Tästä huolimatta kaikissa ryhmäkohtaisissa aineistoissa havainnot keskittyivät jonkin verran nollan ympärille.

Yhdenkään ryhmäkohtaisen jakauman kohdalla havaintojen määrä nollassa ja sen ympärillä ei ollut kuitenkaan niin suuri, että edellä esitetyt mallinnustavoitteet kärsisivät siitä, että nettovaikutusten mallintamiseen käytetään vain *jatkuviin* todennäköisyysjakaumiin perustuvia malleja. Tämä tarkoittaa sitä, että

vaikka nollan kohdalle sijoittuvaa havaintojen ”piikkiä” ei mallinneta erikseen, tämän vaikutus mallinnustuloksiin on niin pieni, ettei sitä juuri huomaa, kun mallien perusteella simuloituja posterioriennusteita verrataan alkuperäisiin havaintoaineistoihin (ks. tulokset, posterioriennustetarkistukset). Jo alkuperäisten havaintoaineistojen jakaumien tarkastelu histogrammeilla osoittaa sen, että nollan ympärille ei ole yhdenkään ryhmän kohdalla keskittynyt ”merkittävää” määrää havaintoja.

Suurin vaikutus nollan ympärille jonkin verran keskittyvistä havainnoista on se, että posterioriennusteiden jakaumissa ennusteet vaihtelevat nollan ympäristössä hieman enemmän, kuin todellisuudessa tapahtuu havaintoaineistojen kohdalla.

4.3.3 Aineistomallin mallinnusaskeleet

4.3.3.1 Tyhjä hierarkkinen normaalimalli (Malli 1)

Aloitan mallinnuksen nk. *tyhjästä mallista*, jossa ei ole lainkaan selittäviä muuttujia mallin vakiota lukuun ottamatta. Malli on hierarkkinen: yksittäiset nettovaikutushavainnot ovat ryvästyneinä henkilöiden alle. Havainnoilla on henkilöittäin järjestysluku t, joka on henkilön maassaoloaika. Mallin kokonaisvaihtelu sijoittuu nyt kahdelle tasolle: henkilö- ja havaintotasolle. Kumpaakin tasoa koskee oma mallinsa: *aineistomalli* on nettovaikutushavaintojen malli, ja *populaatiomalli* tai *satunnaisvaikutusten malli* puolestaan on malli aineistomallin satunnaisvaikutuksille. Koska tyhjässä hierarkkisessa mallissa vain mallin vakio on *satunnainen*, populaatiomalli kertoo, kuinka aineistomallin vakio vaihtelee henkilöiden välillä. Käytän poulaatiomallista myös nimitystä henkilömalli, koska se mallintaa vaihtelua henkilöiden välillä.

Sekä aineistomallille että henkilömallille on määritettävä niiden noudattamat todennäköisyysjakaumat. Mallinvalinta askeleittain etenee osittain jakaumaoletusten laventamisella robustimpaan ja aineistoja enemmän vastaavaan suuntaan.

Sekä aineistomallin että henkilömallin kohdalla on hyvä aloittaa paljon käytetystä ja hyvin tunnetusta normaalijakaumaoletuksesta, koska sekä syntymävaltiokohtaiset nettovaikutusjakaumat että henkilöiden nettovaikutusten keskiarvojen jakaumat kummatkin ovat yksihuippuisia ja melko symmetrisiä.

Sisäkorrelaatiosta. Normaalijakaumaoletus sekä aineisto- että henkilömallille tyhjässä hierarkkisessa mallissa on myös siitä käyttökelpoinen valinta, että näin mallinnustulosten avulla voidaan estimoida nettovaikutusten *sisäkorrelaatio*. Tässä tapauksessa sisäkorrelaatio kertoo, kuinka suuri osa nettovaikutusten jakauman kokonaisvaihtelusta johtuu henkilöiden välisestä vaihtelusta:

$$\begin{aligned} &\textit{sisäkorrelaatio} \\ &= \textit{(henkilöiden "välinen"vaihtelu)}/(\textit{henkilöiden "välinen"vaihtelu} + \textit{henkilöiden ”sisäinen"vaihtelu}). \end{aligned} \tag{4.1}$$

Kuten tuloksista käy ilmi, sisäkorrelaatio on varsin korkea jokaisen maahanmuuttajaryhmän kohdalla: noin 0,40:n ympäristössä. Toisin sanoen noin 40 %:ia kaikesta nettovaikutushavaintojen välisestä vaihtelusta on henkilöiden välisestä vaihtelusta johtuvaa. Suomessa syntyneillä sisäkorrelaatio on vielä

korkeampi, noin 0,50.

1. mallinnusaskel. Tyhjä hierarkkinen normaalimalli, jossa satunnainen vakiotermi henkilötasolla:

$$y_{tim} \sim N(\beta_0 + U_i, \sigma^2) \tag{4.2}$$

ja

$$U_i \sim N(0, \tau^2), \tag{4.3}$$

missä β_0 on nettovaikutusten kaikilla havainnoille yhteinen odotusarvo, U_i on siihen lisättävä henkilötason satunnaisvaikutus, $\sigma^2 \in \mathbb{R}_{>0}$ on havaintojen varianssi ja $\tau^2 \in \mathbb{R}_{>0}$ henkilötason satunnaisen vakiotermin varianssi. Jokaisen henkilön i "satunnainen"vakio on näin ollen: $\beta_0 + U_i$.

Lisäksi mallin varianssiparametrien avulla saadaan ilmaistua sisäkorrelaatio:

$$\textit{sisäkorrelaatio} := \tau^2/(\tau^2 + \sigma^2). \tag{4.4}$$

4.3.3.2 Tyhjä hierarkkinen Studentin t-malli (Malli 2)

Maahanmuuttajaryhmäkohtaisten nettovaikutusjakaumien tarkastelu viittasi siihen, että jakaumat sisältäisivät enemmän poikkeavia arvoja eli paksummat hännät, kuin normaalijakaumaoletus aineistomallille antaisi olettaa. Lisäksi havaintomäärät ovat sen verran suuria aineistoissa, etten näe syytä olettaa, että havaintomäärien kasvaessa rajatta, jakaumat konvergoituisivat normaalijakaumaan. Koska normaalijakaumaan perustuvien mallien regressiokertoimien estimaatit ovat hyvin alttiitta yksittäisillekin poikkeaville havainnoille, päätin kokeilla normaalijakauman tilalla aineistomallissa normaalijakaumaa vastaavaa robustimpaa jakaumaa. Robustimpi jakauma ei ole yhtä herkkä poikkeavien havaintojen suhteen. Myös mallinnustavoitteen (alaluku 4.1.1) nojalla suosin jakaumaa, josta aineisto on todennäköisemmin generoitu kuin normaalijakaumasta.

Normaalijakauman tapauksessa robustimmaksi jakaumavaihtoehdoksi on helppo valita *Studentin t-jakauma*. Se näyttää paljolti normaalijakaumalta, mutta sen paksumpien jakaumahäntien vuoksi se soveltuu normaalijakaumaa paremmin mallinnustilanteisiin, joissa on enemmän poikkeavia havaintoja. Studentin t-jakauma on myös helppo valinta, koska se on usein käytetty ja tunnettu jakauma. Sillä on myös yhteys normaalijakaumaan; Studentin t-jakaumalla on normaalijakaumaan verrattuna yksi parametri enemmän: jakauman vapausasteet. Studentin t-jakauma, jonka vapausaste on ääretön, on normaalijakauma, joten normaalijakauma on Studentin t-jakauman erityistapaus. Lisäksi Studentin t-jakauma voidaan esittää bayesilaisessa yhteydessä – ja siten myös simuloida – posteriorijakaumana, joka on tietyn priorijakauman määrittelemien normaalijakaumien *sekoitusjakauma* (*engl. mixture distribution*) (Geweke 1993, Gelman et al. 2014, 437).

Jatkossa, kun käytän termiä t-jakauma, tarkoitan nimenomaan Studentin t-jakaumaa – erotuksena yleistetymmistä t-jakaumista, joista Studentin t-jakauma on erityistapaus. Käsittelen jäljempänä myös vinoa

Studentin t-jakaumaa: siihen viittaan vinona t-jakaumana.

2. mallinnusaskel. Tyhjä hierarkkinen Studentin t-malli, jossa satunnainen vakiotermi henkilötasolla:

$$y_{tim} \sim T(\beta_0 + U_i, \sigma, \nu) \tag{4.5}$$

ja

$$U_i \sim N(0, \tau^2), \tag{4.6}$$

missä β_0 on nettovaikutusten kaikilla havainnoille yhteinen odotusarvo, U_i on siihen lisättävä henkilötason satunnaisvaikutus, $\sigma \in \mathbb{R}_{>0}$ on havaintojen hajontaparametri, $\nu \in \mathbb{R}_{>0}$ on t-jakauman vapausasteparametri ja $\tau^2 \in \mathbb{R}_{>0}$ henkilötason satunnaisvakion varianssi.

Yllä tyhjässä hierarkkisessa Studentin t-mallissa henkilömallin oletetaan noudattavan normaalijakaumaa. Myös tämä jakaumaoletus voidaan kyseenalaistaa, ja testata robustimpaa vaihtoehtoa. Tämän vuoksi estimoin kaikkien 10 suurimman maahanmuuuttajaryhmän kohdalla siten myös tyhjät mallit, joissa sekä y_{imt} että U_i ovat t-jakautuneita ja joissa niiden vapausasteet ovat estimoitavia parametreja.

Kaikkien 10 suurimman maahanmuuttajaryhmän osalta tulokset olivat samansuuntaisia: aineistomalleissa t-jakauman vapausasteet olivat pieniä, ja niiden jakaumat olivat sijoittuneet suurimmaksi osaksi arvojen 1–3 välille. Sitä vastoin henkilömallien t-jakaumien vapausasteiden jakaumat olivat niin suuria, että ne olivat käytännössä jo hyvin lähellä normaalijakaumaa, joten käytin ryhmäkohtaisten henkilömallien jakaumina normaalijakaumaa Irakissa ja Somaliassa syntyneiden osalta. Irakissa syntyneillä estimoidun t-jakautuneen henkilömallin vapausasteiden odotusarvoestimaatti oli 116, ja keskihajontaestimaatti oli 26. Somaliassa syntyneillä estimoidun t-jakautuneen henkilömallin vapausasteiden odotusarvoestimaatti oli 131 ja, keskihajontaestimaatti oli 27.

Suomessa syntyneiden osalta estimointitulokset t-jakautuneen henkilömallin osalta poikkesivat Somaliassa ja Irakissa syntyneiden tuloksista. Suomessa syntyneillä t-jakautuneen henkilömallin vapausasteiden odotusarvoestimaatti oli 3,5, ja keskihajontaestimaatti oli 0,21. Koska Suomessa syntyneiden osalta sekä t-jakautuneen aineistomallin että t-jakautuneen henkilömallin vapausasteet olivat pieniä, päätin käyttää Suomessa syntyneiden osalta suuremmissakin malleissa t-jakaumaa sekä aineistomallin että henkilömallin osalta.

4.3.3.3 Selittävät muuttujat hierarkkisessa Studentin t-mallissa (Malli 3)

Kun olin määritellyt sekä aineistomallin ja henkilömallin jakaumat, lisäsin seuraavassa askeleessa hierarkkiseen malliin kaikille havainnoille (tai havaintoryppäille) *yhteiset selittävät muuttujat*. ”Havainnoille yhteiset selittävät muuttujat” tarkoittaa käsitteenä sitä, että näiden muuttujien regressiokertoimien arvot

eivät vaihtele ”satunnaisesti” havaintoryppäittäin kuten esimerkiksi henkilöittäin. Edellä ”tyhjissä” malleissa vakiotermin oletettiin vaihtelevan satunnaisesti henkilöittäin normaalijakauman mukaisesti, jonka varianssi on τ^2. Mikäli mallin vakiotermi olisi yhteinen kaikille havainnoille, eli se ei olisi ”satunnainen”, tämä olisi tarkoittanut vakiotermin kohdalla sitä, että sen varianssi oli ollut: $\tau^2 = 0$.

Käsittelen selittäviä muuttujia kolmessa eri ryhmässä: aikaan liittyvissä muuttujissa, henkilöihin liittyvissä muuttujissa sekä selitettävän muuttujan *viivästyneitä arvoja* tai *viipeitä* (*engl. lags*) koskevissa autoregressiivisissa muuttujissa. Tämä jako ei noudata aineiston hierarkkista rakennetta, mutta mallinnustavoitteiden tarkastelun suhteen tällainen jako on selkeämpi. Henkilöihin liittyvien muuttujien ("henkilötason"muuttujat) osalta on syytä huomata, että nämä muuttujat ovat *aikainvariantteja* eli ajassa muuttumattomia: henkilöön liittyvien muuttujien arvot määräytyvät henkilön ensimmäisen Suomeen muuton vuoden arvojen mukaan.

1. Aikamuuttujat

Sisällytän mukaan lopulliseen aineistomalliin aikaan liittyvät selittävät muuttujat, koska se kuuluu mallinnustavoitteisiin (ks. alaluku4.1.3). Aikaan liittyvät muuttujat voivat liittyä joko henkilöiden ikävuosiin, maahanmuuttovuoteen, maassaoloaikaan tai kalenterivuoteen. Päävaikutusten lisäksi aikaan liittyvillä selittävillä muuttujilla voi olla interaktioita henkilöihin liittyvien muuttujien kanssa: tällöin aikamuuttujien regressiokertoimien arvot voivat vaihdella näiden henkilöön liittyvien selittävien muuttujien eri luokkien välillä.

1.1 Ikävuoteen liittyvät muuttujat

Ikävuoden funktioita olevia muuttujia mallissa ovat: ikävuosi, ikävuoden neliö sekä ikävuosiin liittyvät indikaattorimuuttujat kuten alle 20-vuotias -indikaattori ($I(ikä_{ti} < 20)$) tai yli 60-vuotias-indikaattori ($I(ikä_{ti} > 60)$. Ikävuoden neliön sisällytän malliin, sen perusteella, että nettovaikutus vaikuttaa aineistontarkastelun perusteella aluksi muuttuvan positiivisemmaksi iän kasvaessa, mutta laskee jälleen lähestyttäessä eläkeikää. Mainitut ikäluokkaindikaattorit sisällytän puolestaan malleihin, koska alle 20-vuotiaista ja yli 60-vuotiaista suuri osa on iän puolesta työmarkkinoiden ulkopuolella ja siten näiden ikäluokkien odotusarvoiset nettovaikutukset ovat negatiivisemmat kuin työikäisillä 20–60-vuotiailla. Otin syntymävaltiokohtaisten aineistontarkasteluiden perusteella mukaan myös muita ikäluokkaindikaattoreita syntymävaltiokohtaisiin malleihin. Lisäksi malleihin sisältyy indikaattoreita maahanmuuttoikäluokista.

1.2 Maahanmuuttovuoteen liittyvät muuttujat

Maahanmuuttovuoteen liittyen malleissa on mukana maahanmuuttovuosien luokat sekä maahanmuuttovuosien yleiset työttömyysasteet Suomessa Tilastokeskuksen *työvoimatutkimuksen* mukaan (SVT: *työvoimatutkimus*). Maahanmuuttovuodet ovat henkilöiden ensimmäisen muuton vuosia Suomeen. Jos henkilö on muuttanut Suomeen useita kertoja, maahanmuuttovuosi on hänen ensimmäisen muuttonsa kalenterivuosi.

Toisin kuin ikävuosi ja maassaoloaika maahanmuuttovuosi pysyy yhden henkilön kaikkien havaintojen

osalta samanarvoisena, sillä yhdellä henkilöllä voi olla vain yksi ensimmäinen maahanmuuttovuosi Suomeen. Maahanmuuttovuosi voi vaikuttaa julkisen talouden nettovaikutuksiin joko sitä kautta, millaiset olosuhteet Suomessa ovat olleet maahanmuuttovuonna tai vaihtoehtoisesti sitä kautta, millaisia maahanmuuttajia Suomeen on muuttanut tiettynä vuonna. Nämä tekijät voivat tietenkin olla yhteydessä toisiinsa. Koska en tutki mahdollisia syy–seuraussuhteita selittävien ja selitettävän muuttujan välillä, en ota sen tarkemmin kantaa prosessiin, jolla maahanmuuttovuosi mahdollisesti vaikuttaa maahanmuuttajaryhmäkohtaisesti nettovaikutuksiin.

Maahanmuuttovuoden työttömyysasteen otan mukaan malleihin siitä syystä, että muuttovuoden työttömyystilanne voi olla hypoteettisesti yhteydessä siihen, miten maahanmuuttajat työllistyvät ja siten siihen, millaiset heidän elinkaarivaikutuksensa ovat Suomen julkiselle taloudelle. Tällaisen vaikutuksen mallin ennustetarkkuutta parantavasta olemassaolosta – varsinkaan maahanmuuttajaryhmäkohtaisesti – minulla ei ollut tietoa analyysia tehdessäni, mutta halusin varmistaa, että vaikutus on malleissa mukana joka tapauksessa.

1.3 Maassaoloaikaan liittyvät muuttujat

Maassaoloaika Suomessa kertoo, kuinka monta vuotta henkilön ensimmäisestä muutosta Suomeen on kulunut tiettynä kalenterivuonna. Toisin sanoen, mikäli henkilö on muuttanut useammin kuin yhden kerran Suomeen (välissä poismuutto), maassaoloaika lasketaan nettovaikutusmallissa kuitenkin ensimmäisestä muutosta. Aiemmin tässä tutkimuksessa esitetyssä maassaoloajan mallinnuksessa kohteena olivat yksittäiset jaksot Suomessa, eli yhdellä henkilöllä saattoi olla useita erisuuruisia maassaoloajan jaksoja.

Nettovaikutusmallinnuksessa käytettävä maassaoloajan määritelmä helpottaa nettovaikutusennusteiden laatimisen elinkaariennusteita varten, sillä maassaoloaika on deterministisesti laskettavissa nettovaikutusennusteita varten, eikä siihen tarvita erikseen olotilaennusteita.

Koska maassaoloajan positiivinen vaikutus näyttää olevan suurempaa ensimmäisinä maassaoloajan vuosina, käytän malleissa maassaoloajan indikaattoreita sekä näiden indikaattoreiden interaktioita maassaoloajan kanssa. Näin voin määrittää maassaoloajalle erilaisen vaikutuksen esimerkiksi alle 5 vuoden maassaoloajalle sekä maassaoloajoille, jotka ovat 5 vuotta tai sitä suurempia.

1.4 Erityiset maahanmuuttajaryhmäkohtaiset aikamuuttujat

Edellä mainitut aikaan liittyvät muuttujat sisältyvät kaikkiin maahanmuuttajaryhmäkohtaisiin malleihin. Ainoastaan Suomessa syntyneiden osalta olen vaihtanut maassaoloajan kalenterivuoteen, sillä en usko, että maassaoloaika antaisi Suomessa syntyneiden osalta mitään lisäinformaatiota malliin, kun siihen jo sisältyy ikään liittyvät muuttujat.

Lisäsin kuitenkin maahanmuuttajaryhmäkohtaisesti malleihin joitakin juuri näitä ryhmiä koskevia aikaan liittyviä muuttujia sekä ikään ja maassaoloaikoihin liittyen. Lisäksi valitsin Somalialla ja Irakille hieman erilaiset maahanmuuttovuosien luokat. Nämä valinnat tein sekä ryhmäkohtaisiin aineistontarkasteluihin että näitä ryhmiä erityisesti koskeviin tunnettuihin historiallisiin tapahtumiin perustuen. Aineistontarkastelulla viittaan muuttajakohtaisiin hajontakuvioihin selitettävän muuttujan kanssa.

1.5 Interaktiot

Olen käsitellyt edellä aikaan liittyvien muuttujien osalta lähinnä niiden päävaikutuksia. Päävaikutusten lisäksi niillä voi olla interaktiovaikutuksia muiden mallin selittävien muuttujien kanssa. Mikäli aikaan liittyvillä muuttujilla on interaktiovaikutuksia henkilöihin liittyvien muuttujien kanssa, tämä tarkoittaa sitä, että näiden aikaan liittyvien muuttujien regressiokertoimien annetaan vaihdella henkilöön liittyvien muuttujien luokkien välillä. Esimerkiksi jos henkilöön liittyvällä luokitellulla muuttujalla on interaktio maassaoloajan kanssa, tämä tarkoittaa sitä, että maassaoloajan vaikutus ei pysy samana kaikissa henkilöön liittyvän muuttujan luokissa, vaan maassaoloajan vaikutus muuttuu selittävän muuttujan luokan mukaan.

Koska tutkimuksen I osan perusteella useiden eri maahanmuuttajaryhmien kohdalla erot julkisen talouden nettovaikutuksissa ovat suuret erityisesti naisten ja miesten välillä, päätin antaa aikamuuttujien vaikutuksen vaihdella maahanmuuttajaryhmäkohtaisissa malleissa sukupuolen mukaan. Tästä syystä malleissa on interaktiot erityisesti sukupuolen ja aikaan liittyvien muuttujien välillä. Sukupuoli on malleissa määritelty naisindikaattorina ($I(sukupuoli_i = nainen)$).

2. Henkilötason muuttujat maahanmuuttohetkellä

Henkilöihin liittyvät selittävät muuttujat ovat pitkälti samat kuin maassaoloajan mallinnuksen yhteydessä (ks. luku 3.2.3.4). Maassaoloajan mallinnuksen yhteydessä selittävien muuttujien valintakriteerit – käytettävissä olevan FLEED-rekisteriaineiston muuttujista (Tilastokeskus 2016) – määrittelivät valittavat selittävät muuttujat siten, että ne joko olivat yleisiä selittäviä muuttujia yhteiskuntatieteellisissä tutkimuksissa tai mahdollisesti yhteydessä henkilöiden taloudellisiin ja sosiaalisiin siteisiin Suomessa ja siten maassaoloajan pituuteen.

Noudatan nettovaikutusmallien selittävien muuttujien valinnassa pääasiassa linjaa, jonka mukaan ennusteiden luotettavuus (validiteetti) on pieni hinta siitä, että ennusteista saattaa muodostua hieman epätarkempia kuin, jos osa selittävistä muuttujista olisi jätetty kokonaan pois mallista. Erityisesti otan malleihin mukaan kaikki olennaisimmat aikaan liittyvät selittävät muuttujat. Lisäksi selittävien muuttujien kokonaan poisjättäminen mallista vastaa tilannetta, jossa niiden arvot kiinnitettäisiin mallissa tasan nollaan. Tämä on hyvin ”epä-bayesilainen” ratkaisu, sillä bayesilaisessa estimoinnissa estimoitavilla parametreilla on aina oma todennäköisyysjakaumansa. (Gelman et al. 2014, 367).

Koska maassaolomallinnuksen selittävien muuttujien valintakriteerit sopivat sellaisenaan mielestäni myös nettovaikutusmallien selittävien muuttujien valintakriteereiksi, valitsin maahanmuuttajaryhmäkohtaisiin nettovaikutusmalleihin lähes samat henkilöitä koskevat selittävät muuttujat kuin maassaolomalleissa. Ainoat lisätyt henkilötason selittäjät verrattuna maassaolomallinnukseen ovat Uusimaa-indikaattorin sekä mahdollisen puolison nettovaikutuksen sisällyttäminen lopulliseen malliin. Uusimaa-indikaattori kertoo, asuiko henkilö ensimmäisen maassaolovuotensa lopussa Uudellamaalla vai jossain toisessa maakunnassa. Ajatuksena tämän muuttujan mukaanotossa on se, että asuinpaikka Uudellamaalla saattaa kenties olla yhteydessä henkilön mahdollisuuksiin työllistyä ja sitä kautta henkilön julkisen talouden nettovaikutuksiin. Puolison nettovaikutuksia käsittelen vielä erikseen viimeisen mallinnusaskeleen yhteydessä alla luvussa 4.3.3.5.

Harkitsin myös omistusasunto-indikaattorin ottamista mukaan malleihin. Indikaattori kertoo, onko henkilö asunut ensimmäisen muuttovuoden lopussa asuntokunnassa, jonka asunnon asumismuoto on omistusasunto vai jokin muu asumismuoto. Ajatuksena tässä oli, että omistusasuminen kertoisi jotakin asuntokunnan varallisuudesta – ja siten mahdollisesti myös henkilön varallisuudesta – ja että varallisuudella olisi yhteys henkilön nettovaikutuksiin.

Kokeilin omistusasunto-indikaattorin sisällyttämistä joihinkin maahanmuuttajaryhmäkohtaisiin malleihin, mutta kaikkien näiden tuloksissa kyseisen muuttujan posteriorijakauma oli sijoittunut nollan ympärille. Päätin tästä – enemmänkin frekventistiseen päättelyyn perustuvasta syystä – olla ottamatta muuttujaa mukaan malleihin. Sama koski myös alle 18-vuotiaiden lasten määrää asuntokunnassa – malleissa on kuitenkin mukana alle 7-vuotiaiden lasten määrä asuntokunnassa. Muut FLEED-aineiston muuttujat olivat sellaisia, etten uskonut niillä olevan ennustevoimaa henkilön nettovaikutusten suhteen.

Kuten aiemmin jo mainitsin, henkilöitä koskevien muuttujien tiedot koskevat henkilön ensimmäisen Suomeen muuton muuttovuoden tietoja. Tarkasti ottaen tiedot koskevat sen kalenterivuoden, jolloin ensimmäinen muutto Suomeen on tapahtunut, viimeisen päivän tietoja. Tämä tarkoittaa sitä, että henkilötason tiedot eivät muutu ajassa. Tämä mahdollistaa mallien soveltamisen nettovaikutusennusteiden laatimiseen uusille maahanmuuttajille, eli niille henkilöille, jotka eivät sisältyneet mallinnusaineistoon.

Alla on lueteltu nettovaikutusmallien henkilöihin liittyvät selittävät muuttujat (vain päävaikutukset). Selittävät muuttujat sisältävät jokaisen henkilön kohdalla tietoja henkilöstä siltä vuodelta, kun hän muuttaa ensimmäisen kerran Suomeen. Henkilötason selittävät muuttujat, jotka valitsin nettovaikutusmalleihin, olivat:

– *Nainen* (vertailuluokka: mies).

– *Korkein koulutus: 2. asteen koulutus* (vertailuluokka: ei 2. asteen koulutusta tai ei tietoa).

– *Korkein koulutus: 3. asteen koulutus* (vertailuluokka: ei 2. asteen koulutusta tai ei tietoa).

– *Pääasiallinen toiminta: työllinen* (vertailuluokka: muu kuin työllinen tai opiskelija).

– *Pääasiallinen toiminta: opiskelija* (vertailuluokka: muu kuin työllinen tai opiskelija).

– *Puoliso* (vertailuluokka: ei puolisoa).

– *Suomessa syntynyt puoliso* (vertailuluokka: ei Suomessa syntynyttä puolisoa).

– *Alle 7-vuotiaiden lasten lukumäärä* (perheessä).

– *Kotikunta Uudellamaalla* (vertailuluokka: kotikunta ei Uudellamaalla).

3. Autoregressiiviset muuttujat

Kuten jo aiemmin mainitsin, paneeliaineistona kerätyn nettovaikutusaineiston vaihtelu voidaan jakaa

kahteen osaan: henkilöiden väliseen (*engl. between*) vaihteluun sekä henkilöiden ”sisäiseen” (*engl. within*) vaihteluun. Jälkimmäinen voidaan puolestaan jakaa yhdeltä henkilöltä saatujen ajassa perättäisten havaintojen väliseen lineaariseen riippuvuuteen (*engl. serial correlation*) sekä jäljelle jäävään vaihteluun.

On hyvin realistista olettaa, että vaikka sellaisten aikaan liittyvien muuttujien kuten ikä ja maassaoloaika mukaanotto malleihin poistaa ajasta riippuvia trendejä perättäisten havaintojen väliltä, jäljellä jää silti lineaarista riippuvuutta perättäisten havaintojen välille. Tämän korrelaation mallintamista varten malliin on otettava mukaan nk. *autoregressiivisia* muuttujia.

Autoregressiiviset selittävät muuttujat ovat selitettävän muuttujan menneitä arvoja. Kun tietyn henkilön jonkun tietyn maassaolovuoden julkisen talouden nettovaikutuksia selitetään tai ennustetaan, autoregressiivisten muuttujien arvot ovat kyseisen henkilön edellisinä maassaolovuosina havaitut – tai niille ennustetut – julkisen talouden nettovaikutusten arvot. Tilastollisen selvityksen kohteena on tässä yhteydessä, kuinka pitkälle menneisyyteen ulottuvilla nettovaikutusten arvoilla on vielä ennustetarkkuutta lisäävä vaikutus nykyistä arvoa selitettäessä tai tulevia arvoja ennustettaessa, kun muut selittävät muuttujat otetaan myös huomioon. Tämä vastaa likimain sen selvittämistä, kuinka monta maassaoloajan viivästettyä havaintoa (vuotta) on mentävä ajassa taaksepäin nykyisestä nettovaikutuksen arvosta, jotta autokorrelaatio (serial correlation) ei eroa enää nollasta.

Menneiden havaintojen vaikutus voidaan mallintaa joko mallin sovitteen kautta tai virheiden kovarianssimatriisissa. Käytän edellä mainittua menetelmää, jolloin autoregressiiviset vaikutukset ovat mallissa mukana muiden selittävien muuttujien tavoin. Jälkimmäisen menetelmän käyttö tarkoittaisi sitä, että tietyn henkilön ajassa lähellä toisiaan olevat nettovaikutusten *virheet* (*engl. errors*) olisivat korreloituneita autoregressiivisten muuttujien määräämällä tavalla. Perättäisten havaintojen (tai ennusteiden) korreloituneituneisuutta voisi tutkia käyttämieni menneiden nettovaikutushavaintojen eli autoregressiivisten muuttujien (engl. autoregressive, AR) avulla myös menneiden nettovaikutusten virheiden aiheuttamien ”shokkien” kautta (*engl. moving average, MA*).

Käytännössä mallinsin autoregressiiviset vaikutukset maakohtaisissa malleissa mahdollisimman yksinkertaisesti. Tällä tarkoitan sitä, että oletan perättäisten havaintojen autokorrelaatioiden olevan samansuuruisia kaikkien havaintojen välillä, riippumatta esimerkiksi henkilön sukupuolesta. Päämotivaationa autoregressiivisten vaikutusten mallintamisessa minulla oli ottaa huomioon malleissa edes keskimäärin perättäisten havaintojen välinen korrelaatio. En pitänyt autokorrelaatiorakenteen tarkempaa mallintamista tärkeänä mallinnustavoitteiden toteutumisen kannalta. Joka tapauksessa maahanmuuttajaryhmäkohtaisia malleja voisi laajentaa siten, että autokorrelaatiorakenteen antaisi vaihdella esimerkiksi henkilöiden sukupuolen, maassaoloajan tai muiden selittävien muuttujien mukaan.

Mallinnustehtäväksi jäi näin selvittää maahanmuuttajaryhmäkohtaisten mallien osalta autoregressiivistien muuttujien dimensio, eli kuinka monta nettovaikutusten viivästynyttä arvoa malliin sisällytettäisiin jokaisen havainnon osalta henkilöittäin. Selvitin tämän jokaisen maahanmuuttajaryhmän kohdalla niin, että lisäsin malliin, joka sisälsi muut selittävät muuttujat, askeltaen aina yhden viivästyneen arvon, kunnes suurinta viivettä edustavan autoregressiivisen termin estimaatin todennäköisyysjakaumasta yli 95 %:a ei ollut enää nollaa suurempi. Koska mallien ajaminen kaikilla havainnoilla olisi vienyt liikaa aikaa, käytin kunkin ryhmän kohdalla 3 000 satunnaisesti valittua havaintoa, jotka koskivat noin 300 henkilöä. Tulokset olivat hyvin samanlaiset kaikkien ryhmien kohdalla.

Valintamenettelyn mukaan autoregressiivisten vaikutusten dimensio oli lähes jokaisen maan kohdalla kaksi: ensimmäisen viivästyneen arvon autoregressioparametrin jakauma oli yleensä keskittynyt melko suuriin arvoihin, noin $[+0{,}30; +0{,}50]$ välille, mutta jo toisen viivästyneen arvon estimaatin jakauma oli keskittynyt lähelle nollaa. Toisin sanoen henkilöiden tietyn maassaolovuoden nettovaikutuksilla oli keskimäärin ehdollinen lineaarinen yhteys nettovaikutuksiin, jotka olivat siitä korkeintaan kahden (maassaolo)vuoden etäisyydellä. Ehdollistajina toimivat muut nettovaikutusmallin selittävät muuttujat.

3. mallinnusaskel. Kaikille havainnoille yhteiset selittävät muuttujat hierarkkisessa Studentin t-mallissa, jossa satunnainen vakiotermi henkilötasolla:

$$y_{tim} \sim T(\beta_0 + \boldsymbol{X_i}\boldsymbol{\beta_p} + U_i + \boldsymbol{X_m}\boldsymbol{\beta_q} + \boldsymbol{X_t}\boldsymbol{\beta_r} + \boldsymbol{X_{tim}}\boldsymbol{\beta_s} + \boldsymbol{y_{i(t-u)m}}\boldsymbol{\beta_u}, \sigma, \nu) \tag{4.7}$$

ja

$$U_i \sim N(0, \tau^2), \tag{4.8}$$

missä β_0, U_i, σ^2, ν ja τ^2 ovat kuten edellä mallinnusaskeleessa 2.

$\boldsymbol{X_i}$ on puolestaan p kappaletta henkilöittäin vaihtelevia selittäviä muuttujia sisältävä vaakavektori, ja $\boldsymbol{\beta_p}$ on kaikille havainnoille yhteinen regressiokertoimien p-ulotteinen pystyvektori. $\boldsymbol{X_m}$ on q kappaletta maahanmuuttovuosittain vaihtelevia selittäviä muuttujia sisältävä vaakavektori, ja $\boldsymbol{\beta_q}$ on kaikille havainnoille yhteinen regressiokertoimien q-ulotteinen pystyvektori. $\boldsymbol{X_t}$ on r kappaletta maassaolovuosittain vaihtelevia selittäviä muuttujia sisältävä vaakavektori, ja $\boldsymbol{\beta_r}$ on kaikille havainnoille yhteinen regressiokertoimien r-ulotteinen pystyvektori. $\boldsymbol{X_{tim}}$ on s kappaletta maassaolovuosittain, henkilöittäin ja maahanmuuttovuosittain vaihtelevia selittäviä muuttujia sisältävä vaakavektori, ja $\boldsymbol{\beta_s}$ on kaikille havainnoille yhteinen regressiokertoimien s-ulotteinen pystyvektori. $\boldsymbol{y_{(t-u)im}}$ on u kappaletta henkilön nettovaikutuksen viivästyneitä arvoja sisältävä vaakavektori, ja $\boldsymbol{\beta_u}$ on niiden kaikille havainnoille yhteinen autoregressiokertoimien u-ulotteinen pystyvektori.

4.3.3.4 Satunnaisvaikutusten lisääminen hierarkkiseen Studentin t-malliin (Malli 4, osa I)

Edellä esitetyissä mallinnusaskelmissa ainoa mallinnettu satunnaisvaikutus on henkilötason vakiotermi. Tässä mallinnusaskeleessa lisään mallin joustavuutta ottamalla malliin mukaan myös muita satunnaisvaikutuksia. Tämä on olennaisen tärkeää jo sen vuoksi, että tutkimuksen tavoitteena on laatia uskottavia ennusteita uudelle aineistolle eli henkilöille, maahanmuuttovuosille (vuoden 2011 jälkeen) ja maassaoloajoille (yli 23 vuotta), jotka eivät sisälly mallinnusaineistoon. Näin ennusteista tulee realistisempia, sillä niihin sisältyy epävarmuus, joka aiheutuu mainitun uuden aineiston käytöstä. Nämä epävarmuudet estimoidaan aineistosta, joten käsittelen vielä erikseen aineiston ulkopuolisten tekijöiden vaikutusten tekemää lisäepävarmuutta johdannossa (luku 2.3.4) sekä erityisesti päätöksentekoteoriaa käsittelevässä luvussa (luku 12.3.1). Lisäksi satunnaisvaikutusten huomiointi täyttää toisen mallinnustavoitteen eli havaintoaineiston hierarkkisuuden huomioinnin (ks. edellä luku 4.1.2).

Henkilötason satunnaisvaikutukset

Henkilötasolla vakion lisäksi myös *kulmakertoimet* (*engl. slopes*) eli regressiokertoimet voivat vaihdella henkilöittäin. Regressiokertoimet, joihin tässä yhteydessä viittaan, liittyvät iästä ja maassaoloajasta muodostettuihin selittäviin muuttujiin. Koska henkilöt ovat ryvästyneet maahanmuuttovuosien alle, ja jokaisella henkilöllä on vain yksi maahanmuuttovuotta koskeva arvo, maahanmuuttovuoden regressiokerroin ei ole henkilötasolla satunnainen.

Ottamalla mukaan malliin ikään ja maassaoloaikaan liittyvät henkilötason satunnaisvaikutukset, voin huomioida uudelle aineistolla tehtävissä ennusteissa sen epävarmuuden, joka seuraa siitä, että uusilla, eli ei mallinnusaineistoon sisältyvillä, henkilöillä aikamuuttujien vaikutus voi vaihdella satunnaisesti henkilöittäin. Syy tähän mahdolliseen vaihteluun on, että henkilöillä voi olla sellaisia havaitsemattomia vaikutuksia (*havaitsematonta heterogeenisuutta*), joka liittyy aikamuuttujiin.

Ikään liittyviä, eli iän funktiona muodostettavia, selittäviä muuttujia malleissa ovat ikä, iän neliö, ikäluokkaindikaattorit: esimerkiksi alle 20-vuotias- ja yli 60-vuotias -indikaattori. Koska kaksi ensimmäistä (ikä ja iän neliö) koskevat koko varsinaisen työiän (20–60 ikävuotta) intervallia, jossa suurin osa henkilöiden välisistä julkisen talouden elinkaarivaikutusten eroista syntyy, niitä koskevat satunnaisvaikutukset ovat ennusteiden kannalta tärkeimpiä.

Kokeilin aluksi joidenkin maahanmuuttajaryhmäkohtaisten mallien osalta malleja, joissa henkilötason satunnaisen vakion lisäksi malleissa oli henkilötason satunnaisvaikutuksena myös iän regressiokerroin. Vakiotermin ja iän regressiokertoimen satunnaisvaikutukset olivat kuitenkin erittäin korreloituneita, sillä näiden satunnaisvaikutusten välisen korrelaation estimaatin jakauma oli keskittynyt $+0{,}90$:n ja $+1$:n välille. Koska korrelaatio oli keskittynyt lähelle ykköstä – eli singulariteettia mallissa – valitsin estimoitavaksi ikään liittyväksi satunnaisvaikutukseksi iän neliön regressiokertoimen. Iän neliön lisäksi toinen satunnainen henkilötason regressiokerroin malleissa om maassaoloajan regressiokerroin.

Aikatasot: maahanmuuttovuosiin ja maassaoloaikoihin liittyvät satunnaisvaikutukset

Kuten edellä mainitsin, maahanmuuttovuoden ja maassaoloajan satunnaisvaikutukset on syytä sisällyttää malliin, jotta tehtävistä ennusteista tulisi uskottavia. Melkein kaikkien maahanmuuttajaryhmien osalta niiden aineistot sisältävät 24 eri arvoa maahanmuuttovuosille $\{1988, 1989, \ldots, 2010, 2011\}$ sekä maassaoloajalle $\{0, 1, \ldots, 22, 23\}$. Kun näiden maahanmuuttovuosien sekä maassaoloajan vaikutusten annetaan vaihdella satunnaisesti tiettyjen jakaumaoletusten puitteissa mallissa, ne lisäävät mallin kaksi satunnaisvakiota (*engl. random intercept*) lisää, nyt sekä maahanmuuttovuoden että maassaoloajan vaihtelun "tasoille".

Myös maahanmuuttovuoden ja maassaoloajan tasot voivat sisältää satunnaisvakioiden lisäksi myös satunnaisia kulmakertoimia (*engl. random slopes*) eli regressiokertoimia. Malleihin voi sisällyttää näiden aikatasojen satunnaisvaikutuksina esimerkiksi sukupuolen (naisindikaattorin) regressiokertoimen tai muita henkilötason selittävien muuttujien regressiokertoimia. Tämä tarkoittaa sitä, että henkilötason selittävien muuttujien regressiokertoimet voivat muuttua ajassa maahanmuuttovuoden tai maassaoloajan mukaan ”satunnaisesti”. (Lineaarinen muutos on jo malleissa naisindikaattorin ja näiden aikamuuttujien välisinä interaktioina). Koska sukupuoli on aikaan liittyvien vaikutusten ohella tärkein selittävä muuttuja

malleissa, otin sukupuolen regressiokertoimen satunnaisvaikutukseksi sekä maahanmuuttovuoden että maassaoloajan tasoilla.

Malleja voisi laajentaa vielä satunnaisvaikutusten osalta ottamalla malleihin mukaan lisää satunnaisvaikutuksia. Katson kuitenkin, että nyt elinkaariennusteissa käytettävään malliin sisältyvät ainakin tärkeimpien muuttujien satunnaisvaikutukset.

Suomessa syntyneiden kohdalla korvasin maassaoloajan kalenterivuodella, koska ”maassaoloaika” mittaisi ainoastaan aikaa henkilön seuraamisen aloittamisajankohdasta kulloiseenkiin hetkeen kulunutta aikaa. Koska seuraamisen aloitusajankohta on valittu satunnaisesti, maassaoloajan sisällyttäminen malliin ei toisi lisäinformaatioita nettovaikutusten ennustamiseen. Sama koskee ”maahanmuuttovuotta” eli seuranta-ajan aloittamisvuotta.

Satunnaisvaikutusten jakaumaoletukset

Malliin lisättäville satunnaisvaikutuksille on vielä määriteltävä niiden jakaumaoletukset. Olen jo edellä määrittänyt henkilötason satunnaisvakion noudattavan normaalijakaumaa, sillä testaus osoitti, että normaalijakauma toimii tässä yhtä hyvin kuin Studentin t-jakauma. Tästä poikkeuksen tekivät Suomessa syntyneet, joiden mallissa henkilömallin vakio on t-jakautunut.

Oletin, että henkilötason muut satunnaisvaikutukset: iän neliö ja maassaoloaika ovat normaalijakautuneita henkilöiden välillä. En testannut tätä oletusta, sillä en pitänyt poikkeamia oletuksesta erityisen tärkeinä selvitettäväksi, koska uskoin, että satunnaisten regressiokertoimien kohdalla poikkeavia arvoja ei ole niin paljoa kuin satunnaisen vakion kohdalla.

Halusin kuitenkin mallintaa henkilötason kolmen satunnaisvaikutuksen väliset korrelaatiot, jotta voisin selvittää mahdolliset erittäin suuret korrelaatiot näiden vaikutusten välillä sekä näiden suurten korrelaatioiden mahdolliset vaikutukset mallien konvergoitumiseen. Tällä tavalla havaitsin erittäin korkean korrelaation satunnaisvakion ja iän satunnaisvaikutuksen välillä. Samalla satunnaisvaikutusten väliset yhteydet tulevat mukaan ennusteisiin. Lopullisessa mallissa sovelsin henkilötason satunnaisvaikutusten yhteysjakaumalle kolmiulotteista multinormaalijakaumaa. Testasin myös vinon multinormaalijakauman soveltamista henkilötason satunnaisvaikutusten mallinnukseen.

Käytin myös maahanmuuttovuositason ja maassaoloaikatason satunnaisvakioiden jakaumaoletuksina normaalijakaumia. Tässä tapauksessa kaksiulotteisia normaalijakaumia, sillä satunnaisten vakioiden lisäksi malleihin sisältyy satunnainen naisindikaattori. Vaihtoehtona olisi ollut käyttää tässä yhteydessä myös Studentin t-jakaumia, mutta käytännössä tällöin vapausasteita ei olisi estimoitu aineistosta, koska havaintoja aineistossa on sekä maahanmuuttovuoden ja maassaoloajan osalta vain 24 kappaletta. Tällöin olisin kokeillut eri vapausasteita kiinnitettyinä tiettyihin arvoihin ja verrannut näin estimoitujen mallien ennustetarkkuutta.

En kuitenkaan testannut normaalijakaumaoletusta, koska tämä olisi vienyt paljon aikaa. Lisäksi maahanmuuttovuosissa ja tai maassaoloajassa ei vaikuta olevan sellaisia poikkeavia arvoja – toisin kuin henkilöiden välillä – jotka vaatisivat mallintamista varten paksuhäntäistä Studentin t-jakaumaa.

4. mallinnusaskel, osa I. Satunnaisvaikutusten lisääminen hierarkkiseen Studentin t-malliin:

$$y_{tim} \sim T(\beta_0 + \boldsymbol{X_i\beta_p} + \boldsymbol{U_i} + \boldsymbol{X_m\beta_q} + \boldsymbol{U_m} + \boldsymbol{X_t\beta_r} + \boldsymbol{U_t} + \boldsymbol{X_{tim}\beta_s} + \boldsymbol{y_{(t-u)im}\beta_u}, \sigma, \nu) \quad (4.9)$$

ja

$$\boldsymbol{U_i} \sim MN(\boldsymbol{0}, \boldsymbol{\Omega_i}) \quad (4.10)$$

ja

$$\boldsymbol{U_m} \sim MN(\boldsymbol{0}, \boldsymbol{\Omega_m}) \quad (4.11)$$

sekä

$$\boldsymbol{U_t} \sim MN(\boldsymbol{0}, \boldsymbol{\Omega_t}), \quad (4.12)$$

missä $\boldsymbol{U_i}$ on henkilötason satunnaisvaikutusten kolmiulotteinen vektori, joka noudattaa kolmiulotteista multinormaalijakaumaa (MN), jonka odotusarvovektori on $\boldsymbol{0}$ ja jonka kovarianssimatriisi on $\boldsymbol{\Omega_i}$. $\boldsymbol{U_m}$ on maahanmuuttovuositason satunnaisvaikutusten kaksiulotteinen vektori, joka noudattaa kaksiulotteista multinormaalijakaumaa, jonka odotusarvovektori on $\boldsymbol{0}$ ja jonka kovarianssimatriisi on $\boldsymbol{\Omega_m}$. $\boldsymbol{U_t}$ on maassaoloaikatason satunnaisvaikutusten kaksiulotteinen vektori, joka noudattaa kaksiulotteista multinormaalijakaumaa, jonka odotusarvovektori on $\boldsymbol{0}$ ja jonka kovarianssimatriisi on on $\boldsymbol{\Omega_t}$. Muuten parametrit ovat kuin 3. mallinnusaskeleessa edellä.

4.3.3.5 Vino Studentin t-jakauma otosjakaumana ja puolison nettovaikutukset (Malli 4, osa II)

Vino Studentin t-jakauma otosjakaumana

Maahanmuuttajaryhmäkohtaisten julkisen talouden nettovaikutusten havaintojen jakaumien tarkasteluiden perusteella niiden jakaumat eivät ole *symmetrisiä*. Tämä käy ilmi paitsi jakaumien histogrammeja tarkasteltaessa myös siitä, että useimpien jakaumien osalta otoskeskiarvo ja otosmediaani ovat etäällä toisistaan.

Koska mallissa on nyt mukana jo useita eri jakaumia, selvyyden vuoksi kutsun nyt ehdollista todennäköisyysjakaumaa, josta havainnot poimitaan, *otosjakaumaksi* (*engl. sampling distribution*). Nettovaikutusten ensimmäisenä mallinnustavoitteena oli se, että maahanmuuttajaryhmäkohtaiset mallit ovat sellaisia, että aineistot olisivat voineet olla generoituja näistä malleista. Koska tämä tavoite ei toteudu hyvin ilman jakaumien vinouksien mallintamista – eli mallien tuottamat posterioriennustejakaumat poikkeavat ”paljon” havaintoaineiston jakaumasta – päätin yleistää malleissa käytetyt otosjakaumat ei-vinoista Studentin t-jakaumista sellaiseksi Studentin t-jakaumiksi, jotka voivat olla vinoja. Tällöin malleissa on estimoitava vinon t-jakaumien vinousparametri. Otosjakaumat noudattavat vinoja t-jakaumia silloin, kun vinousparametrin arvo poikkeaa nollasta. Kun parametri on nollaa pienempi, t-jakaumat ovat vasemmalle vinoja, ja kun parametri on nolla suurempi, t-jakaumat ovat oikealle vinoja.

Aineistomallin otosjakauman vinouden estimoinnin olisi voinut sisällyttää jo toisen mallinnusaskeleeseen, jossa aineistomallin otosjakauman määritettiin noudattavan (ei-vinoa) Studentin t-jakaumaa. Vino t-jakauma ei sisälly posteriorijakaumien estimointiin käyttämäni *Stan*-ohjelman valmiisiin otosjakaumiin. Tästä syystä sellaisten mallien estimointi, jotka perustuvat vinoon t-jakaumaan, ei ole yhtä suoraviivaista

kuin edellä esitettyjen mallinnusaskelten mallien estimointi. Tämän vuoksi päätin jättää t-jakaumien vinouden estimoinnin viimeiseksi mallinnusaskeleeksi.

Käytin niiden mallien estimointiin, joiden aineistomallien otosjakaumat noudattivat vinoja t-jakaumia, menetelmää, jota voi kutsua *normaalijakaumien odotusarvon ja varianssin sekoitusjakaumaksi* (*engl. normal mean-variance mixture distribution*) (Hu and Kercheval 2006; Roncalli and Lagache 2004; Sahu et al. 2003, 136–140). Samoin kuten ei-vino t-jakauma voidaan esittää bayesilaisessa yhteydessä normaalijakaumien sekoitusjakaumana sopivilla priorijakaumavalinnoilla, tämän menetelmän avulla myös vino t-jakauma voidaan esittää normaalijakaumien sekoitusjakaumana.

Vinon Studentin t-jakauman muodostavien normaalijakautuneiden satunnaismuuttujien jakaumat voidaan esittää Hun ja Kerchevalin artikkelin (2006) merkintöjä noudattaen:

$$X|W \sim N(\mu + W\gamma, W\sigma^2) \tag{4.13}$$

ja

$$W \sim inv.gamma(\nu/2, \nu/2), \tag{4.14}$$

missä μ on t-jakauman odotusarvo ilman vinoutta, γ on vinon t-jakauman vinousparametri, W on apumuuttuja, jolla on käänteisgamma-jakauma (*engl. inverse gamma distribution*) ja ν on t-jakauman vapausasteparametri. Tällöin satunnaismuuttuja X:llä on vino t-jakauma. (Hu and Kercheval 2006, 4–5).

Vino t-jakauma voidaan siten simuloida poimimalla aluksi W:lle arvo käänteis-gammajakaumasta t-jakauman vapausasteen arvolla ν ja sitten poimimalla vinon t-jakauman simuloitu arvo esitetystä W:llä ehdollisesta normaalijakaumasta. Tämä simulaatio toistetaan tarvittava määrä eli esimerkiksi joitakin satoja tai tuhansia kertoja. Esitän tutkimuksen liitteessä B esimerkin siitä, kuinka Stan-ohjelmalla voidaan simuloida vinoja Studentin-t-jakaumia tässä esitetyllä menetelmällä.

Otosjakaumista olisi voinut tehdä vielä käytettyä jakaumaoletusta eli vinoa t-jakaumaa yleisempiä. Tällöin olisin vinon Studentin t-jakauman sijaan käyttänyt jakaumaoletuksena *laajennettua vinoa t-jakaumaa* (*engl. extended skew-t distribution*) (Arellano-Valle and Genton 2010). Posterioriennustetarkistusten perusteella (ks. jäljempänä *havainnot, estimointitulokset ja posterioriennusteet*) käytetty jakaumaoletus aineistomallin otosjakaumalle (vino Studentin t-jakauma) täytti kuitenkin mallinnustavoitteet.

Kokeilin vinoa jakaumaa myös henkilövaikutusten mallin osalta. Tässä tapauksessa jakaumaoletuksen muutos tarkoitti siirtymistä (ei-vinosta) multinormaalijakaumasta vinoon multinormaalijakaumaan. Noudatin estimoinnin osalta menetelmää, jonka ovat esittäneet esimerkiksi Arellano-Valle et al. (2007) ja edelliseen viitaten Lin (2009).

Myös tässä tapauksessa vinon jakauman mallinnus perustuu ei-vinojen (multi)normaalijakaumien sekajakaumaan. Apumuuttujana on (Linin (2009) merkintöjä noudattaen) p-ulotteinen τ_p, joka noudattaa *p-ulotteista standardia puolinormaalijakaumaa* (*engl. p-dimension standard half-normal distribution*): $\tau_p \sim HN_p(0, 1)$. Vino multinormaalijakauma saadaan tällöin multinormaalijakaumasta: $N_p(\xi + \lambda\tau_j, \Omega)$, missä ξ on p-ulotteinen odotusarvovektori ei-vinolle multinormaalijakaumalle, λ on p-ulotteinen vinousparametrien (diagonaali)matriisi, τ_p on p-ulotteinen apumuuttuja ja Ω on multinormaalijakauman kovarianssimatriisi. (Lin 2009, 258 ja 260).

En kuitenkaan käyttänyt lopullisissa malleissa henkilötason satunnaisvaikutusten mallintamiseen vinoa multinormaalijakaumaa. Syynä tähän oli ensisijaisesti se, että henkilöiden nettovaikutusten keskiarvojen jakaumat vaikuttivat tarkasteluissa symmetrisiltä ja siten ei-vinoilta. Toiseksi, omien simulaatiokokeilujeni perusteella, mikäli sekä otosjakauma että henkilöjakauma olivat vinoja, vinous keskittyi otosjakaumaan. Kolmantena syynä olivat vaikeudet saada sellaisia malleja konvergoitumaan Stan-ohjelmalla, jossa oli mallinnettavia (moni)ulotteisia vinouksia sekä havainto- että henkilötasolla.

Puolison nettovaikutukset

Kun laadin nettovaikutusmalleja, eräänä hypoteesina oli, että puolison nettovaikutukset voisivat vaikuttaa henkilön julkisen talouden nettovaikutuksiin. Koska en ollut varma, kuinka huomioida tämä mahdollinen vaikutus, kun laadin mallinnusaskelta, jossa mukaan lisättiin selittävät muuttujat, päätin ratkaista asian vasta mallinnusvaiheen lopussa.

Puolison julkisen talouden nettovaikutukset tarkoittavat tässä yhteydessä mahdollisen puolison nettovaikutuksia sinä vuonna, kun henkilö muutti ensimmäisen kerran Suomeen. Tiedon olemassaolo henkilölle edellyttää kolmen ehdon täyttymistä: henkilöllä oli oltava puoliso, kun henkilö muutti Suomeen ensimmäisen kerran, henkilön ensimmäisen muuton on täytynyt tapahtua vuosina 1995–2011 (vuotta 1995 aiemmille vuosille ei ole nettovaikutustietoja) sekä puolison on täytynyt syntyä ulkomailla. Viimeinen ehto johtuu siitä, että tutkimusaineisto sisältää kaikki ulkomailla syntyneet henkilöt, kun he täyttävät aiemmin tutkimuksessa mainitut väljät ehdot, mutta Suomessa syntyneistä työikäisistä aineistossa on vain noin yksi sadasta henkilöstä.

Tutkin maahanmuuttajaryhmäkohtaisesti korrelaatiokertoimia puolison nettovaikutusten ja selitettävän muuttujan välillä niiden henkilöiden osalta, joille *puolison nettovaikutus* -tieto löytyi. Maahanmuuttajaryhmäkohtaiset korrelaatiokertoimet olivat pieniä, suurin osa välillä $[-0{,}05; +0{,}05]$. Myös Irakissa ja Somaliassa syntyneiden osalta puolison nettovaikutuksilla maahanmuuttovuonna vaikutti olevan heikko lineaarinen yhteys henkilön nettovaikutuksiin, joten jätin sekä Irakissa ja Somaliassa syntyneiden malleista pois puolisoiden nettovaikutustiedon muuttovuonna. Sen lisääminen olisi ollut työläst, koska puolison nettovaikutus olisi pitänyt käytännössä jakaa luokiksi, jotta se olisi voitu mallintaa yhdessä sellaisten henkilöiden kanssa, joilla ei ollut puolisoa muuttovuonna.

4. mallinnusaskel, osa II. Hierarkkinen vino studentin t-malli, jossa satunnaisvaikutuksia:

$$y_{tim} \sim ST(\beta_0 + \boldsymbol{X_i\beta_p} + \boldsymbol{U_i} + \boldsymbol{X_m\beta_q} + \boldsymbol{U_m} + \boldsymbol{X_t\beta_r} + \boldsymbol{U_t} + \boldsymbol{X_{tim}\beta_s} + \boldsymbol{y_{(t-u)im}\beta_u}, \sigma, \nu, \gamma) \tag{4.15}$$

ja

$$\boldsymbol{U_i} \sim MN(\boldsymbol{0}, \boldsymbol{\Omega_i}) \tag{4.16}$$

ja

$$\boldsymbol{U_m} \sim MN(\boldsymbol{0}, \boldsymbol{\Omega_m}) \tag{4.17}$$

sekä

$$\boldsymbol{U_t} \sim MN(\boldsymbol{0}, \boldsymbol{\Omega_t}), \tag{4.18}$$

missä y_{tim}:llä on vino t-jakauma (*engl. skew t-distribution, ST*), jonka vinousparametri on γ. Muuten parametreilla on samat selitteet kuin 4. mallinnusaskeleessa (osa I) edellä.

4.3.3.6 Otosjakauman vinouden mallinnus (Malli 5)

Yllä (4. mallinnusaskel) vaihdoin aineistomallin otosjakauman tavallisesta Studentin t-jakaumasta vinoon Studentin t-jakaumaan. Aineistomalliin tuli näin estimoitavaksi yksi uusi parametri, vinon Studentin t-jakauman vinousparametri γ. Tällaisessa aineistomallissa oletetaan, että kaikki havainnot generoituvat prosessista, jonka todennäköisyysjakauman – eli tässä tapauksessa vinon Studentin t-jakauman – vinous on kaikkien havaintojen kohdalla sama.

Tämä oletus ei kuitenkaan välttämättä pidä paikkaansa, sillä jakauman vinous voi mahdollisesti vaihdella ehdollisesti eri havaintojen kohdalla. ”Ehdollisuus” tarkoittaa tässä sitä, että vinous voi vaihdella sen mukaan, mitä poimittavasta havainnosta tiedetään ennen sen toteutumista. Esimerkiksi, jos tiedetään, että nettovaikutushavainto koskee 20 vuotta maassa ollutta maahanmuuttajaa, niin tämän havainnon vinous voi poiketa havainnosta, joka koskee alle 5 vuotta Suomessa ollut maahanmuuttajaa. Toisin sanoen vinon Studentin t-jakauman vinous voi vaihdella sen mukaan, kuinka kauan maahanmuuttaja on ollut Suomessa. Tämä vinouden ehdollisuus voidaan yleistää koskemaan muitakin tekijöitä kuin maassaoloaikaa. Näin jakauman vinoudelle voidaan luoda oma lineaarinen malli, jossa vinous riippuu käytännössä tutkimusaineistoon sisältyvistä selittävistä muuttujista.

Määritetään nyt otosjakauman ehdollinen vinous vinouden määräävien muuttujien $\boldsymbol{X_{vinous}}$ ja vinouden ”regressio”kertoimien $\boldsymbol{\beta_{vinous}}$ lineaarikombinaationa:

$$\begin{aligned} \text{Vinous} &:= \boldsymbol{X_{vinous}\beta_{vinous}} \\ &= \beta_{vinous,0} + X_{vinous,1}\beta_{vinous,1} + \cdots + X_{vinous,n}\beta_{vinous,n}. \end{aligned} \tag{4.19}$$

Tässä määritelmässä vinouden vakioparametri $\beta_{vinous,0} = \gamma$, eli se vastaa 4. mallinnusaskeleessa esitettyä

jakauman vinousparametria. Mikäli kaikki vinoutta selittävät muuttujat $\boldsymbol{X_{vinous}}$ saavat arvoksi nollan, tällöin Vinous $= \gamma$, eli saadaan 4. mallinnusaskeleessa esitetty otosjakauma.

Kun näin määritetty ehdollinen vinous lisätään otosjakaumaan vinousparametrin γ tilalle, saadaan:

5. mallinnusaskel. Otosjakauman vinouden mallinnus:

$$y_{tim} \sim ST(\beta_0 + \boldsymbol{X_i\beta_p} + \boldsymbol{U_i} + \boldsymbol{X_m\beta_q} + \boldsymbol{U_m} + \boldsymbol{X_t\beta_r} + \boldsymbol{U_t} + \boldsymbol{X_{tim}\beta_s} + \boldsymbol{y_{(t-u)im}\beta_u}, \sigma, \nu, \text{Vinous}) \tag{4.20}$$

ja

$$\text{Vinous} = \boldsymbol{X_{vinous}\beta_{vinous}}. \tag{4.21}$$

Tässä X_{vinous} on niiden tekijöiden vektori, jotka määräävät havainnon y_{tim} otosjakauman vinouden. Muuten parametreilla on samat selitteet, kuin edellä on esitetty aiempien mallinnusaskelten yhteydessä.

Otosjakauman vinouden määräävinä tekijöinä X_{vinous} käytän vain kaikkein tärkeimpiä selittäviä muuttujia nettovaikutusmallinnuksessa eli aikaan liittyviä selittäviä tekijöitä ja sukupuolta. Tällä haluan varmistaa, että ehdollisen vinouden malli pysyy yksinkertaisena ja siten kohtalaisen helposti ymmärrettävänä.

4.3.3.7 Potentiaaliset aineistomallien jatkokehittämiset

Edellä esitettyä lopullista aineistomallia (5. mallinnusaskel) voisi edelleen kehittää. Esitän seuraavaksi joitakin potentiaalisia kehitysmahdollisuuksia lopulliseen käyttämääni aineistomalliin. En sisällyttänyt seuraavia kohtia lopulliseen malliin, koska ne olisivat nähdäkseni vieneet liikaa aikaa siihen nähden, miten ne olisivat parantaneet mallinnustavoitteiden täyttämistä.

1. Lopulliseen aineistomalliin olisi voinut lisätä vielä kaikille havainnoille yhteisiä interaktiovaikutuksia. Näin olisi voinut tehdä esimerkiksi aikamuuttujien ja pääasiallisen toiminnan (työllinen, opiskelija tai muu) välille. Kenties aikamuuttujien vaikutus vaihtelee sen mukaan, mikä henkilön pääasiallinen toiminta on ollut ensimmäisen muuton muuttovuonna. Halusin pitää parametrien määrän kuitenkin muutamassa kymmenessä, joten jätin pois muun muassa tällaisia mahdollisia interaktiovaikutuksia, jotka eivät vaikuttaneet mielestäni merkittäviltä tai suurilta.

2. Posterioriennustetarkistukset eli mallien sopivuuden tarkistukset tämän luvun lopussa osoittavat, että erityisesti sukupuolen ja ikäluokan määritetyt nettovaikutusten malleista luodut posterioriennusteiden keskiarvot ja mediaanit eivät vastaa kovin hyvin havaintoaineiston vastaavia keskiarvoja ja mediaaneita. Yritin korjata tätä tilannetta lisäämällä interaktiotermejä, jotka koostuvat ikäluokista ja sukupuolesta. Tämä ei kuitenkaan parantanut tilannetta juurikaan. Kenties aineistomallin otosjakaumana käytetyn vinon t-jakauman vinous pitäisi mallintaa ehdollisesti käyttämällä ikäluokkia, sukupuolta sekä niiden välisiä interaktioita.

3. Henkilöiden ajassa perättäisten havaintojen muodostaman autokorrelaation (*engl. serial correlation*) tarkempi mallinnus. Käyttämissäni lopullisissa maahanmuuttajaryhmäkohtaisissa malleissa autokorre-

laatioiden oletetaan olevan samansuuruisia kaikille henkilöille riippumatta esimerkiksi henkilöiden ikävuosista. Tämä oletus on melko epärealistinen, ja autokorrelaatioiden suuruuden voisi ehdollistaa eri selittävillä muuttujilla. Autokorrelaatioiden voisi olettaa olevan myös lähempänä nollaa silloin, kun kyse on ääriarvoista, sillä useampi ääriarvo ei vaikuttaisi seuraavaan toistaan aineistoissa. En kuitenkaan mallintanut autokorrelaatiota tarkemmin lopullisessa mallissa, koska autokorrelaatioiden tarkempi rakenne ei ole erityisen mielenkiinnon kohteena tutkimuksessa.

4. Maahanmuuttajaryhmäkohtaiset parannukset aineistomalleissa. Käytännössä havaintoaineistot ovat jokaisen syntymävaltiokohtaisen ryhmän osalta keskittyneet erityisesti enemmän nollan ympäristöön, kuin lopullisen mallin posteriorijakauma on. Tämä on nähdäkseni suurin ero mallien tuottamien posterioriennustejakaumien ja havaintoaineistojen jakaumien välillä. Tämä keskittyminen vaihtelee jonkin verran ryhmittäin. Nollan ympäristö tarkoittaa tässä sitä, että julkisen talouden nettovaikutukset ovat ”lähellä” nollaa euroa. Lähellä tarkoittaa noin paria sataa euroa nollan kummallakin puolella. Keskittymät nollan kohdalla eivät ole kuitenkaan niin merkittäviä tulosten kannalta, että olisin mallintanut niitä erikseen esimerkiksi käyttäen latentteja sekoitusmalleja. Mallien posterioriennustetarkistukset osoittavat (ks.*Havainnot, estimointitulokset ja posterioriennusteet*), etteivät mallinnustavoitteet kärsi siitä, etten mallintanut erikseen nollan ympäristöön sijoittuvia havaintoja.

5. Oletan havaintokohtaisten virhetermien olevan ehdollisesti samoin jakautuneita kaikkien havaintojen kohdalla. Tätä oletusta voisi tutkia mallien tuottamien posterioriennusteiden studentisoituja residuaaleja tarkastelemalla eri selittävien muuttujien arvojen mukaan.

4.3.4 Estimoitavien parametrien priorijakaumat

Koska mallinnus tapahtuu bayesilaisen päättelyn piirissä, kaikille edellä esitettyjen aineistomallien estimoitaville parametreille oli määritettävä priorijakaumat. Kuten muutenkin koko tutkimuksessa, käytin priorijakaumia lähinnä hyödyllisinä mallinnusvälineinä, enkä niinkään välineinä tutkimusaineiston ulkopuolisen informaation sisällyttämisenä parametrien posteriorijakaumaan. Noudattamani ”priorifilosofia” perustuu yleisesti Gelmanin ja Shalizin (2013) esittämään ajatteluun. Käytin yleisesti ottaen tutkimuksessa ”melko epäinformatiivisia jakaumia” tai ”heikosti informatiivisia jakaumia” (*engl. weakly informative distribution*).

Heikosti informatiivinen jakauma on sellainen, jossa parametrin todennäköisyystiheys ei ole toisaalta keskittynyt vain ”kapealle” osalle sen määrittelyjoukkoa, mutta toisaalta todennäköisyystiheys on keskittynyt niin, että ”käytännössä mahdottomilla” osilla sen määrittelyjoukkoa todennäköisyystiheys on nolla tai ainakin hyvin pieni. Käytännössä mahdoton määrittelyjoukon osa tarkoittaa selittävien muuttujien regressiokertoimien osalta esimerkiksi sitä, että selittävän muuttujan ”pieni” muutos vaikuttaisi julkisen talouden nettovaikutusten odotusarvoon miljoonien eurojen verran.

Keskeisin motivaatio heikosti informatiivisten priorijakaumien käyttöön oli, että niiden soveltaminen nopeutti posteriorijakaumien simulointia Stan-ohjelmalla verrattuna epäinformatiivisten priorijakaumien käyttöön (Stan Development Team 2015, 53). Lisäksi monet usein käytetyt epäinformatiiviset priorijakaumat ovat epäsopivia todennäköisyysjakaumia, mikä voi johtaa epäsopiviin posteriorijakaumiin.

Kaikki käyttämäni priorijakaumat olivat sopivia todennäköisyysjakaumia (*engl. proper probability distribution*), mikä varmisti sen, että kaikki maahanmuuttajaryhmäkohtaiset posteriorijakaumat olivat myös sopivia todennäköisyysjakaumia. Lisäksi kaikki käyttämäni priorijakaumat olivat keskenään riippumattomia, koska en nähnyt tarvetta käyttää jakaumia, jotka eivät olisi olleet riippumattomia.

Koska hierarkkisissa malleissa otosjakauman ja priorijakaumien raja on häilyvä, määritin priorijakaumat kaikille estimoitaville parametreille riippumatta siitä, onko parametrin jakauma ehdollinen muilla mallin parametreilla vai ei.

Alla on esitetty lopullisen mallin parametrien käytetyt priorijakaumat. Priorijakaumien yhteisjakauma on näiden alla esitettyjen yksiulotteisten jakaumien tulo, koska yksittäiset jakaumat ovat toisistaan riippumattomia. Regressiokertoimien priorijakauma on myös heikosti informatiivinen, koska käytin sekä selitettävän että selittävien muuttujien osalta standardoituja jakaumia, jolloin vaikutukset ovat pieniä. Standardointi tarkoittaa sitä, että jokaisen muuttujan kohdalla muuttujan arvosta vähennetään aluksi muuttujan (otos)keskiarvo ja tämä erotus jaetaan muuttujan (otos)keskihajonnalla. Standardointi sekä nopeuttaa mallien konvergoitumista niiden estimointiin käytetyssä Stan-ohjelmassa että helpottaa selittävien muuttujien regressiokertoimien priorijakaumien määrittelyä, sillä niiden vaikutukset ovat karkeasti ottaen samassa mittakaavassa. Tästä syystä pystyin määrittämään kaikille regressiokertoimille saman priorijakauman: standardinormaalijakauman.

Parametrien selitykset löytyvät ylhäältä lopullisesta aineistomallista ja sen selostuksesta sekä myös jäljempää mallien parametrien selitetaulukosta (s. 170).

$\beta_0 \sim$ *tasajakauma*$(-3, 3)$, missä β_0 on mallin vakiotermi.

$\beta \sim N(0, 1)$, missä β on mallin regressiokerroin.

$\beta_u \sim \mathcal{U}(-1, 1)$, missä β_u on autoregressiivinen kerroin, jonka viipeen arvo on u.

$\beta_{vinous} \sim N(0, 1)$, missä β_{vinous} on mallinnusaskeleessa 5 vinouden mallintamisen regressiokerroin.

$\sigma \sim$ *cauchy*$(0, 5)$, missä σ on mallin virhetermin keskihajonta.

$\tau_i \sim$ *cauchy*$(0, 5)$, missä τ_i on henkilötason satunnaisvaikutuksen keskihajonta.

$L_1 \sim$ *LKJ*(1), missä L_1 on henkilötason satunnaisvaikutusten välisen korrelaatiomatriisin Cholesky-hajotelma.

$\tau_m \sim$ *cauchy*$(0, 5)$, missä τ_m on maahanmuuttovuositason satunnaisvaikutuksen keskihajonta.

$L_2 \sim$ *LKJ*(1), missä L_2 on maahanmuuttovuositason satunnaisvaikutusten välisen korrelaatiomatriisin Cholesky-hajotelma.

$\tau_t \sim$ *cauchy*$(0, 5)$, missä τ_m on maassaoloajan tason satunnaisvaikutuksen keskihajonta.

$L_3 \sim$ *LKJ*(1), missä L_3 on maassaoloajan tason satunnaisvaikutusten välisen korrelaatiomatriisin Cholesky-hajotelma.

$\nu \sim$ *cauchy*$(0, 10)$, missä ν on t-jakautuneen otosjakauman vapausasteparametri.

$\gamma \sim N(0, 1)$, missä γ on vinon t-jakautuneen otosjakauman vinousparametri. Kaikilla vinoutta selittävien muuttujien kertoimilla on sama priorijakauma.

$W_{tim} \sim$ *inv.gamma*$(\nu/2, \nu/2)$, missä W_{tim} on havaintokohtainen apumuuttuja vinon t-jakauman simuloinnissa.

4.3.5 Estimoitavien parametrien yhteisposteriorijakauma

Olen edellä määrittänyt viisi eri aineistomallia sekä näiden mallien parametreille priorijakaumat. Mallikohtaiset normeeraamattomat posteriorijakaumat saadaan aineistomallien ja priorijakaumien tulona:

posteriorijakauma $\propto$ *aineistomalli* $\times$ *priorijakaumat*.

Eli noudattaen tässä tutkimuksessa käytettyjä merkintätapoja:

$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, missä $\boldsymbol{\theta}$ on mallin estimoitavien parametrien vektori ja $\boldsymbol{y}$ on havaintoaineisto.

Parametreja koskevaa bayesilaista tilastollista päättelyä varten on selvitettävä posteriorijakauma $p(\boldsymbol{\theta}|\boldsymbol{y})$.

4.3.6 Posteriorijakauman simulointi ja konvergoituminen tavoitejakaumaan

4.3.6.1 Posteriorijakauman simulointi Stan-ohjelmalla

Olen edellä esittänyt kaikkien estimoitavien parametrien yhteisen normeerattoman posteriorijakauman. Parametrien estimointia varten tämä normeeraamaton posteriorijakauma on esitettävä joko analyyttisesti (eli jonain tunnettuna todennäköisyysjakaumana), tai se on simuloitava poimimalla posteriorijakaumasta otoksia. Koska estimoitavat posteriorijakaumat eivät noudata mitään tunnettua todennäköisyysjakaumaa, posteriorijakaumat on simuloitava. Koska posteriorijakaumista ei voi poimia suoraan otoksia (*Monte Carlo -simulointi*) on käytettävä *Markov-ketjuihin* perustuvia menetelmiä (*engl. Markov Chain Monte Carlo, MCMC*), kuten aiemmin käytin olotilamallinnuksen yhteydessä.

Käytin posteriorijakaumien simulointiin *Stan*-ohjelman (versiot 2.7.0 ja 2.9.0) R-ohjelmaan sopivaa versiota *RStania* (Goodrich et al. 2015). Valitsin Rstanin posteriorisimulointiin kolmesta syystä:

1. Stanin simulointimenetelmä on tehokas.
2. Stanin simulointimenetelmä johtaa useammin konvergenssiin kuin monet muut menetelmät.
3. Stanilla on ”helppoa” simuloida monimutkaisiakin posteriorijakaumia.

Lisäksi Stan on yhdysvaltalaisen Columbia-yliopiston tilastotieteen professorin Andrew Gelmanin johtaman ryhmän kehittämä (Stan Development Team 2015, iii). Gelman on myös toimittanut *Bayesian Data Analysis* -kirjan, jonka lähestymistapaa bayesilaiseen päättelyyn käytän pääasiassa tässä tutkimuksessa, joten Stan-ohjelman käyttö on myös tässä mielessä sopiva valinta. Kirjassa myös viitataan useasti Stanin käyttöön estimoinnissa.

Stan on varsin uusi ohjelma, jonka ensimmäinen versio on julkaistu vuonna 2012. Se vaikuttaa hyvin kehittyneeltä ohjelmalta bayesilaiseen posteriorisimulointiin, mikä johtuu sen simulointimenetelmän tehokkuudesta ja helppokäyttöisyydestä.

Stanin simulaatiomenetelmä tehokkaampi ja johtaa useammin konvergenssiin kuin yleisemmät MCMC-menetelmät

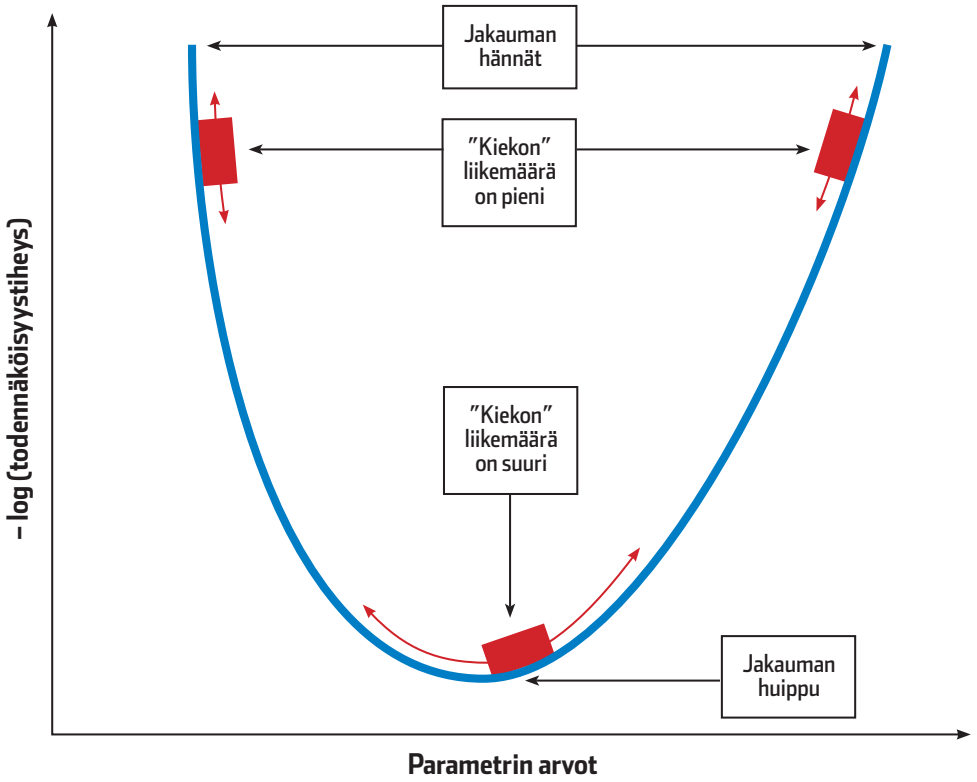
Stan-ohjelman käyttämä posteriorijakaumien estimointimenetelmä perustuu nk. *hamiltonilaiseen Monte Carloon* (*HMC*), jota kutsutaan myös *hybridiksi Monte Carloksi* (*engl. Hamiltonian/Hybrid Monte Carlo, HMC*). Laajemman johdannon HMC:hen antaa Neal (2011), joka sovelsi fysiikassa kehitettyä menetelmää (Duane et al. 1987) ensi kertaa tilastotieteellisiin ongelmiin (Neal 1993). Lyhyemmän johdannon HMC:hen, sekä sen käyttöön Stan-ohjelmassa, antavat Gelman et al. (2014, 300–308) ja Stan-opas (Stan Development Team 2015, 481–488). Kaikki HMC:sta tässä tutkimuksessa esitettävä pohjautuu näihin lähteisiin.

HMC perustuu *hamiltonilaiseen dynamiikkaan*, millä tässä yhteydessä kuvataan systeemin liikemäärän ja sijainnin muutosta ajassa. Systeemin tilaa kuvaa funktio, jonka argumentteja ovat *systeemin liikemäärä* ja *sijainti*, ja jota kutsutaan suomeksi *Hamiltonin funktioksi* (Koskinen and Vainio 2010, 139) (*engl. Hamiltonian*). Systeemin liikemäärän ja sijainnin kehitystä kuvaavia yhtälöitä kutsutaan *Hamiltonin liikeyhtälöiksi* (mt, 140). Systeemin sijainti vastaa estimoitavien parametrien arvoja, ja se on siten d-ulotteinen, missä d on estimoitavien parametrien määrä. Systeemin liikemäärä on myös d-ulotteinen, ja se kuvaa, kuinka systeemi liikkuu ajassa.

HMC:ssa on tarkoitus siten selvittää (simuloida) estimoitavien parametrien ja niitä vastaavien liikemääräparametrien yhteisposteriorijakauma. Estimoitavien parametrien arvot (systeemin paikkavektori) ja liikemäärävektori ovat HMC:ssa yhteydessä toisiinsa: mitä suurempi on estimoitavien parametrien posteriorijakauman log-todennäköisyystiheys parametrien sijainnissa tietyllä hetkellä, sitä nopeammin systeemi liikkuu sillä hetkellä, eli sitä suurempi on sen liikemäärä (ja liike-energia). Mitä pienemmäksi log-todennäköisyystiheys muuttuu liikuttaessa systeemissä, sen pienemmäksi myös liikemäärä (ja liike-energia) muuttuvat. Kun systeemin kineettinen energia pienentyy, sen potentiaalienergia puolestaan kasvaa.

Mikäli estimoitavia parametreja on vain yksi, ja sillä on rajoittamaton määrittelyjoukko sekä yksihuippuinen jakauma, systeemin kehitystä ajassa voisi kuvata mukaillen Nealin (2011) esimerkkiä (s. 114). Tässä esimerkissä – joka on kuvattu alla kuvassa 4.2 – on kitkaton kiekko, joka liikkuu äärettömän syvässä kuopassa sen sivulta toiselle. Kun kiekko on kuopan pohjalla, sen vauhti (ja liike-energia) on suurimmillaan. Kun kiekko lähtee nousemaan jompaakumpaa sen reunoista, sen vauhti lopulta loppuu ja systeemin liike-energia muuntuu kokonaisuudessaan potentiaalienergiaksi. Tällöin kiekko lähtee liukumaan toiseen suuntaan, jne. Kiekko ei tietysti kulje vain deterministisesti kuopan yhdeltä reunalta kuopan toiselle

reunalle, sillä kiekon liikemäärän (ja siten liike-energian) muutokseen vaikuttaa kiekon sijainnin lisäksi satunnainen tekijä, jonka skaala riippuu d-ulotteisesta massamatriisista. Massamatriisi on liikemäärän kovarianssimatriisi, kun sen oletetaan noudattavan d-ulotteista normaalijakaumaa.



Kuva 4.2: *Havainnekuva hamiltonilaiseen Monte Carloon (HMC) perustuvasta yksihuippuisen todennäköisyysjakauman simuloinnista. Jakauman log-todennäköisyystiheyttä kartoittavan "kiekon"liikemäärä on suuri, kun log-todennäköisyystiheys on suuri jakauman huipun läheisyydessä. Liikemäärä on puolestaan pieni, kun log-todennäköisyystiheys on pieni jakauman hännissä. Kuvassa log-todennäköisyystiheyden asteikko on käänteinen havainnollistamisen vuoksi.*

Koska kuoppa vastaa tässä simuloitavan parametrin todennäköisyysjakaumaa, tämä HMC:n ominaisuus takaa sen, että systeemi käy nopeasti parametrin todennäköisyysjakauman eri osissa, eli se tulee simuloitua tehokkaasti. Tämä ominaisuus erottaa HMC:n edukseen muista tavanomaisimmista posteriorioriorijakaumien simulaatiomenetelmistä kuten *Gibbsin otannasta*, jota käytin maassaolomallinnuksen yhteydessä, ja satunnaiskulkuun perustuvasta *Metropolis-Hastings-simuloinnista*. Mikäli käytetään edellistä kuoppavertausta, jälkimmäisten menetelmien tehottomuus verrattuna HMC:hen tarkoittaa sitä, että niissä kiekko jää pitkäksi aikaa vain tiettyyn kuopan alueeseen. Toisin sanoen näiden menetelmien tuottamien perättäisten simulaatioiden (kiekon sijainti perättäisinä ajanhetkinä) välillä on suuri autokorrelaatio.

Lisäksi Rstan muuntaa R-ohjelmassa annetun Stan-ohjelman koodin aluksi *C++* -ohjelmointikielelle, mi-

kä osaltaan nopeuttaa simulaatioita verrattuna siihen, että käytettäisiin suoraan R-ohjelman omaa koodia.

Stan-ohjelmalla helppoa simuloida erilaisia posteriorijakaumia

Edellä mainitun systeemin Hamiltonilaisen funktion kehitys on määritetty jatkuvana ajassa. Mikäli pidättäydytään mainitussa kiekkoesimerkissä, niin sen kulkua seurataan kuitenkin ei-jatkuvin eli diskreetein askelein. Kirjallisuudessa tätä *askelkokoa* (*engl. step size*) merkataan kreikkalaisella ϵ-kirjaimella, ja sen suuruus pitää määrittää.

Hamiltonilaisen systeemin ei-jatkuvaan approksimointiin HMC käyttää nk. *leapfrog*-menetelmää (Neal 2011, 121; Gelman et al. 2014, 301). Siinä jokaiseen leapfrog-askeleeseen kuuluu kolme vaihetta, joissa systeemin liikemäärä päivitetään kaksi kertaa ja sen sijainti kerran. Kirjallisuudessa leapfrog-askeleiden määrää merkataan L-kirjaimella.

HMC:ssa yhden simulaation tuottaminen estimoitavasta posterioriorijakaumasta sisältää L kappaletta leapfrog-hyppyjä. Yhden leapfrog-hypyn koko riippuu puolestaan askelkoosta (ϵ) ja liikemäärän massamatriisista. Kun leapfrog-hyppyjä on tehty L kappaletta, tehdään hyväksy/hylkää-testi (nk. *Metropolis-askel*. Siinä verrataan liikemäärän ja sijainnin *lähtöarvojen* posteriorijakaumaa (eli systeemin todennäköisyyttä lähtöarvoilla) ja leapfrog-hyppyjen *jälkeen* saadun uuden liikemäärän ja sijainnin posteriorijakaumaa (eli systeemin todennäköisyyttä uuden sijainnin arvoilla).

Jotta HMC toimisi tehokkaasti, sekä L:n, ϵ:n että massamatriisin on oltava sopivan suuruisia. Jos L on esimerkiksi liian pieni, edellä mainitun esimerkin ”kiekko” liikkuu hyvin hitaasti kuopassa ja HMC vastaa lähinnä satunnaiskulkuun perustuvaa hidasta Metropolis-Hastings-simulaatiota. Mikäli L on taas liian suuri, ”kiekko” kulkee (turhaan) edestakaisin kuopan laidasta toiseen jokaisella simulaatiokerralla ja tuhlaa näin laskenta-aikaa. Koska HMC:n käyttäjän on varsin työlästä säätää sekä L:n, epsilonin että massamatriisin koot oikeanlaisiksi, Staniin on kehitetty sellaiset menetelmät, jotka hoitavat näiden arvojen säätämisen simuloinnin aikana. Massamatriisi ja ϵ määräytyvät simulaatioiden lämmittelyn (warm-up) aikana ja L puolestaan jokaisen simulaatiokerran yhteydessä (warm-up ja varsinaiset simulaatiot).

Hoffman and Gelman (2014) esittävät nk. *NUTS-otannan* (*No-U-turn sampler*) toimintaperiaatteet, jolla L:n koko määräytyy jokaisen simulaation yhteydessä ilman, että käyttäjän pitää määrätä sen koko (ks. myös Gelman et al. 2014, 304 ja 308;Stan Development Team 2015, 489–490). Nimensä mukaisesti NUTS ei salli sitä, että ”kiekko” pysähtyy kuopan jommallekummalle reunalle ja kääntyy takaisin. L on siten korkeintaan niin suuri, että kiekon takaisinkääntyminen tapahtuisi (vaihtoehtoisesti L saavuttaa sille annetun maksimiarvon): koko kuoppa tulee näin kartoitettua nopeasti, kun sitä ei sahata turhaan edestakaisin.

Hoffman and Gelman (2014) esittävät myös menetelmän, jolla askelkoko ϵ saadaan optimoitua. Sen koko riippuu mm. määritetystä tavoitehyväksymisprosentista uudelle liikemäärän ja sijainnin yhteisposteriorijakaumalle. Stanissa oletusarvo tälle on $0{,}80$, kun käytetään NUTS:ia. Stan jakaa posterioriorisimulaatioiden ketjujen lämmittelyvaiheet (warm-up) kolmeen vaiheeseen, jonka toisen vaiheen aikana Stan etsii liikemäärän massamatriisille sopivan arvon varsinaisia simulaatioita varten.

Koska Stan etsii automaattisesti posteriorisimulaatioon tarvittavat säätöparametrit, sitä on helpompi käyttää kuin sellaisia menetelmiä, joissa tämä jää käyttäjän tehtäväksi. Käyttäjä voi halutessaan kiinnittää näi-

den säätöparametrien että niiden sopivien arvojen etsintään liittyvien apuparametrien arvot itse. Käytin kuitenkin Stanin oletusarvoja. Joissakin posteriorisimulaatioissa tosin nostin L:n maksimimäärä 10:stä 11:een ja tavoitehyväksymisprosentin uusille simuloiduille arvoille 0,80:sta 0,90:ään, jotta posteriorijakauma tulisi kartoitettua nopeammin.

Käyttäjän määritettäväksi jää näin vain estimoitavien parametrien yhteisen posteriorijakauman todennäköisyystiheysfunktio sekä tämän tiheysfunktion logaritmin gradientti, jota tarvitaan päivittämään liikemäärä leapfrog-hypyissä. Mikäli posteriorijakauma noudattaa jotain tunnettua jakaumaa, tai se voidaan muodostaa niiden avulla, nämäkin löytyvät jo valmiiksi Stanista. Asia on näin kaikkien Stanilla estimoitavien posteriorijakaumien kohdalla tässä tutkimuksessa. Posteriorijakaumien estimointi Stanilla on täten hyvin pitkälle automatisoitua.

Stanin käyttö vaatii kuitenkin Stanin oman koodikielen opettelemista. R-ohjelman yhteyteen on kuitenkin tehty paketteja, jotka kirjoittavat Stan-koodia. Käytin tähän tarkoitukseen *BRMS*-pakettia (*Bayesian Regression Models using Stan*, Buerkner 2015). Siinä mallikoodi kirjoitetaan R-ohjelman käyttäjille kenties tunnetummassa muodossa, jota noudattaa esimerkiksi lineaaristen sekamallien estimointiin tarkoitettu *lme*-paketti. Koska BRMS-paketti sisältää vain tietyt malliluokat, lopullisen mallin posteriorijakauman kirjoitin suoraan Stan-koodina. Joka tapauksessa BRMS-paketti toimi hyvänä johdantona Staniin, koska sen kirjoittaman Stan-koodin saa tulostettua.

4.3.6.2 Posteriorisimulaatiot: simulaatioketjuja koskevat valinnat

Posteriorijakaumien simulointeja varten oli vielä määritettävä simulaatioketjujen määrä, simulaatioiden kokonaismäärä yhdessä ketjussa sekä lämmittelyyn käytettävien simulaatioiden määrä yhdessä ketjussa. Käytin useimpien posteriorisimulaatioiden kohdalla kahta ketjua, joissa kummassakin simulaatioiden kokonaismäärä oli 3 000 simulaatiota ja joissa lämmittelyn osuus oli 1 000 ensimmäistä simulaatiota. Posterioripäättelyihin käytettäviä simulaatioita poimin jokaisesta posteriorijakaumasta 2×2 000 simulaatiota = 4 000 simulaatiota yhteensä.

Suhteellisen pieni simulaatioiden määrä on riittävä ketjujen konvergenssia ja Monte Carlo -keskivirheiden pienuutta varten, sillä Stanin edellä esitetty estimointimenetelmä simuloi tehokkaasti posteriorijakaumia. Ketjuja olisi voinut olla enemmänkin kuin kaksi, jotta ketjujen konvergoitumisesta olisi voinut olla vielä varmempi, mutta käytin vain kahta, sillä tämä säästi mallien estimointiin käytettävää aikaa. Joka tapauksessa estimointien konvergenssiin viittaavat niin konvergenssitestit ($\hat{R}$), tulosten järkevyys kuin posterioriennustetarkistukset, joissa estimoitujen posteriorijakaumien pohjalta laadittujen ennusteiden jakaumat vastaavat likimain alkuperäisten havaintoaineistojen jakaumia.

Stan poimii automaattisesti satunnaiset alkuarvot kummankin simulaatioketjun kaikille estimoitaville parametreille. Esimerkiksi regressiokertoimien kohdalla kummankin ketjun jokainen estimoitava regressiokerroin saa satunnaisen alkuarvon tasajakaumasta $U(-2, 2)$. Koska kuitenkin suurin osa regressiokertoimien estimaateista on välillä $[-0{,}1; +0, 1]$ ja niiden keskihajontojen piste-estimaatit ovat useimmiten joitakin sadasosia, automaattisesti annetut alkuarvot vievät posteriorijakauman niin pienelle todennäköisyystiheyden alueelle, että Stan ei erota sitä enää nollasta (tai log-todennäköisyystiheyttä negatiivisesta

äärettömästä). Tällöin Stan ei voi aloittaa ollenkaan posteriorijakauman simulointia.

Niissäkin tapauksissa, joissa log-todennäköisyystiheys ei ole Stanin mukaan negatiivinen ääretön, automaattiset alkuarvot vievät posteriorijakauman niin kauaksi jakauman moodista, että konvergoitumisessa tavoiteposteriorijakaumaan menee joko hyvin kauan, tai simulointiketju jää jumiin (pitkäksi aikaa) hyvin pienen todennäköisyystiheyden alueelle. Jotta välttäisin nämä ongelmat, käytin omia, realistisempia alkuarvoja estimoitavien parametrien kohdalla. Tätä varten ajoin pienillä havaintomäärillä ja simulaatiomäärillä (nopeasti) kokeiluajoja. Kun sain ”järkeviltä” vaikuttavia parametrien posteriorijakaumien estimaatteja, käytin näin saatuja parametrien odotusarvoestimaatteja alkuarvoina niissä simulaatioissa, jotka ajoin koko mallinnusaineistolle mainituilla simulaatiomäärillä.

4.3.6.3 Simulaatioketjujen konvergoituminen tavoiteposteriorijakaumaan ja posterioriestimaattien luotettavuus

On syytä tarkistaa, että Stan-ohjelmalla tuotetut posteriorijakaumien simulaatiot kuvaavat todella tavoiteposteriorijakaumia, eli että ne ovat konvergoituneet edellä määritettyihin posteriorijakaumiin. Tutkin simulaatioiden konvergenssia samoilla menetelmillä kuin maassaoloaikojen mallinuksen yhteydessä (ks. luku 3.2.5.2).

Simulaatioiden konvergenssien lisäksi on syytä tarkastella simulaatioista tehtävien posterioripäättelyiden luotettavuutta silloin, kun kyse on piste-estimaateista. Käytännössä tämä tarkoittaa *Monte Carlo -keskivirheiden*, *MC-kv* (*engl. Monte Carlo Standard Error, MCSE*) esittämistä näille piste-estimaateille. Olen käsitellyt myös Monte Carlo -keskivirheitä tarkemmin maassaoloaikojen mallinnuksen yhteydessä (ks. luku 3.2.5.3). Ilmoitan MC-keskivirheet parametrien odotusarvoestimaateille. Koska nämä MC-keskivirheet ovat hyvin pieniä, oletan, että myös muiden piste-estimaattien (keskihajonnat, prosenttipisteet) kohdalla MC-kv:t ovat pieniä.

Stan-ohjelma tuottaa hieman erisuuruisia MC-keskivirheitä kuin R-ohjelman Coda-paketti, jota käytin maassaoloaikojen mallinuksen yhteydessä. Tämä johtuu siitä, että näissä ohjelmissa käytetään eri menetelmiä Markov-ketjujen autokorrelaation estimointiin. MC-keskivirheet ovat joka tapauksessa kuitenkin niin pieniä, ettei tällä ole käytännön merkitystä.

4.3.7 Mallin sopivuuden tarkistaminen: posterioriennustetarkistukset

Maahanmuuttajaryhmäkohtaisten mallien tulosten tarkistaminen noudatti samoja käytäntöjä kuin maassaoloaikojen mallinnuksen kohdalla (ks. luku 3.2.6). Arvioin heti estimointitulosten valmistuttua niiden järkevyyttä suhteessa mallin ulkopuoliseen informaatioon. Tämä tarkoitti käytännössä sitä, että esimerkiksi yksikään regressiokerroin ei voisi saada odotusarvoestimaateiksi epärealistisia arvoja. Varsinaisesti tarkistin mallien sopivuuden vertaamalla posterioriestimointitulosten pohjalta laadittuja posterioriennusteita alkuperäisiin havaintoaineistoihin. Tarkoitus oli tutkia, miten hyvin posterioriennusteiden jakaumat toistivat alkuperäisten havaintoaineistojen jakaumat.

Vertasin maahanmuuttajaryhmäkohtaisia posterioriennusteita alkuperäiseen aineistoon sekä koko jakauman osalta että mallinnustavoitteiden kannalta tärkeimpien selittävien muuttujien osalta. Selittävät muuttujat tässä yhteydessä olivat sukupuoli (naisindikaattori) sekä aikamuuttujat (ikävuosi, maahanmuuttovuosi ja maassaoloaika). Aikamuuttujien osalta vertasin ennustejakaumia aineiston jakaumiin näiden selittävien muuttujien keskiarvoihin ja mediaaneihin. Koko jakauman vertailun osalta tarkistuksen tavoitteena oli osoittaa, miten käyttökelpoisia ennusteet ovat koko jakauman osalta. Selittävien muuttujien osalta puolestaan, että miten käyttökelpoisia ennusteet ovat selittävien muuttujien luokkien arvojen mukaan. Tämä oli tärkeää tutkimuksessa tehtävien ennusteiden luotettavuuden kannalta, erityisesti maahanmuuttovuosien ja maassaoloaikojen osalta.

Posterioriennusteet voi laatia joko ”vanhalle” aineistolle tai ”uudelle” aineistolle. Ennusteiden laatiminen vanhalle aineistolle tarkoittaa sitä, että ennusteet laaditaan niille henkilöiille, jotka ovat tutkimusaineistossa, jonka pohjalta mallit estimoitiin. Koska osana mallien posterioriestimointia jokaisen henkilön ”oma” satunnaisvaikutus tuli estimoiduksi, näitä arvoja voisi käyttää laadittaessa ennusteita tälle aineistolle. Koska tämän tutkimuksen tavoitteena on kuitenkin laatia käyttökelpoisia ennusteita sellaisille henkilöille, jotka eivät sisältyneet mallinnusaineistoon, laadin malliaineistoja koskevat posterioriennusteet niin, kuin kyse olisi uudesta aineistosta. Tämä tarkoitti käytännössä sitä, että posterioriennustetarkistuksia varten poimittiin ennustetut arvot myös mallin satunnaisvaikutuksille – eikä vain ennustettavia havaintoja otosjakaumasta.

Posterioriennustetarkistukset on esitetty muiden mallinnustulosten ohessa jäljempänä. Posterioriennustetarkistuksissa vertasin posterioriennustejakaumien tunnuslukuja vastaaviin mallinnusaineistojen jakaumien tunnuslukuihin sekä näiden jakaumien histogrammeja toisiinsa. Nämä tunnusluvut olivat jakaumien keskiarvo, keskihajonta, mediaani sekä muut prosenttipisteet. Vertasin prosenttipisteitä graafisesti sekä esittämällä luvut taulukoissa.

Bayesilainen p-arvo

Formaalimpana menetelmänä posterioriennusteiden jakaumien tunnuslukujen vertaamisessa mallinnusaineistojen vastaaviin lukuihin käytin nk. *bayesilaisia p-arvoja* tai *posterioriennusteiden p-arvoja* (*engl. Bayesian p-value, Posterior predictive p-value*). Bayesilaisten p-arvojen tarkoituksena ei ollut käyttää niitä hypoteesien hylkäämiseen (tai jopa mallinvalintaan) niin kuin frekventistisessä päättelyssä toimitaan (frekventististen) p-arvojen kanssa, vaan menetelmänä sen arvioimiseen, kuinka hyvin mallien tuottamat ennusteet vastaavat alkuperäisiä havaintoaineistoja – ja siten kuinka luotettavia mallien pohjalta tehtävät ennusteet ovat.

Bayesilainen p-arvo on havaintoaineistolla ehdollistettu todennäköisyys sille, että posterioriennusteaineistosta saatu testisuureen arvo on suurempi tai yhtä suuri kuin testisuure havaintoaineistossa:

$$\textit{Bayes-p-arvo} := P\left(\left(T(\boldsymbol{y^{enn}}, \boldsymbol{\theta}) \geq T(\boldsymbol{y^{hav}}, \boldsymbol{\theta})\right) | \boldsymbol{y^{hav}}\right), \tag{4.22}$$

missä $\boldsymbol{y^{enn}}$ on posterioriennuste, $\boldsymbol{y^{hav}}$ on havaintoaineisto ja $\boldsymbol{\theta}$ ovat mallin parametrit.

Posteriorijakauman simulointiin perustuvassa estimoinnissa bayesilainen p-arvo kertoo, kuinka todennäköistä on saada yhdestä simulaatiosta testisuureen arvo, joka on suurempi tai yhtä suuri kuin testisuureen

arvo havaintoaineistossa. Yllä olevan lauseen mukainen bayesilainen p-arvo saadaan posteriorisimulaatioista integroimalla posterioriennusteesta malliparametrit ja posterioriennusteet pois yli kaikkien simulaatioiden, jolloin jäljelle jää indikaattori sille, että posterioriennusteen simulaatiokohtainen testisuureen arvo on suurempi tai yhtä suuri kuin aineiston testisuuren arvo. Simulaatioista laskettu bayesilainen p-arvo on siten niiden simulaatioiden osuus kaikista simulaatioita, joissa testisuureen arvo on suurempi tai yhtä suuri kuin havaintoaineistossa. (Gelman et al. 2014, 146).

Testisuureita, joille bayesilaiset p-arvot lasketaan, ovat edellä mainitut jakaumien tunnusluvut. Koska bayesilaiset p-arvot ovat todennäköisyyksiä, ne voivat saada arvoja nollan ja yhden välillä. Arvot, jotka ovat ”lähellä” nollaa tai yhtä viittaavat siihen, että on epätodennäköistä, että havaintoaineiston jakaumatunnusluku olisi syntynyt mallin mukaisesta dataa generoivasta mekanismista. Arvot ”lähellä” 0,5:ttä puolestaan viittaavat siihen, että havaintoaineiston jakaumatunnusluku olisi voinut syntyä mallin mukaisesta dataa generoivasta mekanismista. Luvut lähellä 0,5:ttä ovat siten tavoiteltavia.

Bayesilainen uskottavuusintervalli

Bayesilaisen p-arvon käyttö ei mielestäni kuitenkaan ole tutkimuksen asetelmassa käytännöllinen menetelmä sen arvioimiseksi, kuinka käyttökelpoisia ennusteita lopulliset mallit tuottavat. Tämä johtuu siitä, että monen testisuureen kohdalla mallit tuottavat simulaatioita, jotka ovat ”lähellä” havaintoaineiston arvoa, mutta kuitenkin keskittyneet siten, että bayesilainen p-arvo on joka nolla tai yksi tai lähellä näitä ääripäitä. ”Lähellä” tarkoittaa tässä tapauksessa läheisyyttä havaintoaineiston ja ennusteiden mittakaavassa: esimerkiksi Irakissa syntyneiden mallin kohdalla posterioriennustejakauman odotusarvo on ”vain” 460 euron päässä havaintoaineiston keskiarvosta, mutta silti sen p-arvo on lähes yksi (taulukossa 5.11, s. 302). Koska havainto- ja ennusteaineistossa liikutaan kuitenkin tuhansien ja kymmenien tuhansien eurojen skaalalla, käytännön ennustetarkoituksen kannalta malli tuottaa esimerkissä tehtävänsä, eli tuottaa suhteellisen tarkkoja ennusteita, vaikka sen p-arvo onkin kehno. Äärimmäiset p-arvotkin voidaan hylätä, jos epäsopivuus on pientä verrattuna aineiston sisäiseen vaihteluun, kuten esimerkin tapauksessa on (Gelman et al. 2014, 150).

Arvioidakseni käytännönläheisemmin mallien tuottamien ennusteiden luotettavuutta havaintoaineiston valossa, käytin bayesilaiseen uskottavuusintervalliin perustuvaa testiä:

$$bayes_intervalli : a(euroa) := P\left(\left(\|T(\boldsymbol{y^{enn}}, \boldsymbol{\theta}) - T(\boldsymbol{y^{hav}}, \boldsymbol{\theta})\| \leq a\right) \,|\boldsymbol{y^{hav}}\right), \tag{4.23}$$

missä a on luku, joka edustaa ennusteiden ”toivottavaa” tarkkuutta, $\boldsymbol{y^{enn}}$ on posterioriennuste, $\boldsymbol{y^{hav}}$ on havaintoaineisto ja $\boldsymbol{\theta}$ ovat mallin parametrit. Bayesilainen uskottavuusintervalli kertoo, kuinka suuri osuus ennusteiden tuottamasta testisuureesta sisältyy korkeintaan testattavan "toivottavan"etäisyyden päähän havaintoaineiston testisuureesta. Esimerkiksi, jos tarkasteltava testisuure on keskihajonta ja testattava etäisyys a on 250, bayesilainen intervalli kertoo, kuinka suuri osuus ennustesimulaatioista on sellaisia, että simulaation ennusteiden jakauman keskihajonta on korkeintaan 250 (euron) etäisyydellä havaintoaineiston jakauman estimoidusta keskihajonnasta.

Bayesilaisen intervallin arvot lähellä ykköstä ovat toivottavia, sillä silloin simuloidut posterioriennusteen

testisuureen arvot sijoittuvat toivottavalle etäisyydelle havaitusta testisuureen arvosta. Mitä lähempänä intervallin arvo on nollaa, sitä pienempi osuus ennustesimulaatioiden testisuureen arvoista sijoittuu toivotulle etäisyydelle havaitusta arvosta. Esitän tulosten yhteydessä bayesilaiset intervallit samoille jakaumien tunnusluvuille kuin bayesilaiset p-arvotkin. Esitän tulokset eri suuruisille intervalleille, sillä toivottu ennustetarkkuus voi sekä vaihdella yhden testisuureen kohdalla että myös testisuureiden välillä. Esimerkiksi mediaanin osalta simulaatioiden voi olla toivottavampaa sijoittua euromääräisesti lähemmäksi havaintoaineiston arvoa kuin pienten ja isojen prosenttipisteiden eli jakauman häntien kohdalla.

4.4 Havaintoaineistot, estimointitulokset ja posterioriennustetarkistukset

Esitän tässä luvussa työikäisten nettovaikutuksia koskevat havainnot, syntymävaltiokohtaiset mallinnustulokset sekä posterioriennustetarkistukset. Aloitan esittämällä nettovaikutusmallien parametrien selitetaulukon, joka toimii apuna syntymävaltiokohtaisten mallien tulosten tarkasteluun. Olen esittänyt malliparametrien tarkemman selostuksen sekä valintaperiaatteet edellä.

Koska työikäisten nettovaikutusmallien estimointi tapahtui syntymävaltiokohtaisesti, käyn estimointitulokset ja posterioriennustetarkistukset myös syntymävaltio kerrallaan läpi. Esitän jokaisen syntymävaltion kohdalla aluksi selitettävän muuttujan eli työikäisten nettovaikutushavaintojen jakauman havaintojen kokonaislukumäärän ja tärkeimmät tunnusluvut. Lisäksi esitän jakauman histogrammina. Kaikkien havaintojen yhteisjakauman lisäksi esitän myös nettovaikutusten henkilökeskiarvojen jakaumalle samat luvut sekä histogrammin.

Tämän jälkeen esitän selittävien muuttujien jakaumien tärkeimmät tunnusluvut: keskiarvot, keskihajonnat ja kvartiilit. Selittävien muuttujien jakaumatietojen jälkeen esitän 5 perättäiseen aineistomalliin pohjautuvien parametrien posteriorijakaumien odotusarvoestimaatit. Estimoitavien parametrien odotusarvoestimaattien lisäksi esitän ennustetarkkuutta mittaavien WAIC:n ja LOO:n estimaatit sekä näiden estimaattien keskivirheet kaikille 5 mallille. WAIC:n ja LOO:n pienemmät estimaatit tarkoittavat parempaa ennustarkkuutta. Lisäksi esitän tyhjälle hierarkkiselle normaalijakaumamallille (malli 1) sisäkorrelaation estimaatin. Tämä auttaa hahmottamaan sitä, missä suhteessa nettovaikutusten kokonaisvaihtelu koostuu henkilöiden *välisestä* vaihtelusta ja missä määrin henkilöiden *"sisäisestä"* vaihtelusta.

Esitän lisäksi syntymävaltiokohtaisesti sen mallin osalta, jota käytän elinkaariennusteiden laatimiseen uusille maahanmuuttajille ja Suomessa syntyville, muitakin parametrien posteriorijakaumien estimaatteja kuin odotusarvon. Irakissa ja Somaliassa syntyneiden osalta tämä oli malli 4. Suomessa syntyneiden osalta malli 5. Suomessa syntyneiden osalta mallinnusaskeleet eivät noudata täysin edellä esitettyä mallinnusaskellusta, mihin palaan jäljempänä, kun esitän tulokset Suomessa syntyneitä koskien.

Irakissa ja Somaliassa syntyneiden osalta valitsin elinkaarimallinnukseen mallin 4 mallin 5 sijasta – vaikka jälkimmäisen ennustetarkkuus vaikuttaa olevan hieman parempi – koska mallien 5 tulokset eivät ole tarpeeksi "järkeviä" elinkaariennusteissa. Tämä johtuu siitä, että jos annan elinkaariennusteissa maassaoloajan vaikutuksen pysyä samanlaisena myös aineiston ulkopuolelle menevien maassaolovuosien eli yli 23 maassaolovuoden osalta, jakaumien vinoudet kasvavat (mielestäni) epärealistisen suuriksi, mikä johtaa myös elinkaariennusteissa käytettävien nettovaikutusennusteiden epärealistiseen negatiivisuuteen, kun maassaoloaika kasvattaa jakauman vinoutta vasemmalle ja epärealistiseen positiivisuuteen, kun maassaoloaika kasvattaa jakauman vinoutta oikealle.

Posterioriennustarkistuksissa vertaan posterioriennusteita havaintoaineistoon kaikkien havaintojen osalta, henkilöiden keskiarvojen osalta sekä selittävien muuttujien luokkien mukaan jaoteltuina. Kaikkien havaintojen ja henkilöiden keskiarvojen osalta vertaan posterioriennusteiden tuottamia *jakaumia* havaintoaineiston vastaaviin jakaumiin. Tärkeimpien selittävien muuttujien kuten sukupuolen, iän, maassaoloajan ja maahanmuuttovuosien osalta vertaan posterioriennusteiden *keskiarvoja* ja *mediaaneja* havaintoaineiston vastaaviin keskiarvoihin ja mediaaneihin näiden selittävien muuttujien luokkien osalta. Vertaan keskiarvojen lisäksi mediaaneja, koska nettovaikutusten havaintojen jakaumien hännät ovat leveitä. Tä-

män vuoksi mediaani voi olla keskiarvoa robustimpi mittari jakaumien keskikohdan kuvaamiseen. Tähän viittaa se, että mallinnuksessa otosjakaumiksi sopivat hyvin (vinot) Studentin t-jakaumat, joiden vapausasteet olivat pieniä.

Parametrien selitteet

Taulukossa 4.1 on esitetty työikäisten syntymävaltiokohtaisten mallien estimoitavien parametrien nimet ja niiden selitteet. Taulukossa esitettävät parametrit ovat Somaliassa syntyneiden nettovaikutusmallista, ja parametrit hieman vaihtelevat syntymävaltiokohtaisten mallien välillä. Taulukosta saa kuitenkin yleisen käsityksen siitä, millaisia parametreja malleissa on.

Taulukon 4.1 parametrit, joista ensimmäinen on ”nainen”, ja viimeinen on, ”(Ikä $< 19) \times$ ikä”, ovat selittävien muuttujien regressiokertoimia. Parametrien nimet ja selitteet kertovat, mihin selittävään muuttujaan kukin regressiokerroin liittyy. Indikaattorimuuttujien osalta selitetaulukossa ovat näiden vertailuluokat.

Parametrit ”Nettovaikutus$_{t-1}$” ja ”Nettovaikutus$_{t-2}$” ovat nettovaikutusten autoregressiivisia kertoimia. Ensimmäinen näistä vastaa edellisen vuoden nettovaikutuksen vaikutusta nykyiseen havaintoon ja jälkimmäisessä on kyse kahden vuoden takaisen nettovaikutuksen arvon vaikutuksesta. Ilmoitan estimointitulokset näiden parametrien osalta prosentteina.

Parametrit $\{\tau_{i1}, \tau_{i2}, \tau_{i3}\}$ ovat henkilötason satunnaisvaikutusten keskihajontoja. Parametrit $\{\tau_{m1}, \tau_{m2}\}$ ovat maahanmuuttovuoteen liittyvien satunnaisvaikutusten keskihajontoja, ja $\{\tau_{t1}, \tau_{t2}\}$ ovat puolestaan maassaoloaikaan liittyvien satunnaisvaikutusten keskihajontoja. Loput parametreista ovat σ, eli havaintotason virhetermin hajontaparametri sekä vinon Studentin t-jakauman vapausasteparametri (ν) ja vinoutta selittävät parametrit.

Taulukko 4.1: *Nettovaikutusmallien estimoitavien parametrien nimet ja niiden selitteet. Esimerkkinä Somalia.*

Parametrinimi	**Selite**
Vakio	Odotusarvo, kun kaikkien selittävien muuttujien arvot keskiarvoja.
Nainen	Vertailuluokka: mies.
Ikä	Arvo muuttuu ajassa.
Ikä^2	Iän neliö: $(\text{Ikä} - 39)^2$. Arvo muuttuu ajassa.
Alle 19-vuotias	Vertailuluokka: yli 18-vuotiaat.
Yli 60-vuotias	Vertailuluokka: alle 61-vuotiaat.
Ikä $\times$ nainen	Iän vaikutus: naisilla verrattuna miehiin.
Ikä^2 $\times$ nainen	Iän neliön vaikutus: naisilla verrattuna. miehiin.
Muuttovuosi $<$ 1993	Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2007–2011.

Jatkuu seuraavalla sivulla

Taulukko 4.1 – jatkuu edelliseltä sivulta

Parametrinimi	Selite
Muuttovuosi v. 1993–1996	Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2007–2011.
Muuttovuosi v. 1997–2001	Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2007–2011.
Muuttovuosi v. 2002–2006	Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2007–2011.
Muuttovuoden työttömyys-%	Maahanmuuttovuoden työttömyysprosentti Suomessa.
Maassaolo (vuosia)	Aika vuosina maahanmuutosta. Arvo muuttuu ajassa.
Maassaolo $\times$ nainen	Maassaoloajan vaikutus: naisilla verrattuna miehiin.
Maassaolo $\times$ (maassaolo < 5)	Maassaoloajan vaikutus: alle 5 vuotta maassa verrattuna yli 4 vuotta maassa.
Maassaolo $\times$ (maassaolo < 5) $\times$ nainen	Maassaoloajan vaikutus: alle 5 vuotta maassa verrattuna yli 4 vuotta maassa. Vaikutus naisilla verrattuna miehiin.
Maassaolo $\times$ (muuttovuosi $<$ 1993)	Maassaoloajan vaikutus: ennen 1993 muuttaneilla verrattuna myöhemmin muuttaneilla.
Maassaolo < 5	Vertailuluokka: maassaoloaika yli 4 vuotta.
(Maassaolo < 5) $\times$ nainen	Maassaoloaika alle 5 vuotta: vaikutus naisilla verrattuna miehiin.
Maassaolo alle 1 vuosi	Vertailuluokka: maassaoloaika vähintään 1 vuoden.
Muutto alle 20-v.	Maahanmuuton ikäluokka. Vertailuluokka: muutto yli 39-v.
Muutto 20–29-v.	Maahanmuuton ikäluokka. Vertailuluokka: muutto yli 39-v.
Muutto 30–39-v.	Maahanmuuton ikäluokka. Vertailuluokka: muutto yli 39-v.
Muutto alle 20-v. $\times$ maassaolo	Maassaoloajan vaikutus: maahanmuutto alle 20-v. verrattuna muutto yli 39-v.
Muutto 20–29-v. $\times$ maassaolo	Maassaoloajan vaikutus: maahanmuutto 20–29-v. verrattuna muutto yli 39-v.
Muutto 30–39-v. $\times$ maassaolo	Maassaoloajan vaikutus: maahanmuutto 30–39-v. verrattuna muutto yli 39-v.
Kotouttamismenoja	Maahanmuutto v. 2002 jälkeen ja maassaoloaika alle 5 vuotta.
(Kalenterivuosi > 2000)	Vertailuluokka: kalenterivuosi ennen vuotta 2001.
Pääasiallinen toiminta: työllinen	Vertailuluokka: pääasiallinen toiminta muuttovuonna: muu.
Pääasiallinen toiminta: opiskelija	Vertailuluokka: pääasiallinen toiminta muuttovuonna: muu.

Jatkuu seuraavalla sivulla

Taulukko 4.1 – jatkuu edelliseltä sivulta

Parametrinimi	Selite
Korkein koulutus: 2. asteen koulutus	Vertailuluokka: muuttovuonna ei 2.asteen koulutusta/ei tietoa.
Korkein koulutus: 3. asteen koulutus	Vertailuluokka: muuttovuonna ei 2.asteen koulutusta/ei tietoa.
Alle 7-vuotiaiden lasten lukumäärä	Lasten määrä maahanmuuttovuonna.
Puoliso	Vertailuluokka: ei puolisoa maahanmuuttovuonna.
Kotikunta Uudellamaalla	Vertailuluokka: kotikunta ei Uudellamaalla maahanmuuttovuonna.
Ikä > 55	Vertailuluokka: ikä alle 56 vuotta.
(Ikä > 55) $\times$ ikä	Iän vaikutus: yli 55-vuotiailla verrattuna alle 56-vuotiaat.
(Ikä < 40) $\times$ ikä $\times$ nainen	Iän vaikutus: alle 40-v. verrattuna yli 39-v. Vaikutus naisilla verrattuna miehiin.
(Ikä < 40) $\times$ ikä	Iän vaikutus: alle 40-v. verrattuna yli 39-v.
(Ikä > 50) $\times$ ikä $\times$ nainen	Iän vaikutus: yli 50-v. verrattuna alle 51-v. Vaikutus naisilla verrattuna miehiin.
(Ikä > 50) $\times$ ikä	Iän vaikutus: yli 50-v. verrattuna alle 51-v.
(Ikä < 19) $\times$ ikä $\times$ nainen	Iän vaikutus: alle 19-v. verrattuna yli 19-v. Vaikutus naisilla verrattuna miehiin.
(Ikä < 19) $\times$ ikä	Iän vaikutus: alle 19-v. verrattuna yli 19-v.
Nettovaikutus$_{t-1}$ (%)	Autoregressiokerroin prosentteina. Nettovaikutus yhden vuoden ennen havaintoa.
Nettovaikutus$_{t-2}$ (%)	Autoregressiokerroin prosentteina. Nettovaikutus kaksi vuotta ennen havaintoa.
τ_{i1}	Henkilötason satunnaisvaikutuksen keskihajonta: vakiotermi.
τ_{i2}	Henkilötason satunnaisvaikutuksen keskihajonta: maassaoloaika.
τ_{i3}	Henkilötason satunnaisvaikutuksen keskihajonta: iän neliö.
τ_{m1}	Maahanmuuttovuositason satunnaisvaikutuksen keskihajonta: vakiotermi.
τ_{m2}	Maahanmuuttovuositason satunnaisvaikutuksen keskihajonta: naisindikaattori.
τ_{t1}	Maassaolotason satunnaisvaikutuksen keskihajonta: vakiotermi.
τ_{t2}	Maassaolotason satunnaisvaikutuksen keskihajonta: naisindikaattori.
σ	Havaintotason virhetermin hajontaparametri.
ν	Studentin t-jakauman vapausasteparametri.

Jatkuu seuraavalla sivulla

Taulukko 4.1 – jatkuu edelliseltä sivulta

Parametrinimi	Selite
γ/Vinous: vakio	Vinon Studentin t-jakauman vinousparametri./Vinousmallin vakio.
Vinous: nainen	Jakauman vinous. Vertailuluokka: mies.
Vinous: ikä	Jakauman vinous. Arvo muuttuu ajassa.
Vinous: ikä^2	Jakauman vinous. Arvo muuttuu ajassa.
Vinous: muuttovuosi $<$ 1993	Jakauman vinous. Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2002–2011.
Vinous: muuttovuosi v. 1993–2001	Jakauman vinous. Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2002–2011.
Vinous: maassaolo (vuosia)	Jakauman vinous. Aika vuosina maahanmuutosta. Arvo muuttuu ajassa.
Vinous: (muuttovuosi $<$ 1993) $\times$ maassaolo	Jakauman vinous. Maassaoloajan vaikutus: ennen 1993 muuttaneilla verrattuna myöhemmin muuttaneet.
Vinous: maassaolo $<$ 5	Jakauman vinous. Vertailuluokka: maassaoloaika yli 4 vuotta.
Vinous: (maassaolo $<$ 5)$\times$ maassaolo	Jakauman vinous.Maassaoloajan vaikutus: alle 5 vuotta maassa verrattuna yli 4 vuotta maassa.

4.4.1 Irak

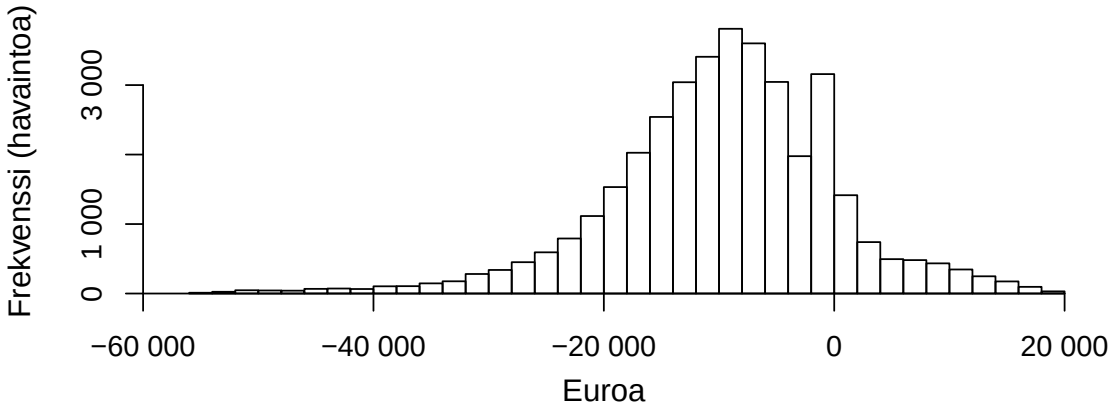
4.4.1.1 Irak: havaintoaineisto

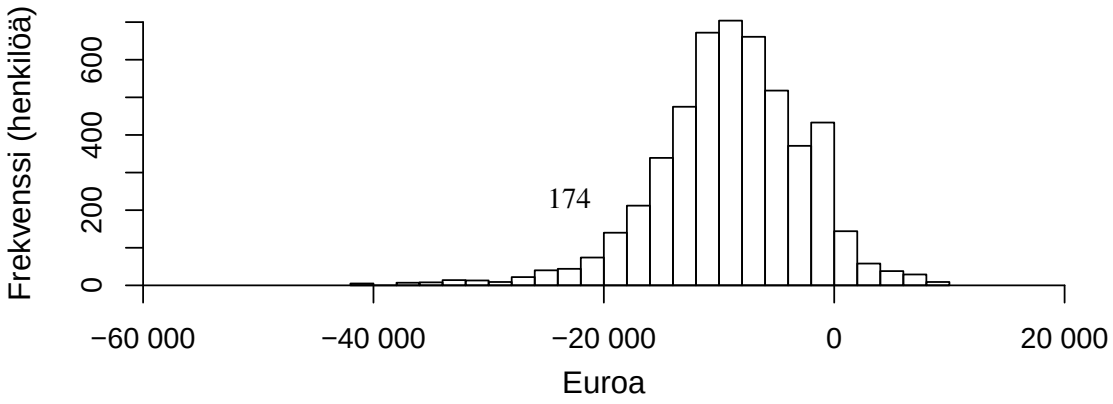
Taulukossa 4.2 on esitetty Irakissa syntyneiden 16–62-vuotiaiden nettovaikutushavaintojen lukumäärä ja havaintojen jakauman keskeisimmät tunnusluvut. Taulukossa on esitetty samat tiedot myös havaintoaineiston henkilöiden nettovaikutushavaintojen keskiarvoille. Mallinnusaineistoon ei sisälly alle 16-vuotiaana muuttaneita. Nämä nettovaikutushavainnot muodostavat selitettävän muuttujan nuorten ja työikäisten julkisen talouden nettovaikutusten mallinnuksessa. Taulukon 4.2 tietojen mukaan esimerkiksi kaikkien havaintojen keskiarvo on -10 200 euroa ja mediaani -9 200 euroa. Havaintoaineiston henkilöiden (5 147 henkilöä) nettovaikutushavaintojen keskiarvojen mediaani on -8 800 euroa ja keskihajonta 11 600 euroa.

Kuvassa 4.3 on esitetty histogrammit sekä kaikkien nettovaikutushavaintojen jakaumalle (ylempi histogrammi) että henkilöiden nettovaikutusten keskiarvojen jakaumalle (alempi histogrammi). Kummastakin jakaumasta on katkaistu pois niiden ylin ja alin prosentti pois. Kaikkien havaintojen histogrammista näkee, että jakauma on vasemmalle vino. Lisäksi siitä näkee, että osa havainnoista on kertynyt nollan lähelle (pylväs heti nollan vasemmalla puolella). Henkilöiden keskiarvojen jakauma näyttää histogrammin perusteella silmämääräisesti melko normaalijakautuneelta lukuun ottamatta havaintoja noin -30 000 eurosta vasemmalle.

Taulukko 4.2: *Irakissa syntyneet 16–62-vuotiaat. Nettovaikutushavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa sekä henkilöiden lukumäärä ja henkilöiden nettovaikutushavaintojen keskiarvojen jakauman tunnusluvut mallinnusaineistossa. Euromääräiset tunnusluvut pyöristetty lähimpään sataan euroon.*

	Havainnot	Henkilöt
Kokonaislukumäärä (N)	37 830	5 147
Keskiarvo	-10 200	-9 500
Keskihajonta	16 400	11 600
1. prosenttipiste (1 %)	-54 900	-42 900
5. prosenttipiste (5 %)	-28 100	-20 900
10. prosenttipiste (10 %)	-21 900	-17 200
Alakvartiili (25 %)	-15 100	-12 700
Mediaani (50 %)	-9 200	-8 800
Yläkvartiili (75 %)	-3 500	-4 600
90. prosenttipiste (90 %)	1 400	-500
95. prosenttipiste (95 %)	7 200	600
99. prosenttipiste (99 %)	18 600	9 200





Taulukossa 4.3 on esitetty Irakissa syntyneiden nettovaikutusmallien selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden kaikkien havaintojen joukossa. Edellä parametrien selitetaulukossa (s. 170) on lyhyitä selitteitä selittävistä muuttujista.

Taulukko 4.3: *Irakissa syntyneet 16–62-vuotiaat. Nettovaikutusmallin selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta).*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,39	0,49	0	0	1
Ikä	35,4	10,2	27	35	42
Ikä^2	1358	768	729	1225	1764
Alle 19-vuotias	0,02	0,15	0	0	0
Yli 60-vuotias	0,01	0,08	0	0	0
Ikä $\times$ nainen	13,6	18,2	0	0	30
Ikä^2 $\times$ nainen	516	806	0	0	900
Muuttovuosi $<$ 1992	0,07	0,25	0	0	0
Muuttovuosi v. 1992 tai 1995–1997	0,30	0,46	0	0	1
Muuttovuosi v. 1993–1994	0,11	0,32	0	0	0
Muuttovuosi v. 1998–2001	0,21	0,40	0	0	0
Muuttovuosi v. 2002–2006	0,15	0,35	0	0	0
Muuttovuosi v. 2007–2011	0,17	0,38	0	0	0
Muuttovuoden työttömyys-%	10,9	3,5	8	10	15
Maassaolo (vuosia)	6,1	5,0	2	5	10
Maassaolo $\times$ nainen	2,4	4,2	0	0	3
Maassaolo $\times$ (maassaolo $<$ 8)	1,9	2,3	0	1	4
Maassaolo $\times$ (maassaolo $<$ 8) $\times$ nainen	0,8	1,8	0	0	0
Maassaolo $\times$ (muuttovuosi $<$ 1992)	0,87	3,50	0	0	0
Maassaolo $<$ 8	0,65	0,48	0	1	1
(Maassaolo $<$ 8) $\times$ nainen	0,25	0,43	0	0	1
Maassaolo alle 1 vuosi	0,12	0,33	0	0	0
Muutto alle 20-v.	0,14	0,34	0	0	0
Muutto 20–29-v.	0,42	0,49	0	0	1
Muutto 30–39-v.	0,30	0,46	0	0	1
Muutto yli 39-v.	0,14	0,34	0	0	0
Muutto alle 20-v. $\times$ maassaolo	0,73	2,51	0	0	0
Muutto 20–29-v. $\times$ maassaolo	2,58	4,51	0	0	4
Muutto 30–39-v. $\times$ maassaolo	2,04	4,21	0	0	1
Muutto yli 39-v. $\times$ maassaolo	0,75	2,54	0	0	0
Kalenterivuosi	2005	4,6	2002	2006	2009

Jatkuu seuraavalla sivulla

Taulukko 4.3 – jatkuu edelliseltä sivulta

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Kalenterivuosi × nainen	779	977	0	0	2003
Kalenterivuosi < 2001	0,20	0,40	0	0	0
Pääasiallinen toiminta: työllinen	0,04	0,20	0	0	0
Pääasiallinen toiminta: opiskelija	0,11	0,31	0	0	0
Pääasiallinen toiminta: muu	0,85	0,36	1	1	1
Korkein koulutus: 2. asteen koulutus	0,10	0,30	0	0	0
Korkein koulutus: 3. asteen koulutus	0,08	0,27	0	0	0
Korkein koulutus: ei 2.astetta/ei tietoa	0,82	0,38	1	1	1
Alle 7-vuotiaiden lasten lukumäärä	0,54	0,95	0	0	1
Puoliso	0,50	0,50	0	1	1
Kotikunta Uudellamaalla	0,32	0,47	0	0	1
Ikä > 55	0,04	0,19	0	0	0
(Ikä > 55) × ikä	2,1	10,9	0	0	0
(Ikä > 45) × ikä × nainen	3,4	12,8	0	0	0
(Ikä > 45) × ikä	0,06	0,25	0	0	0
(Ikä > 55) × ikä × nainen	0,8	7,0	0	0	0
(Ikä > 55) × ikä	0,01	0,12	0	0	0
(Ikä < 19) × ikä × nainen	0,1	1,6	0	0	0
(Ikä < 19) × ikä	0,01	0,09	0	0	0

4.4.1.2 Irak: estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Taulukossa 4.4 on esitetty perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit Irakissa syntyneille 16–62-vuotiaille. Olen käsitellyt viiden perättäisen mallin rakentamista mallinnusaskel kerrallaan luvussa 4.3.3. Estimoitavien parametrien selitteet on puolestaan esitetty parametrien selitetaulukossa s. 170.

Mallin vakiotermiä, regressiokertoimia, satunnaisvaikutusten ja virhetermin keskihajontoja (τ-parametrit, σ) sekä vinousparametreja koskevat odotusarvoestimaatit ovat euromääräisiä. Autoregressiiviset kertoimet on ilmaistu prosentteina ja t-jakauman vapausasteilla (ν) ei puolestaan ole mitään erillistä yksikköä.

Malli 1. Käyn lyhyesti läpi joitakin mallikohtaisia keskeisimpiä tuloksia taulukosta 4.4. Mallissa 1 eli tyhjässä normaalijakaumaan perustuvassa mallissa on vain kolme parametria: kaikkien havaintojen yhteinen vakio, vakiotermin henkilötason satunnaisvaikutuksen keskihajonta (τ_{i1}) sekä havaintojen virhetermin keskihajonta (σ). Vakio on estimaatti kaikkien havaintojen jakauman odotusarvolle, ja hajontapa-

rametrien avulla voidaan tässä tyhjässä mallissa ilmaista, kuinka suuri osuus mallin kokonaisvaihteluista on henkilöiden *välistä* (τ_{i1}) sekä kuinka suuri osuus vaihtelusta on henkilöiden ”*sisäistä*” vaihtelua (σ). Näiden hajontaparametrien avulla saadaan laskettua sisäkorrelaatio, jonka odotusarvo on noin 0,4. Tämä tarkoittaa sitä, että Irakissa syntyneillä noin 40 %:ia nettovaikutushavaintojen välisestä kokonaisvaihtelusta aiheutuu henkilöiden välisestä vaihtelusta. Sisäkorrelaatio on samaa suuruusluokkaa, noin 0,4–0,5, myös muiden syntymävaltioiden osalta.

Malli 2. Malli 2 on muuten kuin malli 1, mutta nyt aineistomallin otosjakaumana on normaalijakauman sijaan sitä vastaava robustimpi jakauma, Studentin t-jakauma. Tämä tuo malliin yhden uuden parametrin, t-jakauman vapausasteet, ν. Tämän parametrin odotusarvo on n. 2, eli vapausasteiden lukumäärä on ”pieni”, mikä johtuu nettovaikutusten jakauman leveistä hännistä. T-jakaumaan perustuvan mallin ennustetarkkuuden mittarit Waic ja LOO ovat paremmat kuin normaalijakaumaan perustuvan mallin 1. Mallin 2 Waic ja LOO ovat noin 60 000, ja mallin 1 puolestaan noin 97 000. Pienempi luku tarkoittaa parempaa ennustetarkkuutta.

Taulukko 4.4: *Perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit, mallikohtaiset ennustetarkkuuden estimaatit ja keskihajonnat (Waic ja LOO) sekä sisäkorrelaatioestimaatti. Mallien valinta on esitetty luvussa 4.3.3.*

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
Vakio	-9 860	-8 320	-9 070	-7 840	-7 860
Nainen			-3 760	-5 040	-3 630
Ikä			-1 580	-1 250	-870
Ikä^2			1 610	1 020	760
Alle 19-vuotias			1 440	1 210	1 120
Yli 60-vuotias			-50	-230	-220
Ikä $\times$ nainen			-3 830	-4 360	-4 190
Ikä^2 $\times$ nainen			4 660	5 200	4 940
Muuttovuosi $<$ 1992			1 510	1 900	2 690
Muuttovuosi v. 1992 tai 1995–1997			2 060	2 410	2 640
Muuttovuosi v. 1993–1994			760	1 020	1 460
Muuttovuosi v. 1998–2001			740	1 000	1 410
Muuttovuosi v. 2002–2006			760	780	730
Muuttovuoden työttömyys-%			-110	-160	-120
Maassaolo (vuosia)			2 320	1 940	1 140
Maassaolo $\times$ nainen			-50	-440	-280
Maassaolo $\times$ (maassaolo $<$ 8)			4 210	2 370	910
Maassaolo $\times$ (maassaolo $<$ 8) $\times$ nainen			140	160	400
Maassaolo $\times$ (muuttovuosi $<$ 1992)			40	340	240
Maassaolo $<$ 8			1 830	880	710

Jatkuu seuraavalla sivulla

Taulukko 4.4 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
(Maassaolo < 8) $\times$ nainen			280	620	730
Maassaolo alle 1 vuosi			8 500	7 210	6 870
Muutto alle 20-v.			760	670	520
Muutto 20–29-v.			900	560	520
Muutto 30–39-v.			720	530	540
Muutto alle 20-v. $\times$ maassaolo			230	20	-40
Muutto 20–29-v. $\times$ maassaolo			10	-150	-160
Muutto 30–39-v. $\times$ maassaolo			230	270	250
Kalenterivuosi < 2001			100	0	-10
Pääasiallinen toiminta: työllinen			3 000	4 110	4 240
Pääasiallinen toiminta: opiskelija			-230	-1 070	-980
Korkein koulutus: 2. asteen koulutus			10	-100	-120
Korkein koulutus: 3. asteen koulutus			950	670	640
Alle 7-vuotiaiden lasten lukumäärä			-740	-1 090	-1 060
Puoliso			50	-160	-180
Kotikunta Uudellamaalla			590	1 010	990
Ikä > 55			-1 290	-3 220	-3 460
(Ikä > 55) $\times$ ikä			300	1 470	1 590
(Ikä > 45) $\times$ ikä $\times$ nainen			-1 770	-1 670	-1 590
(Ikä > 45) $\times$ ikä			1 910	1 940	1 860
(Ikä > 55) $\times$ ikä $\times$ nainen			-1 230	510	600
(Ikä > 55) $\times$ ikä			3 050	-420	-540
(Ikä < 19) $\times$ ikä $\times$ nainen			4 700	8 030	8 250
(Ikä < 19) $\times$ ikä			8 430	13 510	13 700
Nettovaikutus$_{t-1}$ (%)			46,1	29,5	28,1
Nettovaikutus$_{t-2}$ (%)			14,1	5,1	5,4
τ_{h1}	9 120	5 770	2 090	3 820	3 830
τ_{h2}				2 930	2 880
τ_{h3}				560	500
τ_{v1}				820	810
τ_{v2}				390	380
τ_{a1}				320	270
τ_{a2}				140	120
σ	13 600	4 910	4 490	4 140	4 090
ν		2,1	2,08	2,32	2,39
γ/Vinous: vakio				-620	-740
Vinous: nainen					-660
Vinous: ikä					-290
Vinous: ikä^2					190

Jatkuu seuraavalla sivulla

Taulukko 4.4 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
Vinous: muuttovuosi $<$ 1992					-410
Vinous: muuttovuosi v. 1992–2001					-170
Vinous: maassaolo (vuosia)					500
Vinous: (muuttovuosi $<$ 1993)$\times$ maassaolo					70
Vinous: maassaolo $<$ 8					-90
Vinous: (maassaolo $<$ 8)$\times$ maassaolo					680
Sisäkorrelaatio	0,401				
Waic (estimaatti)	97 023	59 876	50 923	35 305	33 572
Waic (estimaatin keskivirhe)	(4 387)	(470)	(477)	(336)	(330)
LOO (estimaatti)	96 299	59 911	50 926	40 300	38 548
LOO (estimaatin keskivirhe)	(4 067)	(470)	(477)	(378)	(371)

Malli 3. Malli 3 on muuten kuin malli 2, mutta nyt olen lisännyt mukaan kaikille havainnoille yhteiset selittävät muuttujat. Taulukossa 4.4 on esitetty näitä selittäviä muuttujia vastaavien regressiokertoimien odotusarvot, jotka ovat euromääräisiä.

Olen skaalannut regressiokertoimien odotusarvoestimaatit niin, että *indikaattorimuuttujien* osalta estimaatit kertovat, mikä on nettovaikutushavainnon odotusarvon muutos, kun indikaattorimuuttujan arvo vaihtuu nollasta ykköseksi eli vertailuluokasta taulukossa 4.4 ilmoitettuun arvoon, kun muiden muuttujien arvot pidetään vakioina. Esimerkiksi indikaattorimuuttuja ”nainen” kertoo, mikä on (vuosittaisen) nettovaikutusten odotusarvo naisilla verrattuna miehiin. Mallin 3 tulosten perusteella Irakissa syntyneiden naisten nettovaikutusten odotusarvo, kun muut muuttujat pidetään vakioina, on -3 760 euroa verrattuna miehiin.

Samaten esimerkiksi ”pääasiallinen toiminta: työllinen”-muuttujan regressiokertoimen estimaatti kertoo, että niillä Irakissa syntyneillä, joilla ensimmäisen suomessaolovuoden pääasiallinen toiminta oli työllinen, vuosittaisten nettovaikutusten odotusarvo on 3 000 euroa korkeampi, kuin jos pääasiallinen toiminta olisi saanut vertailuluokan arvon (”pääasiallinen toiminta: muu”). Edellä parametrien selitetaulukossa (s. 170) on esitetty, mitkä muuttujat ovat indikaattorimuuttujia ja mitkä näiden muuttujien arvojen vertailuluokat ovat.

Muut regressiokertoimet – kuin indikaattorimuuttujat – on skaalattu niin, että estimaatit kertovat, mikä on muutos nettovaikutusten odotusarvossa, kun selittävän muuttujan arvo kasvaa yhden keskihajonnan verran ja muiden selittävien muuttujien arvot pidetään samoina. Esimerkiksi selittävän muuttujan ”ikä” kohdalla tämä tarkoittaa sitä, että Irakissa syntyneillä nettovaikutusten odotusarvo muuttuu ikää koskevan regressiomuuttujan vuoksi -1 580 euroa, kun ikä kasvaa yhdellä keskihajonnalla (tässä ei oteta huomioon iän neliön vaikutusta eikä muita ikään liittyviä vaikutuksia). Edellä esitetystä selittävien muuttujien taulukossa (175) voidaan todeta, että iän keskihajonta on 10,2 (ikä)vuotta, joten mainittu muutos (-1 580

e) nettovaikutusten odotusarvossa tapahtuu, kun henkilön ikä kasvaa 10,2 vuotta.

Esitän interaktiotermien, jossa yksi muuttujista on indikaattorimuuttuja ja toinen muu kuin indikaattorimuuttuja, tulkinnan esimerkin kautta. Interaktiotermin ”ikä $\times$ nainen” regressiokertoimen odotusarvon estimaatti -3 830 (euroa) kertoo, kuinka suuri ero iän vaikutuksessa on sukupuolten välillä, kun muut selittävät muuttujat pidetään vakioina. Tässä tapauksessa siis naisilla yhden keskihajonnan muutos iässä (10,2 ikävuotta) vaikuttaa miehiin *verrattuna* -3 830 euroa. Kääntäen, miehillä yhden keskihajonnan muutos iässä (10,2 ikävuotta) tarkoittaa nettovaikutusten odotusarvoon +3 830 suuruista muutosta *verrattuna* naisiin.

Malli 4. Malli 4 on muuten kuin malli 3, mutta nyt olen lisännyt muutkin satunnaisvaikutukset malliin – malleissa 1–3 olleen henkilöiden välillä vaihtelevan mallin satunnaistermin lisäksi. Taulukossa on esitetty satunnaisvaikutusten keskihajonnat (edellä s. 170 löytyvät satunnaisvaikutusten lyhyet selitteet ja mallinrakennusosassa tarkemmat selosteet). Mallissa 4 olen vaihtanut myös otosjakauman ei-vinosta/tavallisesta Studentin t-jakaumasta *vinoon* Studentin t-jakaumaan. Tästä syystä mallin 4 tuloksissa on vinousparametrin (γ) odotusarvon estimaatti, -620 euroa. Koska tämä estimaatti on negatiivinen, tämä viittaa siihen, että nettovaikutusten jakauma on vasemmalle vino – myös silloin kun kaikki mallissa 4 olevat tekijät huomioidaan.

Malli 5. Malli 5 on muuten kuin malli 4, mutta nyt olen lisännyt jakauman vinouteen vaikuttavia tekijöitä malliin. Vinoutta ”selittävien” muuttujien kertoimilla on sama tulkinta vinouden suhteen kuin edellä regressiokertoimilla nettovaikutusten odotusarvon suhteen. Indikaattorimuuttujien kertoimet kertovat, kuinka paljon vinous muuttuu luokkien välillä. Muiden kuin indikaattorimuuttujien kohdalla puolestaan kertoimet ovat muutoksia vinouteen, kun vinoutta selittävän muuttujan arvo muuttuu yhdellä sen keskihajonnalla. Esimerkiksi ”vinous: ikä”-muuttujan kertoimen estimaatti -290 tarkoittaa sitä, että jakauman vinous muuttuu -290 (euroa), kun ikä muuttuu yhden keskihajonnan positiivisemmaksi eli 10,2 vuotta.

Taulukon 4.4 ennustemittarien Waic ja LOO mukaan isommat mallit ovat parempia ennustajia kuin pienemmät mallit. Ennustemittarien Waic ja LOO estimaatit vastaavat hyvin toisiaan mallien 1–3 osalta, jolloin otosjakaumina ovat normaalijakauma ja Studentin t-jakauma. Mallien 4 ja 5 kohdalla näin ei ole, ja tämä johtuu siitä, että otosjakaumana näissä malleissa on vino Studentin t-jakauma.

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” malli on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän uusien maahanmuuttajien nuorten ja työikäisten nettovaikutusten ennustamiseen Irakissa syntyneiden elinkaariennusteissa. Valitsin lopulliseksi malliksi mallin 4, vaikka ennustetarkkuuden mittareiden puolesta malli 5 olisikin ollut parempi. Koska mallin 5 parametrin, ”vinous: (maassaolo $<$ 8)$\times$ maassaolo”, odotusarvo on positiivinen, tämä tarkoittaa elinkaariennusteiden kannalta sitä, että Irakissa syntyneiden nettovaikutusennusteiden jakauman vinous vasemmalle kasvaisi rajatta kahdeksasta maassaolovuodesta lähtien. Tämä tarkoittaisi myös nettovaikutusten odotusarvon pienenemistä ennusteissa. Koska on vaikeaa perustella, että ennustejakauman vinous vasemmalle kasvaisi rajatta todellisuudessa – ja koska tämä tuottaisi siten epärealistisen negatiivisia ennusteita Irakissa syntyneille uusille maahanmuuttajille – käytän elin-

kaariennusteissa nettovaikutusten ennustamiseen mallia 4.

Taulukossa 4.5 on esitetty Irakissa syntyneiden nettovaikutusmallin (malli 4) posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvojen (OA), keskihajontojen (KH), kvartiilien sekä jakaumien hännistä prosenttipisteiden 2,5 (%) ja 97,5 (%) estimaatit. Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli ykköstä, en käsittele niitä tarkemmin.

Taulukko 4.5: *Irakissa syntyneiden16–62-vuotiaiden nettovaikutusten lopullinen malli (malli 4). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin Monte Carlo-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakio	-7 840	560	-8 930	-8 220	-7 840	-7 460	-6 770	14	1,00
Nainen	-5 040	810	-6 670	-5 570	-5 040	-4 500	-3 500	20	1,00
Ikä	-1 250	420	-2 060	-1 530	-1 260	-970	-400	9	1,00
Ikä^2	1 020	410	220	750	1 040	1 310	1 810	10	1,00
Alle 19-vuotias	1 210	250	730	1 040	1 210	1 380	1 700	4	1,00
Yli 60-vuotias	-230	550	-1 310	-610	-230	150	840	9	1,00
Ikä $\times$ nainen	-4 360	750	-5 840	-4 860	-4 370	-3 860	-2 850	19	1,00
Ikä^2 $\times$ nainen	5 200	820	3 590	4 640	5 210	5 730	6 800	22	1,00
Muuttovuosi $<$ 1992	1 900	960	-40	1 300	1 900	2 540	3 770	27	1,00
Muuttovuosi v. 1992 tai 1995–1997	2 410	2 070	-1 570	1 070	2 330	3 650	6 820	76	1,00
Muuttovuosi v. 1993–1994	1 020	1 330	-1 580	150	1 010	1 800	3 800	41	1,00
Muuttovuosi v. 1998–2001	1 000	860	-720	450	980	1 510	2 790	31	1,00
Muuttovuosi v. 2002–2006	780	640	-470	370	760	1 170	2 080	21	1,00
Muuttovuoden työttömyys-%	-160	220	-630	-290	-150	-20	250	8	1,00
Maassaolo (vuosia)	1 940	390	1 170	1 690	1 930	2 180	2 740	11	1,00
Maassaolo $\times$ nainen	-440	340	-1 080	-650	-440	-220	240	9	1,00
Maassaolo $\times$ (maassaolo $<$ 8)	2 370	880	690	1 800	2 340	2 890	4 230	24	1,00
Maassaolo $\times$ (maassaolo $<$ 8) $\times$ nainen	160	710	-1 310	-270	180	620	1 530	16	1,00
Maassaolo $\times$ (muuttovuosi $<$ 1992)	340	290	-220	140	340	540	900	8	1,00
Maassaolo $<$ 8	880	640	-380	480	870	1 280	2 170	17	1,00
(Maassaolo $<$ 8) $\times$ nainen	620	600	-570	230	610	1 000	1 800	13	1,00
Maassaolo alle 1 vuosi	7 210	570	6 080	6 850	7 210	7 580	8 380	14	1,00
Muutto alle 20-v.	670	610	-500	260	660	1 060	1 860	17	1,00
Muutto 20–29-v.	560	480	-420	250	560	880	1 510	15	1,00
Muutto 30–39-v.	530	380	-240	290	540	770	1 270	12	1,00

Jatkuu seuraavalla sivulla

Taulukko 4.5 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Muutto alle 20-v. × maassaolo	20	340	-670	-210	30	260	680	10	1,00
Muutto 20–29-v. × maassaolo	-150	260	-660	-340	-150	20	370	8	1,00
Muutto 30–39-v. × maassaolo	270	240	-200	110	270	430	740	7	1,00
Kalenterivuosi < 2001	0	130	-250	-80	0	90	270	2	1,00
Pääasiallinen toiminta: työllinen	4 110	290	3 530	3 910	4 110	4 310	4 680	6	1,00
Pääasiallinen toiminta: opiskelija	-1 070	200	-1 460	-1 210	-1 070	-940	-690	4	1,00
Korkein koulutus: 2. asteen koulutus	-100	190	-470	-230	-100	20	270	4	1,00
Korkein koulutus: 3. asteen koulutus	670	220	240	520	670	820	1 110	4	1,00
Alle 7-vuotiaiden lasten lukumäärä	-1 090	80	-1 250	-1 140	-1 090	-1 030	-930	2	1,00
Puoliso	-160	140	-440	-250	-160	-70	110	3	1,00
Kotikunta Uudellamaalla	1 010	120	780	940	1 010	1 090	1 230	2	1,00
Ikä > 55	-3 220	2 750	-8 680	-5 050	-3 230	-1 350	2 060	49	1,00
(Ikä > 55) × ikä	1 470	1 300	-1 040	570	1 460	2 340	4 060	23	1,00
(Ikä > 45) × ikä × nainen	-4 390	1 010	-6 350	-5 050	-4 400	-3 730	-2 340	24	1,00
(Ikä > 45) × ikä	5 100	1 090	2 870	4 380	5 100	5 810	7 220	25	1,00
(Ikä > 55) × ikä × nainen	510	2 010	-3 470	-850	490	1 860	4 490	38	1,00
(Ikä > 55) × ikä	-420	4 260	-8 920	-3 220	-410	2 520	7 940	80	1,00
(Ikä < 19) × ikä × nainen	8 030	4 790	-1 570	4 730	8 090	11 250	17 290	76	1,00
(Ikä < 19) × ikä	13 510	8 460	-3 310	7 710	13 630	19 080	29 910	134	1,00
Nettovaikutus$_{t-1}$ (%)	29,5	0,7	28,1	29	29,5	29,9	30,8	0,02	1,00
Nettovaikutus$_{t-2}$ (%)	5,1	0,5	4,2	4,8	5,1	5,4	6	0,01	1,00
τ_{i1}	3 820	130	3 570	3 740	3 820	3 910	4 080	5	1,00
τ_{i2}	2 930	100	2 730	2 860	2 930	3 000	3 130	5	1,00
τ_{i3}	560	120	320	480	560	630	780	10	1,00
τ_{m1}	820	230	450	650	790	960	1 350	7	1,00
τ_{m2}	390	100	220	320	390	450	630	2	1,00
τ_{t1}	320	80	210	270	310	370	510	2	1,00
τ_{t2}	140	60	30	100	140	180	280	2	1,00
σ	4 140	40	4 070	4 120	4 140	4 170	4 220	2	1,01

Jatkuu seuraavalla sivulla

Taulukko 4.5 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
ν	2,32	0,04	2,24	2,29	2,31	2,34	2,39	0,002	1,01
γ	-620	30	-690	-650	-620	-600	-570	2	1,01

Listaan lyhyesti joitakin huomioita keskeisimmistä vaikutuksista ja estimointituloksista taulukossa 4.5. Kun kirjoitan parametrin sijoittuvan jollekin tietylle intervallille, tarkoitan tällöin kyseisen parametrin posteriorijakauman 95%:n uskottavuusintervallia 2,5 %:sta 97,5 %:iin.

- **Sukupuoli**. Päävaikutus naisilla verrattuna miehiin noin -6 700:n ja -3 500:n välillä. Myös interaktiovaikutukset ”suuria”, erityisesti ikää koskevien muuttujien kanssa.
- **Maahanmuuttovuosi**. Maahanmuuttovuosiluokkaan 2007–2011 verrattuna jokaisella maahanmuuttoluokalla nettovaikutusten odotusarvo on vähintään 75 %:n todennäköisyydellä suurempi tutkimusaineiston perusteella.
- **Maassaoloaika (1)**. Päävaikutus on odotetusti positiivinen, eli 5 vuoden lisäys maassaolossa (maassaoloajan keskihajonta) nostaa nettovaikutusten odotusarvoa noin 1 200–2 700 eurolla.
- **Maassaoloaika (2)**. Maassaoloajan positiivinen vaikutus on kuitenkin ensimmäisinä maassaolon vuosina (alle 8 vuotta maassa) esitettyä maassaoloajan keskimääräistä vaikutusta suurempi ja sen jälkeen puolestaan huomattavasti pienempi. Tämä johtuu siitä, että interaktion, Maassaolo $\times$ (maassaolo < 8), odotusarvo on yli 2000 euroa. Siten odotusarvoisesti alle 8 vuotta maassa olleilla 5 vuoden lisäys maassaolossa nostaa nettovaikutusten odotusarvoa noin 2 800 eurolla. Kahdeksan vuotta ja pitempään maassa olleilla 5 lisävuotta nostaa odotusarvoa vain noin 400 eurolla.
- **Muut regressiokertoimet**. Jos on asunut maahanmuuttovuonna Uudellamaalla tai pääasiallinen toiminta ollut työllinen, niin nettovaikutusten odotusarvo on positiivisempi kuin vertailuluokissa. Lisäksi mitä enemmän on alle 7-vuotiaita lapsia muuttovuonna, sitä negatiivisempi on nettovaikutusten odotusarvo.
- **Autoregressiiviset termit**. Ensimmäisen autoregressiivinen termin jakauma keskittynyt 0,30:n ympäristöön. Toisen jakauma keskittynyt lähelle nollaa.
- **Satunnaisvaikutukset**. Henkilöittäin vaihtelevan mallin vakiotermin keskihajonnan (τ_{i1}) odotusarvo noin 3 800 euroa, maahanmuuttovuoden (”vakion”) satunnaisvaikutuksen keskihajonnan (τ_{m1}) odotusarvo noin 820 euroa ja maassaoloajan (”vakion”) satunnaisvaikutuksen keskihajonnan (τ_{t1}) odotusarvo noin 320 euroa. Näistä satunnaisvaikutuksista viimeinen – eli maassaoloaikaan liittyvä – aiheuttaa näin odotusarvoisesti näistä kolmesta satunnaisvaikutuksesta eniten kumulatiivisia epävarmuuksia elinkaariennusteisiin, kun uusi maahanmuuttaja on ollut Suomessa yli 11 vuotta.
- **Muut parametrit**. Studentin t-jakauman vapausasteet ovat ”pieniä”, sillä ν-parametrin posteriorijakauma on keskittynyt noin 2,3 ympäristöön. Nettovaikutusten otosjakauma on vasemmalle vino, koska γ on negatiivinen.

4.4.1.3 Irak: posterioriennustetarkistukset

Posterioriennusteet on muodostettu yllä esitetyn mallin 4 (eli elinkaariennusteissa käytettävän) parametrien posteriorijakaumien estimaateista ja havaintoaineistosta niin, kuin on esitetty maassaoloaikojen

mallinnuksen yhteydessä luvussa 3.2.6. Olen laatinut posterioriennusteet niin, kuin kyse olisi uusien maahanmuuttajien aineistosta, joten poimin posterioriennusteita varten myös ennusteet uusien maahanmuuttajien satunnaisvaikutuksille.

Posterioriennustetarkistus: Kaikkien nettovaikutushavaintojen yhteisjakauma ja henkilöiden nettovaikutusten keskiarvojen yhteisjakauma.

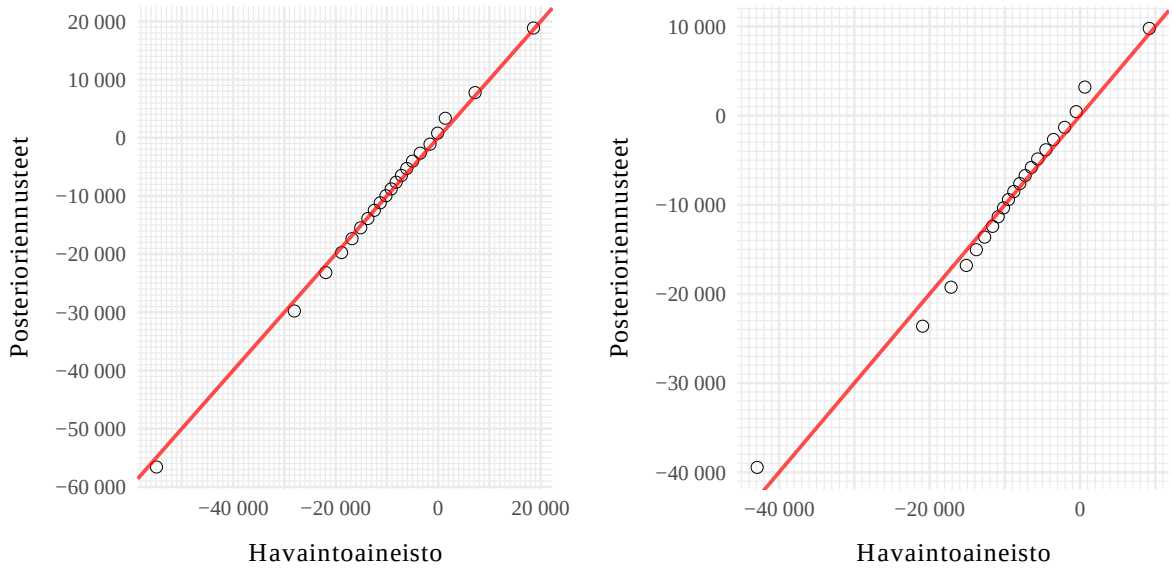
Aloitan posterioriennustetarkistukset vertaamalla havaintoaineiston kaikkien nettovaikutushavaintojen muodostamaa yhteisjakaumaa posterioriennustejakauman kaikkien ennusteiden muodostamaan yhteisjakaumaan. Ensimmäisen mallinnustavoitteen mukaan näiden jakaumien pitäisi vastata mahdollisimman hyvin toisiaan, erityisesti jakaumien tärkeimpien tunnuslukujen kuten keskiarvojen, keskihajontojen ja kvartiilien suhteen.

Tämän lisäksi vertaan kaikkien havaintoaineiston henkilöiden nettovaikutusten keskiarvojen jakaumaa ennustettuihin henkilöiden nettovaikutusten keskiarvojen jakaumaan. Nämä jakaumat muodostetaan havaintoaineistossa laskemalla nettovaikutusten keskiarvot kaikille aineiston henkilöille ja muodostamalla näistä jakauma. Posterioriennusteiden kohdalla lasketaan ennustettujen nettovaikutusten keskiarvot kaikille henkilöille, ja muodostetaan näistä jakauma. Henkilötasoisten jakaumien vertailun tarkoituksena on osoittaa, että elinkaariennusteissa käytettävä nettovaikutusten malli tuottaa luotettavia ennusteita paitsi havaintotasolla niin myös henkilötasolla.

Tarkistus kvantiilikuviolla. Vertaan alla kuviossa 4.4 havaintoaineiston jakaumien prosenttipisteitä vastaavien posterioriennustejakaumien prosenttipisteisiin. Jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste. Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeamat vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).

Kuva 4.4 osoittaa, että posterioriennusteiden tuottamat jakaumat vastaavat hyvin havaintoaineiston jakaumia sekä kaikkien havaintojen osalta että henkilöiden nettovaikutusten keskiarvojen osalta. Kaikkien havaintojen ja ennusteiden vertailussa suurin poikkeama on havaintoaineistossa nollan kohdalla (90. prosenttipiste). Posterioriennusteiden jakauma on tässä noin 2 000 euroa liian suuri, kuten käy ilmi alla esitettävästä taulukosta 5.11. Tämä johtuu havaintojen (suhteellisen pienestä) keskittymästä nollan ympärille. Osa nollan ympärille keskittyneistä havainnoista lienee vuosilta 2010 ja 2011 sellaisilta henkilöiltä, jotka eivät todellisuudessa olleet Suomessa mainittuina vuosina. Johtuen tutkimuksen maassaolokriteeristä ja havaintoaineiston ulottumisesta vuoteen 2011, en ole voinut poistaa tällaisia havaintoja nettovaikutusten havaintoaineistosta. Joka tapauksessa, kuten olen jo edellä kirjoittanut, pidän havaintojen keskittymistä nollan ympäristöön sen verran pienenä, etten lähtenyt mallintamaan sitä erikseen.

Henkilöiden nettovaikutusten keskiarvoissa ei ole samanlaista havaintojen keskittymää nollan tai min-



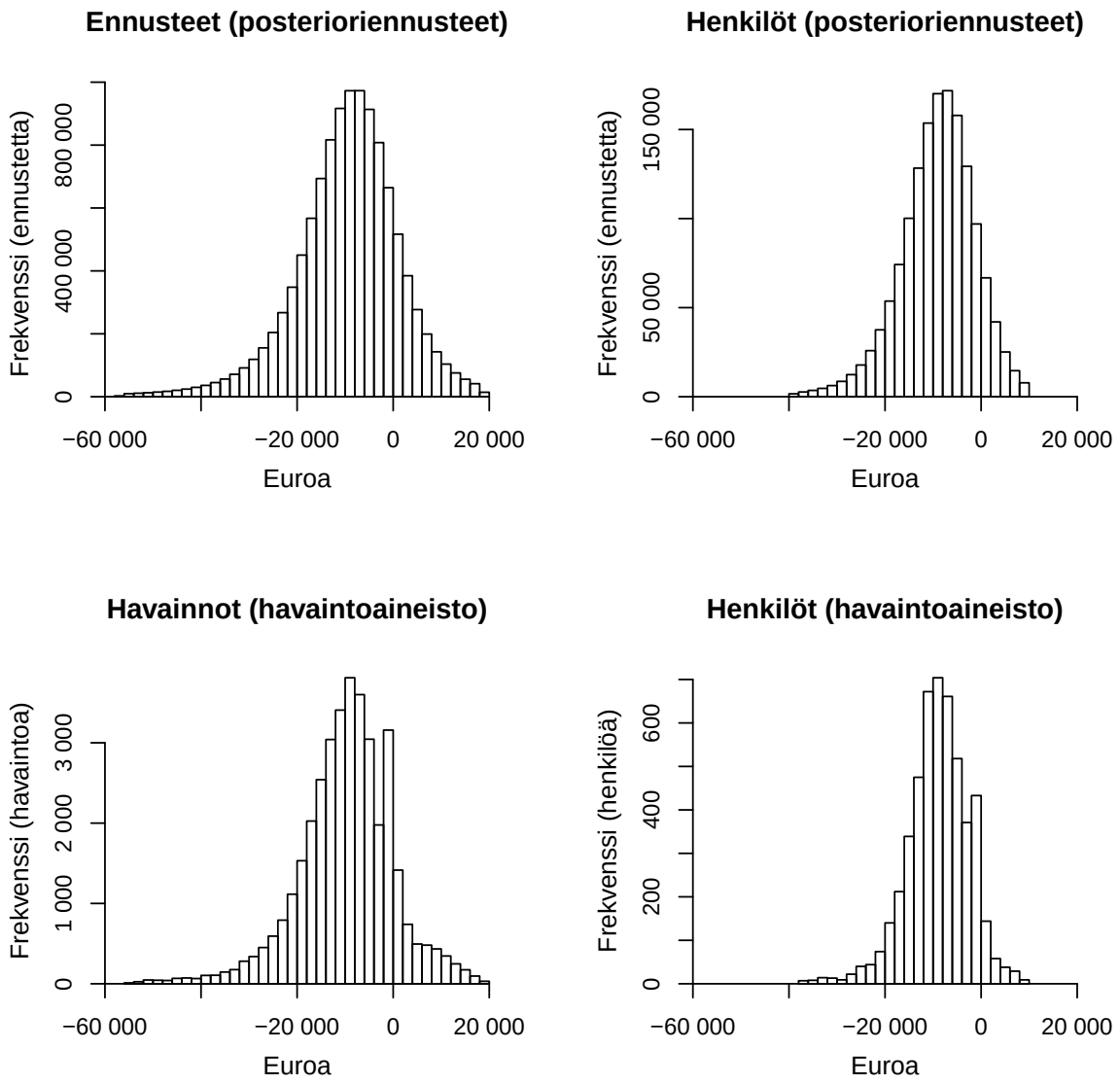
(a) *Kaikki havainnot ja ennusteet.*

(b) *Henkilöiden keskiarvot.*

Kuva 4.4: *Havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

kään muun yksittäisen nettovaikutusten arvon ympärille. Posterioriennusteiden henkilöiden keskiarvojen jakauma vastaa hyvin havaintoaineiston jakaumaa.

Tarkistus histogrammeilla. Verrataan nyt havaintoaineiston ja posterioriennusteiden jakaumia prosenttipisteiden vertailukuvioiden (Q-Q plot) lisäksi histogrammien avulla. Nämä on esitetty kuvassa 4.5. Kuvan ylärivillä on kaksi posterioriennusteista muodostettua jakaumaa, ja alarivillä puolestaan niitä vastaavat havaintoaineiston jakaumat. Histogrammien vertailu osoittaa, että posterioriennustejakaumat vastaavat hyvin havaintoaineiston jakaumia. Kaikkien havaintojen jakaumassa (vasen alakulma) on histogrammissa nähtävissä havaintojen keskittyminen nollan ympäristöön. Kuvasta käy myös ilmi, että henkilöiden keskiarvojen jakauma on kapeampi kuin kaikkien nettovaikutushavaintojen jakauma.



Kuva 4.5: *Posterioriennusteiden ja havaintoaineiston jakaumat histogrammeina. Posterioriennusteiden ennusteet ja ennustetut henkilöiden nettovaikutusten keskiarvot (ylärivi). Havaintoaineiston havainnot ja havaitut henkilöiden nettovaikutusten keskiarvot (alarivi). Kaikista jakaumista on katkaistu pois alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Tarkistus testisuureiden avulla. Alla taulukossa 5.11 on esitetty havaintoaineiston ja posterioriennusteiden jakaumien vertailut euromääräisinä lukuina sekä *Bayes-p-arvojen* ja eri euromääräisten *Bayes-intervallien* avulla. Olen käsitellyt näiden testisuureiden merkitystä ja laskemista alaluvussa 4.3.7. Tau-

lukossa 5.11 verrataan posterioriennustejakaumien (havaintojen ja henkilöiden keskiarvojen jakaumat) keskiarvoja, keskihajontoja sekä joitakin prosenttipisteitä havaintoaineiston vastaaviin jakauman tunnuslukuihin.

Taulukon 5.11 ensimmäisellä ja toisella rivillä ovat posterioriennustejakaumien ja havaintoaineiston jakaumien tunnuslukujen estimaatit. Bayes-p-arvot ja Bayes-intervallit eri euromäärille ovat testisuureita estimaattien vastaavuuksille havaintoaineiston ja posterioriennusteiden välillä. Bayes-intervalli ilmaisee todennäköisyyden sille, että posterioriennustejakauman estimaatti poikkeaa korkeintaan ilmoitetun euromäärän verran havaintoaineiston estimaatista. Esimerkiksi 2 000 euron Bayes-intervallin sisään mahtuvat kaikkien havaintojen jakauman osalta varsin suurella todennäköisyydellä (yli 60 ‰) kaikki jakaumapisteet – ja jos jätetään keskittymä nollassa sekä jakauman vasen häntä huomiotta, niin muut esitetyt prosenttipisteiden estimaatit mahtuvat lähes varmasti 2 000 euron Bayes-intervalliin.

Taulukko 4.6: *Havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien keskiarvot (KA), keskihajonnat (KH) ja prosenttipisteet:* $1, 5, 10, 25, 50, 75, 90, 95, 99$. *Vertailtavien estimaattien Bayes-p-arvot ja Bayes-intervallit eri euromäärillä.*

Kaikki havainnot ja ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %
Posterioriennusteet	-10 110	16 680	-56 590	-29 770	-23 180	-15 470	-8 800	-2 650	3 350	7 760	18 850
Aineiston havainnot	-10 250	16 410	-54 950	-28 050	-21 920	-15 100	-9 160	-3 490	1 390	7 200	18 590
Bayes-p-arvo	0,71	0,67	0,07	0	0	0,11	0,94	1	1	0,92	0,62
Bayes-voima: 2 000 euroa	1	1	0,65	0,72	0,96	1	1	1	0,61	1	0,97
Bayes-voima: 1 000 euroa	1	0,91	0,31	0,07	0,28	0,97	1	0,73	0	0,86	0,74
Bayes-voima: 500 euroa	0,92	0,62	0,14	0	0,01	0,69	0,72	0,07	0	0,48	0,45
Bayes-voima: 250 euroa	0,6	0,33	0,07	0	0	0,36	0,31	0	0	0,21	0,2

Henkilöiden keskiarvot ja niiden ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %
Posterioriennusteet	-9 400	10 910	-39 450	-23 620	-19 250	-13 630	-8 510	-3 830	440	3 200	9 780
Aineiston havainnot	-9 470	11 620	-42 920	-20 930	-17 150	-12 680	-8 810	-4 560	-520	630	9 190
Bayes-p-arvo	0,61	0,19	0,97	0	0	0	0,83	0,97	0,99	1	0,76
Bayes-voima: 2 000 euroa	1	0,89	0,2	0,08	0,42	1	1	0,99	0,99	0,09	0,97
Bayes-voima: 1 000 euroa	1	0,41	0,07	0	0	0,61	0,99	0,79	0,56	0	0,77
Bayes-voima: 500 euroa	0,88	0,18	0,05	0	0	0,04	0,72	0,26	0,11	0	0,45
Bayes-voima: 250 euroa	0,58	0,11	0,02	0	0	0,01	0,37	0,11	0,03	0	0,2

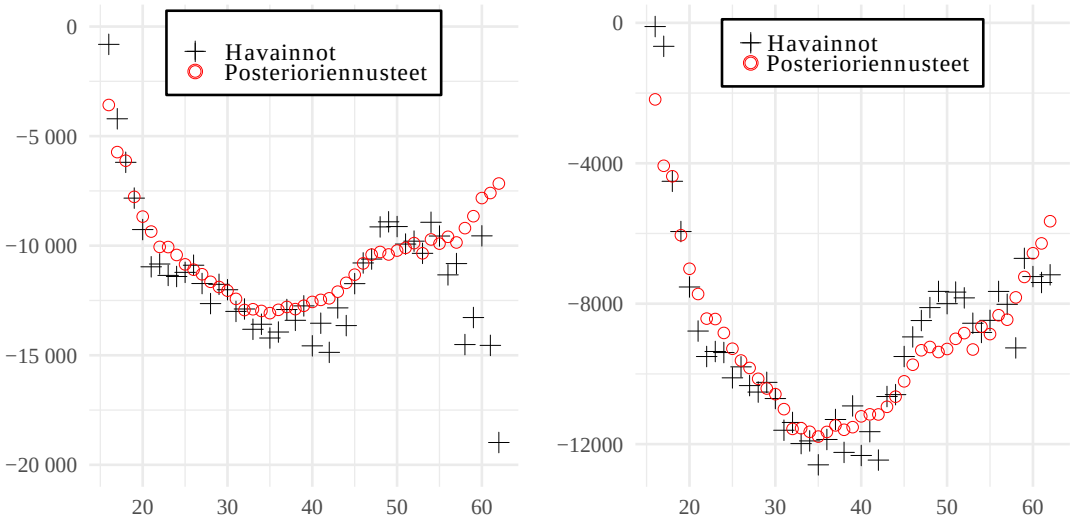
Posterioriennustarkistukset selittävien muuttujien luokkien mukaan.

Taulukon 5.11 jälkeen vertaan posterioriennusteita havaintoaineistoon tärkeimpien selittävien muuttujien luokkien mukaan. Tärkeimpiä selittäviä muuttujia ovat sukupuoli sekä aikaan liittyvät muuttujat: ikä, maahanmuuttovuosi, maassaoloaika ja kalenterivuosi. Viimeinen näistä muuttujista ei ollut lainkaan mallissa, koska maahanmuuttovuosi, maassaoloaika ja kalenterivuosi eivät voi olla samaan aikaan lineaarisessa mallissa.

Esitän posterioriennustetarkistuksen tästä huolimatta myös kalenterivuoden suhteen. Esitän posterioriennustetarkistukset maassaoloajan suhteen erikseen vuonna 1992 ja sitä myöhemmin Suomeen muuttaneille sekä ennen vuotta 1992 Suomeen muuttaneille. Syynä tähän on se, että nettovaikutusten havaintojen aineistotarkastelun ja estimointitulosten perusteella ennen vuotta 1992 Suomeen muuttaneilla Irakissa syntyneillä on ollut positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla – senkin jälkeen, kun maassaoloaika huomioidaan.

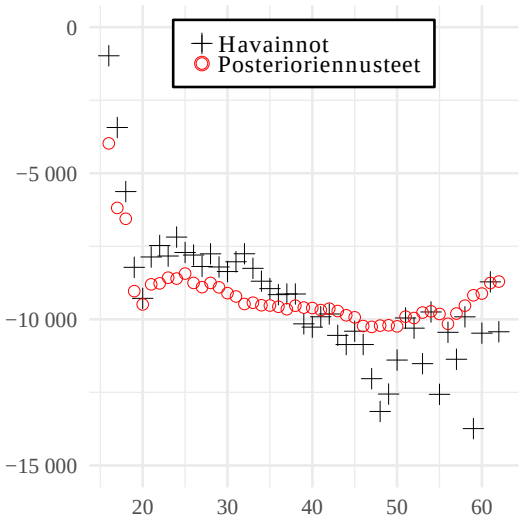
Vertaan posterioriennusteiden keskiarvoja ja mediaaneja selittävien luokkien mukaan vastaaviin havaintoaineiston keskiarvoihin ja mediaaneihin. Esimerkiksi iän kohdalla vertaan keskiarvoja ja mediaaneja niin, että keskiarvot ja mediaanit on laskettu posterioriennusteista jokaiselle mallinnusaineiston iälle erikseen 16-vuotiaista 62-vuotiaisiin. Keskiarvojen lisäksi esitän mediaanit, koska nettovaikutusten jakauman leveiden häntien vuoksi mediaani on oletettavasti keskiarvoa robustimpi jakaumien keskiluku. Kuvioiden jälkeen esitän jokaiseen selittävien muuttujien luokkien suhteen tehtyyn posterioriennustetarkistukseen liittyen tärkeimpiä huomioita.

Posterioriennustetarkistus iän ja sukupuolen mukaan.

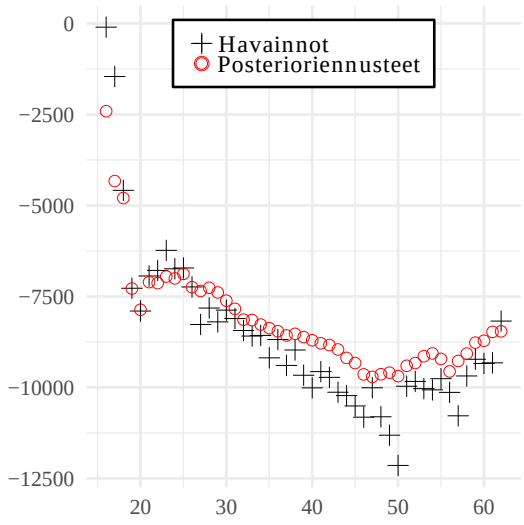


(a) *Keskiarvot iän mukaan. Naiset.*

(b) *Mediaanit iän mukaan. Naiset.*



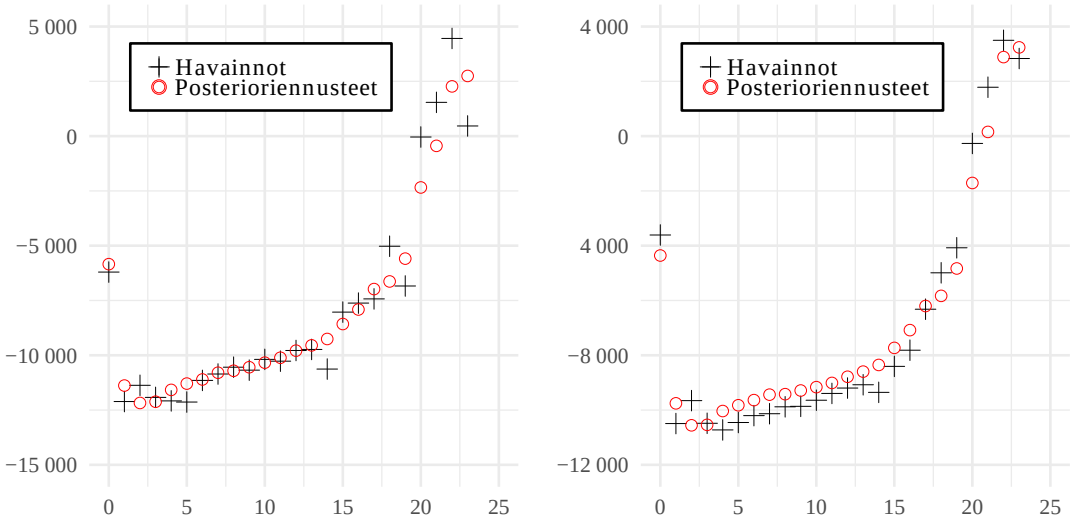
(c) *Keskiarvot iän mukaan. Miehet.*



(d) *Mediaanit iän mukaan. Miehet.*

Kuva 4.6: *Havainnot ja posterioriennusteet iän ja sukupuolen mukaan.*

Posteriorienustetarkistus maassaoloajan mukaan.

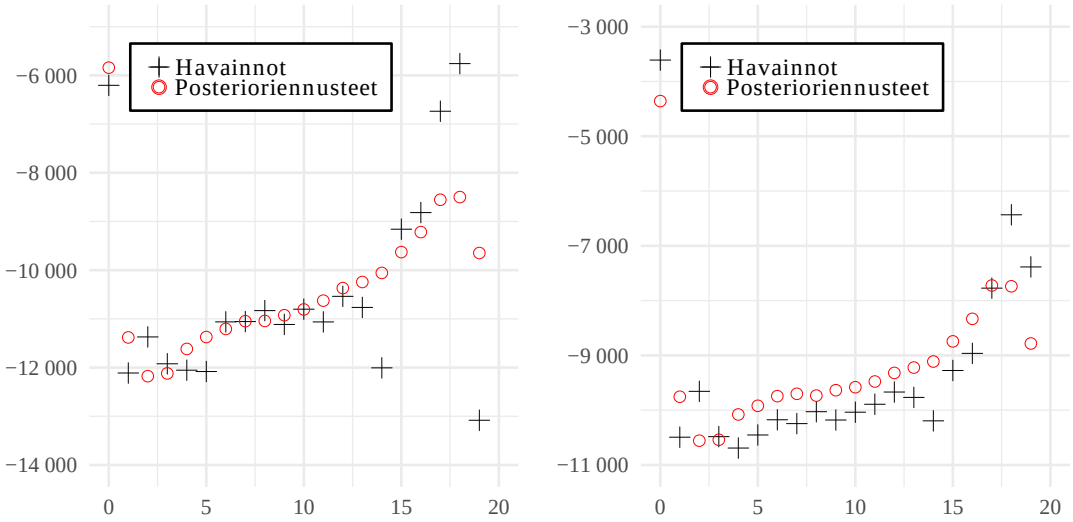


(a) *Keskiarvot maassaoloajan mukaan.*

(b) *Mediaanit maassaoloajan mukaan.*

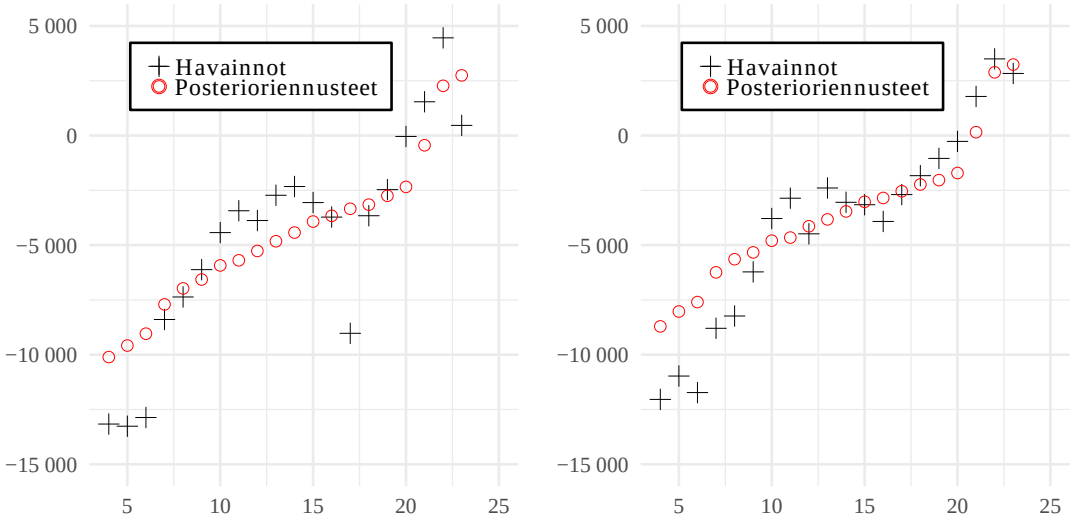
Kuva 4.7: *Havainnot ja posteriorienusteet maassaoloajan (vuosia) mukaan.*

Posterioriennustetarkistus maassaoloajan ja maahanmuuttovuosiluokan mukaan. Maassaoloaika vuonna 1992 ja myöhemmin muuttaneilla verrattuna ennen vuotta 1992 muuttaneisiin.



(a) *Keskiarvot maassaoloajan mukaan. Vuonna 1992 ja myöhemmin Suomeen muuttaneet.*

(b) *Mediaanit maassaoloajan mukaan. Vuonna 1992 ja myöhemmin Suomeen muuttaneet.*

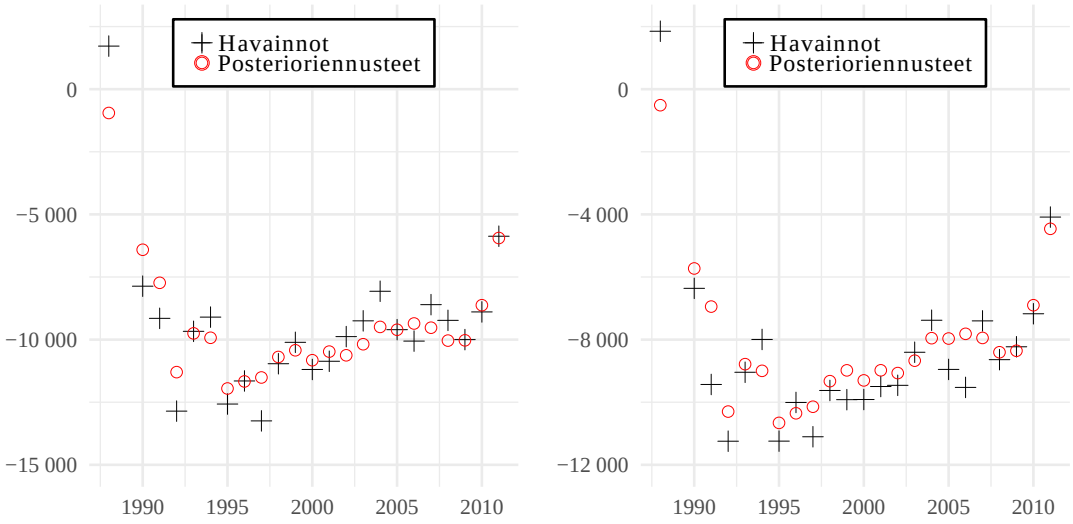


(c) *Keskiarvot maassaoloajan mukaan. Ennen vuotta 1992 Suomeen muuttaneet.*

(d) *Mediaanit maassaoloajan mukaan. Ennen vuotta 1992 Suomeen muuttaneet.*

Kuva 4.8: *Havainnot ja posterioriennusteet maassaoloajan ja maahanmuuttovuosiluokan mukaan.*

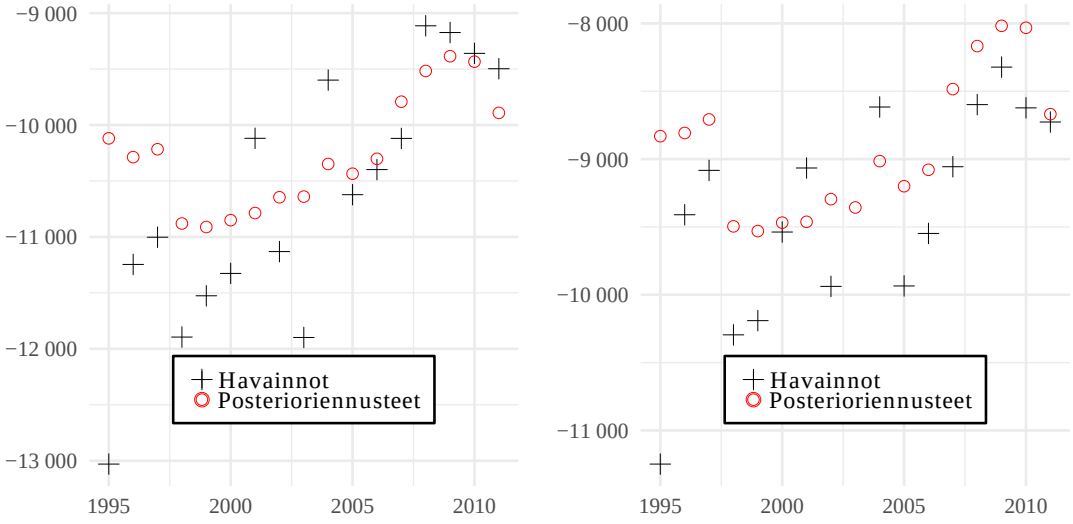
Posterioriennustetarkistus maahanmuuttovuoden mukaan.



(a) *Keskiarvot maahanmuuttovuoden mukaan.* (b) *Mediaanit maahanmuuttovuoden mukaan.*

Kuva 4.9: *Havainnot ja posterioriennusteet maahanmuuttovuoden mukaan.*

Posterioriennustetarkistus kalenterivuoden mukaan.



(a) *Keskiarvot kalenterivuoden mukaan.* (b) *Mediaanit kalenterivuoden mukaan.*

Kuva 4.10: *Havainnot ja posterioriennusteet kalenterivuoden mukaan.*

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan. Havaintoja kuvista 4.6–4.10.

- **Yleistä (1)**. Posterioriennusteiden mediaanit näyttävät vastaavan havaintoaineiston mediaaneita paremmin kuin keskiarvojen osalta – sekä kaikkien havaintojen osalta (taulukko 5.11) että selittävien muuttujien luokkien mukaan tarkasteltuna. Erityisesti tämä tulee esille iän ja sukupuolen mukaan esitetyssä posterioriennustetarkastuksessa (kuva 4.6), jossa varsinkin noin 60-vuotiailla naisilla posterioriennusteiden mediaanit vastaavat hyvin havaintoaineiston mediaaneita, toisin kuin keskiarvot (vrt. kuvan 4.6 osakuvat (a) ja (b)).

- **Yleistä (2)**. Kuvat 4.6–4.10 osoittavat, että yleisesti ottaen posterioriennusteiden keskiarvot ja mediaanit vastaavat hyvin havaintoaineiston keskiarvoja ja mediaaneita, kun vertailu tehdään tärkeimpien selittävien muuttujien luokkien suhteen.

- **Ikä sukupuolen mukaan** (kuva 4.6). Naisilla nettovaikutukset vaihtelevat ”paljon” iän mukaan siten, että ikävuosikohtaisten nettovaikutusten keskiarvojen ja mediaanien minimi on noin 35–40-vuotiaana. Noin 60-vuotiailla naisilla posterioriennusteiden ikävuosikohtaiset mediaanit vastaavat havaintojen mediaaneja paremmin kuin posterioriennusteiden keskiarvot havaintoaineiston keskiarvoja. Miehillä iällä ei vaikuta kuvan 4.6 perusteella olevan niin suurta vaikutusta nettovaikutusten tasoon kuin naisilla.

- **Maassaoloaika (kaikki)** (kuva 4.7). Nettovaikutusten keskiarvoissa ja mediaaneissa maassaoloajan suhteen on nähtävissä havaintojen perusteella kuvassa 4.7 neljä periodia: ensimmäinen maassaolovuosi (0 maassaolovuotta), alle 5 vuotta maassa (1–4 maassaolovuotta), 5–15 maassaolovuotta ja yli 15 maassaolovuotta. Ensimmäisenä vuonna nettovaikutukset ovat keskimäärin positiivisemmat kuin 1–4 maassaolovuonna. Maassaolovuosina 5–15 vuotta nettovaikutukset ovat positiivisemmat kuin ensimmäisinä maassaolovuosina, yli 15 maassaolovuoden jälkeen nettovaikutukset näyttävät muuttuvan nopeasti positiivisemmiksi. Posterioriennusteet seuraavat tätä kehitystä melko hyvin sekä keskiarvojen että mediaanien osalta.

- **Maassaoloaika (v. 1992 ja myöhemmin muuttaneet, ennen v. 1992 muuttaneet) (1)** (kuva 4.8). Keskiarvojen ja mediaanien nopea nousu 15 maassaolovuoden jälkeen (kuva 4.7) johtuu ainakin osittain siitä, että ennen vuotta 1992 muuttaneilla on positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla. Erityisesti 1980-luvulla muuttaneilla Irakissa syntyneillä on keskimäärin tuhansia euroja positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla. Tämä johtuu osaltaan suoraan siitä, että 1980-luvulla muuttaneilla pääsiallinen toiminta oli noin puolella väestöstä työllinen, kun myöhemmin muuttaneilla se on ollut vain muutaman prosentin (koko mallinnusväestön keskiarvo on 4 prosenttia). Tämän taustalla puolestaan on nähdäkseni se, että vielä 1980-luvulla Irakista on muutettu Suomeen muista kuin nk. humanitaarisista syistä (ml. perheenyhdistämiset), toisin kuin 1990-luvun alusta lähtien.

- **Maassaoloaika (v. 1992 ja myöhemmin muuttaneet, ennen v. 1992 muuttaneet) (2)** (kuva 4.8). Kuten kuvasta 4.8 ilmenee, ennen vuotta 1992 muuttaneilla (osakuvat (c) ja (d)) keskimääräiset nettovaikutukset ovat olleet maassaoloajan mukaan tarkasteltuina monta tuhatta euroa positiivisemmat kuin myöhemmin muuttaneilla, ainakin kun tarkastellaan muita kuin muutamaa aivan ensimmäista

maassaolovuotta. Taulukon 4.5 tulosten mukaan maassaoloaika vaikuttaakin positiivisemmin nettovaikutusten kehitykseen ennen vuotta 1992 muuttaneilla kuin myöhemmin muuttaneilla. Lisäksi ennen vuotta 1992 muuttaneista suurempi osuus on ollut työllisiä ensimmäisenä maassaolovuonna Suomessa kuin myöhemmin muuttaneista, ja työllisyys maahanmuuttovuonna nostaa nettovaikutusten odotusarvoa useilla tuhansilla euroilla.

- **Maahanmuuttovuosi (1)** (kuva 4.9). Käsittelin yllä maassaolovuosien vaikutusten yhteydessä jo maahanmuuttovuoden vaikutusten nettovaikutuksiin. Kuvasta 4.9 käy selvästi ilmi, kuinka 1980-luvulla muuttaneilla sekä nettovaikutusten keskiarvo ja mediaani on selvästi positiivisempi kuin myöhemmin muuttaneilla. Edelleen vuosina 1990–1994 muuttaneilla vaikuttaa olevan positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla, vaikka maassaoloaika otettaisiin huomioon, kuten käy ilmi edellä käsitellyistä maassaoloajan kuvioista. Ainakin 1980-luvulla muuttaneiden osalta vaikuttaa selvältä, että positiivisemmat nettovaikutukset johtuvat erilaisesta Irakissa syntyneiden maahanmuuton rakenteesta, eli siitä, että 1980-luvulla Irakissa syntyneet muuttivat Suomeen muista kuin humanitaarisista syistä. Posterioriennusteiden keskiarvot ja mediaanit vastaavat joka tapauksessa hyvin havaintoaineiston vastaavia tunnuslukuja maahanmuuttovuoden suhteen. Koska vuonna 1992 muuttaneilla näyttää olevan keskimäärin negatiivisemmat nettovaikutukset kuin vuosina 1993–1994, liitin vuonna 1992 muuttaneet samaan maahanmuuttovuosien luokkaan kuin 1995–1997 Suomeen ensimmäisen kerran muuttaneet.

- **Maahanmuuttovuosi (2)** (kuva 4.9). Vuonna 2011 muuttaneilla nettovaikutukset ovat positiivisemmat, koska heitä koskien on olemassa vain ensimmäisen vuosi maassa, jolloin nettovaikutukset ovat positiivisemmat (koska keskimäärin Suomessa ollaan vain osan kalenterivuotta maahanmuuttovuonna). Sama koskee lievemmin esimerkiksi vuonna 2010 muuttaneita, joilla noin puolet havainnoista koskevat maahanmuuttovuotta. Posterioriennusteiden keskiarvot ja mediaanit vastaavat joka tapauksessa hyvin havaintoaineiston vastaavia tunnuslukuja maahanmuuttovuoden suhteen.

- **Kalenterivuosi** (kuvio 4.10). Kalenterivuosien osalta erityisesti ennusteiden mediaanit vastaavat melko hyvin havaintoaineiston mediaaneita. Nettovaikutusten taso kalenterivuoden suhteen vaihtelee Irakissa syntyneiden kohdalla ainakin sen mukaan, kuinka pitkään Suomessa olleista Irakissa syntyneiden väestö koostuu kunakin kalenterivuonna. Mikäli väestössä on suhteellisesti enemmän juuri Suomeen muuttaneita, nettovaikutukset ovat tuolloin negatiivisemmat kuin sellaisina vuosina, jolloin juuri muuttaneiden osuus on pienempi – kun pidetään muut vaikuttavat tekijät vakioina.

4.4.2 Somalia

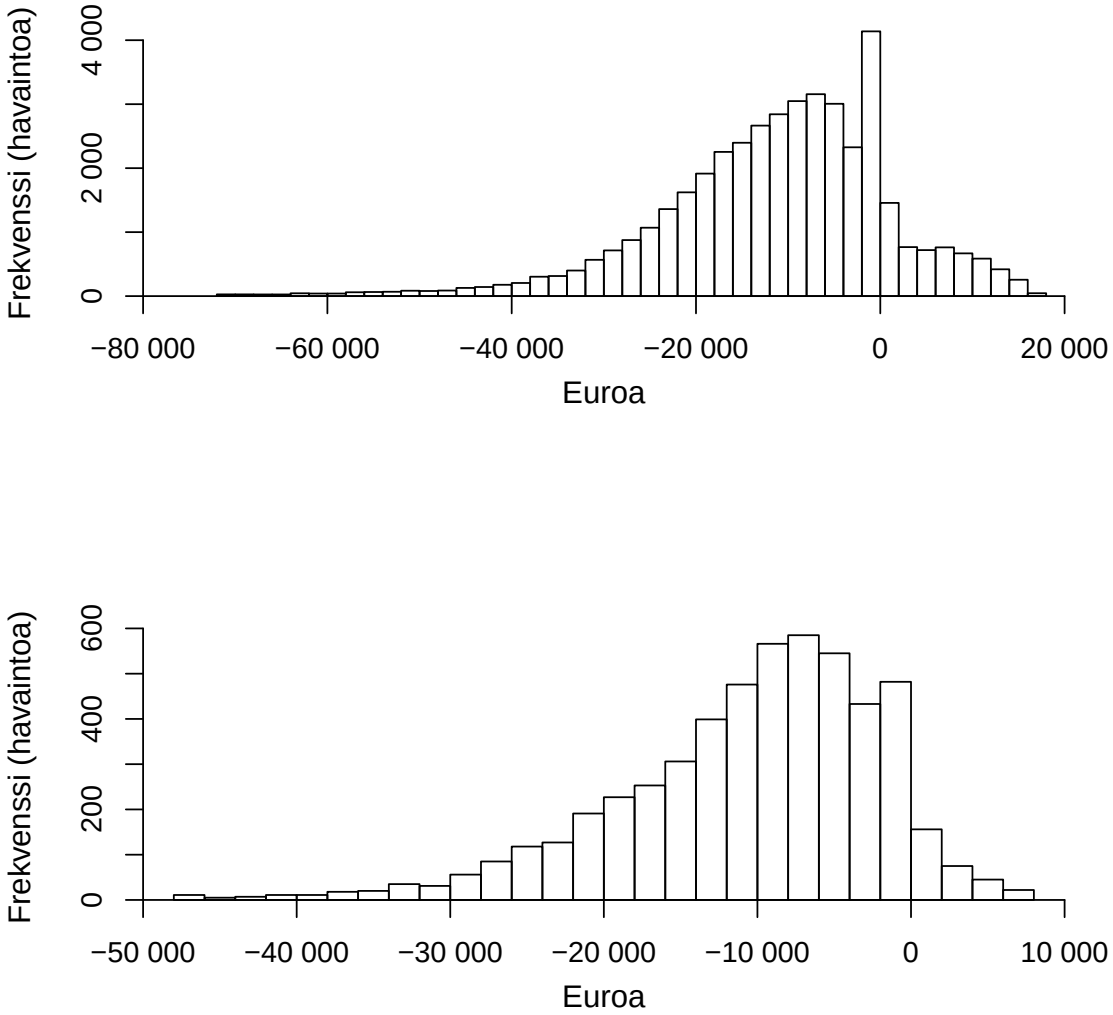
4.4.2.1 Somalia: havaintoaineisto

Taulukossa 4.7 on esitetty Somaliassa syntyneiden 16–62-vuotiaiden nettovaikutushavaintojen lukumäärä ja havaintojen jakauman keskeisimmät tunnusluvut. Taulukossa on esitetty samat tiedot myös havaintoaineiston henkilöiden nettovaikutushavaintojen keskiarvoille. Mallinnusaineistoon ei sisälly alle 16-vuotiaana Suomeen muuttaneita. Nämä nettovaikutushavainnot muodostavat selitettävän muuttujan nuorten ja työikäisten julkisen talouden nettovaikutusten mallinnuksessa. Taulukon 4.7 tietojen mukaan esimerkiksi kaikkien havaintojen (42 835 kappaletta) keskiarvo on -12 000 euroa ja mediaani -9 800 euroa. Havaintoaineiston henkilöiden (5 412 henkilöä) nettovaikutushavaintojen keskiarvojen mediaani on -9 100 euroa ja keskihajonta 14 000 euroa.

Taulukko 4.7: *Somaliassa syntyneet 16–62-vuotiaat. Nettovaikutushavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa sekä henkilöiden lukumäärä ja henkilötasoisten nettovaikutushavaintojen keskiarvojen jakauman tunnusluvut mallinnusaineistossa. Euromääräiset tunnusluvut pyöristetty lähimpään sataan euroon.*

	Havainnot	Henkilöt
Kokonaislukumäärä (N)	42 835	5 412
Keskiarvo	-12 000	-11 200
Keskihajonta	19 300	14 000
1. prosenttipiste (1 %)	-72 200	-51 600
5. prosenttipiste (5 %)	-35 400	-27 800
10. prosenttipiste (10 %)	-27 400	-22 900
Alakvartiili (25 %)	-18 200	-15 400
Mediaani (50 %)	-9 800	-9 100
Yläkvartiili (75 %)	-2 400	-4 300
90. prosenttipiste (90 %)	2 900	-200
95. prosenttipiste (95 %)	8 800	800
99. prosenttipiste (99 %)	16 400	7 000

Kuvassa 4.11 on esitetty histogrammit sekä kaikkien nettovaikutushavaintojen jakaumalle (ylempi histogrammi) että henkilöiden nettovaikutusten keskiarvojen jakaumalle (alempi histogrammi). Kummastakin jakaumasta on katkaistu pois niiden ylin ja alin prosentti pois. Kaikkien havaintojen histogrammista näkee, että jakauma on vasemmalle vino. Lisäksi siitä näkee, että osa havainnoista on kertynyt nollan lähelle (pylväs heti nollan vasemmalla puolella). Henkilöiden keskiarvojen jakauma näyttää histogrammin perusteella vasemmalle vinolta. Keskiarvojen osalta nollan ympäristössä ei ole ainakaan yhtä selvää havaintojen keskittymää kuin kaikkien havaintojen jakaumassa.



Kuva 4.11: *Somaliassa syntyneet 16–62-vuotiaat. Nettovaikutushavaintojen katkaistun jakauman histogrammi mallinnusaineistossa (yllä) sekä henkilöiden nettovaikutushavaintojen keskiarvojen katkaistun jakauman histogrammi (alla). Jakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Taulukossa 4.8 on esitetty Somaliassa syntyneiden nettovaikutusmallien selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden kaikkien havaintojen joukossa. Edellä parametrien selitetaulukossa (s. 170) on lyhyitä selitteitä selittävistä muuttujista.

Taulukko 4.8: *Somaliassa syntyneet 16–62-vuotiaat. Nettovaikutusmallien selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta).*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,48	0,50	0	0	1
Ikä	34,0	10,1	26	33	41
Ikä^2	1254	744	676	1089	1681
Alle 19-vuotias	0,04	0,19	0	0	0
Yli 60-vuotias	0,01	0,09	0	0	0
Ikä $\times$ nainen	16,0	18,1	0	0	31
Ikä^2 $\times$ nainen	586	801	0	0	961
Muuttovuosi $<$ 1993	0,30	0,46	0	0	1
Muuttovuosi v. 1993–1996	0,27	0,44	0	0	1
Muuttovuosi v. 1997–2001	0,18	0,39	0	0	0
Muuttovuosi v. 2002–2006	0,12	0,32	0	0	0
Muuttovuosi v. 2007–2011	0,13	0,34	0	0	0
Muuttovuoden työttömyys-$\%$	11,0	3,5	7,8	11,4	14,6
Maassaolo (vuosia)	7,10	5,36	3	6	11
Maassaolo $\times$ nainen	3,14	4,80	0	0	5
Maassaolo $\times$ (muuttovuosi $<$ 1993)	3,26	5,66	0	0	5
Maassaolo $<$ 5	0,39	0,49	0	0	1
(Maassaolo $<$ 5)$\times$ nainen	0,20	0,40	0	0	0
Maassaolo $\times$(maassaolo $<$ 5)	0,74	1,29	0	0	1
Maassaolo $\times$ (maassaolo $<$ 5) $\times$ nainen	0,38	0,99	0	0	0
Maassaolo $>$ 15	0,09	0,29	0	0	0
(Maassaolo $>$ 15)$\times$ nainen	0,03	0,17	0	0	0
Maassaolo $\times$ maassaolo $>$ 15	1,59	5,05	0	0	0
(Maassaolo $\times$) maassaolo $>$ 15 $\times$ nainen	0,52	2,97	0	0	0
Maassaolo alle 1 vuosi	0,09	0,29	0	0	0
Muutto alle 20-v.	0,22	0,42	0	0	0
Muutto 20–29-v.	0,49	0,50	0	0	1
Muutto 30–39-v.	0,18	0,38	0	0	0
Muutto yli 39-v.	0,11	0,31	0	0	0
Muutto alle 20-v. $\times$ maassaolo	1,36	3,46	0	0	0
Muutto 20–29-v. $\times$ maassaolo	3,94	5,56	0	0	7
Muutto 30–39-v. $\times$ maassaolo	1,27	3,58	0	0	0
Muutto yli 39-v.$\times$ maassaolo	0,53	2,07	0	0	0
Kalenterivuosi	2004	5,0	2000	2005	2009
Kalenterivuosi < 2001	0,27	0,44	0	0	1

Jatkuu seuraavalla sivulla

Taulukko 4.8 – jatkuu edelliseltä sivulta

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Pääasiallinen toiminta: työllinen	0,02	0,13	0	0	0
Pääasiallinen toiminta: opiskelija	0,09	0,29	0	0	0
Pääasiallinen toiminta: muu	0,89	0,31	1	1	1
Korkein koulutus: 2. asteen koulutus	0,06	0,23	0	0	0
Korkein koulutus: 3. asteen koulutus	0,01	0,12	0	0	0
Korkein koulutus: Ei 2.astetta/Ei tietoa	0,93	0,26	1	1	1
Alle 7-vuotiaiden lasten lukumäärä	0,46	0,91	0	0	1
Puoliso	0,37	0,48	0	0	1
Kotikunta Uudellamaalla	0,59	0,49	0	1	1
Ikä > 55	0,03	0,18	0	0	0
(Ikä > 55) $\times$ ikä	1,9	10,3	0	0	0
(Ikä < 40) $\times$ ikä $\times$ nainen	10,2	14,2	0	0	25
(Ikä < 40) $\times$ ikä	0,36	0,48	0	0	1
(Ikä > 50) $\times$ ikä $\times$ nainen	1,9	10,2	0	0	0
(Ikä > 50) $\times$ ikä	0,04	0,18	0	0	0
(Ikä < 19) $\times$ ikä $\times$ nainen	0,3	2,3	0	0	0
(Ikä < 19) $\times$ ikä	0,02	0,13	0	0	0

4.4.2.2 Somalia: estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Taulukossa 4.9 on esitetty perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit Somaliassa syntyneille 16–62-vuotiaille. Olen käsitellyt viiden perättäisen mallin rakentamista mallinnusaskel kerrallaan alaluvussa 4.3.3. Estimoitavien parametrien selitteet on puolestaan esitetty parametrien selitetaulukossa s. 170.

Mallin vakiotermiä, regressiokertoimia, satunnaisvaikutusten ja virhetermin keskihajontoja (τ-parametrit, σ) sekä vinousparametreja koskevat odotusarvoestimaatit ovat euromääräisiä. Autoregressiiviset kertoimet on ilmaistu prosentteina, ja t-jakauman vapausasteilla (ν) ei puolestaan ole mitään erillistä yksikköä. En käsittele tässä yhteydessä tarkemmin taulukossa 4.9 esitettyjen parametrien odotusarvojen tulkintaa, sillä olen tehnyt tämän käytännössä jo Irakissa syntyneiden nettovaikutusmallinnustulosten yhteydessä (ks. taulukko 4.4, s. 177, sekä taulukon ylä- ja alapuolella oleva teksti koskien taulukon tulosten tulkintaa).

Taulukon 4.9 mallien ennustetarkkuuden mittarien (Waic ja LOO) mukaan parametrimäärältään suuremmat mallit ovat parempia kuin pienemmät mallit. Ennustarkkuuksien mittareiden Waic ja LOO estimaatit ovat lähellä toisiaan kolmen ensimmäisen mallin osalta. Mallien 4 ja 5 osalta, joissa otosjakaumana on vino Studentin t-jakauma, arvot kuitenkin poikkeavat toisistaan, kuten ne poikkesivat myös Irakissa

syntyneiden estimointituloksissa.

Taulukko 4.9: *Perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit, mallikohtaiset ennustetarkkuuden estimaatit ja keskihajonnat (Waic ja LOO) sekä sisäkorrelaatioestimaatti. Mallien valinta on esitetty luvussa 4.3.3.*

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
Vakio	-11 540	-9 350	-11 270	-9 560	-9 400
Nainen			-10 580	-13 730	-12 220
Ikä			-2 800	-3 380	-3 300
Ikä^2			2 480	2 820	2 760
Alle 19-vuotias			670	650	620
Yli 60-vuotias			50	140	140
Ikä $\times$ nainen			-6 630	-6 560	-6 110
Ikä^2 $\times$ nainen			7 770	7 820	7 300
Muuttovuosi $<$ 1993			880	1 130	1 920
Muuttovuosi v. 1993–1996			690	520	960
Muuttovuosi v. 1997–2001			1 350	1 510	1 960
Muuttovuosi v. 2002–2006			770	320	440
Muuttovuoden työttömyys-$\%$			30	50	50
Maassaolo (vuosia)			1 920	1 350	880
Maassaolo $\times$ nainen			-750	-2 180	-2 100
Maassaolo $\times$ (muuttovuosi $<$ 1993)			1 070	1 660	1 680
Maassaolo $<$ 5			1 660	40	140
(Maassaolo $<$ 5) $\times$ nainen			20	-280	-170
Maassaolo $\times$ (maassaolo $<$ 5)			3 500	850	940
Maassaolo $\times$ (maassaolo $<$ 5) $\times$ nainen			-1 780	-3 270	-3 170
Maassaolo $>$ 15			1 830	3 750	3 810
(Maassaolo $>$ 15) $\times$ nainen			-2 750	-200	-170
Maassaolo $\times$ (maassaolo $>$ 15)			-1 030	-2 080	-2 090
Maassaolo $\times$ (maassaolo $>$ 15) $\times$ nainen			1 370	170	190
Maassaolo alle 1 vuosi			9 650	7 190	7 090
Muutto alle 20-v.			1 460	2 920	2 800
Muutto 20–29-v.			930	1 690	1 580
Muutto 30–39-v.			1 120	1 860	1 820
Muutto alle 20-v. $\times$ maassaolo			800	2 150	2 130
Muutto 20–29-v. $\times$ maassaolo			410	1 070	1 050
Muutto 30–39-v. $\times$ maassaolo			530	1 180	1 180
Kalenterivuosi < 2001			-40	-470	-500
Pääasiallinen toiminta: työllinen			2 700	3 800	3 760
Pääasiallinen toiminta: opiskelija			-780	-1 380	-1 400

Jatkuu seuraavalla sivulla

Taulukko 4.9 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
Korkein koulutus: 2. asteen koulutus			10	20	50
Korkein koulutus: 3. asteen koulutus			1 900	2 060	2 020
Alle 7-vuotiaiden lasten lukumäärä			-620	-980	-980
Puoliso			240	390	390
Kotikunta Uudellamaalla			590	1 200	1 200
Ikä > 55			3 030	2 140	1 750
(Ikä > 55) $\times$ ikä			-1 330	-810	-610
(Ikä < 40) $\times$ ikä $\times$ nainen			1 170	-120	-380
(Ikä > 50) $\times$ ikä $\times$ nainen			-6 600	-4 970	-4 760
(Ikä < 19) $\times$ nainen			7 940	15 790	15 260
(Ikä < 40) $\times$ nainen			-640	230	390
(Ikä > 50) $\times$ nainen			11 330	8 550	8 220
(Ikä < 19) $\times$ ikä $\times$ nainen			5 170	10 240	9 930
Nettovaikutus$_{t-1}$			46,1	22,4	21,9
Nettovaikutus$_{t-2}$			12,2	5,4	5,4
τ_{i1}	12 120	7 660	2 600	4 750	4 810
τ_{i2}				3 170	3 200
τ_{i3}				270	290
τ_{m1}				570	580
τ_{m2}				280	270
τ_{t1}				140	130
τ_{t2}				250	250
σ	15 040	5 950	5 200	4 850	4 830
ν		2,22	2,1	2,8	2,84
γ/Vinous: vakio				-1 550	-1 640
Vinous: nainen					-780
Vinous: ikä					-30
Vinous: ikä^2					-30
Vinous: muuttovuosi $<$ 1993					-500
Vinous: muuttovuosi v. 1993–2002					-330
Vinous: maassaolo (vuosia)					280
Vinous: (muuttovuosi $<$ 1993)$\times$maassaolo					-40
Vinous: maassaolo $<$ 5					-170
Vinous: (maassaolo $<$ 5)$\times$maassaolo					-190
Sisäkorrelaatio	0,446				
Waic (estimaatti)	104 707	67 583	56 059	35 654	34 931
Waic (estimaatin keskivirhe)	(2 540)	(495)	(511)	(337)	(338)
LOO (estimaatti)	104 415	67 654	56 063	41 343	40 587
LOO (estimaatin keskivirhe)	(2 434)	(495)	(511)	(380)	(379)

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” malli on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän uusien maahanmuuttajien elinkaariennusteissa nuorten ja työikäisten nettovaikutusten ennustamiseen Somaliassa syntyneiden osalta. Valitsin lopulliseksi malliksi mallin 4, vaikka ennustetarkkuuden mittareiden puolesta malli 5 olisikin ollut parempi. Mallin 5 osalta ongelmallista olisi elinkaariennusteiden osalta oletus siitä, että maassaoloajan vaikutus jatkuu yli 23 maassaolovuoden jälkeen elinkaariennusteissa samansuuruisena kuin maassaoloajoilla, jotka ovat 23 vuotta pienempiä.

Kuten esitin Irakissa syntyneiden mallinnustulosten yhteydessä, Irakissa syntyneiden osalta tämä oletus olisi tarkoittanut mallin 5 tuottamisen ennusteiden osalta, että ennustejakauman vinous vasemmalle olisi kasvanut rajatta – ja samalla ennusteiden odotusarvo muuttunut negatiivisemmiksi. Koska on vaikeaa perustella, että ennustejakauman vinous vasemmalle kasvaisi rajatta todellisuudessa – ja koska tämä tuottaisi siten epärealistisen negatiivisia ennusteita Irakissa syntyneille uusille maahanmuuttajille – päätin käyttää elinkaariennusteissa Irakissa syntyneiden osalta nettovaikutusten ennustamiseen mallia 4.

Somaliassa syntyneiden osalta mallin 5 käyttö johtaisi – päinvastoin kuin Irakissa syntyneillä – siihen että ennustejakauman vinous kasvaisi rajatta *oikealla* – ja siten ennustejakauman odotusarvo muuttuisi positiivisemmaksi vinouden muutoksen johdosta. Koska tässäkin tapauksessa on vaikeaa perustella, että ennustejakauman vinous oikealle kasvaisi rajatta todellisuudessa – ja ollakseni johdonmukainen Irakissa ja Somaliassa syntyneiden mallien osalta – käytän myös Somaliassa syntyneiden kohdalla mallia 4 lopullisena mallina, eli mallina, jota käytetään uusien maahanmuuttajien elinkaariennusteissa nettovaikutusten ennustamiseen 16–62-vuotiaina.

Taulukossa 4.10 on esitetty Somaliassa syntyneiden nettovaikutusmallin (malli 4) posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvojen (OA), keskihajontojen (KH), kvartiilien sekä jakaumien hännistä prosenttipisteiden 2,5 (%) ja 97,5 (%) estimaatit. Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli ykköstä, en käsittele niitä tarkemmin.

Taulukko 4.10: *Somaliassa syntyneiden16–62-vuotiaiden nettovaikutusten lopullinen malli (malli 4). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin Monte Carlo -keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakio	-9 560	520	-10 600	-9 900	-9 560	-9 220	-8 540	12	1,00
Nainen	-13 730	1 100	-15 800	-14 460	-13 760	-13 020	-11 530	31	1,00
Ikä	-3 380	520	-4 390	-3 740	-3 370	-3 030	-2 350	12	1,00
Ikä^2	2 820	470	1 900	2 510	2 810	3 140	3 720	11	1,00
Alle 19-vuotias	650	230	210	490	640	800	1 130	4	1,00
Yli 60-vuotias	140	590	-1 040	-240	140	530	1 270	9	1,00
Ikä $\times$ nainen	-6 560	1 150	-8 800	-7 330	-6 590	-5 810	-4 260	34	1,00
Ikä^2 $\times$ nainen	7 820	1 340	5 150	6 960	7 840	8 720	10 430	40	1,00
Muuttovuosi $<$ 1993	1 130	540	100	790	1 130	1 480	2 180	14	1,00
Muuttovuosi v. 1993–1996	520	900	-1 150	-90	490	1 110	2 320	25	1,00
Muuttovuosi v. 1997–2001	1 510	520	490	1 170	1 500	1 840	2 570	14	1,00
Muuttovuosi v. 2002–2006	320	450	-570	30	330	620	1 220	14	1,00
Muuttovuoden työttömyys-%	50	100	-160	-20	50	110	230	3	1,00
Maassaolo (vuosia)	1 350	260	830	1 180	1 350	1 520	1 840	7	1,00
Maassaolo $\times$ nainen	-2 180	460	-3 090	-2 480	-2 180	-1 890	-1 280	11	1,00
Maassaolo $\times$ (muuttovuosi $<$ 1993)	1 660	210	1 230	1 520	1 660	1 790	2 070	4	1,00
Maassaolo $<$ 5	40	440	-850	-240	30	300	940	11	1,00
(Maassaolo $<$ 5) $\times$ nainen	-280	1 000	-2 230	-920	-270	360	1 710	25	1,00
Maassaolo $\times$ (maassaolo $<$ 5)	850	520	-200	520	850	1 180	1 890	12	1,00
Maassaolo $\times$ (maassaolo $<$ 5) $\times$ nainen	-3 270	1 040	-5 410	-3 930	-3 250	-2 580	-1 270	25	1,00
Maassaolo $>$ 15	3 750	1 070	1 570	3 050	3 780	4 490	5 760	22	1,00
(Maassaolo $>$ 15) $\times$ nainen	-200	2 550	-5 260	-1 890	-190	1 510	4 620	64	1,00
Maassaolo $\times$ (maassaolo $>$ 15)	-2 080	570	-3 150	-2 480	-2 090	-1 690	-920	12	1,00
Maassaolo $\times$ (maassaolo $>$ 15) $\times$ nainen	170	1 340	-2 370	-740	180	1 060	2 830	34	1,00
Maassaolo alle 1 vuosi	7 190	300	6 560	7 020	7 190	7 380	7 810	7	1,00

Jatkuu seuraavalla sivulla

Taulukko 4.10 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Muutto alle 20-v.	2 920	690	1 560	2 450	2 900	3 380	4 290	17	1,00
Muutto 20–29-v.	1 690	570	580	1 320	1 680	2 070	2 840	15	1,00
Muutto 30–39-v.	1 860	460	960	1 550	1 860	2 170	2 780	12	1,00
Muutto alle 20-v. × maassaolo	2 150	380	1 420	1 890	2 130	2 400	2 920	9	1,00
Muutto 20–29-v. × maassaolo	1 070	320	460	860	1 070	1 290	1 690	8	1,00
Muutto 30–39-v. × maassaolo	1 180	320	570	960	1 180	1 380	1 820	8	1,00
Kalenterivuosi < 2001	-470	150	-760	-560	-460	-370	-180	2	1,00
Pääasiallinen toiminta: työllinen	3 800	480	2 880	3 480	3 800	4 120	4 770	9	1,00
Pääasiallinen toiminta: opiskelija	-1 380	240	-1 850	-1 540	-1 380	-1 220	-900	5	1,00
Korkein koulutus: 2. asteen koulutus	20	290	-550	-170	20	210	580	6	1,00
Korkein koulutus: 3. asteen koulutus	2 060	560	930	1 690	2 090	2 450	3 150	10	1,00
Alle 7-vuotiaiden lasten lukumäärä	-980	70	-1 130	-1 030	-980	-930	-840	1	1,00
Puoliso	390	160	80	280	390	500	700	3	1,00
Kotikunta Uudellamaalla	1 200	140	930	1 100	1 200	1 290	1 470	3	1,00
Ikä > 55	2 140	3 330	-4 380	-100	2 160	4 350	8 660	64	1,00
(Ikä > 55) × ikä	-810	1 450	-3 650	-1 770	-820	170	1 980	28	1,00
(Ikä < 40) × ikä × nainen	-120	1 140	-2 390	-870	-90	660	2 070	31	1,00
(Ikä > 50) × ikä × nainen	-4 970	1 310	-7 460	-5 850	-4 980	-4 100	-2 450	32	1,00
(Ikä < 19) × nainen	15 790	6 880	2 270	11 310	15 830	20 410	29 120	132	1,00
(Ikä < 40) × nainen	230	710	-1 120	-270	200	710	1 640	16	1,00
(Ikä > 50) × nainen	8 550	2 280	4 160	7 000	8 560	10 090	13 000	53	1,00
(Ikä < 19) × ikä × nainen	10 240	4 210	1 940	7 550	10 220	13 080	18 390	81	1,00
Nettovaikutus$_{t-1}$	22,4	0,6	21,4	22,1	22,4	22,8	23,6	0,02	1,00
Nettovaikutus$_{t-2}$	5,4	0,4	4,7	5,1	5,4	5,7	6,2	0,01	1,00
τ_{i1}	4 750	120	4 520	4 670	4 750	4 830	5 000	6	1,00
τ_{i2}	3 170	90	2 990	3 100	3 160	3 220	3 350	4	1,00
τ_{i3}	270	150	20	160	250	360	580	16	1,04
τ_{m1}	570	160	330	460	550	660	950	4	1,00
τ_{m2}	280	110	90	210	270	340	530	3	1,00

Jatkuu seuraavalla sivulla

Taulukko 4.10 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
τ_{t1}	140	70	20	90	140	190	300	3	1,00
τ_{t2}	250	70	140	200	240	290	410	2	1,00
σ	4 850	30	4 780	4 820	4 850	4 870	4 920	1	1,00
ν	2,8	0,05	2,7	2,77	2,8	2,83	2,9	0,004	1,00
γ	-1 550	60	-1 680	-1 590	-1 550	-1 510	-1 420	5	1,00

Listaan lyhyesti joitakin huomioita keskeisimmistä vaikutuksista ja estimointituloksista taulukossa 4.10. Kun kirjoitan parametrin sijoittuvan jollekin tietylle intervallille, tarkoitan tällöin kyseisen parametrin posteriorijakauman 95%:n uskottavuusintervallia 2,5 %:sta 97,5 %:iin.

- **Sukupuoli**. Päävaikutus naisilla verrattuna miehiin on noin -15 800:n ja -11 500:n välillä. Myös interaktiovaikutukset ovat ”suuria”, erityisesti ikää ja maassaoloaikaa koskevien muuttujien kanssa. Naisilla maassaoloajan vaikutus on negatiivisempi kuin miehillä ensimmäisen viiden vuoden ajan.
- **Maahanmuuttovuosi (1)**. Vuosina 2007–2011 muuttaneisiin verrattuna ennen vuotta 1993 muuttaneilla 100–2 200 euroa positiivisemmat nettovaikutukset. Myös vuosina 1997–2001 muuttaneilla nettovaikutukset ovat positiivisempia kuin vuosina 2007–2011 muuttaneilla, kun muiden muuttujien vaikutus huomioidaan. Vuosina 1993–1996 ja 2002–2006 muuttaneilla nettovaikutusten ero vuosina 2007–2011 muuttaneisiin ei ole näin selvä: kummankin maahanmuuttovuosiluokan osalta on noin 25 %:n todennäköisyys, että 2007–2011 muuttaneiden nettovaikutukset ovatkin positiivisemmat.
- **Maassaoloaika (1)**. Päävaikutus odotetusti positiivinen eli 7,1 vuoden lisäys maassaolossa (maassaoloajan keskihajonta) nostaa nettovaikutusten odotusarvoa 830–1 840 eurolla.
- **Maassaoloaika (2)**. Maassaoloajan positiivinen vaikutus on kuitenkin ensimmäisinä maassaolon vuosina (alle 5 vuotta maassa) maassaoloajan yleistä päävaikutusta suurempi, ja sen jälkeen puolestaan pienempi.
- **Maassaoloaika (3)**. Maassaoloajan vaikutus ennen vuotta 1993 muuttaneilla on odotusarvoisesti 1 660 euroa positiivisempi (per 7,1 vuotta maassa) kuin myöhemmin muuttaneilla, ks. myös kuva 4.16, s. 217.
- **Muut regressiokertoimet**. Jos on asunut maahanmuuttovuonna Uudellamaalla tai pääasiallinen toiminta on ollut työllinen, niin nettovaikutusten odotusarvo on positiivisempi kuin vertailuluokassa "muut", joka sisältää myös työttömänä olleet. Opiskelijana olleilla nettovaikutukset ovat puolestaan negatiivisemmat kuin vertailuluokassa. Lisäksi mitä enemmän on alle 7-vuotiaita lapsia muuttovuonna, sitä negatiivisempi on nettovaikutusten odotusarvo. Näiden vaikutusten osalta tulokset ovat hyvin samanlaiset kuin Irakissa syntyneillä.
- **Autoregressiiviset termit**. Ensimmäisen autoregressiivinen termin jakauma keskittynyt 0,22:n ympäristöön. Toisen termin jakauma keskittynyt 0,05:n ympäristöön eli lähelle nollaa.
- **Satunnaisvaikutukset**. Henkilöittäin vaihtelevan mallin vakiotermin keskihajonnan (τ_{i1}) odotusarvo noin 4 750 euroa, maahanmuuttovuoden (”vakion”) satunnaisvaikutuksen keskihajonnan (τ_{m1}) odotusarvo noin 570 euroa ja maassaoloajan (”vakion”) satunnaisvaikutuksen keskihajonnan (τ_{t1}) odotusarvo noin 140 euroa. Näistä satunnaisvaikutuksista ”satunnainen” henkilötasoinen vakio aiheuttaa odotusarvoisesti suunnilleen yhtä paljon epävarmuuksia elinkaariennusteisiin kuin maassaoloajan (”vakion”) satunnaisvaikutus, jos uusi maahanmuuttaja on koko työiän (ikävuodet 20–62) Suomessa.
- **Muut parametrit**. Studentin t-jakauman vapausasteet ”pieniä”, koska ν-parametrin posteriorijakauma on keskittynyt noin 2,8 ympäristöön. Nettovaikutusten otosjakauma on vasemmalle vino,

koska γ on negatiivinen.

4.4.2.3 Somalia: posterioriennustetarkistukset

Posterioriennusteet on muodostettu yllä esitetyn mallin 4 (eli elinkaariennusteissa käytettävän) parametrien posteriorijakaumien estimaateista ja havaintoaineistosta niin, kuin on esitetty maassaoloaikojen mallinnuksen yhteydessä luvussa 3.2.6. Olen laatinut posterioriennusteet niin, kuin kyse olisi uusien maahanmuuttajien aineistosta, joten poimin posterioriennusteita varten myös ennusteet uusien maahanmuuttajien satunnaisvaikutuksille.

Posterioriennustetarkistus: Kaikkien nettovaikutushavaintojen yhteisjakauma ja henkilöiden nettovaikutusten keskiarvojen yhteisjakauma.

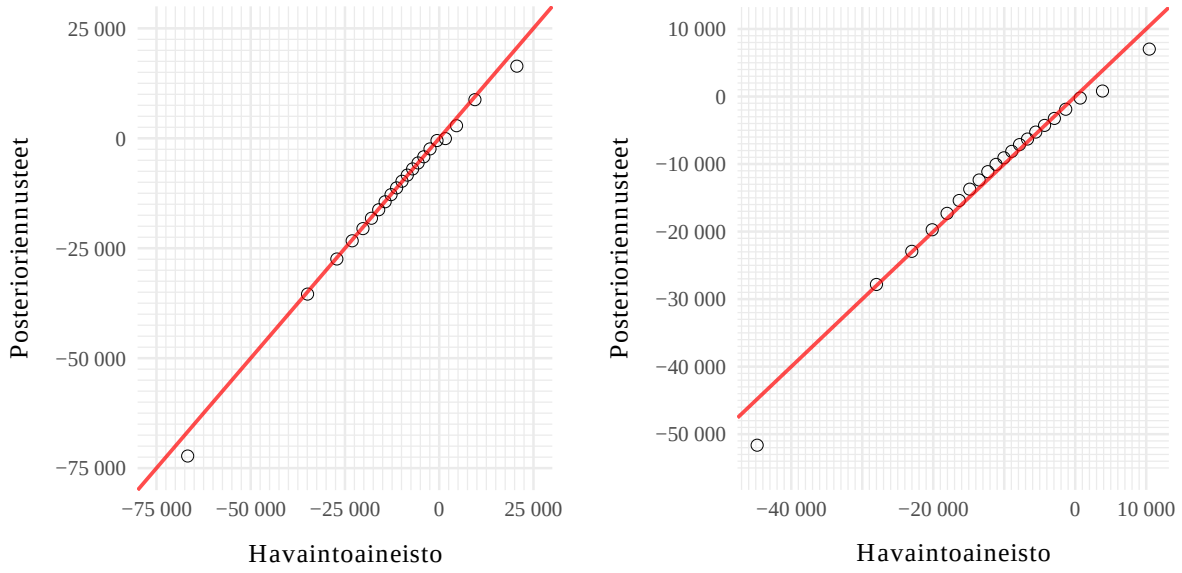
Aloitan posterioriennustetarkistukset vertaamalla havaintoaineiston kaikkien nettovaikutushavaintojen muodostamaa yhteisjakaumaa posterioriennustejakauman kaikkien ennusteiden muodostamaan yhteisjakaumaan. Ensimmäisen mallinnustavoitteen mukaan näiden jakaumien pitäisi vastata mahdollisimman hyvin toisiaan, erityisesti jakaumien tärkeimpien tunnuslukujen kuten keskiarvojen, keskihajontojen ja kvartiilien suhteen.

Tämän lisäksi vertaan kaikkien havaintoaineiston henkilöiden nettovaikutusten keskiarvojen jakaumaa ennustettuihin henkilöiden nettovaikutusten keskiarvojen jakaumaan. Nämä jakaumat muodostetaan havaintoaineistossa laskemalla nettovaikutusten keskiarvot kaikille aineiston henkilöille ja muodostamalla näistä jakauma. Posterioriennusteiden kohdalla lasketaan ennustettujen nettovaikutusten keskiarvot kaikille henkilöille, ja muodostetaan näistä jakauma. Henkilötasoisten jakaumien vertailun tarkoituksena on osoittaa, että elinkaariennusteissa käytettävä nettovaikutusten malli tuottaa luotettavia ennusteita paitsi havaintotasolla niin myös henkilötasolla.

Tarkistus kvantiilikuviolla. Vertaan alla kuviossa 4.12 havaintoaineiston jakaumien prosenttipisteitä vastaavien posterioriennustejakaumien prosenttipisteisiin. Jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeama vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).

Kuva 4.12 osoittaa, että posterioriennusteiden tuottamat jakaumat vastaavat hyvin havaintoaineiston ja-



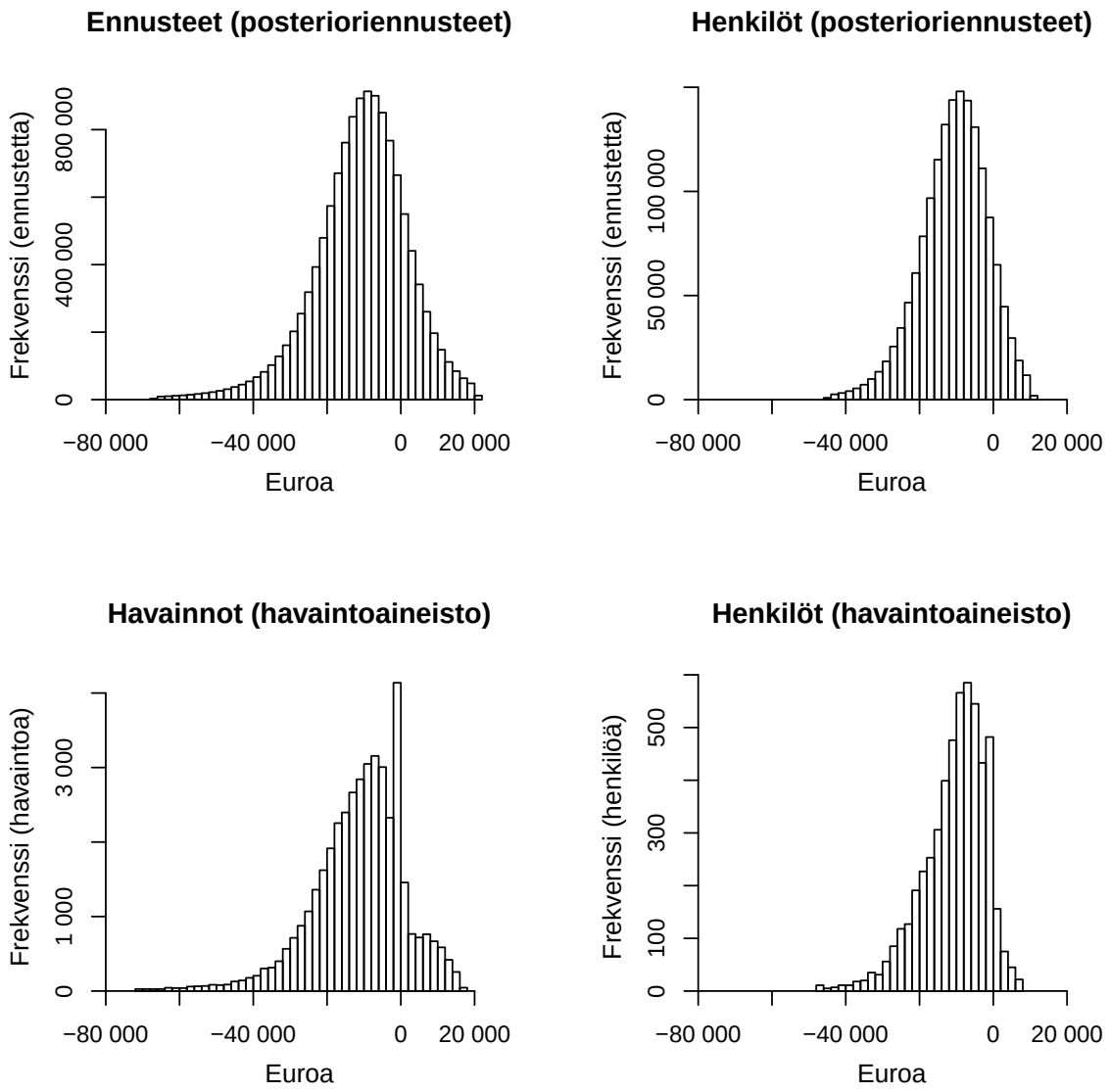
(a) *Kaikki havainnot ja ennusteet.*

(b) *Henkilöiden keskiarvot.*

Kuva 4.12: *Havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

kaumia sekä kaikkien havaintojen osalta että henkilöiden nettovaikutusten keskiarvojen osalta. Kuten Irakissa syntyneiden kohdalla, myös Somaliassa syntyneillä nollan ympäristössä on havaintojen keskittymä, minkä vuoksi prosenttipisteet kuvassa 4.12 poikkeavat hieman punaisesta suorasta havaintoaineiston nollan kohdalla. Sekä kaikkien havaintojen että henkilöiden keskiarvojen jakaumissa esitetyt alimmat ja ylimmät (1. ja 99.) prosenttipisteiden arvot poikkeavat posterioriennusteiden ja havaintoaineiston välillä muutamia tuhansia euroja toisistaan.

Tarkistus histogrammeilla. Verrataan nyt havaintoaineiston ja posterioriennusteiden jakaumia prosenttipisteiden vertailukuvioiden (Q-Q plot) lisäksi histogrammien avulla. Nämä on esitetty kuvassa 4.13. Kuvan ylärivillä on kaksi posterioriennusteista muodostettua jakaumaa, ja alarivillä puolestaan niitä vastaavat havaintoaineiston jakaumat. Histogrammien vertailu osoittaa, että posterioriennustejakaumat vastaavat hyvin havaintoaineiston jakaumia. Kaikkien havaintojen jakaumassa (vasen alakulma) on histogrammissa nähtävissä havaintojen keskittyminen nollan ympäristöön. Kuvasta käy myös ilmi, että henkilöiden keskiarvojen jakauma on kapeampi kuin kaikkien nettovaikutusaineistojen jakauma.



Kuva 4.13: *Posterioriennusteiden ja havaintoaineiston jakaumat histogrammeina. Posterioriennusteiden ennusteet ja ennustetut henkilöiden nettovaikutusten keskiarvot (ylärivi). Havaintoaineiston havainnot ja havaitut henkilöiden nettovaikutusten keskiarvot (alarivi). Kaikista jakaumista on katkaistu pois alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Tarkistus testisuureiden avulla. Alla taulukossa 4.11 on esitetty havaintoaineiston ja posterioriennusteiden jakaumien vertailut euromääräisinä lukuina sekä *Bayes-p-arvojen* ja eri euromääräisten *Bayes-intervallien* avulla. Olen käsitellyt näiden testisuureiden merkitystä ja laskemista alaluvussa 4.3.7. Tau-

lukossa 4.11 verrataan posterioriennustejakaumien (havaintojen ja henkilöiden keskiarvojen jakaumat) keskiarvoja, keskihajontoja sekä joitakin prosenttipisteitä havaintoaineiston vastaaviin jakauman tunnuslukuihin.

Taulukon 4.11 ensimmäisellä ja toisella rivillä ovat posterioriennustejakaumien ja havaintoaineiston jakaumien tunnuslukujen estimaatit. Bayes-p-arvot ja Bayes-intervallit eri euromäärille ovat testisuureita estimaattien vastaavuuksille havaintoaineiston ja posterioriennusteiden välillä. Bayes-intervalli ilmaisee todennäköisyyden sille, että posterioriennustejakauman estimaatti poikkeaa korkeintaan ilmoitetun euromäärän verran havaintoaineiston estimaatista. Kaikkien havaintojen jakauman osalta posterioriennusteiden estimaatit eivät melko suurilla todennäköisyyksillä poikkea havaintoaineiston estimaateista 1 000 euroa enempää – lukuun ottamatta alimpia ja ylimpiä prosenttipisteitä sekä nollan ympäristöä (90. prosenttipiste).

Taulukko 4.11: *Havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien keskiarvot (KA), keskihajonnat (KH) ja prosenttipisteet:* $1, 5, 10, 25, 50, 75, 90, 95, 99$. *Vertailtavien estimaattieen Bayes-p-arvot ja Bayes-intervallit eri euromäärillä.*

Kaikki havainnot ja ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %
Posterioriennusteet	-11 510	18 640	-66 730	-34 920	-27 160	-17 970	-9 870	-2 440	4 590	9 470	20 590
Aineiston havainnot	-12 020	19 280	-72 240	-35 420	-27 440	-18 180	-9 760	-2 410	2 870	8 790	16 410
Bayes-p-arvo	0,98	0,02	1	0,91	0,85	0,8	0,31	0,45	1	0,9	1
Bayes-voima: 2 000 euroa	1	1	0	1	1	1	1	1	0,78	0,97	0,01
Bayes-voima: 1 000 euroa	0,95	0,88	0	0,9	0,99	1	1	1	0,02	0,76	0
Bayes-voima: 500 euroa	0,48	0,34	0	0,46	0,75	0,9	0,95	0,9	0	0,4	0
Bayes-voima: 250 euroa	0,15	0,11	0	0,17	0,4	0,57	0,64	0,59	0	0,21	0

Henkilöiden keskiarvot ja niiden ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %
Posterioriennusteet	-11 070	11 940	-44 810	-28 000	-23 020	-16 340	-10 040	-4 320	710	3 840	10 430
Aineiston havainnot	-11 190	13 960	-51 630	-27 830	-22 920	-15 390	-9 070	-4 280	-220	810	7 020
Bayes-p-arvo	0,68	0,04	1	0,35	0,36	0	0	0,38	1	1	1
Bayes-voima: 2 000 euroa	1	0,38	0,01	1	1	1	1	1	1	0	0,02
Bayes-voima: 1 000 euroa	1	0,08	0	0,97	1	0,6	0,55	1	0,61	0	0
Bayes-voima: 500 euroa	0,93	0,04	0	0,73	0,88	0,02	0,01	0,93	0,08	0	0
Bayes-voima: 250 euroa	0,65	0,02	0	0,42	0,59	0	0	0,66	0,01	0	0

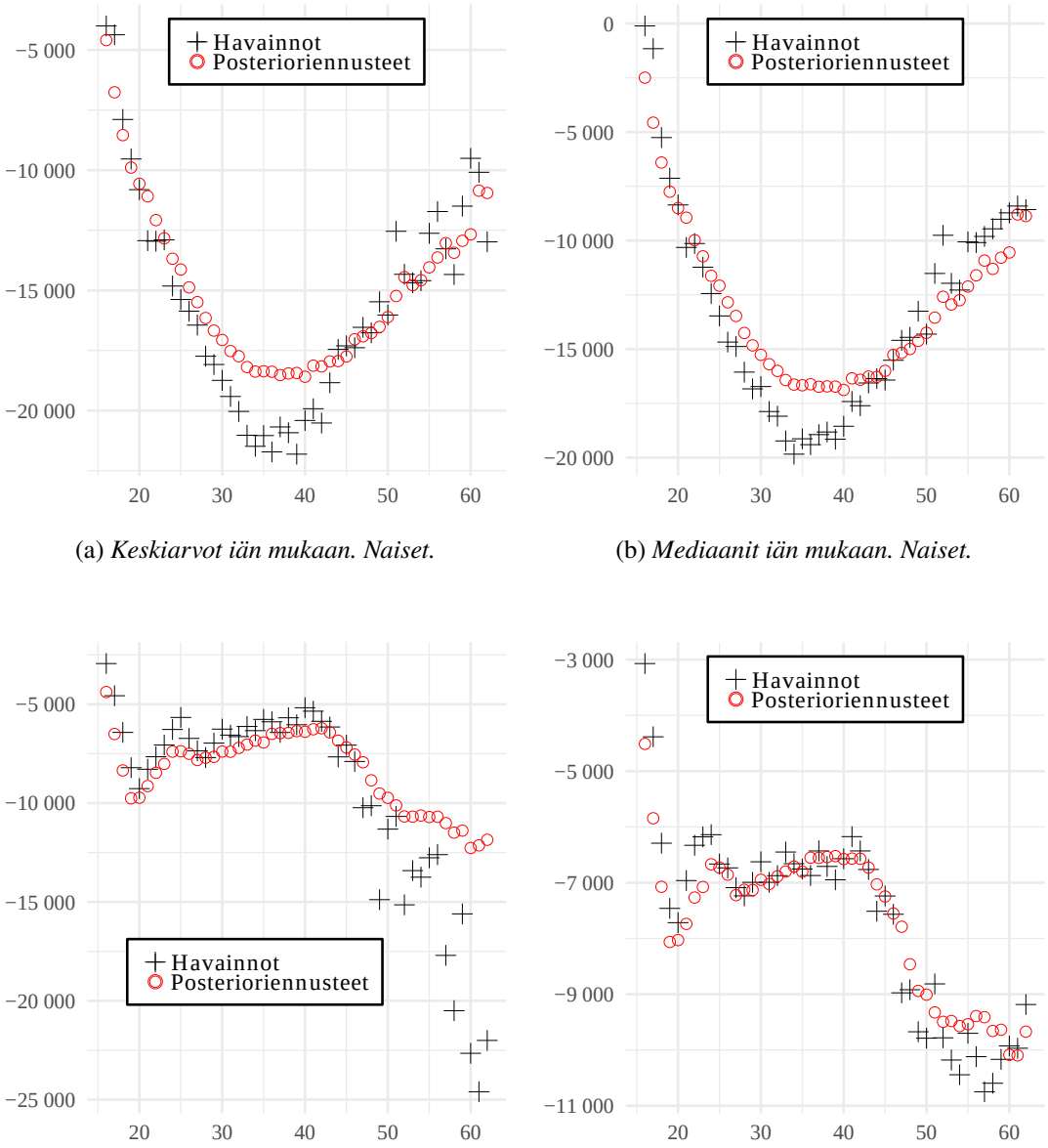
Posterioriennustarkistukset selittävien muuttujien luokkien mukaan.

Taulukon 4.11 jälkeen vertaan posterioriennusteita havaintoaineistoon tärkeimpien selittävien muuttujien luokkien mukaan. Tärkeimpiä selittäviä muuttujia ovat sukupuoli sekä aikaan liittyvät muuttujat: ikä, maahanmuuttovuosi, maassaoloaika ja kalenterivuosi. Viimeinen näistä muuttujista ei ollut lainkaan mallissa, koska maahanmuuttovuosi, maassaoloaika ja kalenterivuosi eivät voi olla samaan aikaan lineaarisessa mallissa. Esitän posterioriennustetarkistuksen tästä huolimatta myös kalenterivuoden suhteen.

Vertaan posterioriennusteiden keskiarvoja ja mediaaneja selittävien luokkien mukaan vastaaviin havaintoaineiston keskiarvoihin ja mediaaneihin. Esimerkiksi iän kohdalla vertaan keskiarvoja ja mediaaneja niin, että keskiarvot ja mediaanit on laskettu posterioriennusteista jokaiselle mallinnusaineiston iälle erikseen 16-vuotiaista 62-vuotiaisiin. Keskiarvojen lisäksi esitän mediaanit, koska nettovaikutusten jakauman leveiden häntien vuoksi mediaani on oletettavasti keskiarvoa robustimpi jakaumien keskiluku. Kuvioiden jälkeen esitän jokaiseen selittävien muuttujien luokkien suhteen tehtyyn posterioriennustetarkistukseen liittyen tärkeimpiä huomioita.

Maassaoloajan suhteen esitän posterioriennustetarkistukset erikseen vuonna 1993 ja sitä myöhemmin Suomeen muuttaneille sekä ennen vuotta 1993 muuttaneille. Syynä tähän on se, että nettovaikutusten havaintojen tarkastelun perusteella ennen vuotta 1993 Suomeen muuttaneilla Somaliassa syntyneillä on ollut positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla. Taulukossa 4.10 esitettyjen estimointitulosten perusteella tämä ero selittyy ainakin sillä, että ennen vuotta 1993 muuttaneilla maassaoloajan vaikutus on positiivisempi kuin myöhemmin muuttaneilla. Ennen vuotta 1993 muuttaneiden positiivisemmat nettovaikutukset selittyvät aineistontarkastelun perusteella lisäksi sillä, että ennen vuotta 1993 muuttaneiden joukossa on enemmän miehiä suhteessa naisiin kuin myöhemmin muuttaneilla. Miesten nettovaikutukset ovat tuhansia euroja positiivisemmat kuin naisten, ja myös maassaoloaika vaikuttaa miehillä positiivisemmin kuin naisilla.

Posterioriennustetarkistus iän ja sukupuolen mukaan.



(a) *Keskiarvot iän mukaan. Naiset.*

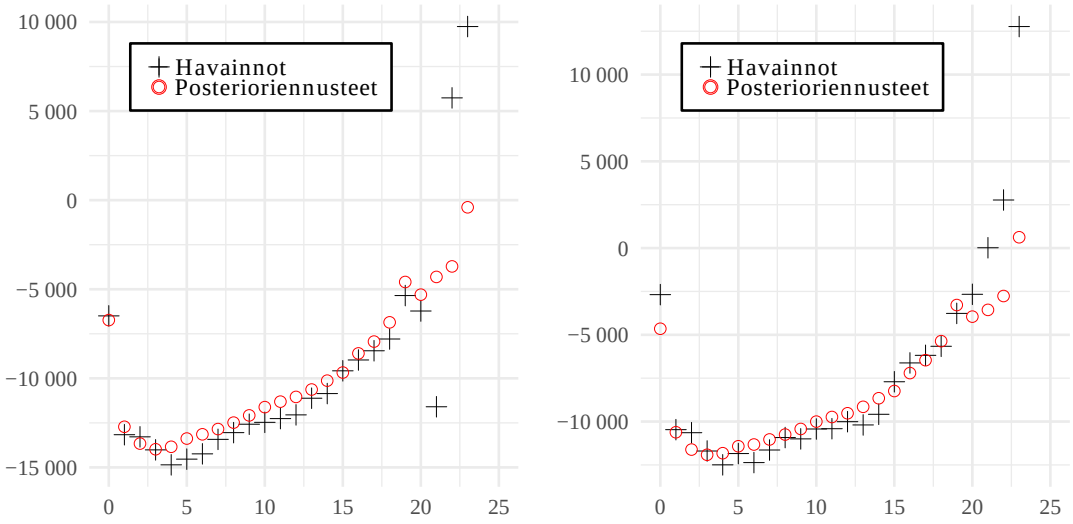
(b) *Mediaanit iän mukaan. Naiset.*

(c) *Keskiarvot iän mukaan. Miehet.*

(d) *Mediaanit iän mukaan. Miehet.*

Kuva 4.14: *Havainnot ja posterioriennusteet iän ja sukupuolen mukaan.*

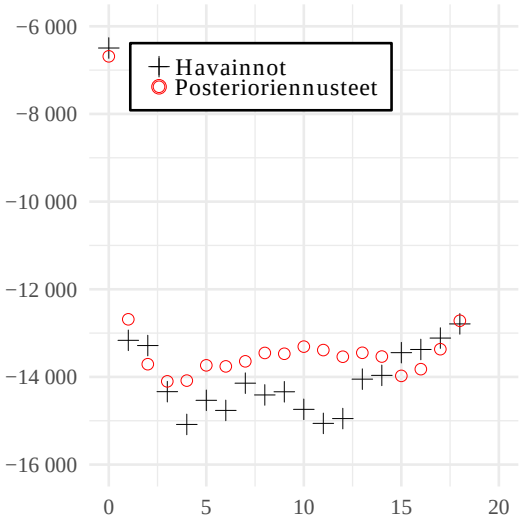
Posteriorienustetarkistus maassaoloajan mukaan.



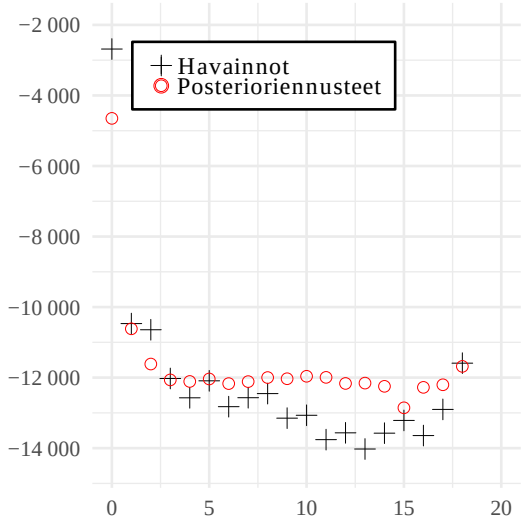
(a) *Keskiarvot maassaoloajan mukaan.* (b) *Mediaanit maassaoloajan mukaan.*

Kuva 4.15: *Havainnot ja posteriorienustteet maassaoloajan (vuosia) mukaan.*

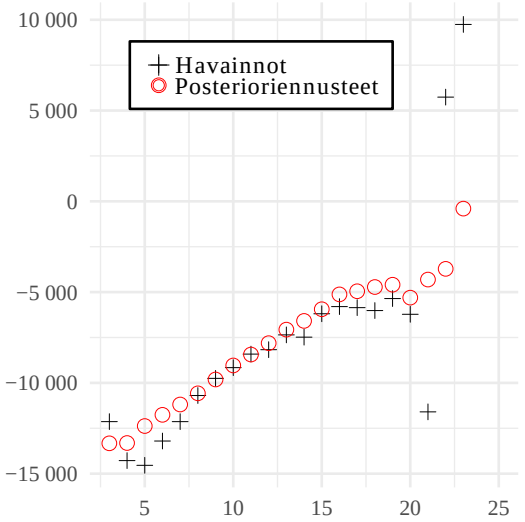
Posterioriennustetarkistus maassaoloajan ja maahanmuuttovuosiluokan mukaan. Maassaoloaika vuonna 1993 ja myöhemmin muuttaneilla verrattuna ennen vuotta 1993 muuttaneisiin.



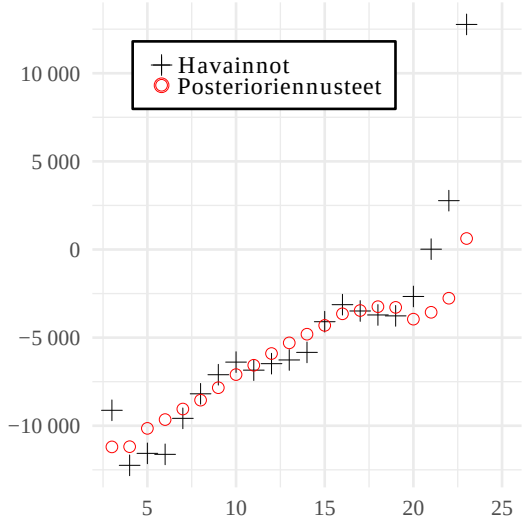
(a) *Keskiarvot maassaoloajan mukaan. Vuonna 1993 ja myöhemmin Suomeen muuttaneet.*



(b) *Mediaanit maassaoloajan mukaan. Vuonna 1993 ja myöhemmin Suomeen muuttaneet.*



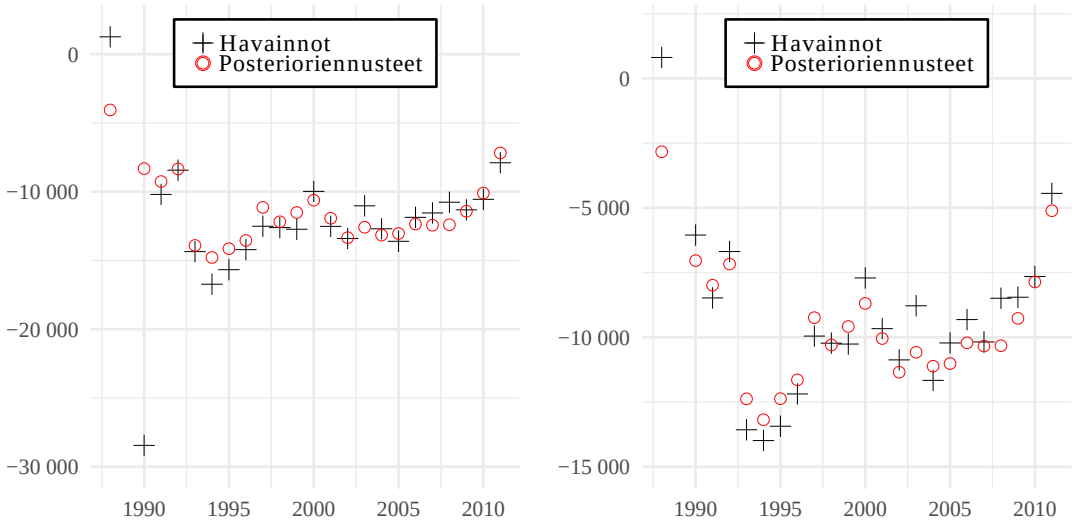
(c) *Keskiarvot maassaoloajan mukaan. Ennen vuotta 1993 Suomeen muuttaneet.*



(d) *Mediaanit maassaoloajan mukaan. Ennen vuotta 1993 Suomeen muuttaneet.*

Kuva 4.16: *Havainnot ja posterioriennusteet maassaoloajan ja maahanmuuttovuosiluokan mukaan.*

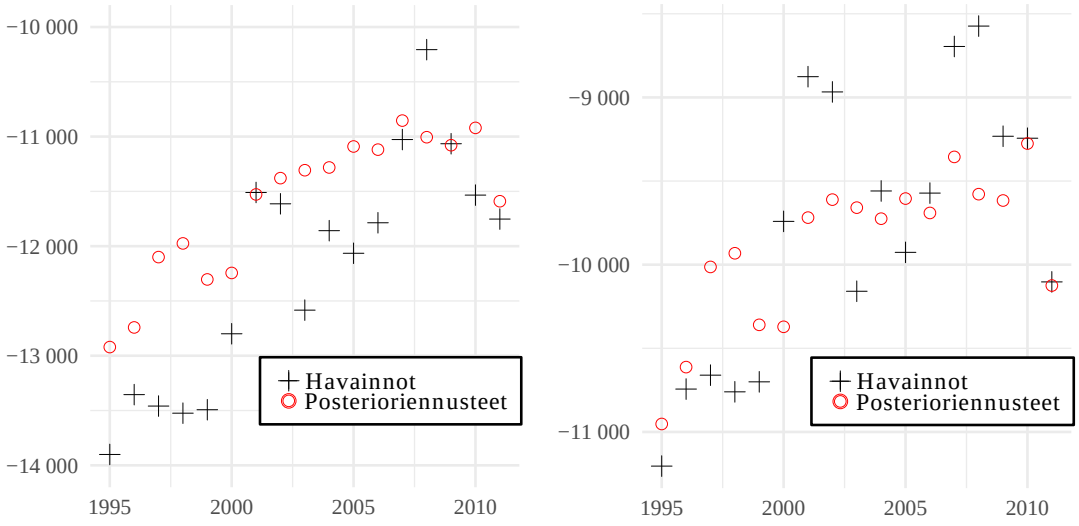
Posterioriennustetarkistus maahanmuuttovuoden mukaan.



(a) *Keskiarvot maahanmuuttovuoden mukaan.* (b) *Mediaanit maahanmuuttovuoden mukaan.*

Kuva 4.17: *Havainnot ja posterioriennusteet maahanmuuttovuoden mukaan.*

Posterioriennustetarkistus kalenterivuoden mukaan.



(a) *Keskiarvot kalenterivuoden mukaan.* (b) *Mediaanit kalenterivuoden mukaan.*

Kuva 4.18: *Havainnot ja posterioriennusteet kalenterivuoden mukaan.*

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan. Havaintoja kuvista 4.14–4.18.

- **Yleistä (1)**. Posterioriennusteiden mediaanit näyttävät vastaavan havaintoaineiston mediaaneita paremmin kuin keskiarvojen osalta – sekä kaikkien havaintojen osalta (taulukko 4.11) että selittävien muuttujien luokkien mukaan tarkasteltuna. Erityisesti tämä tulee esille iän ja sukupuolen mukaan esitetyssä posterioriennustetarkastuksessa (kuva 4.14), jossa varsinkin yli 50-vuotiailla miehillä posterioriennusteiden mediaanit vastaavat hyvin havaintoaineiston mediaaneita, toisin kuin keskiarvot (vrt. kuvan 4.14 osakuvat (c) ja (d)).

- **Yleistä (2)**. Kuvat 4.14–4.18 osoittavat, että yleisesti ottaen posterioriennusteiden keskiarvot ja mediaanit vastaavat hyvin havaintoaineiston keskiarvoja ja mediaaneita, kun vertailu tehdään tärkeimpien selittävien muuttujien luokkien suhteen.

- **Ikä sukupuolen mukaan** (kuva 4.14). Naisilla nettovaikutukset vaihtelevat ”paljon” iän mukaan siten, että ikävuosikohtaisten nettovaikutusten keskiarvojen ja mediaanien minimi on noin 35–40-vuotiaana. Posterioriennusteet eivät tavoita edes mediaanien osalta 25–40-vuotiaiden naisten negatiivisia nettovaikutuksia, vaan ennusteet ovat liian positiivisia. Miehillä nettovaikutukset vaikuttavat kuvan 4.14 perusteella pysyvän suunnilleen samalla tasolla noin 25-vuotiaasta vähän yli 40-vuotiaaksi. Tämän jälkeen ikäluokittaiset keskiarvot ja mediaanit muuttuvat nopeasti negatiivisemmiksi. Mediaanien osalta posterioriennusteet tavoittavat tämän nettovaikutusten muuttumisen negatiivisemmiksi.

- **Maassaoloaika (kaikki)** (kuva 4.15). Nettovaikutusten keskiarvoissa ja mediaaneissa maassaoloajan suhteen on nähtävissä havaintojen perusteella kuvassa 4.15 neljä periodia melkein samoin kuin Irakissa syntyneillä: ensimmäinen maassaolovuosi (0 maassaolovuotta), alle 5 vuotta maassa (1–4 maassaolovuotta), 5–14 maassaolovuotta ja yli 14 maassaolovuotta. Tosin keskiarvojen perusteella nämä luokat eivät ole niin selvärajaisia kuin mediaanien perusteella luokiteltuna. Ensimmäisenä vuonna nettovaikutukset ovat keskimäärin positiivisemmat kuin 1–4 maassaolovuonna. Maassaolovuosina 5–14 vuotta nettovaikutukset ovat positiivisemmat kuin ensimmäisinä maassaolovuosina, ja yli 14 maassaolovuoden jälkeen nettovaikutukset näyttävät muuttuvan nopeasti positiivisemmiksi. Posterioriennusteet seuraavat tätä kehitystä melko hyvin sekä keskiarvojen että mediaanien osalta.

- **Maassaoloaika (v. 1993 ja myöhemmin muuttaneet, ennen v. 1993 muuttaneet) (1)** (kuva 4.16). Keskiarvojen ja mediaanien nopea nousu 14 maassaolovuoden jälkeen (kuva 4.15) johtuu siitä, että ennen vuotta 1993 muuttaneilla on positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla. Kuten olen jo aiemmin selventänyt, tämä johtuu siitä, että ennen vuotta 1993 muuttaneilla maassaolon vaikutus on positiivisempi kuin myöhemmin muuttaneilla ja ennen vuotta 1993 muuttaneiden joukossa on enemmän miehiä kuin myöhemmin muuttaneissa. Viitatun taulukon 4.10 estimaattien perusteella ennen 1993 muuttaneiden nettovaikutukset eivät ole odotusarvoisesti juurikaan kaikkia muita maahanmuuttovuosiluokkia positiivisempia, kun muut selittävät kuten maassaoloajan ja sukupolven vaikutukset huomioidaan.

- **Maassaoloaika (v. 1993 ja myöhemmin muuttaneet, ennen v. 1993 muuttaneet) (2)** (kuva 4.16). Kuvan 4.16 osakuvien (a) ja (b) perusteella nettovaikutusten keskiarvot ja mediaanit eivät vaiku-

ta juurikaan muuttuvan positiivisemmiksi maassaolovuosien suhteen. Todellisuudessa, kun muut selittävät tekijät otetaan huomioon, maassaoloajalla on positiivinen vaikutus 15 maassaolovuoteen asti, kuten selvitin edellä taulukon 4.10 tuloksiin viitaten. Tämän jälkeen maassaoloajan vaikutus on lähellä nollaa. Maassaoloajan vaikutus on negatiivisempi naisilla kuin miehillä.

- **Maahanmuuttovuosi (1)** (kuva 4.17). Käsittelin yllä maassaolovuosien vaikutusten yhteydessä jo maahanmuuttovuoden vaikutuksen nettovaikutuksiin. Kuvasta 4.17 käy selvästi ilmi, kuinka ennen 1993 muuttaneilla on ollut positiivisemmat nettovaikutukset kuin myöhemmin muuttaneilla. Parametrien estimaattitaulukossa 4.10 positiivisin nettovaikutusten odotusarvo on 1997–2001 muuttaneilla, kun muiden selittävien muuttujien vaikutukset huomioidaan. Myös viitatussa kuvassa vuoden 2000 ympärillä muuttaneilla vaikutukset ovat jonkin verran positiivisemmat kuin vuoden 1995 tai vuoden 2005 ympärillä muuttaneilla.

- **Maahanmuuttovuosi (2)** (kuva 4.17). Vuonna 2011 muuttaneilla nettovaikutukset ovat positiivisemmat, koska heitä koskien on olemassa vain ensimmäisen vuosi maassa, jolloin nettovaikutukset ovat positiivisemmat (koska keskimäärin Suomessa ollaan vain osan kalenterivuotta maahanmuuttovuonna). Sama koskee lievemmin esimerkiksi vuonna 2010 muuttaneita, joilla noin puolet havainnoista koskevat maahanmuuttovuotta. Posterioriennusteiden keskiarvot ja mediaanit vastaavat joka tapauksessa hyvin havaintoaineiston vastaavia tunnuslukuja maahanmuuttovuoden suhteen.

- **Kalenterivuosi** (kuvio 4.18). Kalenterivuosien osalta erityisesti ennusteiden mediaanit vastaavat melko hyvin havaintoaineiston mediaanien keskiarvoa vuoden 2000 jälkeen. Nettovaikutusten taso kalenterivuoden suhteen vaihtelee Somaliassa syntyneiden kohdalla sen mukaan, kuinka pitkään Suomessa olleista Somaliassa syntyneiden väestö koostuu kunakin kalenterivuonna. Mikäli väestössä on suhteellisesti enemmän juuri Suomeen muuttaneita, nettovaikutukset ovat tuolloin negatiivisemmat kuin sellaisina vuosina, jolloin juuri muuttaneiden osuus on pienempi – kun pidetään muut vaikuttavat tekijät vakioina.

4.4.3 Suomi

4.4.3.1 Suomi: havaintoaineisto

Taulukossa 4.12 on esitetty Suomessa syntyneiden 20–62-vuotiaiden nettovaikutushavaintojen lukumäärä ja havaintojen jakauman keskeisimmät tunnusluvut. Suomessa syntyneiden mallinnusaineistoa varten poimin kokonaishavaintoaineistosta yksinkertaisella satunnaisotannolla niin monta henkilöä ja heitä koskevat nettovaikutushavainnot, jotta mallinnusaineiston kokonaishavaintomääräksi tulisi 50 000. Henkilöiden määräksi tuli mallinnusaineistoon näin 3 990 henkilöä.

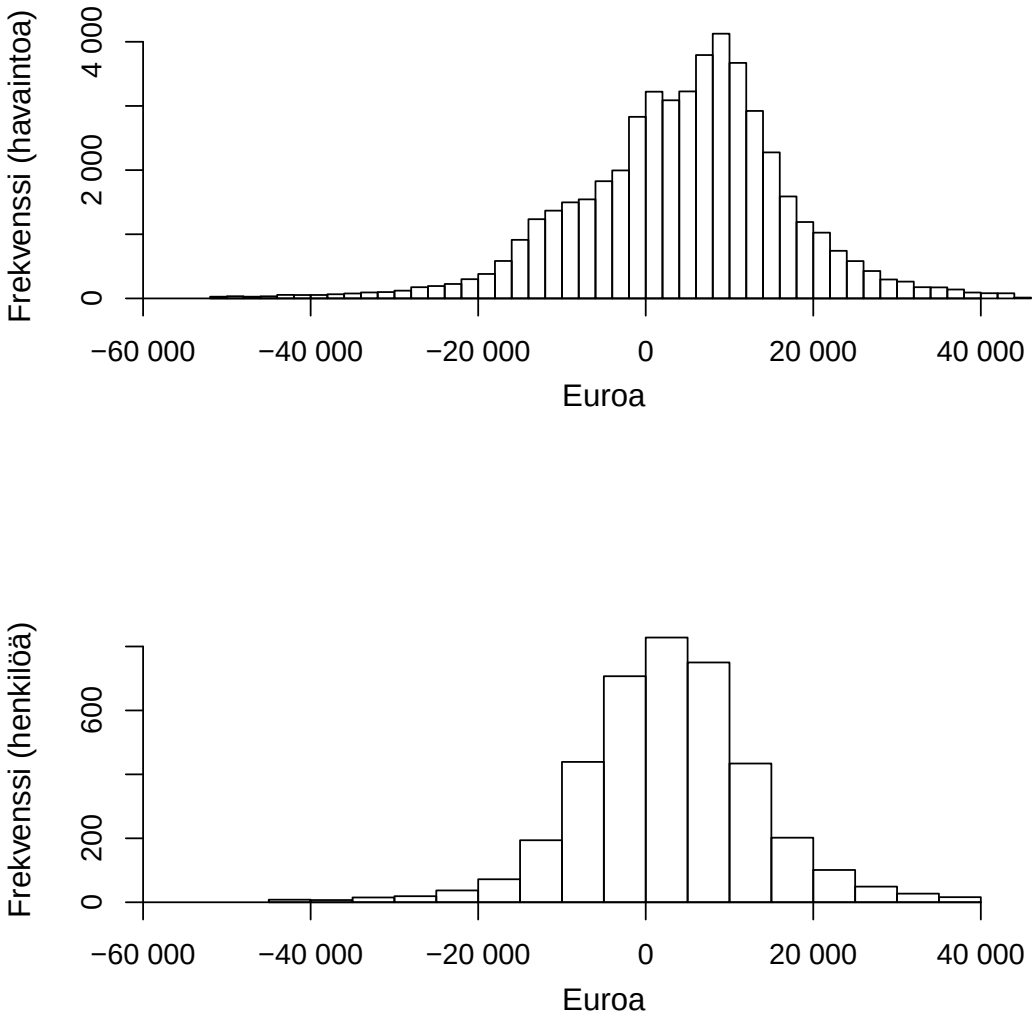
En käyttänyt alkuperäistä kokonaishavaintoaineistoa, joka olisi käsittänyt satoja tuhansia havaintoja, ajansäästösyistä. Koska 50 000 havainnolla suurimman mallin estimointiin kului useita vuorokausia, kokonaishavaintoaineistolla suurimman mallin estimointiin olisi saattanut kulua kuukausi. Saatan nostaa vielä mallinnuksessa käytettävää havaintomäärää tämän tutkimuksen II osan mahdollisissa päivitetyissä versioissa. Suurempi havaintomäärä mallinnusaineistossa tarkoittaa käytännössä parametrien posteriorijakaumien keskihajontojen pienenemistä.

Nettovaikutushavainnot muodostavat selitettävän muuttujan työikäisten julkisen talouden nettovaikutusten mallinnuksessa. Taulukon 4.12 tietojen mukaan esimerkiksi kaikkien havaintojen keskiarvo on +3 900 euroa ja mediaani +5 500 euroa. Taulukossa 4.12 on esitetty samat tiedot myös havaintoaineiston henkilöiden nettovaikutushavaintojen keskiarvoille. Havaintoaineiston henkilöiden (3 990 henkilöä) nettovaikutushavaintojen keskiarvojen mediaani oli +2 700 euroa ja keskihajonta 20 900 euroa.

Taulukko 4.12: *Suomessa syntyneet 20–62-vuotiaat. Nettovaikutushavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa sekä henkilöiden lukumäärä ja henkilötasoisten nettovaikutushavaintojen keskiarvojen jakauman tunnusluvut mallinnusaineistossa. Euromääräiset tunnusluvut pyöristetty lähimpään sataan euroon.*

	Havainnot	Henkilöt
Kokonaislukumäärä (N)	50 000	3 990
Keskiarvo	3 900	2 800
Keskihajonta	23 900	20 900
1. prosenttipiste (1 %)	-52 600	-47 100
5. prosenttipiste (5 %)	-18 100	-14 900
10. prosenttipiste (10 %)	-12 400	-9 900
Alakvartiili (25 %)	-2 900	-3 700
Mediaani (50 %)	5 500	2 700
Yläkvartiili (75 %)	12 000	9 100
90. prosenttipiste (90 %)	19 300	15 700
95. prosenttipiste (95 %)	25 000	21 400
99. prosenttipiste (99 %)	44 500	40 300

Kuvassa 4.19 on esitetty histogrammit sekä kaikkien nettovaikutushavaintojen jakaumalle (ylempi histogrammi) että henkilöiden nettovaikutusten keskiarvojen jakaumalle (alempi histogrammi). Kummastakin jakaumasta on katkaistu pois niiden ylin ja alin prosentti pois.



Kuva 4.19: *Suomessa syntyneet 20–62-vuotiaat. Nettovaikutushavaintojen katkaistun jakauman histogrammi mallinnusaineistossa (yllä) sekä henkilöiden nettovaikutushavaintojen keskiarvojen katkaistun jakauman histogrammi (alla). Jakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Histogrammin perusteella vaikuttaa siltä, että kaikkien havaintojen jakauma on vasemmalle vino. Tähän viittaa myös se, että taulukossa 4.12 havaintojen keskiarvo (3 900 e) on pienempi kuin havaintojen mediaani (5 500 e). Henkilöiden keskiarvojen jakauma vaikuttaa histogrammin ja taulukon 4.12 tietojen perusteella melko symmetriseltä.

Taulukossa 4.13 on esitetty Suomessa syntyneiden nettovaikutusmallien selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden kaikkien havaintojen joukossa. Edellä parametrien selitetaulukossa (s. 170) on lyhyitä selityksiä selittävistä muuttujista.

Taulukko 4.13: *Suomessa syntyneet 20–62-vuotiaat. Nettovaikutusmallien selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta).*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,48	0,50	0	0	1
Ikä	41,7	12,2	31	42	52
Ikä^2	1883	1005	961	1764	2704
Ikä $\times$ nainen	20,1	22,5	0	0	42
Ikä^2 $\times$ nainen	912	1178	0	0	1764
Kalenterivuosi	2003	4,9	1999	2003	2007
Kalenterivuosi $\times$ nainen	963	1001	0	0	2003
Kalenterivuosi < 2001	0,36	0,48	0	0	1
Kalenterivuosi $\times$ (kalenterivuosi < 2001)	711	956	0	0	1996
Ikä $<$ 28	0,17	0,38	0	0	0
(Ikä $<$ 28) $\times$ ikä	4,0	8,9	0	0	0
(27 $<$ Ikä $<$ 37) $\times$ nainen	0,1	0,3	0	0	0
(27 $<$ Ikä $<$ 37) $\times$ ikä $\times$ nainen	3,8	10,6	0	0	0
(36 $<$ Ikä $<$ 47) $\times$ nainen	0,1	0,3	0	0	0
(36 $<$ Ikä $<$ 47) $\times$ ikä $\times$ nainen	4,8	13,4	0	0	0
(46 $<$ Ikä $<$ 58)	0,29	0,45	0	0	1
Ikä $>$ 57	0,11	0,31	0	0	0
(46 $<$ Ikä $<$ 58)$\times$ ikä	14,9	23,5	0	0	48
(Ikä $>$ 57)$\times$ ikä	6,6	18,7	0	0	0
(Ikä $<$ 28) $\times$ nainen	0,1	0,3	0	0	0
(Ikä $<$ 28) $\times$ ikä $\times$ nainen	1,9	6,4	0	0	0

4.4.3.2 Suomi: estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Esitän lyhyesti Suomessa syntyneiden nettovaikutusten mallinnusaskeleet. Suomessa syntyneiden nettovaikutusten mallinnusaskeleet poikkeavat hieman siitä, mitä on esitetty alaluvussa 4.3.3. Tämä aiheuttaa joitakin muutoksia myös nettovaikutusmallien parametreissa verrattuna Irakissa ja Somaliassa syntyneiden malleihin. Alla taulukossa 4.14 esitetyt Suomessa syntyneiden nettovaikutusten mallikohtaiset tulokset noudattavat seuraavaa rakennetta:

- **Malli 1**: Tyhjä hierarkkinen normaalimalli. Eli on sama kuin luvussa 4.3.3.1 esitetty malli 1.
- **Malli 2**: Tyhjä hierarkkinen Studentin t-malli. Eli on sama kuin luvussa 4.3.3.2 esitetty malli 2.
- **Malli 3**: Tyhjä hierarkkinen malli, jossa sekä havainnoilla (otosjakauma) että henkilötason satunnaisella vakiolla on Studentin t-jakauma. Mallissa 1 ja 2 satunnainen vakio on normaalijakautunut. Suomessa syntyneillä malli 3 poikkeaa siitä, mitä esitetty luvussa 4.3.3.3 kolmanneksi mallinnusaskeleeksi. Muiden kuin Suomessa syntyneiden osalta henkilötason satunnainen vakio on normaalijakautunut.
- **Malli 4**: Selittävät muuttujat hierarkkisessa Studentin t-mallissa ja vino Studentin t-jakauma otosjakaumana. Vastaa luvussa 4.3.3.3 esitettyä mallia 3, paitsi siltä osin että henkilötason satunnaisvakio on t-jakautunut ja että otosjakaumana on vino t-jakauma.
- **Malli 5**: Satunnaisvaikutusten lisääminen malliin. Vastaa luvuissa 4.3.3.4 ja 4.3.3.5 esitettyä mallia 4, paitsi siltä osin että henkilötason satunnaisvakio on t-jakautunut.

Suomessa syntyneiden osalta mallinnusaskeleet poikkeavat täten muiden syntymävaltioiden mallinnusaskeleista siinä, että Suomessa syntyneillä henkilöiden välillä vaihteleva vakiotermi on t-jakautunut, kun muiden maiden kohdalla tämä satunnaisvaikutus on jäänyt testauksen jälkeen normaalijakautuneeksi (ks. mallinnusaskel 2, luku 4.3.3.2).

Suomessa syntyneillä näin ollen mallinnusaskeleessa 2 henkilötason satunnaisvaikutus (on muiden syntymävaltioiden tavoin) normaalijakautunut, mutta Suomen osalta mallinnusaskel 3 on pitänyt käyttää siihen, että tyhjään malliin saadaan tämä satunnaisvaikutus myös t-jakautuneena. Suomessa syntyneiden osalta en myöskään mallintanut otosjakaumana toimivan vinon Studentin t-jakauman selittävillä muuttujilla ehdollista vinoutta, kuten tein Irakissa ja Somaliassa syntyneiden kohdalla (Irakin ja Somalian mallien mallinnusaskel 5). Syynä tähän oli, että koska en Irakin ja Somalian kohdalla päätynyt käyttämään ehdollisen vinouden mallia uusien maahanmuuttajien elinkaariennusteissa, ajattelin, etten tee myöskään niin Suomessa syntyneiden elinkaariennusteiden kohdalla.

Suomessa syntyneiden mallinnus poikkeaa myös siinä muiden syntymävaltioiden nettovaikutusten malleista, että maassaoloaikaa ja maahanmuuttovuotta ei ole lainkaan mallissa. Sen sijaan Suomen osalta selittäviä muuttujia sisältävissä nettovaikutusmalleissa on selittävänä muuttujana kalenterivuosi. Lopullinen Suomessa syntyneiden nettovaikutusmalli sisältää myös satunnaisvaikutuksia kalenterivuosien ”ta-

solla”.

Alla taulukossa 4.14 on esitetty perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit Suomessa syntyneille 20–62-vuotiaille. Suomessa syntyneiden nettovaikutusmallin estimoitavien parametrien selitteet on puolestaan suurimmaksi osaksi esitetty parametrien selitetaulukossa s. 170. Parametrien selitetaulukkoon verrattuna Suomessa syntyneiden nettovaikutusmallinnuksessa on seuraavat parametrit:

- **Satunnaisvaikutusten hajontaparametrit (τ):**
  - τ_{i1}: henkilötason t-jakautuneen satunnaisvaikutuksen hajontaparametri: vakio.
  - τ_{i2}: henkilötason satunnaisvaikutuksen keskihajonta: ikä^2.
  - τ_{i3}: henkilötason satunnaisvaikutuksen keskihajonta: kalenterivuosi.
  - τ_{k1}: kalenterivuositason satunnaisvaikutuksen keskihajonta: vakio.
  - τ_{k2}: kalenterivuositason satunnaisvaikutuksen keskihajonta: nainen.
  - τ_{k3}: kalenterivuositason satunnaisvaikutuksen keskihajonta: ikä^2.
- **Studentin t-jakaumien vapausasteparametrit (ν):**
  - ν_y: t-jakautuneen otosjakauman vapausasteparametri.
  - ν_{hlo}: t-jakautuneen henkilötason satunnaisen vakiotermin vapausasteparametri.

Mallin vakiotermiä, regressiokertoimia, satunnaisvaikutusten ja virhetermin keskihajontoja (τ-parametrit, σ) sekä vinousparametreja koskevat odotusarvoestimaatit ovat euromääräisiä. Autoregressiiviset kertoimet on ilmaistu prosentteina, ja t-jakaumien vapausasteilla (ν_y ja ν_{hlo}) ei puolestaan ole mitään erillistä yksikköä. En käsittele tässä yhteydessä tarkemmin taulukossa 4.14 esitettyjen parametrien odotusarvojen tulkintaa, sillä olen tehnyt tämän käytännössä jo Irakissa syntyneiden nettovaikutusmallinnustulosten yhteydessä (ks. taulukko 4.4, s. 177).

Taulukon 4.14 mallien ennustetarkkuuden mittarien (Waic ja LOO) mukaan parametrimäärältään suuremmat mallit ovat parempia kuin pienemmät mallit. Ennustarkkuuksien mittareiden Waic ja LOO estimaatit ovat lähellä toisiaan kolmen ensimmäisen mallin osalta. Mallien 4 ja 5 osalta, joissa otosjakaumana on vino Studentin t-jakauma, arvot lisätään möyhemmin viitattuun taulukkoon. Ne joka tapauksessa poikkeavat toisistaan, kuten ne poikkesivat myös Irakissa ja Somaliassa syntyneiden estimointituloksissa.

Taulukko 4.14: *Perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit, mallikohtaiset ennustetarkkuuden estimaatit ja keskihajonnat (Waic ja LOO) sekä sisäkorrelaatioestimaatti.*

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
Vakio	2 990	5 100	4 930	9 630	10 530
Nainen				-1 740	-3 350
Ikä				1 360	1 830
Ikä^2				-160	-590
Ikä $\times$ nainen				290	260
Ikä^2 $\times$ nainen				300	320
Kalenterivuosi				220	450
Kalenterivuosi $\times$ nainen				130	-40
Kalenterivuosi < 2001				-350	-130
Kalenterivuosi $\times$ (kalenterivuosi < 2001)				-100	490
Ikä $<$ 28				12 650	11 960
(Ikä $<$ 28) $\times$ ikä				10 070	9 750
(27 $<$ Ikä $<$ 37) $\times$ nainen				-1 690	-1 950
(27 $<$ Ikä $<$ 37) $\times$ ikä $\times$ nainen				620	-440
(36 $<$ Ikä $<$ 47) $\times$ nainen				-1 220	-900
(36 $<$ Ikä $<$ 47) $\times$ ikä $\times$ nainen				2 250	1 430
(46 $<$ Ikä $<$ 58)				290	380
Ikä $>$ 57				5 490	4 430
(46 $<$ Ikä $<$ 58) $\times$ ikä				-1 230	-1 700
(Ikä $>$ 57) $\times$ ikä				-8 110	-8 450
(Ikä $<$ 28) $\times$ nainen				860	140
(Ikä $<$ 28) $\times$ ikä $\times$ nainen				2 270	580
Nettovaikutus$_{t-1}$				44,3	24,9
Nettovaikutus$_{t-2}$				4,1	2,1
τ_{i1}	17 560	10 570	7 940	3 410	4 330
τ_{i2}					1 110
τ_{i3}					2 770
τ_{k1}					320
τ_{k2}					140
τ_{k3}					300
σ	18 390	3 140	3 170	2 900	2 260
ν_y		1,22	1,23	1,33	1,2
ν_{hlo}			3,92	3,04	2,88
γ				-100	-90
Sisäkorrelaatio	0,488				

Jatkuu seuraavalla sivulla

Taulukko 4.14 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3	Malli 4	Malli 5
Waic (estimaatti)	120 989	36 789	36 498	–	–
Waic (estimaatin keskivirhe)	(5 346)	(657)	(655)	(–)	(–)
LOO (estimaatti)	119 482	37 147	36 891	–	–
LOO (estimaatin keskivirhe)	(4 879)	(657)	(652)	(–)	(–)

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” mallin on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän Suomessa syntyvien elinkaariennusteissa työikäisten nettovaikutusten ennustamiseen. Valitsin lopulliseksi malliksi mallin 5.

Taulukossa 4.15 on esitetty Suomissa syntyneiden nettovaikutusmallin (malli 5) posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvojen (OA), keskihajontojen (KH), kvartiilien sekä jakaumien hännistä prosenttipisteiden 2,5 (%) ja 97,5 (%) estimaatit. Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli yhtä, en käsittele niitä tarkemmin.

Taulukko 4.15: *Suomessa syntyneiden 20–62-vuotiaiden nettovaikutusten lopullinen malli (malli 5). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin Monte Carlo -keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakio	10 530	410	9 730	10 260	10 540	10 820	11 310	15	1,00
Nainen	-3 350	490	-4 310	-3 680	-3 360	-3 040	-2 370	24	1,01
Ikä	1 830	260	1 300	1 650	1 830	2 010	2 330	9	1,00
Ikä^2	-590	180	-940	-710	-590	-470	-250	6	1,00
Ikä $\times$ nainen	260	220	-180	110	260	420	700	7	1,00
Ikä^2 $\times$ nainen	320	420	-490	40	320	600	1 140	16	1,01
Kalenterivuosi	450	120	210	380	450	530	700	4	1,00
Kalenterivuosi $\times$ nainen	-40	140	-320	-140	-40	50	220	5	1,00
Kalenterivuosi < 2001	-130	280	-710	-310	-130	50	440	8	1,00
Kalenterivuosi $\times$ (kalenterivuosi < 2001)	490	370	-250	250	490	730	1 220	19	1,00
Ikä $<$ 28	11 960	720	10 590	11 450	11 960	12 440	13 420	32	1,00
(Ikä $<$ 28) $\times$ ikä	9 750	600	8 620	9 330	9 750	10 150	10 970	27	1,00
(27 $<$ Ikä $<$ 37) $\times$ nainen	-1 950	490	-2 880	-2 290	-1 950	-1 620	-970	14	1,00
(27 $<$ Ikä $<$ 37) $\times$ ikä $\times$ nainen	-440	1 710	-3 630	-1 620	-440	710	2 980	67	1,01
(36 $<$ Ikä $<$ 47) $\times$ nainen	-900	450	-1 780	-1 200	-890	-590	-10	16	1,01
(36 $<$ Ikä $<$ 47) $\times$ ikä $\times$ nainen	1 430	1 000	-530	750	1 440	2 100	3 410	38	1,01
(46 $<$ Ikä $<$ 58)	380	200	0	240	380	510	770	7	1,00
Ikä $>$ 57	4 430	2 770	-1 040	2 510	4 430	6 320	9 880	94	1,00
(46 $<$ Ikä $<$ 58)$\times$ ikä	-1 700	460	-2 620	-2 020	-1 700	-1 390	-810	17	1,00
(Ikä $>$ 57)$\times$ ikä	-8 450	2 720	-13 950	-10 260	-8 480	-6 580	-3 170	103	1,00
(Ikä $<$ 28) $\times$ nainen	140	1 330	-2 510	-730	150	1 040	2 740	55	1,00
(Ikä $<$ 28) $\times$ ikä $\times$ nainen	580	2 390	-4 050	-1 060	610	2 180	5 410	101	1,01
Nettovaikutus$_{t-1}$	24,9	0,5	23,9	24,6	24,9	25,3	26	0,04	1,02
Nettovaikutus$_{t-2}$	2,1	0,3	1,6	1,9	2,1	2,3	2,6	0,01	1,00
τ_{i1}	4 330	120	4 080	4 250	4 330	4 410	4 550	11	1,02

Jatkuu seuraavalla sivulla

Taulukko 4.15 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
τ_{i2}	1 110	80	950	1 070	1 110	1 160	1 260	6	1,02
τ_{i3}	2 770	60	2 650	2 730	2 770	2 810	2 890	4	1,01
τ_{k1}	320	110	170	250	300	380	570	5	1,00
τ_{k2}	140	40	90	120	140	160	230	1	1,00
τ_{k3}	300	70	190	250	290	340	440	2	1,00
σ	2 260	20	2 220	2 250	2 260	2 280	2 310	1	1,00
ν_y	1,2	0,01	1,17	1,19	1,2	1,21	1,23	0,001	1,00
ν_{hlo}	2,88	0,17	2,56	2,76	2,88	2,99	3,23	0,015	1,04
γ	-91	5	-101	-95	-91	-88	-82	0,4	1,02

Listaan lyhyesti joitakin huomioita keskeisimmistä vaikutuksista ja estimointituloksista taulukossa 4.15. Kun kirjoitan parametrin sijoittuvan jollekin tietylle intervallille, tarkoitan tällöin kyseisen parametrin posteriorijakauman 95%:n uskottavuusintervallia 2,5 %:sta 97,5 %:iin.

- **Sukupuoli**. Päävaikutus naisilla verrattuna miehiin noin -4 300:n ja -2 400:n välillä.
- **Kalenterivuosi**. Kalenterivuoden päävaikutuksen mukaan nettovaikutusten odotusarvo kasvaa noin 5 vuodessa (kalenterivuoden keskihajonta aineistossa) noin 210–700 euroa. Ennen vuotta 2001 nettovaikutusten odotusarvot ovat olleet odotusarvoisesti negatiivisempia kuin vuonna 2001 ja sen jälkeen. Posterioriennustetarkistusten alla esitettävien kuvioiden perusteella kalenterivuosiin liittyvä suurin vaikutus on se, onko kalenterivuosi ennen vuotta 2001 vai ei. Nettovaikutusten trendiä kalenterivuosien suhteen ei vaikuttaisi havaintoaineistoon perustuvien kuvioiden perusteella olevan.
- **Muut regressiokertoimet**. Mallissa on ikäluokkaindikaattoreita sekä niiden interaktioita iän ja naisindikaattorin kanssa. Iän ja sukupuolen yhteisvaikutuksia havainnollistavat posterioriennustetarkistusten yhteydessä esitettävät kuvat.
- **Autoregressiiviset termit**. Ensimmäisen autoregressiivinen termin jakauma keskittynyt 0,25:n ympäristöön. Toisen termin jakauma keskittynyt 0,02:n ympäristöön eli lähelle nollaa.
- **Satunnaisvaikutukset**. Henkilöittäin vaihtelevan t-jakautuneen mallin vakiotermin hajontaparametrin (τ_{i1}) odotusarvo on noin 4 330 euroa, mikä tarkoittaa t-jakauman ominaisuudet huomioiden sitä, että kyseessä olevan satunnaisvaikutuksen keskihajonnan odotusarvo on noin 14 170 euroa. Kalenterivuosiin liittyvän satunnaisen normaalijakautuneen ”vakion” odotusarvo on puolestaan noin 320 euroa. Tämä tarkoittaa sitä, että henkilöiden välinen havaitsematon heterogeenisuus nettovaikutusten tasossa aiheuttaa elinkaariennusteisiin hieman enemmän epävarmuutta kuin ajassa kumuloituva havaitsematon heterogeenisuus kalenterivuosien nettovaikutusten tasossa. Tämä seuraa siitä, että työikäisten nettovaikutusten ennusteet laaditaan 43 (ikä)vuodelle, ja 43 $\times$ 320 (euroa) = 13 760 (euroa) on samassa kokoluokassa oleva kumuloitunut epävarmuuden määrä kuin henkilöiden välisistä havaitsemattomista eroista odotusarvoisesti johtuva 14 170 euroa.
- **Muut parametrit**. Sekä otosjakauman että henkilötason satunnaisen vakion Studentin t-jakauman vapausasteet ovat ”pieniä”: ν_y-parametrin posteriorijakauma on keskittynyt noin 1,20 ympäristöön, ja ν_{hlo} on 2,56:n ja 3,23:n välillä. Otosjakauma on näin ollen lähestulkoon Cauchy-jakautunut, sillä Studentin t-jakauma vakioarvolla ν=1 on Cauchy-jakauma. Nettovaikutusten otosjakauma on vasemmalle vino, koska γ on negatiivinen.

4.4.3.3 Suomi: posterioriennustetarkistukset

Posterioriennusteet on muodostettu yllä esitetyn mallin 5 (eli elinkaariennusteissa käytettävän) parametrien posteriorijakaumien estimaateista ja havaintoaineistosta niin, kuin on esitetty maassaoloaikojen mallinnuksen yhteydessä luvussa 3.2.6. Olen laatinut posterioriennusteet niin, kuin kyse olisi uusien maahanmuuttajien aineistosta, joten poimin posterioriennusteita varten myös ennusteet syntyvien eli uusien suomalaisten satunnaisvaikutuksille.

Rajoitteen asettaminen apumuuttujalle W, jotta nettovaikutusten posterioriennustejakaumalla olisi odotusarvo. Taulukon 4.15 parametrien posterioriestimaattien mukaan vinoa Studentin t-jakaumaa noudattavan otosjakauman vapausasteparametrin ν_y posteriorijakauma on keskittynyt 1,2:n ympäristöön. Aiemmin luvussa 4.3.3.5 esitettiin, miten vino Studentin t-jakauma saadaan muodostettua apumuuttujan W avulla, jolla on käänteisgamma-jakauma: $W \sim inv.gamma(\nu_y/2, \nu_y/2)$ (ks. lauseke 4.13, s. 154).

Käänteisgamma-jakauman ominaisuuksia ja *iteroidun odotusarvon* kaavaa soveltamalla huomataan, että vinolla Stundentin t-jakaumalla ei ole odotusarvoa, mikäli $\nu_y < 2$. Koska taulukon 4.15 tulosten mukaan ν_y:n jakauma on keskittynyt 1,2:n ympäristöön (eli alle 2:n), näin ollen Suomessa syntyneiden nettovaikutusten posterioriennustejakaumalla ei ole odotusarvoa. Samalla myöskään Suomessa syntyvien henkilöiden elinkaariennusteilla ei ole odotusarvoa.

Jotta Suomessa syntyneiden työikäisten nettovaikutusten posterioriennusteiden posterioriennustejakaumalla (ja siten elinkaariennusteille) saataisiin odotusarvo, asetin apumuuttajan W jakaumalle rajoitteen. Kun käänteisgamma-jakautuneelle apumuuttuja W:lle asetetaan rajoite, sille saadaan odotusarvo (vaikka $\nu_y < 2$), ja näin ollen iteroidun odotusarvon perusteella myös nettovaikutusten posterioriennustejakaumalla on odotusarvo.

Rajoitin W:n jakaumaa niin, että posterioriennustejakaumassa mahdolliset W:n arvot olivat pienempiä kuin W:n nettovaikutusten lopullisessa mallissa estimoidun posterioriennustejakauman 99,95:s prosenttipiste (mikä tarkoittaa W:n arvona noin 3 800:aa). Leikkasin W:n estimoidusta jakaumasta siten ylimmän noin puolen promillen pois. Päädyin tähän rajoitteeseen kokeilemalla sitä, miten hyvin eri W:n rajoitteen tuottamat nettovaikutusten posterioriennustejakaumat vastaavat nettovaikutusten havaintoaineiston jakaumaa.

Käytännössä rajoittamattoman W:n jakauma näkyi posterioriennustetarkistuksissa niin, että nettovaikutusten posterioriennustejakauma sisälsi havaintoaineistoa enemmän äärimmäisiä nettovaikutusennusteita/-havaintoja. Kun W:lle asetettiin mainittu rajoite, myös nettovaikutusten posterioriennustejakauman hännät vastasivat melko hyvin nettovaikutusten havaintoaineiston jakaumaa alle 0,1:nnen prosenttipisteen ja yli 99,9:nnen prosenttipisteen alueilla.

Posterioriennustetarkistus: Kaikkien nettovaikutushavaintojen yhteisjakauma ja henkilöiden nettovaikutusten keskiarvojen yhteisjakauma.

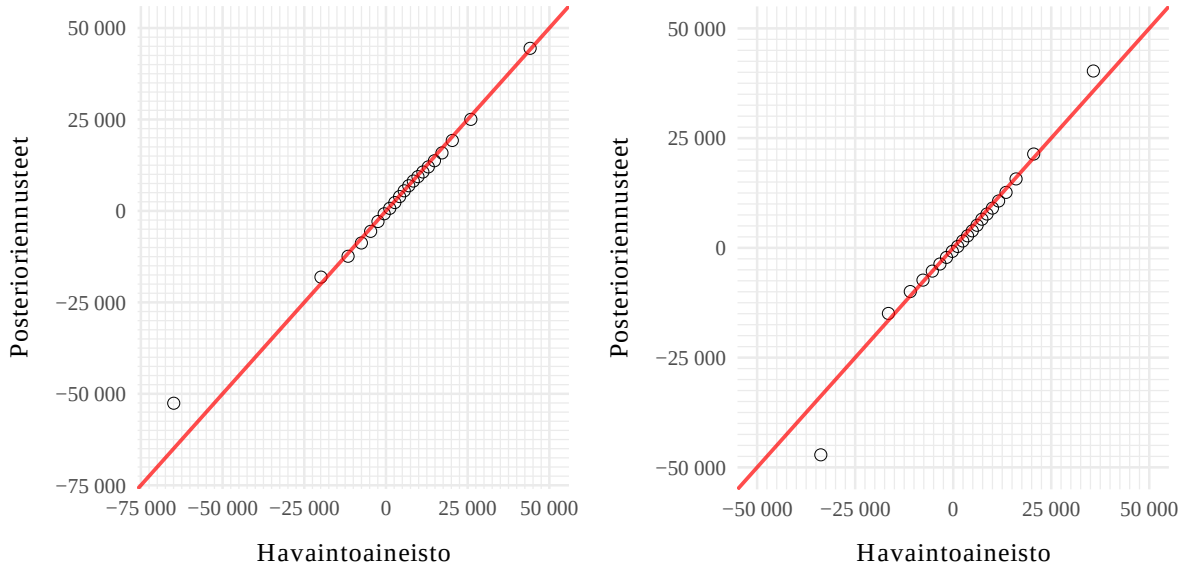
Aloitan posterioriennustetarkistukset vertaamalla havaintoaineiston kaikkien nettovaikutushavaintojen

muodostamaa yhteisjakaumaa posterioriennustejakauman kaikkien ennusteiden muodostamaan yhteisjakaumaan. Ensimmäisen mallinnustavoitteen mukaan näiden jakaumien pitäisi vastata mahdollisimman hyvin toisiaan, erityisesti jakaumien tärkeimpien tunnuslukujen kuten keskiarvojen, keskihajontojen ja kvartiilien suhteen.

Tämän lisäksi vertaan kaikkien havaintoaineiston henkilöiden nettovaikutusten keskiarvojen jakaumaa ennustettuihin henkilöiden nettovaikutusten keskiarvojen jakaumaan. Nämä jakaumat muodostetaan havaintoaineistossa laskemalla nettovaikutusten keskiarvot kaikille aineiston henkilöille ja muodostamalla näistä jakauma. Posterioriennusteiden kohdalla lasketaan ennustettujen nettovaikutusten keskiarvot kaikille henkilöille, ja muodostetaan näistä jakauma. Henkilötasoisten jakaumien vertailun tarkoituksena on osoittaa, että elinkaariennusteissa käytettävä nettovaikutusten malli tuottaa luotettavia ennusteita paitsi havaintotasolla niin myös henkilötasolla.

Tarkistus kvantiilikuvioilla. Vertaan alla kuviossa 4.20 havaintoaineiston jakaumien prosenttipisteitä vastaavien posterioriennustejakaumien prosenttipisteisiin. Jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeama vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).



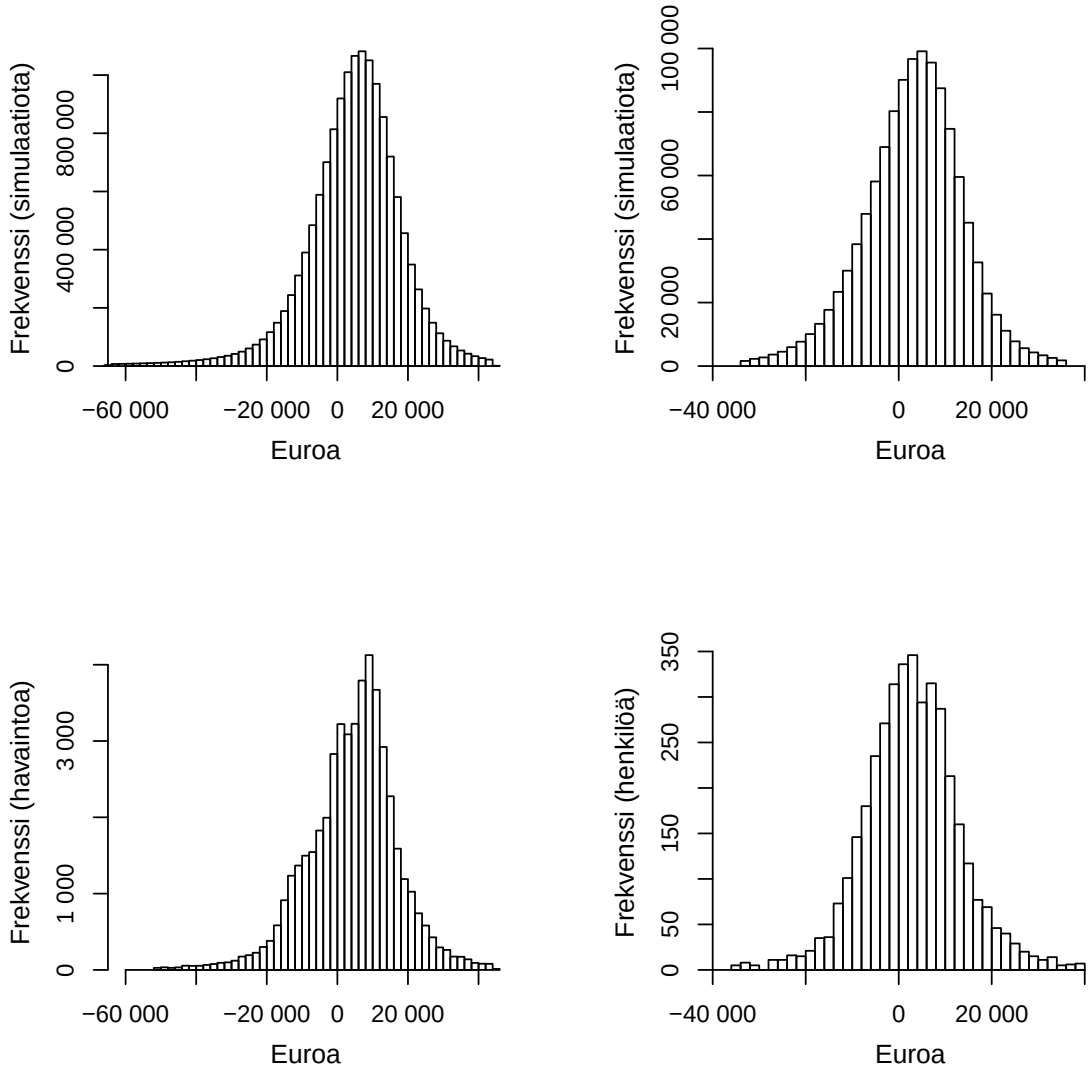
(a) *Kaikki havainnot ja ennusteet.*

(b) *Henkilöiden keskiarvot.*

Kuva 4.20: *Havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

Kuva 4.20 osoittaa, että posterioriennusteiden tuottamat jakaumat vastaavat hyvin havaintoaineiston jakaumia sekä kaikkien havaintojen osalta että henkilöiden nettovaikutusten keskiarvojen osalta lukuun ottamatta sekä henkilöiden että havaintojen osalta alimpia prosenttipisteitä sekä henkilöiden keskiarvojen jakaumaa koskien myös ylintä prosenttipistettä.

Tarkistus histogrammeilla. Verrataan nyt havaintoaineiston ja posterioriennusteiden jakaumia prosenttipisteiden vertailukuvioiden (Q-Q plot) lisäksi histogrammien avulla. Nämä on esitetty kuvassa 4.21. Kuvan ylärivillä on kaksi posterioriennusteista muodostettua jakaumaa, ja alarivillä puolestaan niitä vastaavat havaintoaineiston jakaumat. Histogrammien vertailu osoittaa, että posterioriennustejakaumat vastaavat hyvin havaintoaineiston jakaumia. Tosin sekä kaikkien ennusteiden että henkilöiden keskiarvojen ennusteiden katkaistuissa jakaumissa vasen häntä on hieman ”pitempi” kuin havaintoaineiston vastaavissa jakaumissa. Yllä kuvassa 4.20 tämä näkyy ennustettujen alimpien prosenttipisteiden poikkeamisena havaituista.



Kuva 4.21: *Posterioriennusteiden ja havaintoaineiston jakaumat histogrammeina. Posterioriennusteiden ennusteet ja ennustetut henkilöiden nettovaikutusten keskiarvot (ylärivi). Havaintoaineiston havainnot ja havaitut henkilöiden nettovaikutusten keskiarvot (alarivi). Kaikista jakaumista on katkaistu pois alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Tarkistus testisuureiden avulla. Alla taulukossa 5.16 on esitetty havaintoaineiston ja posterioriennusteiden jakaumien vertailut euromääräisinä lukuina sekä *Bayes-p-arvojen* ja eri euromääräisten *Bayes-intervallien* avulla. Olen käsitellyt näiden testisuureiden merkitystä ja laskemista alaluvussa 4.3.7. Taulukossa 5.16 verrataan posterioriennustejakaumien (havaintojen ja henkilöiden keskiarvojen jakaumat) keskiarvoja, keskihajontoja sekä joitakin prosenttipisteitä havaintoaineiston vastaaviin jakauman tunnuslukuihin.

Taulukon 5.16 ensimmäisellä ja toisella rivillä ovat posterioriennustejakaumien ja havaintoaineiston jakaumien tunnuslukujen estimaatit. Bayes-p-arvot ja Bayes-intervallit eri euromäärille ovat testisuureita estimaattien vastaavuuksille havaintoaineiston ja posterioriennusteiden välillä. Bayes-intervalli ilmaisee todennäköisyyden sille, että posterioriennustejakauman estimaatti poikkeaa korkeintaan ilmoitetun euromäärän verran havaintoaineiston estimaatista. Kaikkien havaintojen jakauman osalta posterioriennusteiden estimaatit eivät poikkea yli 0,50:n todennäköisyydellä havaintoaineiston estimaateista 1 000 euroa enempää lukuun ottamatta 1:stä, 5:ttä ja 99:ttä prosenttipistettä.

Taulukko 4.16: *Havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien keskiarvot (KA), keskihajonnat (KH) ja prosenttipisteet:* $1, 5, 10, 25, 50, 75, 90, 95, 99$. *Vertailtavien estimaattieen Bayes-p-arvot ja Bayes-intervallit eri euromäärillä.*

Kaikki havainnot ja ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %
Posterioriennusteet	4 180	24 440	-64 950	-19 890	-11 600	-2 450	5 580	12 950	20 290	25 970	44 110
Aineiston havainnot	3 900	23 920	-52 560	-18 080	-12 390	-2 890	5 500	12 040	19 270	25 030	44 460
Bayes-p-arvo	0,8	0,53	0	0	0,99	0,99	0,67	1	1	0,97	0,41
Bayes-voima: 2 000 euroa	1	0,58	0	0,62	1	1	1	1	1	0,98	0,66
Bayes-voima: 1 000 euroa	0,98	0,32	0	0,06	0,72	1	1	0,68	0,46	0,57	0,37
Bayes-voima: 500 euroa	0,75	0,16	0	0,01	0,18	0,62	1	0,02	0,04	0,2	0,19
Bayes-voima: 250 euroa	0,44	0,08	0	0	0,02	0,17	0,82	0	0	0,07	0,08

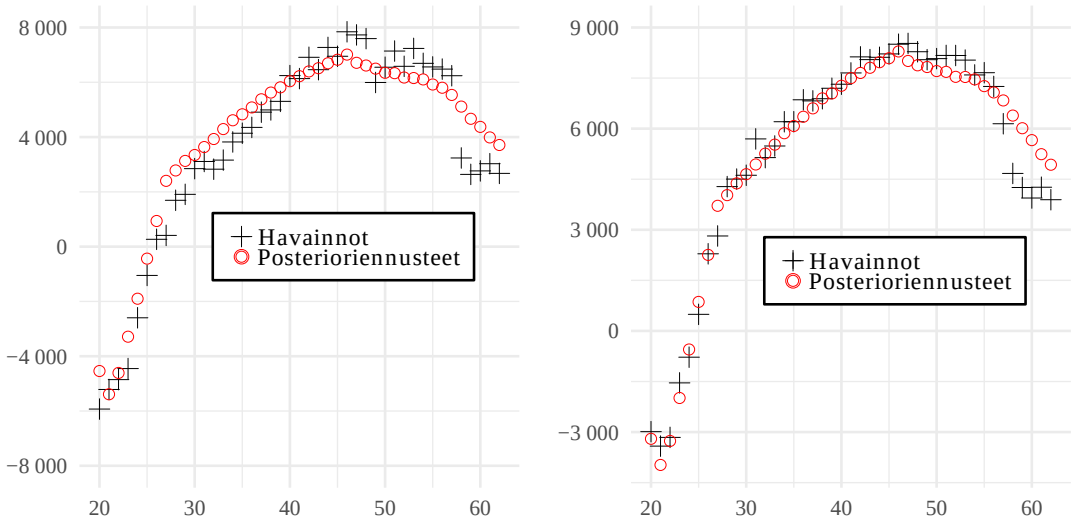
Henkilöiden keskiarvot ja niiden ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %
Posterioriennusteet	3 240	18 360	-33 770	-16 520	-10 950	-3 390	3 660	10 010	16 020	20 520	35 740
Aineiston havainnot	2 810	20 860	-47 140	-14 930	-9 940	-3 670	2 750	9 060	15 710	21 380	40 300
Bayes-p-arvo	0,92	0,16	1	0	0,01	0,86	1	1	0,84	0,05	0,03
Bayes-voima: 2 000 euroa	1	0,33	0	0,77	0,99	1	1	1	1	0,99	0,1
Bayes-voima: 1 000 euroa	0,96	0,17	0	0,18	0,47	1	0,68	0,59	0,98	0,59	0,04
Bayes-voima: 500 euroa	0,61	0,07	0	0,03	0,1	0,79	0,04	0,01	0,73	0,21	0,02
Bayes-voima: 250 euroa	0,27	0,05	0	0,02	0,04	0,41	0	0	0,39	0,08	0,01

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan.

Taulukon 5.16 jälkeen vertaan posterioriennusteita havaintoaineistoon tärkeimpien selittävien muuttujien luokkien mukaan.

Vertaan posterioriennusteiden keskiarvoja ja mediaaneja selittävien luokkien mukaan vastaaviin havaintoaineiston keskiarvoihin ja mediaaneihin. Esimerkiksi iän kohdalla vertaan keskiarvoja ja mediaaneja niin, että keskiarvot ja mediaanit on laskettu posterioriennusteista jokaiselle mallinnusaineiston iälle erikseen 20-vuotiaista 62-vuotiaisiin. Keskiarvojen lisäksi esitän mediaanit, koska nettovaikutusten jakauman leveiden häntien vuoksi mediaani on oletettavasti keskiarvoa robustimpi jakaumien keskiluku. Kuvioiden jälkeen esitän jokaiseen selittävien muuttujien luokkien suhteen tehtyyn posterioriennustetarkistukseen liittyen tärkeimpiä huomioita.

Posterioriennustetarkistus iän mukaan.

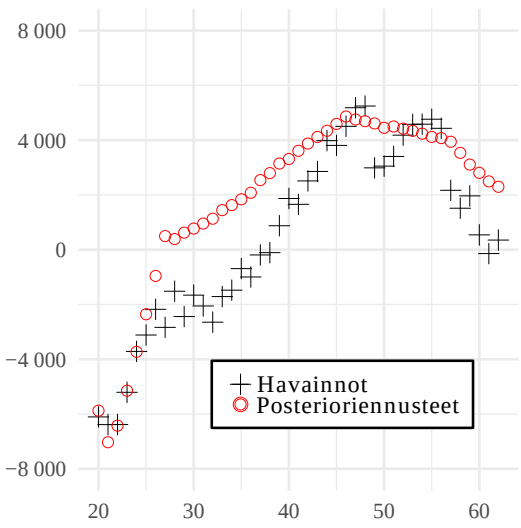


(a) *Keskiarvot iän mukaan.*

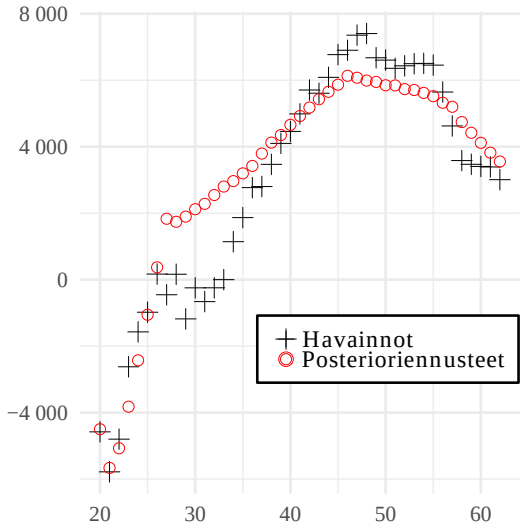
(b) *Mediaanit iän mukaan.*

Kuva 4.22: *Havainnot ja posterioriennusteet iän mukaan.*

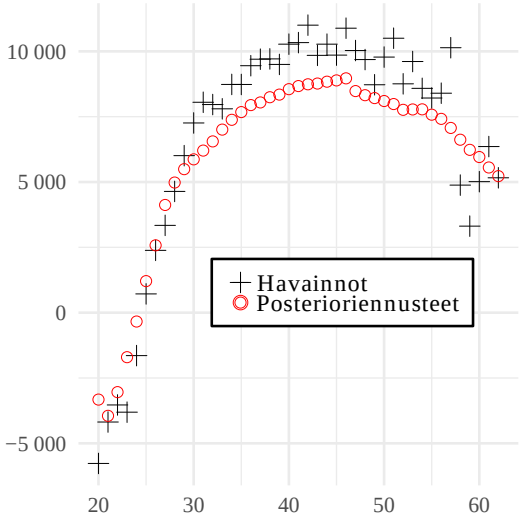
Posterioriennustetarkistus iän ja sukupuolen mukaan.



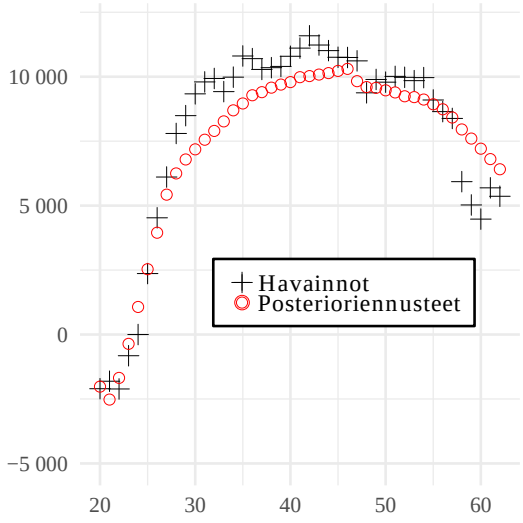
(a) *Keskiarvot iän mukaan. Naiset.*



(b) *Mediaanit iän mukaan. Naiset.*



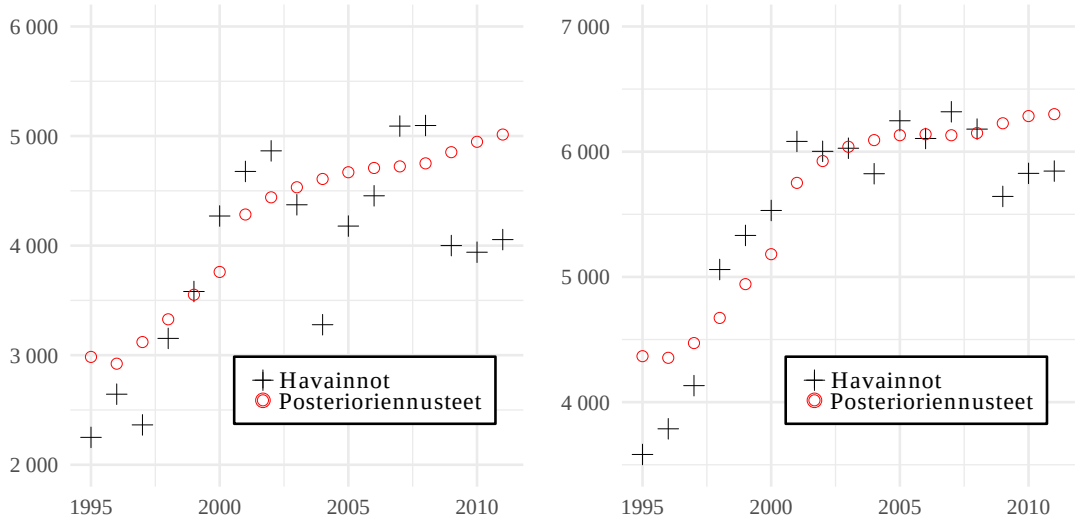
(c) *Keskiarvot iän mukaan. Miehet.*



(d) *Mediaanit iän mukaan. Miehet.*

Kuva 4.23: *Havainnot ja posterioriennusteet iän ja sukupuolen mukaan.*

Posterioriennustetarkistus kalenterivuoden mukaan.



(a) *Keskiarvot kalenterivuoden mukaan.* (b) *Mediaanit kalenterivuoden mukaan.*

Kuva 4.24: *Havainnot ja posterioriennusteet kalenterivuoden mukaan.*

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan. Havaintoja kuvista 4.22–4.24.

- **Iän mukaan** (kuva 4.22). Posterioriennusteet vastaavat hyvin erityisesti havaintoaineiston ikävuosikohtaisia mediaaneita. Tosin ennustetut mediaanit ovat havaittuja mediaaneita noin 1 000 euroa pienempiä 58–62-vuotiailla.
- **Posterioriennustetarkistus iän ja sukupuolen mukaan** (kuva 4.23). Naisten osalta ikäluokittaiset ennusteiden keskiarvot ja ennusteet ovat liian suuria verrattuna havaintoaineistoon n. 25–45-vuotiaina. Yli 56-vuotiailla naisilla ennustetut keskiarvot ovat havaittuja suurempia. Miehillä ennustetut keskiarvot jäävät havaittuja pienemmiksi 30–55-vuotiailla, ja myös mediaanit havaittuja pienemmiksi 27–45-vuotiailla Mallia voisi kehittää edelleen näiden poikkeamien osalta: Koska malliin on jo lisätty erilaisia ikäluokittaisia sukupuolen mukaan vaihtelevia regressiokertoimia, parannusta voisi etsiä esimerkiksi vinouden mallinnuksesta ikäluokan ja sukupuolen mukaan.
- **Posterioriennustetarkistus kalenterivuoden mukaan (1)** (kuva 4.24). Kalenterivuosittaisissa keskiarvoissa havainnot jakaantuvat nettovaikutusten tason perusteella kahteen joukkoon: keskiarvoihin vuonna 2000 ja sen jälkeen sekä keskiarvoihin ennen vuotta 2000. Kalenterivuosittaiset mediaanit ovat puolestaan pysyneet suunnilleen samassa tasossa vuosina 2001–2011. Toisin sanoen ainakaan 2000-luvulla aineiston perusteella ei ole havaittavissa mitään (nousevaa tai laskevaa) trendiä kalenterivuosien suhteen. Lopullisen mallin mukaan (taulukko 4.15) nettovaikutuksilla olisi

olemassa nouseva trendi kalenterivuoden päävaikutuksen mukaan, mutta kun huomioidaan kalenterivuosien interaktiotermi (alle vuoden 2001 olevien kalenterivuosien luokan kanssa), niin myös näiden estimointitulosten mukaan positiivinen trendi keskittyy kalenterivuosiin ennen vuotta 2000. Tämä näkyy kuviossa 4.24 siinä, että sekä havaitut että ennustetut nettovaikutusten keskiarvot ja mediaanit vaihtelevat suunnilleen saman tason ympärillä vuosina 2001–2011 – eli kalenterivuosien suhteen ei ole trendiä.

- **Posterioriennustetarkistus kalenterivuoden mukaan (2)** (kuva 4.24). Se, että nettovaikutusten kalenterivuosittaiset keskiarvot ja mediaanit ovat olleet 2000-luvulla korkeammat kuin ennen vuotta 2001, selittyy tutkimusaineistoon liittyvillä tekijöillä. Tutkimuksen I osan *saatujen tulonsiirtojen* aikasarjakuvion perusteella saatujen tulonsiirtojen mediaani on äkisti pienentynyt pysyvästi vuosien 2000 ja 2001 välillä lähes 2 000 eurolla. Lisäksi saatujen tulonsiirtojen keskiarvo on pienentynyt noin 1 000 eurolla. (Salminen (2015a), kuvio 8, s. 25).

  Nämä järjestelmälliset muutokset saaduissa tulonsiirroissa – ja siten julkisen talouden nettovaikutuksissa – vastaavat täsmällisesti kuvassa 4.24 nähtäviä järjestelmällisiä eroja nettovaikutusten tasossa (mediaanien osalta vuodet 2001–2011 verrattuna ennen vuotta 2001). En tiedä, mikä tekijä tutkimusaineistossa aiheuttaa hyppäyksen saaduissa tulonsiirroissa ja siten nettovaikutuksissa, mutta mikäli tämä ”tämä” hyppäys ei ole todellinen (vaan on selitettävissä esimerkiksi saatuja tulonsiirtoja koskevilla muutoksilla Tilastokeskuksen rekistereissä), julkisen talouden nettovaikutukset ovat pysyneet suunnilleen samalla tasolla koko tutkimusaineiston havaintojakson ajan eli vuosina 1995–2011.

Luku 5

Palkkasummien mallinnus

5.1 Mallinnustavoitteet

Työikäisten 18–62-vuotiaiden palkkasummien mallintamisen tavoitteena on laatia maahanmuuttajaryhmäkohtaiset mallit, joita voidaan käyttää julkisen talouden elinkaarivaikutusten maahanmuuttajaryhmäkohtaisissa uskottavissa ennusteissa. Palkkasummalla tarkoitetaan tässä yhteydessä henkilön (palkansaajan) vuoden aikana saamien niiden palkkojen summaa, josta maksetaan pakolliset työeläkemaksut. Pakollisilla työeläkemaksuilla tarkoitetaan tässä yhteydessä puolestaan henkilöiden (palkansaajien) sekä heidän työnantajiensa maksamien pakollisten eli lakisääteisten työeläkemaksujen summaa. Käytän pakollisista eläkemaksuista myös lyhennettyä ilmaisua: eläkemaksut. Lisäksi ”henkilöillä” viittaan palkansaajiin eläkemaksujen yhteydessä.

Käytän palkkasummaa koskevia mallinnustuloksia lakisääteisten työeläkemaksujen ennustamiseen. Lakisääteisiä työeläkemaksuja koskevia ennusteita käytän puolestaan eläkeikäisten (yli 62-vuotiaat) julkisen talouden nettovaikutusten ennustamiseen. Lisäksi käytän tuloksia julkisen talouden nettovaikutusten ennustamiseen työikäisille 18–62-vuotiaille työantajien maksamien eläkemaksujen ja muiden sosiaalivakuutusmaksujen osalta, sillä ne eivät sisälly julkisen talouden nettovaikutusten aineistoon ja sen pohjalta tehtyihin malleihin.

Lasken havaitut palkkasummat lakisääteisiä eläkemaksuja koskevasta rekisteriaineistosta, kuten jäljempänä esitän. En mallinna lakisääteisiä eläkevakuutusmaksuja kuitenkaan suoraan, sillä niiden arvo riippuu paitsi palkkojen kehityksestä myös vaihtuvista työeläkemaksuprosenteista, eli niistä osuuksista, joilla palkoista maksetaan työeläkemaksuja. Ennustan tulevat työeläkemaksut palkkasummien, niiden reaalisen kasvun sekä työeläkemaksuprosenttien ennusteiden avulla niin kuin esitän luvussa 6. Tässä luvussa keskityn kuitenkin pelkkien palkkasummien mallintamiseen.

”Irrotan” palkkasummahavainnot aineistossa niiden reaalisesta kehityksestä Tilastokeskuksen ansiotasoindekseillä. Toisin sanoen muunnan vuosina 1995–2014 havaitut palkkasummat vuoden 2011 palkkatasoon. Näin voin laatia erikseen ennusteet palkkasummille, palkkasummien reaaliselle kasvulle sekä eläkemaksuprosenteille.

Uusien maahanmuuttajien palkkasummaennusteet. Laadin elinkaarivaikutusten ennusteet siten, että ulkomailla syntyneiden osalta ennusteet tehdään henkilöiden Suomeen muuton vuoden tietojen perusteella ja Suomessa syntyneiden osalta koko eliniälle syntymästä lähtien. Ulkomailla syntyneiden osalta ennusteet perustuvat siten sille, mitä heistä tiedetään, kun he muuttavat ensimmäisen kerran Suomeen. Koska käytän ennusteita palkkasummista – ja siten maksetuista pakollisista työeläkemaksuista – osana elinkaarivaikutusten ennustamista, myös pakollisten työeläkemaksujen ennusteet perustuvat siihen, mitä ulkomailla syntyneistä tiedetään, kun he muuttavat ensimmäisen kerran Suomeen. Palkkasummamallien aikainvariantit – eli ajassa muuttumattomat – selittävät muuttujat perustuvat siten ensimmäisen muuttovuoden tietoihin.

Palkkasummamallinnuksen tavoitteet. Esitän seuraavaksi palkkasummamalleille kriteerit, jotka niiden tulee täyttää, jotta ne voivat toimia uskottavina malleina palkkasummien ennusteille. Kriteerit ovat käytännössä samat kuin nettovaikutusmallien yhteydessä: palkkasummamallien on toistettava alkuperäinen aineisto, mallien on huomioitava otanta-asetelma ja malleissa on oltava mukana aikaan liittyvät tärkeimmät vaikutukset. Koska nämä mallinnustavoitteet ovat samanlaiset kuin nettovaikutusmallien yhteydessä (edellä luvussa 4.1), en toista niitä tässä.

Suurin ero mallinnustavoitteissa liittyy siihen, että mallien on toistettava alkuperäinen aineisto. Nettovaikutusaineistot olivat symmetrisiä ja yksihuippuisia, joten niiden kohdalla tärkeintä oli toistaa alkuperäinen jakauma erityisesti jakaumien huippujen ympäristössä, johon suurin osa havainnoista oli keskittynyt.

Palkkasummien ja eläkemaksujen aineistojen jakaumat sen sijaan koostuvat epäjatkuvasta sekä jatkuvasta komponentista. Epäjatkuva komponentti koostuu niistä havainnoista, joissa henkilöiden vuosittainen eläkemaksujen summa on tasan nolla euroa, eli maksuja ei ole. Jatkuva komponentti puolestaan koostuu niistä havainnoista, joissa vuosittainen palkkasummien ja eläkemaksujen summa on positiivinen eli yli nolla euroa.

Yli nollan olevien havaintojen jakauman massa on keskittynyt sen vasempaan laitaan eli positiivisiin reaalilukuihin nollan ”lähellä”. Tästä syystä palkkasummien kohdalla tavoite, jonka mukaan mallien on toistettava alkuperäinen aineisto, tarkoittaa ensisijaisesti sitä, että nollien osuus tulee mallinnettua oikein sekä sitä, että yli nollien osalta mallinnetun jakauman tuottamat ennusteet vastaavat havaintoaineiston jakaumaa lukuun ottamatta oikean hännän ääriarvoja. Käsittelen tarkemmin palkkasummien havaittuja jakaumia sekä niiden mallinnusta jäljempänä. Maahanmuuttajaryhmäkohtaiset havaintoaineistot on esitetty mallinnustulosten yhteydessä.

5.2 Palkkasummien havaintoaineisto

5.2.1 Palkkasummien laskeminen eläkemaksuja koskevista rekisteritiedoista

Laskin palkkasummat pakollisia eläkemaksuja koskevista rekisteritiedoista. Tiedot tutkimusaineistoon sisältyvien henkilöiden maksamien pakollisten vuosittaisten eläkevakuutusmaksujen summista tulevat Tilastokeskuksen hallinnoimista rekistereistä. Tutkimuskäytössäni ollut rekisteritieto sisälsi henkilöiden kaikkien nykyisten työeläkelakien mukaan maksamat pakolliset eläkevakuutusmaksut ja työttömyysvakuutusmaksut. Rekisteritieto on peräisin työnantajien vuosittain ilmoittamista tai verotoimiston tallentamista (ammatinharjoittajat, freelancerit) palkansaajien maksamista eläke- ja työttömyysvakuutusmaksuista koskien pää- ja sivutoimen palkkoja, merityötuloa, sijaismaksajan maksamaa palkkaa tai palkkaturvaa sekä esimerkiksi koskien kunnallisen perhepäivähoitajan tai omaishoitajan palkkoja. Yrittäjien YEL- ja MyEL-lakien mukaan vakuutetuilta yrittäjiltä tietoon sisältyy työtuloista maksetut pakolliset eläke- ja työttömyysvakuutusmaksut. (Tosin viitatussa keskustelussa Tilastokeskuksen virkamies kirjoittaa tässä yhteydessä yrittäjän "palkasta", mutta oletan, että kyse on sittenkin yrittäjän työtuloista). (Matala (2016)).

Rekisteritiedot ovat aineistossa käytössä koskien vuosia 2005–2014. Tiedot ennen vuotta 2005 ovat hyvin puutteellisia, joten en sisällyttänyt niitä mallinnusaineistoon. Puutteet rekisteritiedoissa ennen vuotta 2005 johtuvat Tilastokeskuksen mukaan todennäköisesti siitä, että ennen vuotta 2005 tieto kattoi vain verovelvollisten itse ilmoittamat maksut. Vuodesta 2005 lähtien tieto on sisältänyt myös työnantajien ilmoittamat maksut. (Matala (2016)).

Koska palkansaajan lakisääteisiä maksuja koskeva rekisteritieto sisältää sekä eläke- ja työttömyysvakuutusmaksu, siitä on palkkasumman laskemiseksi erotettava osuus, joka koskee vain eläkemaksuja. Tämä palkansaajan vuosittain maksama eläkevakuutusmaksu saadaan eläke- ja työttömyysmaksuja koskevasta rekisteritiedosta, $maksu_{rek}$, seuraavasti:

$$\textit{eläkemaksu} = maksu_{rek} \times \frac{osuus_{el}(i)}{osuus_{el}(i) + osuus_{tvm}}. \tag{5.1}$$

Tässä yhtälössä *eläkemaksu* on vuosittainen palkansaajan (henkilön) maksama eläkevakuutusmaksu, $maksu_{rek}$ on alkuperäinen rekisteritieto, joka sisältää sekä palkansaajan vuosittaisen eläkemaksu tiedon että työttömyysvakuutusmaksutiedon. $osuus_{el}$ on palkansaajan eläkevakuutusmaksujen osuus palkasta ja $osuus_{tvm}$ on palkansaajan työttömyysvakuutusmaksujen osuus palkasta. Palkansaajan eläkevakuutusmaksujen osuus palkasta riippuu palkansaajan ikäluokasta i.

Kun vuosittaiset palkansaajien eläkemaksutiedot on erotettu työttömyysvakuutusmaksuista, henkilöiden vuosittaiset palkkasummat saadaan laskettua muodostetuista eläkemaksutiedoista:

$$palkkasumma = \textit{eläkemaksu} \times \frac{1}{osuus_{el}(i)}. \tag{5.2}$$

Tässä yhtälössä $osuus_{el}(i)$ on jälleen palkansaajan ikäluokasta i riippuva osuus palkkasummasta, joka menee eläkemaksuihin.

Muunsin vielä vuosittaiset lasketut palkkasummat vuoden 2011 palkkatasoon käyttämällä muunnokseen *palkansaajien ansiotasoindeksin (ATI)* ja *säännöllisten ansioiden indeksin (SANI)* välisiä vuosittaisia keskiarvoja (SVT: *Ansiotasoindeksi*). Koska kummankin indeksin osalta vuosittaiset arvot ovat hyvin lähellä toisiaan, myös niiden keskiarvo on lähellä kummankin indeksin lukuja.

Koska oletan (VATT (2014) tavoin) elinkaariennusteissa kaikkien henkilöiden jäävän eläkkeelle sen vuoden alussa, kun he täyttävät 63 vuotta, rajoitin palkkasummamallinnuksen koskemaan vain niitä henkilövuositasoisia rivejä aineistossa, joissa henkilöt olivat 18—62-vuotiaita. Lisäehtona sille, että henkilövuositasoinen rivi sisältyi aineistoon, oli, että havaintoa koskeva henkilö oli Suomessa havaintoa koskevana vuonna tutkimuksessa käytettävän maassaolokriteerin mukaisesti.

Laskettujen palkkasummaestimaattien tarkistus eläkemaksujen totaalin estimaatin avulla

Esitin yllä, kuinka henkilövuosittaiset palkkasummahavainnot lasketaan rekisteritiedoista, jotka koskevat henkilöiden vuosittain maksamia eläke- ja työttömyysvakuutusmaksuja. Koska palkkasummat lasketaan eläkemaksuja koskevista rekisteritiedoista, päätin tarkistaa eläkemaksuja koskevien rekisteritietojen kattavuuden – ja siten laskettujen palkkasummaestimaattien luotettavuuden – vertaamalla rekisteriaineistosta estimoitua eläkemaksujen totaalia todelliseen eläkemaksujen kokonaiskarttumaan. Toisin sanoen laskin eläkemaksuja koskevista rekisteritiedoista estimaatin koko Suomen väestön (Suomessa syntyneet ja ulkomailla syntyneet yhteensä) maksamien eläkevakuutusmaksujen kokonaissummalle vuonna 2011.

Tähän vuoden 2011 eläkemaksujen totaalin estimaatin laskemiseen käytin kaikkia 18–70-vuotiaita koskevia havaintoja tutkimusaineistossa. Palkansaajien eläkemaksut laskin edellä esitetyn lausekkeen 5.1 mukaisesti. Työnantajien keskimääräiset eläkemaksuosuudet vaihtelevat sen mukaan, minkä eläkelain piirissä maksut kerätään. Koska tutkimusaineisto ei sisällä tietoa siitä, minkä eläkelain piirissä eläkemaksut on maksettu, käytin työnantajien maksuosuutena painotettua keskiarvoa eri eläkelakien maksuosuuksien välillä. Painoina käytin eläkelakikohtaisia maksutuloja vuonna 2011. Työnantajien maksujen laskentapa esitetään tarkemmin muita ennusteparametreja ja eläkevaikutusten ennustamista käsittelevän luvun alaluvussa 6.1.2. Otin totaalin estimaatin laskemisessa lisäksi huomioon tutkimusaineistoon sisältyvien Suomessa syntyneiden 18–70-vuotiaiden osuuden koko Suomessa syntyneen väestön samanikäisten osuudesta.

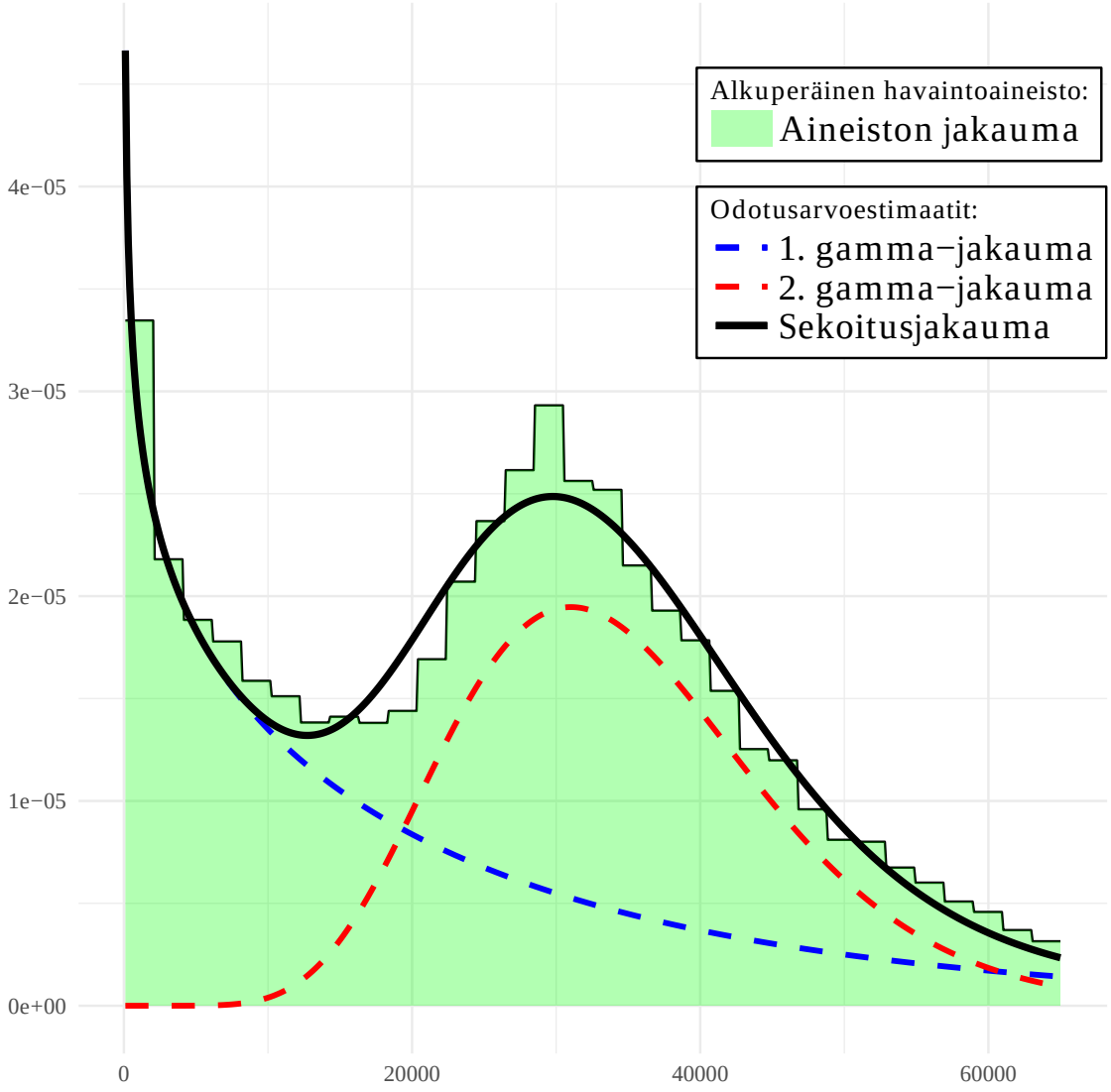
Esitellyllä tavalla aineistosta laskettujen palkansaajien ja työnantajien eläkemaksujen estimoitu kokonaissumma, eli eläkemaksujen totaalin estimaatti, oli vuonna 2011 noin 17,7 miljardia euroa. Todellinen luku vuonna 2011 oli 18,5 miljardia euroa (ETK 2012, 25). Koska estimoitu luku oli noin 96 %:ia todellisesta luvusta, eli samassa mittakaavassa, voin pitää palkansaajien pakollisia työttömyys- ja eläkevakuutusmaksuja koskevista rekisteritiedoista laskettuja henkilötasoisia palkkasummaestimaatteja likipitäen oikeina. Lisäksi, mikäli 18,5 miljardin euron kokonaissummasta poistetaan eläkelakien luokka "Muut", summa oli 18,25 miljardia, jolloin saamani estimaatti olisi noin 97 %:ia todellisesta luvusta. Luokka "Muut"sisältää esimerkiksi Suomen Pankin ja Ortodoksisen kirkon eläkemaksut, jotkä eivät välttämättä sisälly rekisteriaineistotietoihin.

Koska estimoidut palkkasummat ovat tämän tarkistuksen perusteella keskimäärin noin 96%:ia todellisista palkkasummista, kerron *kaikki* elinkaariennusteissa käytetyt palkkasummaennusteet 1/0,96:lla. Tässä palkkasummien mallinnusta käsittelevässä luvussa käytän kuitenkin suoraan edellä esitetyllä me-

netelmällä laskettuja palkkasummaestimaatteja. Pyydän lukijaa huomioimaan, että tämän luvun kaikki palkkasummaluvut ovat siten kesimäärin noin 96%:ia todellisista palkkasummista.

5.2.2 Palkkasumma-aineistojen kuvaus

Suomessa syntyneiden palkkasummien jakaumasta käyvät ilmi keskeiset elementit, jotka koskevat myös kaikkia maahanmuuttajaryhmäkohtaisia aineistoja. Osa havainnoista on keskittynyt nollaan, joka muodostaa havaintojen yhteisjakauman epäjatkuvan osan. Yli nollan olevien havaintojen (alla kuvassa 5.1 Suomessa syntyneitä 18–67-vuotiaita koskien) jakaumassa havainnot ovat paitsi keskittyneet jakauman vasempaan laitaan lähelle nollaa, niin jakaumassa on kuitenkin paikallinen huippu noin 30 000 euron kohdalla.



Kuva 5.1: ***Yli nollan olevien palkkasummien havaittu jakauma ja sen mallinnus***. *Vihreä histogrammi ("Aineiston jakauma"): Suomessa syntyneen kantaväestön 18–67-vuotiaiden henkilöiden vuosittaisten yli nollan euron olevien palkkasummien jakauma vuosina 2005–2014. Palkkasummat ovat vuoden 2011 ansiotasossa. Jakaumasta on katkaistu pois sen yläpää yksityisyydensuojan säilyttämiseksi. Lähde: Tilastokeskuksen rekisteri maksetuista lakisääteisistä eläkemaksuista, omat laskelmat. Jakaumien viivat ("Odotusarvoestimaatit"): Jäljempänä luvussa 5.3.1.2 esitetyn tilastollisen mallin mukaisten parametrien odotusarvoestimaattien pohjalta simuloidut jakaumat palkkasummien yli nollien kahdelle komponentille sekä niiden sekoitusjakaumalle. "1. gamma-jakauma" on vain osan vuotta töissä olevien palkkasummille mallinnettu jakauma, "2. gamma-jakauma" on koko vuoden kokoaikaisessa töissä olleiden palkkasummille mallinnettu jakauma ja "Sekoitusjakauma" on edellisen kahden jakauman sekoitusjakauma eli kaikkien yli nollan olevien palkkasummien mallinnettu jakauma. Kuvan vaaka-akselin yksikkö on euro, ja pystyakselilla on kahden gamma-jakauman sekoitusjakauman todennäköisyystiheys.*

Tulkitsen yli nollan olevien havaintojen koostuvan kahden jakauman sekoitusjakaumasta (kuvassa 5.1 tämän jakauman estimaatti on mustalla yhtenäisellä viivalla piirretty ”Sekoitusjakauma”): koko vuoden kokoaikaisessa töissä olleiden jakaumasta, jonka huippu on mainitun noin 30 000 euron kohdalla (kuvassa 5.1 tämän jakauman estimaatti on punaisella katkoviivalla piirretty ”2. gamma-jakauma”), sekä toisaalta niiden henkilöiden jakaumasta, jotka ovat olleet vain osan vuotta töissä (kuvassa 5.1 tämän jakauman estimaatti on sinisellä katkoviivalla piirretty ”1. gamma-jakauma”). Sinänsä tulkinnan oikeellisuudella ei ole merkitystä, sillä yli nollan olevien havaintojen käsittäminen kahden jakauman sekoitusjakaumana on lähinnä mallinnuksellinen apuväline.

Lisäksi yli nollan olevien havaintojen kuulumista jompaankumpaan näistä jakaumista ei tiedetä varmasti, sillä oletan kummankin jakauman määrittelyjoukon koostuvan positiivisista reaaliluvuista. Vaikka tätä havaintojen kuulumista jompaankumpaan sekoitusmallin jakaumista ei tiedetä täydellä varmuudella – eli todennäköisyydellä yksi – kuulumiselle voidaan määrittää havaintokohtaiset todennäköisyydet. Tämä todennäköisyys voi olla ehdollista, eli riippua selittävien muuttujien arvoista. Koska yli nollan olevien havaintojen kuulumista sekoitusjakauman kahteen komponenttiin ei tiedetä varmasti, näiden kahden jakauman sekoitusjakaumaa kutsutaan *latentiksi sekoitusjakaumaksi* (*engl. latent mixture distribution*).

Havaintojen kuuluminen joko niiden niiden havaintojen joukkoon, jotka saavat arvoksi tasan nolla, tai niiden havaintojen joukkoon, jotka saavat arvoksi yli nollan, tiedetään aineistossa puolestaan varmasti. Tästä syystä nollassa olevien havaintojen jakauma ja yli nollan olevien havaintojen jakauma muodostavat sekoitusjakauman, joka ei kuitenkaan ole *latentti*. Tehtävien eläkemaksuennusteiden osalta havainnon kuuluminen nollahavaintojen joukkoon on kuitenkin latenttia, koska tätä ei tiedetä varmasti etukäteen.

Se osuus kaikista havainnoista, joka saa arvoksi nollan, vaihtelee huomattavasti syntymävaltion mukaan. Lisäksi yli nollia koskevan latentin sekoitusjakauman osalta vaihtelua on syntymävaltioiden mukaan siinä, kuinka suuren huipun koko vuoden työssä olleiden jakauma muodostaa, sekä siinä, missä tämän jakauman huippu sijaitsee. Myös vain osan vuodesta töissä olleen jakaumissa on eroja. Maahanmuuttajaryhmäkohtaiset palkkasumma-aineistojen jakaumat on esitetty jäljempänä estimointitulosten ja posterioriennustetarkistusten yhteydessä (ks. luku 5.4).

5.3 Palkkasummien mallinnus

Määritin palkkasummien yhteisjakaumalle tilastollisen mallin, joka riippuu valituista malliparametreista ($\boldsymbol{\theta}$). Tätä havaintoaineistoa ($\boldsymbol{y}$) koskevaa mallia kutsun aineistomalliksi, $p(\boldsymbol{y}|\boldsymbol{\theta})$, samoin kuin nettovaikutusmallinnuksen yhteydessä. Koska mallinnus tapahtuu bayesilaisen päättelyn piirissä, aineistomallin estimoitaville parametreille on määritettävä omat tilastolliset jakaumansa eli priorijakaumat, $p(\boldsymbol{\theta})$. Mallin estimoitavien parametrien havaintoaineistolla ehdollistettu yhteisjakauma eli posteriorijakauma on Bayesin kaavan mukaan verrannollinen aineistojakauman ja priorijakauman tulolle: $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

Alaluvussa 5.3.1 määritän aineistomallin ja alaluvussa 5.3.2 määritän puolestaan estimoitavien parametrien priorijakaumat. Määritän posteriorijakauman vielä uudestaan alaluvussa 5.3.3 ja alaluvussa 5.3.4 käsittelen näin määritetyn posteriorijakauman selvittämistä simuloimalla.

5.3.1 Palkkasummien aineistomallit

Määritin jokaiselle maahanmuuttajaryhmälle neljä perättäistä aineistomallia sekä näiden mallien estimoitavien parametrien priorijakaumat. Tuloksissa (luku 5.4) on esitetty näin muodostettujen posteriorijakaumien estimointitulokset. Perättäisillä aineistomalleilla tarkoitan sitä, että pienempien mallien parametrit sisältyvät aina suurempaan malliin.

Kahden ensimmäisen aineistomallin tarkoitus on määrittää otosjakauman luokka, eli millaisia todennäköisyysjakaumia aineiston mallintamiseen käytetään. Kolmannessa aineistomallissa lisään malliin kaikille havainnoille yhteiset selittävät muuttujat. Tähän asti oletan havaintojen olevan riippumattomia, mikä ei todellisuudessa pidä paikkaansa johtuen havaintojen ryvästymisestä henkilöittäin.

En huomioinut ryvästymistä vielä tässä vaiheessa, sillä ryvästymisen huomioivien satunnaisvaikutuksia sisältävien hierarkkisten mallien estimointi olisi vienyt huomattavasti enemmän aikaa kuin niiden mallien estimointi, joissa ryvästymistä ei huomioida satunnaisvaikutuksilla. Huomioin havaintojen ryvästymisen henkilöittäin kuitenkin lopullisessa eli neljännessä aineistomallissa. Olen käsitellyt työikäisten nettovaikutusmallinnuksen yhteydessä enemmän hierarkkisten mallien käyttöä paneeliaineistojen mallintamisessa (ks. luku4.3.1).

Ennustetarkkuuden mittarit: WAIC ja LOO. Parametriestimaattien lisäksi ilmoitan tuloksissa kaikkia neljää mallia koskien niiden ennustetarkkuutta mittaavat LOO:n ja WAIC:n (ks. ennustetarkkuuden mittareista luku 4.3.2.3). Nämä ennustetarkkuuden mittarit kertovat tuloksissa, että parametrimäärältään suuremmat mallit ennustavat tarkemmin *käytettävissä olevaa* havaintoaineistoa kuin pienemmät mallit jokaisen maahanmuuttajaryhmän sekä Suomessa syntyneiden mallien kohdalla. Mallinvalinta neljän perättäisen aineistomallin välillä ei kuitenkaan perustu LOO- ja WAIC-lukuihin vaan mallinnustavoitteisiin: havaintoaineiston jakauman hyvin toistavan otosjakauman valintaan, aikamuuttujien sisällyttämiseen mukaan ja havaintojen ryvästymisen huomiointiin.

LOO:n ja WAIC:n käytössä mallinvalinnassa on omat puutteensa, jos niitä käytetään erityisesti valitsemaan ennustemallia *uudelle aineistolle*, mikä on tämän tutkimuksen pääasiallinen tarkoitus. Tämä johtuu siitä, että WAIC ja LOO perustuvat mallinvalintakriteereinä käytettävissä olevaan aineistoon, ja siten ne tuottavat mallinvalinnassa malleja, jotka ovat ylisopivia suhteessa uuteen aineistoon, koska ne perustuvat käsillä olevan aineiston vertailuihin. (Gelman et al. 2014, 181; Vehtari et al. 2015, 18).

5.3.1.1 Tyhjä hurdle-gamma (Malli 1)

Aloitin mallinnuksen jokaisen maahanmuuttajaryhmän sekä Suomessa syntyneiden kohdalla tyhjällä *hurdle-gamma*-mallilla. “Tyhjä” viittaa siihen, että malleissa ei ollut vakiotermien lisäksi mukana lainkaan selittäviä muuttujia, sillä mallinnuksen tarkoitus oli vain selvittää otosjakauman luokka eli selitettävän muuttujan jakaumaa vastaava jakauma. *Hurdle* (*suom. este*) viittaa palkkasummien jakauman siihen epäjatkuvaan osaan, joka sijaitsee nollassa. Käytän englanninkielistä nimitystä “hurdle”, sillä mallinnuskäsitteellä ei ole tietääkseni vakiintunutta suomalaista nimitystä, ja suomenkielisen “esteen” käyttäminen

terminä saattaisi hämmentää lukijaa.

Esimerkiksi Tutz (2012) esittää hurdle-mallien perusteet. Hurdle-mallien perusideana on, että aineisto on generoitunut kahden satunnaisprosessin tuloksena: ensimmäisen prosessin tuloksena selviää, saako havainto arvoksi tasan nollan vai nollaa suuremman arvon. Toisen prosessin tuloksena havainto saa nollaa suuremman arvon. Nimitys hurdle (este) tulee siitä, että havainto saa jonkin nollaa suuremman arvon, kunhan aluksi "päästään" nollassa olevan esteen "yli". (Tutz 2012, 200–204).

Yli nollien olevien palkkasummien mallintamiseen käytin aluksi yhtä gamma-jakaumaa. Gamma-jakaumat olivat potentiaalisesti sopivia yli nollan olevien palkkasummien mallintamiseen hurdlen kanssa, sillä gamma-jakaumien määrittelyjoukko käsittää vain positiiviset reaaliluvut. Erityisesti gamma-jakaumat, joiden muoto-parametri on arvoltaan alle 1 (tai lähellä ykköstä), vaikuttivat olevan sopivia jakaumia yli nollien olevien havaintojen mallintamiseen. Näillä gamma-jakaumilla todennäköisyystiheys on suurimmillaan lähestyttäessä nollaa oikealta, ja niiden todennäköisyystiheydet palkkasummien suhteen ovat aidosti laskevia funktioita lukuun ottamatta epäjatkuvuuskohtaa nollassa.

Tämä vastaa havaintoa siitä, että suhteellisesti eniten palkkasummia on lähellä nollaa ja mitä kauemmaksi nollasta mennään oikealle, niin sitä vähemmän palkkasummahavaintoja vaikuttaa olevan. Koska tämä ei kuitenkaan pidä paikkaansa siltä osin, että palkkasummien jakaumalla on paikallinen huippu noin 30 000 euron kohdalla, seuraavassa mallinnusaskeleessa sovitan yli nollien havaintoihin kahden gamma-jakauman sekoitusjakaumaa, missä tämä paikallinen huippu otetaan huomioon. Aloitan kuitenkin mallinnuksen yhdestä gamma-jakaumasta. Erityisesti Irakissa ja Somaliassa syntyneiden kohdalla paikallinen huippu palkkasummissa on muutenkin niin pieni, että hypoteesi yhden gamma-jakauman riittävyydestä yli nollien olevien havaintojen mallintamiseen ei vaikuttanut mahdottomalta näiden syntymävaltioiden kohdalla. Tämä tarkoittaa sitä, että havaintoaineistossa Irakissa ja Somaliassa syntyneiden yli nollien havainnoista suurin osa on luultavasti sellaisia vuosittaisia henkilöitä koskevia palkkasummia, joissa henkilö ei ole ollut koko vuotta työllisenä.

Hurdle-gamman käyttö aineistomallina oli lisäksi helppoa siitä syystä, että tämä todennäköisyysjakaumien luokka löytyy Stan-ohjelmaa käyttävästä R-ohjelman BRMS-paketista (Buerkner 2015). Jäljempänä esitettävien laajempien mallien estimointiin käytettävät Stan-koodit perustuvat BRMS-ohjelman Stan-koodiin, jolla simuloin hurdle-gamma-aineistomalliin perustuvia posteriorijakaumia tässä mallinnusaskeleessa 1.

Hurdle-gamma-todennäköisyysjakauma

Edellä olevan perusteella havaituilla palkkasummilla (y) ja hurdlella (H) on yhteisjakauma (mukaillen Tutzia 2012, 201):

$$p(y,H) = \begin{cases} p(y,H=1) = p(y|H=1)p(H=1) = 1 \times p(H=1) = p(H=1), & \text{kun } H=1 \\ p(y,H=0) = p(y|H=0)p(H=0), & \text{kun } H=0 \end{cases} \tag{5.3}$$

missä $y = 0$, kun $H = 1$ ja $y > 0$, kun $H = 0$ sekä $p(H = 1) = 1 - p(H = 0)$.

Huomautan, että käytän tässä yhteydessä puutteellista ilmaisutapaa palkkasummien havaintojen ja hurdlen todennäköisyysjakaumista (ks. selvennys merkintätapoja koskien: johdannon luku 2.2.4). Korrektimpi ilmaisutapa tarkoittaisi esimerkiksi sitä, että havaintoihin satunnaismuuttujana viitattaisiin isolla kirjaimella Y ja tämän satunnaismuuttujan yksittäisiin realisoituneisiin arvoihin pienellä kirjaimella y. Havaintojen todennäköisyysfunktio voitaisiin tällöin lausua korrektimmin: $f_Y(y)$ tai $f_Y(Y = y)$.

Summaamalla hurdlen (H) kahden mahdollisen arvon ($H \in \{0, 1\}$) yli saadaan palkkasummien marginaalijakauma yllä olevasta y:n ja H:n yhteisjakaumasta:

$$p(y) = P(H = 1) + p(y|H = 0)p(H = 0). \tag{5.4}$$

Hurdle (H) on *Bernoulli*-jakautunut yhden havainnon j osalta:

$$p(H_j) \sim Ber(p_{hurdle}), \tag{5.5}$$

missä, $0 < p_{hurdle} < 1$, on todennäköisyys sille, että palkkasummahavainto $y_j = 0$.

Mikäli havainto j on suurempi kuin nolla, jolloin $H_j = 0$, eli päästään hurdlen "yli", palkkasummahavainnon j ehdollinen todennäköisyysjakauma noudattaa gamma-jakaumaa :

$$p(y_j|H = 0) \sim gamma(k, k/\mu_{gamma_j}), \tag{5.6}$$

missä $k > 0$ on gamma-jakauman muotoparametri ja $k/\mu_{gamma_j} > 0$ on havainnon j gamma-jakauman uudelleenparametroitu rate-parametri (skaalaparametrin käänteisluku). Uudelleenparametroidussa rate-parametrissa:

$$\mu_{gamma_j} = exp(\boldsymbol{x}'_j \boldsymbol{\beta}_{gamma}), \tag{5.7}$$

missä eksponenttifunktion argumenttina on lineaarinen (affiini) sovite, jonka regressiokertoimien pystyvektori on $\boldsymbol{\beta}_{gamma}$ ja $\boldsymbol{x}'_j$ on havainnon j selittävien muuttujien vaakavektori. Gamma-jakauman käänteinen skaalaparametri on uudelleenparametroitu yllä niin, että gamma-jakauman odotusarvo $k \times (\mu_{gamma_j}/k) = \mu_{gamma_j}$ riippuu vain sovitteesta μ_{gamma_j}.

Tässä mallinnusaskeleessa 1 esitetyssä tyhjässä mallissa $\boldsymbol{\beta}_{gamma}$ sisältää vain yhden komponentin eli vakiotermin β_{gamma_0}. Samalla x_j sisältää vain yhden komponentin, joka on luku yksi. Yli nollia koskeva malli on toisin sanoen *yleistetty lineaarinen malli* (*engl. generalized linear model*), missä niin kutsuttuna *linkkifunktiona* (*engl. link function*) toimii eksponenttifunktion μ_{gamma_j}:n käänteisfunktio eli logaritmifunktio. Logaritmifunktio on sopiva valinta linkkifunktioksi, sillä näin μ_{gamma_j} saa käytännössä palkkasummamalleissa arvoksi vain positiivisia reaalilukuja, mitä gamma-jakauman rate-parametri vaatii mahdollisiksi arvoikseen.

5.3.1.2 Tyhjä hurdle-gamma-sekoitusjakauma (Malli 2)

Koska havaituissa palkkasumma-aineistoissa on paikallinen huippu yli nollien osalta noin 30 000 euron kohdalla, edellä esitetty gamma-jakauma ei ole kovin hyvä aineistomallina, koska sen avulla paikallista huippua ei saada mallinnettua. Tämä puolestaan rikkoo ensimmäistä mallinnustavoitetta, jonka mukaan mallien tuottamien posterioriennustejakaumien on toistettava alkuperäiset aineistot.

Jotta olisin saanut mallinnettua yli nollan olevat palkkasummat paremmin, käytin yli nollien mallintamiseen kahden gamma-jakauman latenttia sekoitusjakaumaa. Havaintojen osalta niiden kuulumista jompaankumpaan näistä jakaumista ei tiedetä varmasti, mutta sille voidaan määrittää oma todennäköisyytensä. Jakaumakuulumista kuvaa satunnaismuuttuja θ.

Ensimmäinen gamma-jakauma on muodoltaan samanlainen kuin gamma-jakauma yllä hurdle-gamma-mallin yhteydessä: sen muotoparametri saa arvoksi alle ykkösen (tai arvon ykkösen lähellä). Toisen – lisättävän gamma-jakauman – muotoparametri saa puolestaan arvoksi yli ykkösen. Palaan näiden gamma-jakaumien muotoparametrien todennäköisyysjakaumien määrittämiseen jäljempänä parametrien priorijakaumien määrittämisen yhteydessä. Toinen gamma-jakauma eroaa ensimmäisestä siinä, että sillä on yksi huippu, sekä siinä, että sen todennäköisyysmassa on keskittynyt huipun ympärille eikä nollan läheisyyteen. Näin ollen se vaikuttaa hyvältä kandidaatilta mallinnettaessa palkkasummien jakauman paikallista huippua. Gamma-jakaumissa on myös se hyvä ominaisuus, että jakaumat ovat oikealle vinoja. Tämä johtuu siitä, että tyypillisten tulojakaumien tavoin myös palkkasummien jakaumat ovat oikealle vinoja.

Mallinnusaskeleeseen 1 (hurdle-gamma-malli) tehtävien esitettyjen muutosten jälkeen havainnon j todennäköisyysjakauma yli nollan euron osalta noudattaa kahden gamma-jakauman ($\mathcal{G}$) latenttia sekoitusjakaumaa:

$$\begin{aligned} p(y_j|H_j = 0) \sim\ & \mathcal{G}(k_{gamma1}, k_{gamma1}/\mu_{gamma1_j}) \times \theta \\ & \times \mathcal{G}(k_{gamma2}, k_{gamma2}/\mu_{gamma2_j}) \times (1-\theta), \end{aligned} \tag{5.8}$$

missä gamma-jakaumien ($\mathcal{G}$) parametrit ovat kuten yksittäisessä gamma-jakaumassa edellä (lausekkeessa 5.6) ja $\theta \in \{0, 1\}$ on satunnaismuuttuja, joka saa arvoksi ykkösen, kun havainto kuuluu ensimmäiseen gamma-jakaumaan, ja nollan, kun havainto kuuluu toiseen gamma-jakaumaan.

Parametrilla θ on Bernoulli-jakauma:

$$p(\theta) \sim Ber(p_\theta), \tag{5.9}$$

missä parametri $0 < p_\theta < 1$ on havainnon todennäköisyys kuulua ensimmäiseen gamma-jakaumaan ja $1 - p_\theta$ on havainnon todennäköisyys kuulua toiseen gamma-jakaumaan (kun tiedetään, että havainto on yli nollan).

5.3.1.3 Selittävät muuttujat hurdle-gamma-sekoitusjakaumassa (Malli 3)

Kolmannessa mallinnusvaiheessa valitsin ja lisäsin maahanmuuttajaryhmäkohtaisiin malleihin kaikille maahanmuuttajaryhmäkohtaisille havainnoille yhteiset selittävät muuttujat. Näiden selittävien muuttujien regressiokertoimien arvot eivät toisin sanoen vaihtele malleissa havaintoryppäittäin eli esimerkiksi henkilöittäin. Viimeisessä mallinnusvaiheessa sallin malliin sisältyvien vakioiden vaihdella "satunnaisesti" henkilöiden välillä. Esitän aluksi, miten eläkemalliin sisältyvät selittävät muuttujat ja niiden arvot tulkitaan. Sen jälkeen esitän selittävien muuttujien valintaprosessin maahanmuuttajaryhmäkohtaisiin malleihin. Selittävien muuttujien valinta palkkasummamalleihin poikkeaa siitä, miten valitsin selittävät muuttujat olotila- ja nettovaikutusmalleihin.

Hurdle-gamma-sekoitusjakauman neljä regressiomallia

Selittävien muuttujien muodostamia lineaarikombinaatioita eli mallisovitteita on maahanmuuttajaryhmäkohtaisissa malleissa neljä kappaletta. Edellä esitetyn perusteella hurdle-gamma-sekoitusjakaumaa noudattava palkkasummamalli voidaan kirjoittaa havainnolle j muodossa:

$$\begin{aligned} p(H_j) &\sim Ber(\mu_{hurdle_j}), \\ p(y_j|H_j = 0) &\sim \mathcal{G}(k_{gamma1}, k_{gamma1}/\mu_{gamma1_j}) \times \theta_j \\ &\quad \times \mathcal{G}(k_{gamma2}, k_{gamma2}/\mu_{gamma2_j}) \times (1 - \theta_j). \end{aligned} \tag{5.10}$$

Hurdlen arvon selittäminen muodostaa ensimmäisen selittävistä muuttujista muodostetun mallisovitteen, μ_{hurdle_j}. Ensimmäiseen gamma-jakaumaan kuulumisen indikaattorin ($\theta_j = 1$) selittäminen muodostaa toisen mallisovitteen. Kummankin gamma-jakauman regressiomallit sisältävät kolmannen ja neljännen mallisovitteen: μ_{gamma1_j} ja μ_{gamma2_j}. Käyn nämä sovitteet sisältävät regressiomallit vielä yksitellen läpi.

1. Hurdlen logistinen regressiomalli:

$$p(H_j) \sim Ber(\mu_{hurdle_j}) \tag{5.11}$$

ja

$$\mu_{hurdle_j} = logit^{-1}(\boldsymbol{x}'_{hurdle_j}\boldsymbol{\beta}_{hurdle}). \tag{5.12}$$

Logit-funktion käänteisfunktion eli *logistisen funktion* argumenttina on lineaarinen (affiini) sovite, jonka regressiokertoimien pystyvektori on $\boldsymbol{\beta}_{hurdle}$ ja $\boldsymbol{x}'_{hurdle_j}$ on hurdlea selittävien muuttujien vaakavektori havainnon j kohdalla. Regressiokertoimien vektorin ensimmäinen alkio on vakiotermi β_{hurdle_0}, joka tuli esille jo edellä tyhjissä malleissa. Hurdlea – eli nollahavaintoja – koskeva malli on toisin sanoen yleistetty lineaarinen malli (engl. generalized linear model), missä nk. linkkifunktiona (engl. link function) on μ_{hurdle_j}:n käänteisfunktio eli logit-funktio. Hurdle-malli on toisin sanoen *logistinen regressiomalli*. Regressiokertoimien estimaatit kertovat, kuinka yhden yksikön muutos yhdessä selittävässä muuttujassa, kun muiden muuttujien arvot pidetään vakioina, vaikuttaa hurdlen *log-oddsiin* eli *log-vedonlyöntisuhteeseen*, $log(p/(1-p))$, missä p on todennäköisyys sille, että hurdle saa arvoksi ykkösen.

Tämä tarkoittaa sitä, että mikäli regressiokertoimen estimaatti on positiivinen, tällöin todennäköisyys sille, että palkkasumma on tasan nolla, kasvaa, kun regressiokerrointa koskevan selittävän muuttujan arvo kasvaa.

2. Ensimmäiseen gamma-jakaumaan kuulumisen indikaattorin, θ:n, logistinen regressiomalli:

$$p(\theta_j) \sim Ber(\mu_{\theta_j}) \tag{5.13}$$

ja

$$\mu_{\theta_j} = logit^{-1}(\boldsymbol{x}'_{\theta_j}\boldsymbol{\beta}_\theta). \tag{5.14}$$

Logit-funktion käänteisfunktion eli logistisen funktion argumenttina on lineaarinen (affiini) sovite, jonka regressiokertoimien pystyvektori on $\boldsymbol{\beta}_\theta$ ja $\boldsymbol{x}'_{\theta_j}$ on gamma-jakaumaan kuulumista selittävien muuttujien vaakavektori havainnon j kohdalla. Regressiokertoimien vektorin ensimmäinen alkio on vakiotermi β_{θ_0}, joka vastaa edellä tyhjissä malleissa esitettyä p_θ-parametria. θ:aa – eli kuulumista ensimmäiseen gamma-jakaumaan – koskeva malli on toisin sanoen yleistetty lineaarinen malli (engl. generalized linear model), missä linkkifunktiona (engl. link function) on μ_{theta_j}:n käänteisfunktio eli logit-funktio. θ-malli on toisin sanoen logistinen regressiomalli.

Regressiokertoimien estimaatit kertovat, kuinka yhden yksikön muutos yhdessä selittävässä muuttujassa, kun muiden muuttujien arvot pidetään vakioina, vaikuttaa θ:n arvon *log-oddsiin* eli *log-vedonlyöntisuhteeseen*, $log(p/(1-p))$, missä p on todennäköisyys sille, että θ saa arvoksi ykkösen. Tämä tarkoittaa sitä, että mikäli regressiokertoimen estimaatti on positiivinen, tällöin todennäköisyys sille, että palkkasumma kuuluu ensimmäiseen gamma-jakaumaan, kasvaa, kun regressiokerrointa koskevan selittävän muuttujan arvo kasvaa.

3. ja 4. Gamma-jakaumien regressiomallit:

$$\mu_{gamma1_j} = exp(\boldsymbol{x}'_{gamma1_j}\boldsymbol{\beta}_{gamma1}). \tag{5.15}$$

Eksponenttifunktion argumenttina on lineaarinen (affiini) sovite, jonka regressiokertoimien pystyvektori on $\boldsymbol{\beta}_{gamma1}$ ja $\boldsymbol{x}'_{gamma1_j}$ on selittävien muuttujien vaakavektori. Kuten jo esitin 2. mallinnusaskeleen yhteydessä, kyseessä on yleistetty lineaarinen malli, jonka linkkifunktiona on logaritmifunktio eli eksponenttifunktion käänteisfunktio. Toisen gamma-jakauman regressiomalli on muodoltaan samanlainen. Regressiokertoimien estimaatit kertovat, kuinka yhden yksikön muutos yhdessä selittävässä muuttujassa vaikuttaa selittävää muuttujaa koskevan gamma-jakauman odotusarvon logaritmiin, kun muiden muuttujien arvot pidetään vakioina. Tämä tarkoittaa sitä, että mikäli regressiokertoimen estimaatti on positiivinen, tällöin palkkasummien hurdlella ehdollinen odotusarvo kasvaa, kun regressiokerrointa koskevan selittävän muuttujan arvo kasvaa.

Esitän regressiokertoimia koskevat estimointitulokset niin, että tässä luvussa esitetyistä (alkuperäisistä) regressiokertoimista on otettu eksponentit. Palaan näin saatujen regressiokertoimien tulkintaan ja estimointituloksiin luvussa 5.4.

Selittävien muuttujien valinnat malleihin

Valitsin mallien selittävät muuttujat likimain samasta joukosta rekisteritietoja kuin nettovaikutusmalleissa. Tiedot koskevat henkilöiden tietoja heidän muuttaessaan Suomeen ensimmäistä kertaa. Havaintoaineistoon sisältyvät vain henkilöt, jotka olivat muuttaneet Suomeen ensimmäisen kerran 16–62-vuotiaina. Suomessa syntyneillä tiedot koskevat sitä vuotta, josta saakka heistä on havaintoja aineistossa. Poikkeuksen tästä muodostavat aikamuuttujat ikään, maassaoloaikaan (vuodet ensimmäisestä muutosta) ja kalenterivuoteen liittyen, jotka voidaan määrittää deterministisesti kaikille vuosille ensimmäisestä maahanmuuttovuodesta lähtien.

Irakissa ja Somaliassa syntyneiden mallit maassaolovuoden tiedoilla:

1. ”Perusmuuttujat”: Hurdlen ja gamma1-jakauman regressiomalleihin.
2. Mallintarkistusten perusteella valittujen ikäluokkiin ja maassaoloaikaan liittyvien muuttujien lisäys hurdlen ja gamma1-jakauman regressiomalleihin..
3. Mahdollisuuksien mukaan sama gamma2-jakauman ja θ:n regressiomalleja koskien. Ensisijaisesti mukaan aikaan liittyvät muuttujat.
4. "Mahdollisuuksilla"viitataan tässä estimointiongelmiin, joita ilmaantui gamma2-jakauman ja θ:n regressiomalleihin lisättäessä regressiokertoimia. Nämä johtuivat luultavasti siitä, että gamma2-jakauma on lähes olematon, koska vain pieni osuus Irakia ja Somaliaa koskevista havainnoista koskee koko vuoden kokoaikaisessa töissä olleita.

Suomessa syntyvien henkilöiden mallinnus:

1. ”Perusmuuttujat” kaikkiin neljään malliin: sukupuoli, ikä, iän neliö ja kalenterivuosi.
2. Mallintarkistusten perusteella lisätään vielä joitakin ikäluokkaindikaattorimuuttujia.

5.3.1.4 Satunnaisvaikutukset hurdle-gamma-sekoitusjakaumassa (Malli 4)

Viimeisessä mallinnusvaiheessa otin huomioon palkkasumma-aineiston hierarkkisen rakenteen. Kuten yksittäiset havainnot nettovaikutusaineiston kohdalla, myös yksittäiset palkkasummahavainnot ovat ryvästyneet henkilöittäin. Havainnot eivät näin ollen ole riippumattomia, vaan yksittäisen henkilön havainnot muistuttavat toisiaan enemmän kuin sellaisessa tapauksessa, jossa kaikki havainnot olisivat riippumattomia. Koska (henkilötasoiset) selittävät muuttujat eivät sisällä kaikkia niitä vaikutuksia, jotka vaikuttavat palkkasummien määrään, yksittäisten henkilöiden havainnot muistuttavat edelleen toisiaan, vaikka palkkasummamalleihin lisätään kaikille havainnoille yhteiset selittävät muuttujat edellä mallissa 3.

Havainnot eivät ole siis ehdollisestikaan riippumattomia, kun ehdollistajina ovat palkkasummamallin havaitut selittävät muuttujat. Tämän havaitsemattoman heterogeenisuuden huomioon ottamiseksi annan palkkasummamallien vakiotermien vaihdella “satunnaisesti” henkilöiden välillä.

Vakiotermejä on palkkasummamallissa neljä kappaletta eli yksi jokaista yksittäistä regressiomallia kohti, jotka on esitetty edellä mallissa 3:

1. Hurdlen vakio: $\beta_{hurdle_0} = \beta_{hurdle_0_yht} + U_{hurdle_0}$.
2. Ensimmäisen gamma-jakauman vakio: $\beta_{gamma1_0} = \beta_{gamma1_0_yht} + U_{gamma1_0}$.
3. Toisen gamma-jakauman vakio: $\beta_{gamma2_0} = \beta_{gamma2_0_yht} + U_{gamma2_0}$.
4. θ:n vakio: $\beta_{\theta_0} = \beta_{\theta_0_yht} + U_{\theta_0}$.

Kun vakiotermit sisältävät satunnaisvaikutuksia, ne muodostuvat kahdesta yhteenlaskettavasta. Ensimmäinen näistä yhteenlaskettavista on kaikille havainnoille (ja henkilöille) *yhteinen* vakiotermin osa, jonka alaindeksissä on ”yht” ilmaisemassa sitä, että se on kaikille havainnoille yhteinen. Nollat alaindekseissä puolestaan viittaavat siihen, että kyse on vakiotermeistä. Esimerkiksi hurdlen vakiotermin, β_{hurdle_0}, kohdalla tämä kaikille havainnoille yhteinen osa on $\beta_{hurdle_0_yht}$. Toinen yhteenlaskettava jokaisen vakiotermin kohdalla on satunnaisvaikutus, jotka on ilmaistu isolla U-kirjaimella. Esimerkiksi ensimmäisen gamma-jakauman (vain osan vuodesta töissä) vakiotermiä, β_{gamma1_0}, koskeva satunnaisvaikutus on U_{gamma1_0}.

Oletan jokaisen vakiotermiin liittyvän neljän satunnaisvaikutuksen noudattavan gamma-jakaumia seuraavasti (ks. Sutradhar 2003 valitun todennäköisyysjakaumien luokan, $\mathcal{G}(1/\lambda, 1/\lambda)$, suotuisuudesta satunnaisvaikutusten jakaumien mallinnuksessa):

1. $U_{hurdle_0} \sim \mathcal{G}(1/\lambda_{hurdle_0}, 1/\lambda_{hurdle_0})$.
2. $U_{gamma1_0} \sim \mathcal{G}(1/\lambda_{gamma1_0}, 1/\lambda_{gamma1_0})$.
3. $U_{gamma2_0} \sim \mathcal{G}(1/\lambda_{gamma2_0}, 1/\lambda_{gamma2_0})$.
4. $U_{\theta_0} \sim \mathcal{G}(1/\lambda_{\theta_0}, 1/\lambda_{\theta_0})$.

Gamma-jakaumat muodostuvat muoto- (*engl. shape*) ja rate-parametrista. Esitetyissä satunnaisvaikutusten gamma-jakaumissa on kuitenkin vain yksi estimoitava parametri jokaisessa jakaumassa. Esimerkiksi θ-parametriin liittyvän satunnaisvaikutuksen, U_{θ_0}, gamma-jakauman ainoa estimoitava parametri on λ_{θ_0}. Tällä satunnaisvaikutusten gamma-jakaumien määritelmällä niiden odotusarvo ja varianssi ovat yksinkertaisia: jokaisen jakauman odotusarvo on 1, ja jokaisen jakauman varianssi on jakauman estimoitava parametri. Esimerkiksi θ:n satunnaisvaikutuksen varianssi on λ_{θ_0}.

Kokeilin Suomessa syntyneiden osalta satunnaisvaikutusten jakaumaoletuksena myös normaalijakaumia (ja niiden muodostamaa multinormaalijakaumaa), mutta gamma-jakaumat vaikuttivat sopivan paremmin aineistoon. Satunnaisvaikutuksille olisi voinut kokeilla myös muita jakaumaoletuksia samoin, kuten esimerkiksi nettovaikutusmallien yhteydessä testattiin Studentin t-jakauman sopivuutta henkilötason ”satunnaisen” vakiotermin jakaumaksi. Tyydyin kuitenkin käyttämään usein käytettyä gamma-jakaumaa, sillä posterioriennustetarkistusten perusteella en nähnyt tälle kokeilulle tarvetta. Tosin palkkasummien

mallia voi vielä yrittää parantaa esimerkiksi hurdlen (nollien osuuden) osalta kokeilemalla hurdlen vakion satunnaisvaikutuksen jakaumana muitakin jakaumia kuin nyt käytetyn gamma-jakauman.

Satunnaisvaikutusten riippumattomuus. Nyt käytetty jakaumaoletus vakiotermien satunnaisvaikutuksille olettaa, että esitetyt neljä satunnaisvaikutusta ovat toisistaan riippumattomia. Tämä ei kuitenkaan pidä paikkaansa multinormaalijakaumalla tekemäni mallinnuskokeilujen perusteella. Käyttämällä satunnaisvaikutuksille yhteistä moniulotteista todennäköisyysjakaumaa malliin saataisiin sisällytettyä tilastolliset yhteydet niiden havaitsemattomien tekijöiden välillä, jotka yhtäältä vaikuttavat siihen, kuuluuko havainto nollaan eli hurdleen, ja toisaalta siihen, mikä on palkkasummien odotusarvo yli nollien joukossa. Satunnaisvaikutuksia sisältävää mallia voisikin jatkossa laajentaa niin, että yksiulotteiset gamma-jakaumat korvattaisiin moniulotteisella gamma-jakaumalla (jonka esittely löytyy esimerkiksi tekstistä Furman 2007).

Mainitsen esimerkkinä Yaun et el. (2003) artikkelista, jossa sovelletaan normaalijakautuneita satunnaisvakiotermejä samankaltaisessa mallinnusasetelmassa, kuin tässä on kyse. Artikkelissa on kahden gammajakauman sijaan kahden normaalijakauman sekoitusjakauma ja hurdlea ei ole, mutta muuten mallinnusasetelma muistuttaa tässä esitettyä.

Mahdolliset muut satunnaisvaikutukset. Palkkasummamallien osalta – kuten nettovaikutusmallien yhteydessä – oletan, että henkilöiden palkkasummat ovat ehdollisesti riippumattomia toisistaan, kun selittävät muuttujat, henkilöiden perättäisten havaintojen riippuvaisuus toisistaan ja satunnaisvaikutukset huomioidaan.

Lopullisissa palkkasummamalleissa satunnaisia ovat henkilöiden kohdalla vain tasot eli vakiotermit, mutta eivät kulmakertoimet eli muut regressiokertoimet. En mallintanut kulmakertoimia satunnaisina, enkä sisällyttänyt malliin satunnaista vakiotermiä maahanmuuttovuositasolla, sillä nämä olisivat vieneet liikaa aikaa, koska posteriorien simulointiaika Stan-ohjelmalla olisi kasvanut moninkertaiseksi.

Tämän seurauksena oletan malleissa, että kaikkien regressiokertoimien vaikutukset ovat samanlaiset kaikkien henkilöiden osalta sekä että maahanmuuttovuosien ja maassaoloaikojen arvot eivät ole yhteydessä sellaisiin havaitsemattomiin tekijöihin, jotka vaikuttavat palkkasummien määrään. Koska viimeksi mainittu asia on varsin epärealistinen, eläkemallien jatkokehittämisen ensimmäisen askel olisikin sisällyttää malleihin maahanmuuttovuositason ja maassaoloajan tason satunnaiset vakiotermit. Toinen parannus olisi sisällyttää malleihin myös satunnaisvaikutuksia regressiokertoimien kohdalla, kuten nettovaikutusmallien osalta tein henkilötasolla.

5.3.2 Estimoitavien parametrien priorijakaumat

Edellä luvussa 5.3.1 määritin palkkasummien aineistomallit, joissa palkkasummien ehdolliset todennäköisyystiheysfunktiot oli ehdollistettu estimoitavilla parametreilla. Näille estimoitaville parametreille on määritettävä vielä priorijakaumat, jotta parametreille saadaan määritettyä niiden normeeraamaton yhteisposteriorijakauma.

Määritin parametreille priorijakaumat, jotka ovat toisistaan riippumattomia, koska en nähnyt omaavani sellaista tutkimusaineiston ulkopuolista informaatiota, jonka perusteella olisin voinut määrittää parametreille toisistaan riippuvia priorijakaumia. Näin priorijakaumien määrittäminen oli myös helpompaa. Kaikki priorijakaumat olivat sopivia (*engl. proper*) todennäköisyysjakaumia, joten kaikkien parametrien yhteinen posteriorijakauma oli myös sopiva todennäköisyysjakauma.

Määritin gamma-jakaumien muotoparametreja lukuun ottamatta kaikille muille parametreille heikosti informatiiviset priorijakaumat. “Heikosti informatiivinen” tarkoittaa tässä yhteydessä sitä, että priorijakaumien todennäköisyystiheys on keskittynyt parametrien osalta “järkevälle” alueelle. Regressiokertoimien osalta tämä järkevyys tarkoittaa sitä, ettei yksikään yhden yksikön muutos selittävän muuttujan arvossa tuota esimerkiksi miljoonien eurojen muutosta henkilön vuosittaisen palkkasumman odotusarvossa. Priorijakaumat ovat kuitenkin sen verran leveitä, että aineistomallin tulokset käytännössä määrittävät näiden parametrien posteriorijakaumat, koska havaintomäärät ovat syntymävaltiokohtaisissa aineistoissa ”suuria” eli joitakin kymmeniä tuhansia.

Regressiokertoimien (ml. vakiotermit) osalta käytin priorijakaumia, jotka olivat normaalijakaumia, joiden odotusarvot olivat nollia ja keskihajonnat arvoltaan kymmenen. Poikkeuksen tästä tein θ-parametrin arvon sovitteen vakion kohdalla, jossa priorijakauman keskihajonta oli kaksi. Käytin selittävien muuttujien osalta malleissa niiden standardoituja arvoja, minkä johdosta selittävien muuttujien vaikutukset regressiokertoimien kautta olivat likimain samassa mittakaavassa. Priorijakaumien keskihajonnat, jotka olivat arvoltaan kymmenen, olivat ”heikosti informatiivisia”, koska regressiokertoimien posteriorijakaumien estimointitulosten (luku 5.4) perusteella priorijakaumien keskihajonnat olivat paljon suurempia kuin regressiokertoimien posteriorijakaumien keskihajonnat. Hajontatermien osalta käytin $Cauchy(0, 10)$-jakaumia, jotka ovat melko epäinformatiivisia varianssiparametrien arvojen suhteen.

Gamma-jakaumien muotoparametrien melko informatiiviset priorijakaumat. Muista parametreista poiketen määritin gamma-jakaumien muotoparametreille melko informatiiviset priorit. Tämä johtuu siitä, että sekajakaumilla on niiden posteriorijakaumien simulaatioihin liittyen kaksi ominaisuutta, jonka vuoksi posteriorijakaumien simulaatiot eivät konvergoidu oikein, mikäli sekoitteen muodostavien yksittäisten jakaumien (sekoitteen komponentit) sijainnin määräävien parametrien priorijakaumat ovat epäinformatiivisia.

Ensiksi, sekoitusjakauman komponenttien posteriorijakaumat voivat vaihtua (*engl. label switching in mixture models*) keskenään simulaatioketjuittain. Tämä tapahtuu helposti, mikäli gamma-jakaumien muotoparametreilla on epäinformatiiviset priorijakaumat. Palkkasummamallin tapauksessa tämä tarkoittaa sitä, että joissakin simulaatioketjuissa ensimmäiseksi gamma-jakaumaksi määritetty komponentti saisi posteriorijakaumakseen sekoitusjakauman vasemmanpuoleisen jakauman ja toisissa simulaatioketjuissa puo-

lestaan sekoitusjakauman oikeanpuoleisen jakauman. Toiseksi gamma-jakaumaksi määritetty jakauma puolestaan saisi simulaatiokohtaisesti posteriorijakaumakseen aina päinvastaisen jakauman kuin ensimmäiseksi gamma-jakaumaksi määritetty. Tavoitteena simulaatiossa tietenkin on, että jakaumakomponentit saisivat jokaisessa simulaatioketjussa aina saman posteriorijakauman kahden jakaumakomponentin väliltä. (Stan Development Team 2015, 202–204).

Toiseksi, mikäli gamma-jakaumien muotoparametreilla on epäinformatiiviset jakaumat, gamma-jakaumien sekoitusjakauma voi ”romahtaa” yhdeksi gamma-jakaumaksi. Tässä tapauksessa tämä tarkoittaa sitä, että yli nollien jakauma romahtaisi mallinnusaskeleessa 2. esitetystä kahden gamma-jakauman sekoitusjakaumasta mallinnusaskeleen 1. yhdeksi gamma-jakaumaksi. (Stan Development Team 2015, 202–204).

Kuten nettovaikutusmallien yhteydessä, simuloin myös palkkasummamallien parametrien maahanmuuttajaryhmäkohtaiset yhteisposteriorijakaumat Stan-ohjelmaa käyttäen. Stan-ohjelman oppaassa suositellaan ratkaisemaan mainitut kaksi sekoitusjakaumiin liittyvää simulaatio-ongelmaa kolmella mahdollisella tavalla. Ensimmäinen vaihtoehto on määrittää jakaumien sijainnin määrittävien parametrien priorijakaumat niin, että jakaumakomponenteilla on suuruusjärjestys. Toinen vaihtoehto on yksinkertaisesti määrittää jakaumakomponenttien sijaintien määrittävien parametrien priorijakaumat tarkemmiksi (informatiivisemmiksi). Kolmas vaihtoehto on puolestaan säätää Stanin simulointiparametreja. Koska pelkkä jakaumien sijainnin määrittävien parametrien priorijakaumien suuruusjärjestys ei taannut palkkasummien osalta haluttua sekoitusjakaumaa posterioriestimaattien osalta, päätin käyttää gamma-jakaumien muotoparametrien priorijakaumina melko informatiivisia jakaumia. (Stan Development Team 2015, 203–204).

Käytännössä tämä tarkoitti sitä, että simuloin aluksi useita ketjuja posteriorijakaumista melko epäinformatiivisin gamma-jakaumien muotoparametrein, jotta pystyin selvittämään kummankin gamma-jakaumakomponentin muotoparametrin posteriorijakaumat. Käytin sitten varsinaisessa posteriorisimuloinnissa gamma-jakaumien muotoparametrien priorijakaumina gamma-jakaumia, jotka olivat leveämpiä kuin kokeilusimulaatioiden posteriorijakaumat muotoparametreille, mutta kuitenkin sen verran informatiiviset, että ne määrittivät likimain muotoparametrien sijainnin niiden mahdollisten arvojen määrittelyjoukossa.

Käytin mallien estimoinneissa palkkasummista arvoja, joissa ne oli jaettu tuhannella.

Parametrien priorijakaumat

Irakissa syntyneiden osalta lopullisen mallin (malli 3.) priorijakaumat olivat jakautuneet likimain seuraavasti, missä $\mathcal{G}$ viittaa gamma-jakaumaan:

Ensimmäisen gamma-jakauman muotoparametri: $k_{gamma1} \sim \mathcal{G}(4, 4)$.

Toisen gamma-jakauman muotoparametri: $k_{gamma2} \sim \mathcal{G}(80, 2)$.

Gamma-jakaumien ja hurdlen regressiokertoimet: $\beta \sim N(0, 100)$.

θ-parametrin arvon sovitteen vakio: $\beta_{\theta_0} \sim N(0, 2)$.

θ-parametrin arvon sovitteen regressiokertoimet: $\beta_\theta \sim N(0, 100)$.

Lisäksi Suomessa syntyneen kantaväestön lopullisessa mallissa (malli 4.) satunnaisvaikutusten jokaisella neljällä parametreilla λ oli sama priorijakauma:

Gamma-jakautuneiden satunnaisvaikutusten parametrit: $\lambda \sim Cauchy(0, 10)$.

Tässä Suomessa syntyneen kantaväestön lopullisessa mallissa jouduin käytännössä kiinnittämään kahden parametrin arvot priorijakaumien avulla. Nämä olivat 1. gamma-jakauman ja 2. gamma-jakauman vakioparametrit. Ilman tätä kiinnitystä 1. gamma-jakauma ja 2. gamma-jakauma vaihtuivat keskenään simulaatioketjuittain.

5.3.3 Estimoitavien parametrien yhteisposteriorijakauma

Olen edellä määrittänyt neljä eri aineistomallia sekä näiden mallien parametreille priorijakaumat. Malikohtaiset posteriorijakaumat ovat suoraan verrannollisia aineistomallien ja priorijakaumien tuloille:

posteriorijakauma $\propto$ *aineistomalli* $\times$ *priorijakaumat*.

Eli noudattaen tässä tutkimuksessa käytettyjä merkintätapoja:

$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, missä $\boldsymbol{\theta}$ ovat mallin estimoitavat parametrit ja $\boldsymbol{y}$ on havaintoaineisto.

Parametreja koskevaa bayesilaista tilastollista päättelyä varten on selvitettävä posteriorijakauma $p(\boldsymbol{\theta}|\boldsymbol{y})$.

5.3.4 Posteriorijakauman simulointi ja konvergoituminen tavoitejakaumaan

5.3.4.1 Posteriorijakauman simulointi Stan-ohjelmalla

Koska palkkasummamallien parametrien yhteistodennäköisyysjakauma ei noudata mitään tunnettua todennäköisyysjakaumaa – siten että sillä olisi analyyttinen ratkaisu, tai niin, että siitä voisi poimia suoraan arvoja arvoja Monte Carlo -simuloinnilla – simuloin posteriorijakaumat MCMC-menetelmiä käyttäen eli käytännössä Stan-ohjelman avulla (Goodrich et al. 2015). Koska olen esitellyt Stan-ohjelman perusteet nettovaikutusmallinnuksen yhteydessä (luvussa 4.3.6.1), en tee sitä enää uudelleen tässä.

Tyhjän hurdle-gamma-mallin estimoinnissa käytin nettovaikutusmallinnuksen yhteydessä lyhyesti esiteltyä (luku 4.3.6.1) R-ohjelman BRMS-pakettia (Buerkner 2015). BRMS-ohjelma tuottaa Stan-ohjelman koodia ja käyttää Stania posteriorijakaumien simulointiin. Koska Brms-ohjelman aineistomallien jakaumaperheistä löytyy hurdle-gamma-jakauma, käytin BRMS-ohjelmaa tyhjään hurdle-gammajakaumaan perustuvan posteriorin simulointiin. Kehitin myös laajempien mallien Stan-koodit BRMS:n hurdle-gamma-jakauman koodin pohjalta. Esitän tutkimuksen liitteessä C, kuinka simuloida hurdle-gamma-sekoitusjakaumia Stan-ohjelmalla.

5.3.4.2 Posteriorisimulaatiot: simulaatioketjuja koskevat valinnat

Rakensin mallinnusaskeleen 3 (alaluku 5.3.1.3) lisäämällä malliin aina muutaman selittävän muuttujan kerrallaan. Nopeutin tämän mallinnusvaiheen simulaatioita käyttämällä aina pienemmän mallin parametrien odotusarvojen posterioriestimaatteja seuraavan suuremman mallin simulaatioiden alkuarvoina. Lisättävien regressiokertoimien alkuarvoina käytin nollia.

Käytin jokaisen maahanmuuttajaryhmän kohdalla koko aineistolla estimoidun neljän posteriorijakauman estimointiin kahta simulaatioketjua. Kuten nettovaikutusten kohdalla, simulaatioketjuja olisi voinut olla enemmän, mutta ajan säästämiseksi käytin kahta. Simulaatioketjukohtaisten simulaatioiden kokonaislukumäärät ja lämmittelysimulaatioiden (*engl. warm-up*) määrät olivat maahanmuuttajaryhmittäin hieman suuremmat kuin nettovaikutusmallinnuksen yhteydessä. Lämmittelysimulaatioiden määrä lopullisissa malleissa oli noin 800 simulaatiota ja kokonaissimulaatiomäärä oli noin 2 800 simulaatiota per simulaatioketju. Posterioripäättelyihin käytettävä simulaatioiden määrä kahdesta ketjusta oli täten noin 4 000 simulaatiota.

5.3.4.3 Simulaatioketjujen konvergoituminen tavoiteposteriorijakaumaan ja posterioriestimaattien luotettavuus

Posteriorijakaumien simulaatioiden konvergoitumisen tarkistaminen ja MC-keskivirheiden määrittäminen olivat palkkasummien osalta samankaltaiset kuin olotilamallinnuksessa ja nettovaikutusmallinnuksessa, joten en toista niitä tässä (ks. Monte Carlo -keskivirheiden lyhyt esittely luvussa 3.2.5.3).

5.3.5 Mallin sopivuuden tarkistaminen: posterioriennustetarkistukset

Posterioriennustejakauman simulointi ja posterioriennustetarkistusten tekeminen noudattivat palkkasummien osalta olotilamallinnuksen ja nettovaikutusmallinnuksen vastaavia menettelyitä (ks. posterioriennustetarkistuksista luku 3.2.6). Palkkasummien yhteydessä posterioriennusteet poimittiin Irakin ja Somalian osalta mallin 3 määrittämästä aineistomallista ja Suomen osalta mallin 4 määrittämästä aineistomallista siten, että parametrien priorijakaumina käytettiin posteriorisimulaatioiden estimointituloksia.

5.4 Havainnot, estimointitulokset ja posterioriennustetarkistukset

Esitän tässä luvussa työikäisten palkkasummia koskevat havainnot, syntymävaltiokohtaiset mallinnustulokset sekä posterioriennustetarkistukset. Aloitan esittämällä palkkasummamallien parametrien selitetaulukon, joka toimii apuna syntymävaltiokohtaisten mallien tulosten tarkasteluun. Olen esittänyt malliparametrien tarkemman selostuksen sekä valintaperiaatteet edellä.

Koska palkkasummamallien estimointi tapahtui syntymävaltiokohtaisesti, käyn estimointitulokset ja posterioriennustetarkistukset myös syntymävaltio kerrallaan läpi. Esitän jokaisen syntymävaltion kohdalla aluksi selitettävän muuttujan eli palkkasummahavaintojen jakauman havaintojen kokonaislukumäärän ja tärkeimmät tunnusluvut. Lisäksi esitän palkkasummahavaintojen jakauman histogrammina. Kaikkien havaintojen yhteisjakauman lisäksi esitän myös nettovaikutusten henkilökeskiarvojen jakaumalle samat luvut sekä histogrammin. Koska palkkasummien havainnot ovat jakautuneet kahteen osaan, nolliin ja yli nolliin, esitän sekä palkkasummahavaintojen että posterioriennusteiden jakaumat sekä valikoidut jakaumien tunnusluvut erikseen sekä nollien osuuksille että yli nollien jakaumille.

Palkkasummahavaintojen jälkeen esitän selittävien muuttujien jakaumien tärkeimmät tunnusluvut: keskiarvot, keskihajonnat ja kvartiilit. Selittävien muuttujien jakaumatietojen jälkeen esitän 3/4 perättäiseen aineistomalliin pohjautuvien parametrien posteriorijakaumien odotusarvoestimaatit. Estimoitavien parametrien odotusarvoestimaattien lisäksi esitän ennustetarkkuutta mittaavien WAIC:n ja LOO:n estimaatit sekä näiden estimaattien keskivirheet kaikille 3/4 mallille. WAIC:n ja LOO:n pienemmät estimaatit tarkoittavat parempaa ennustarkkuutta. Esitän lisäksi syntymävaltiokohtaisesti sen mallin osalta, jota käytän elinkaariennusteiden laatimiseen uusille maahanmuuttajille ja Suomessa syntyville, muitakin parametrien posteriorijakaumien estimaatteja kuin odotusarvon. Irakissa ja Somaliassa syntyneiden osalta tämä oli malli 3 ja Suomessa syntyneiden osalta malli 4.

Posterioriennustarkistuksissa vertaan posterioriennusteita havaintoaineistoon kaikkien havaintojen osalta, henkilöiden keskiarvojen osalta sekä selittävien muuttujien luokkien mukaan jaoteltuina. Kaikkien havaintojen ja henkilöiden keskiarvojen osalta vertaan posterioriennusteiden tuottamia *jakaumia* havaintoaineiston vastaaviin jakaumiin. Tärkeimpien selittävien muuttujien kuten sukupuolen, iän, maassaoloajan ja maahanmuuttovuosien osalta vertaan posterioriennusteiden *keskiarvoja* havaintoaineiston vastaaviin keskiarvoihin näiden selittävien muuttujien luokkien osalta. Teen vertailun erikseen sekä nollahavaintojen osuuksille sekä yli nollia koskeville jakaumille.

Parametrien selitteet

Taulukossa 5.1 on esitetty palkkasummien syntymävaltiokohtaisten mallien estimoitavien parametrien nimet ja niiden selitteet. Taulukossa esitettävät selittäviä muuttujia koskevat regressiokertoimet ovat Somaliassa syntyneiden palkkasummamallista, ja nämä vaihtelevat syntymävaltiokohtaisten mallien välillä. Taulukosta saa kuitenkin yleisen käsityksen siitä, millaisia parametreja malleissa on. Gamma-jakautuneiden satunnaisvaikutusten parametrit (λ) koskevat vain Suomessa syntyneiden palkkasummamallia (tutkimuksen tässä versiossa).

Taulukon 5.1 parametrit, joista ensimmäinen on ”nainen”, ja viimeinen on, ”Muutto 20–39-v.”, ovat selittävien muuttujien regressiokertoimia. Parametrien nimet ja selitteet kertovat, mihin selittävään muuttujaan kukin regressiokerroin liittyy. Indikaattorimuuttujien osalta selitetaulukossa ovat näiden vertailuluokat. Gamma-jakaumien muotoparametreja on jokaisen syntymävaltion malleissa kaksi kappaletta lukuun ottamatta mallia 1. Muotoparametreja on yksi kumpaakin gamma-sekoitusjakauman gamma-jakautunutta komponenttia kohti.

Taulukko 5.1: *Palkkasummamallien estimoitavien parametrien nimet ja niiden selitteet. Esimerkkinä Somalia.*

Parametrinimi	**Selite**
Parametrinimi (H)	Parametri liittyy ”hurdleen”, joka kertoo, saako havainto arvoksi nollan (eikä yli nollaa).
Parametrinimi ($\mathcal{G}_1$)	Parametri liittyy gamma1-jakaumaan.
Parametrinimi ($\mathcal{G}_2$)	Parametri liittyy gamma2-jakaumaan.
Parametrinimi (θ)	Parametri liittyy θ:aan eli indikaattorimuuttujaan, joka kertoo, kuuluuko havainto gamma1-jakaumaan (eikä gamma2-jakaumaan).
Vakio	Lineaarisen sovitteen arvo, kun kaikkien (standardoitujen) selittävien muuttujien arvot ovat keskiarvoja.
Nainen	Vertailuluokka: mies.
Ikä	Arvo muuttuu ajassa.
Ikä^2	Iän neliö: $(\text{Ikä} - 39)^2$. Arvo muuttuu ajassa.
Maassaolo (vuosia)	Aika vuosina maahanmuutosta. Arvo muuttuu ajassa.
Maassaolo < 6	Vertailuluokka: maassaoloaika yli 5 vuotta.
Maassaolo > 9	Vertailuluokka: maassaoloaika alle 10 vuotta.
Maassaolo > 15	Vertailuluokka: maassaoloaika alle 16 vuotta.
Maassaolo $\times$ (maassaolo < 6)	Maassaoloajan vaikutus: alle 6 vuotta maassa verrattuna yli 5 vuotta maassa.
Maassaolo $\times$ (maassaolo > 9)	Maassaoloajan vaikutus: yli 9 vuotta maassa verrattuna alle 10 vuotta maassa.
Maassaolo $\times$ (maassaolo > 15)	Maassaoloajan vaikutus: yli 15 vuotta maassa verrattuna alle 16 vuotta maassa.
Ikä < 30	Vertailuluokka: ikä yli 29 vuotta.
(Ikä < 30) $\times$ ikä	Iän vaikutus: alle 30-vuotiailla verrattuna yli 29-vuotiaat.
(Ikä < 30) $\times$ maassaolo	Maassaoloajan vaikutus: alle 30-vuotiailla verrattuna yli 29-vuotiaat.
Ikä > 50	Vertailuluokka: ikä alle 51 vuotta.
(Ikä > 50) $\times$ ikä	Iän vaikutus: yli 50-vuotiailla verrattuna alle 51-vuotiaat.
Muuttovuosi v. 1993–2001	Maahanmuuttovuosiluokka. Vertailuluokka: muutto v. 2002–2014.
Korkein koulutus: 2. asteen koulutus	Vertailuluokka: muuttovuonna ei 2.asteen koulutusta/ei tietoa.
Korkein koulutus: 3. asteen koulutus	Vertailuluokka: muuttovuonna ei 2.asteen koulutusta/ei tietoa.
Pääasiallinen toiminta: opiskelija	Vertailuluokka: pääasiallinen toiminta muuttovuonna: muu (ml. työtön).

Jatkuu seuraavalla sivulla

Taulukko 5.1 – jatkuu edelliseltä sivulta

Parametrinimi	Selite
Pääasiallinen toiminta: työllinen	Vertailuluokka: pääasiallinen toiminta muuttovuonna: muu (ml. työtön).
Alle 7-vuotiaiden lasten lukumäärä	Lasten määrä henkilön asuntokunnassa maahanmuuttovuonna.
Kotikunta Uudellamaalla	Vertailuluokka: kotikunta ei Uudellamaalla maahanmuuttovuonna.
Muutto alle 20-v.	Maahanmuuton ikäluokka. Vertailuluokka: muutto yli 39-v.
Muutto 20–39-v.	Maahanmuuton ikäluokka. Vertailuluokka: muutto yli 39-v.
Muotoparametri	Gamma-jakauman muotoparametri.
λ	Gamma-jakautuneen satunnaisvaikutuksen parametri.

5.4.1 Somalia

5.4.1.1 Somalia: havaintoaineisto

Taulukossa 5.2 on esitetty Somaliassa syntyneiden 18–62-vuotiaiden palkkasummahavaintojen lukumäärä ja havaintojen jakauman keskeisimmät tunnusluvut. Havainnot koskevat vuosia 2005–2014. Taulukossa on esitetty samat tiedot myös havaintoaineiston henkilöiden nettovaikutushavaintojen keskiarvoille. Mallinnusaineistoon ei sisälly alle 16-vuotiaana muuttaneita. Nämä palkkasummahavainnot muodostavat selitettävän muuttujan palkkasummien mallinnuksessa. Taulukossa 5.2 on esitetty kaikkien palkkasummahavaintojen jakaumatietojen lisäksi vielä erikseen niiden palkkasummahavaintojen, jotka ovat yli nollan, jakauman keskiarvo ja keskihajonta. Lisäksi taulukon alimmalla rivillä on esitetty nollien osuudet havainnoista.

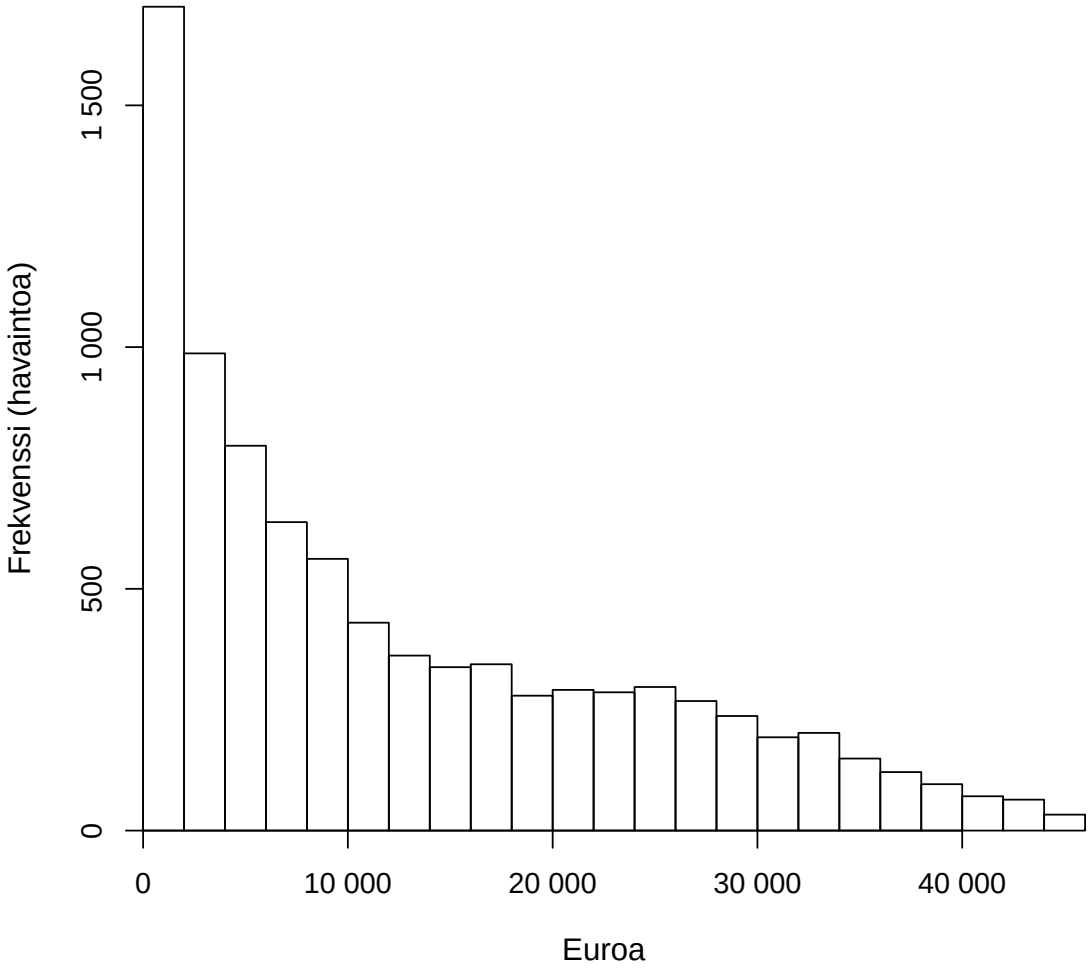
Somaliassa syntyneiden palkkasummahavaintojen mallinnusaineisto koostuu 29 817 henkilövuositasoisesta palkkasummahavainnosta, jotka on poimittu paneeliaineistona/pitkittäisaineistona 4 951 henkilöltä. Taulukon 5.2 tietojen mukaan esimerkiksi kaikkien palkkasummahavaintojen keskiarvo on 3 880 euroa, keskihajonta on 8 940 euroa ja viisi prosenttia kaikista havainnoista on 26 210 euroa suurempia. Havaintoaineiston henkilöiden palkkasummahavaintojen keskiarvojen keskiarvo on 2 720 euroa ja keskihajonta 6 290 euroa. Yli nollien keskiarvo kaikkien havaintojen joukossa on 13 080 euroa ja henkilöiden yli nollien keskiaro on puolestaan 6 310 euroa. Nollien osuus koko havaintoaineistossa on 70 prosenttia.

Havaintoaineistossa noin 30 %:ia havainnoista on yli nollan. Kuvassa 5.4.1.2 on esitetty näiden yli nollien olevien palkkasummien jakauman histogrammi. Jakaumasta on katkaistu pois ylin prosentti.

Taulukko 5.2: *Somaliassa syntyneet 18–62-vuotiaat vuosina 2005–2014. Palkkasummahavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa sekä henkilöiden lukumäärä ja henkilötasoisten palkkasummahavaintojen keskiarvojen jakauman tunnusluvut mallinnusaineistossa. Havaintojen osalta "ylinollat" käsittävät vain ne havainnot, joissa palkkasumma on yli nolla euroa; henkilöiden osalta ne henkilöt, joiden keskiarvo on yli nollan. Nollien osuudet kertovat nollien osuudet kaikkien havaintojen sekä kaikkien henkilöiden keskiarvojen joukoissa. Euromääräiset tunnusluvut pyöristetty lähimpään kymmeneen euroon.*

	Havainnot	Henkilöt
Kokonaislukumäärä (N)	29 817	4 951
Keskiarvo	3 880	2 720
Keskihajonta	8 940	6 290
1. prosenttipiste (1 %)	0	0
5. prosenttipiste (5 %)	0	0
10. prosenttipiste (10 %)	0	0
Alakvartiili (25 %)	0	0
Mediaani (50 %)	0	0
Yläkvartiili (75 %)	1 350	1 900
90. prosenttipiste (90 %)	16 210	9 050
95. prosenttipiste (95 %)	26 210	17 230
99. prosenttipiste (99 %)	39 140	30 190
Ylinollat: keskiarvo	13 080	6 310
Ylinollat: keskihajonta	12 220	8 330
Nollien osuus	0,70	0,57

Histogrammista näkee, että havaintojen suurin tiheys (frekvenssi) on lähellä nollaa ja laskee oikealle päin mentäessä. Noin 25 000 euron kohdalla yli nollien jakaumassa on kuitenkin paikallinen huippu. Mallinnuksessa toisen gamma-jakauman moodi sijoittuu suunnilleen tähän kohtaan. Tämä paikallinen huippu on helpommin nähtävissä Suomessa syntyneiden vastaavassa histogrammissa (ks. histogrammi, s. 314).



Kuva 5.2: *Somaliassa syntyneet 18–62-vuotiaat. Yli nollan olevien palkkasummahavaintojen katkaistun jakauman histogrammi mallinnusaineistossa. Jakaumasta on poistettu ylin prosentti yksityisyydensuojan varmistamiseksi.*

Taulukossa 5.3 on esitetty Somaliassa syntyneiden palkkasummamallien selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden kaikkien havaintojen joukossa. Edellä parametrien selitetaulukossa (s. 262) on lyhyitä selityksiä selittävistä muuttujista.

Taulukko 5.3: *Somaliassa syntyneet 18–62-vuotiaat. Palkkasummamallien selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta).*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,54	0,50	0	1	1
Ikä	34,3	10,6	26	33	41
Ikä^2	1287	795	676	1089	1681
Maassaolo (vuosia)	6,8	5,7	2	5	11
Maassaolo < 6	0,52	0,50	0	1	1
Maassaolo > 9	0,31	0,46	0	0	1
Maassaolo > 15	0,10	0,30	0	0	0
Maassaolo $\times$ maassaolo < 6	1,17	1,62	0	0	2
Maassaolo $\times$ maassaolo > 9	4,44	6,78	0	0	11
Maassaolo $\times$ maassaolo > 15	1,80	5,39	0	0	0
Ikä < 30	0,38	0,49	0	0	1
Ikä $< 30 \times$ ikä	9,14	11,82	0	0	22
Ikä $< 30 \times$ maassaolo	1,38	2,65	0	0	2
Ikä > 50	0,09	0,28	0	0	0
Ikä $> 50 \times$ ikä	4,91	15,80	0	0	0
Muuttovuosi v. 1993–2001	0,36	0,48	0	0	1
Korkein koulutus: 2. asteen koulutus	0,05	0,21	0	0	0
Korkein koulutus: 3. asteen koulutus	0,01	0,11	0	0	0
Korkein koulutus: ei 2.astetta/ei tietoa	0,94	0,23	1	1	1
Pääasiallinen toiminta: opiskelija	0,09	0,29	0	0	0
Pääasiallinen toiminta: työllinen	0,03	0,16	0	0	0
Pääasiallinen toiminta: muu	0,88	0,32	1	1	1
Alle 7-vuotiaiden lasten lukumäärä	0,49	0,92	0	0	1
Kotikunta Uudellamaalla	0,58	0,49	0	1	1
Muutto alle 20-v.	0,26	0,44	0	0	1
Muutto 20–39-v.	0,60	0,49	0	1	1
Muutto yli 39-v.	0,14	0,35	0	0	0

5.4.1.2 Somalia: estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Taulukossa 5.4 on esitetty perättäisten palkkasummamallien parametrien odotusarvoestimaatit Somaliassa syntyneille 18–62-vuotiaille. Olen käsitellyt perättäisten mallien rakentamista mallinnusaskel kerrallaan alaluvussa 5.3.1. Estimoitavien parametrien selitteet on puolestaan esitetty parametrien selitetaulukossa

s. 262. Selitetaulukossa on myös selvennetty lyhyesti, mitä tarkoittaa, että parametri liittyy joko hurdleen (H), ensimmäiseen gamma-jakaumaan ($\mathcal{G}_1$), toiseen gamma-jakaumaan ($\mathcal{G}_2$) tai θ-indikaattoriin, joka kertoo, kuuluuko havainto ensimmäiseen gamma-jakaumaan (eikä toiseen) kahden gamma-jakauman sekoitusjakaumassa

Vakioparametrien tulkinta

Taulukossa 5.4 on esitetty neljä vakiotermiä, yksi jokaista mallin neljää komponenttia ($H, \mathcal{G}_1, \mathcal{G}_2, \theta$) kohti. Gamma-jakaumia koskevat vakiotermit ovat euromääräisiä, ja hurdlea ja θ:a koskevat vakiotermit puolestaan englanninkielistä termiä käyttäen *odds*eja. Odds on $\frac{p}{1-p}$, missä p on hurdlen tapauksessa todennäköisyys, että havainto saa arvokseen nollan, ja θ:n tapauksessa puolestaan todennäköisyys, että (yli nollan oleva) havainto kuuluu ensimmäiseen gamma-jakaumaan (eikä toiseen).

Malli 1 koostuu hurdlesta (nollien osuus) sekä yhdestä gamma-jakaumasta. Taulukon 5.4 mallia 1 koskevien parametrien odotusarvoestimaattien mukaan hurdlen eli nollahavainnon todennäköisyys on $\frac{2{,}37}{1+2{,}37} \approx 0{,}70$, koska hurdlen odds $\frac{p}{1-p} = 2{,}37$. Todennäköisyys nollahavainnolle vastaa nollien määrää havaintoaineistossa eli 70 %:ia havainnoista. Gamma-jakauman odotusarvo ”Vakio ($\mathcal{G}_1$)” on 13 080 euroa. Se on samalla yli nollien olevien palkkasummien odotusarvo mallissa 1. Tämä vastaa täysin yli nollien olevien havaintojen keskiarvoa, joka on esitetty edellä taulukossa 5.2. Gamma-jakauman muotoparametrin odotusarvo on 0,82, mikä tarkoittaa sitä, että gamma-jakauman todennäköisyystiheysfunktio on aidosti laskeva määrittelyjoukossaan eli positiivisilla reaaliluvuilla.

Mallissa 2 yli nollien mallinnus koostuu kahden gamma-jakauman sekoitusjakaumasta. Ensimmäisen gamma-jakauman odotusarvo on 9 030 euroa ja toisen puolestaan 27 800 euroa. Odotusarvo todennäköisyydelle kuulua ensimmäiseen gamma-jakaumaan on $\frac{3{,}64}{1+3{,}64} \approx 0{,}78$. Esitin aineistomallien rakentamisen yhteydessä mallinnukseen liittyvän skemaattisen oletuksen siitä, että ensimmäiseen gamma-jakaumaan kuuluvat palkkasummat, jotka koskevat havaintoja, joissa henkilö ei ole ollut koko vuotta töissä. Jos oletus pitää paikkansa, niin havaintoaineistossa noin 78 %:ia niistä, joilla oli palkkatuloja, eivät olleet kuitenkaan koko vuotta töissä. Oletuksen paikkansapitävyys ei ole olennaista mallinnuksen tavoitteiden kannalta, mutta saattaa auttaa hahmottamaan tuloksia. Oletuksen puitteissa 9 030 euroa olisi palkkasumman odotusarvo niille, jotka eivät olleet koko vuotta töissä, ja 27 800 euroa puolestaan niille, jotka olivat koko vuoden töissä. Toisen gamma-jakauman odotusarvoa (27 800e) voi vielä verrat yli nollien histogrammiin edellä (kuva), missä suunnilleen tämän odotusarvon kohdalla on havaittavissa yli nollien jakauman paikallinen huippu.

Mallin 3 vakiotermit poikkeavat mallin 2 vakiotermeistä lähinnä sen vuoksi, että olen laskenut iän neliön ikähavainnoista, jotka on keskistetty 39 vuoteen sen sijaan, että ne olisi keskistetty iän keskiarvoon havaintoaineistossa (ks. parametrien selitteet, s. 262).

Taulukko 5.4: *Perättäisten palkkasummamallien parametrien odotusarvoestimaatit, mallikohtaiset ennustetarkkuuden estimaatit ja keskihajonnat (Waic ja LOO). Mallien valinta on esitetty alaluvussa 5.3.1. H:n ja θ:n vakioparametrit ovat oddseja, regressiokertoimet vedonlyöntisuhteita (engl. odds ratio). Gamma-jakaumien ($\mathcal{G}_1$ ja $\mathcal{G}_2$) vakioparametrit ovat euroja, regressiokertoimet odotusarvon lineaarisia kertoimia. Indikaattorimuuttujien osalta regressiokertoimet ovat muutoksia suhteessa vertailuryhmään, ja jatkuvien muuttujien osalta regressiokertoimet ovat muutoksia, kun muuttujan arvo kasvaa yhdellä keskihajonnalla.*

	Malli 1	Malli 2	Malli 3
Vakioparametrit:			
Vakio (H)	0,82	2,37	0,97
Vakio ($\mathcal{G}_1$)	13 080	9 030	13 230
Vakio ($\mathcal{G}_2$)		27 800	26 760
Vakio (θ)		3,65	3,93
Hurdle-regressiokertoimet (H):			
Nainen (H)			3,29
Ikä (H)			1,03
Ikä^2 (H)			0,95
Maassaolo (vuosia) (H)			0,30
Maassaolo $<$ 6 (H)			0,53
Maassaolo $\times$ (maassaolo $<$ 6) (H)			0,12
Maassaolo $>$ 15 (H)			1,03
Maassaolo $\times$ (maassaolo $>$ 15) (H)			0,96
Ikä $<$ 30 (H)			0,69
(Ikä $<$ 30) $\times$ ikä (H)			1,04
(Ikä $<$ 30) $\times$ maassaolo (H)			0,83
Ikä $>$ 50 (H)			0,13
Ikä $>$ 50 $\times$ ikä (H)			4,75
Muuttovuosi v. 1993–2001 (H)			0,71
Korkein koulutus: 2. asteen koulutus (H)			0,79
Korkein koulutus: 3. asteen koulutus (H)			0,71
Pääasiallinen toiminta: opiskelija (H)			0,77
Pääasiallinen toiminta: työllinen (H)			0,15
Alle 7-vuotiaiden lasten lukumäärä (H)			1,12
Kotikunta Uudellamaalla (H)			0,91
Muutto alle 20-v. (H)			0,29
Muutto 20–39-v. (H)			0,46
1. gamma-jakauman regressiokertoimet ($\mathcal{G}_1$):			
Nainen ($\mathcal{G}_1$)			0,91
Ikä ($\mathcal{G}_1$)			1,47

Jatkuu seuraavalla sivulla

Taulukko 5.4 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3
Ikä^2 ($\mathcal{G}_1$)			0,71
Maassaolo (vuosia) ($\mathcal{G}_1$)			2,05
Maassaolo $<$ 6 ($\mathcal{G}_1$)			1,77
Maassaolo $\times$ (maassaolo $<$ 6) ($\mathcal{G}_1$)			3,26
Maassaolo $\times$ (maassaolo $>$ 9) ($\mathcal{G}_1$)			0,85
Ikä $<$ 30 ($\mathcal{G}_1$)			1,60
(Ikä $<$ 30) $\times$ maassaolo ($\mathcal{G}_1$)			0,86
Ikä $>$ 50 ($\mathcal{G}_1$)			1,28
Muuttovuosi v. 1993–2001 ($\mathcal{G}_1$)			1,22
Korkein koulutus: 2. asteen koulutus ($\mathcal{G}_1$)			1,06
Korkein koulutus: 3. asteen koulutus ($\mathcal{G}_1$)			2,13
Pääasiallinen toiminta: opiskelija ($\mathcal{G}_1$)			1,06
Pääasiallinen toiminta: työllinen ($\mathcal{G}_1$)			2,43
Kotikunta Uudellamaalla ($\mathcal{G}_1$)			1,19
Muutto alle 20-v. ($\mathcal{G}_1$)			1,34
2. gamma-jakauman regressiokertoimet ($\mathcal{G}_2$):			
Ikä ($\mathcal{G}_2$)			1,05
Ikä^2 ($\mathcal{G}_2$)			0,97
Maassaolo (vuosia) ($\mathcal{G}_2$)			1,19
Maassaolo $<$ 6 ($\mathcal{G}_2$)			1,00
Maassaolo $\times$ (maassaolo $<$ 6) ($\mathcal{G}_2$)			1,11
Maassaolo $>$ 9 ($\mathcal{G}_2$)			1,08
Maassaolo $\times$ (maassaolo $>$ 9) ($\mathcal{G}_2$)			0,86
Ikä $<$ 30 ($\mathcal{G}_2$)			1,01
(Ikä $<$ 30) $\times$ maassaolo ($\mathcal{G}_2$)			0,95
Muuttovuosi v. 1993–2001 ($\mathcal{G}_2$)			1,04
Kuuluminen 1. gamma-jakaumaan, regressiokertoimet (θ):			
Nainen (θ)			9,39
Ikä (θ)			1,64
Ikä^2 (θ)			1,09
Maassaolo (vuosia) (θ)			0,18
Maassaolo $<$ 6 (θ)			0,77
Maassaolo $\times$ (maassaolo $<$ 6) (θ)			0,25
Maassaolo $>$ 9 (θ)			1,11
Maassaolo $\times$ (maassaolo $>$ 9((θ)			1,40
Ikä $<$ 30 (θ)			1,23
(Ikä $<$ 30) $\times$ maassaolo (θ)			0,60
Muuttovuosi v. 1993–2001 (θ)			1,95
Muutto alle 20-v. (θ)			1,43

Jatkuu seuraavalla sivulla

Taulukko 5.4 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3
Gamma-jakaumien muotoparametrit:			
Muotoparametri($\mathcal{G}_1$)	0,82	0,79	0,89
Muotoparametri($\mathcal{G}_2$)		11,8	10,4
Mallien ennustetarkkuudet Waic ja LOO:			
Waic (estimaatti)	75 964	75 267	66 607
Waic (estimaatin keskivirhe)	(526)	(517)	(475)
LOO (estimaatti)	75 964	75 267	66 607
LOO (estimaatin keskivirhe)	(526)	(517)	(475)

Regressiokertoimien tulkinta

H:n ja θ:n regressiokertoimet vedonlyöntisuhteita (engl. odds ratio) ja gamma-jakaumien ($\mathcal{G}_1$ ja $\mathcal{G}_2$) regressiokertoimet ovat puolestaan jakaumien odotusarvojen lineaarisia kertoimia. Indikaattorimuuttujien osalta regressiokertoimet ovat muutoksia suhteessa vertailuryhmään, ja jatkuvien muuttujien osalta regressiokertoimet ovat muutoksia, kun muuttujan arvo kasvaa yhdellä keskihajonnalla.

Hurdlen (ja θ:n) regressiokertoimien tulkinta. Palkkasummien havainto saa arvoksi nollan, kun hurdle-indikaattori (H) saa arvoksi ykkösen. Olkoon H:n todennäköisyys saada arvoksi yksi: $p(H = 1)$. Olkoon lisäksi ehdollinen odotusarvo $p(H = 1|nainen = 1)$, todennäköisyys naiselle saada hurdlen arvoksi 1 (ja palkkasumman arvoksi 0), ja $p(H = 1|nainen = 0)$ puolestaan miehelle todennäköisyys saada hurdlen arvoksi 1 (ja palkkasumman arvoksi 0). Tällöin taulukossa 5.4 esitetyn indikaattorimuuttujan ”Nainen (H)” regressiokertoimen odotusarvolla 3,29 on seuraava tulkinta, kun muiden selittävien muuttujien arvot pidetään vakioina:

$$\frac{p(H = 1|\text{nainen} = 1)}{1 - p(H = 1|\text{nainen} = 1)} = 3{,}29 \times \frac{p(H = 1|\text{nainen} = 0)}{1 - p(H = 1|\text{nainen} = 0)}.$$

Naisilla odds saada hurdlen arvoksi 1 on täten odotusarvoisesti 3,29-kertainen miehiin verrattuna, kun muiden selittävien muuttujien arvot pidetään vakioina. Tällöin naisilla on korkeampi todennäköisyys saada palkkasumman arvoksi nolla kuin miehillä.

Ikä on jatkuva selittävä muuttuja, jonka keskihajonta mallinnusaineistossa on taulukon 5.3 mukaan 10,6 ikävuotta. Merkitään ikävuosien saamaa arvoa x:llä siten, että $x \in \{18, 19, \ldots, 61, 62\}$. Tällöin taulukossa 5.4 esitetyn ”Ikä (H)” regressiokertoimen odotusarvolla 1,03 on seuraava tulkinta, kun muiden selittävien muuttujien arvot pidetään vakioina:

$$\frac{p(H = 1|\text{ikä} = x)}{1 - p(H = 1|\text{ikä} = x)} = 1{,}03 \times \frac{p(H = 1|\text{ikä} = x + 10{,}6)}{1 - p(H = 1|\text{ikä} = x + 10{,}6)}.$$

Tämä tarkoittaa sitä, että odds saada hurdlen arvoksi 1 kasvaa täten odotusarvoisesti 1,03-kertaiseksi, kun ikävuosien määrä kasvaa 10,6 vuodella, kun muiden selittävien muuttujien arvot pidetään vakioina. Ikävuosien keskihajonta aineistossa on 10,6 vuotta. Jatkuvien muuttujien kohdalla regressiokertoimet kuvaavat sitä, kuinka paljon odds muuttuu, kun jatkuvan selittävän muuttujan arvo muuttuu yhdellä keskihajonnalla. Hurdlea koskevien regressiokertoimien tulkinnat ovat samanlaiset myös θ-indikaattoria koskevien regressiokertoimien osalta.

Gamma-jakaumien regressiokertoimien tulkinta. Gamma-jakaumien ($\mathcal{G}_1$ ja $\mathcal{G}_2$) regressiokertoimet ovat näiden kahden jakauman odotusarvon lineaarisia kertoimia. Indikaattorimuuttujien osalta regressiokertoimet ovat muutoksia suhteessa vertailuryhmään, ja jatkuvien muuttujien osalta regressiokertoimet ovat muutoksia, kun muuttujan arvo kasvaa yhdellä keskihajonnalla.

Olkoon ensimmäisen gamma-jakauman ($\mathcal{G}_1$) odotusarvo $E(\mathcal{G}_1)$. Olkoon nyt tämä odotusarvo naisille $E(\mathcal{G}_1|nainen = 1)$ ja miehille $E(\mathcal{G}_1|nainen = 0)$. Taulukossa 5.4 esitetyn indikaattorimuuttujan ”Nainen ($\mathcal{G}_1$)” regressiokertoimen odotusarvolla 0,91 on seuraava tulkinta, kun muiden selittävien muuttujien arvot pidetään vakioina:

$E(\mathcal{G}_1|nainen = 1) = 0{,}91 \times E(\mathcal{G}_1|nainen = 0)$.

Naisilla ensimmäisen gamma-jakauman euromääräinen odotusarvo on täten 0,91-kertainen miehiin verrattuna, kun muiden selittävien muuttujien arvot pidetään vakioina. Tämä tarkoittaa sitä, että jos miehillä ensimmäisen gamma-jakauman odotusarvo on esimerkiksi 10 000 euroa, niin naisilla tämä odotusarvo on 0,91×10 000 euroa = 9 100 euroa – kun muiden selittävien muuttujien arvot pidetään vakioina. Naisindikaattoria koskeva regressiokertoimen tulkinta pätee myös muille indikaattorimuuttujille.

Käsitellään nyt ikää koskevaa regressiokerrointa esimerkkinä jatkuvien muuttujien regressiokertoimien tulkinnasta. Ikä on jatkuva selittävä muuttuja, jonka keskihajonta mallinnusaineistossa on taulukon 5.3 mukaan 10,6 ikävuotta. Merkitään ikävuosien saamaa arvoa x:llä siten, että $x \in \{18, 19, \ldots, 61, 62\}$. Tällöin taulukossa 5.4 esitetyn ”Ikä ($\mathcal{G}_1$)” regressiokertoimen odotusarvolla 1,47 on seuraava tulkinta, kun muiden selittävien muuttujien arvot pidetään vakioina:

$E(\mathcal{G}_1|\text{ikä} = x) = 1{,}47 \times E(\mathcal{G}_1|\text{ikä} = x + 10{,}6)$.

Tämä tarkoittaa sitä, että ensimmäisen gamma-jakauman euromääräinen odotusarvo kasvaa 1,47-kertaiseksi, kun ikävuosien määrä kasvaa 10,6 vuodella ja muiden selittävien muuttujien arvot pidetään vakioina. Ikävuosien keskihajonta aineistossa on 10,6 vuotta. Jatkuvien muuttujien kohdalla regressiokertoimet ovat kertoimia gamma-jakauman odotusarvon muutokselle, kun jatkuvan selittävän muuttujan arvo muuttuu yhdellä keskihajonnalla. Toista gamma-jakaumaa ($\mathcal{G}_2$) koskevien regressiokertoimien tulkinnat ovat samanlaiset kuin tässä esitetyt ensimmäistä gamma-jakaumaa ($\mathcal{G}_1$) koskevat tulkinnat.

Käsittelen mallin 3 regressiokertoimien tuloksia tarkemmin lopullisen mallin – eli mallin 3 – parametrien posteriorijakaumaestimaattien tulosten yhteydessä alla.

Taulukon 5.4 ennustemittarien Waic ja LOO mukaan isommat mallit ovat parempia ennustajia kuin pienemmät mallit. Tosin ero mallin 1 ja mallin 2 välillä ei ole suuri, mikä johtuu siitä, että Somaliassa

syntyneiden yli nollien havaintoaineistoa voisi kuvata melko hyvin vain yhdellä gamma-jakaumalla kahden gamma-jakauman sekoitusjakauman sijaan, sillä ”koko vuoden töissä olleiden” määrä on suhteellisen pieni, mikä näkyy jakauman paikallisen huipun pienuutena noin 25 000 euron kohdalla (ks. jakauman histogrammi, kuvio 5.4.1.2).

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” mallin on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän uusien maahanmuuttajien elinkaariennusteissa palkkasummien ennustamiseen Somaliassa syntyneiden osalta. Somalian osalta lopullinen mallin on malli 3, joka sisältää kaikille havainnoille yhteiset selittävät muuttujat. Tutkimuksen tässä versiossa Somaliassa syntyneille ei ole vielä olemassa mallia 4, joka sisältäisi palkkasummien mallin neljän vakiotermin satunnaisvaikutukset. Selittävien muuttujien vakiotermien ja regressiokertoimien tulkinnat on esitetty yllä, joten en toista niitä tässä.

Taulukossa 5.5 on esitetty Somaliassa syntyneiden lopullisen palkkasummamallin (malli 3) posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat Monte Carlo-keskivirheet (MC-keskivirheet) sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli ykköstä, en käsittele niitä tarkemmin.

Taulukko 5.5: *Somaliassa syntyneiden 18–62-vuotiaiden palkkasummien lopullinen malli (malli 3). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakioparametrit:									
Vakio (H)	0,97	0,08	0,83	0,91	0,96	1,02	1,12	0,002	1,00
Vakio ($\mathcal{G}_1$)	13 230	770	11 760	12 700	13 200	13 740	14 780	18	1,00
Vakio ($\mathcal{G}_2$)	26 760	1 000	24 840	26 060	26 740	27 430	28 770	23	1,00
Vakio (θ)	3,93	0,74	2,75	3,40	3,86	4,36	5,64	0,0159	1,00
Regressiokertoimet (H):									
Nainen (H)	3,29	0,10	3,10	3,23	3,29	3,36	3,49	0,0008	1,00
Ikä (H)	1,03	0,06	0,92	0,99	1,03	1,07	1,15	0,0012	1,00
Ikä^2 (H)	0,95	0,08	0,80	0,89	0,94	1,00	1,12	0,0017	1,00
Maassaolo (vuosia) (H)	0,30	0,02	0,27	0,29	0,30	0,31	0,33	0,0003	1,00
Maassaolo < 6 (H)	0,53	0,04	0,46	0,51	0,53	0,56	0,61	0,0003	1,00
Maassaolo $\times$ (maassaolo < 6) (H)	0,12	0,01	0,10	0,11	0,12	0,12	0,14	0,0001	1,00
Maassaolo > 15 (H)	1,03	0,29	0,56	0,82	0,99	1,19	1,72	0,0016	1,00
Maassaolo $\times$ (maassaolo > 15) (H)	0,96	0,15	0,70	0,85	0,95	1,05	1,29	0,0008	1,00
Ikä < 30 (H)	0,69	0,07	0,56	0,64	0,69	0,74	0,85	0,0008	1,00
(Ikä < 30) $\times$ ikä (H)	1,04	0,23	0,66	0,87	1,01	1,18	1,57	0,0025	1,00
(Ikä < 30) $\times$ maassaolo (H)	0,83	0,06	0,72	0,79	0,83	0,87	0,94	0,0004	1,00
Ikä > 50 (H)	0,13	0,06	0,04	0,08	0,11	0,16	0,28	0,0004	1,00
(Ikä > 50) $\times$ ikä (H)	4,75	1,41	2,59	3,74	4,55	5,54	7,91	0,0086	1,00
Muuttovuosi v. 1993–2001 (H)	0,71	0,04	0,64	0,68	0,71	0,73	0,78	0,0003	1,00
Korkein koulutus: 2. asteen koulutus (H)	0,79	0,05	0,69	0,75	0,78	0,82	0,89	0,0002	1,00
Korkein koulutus: 3. asteen koulutus (H)	0,71	0,09	0,55	0,64	0,70	0,76	0,90	0,0002	1,00
Pääasiallinen toiminta: opiskelija (H)	0,77	0,04	0,69	0,74	0,76	0,79	0,84	0,0002	1,00
Pääasiallinen toiminta: työllinen (H)	0,15	0,01	0,12	0,14	0,15	0,15	0,17	0,0000	1,00
Alle 7-vuotiaiden lasten lukumäärä (H)	1,12	0,02	1,08	1,11	1,12	1,13	1,16	0,0003	1,00

Jatkuu seuraavalla sivulla

Taulukko 5.5 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Kotikunta Uudellamaalla (H)	0,91	0,03	0,85	0,89	0,91	0,93	0,96	0,0002	1,00
Muutto alle 20-v. (H)	0,29	0,03	0,23	0,27	0,29	0,31	0,36	0,0003	1,00
Muutto 20–39-v. (H)	0,46	0,04	0,39	0,44	0,46	0,49	0,55	0,0003	1,00
Regressiokertoimet ($\mathcal{G}_1$):									
Nainen ($\mathcal{G}_1$)	0,91	0,05	0,82	0,87	0,90	0,94	1,00	0,0004	1,00
Ikä ($\mathcal{G}_1$)	1,47	0,10	1,29	1,40	1,47	1,54	1,67	0,0024	1,00
Ikä^2 ($\mathcal{G}_1$)	0,71	0,03	0,64	0,68	0,71	0,73	0,78	0,0009	1,00
Maassaolo (vuosia) ($\mathcal{G}_1$)	2,05	0,15	1,76	1,95	2,05	2,15	2,36	0,0034	1,00
Maassaolo $<$ 6 ($\mathcal{G}_1$)	1,77	0,17	1,47	1,65	1,76	1,88	2,12	0,0015	1,00
Maassaolo $\times$ (maassaolo $<$ 6) ($\mathcal{G}_1$)	3,26	0,78	2,00	2,71	3,18	3,72	5,05	0,0074	1,00
Maassaolo $\times$ (maassaolo $>$ 9) ($\mathcal{G}_1$)	0,85	0,17	0,57	0,73	0,83	0,95	1,23	0,0015	1,00
Ikä $<$ 30 ($\mathcal{G}_1$)	1,60	0,13	1,35	1,50	1,59	1,68	1,86	0,0013	1,00
(Ikä $<$ 30) $\times$ maassaolo ($\mathcal{G}_1$)	0,86	0,07	0,72	0,81	0,85	0,90	1,01	0,0007	1,00
Ikä $>$ 50 ($\mathcal{G}_1$)	1,28	0,14	1,04	1,19	1,28	1,37	1,57	0,0006	1,00
Muuttovuosi v. 1993–2001 ($\mathcal{G}_1$)	1,22	0,08	1,07	1,17	1,22	1,27	1,38	0,0006	1,00
Korkein koulutus: 2. asteen koulutus ($\mathcal{G}_1$)	1,06	0,07	0,93	1,01	1,05	1,10	1,21	0,0002	1,00
Korkein koulutus: 3. asteen koulutus ($\mathcal{G}_1$)	2,13	0,25	1,70	1,96	2,11	2,29	2,66	0,0004	1,00
Pääasiallinen toiminta: opiskelija ($\mathcal{G}_1$)	1,06	0,05	0,96	1,02	1,06	1,10	1,17	0,0002	1,00
Pääasiallinen toiminta: työllinen ($\mathcal{G}_1$)	2,43	0,16	2,12	2,32	2,43	2,54	2,77	0,0004	1,00
Kotikunta Uudellamaalla ($\mathcal{G}_1$)	1,19	0,04	1,11	1,16	1,18	1,21	1,27	0,0003	1,00
Muutto alle 20-v. ($\mathcal{G}_1$)	1,34	0,09	1,17	1,28	1,34	1,40	1,53	0,0009	1,00
Regressiokertoimet ($\mathcal{G}_2$):									
Ikä ($\mathcal{G}_2$)	1,05	0,03	0,99	1,03	1,05	1,06	1,10	0,0006	1,00
Ikä^2 ($\mathcal{G}_2$)	0,97	0,02	0,93	0,95	0,97	0,98	1,02	0,0005	1,00
Maassaolo (vuosia) ($\mathcal{G}_2$)	1,19	0,07	1,06	1,15	1,19	1,24	1,33	0,0015	1,00
Maassaolo $<$ 6 ($\mathcal{G}_2$)	1,00	0,06	0,88	0,95	1,00	1,04	1,13	0,0007	1,00
Maassaolo $\times$ (maassaolo $<$ 6) ($\mathcal{G}_2$)	1,11	0,15	0,85	1,00	1,10	1,20	1,42	0,0016	1,00
Maassaolo $>$ 9 ($\mathcal{G}_2$)	1,08	0,05	0,99	1,05	1,08	1,11	1,17	0,0003	1,00
Maassaolo $\times$ (maassaolo $>$ 9) ($\mathcal{G}_2$)	0,86	0,08	0,71	0,80	0,85	0,91	1,03	0,0007	1,00

Jatkuu seuraavalla sivulla

Taulukko 5.5 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Ikä < 30 ($\mathcal{G}_2$)	1,01	0,04	0,94	0,99	1,01	1,04	1,10	0,0004	1,00
(Ikä < 30) $\times$ maassaolo ($\mathcal{G}_2$)	0,95	0,04	0,87	0,92	0,95	0,98	1,03	0,0003	1,00
Muuttovuosi v. 1993–2001 ($\mathcal{G}_2$)	1,04	0,03	0,98	1,02	1,04	1,05	1,09	0,0002	1,00
Regressiokertoimet (θ):									
Nainen (θ)	9,39	1,79	6,58	8,11	9,12	10,44	13,66	0,0154	1,00
Ikä (θ)	1,64	0,29	1,14	1,43	1,61	1,82	2,26	0,0072	1,00
Ikä^2 (θ)	1,09	0,13	0,85	0,99	1,08	1,17	1,37	0,0030	1,00
Maassaolo (vuosia) (θ)	0,18	0,04	0,11	0,15	0,17	0,20	0,27	0,0009	1,00
Maassaolo < 6 (θ)	0,77	0,19	0,47	0,64	0,75	0,89	1,21	0,0020	1,00
Maassaolo $\times$ (maassaolo < 6) (θ)	0,25	0,16	0,06	0,14	0,21	0,32	0,64	0,0016	1,00
Maassaolo > 9 (θ)	1,11	0,22	0,74	0,95	1,09	1,25	1,59	0,0016	1,00
Maassaolo $\times$ (maassaolo > 9) (θ)	1,40	0,70	0,50	0,91	1,25	1,71	3,17	0,0066	1,00
Ikä < 30 (θ)	1,23	0,27	0,79	1,04	1,20	1,39	1,83	0,0027	1,00
(Ikä < 30) $\times$ maassaolo (θ)	0,60	0,14	0,37	0,50	0,58	0,68	0,91	0,0016	1,00
Muuttovuosi v. 1993–2001 (θ)	1,95	0,31	1,42	1,73	1,92	2,14	2,63	0,0026	1,00
Muutto alle 20-v. (θ)	1,43	0,24	1,02	1,26	1,41	1,58	1,95	0,0020	1,00
Muotoparametrit:									
Muotoparametri($\mathcal{G}_1$)	0,89	0,01	0,86	0,88	0,89	0,90	0,92	0,0002	1,00
Muotoparametri($\mathcal{G}_2$)	10,4	0,7	9,0	9,9	10,4	10,9	11,9	0,012	1,00

Listaan lyhyesti joitakin huomioita keskeisimmistä vaikutuksista ja estimointituloksista taulukossa 5.5. Kun kirjoitan parametrin sijoittuvan jollekin tietylle intervallille, tarkoitan tällöin kyseisen parametrin posteriorijakauman 95%:n uskottavuusintervallia 2,5 %:sta 97,5 %:iin. Käytän lisäksi vedonlyöntisuhteesta sen englanninkieliseen nimitykseen, odds ratio, pohjautuvaa lyhennettä *OR*.

- **Sukupuoli:**
  - **Hurdle (H)**. Naisindikaattorin OR on noin 3,10–3,50 eli naisilla nollapalkkasumman odds yli 3-kertainen miehiin verrattuna. Yli ykkösen oleva OR tarkoittaa, että nollapalkkasumman todennäköisyys on suurempi kuin vertailuluokassa. Mikäli OR on pienempi kuin yksi, niin todennäköisyys on puolestaan pienempi kuin vertailuluokassa.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$)**. Naisindikaattorin kerroin ensimmäisessä gamma-jakaumassa on noin 0,82–1 miehiin verrattuna. Todennäköisyys sille, että tämä vaikutus on ykköstä pienempi on noin 97,5 %:ia. Ensimmäiseen gamma-jakaumaan kuuluvat palkkasummat syntyvät (luultavasti) henkilöiltä, jotka ovat vain osan vuotta töissä. Naisilla, jotka ovat vain osan vuotta töissä, palkkasummat ovat näin ollen noin 82–100 %:ia miesten palkkasummista.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ)**. Naisindikaattorin OR noin 6,6–13,7 eli naisilla yli nollan olevan palkkasumman odds kuulua ensimmäiseen gamma-jakaumaan (eikä toiseen gamma-jakaumaan) on ”paljon” suurempi kuin miehillä. Yli ykkösen oleva OR tarkoittaa, että ensimmäiseen gamma-jakaumaan kuulumisen todennäköisyys on suurempi kuin vertailuluokassa. Mikäli OR on pienempi kuin yksi, niin todennäköisyys on puolestaan pienempi kuin vertailuluokassa. Mikäli ensimmäinen gamma-jakauma vastaa vain osan vuotta työssä olevien palkkasummia (ja toinen gamma-jakauma koko vuoden töissä olevien), niin tämän tuloksen perusteella naisilla on miehiä suurempi todennäköisyys olla vain osan vuotta töissä verrattuna siihen, että olisi koko vuoden töissä.
- **Ikä:**
  - **Hurdle (H)**. Iän päävaikutuksen ja iän neliön vaikutuksen jakaumat ovat keskittyneet ykkösen ympäristöön. Alle 30-vuotiailla nollapalkkasumman todennäköisyys on pienempi kuin yli 29-vuotiailla. Interaktioista: Alle 30-vuotiailla maassaoloajan lisääntyminen vähentää nollan todennäköisyyttä enemmän kuin yli 29-vuotiailla. Yli 50-vuotiailla iän lisääntyminen kasvattaa nopeasti nollapalkkasumman todennäköisyyttä.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$)**. Iän päävaikutus yli yhden, joten iän lisääntyminen tämän vaikutuksen perusteella nostaa palkkasummia. Toisaalta iän neliön vaikutus aiheuttaa sen, että aluksi palkkasummat nousevat työiässä, kunnes ne laskevat lähestyttäessä eläkeikää.
  - **Toinen gamma-jakauma ($\mathcal{G}_2$)**. Ikään liittyvien vaikutusten jakaumat ovat keskittyneet ykkösen ympäristöön.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ)**. Ikä vaikuttaa lisäävän todennäköisyyttä kuulua ensimmäiseen gamma-jakaumaan. Tämä todennäköisyys on lisäksi alle 30-vuotiailla muita korkeampi. Toisaalta maassaoloaika vähentää todennäköisyyttä alle 30-vuotiailla enemmän kuin muilla.

- **Maassaoloaika:**
  - **Hurdle (H).** Maassaoloajan kasvu pienentää todennäköisyyttä nollapalkkasummalle. Maassaoloaika pienentää tätä todennäköisyyttä enemmän alle kuusi vuotta maassa olleilla kuin enemmän maassa olleilla. Nollan todennäköisyys pienenee siten aluksi jyrkemmin, mutta tasoittuu sittemmin maassaoloajan kasvaessa.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$).** Maassaoloajan kasvu nostaa ensimmäiseen gamma-jakaumaan kuuluvien palkkasummien odotusarvoa. Nousu on suurinta alle kuusi vuotta maassa olleilla. Yli yhdeksän vuotta maassa olleilla maassaoloajan vaikutus on ainakin 75 %:n todennäköisyydellä pienempää kuin vähemmän aikaa maassa olleilla.
  - **Toinen gamma-jakauma ($\mathcal{G}_2$).** Maassaoloajan nousu yhdellä keskihajonnalla (5,7 vuotta) nostaa toiseen gamma-jakaumaan kuuluvien palkkasummien odotusarvoa noin 1,06–1,33-kertaisiksi.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ).** Maassaoloaika vaikuttaa θ-parametriin samoin kuin hurdleen ja 1. ja 2. gamma-jakaumaan: Maassaoloajan nousu pienentää yli nollan olevan palkkasumman todennäköisyyttä kuulua ensimmäiseen gamma-jakaumaan – ja nostaa todennäköisyyttä kuulua toiseen. Todennäköisyys kuulua ensimmäiseen gamma-jakaumaan pienenee nopeiten alle 6 vuotta maassa olleilla. Mikäli ensimmäinen gamma-jakauma vastaa vain osan vuotta työssä olevien palkkasummia (ja toinen gamma-jakauma koko vuoden töissä olevien), niin tämän tuloksen perusteella maassaoloaika kasvattaa todennäköisyyttä olla koko vuosi töissä – verrattuna siihen, että olisi vain osan vuodesta töissä.
- **Maahanmuuttovuosiluokka:**
  - **Hurdle (H).** Vuosina 1993–2001 muuttaneilla on pienempi todennäköisyys saada nollapalkkasumma kuin vertailuluokassa eli vuosina 2002–2014 muuttaneilla – kun muut vaikutukset, eli esimerkiksi maassaoloaika, otetaan huomioon.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$).** Vuosina 1993–2001 muuttaneilla on ensimmäisen gamma-jakauman euromääräinen palkkasumma on noin 1,07–1,38-kertainen kuin vertailuluokassa eli vuosina 2002–2014 muuttaneilla – kun muut vaikutukset, eli esimerkiksi maassaoloaika, otetaan huomioon.
  - **Toinen gamma-jakauma ($\mathcal{G}_2$).** Vuosina 1993–2001 muuttaneilla on toisen gamma-jakauman euromääräinen palkkasumma on suunnilleen samansuuruinen kuin vertailuluokassa eli vuosina 2002–2014 muuttaneilla – kun muut vaikutukset, eli esimerkiksi maassaoloaika, otetaan huomioon.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ).** Vuosina 1993–2001 muuttaneilla todennäköisyys kuulua ensimmäiseen gamma-jakaumaan on suurempi kuin vertailuluokassa eli vuosina 2002–2014 muuttaneilla – kun muut vaikutukset, eli esimerkiksi maassaoloaika, otetaan huomioon. Kaiken kaikkiaan vuosina 1993–2001 muuttaneilla vaikuttaa tulosten perusteella olevan pienempi todennäköisyys saada nollapalkkasummat kuin myöhemmin eli vuosina 2002–2014 muuttaneilla. Myöhemmin muuttaneista kuitenkin suurempi osa koko vuoden töissä kuin aiemmin muuttaneista. Palkkasummat ovat suurempia aiemmin muuttaneilla, kun katsotaan vain osan vuotta töissä olevia. Koko vuoden töissä olevien osalta

palkkasummien suuruudessa ei ole tulosten perusteella ainakaan mitään ”suurta” eroa näiden maahanmuuttovuosiluokkien välillä.

5.4.1.3 Somalia: posterioriennustetarkistukset

Palkkasummien posterioriennusteet on muodostettu yllä esitetyn mallin 3 (eli elinkaariennusteissa käytettävän) parametrien posteriorijakaumien estimaateista ja havaintoaineistosta niin, kuin on esitetty maassaoloaikojen mallinnuksen yhteydessä luvussa 3.2.6.

Posterioriennustetarkistus: Kaikkien palkkasummahavaintojen yhteisjakauma ja henkilöiden palkkasummien keskiarvojen yhteisjakauma.

Aloitan posterioriennustetarkistukset vertaamalla havaintoaineiston kaikkien palkkasummahavaintojen muodostamaa yhteisjakaumaa posterioriennustejakauman kaikkien ennusteiden muodostamaan yhteisjakaumaan. Ensimmäisen mallinnustavoitteen mukaan näiden jakaumien pitäisi vastata mahdollisimman hyvin toisiaan, erityisesti jakaumien tärkeimpien tunnuslukujen kuten keskiarvojen, keskihajontojen ja kvartiilien suhteen. Koska palkkasummat ovat jakautuneet nolliin ja yli nolliin, mallinnustavoitteiden tulisi täyttyä sekä nollien osuuden että ylinollien jakauman osalta.

Kaikkien havaintojen jakauman lisäksi vertaan kaikkien havaintoaineiston henkilöiden palkkasummien keskiarvojen jakaumaa ennustettuihin henkilöiden palkkasummien keskiarvojen jakaumaan. Nämä jakaumat muodostetaan havaintoaineistossa laskemalla palkkasummien keskiarvot kaikille aineiston henkilöille ja muodostamalla näistä jakauma. Posterioriennusteiden kohdalla lasketaan ennustettujen palkkasummien keskiarvot kaikille henkilöille, ja muodostetaan näistä jakauma. Henkilötasoisten jakaumien vertailun tarkoituksena on osoittaa, että elinkaariennusteissa käytettävä nettovaikutusten malli tuottaa luotettavia ennusteita paitsi havaintotasolla niin myös henkilötasolla. Tämä tavoite ei kuitenkaan toteudu käytännössä niin kauan, kuin Somaliassa syntyneiden lopullisena palkkasummamallina ei ole satunnaisvaikutuksia sisältävää mallia. Satunnaisvaikutukset ovat tässä tapauksessa henkilötason satunnaisvaikutuksia mallin neljässä vakiotermissä (vrt. tulokset Suomessa syntyneiden osalta, jossa nämä satunnaisvaikutukset ovat mukana lopullisessa mallissa).

Taulukossa 5.11 on esitetty yllä kuvatun posterioriennustetarkistuksen tulokset. Ylemmässä taulukon osassa verrataan kaikkien palkkasummahavaintojen jakauman tunnuslukujen estimaatteja vastaaviin posterioriennustejakauman estimaatteihin. Bayes-p-arvot ja Bayes-intervallit eri euromäärille ovat testisuureita estimaattien vastaavuuksille havaintoaineiston ja posterioriennusteiden välillä. Bayes-intervalli ilmaisee todennäköisyyden sille, että posterioriennustejakauman estimaatti poikkeaa korkeintaan ilmoitetun euromäärän verran havaintoaineiston estimaatista. Esimerkiksi 2 000 euron Bayes-intervallin sisään mahtuvat kaikkien havaintojen jakauman osalta varsin suurella todennäköisyydellä kaikki jakaumapisteet lukuun ottamatta nollan ympäristöön havaintoaineistossa sijoittuvaa 90. prosenttipistettä. Bayes-p-arvo ja Bayes-intervallit on määritetty tarkemmin työikäisten nettovaikutusmallinnuksen yhteydessä (ks. luku

4.3.7).

Taulukon ylemmästä osasta käy ilmi, että kaikkien havaintojen ja posterioriennusteiden jakaumat vastaavat hyvin toisiaan erityisesti jakaumien havaittujen keskiarvojen ja nollaosuuksien osalta. Ylinollien keskiarvo on posterioriennusteiden jakaumassa noin 350 euroa korkeampi kuin aineistossa. Tämä johtuu siitä, että erityisesti palkkasummien jakauman oikeassa hännässä (vrt. 99 %-piste) posterioriennusteet saavat suurempia arvoja kuin havaintoaineiston jakaumassa. Tämä saa myös aikaan sen, että posterioriennusteiden jakauman keskihajonta on noin 600 euroa suurempi kuin havaintoaineiston jakaumassa.

Posterioriennustejakauman sopivuutta voisi vielä parantaa – paitsi lisäämällä lopulliseen malliin satunnaisvaikutuksia – erityisesti maassaoloajan ja sukupuolen vaikutusten suhteen. Kuten alempana selviää, nyt erityisesti 15 maassaolovuoden jälkeen malli tuottaa keskimäärin liian suuria ylinollien palkkasummaennnusteita verrattuna havaintoaineistoon. Lisäksi erityisesti naisten osalta malli tuottaa suurempia ennusteita kuin havainnot jakauman oikeassa päässä, kuten myös selviää jäljempänä. Toisaalta nyt elinkaariennusteissa käytettävä malli ei ainakaan tuota liian pieniä palkkasummaennusteita suhteessa havaintoaineistoon.

Taulukon 5.11 alemmasta osasta selviää, miksi lopulliseen malliin tulisi vielä sisällyttää ainakin mallin vakiotermien osalta henkilötason satunnaisvaikutukset. Henkilöiden keskiarvojen osalta vain posterioriennusteiden keskiarvo on lähellä havaintoaineiston keskiarvoa. Nollaosuuksista huomaa selvästi, että havaintoaineistossa nollat keskittyvät selvästi tiettyihin henkilöihin syistä, joita mallissa olevat selittävät muuttujat eivät tavoita – toisin sanoen havaintoaineistossa olevasta *havaitsemattomasta heterogeenisuudesta* johtuen. Samasta syystä myös ylinollien keskihajonta on havaintoaineistossa huomattavasti suurempaa kuin henkilöiden keskiarvojen osalta posterioriennusteissa.

Taulukko 5.6: *Palkkasummien havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Kaikkien ennusteiden jakauma ja ennustettujen henkilötasoisten keskiarvojen jakauma. Jakaumien keskiarvot (KA), keskihajonnat (KH), prosenttipisteet:* $1, 5, 10, 25, 50, 75, 90, 95, 99$*, ylinollien keskiarvot (KA(yn)), ylinollien keskihajonnat (KH(yn)) ja nollien osuudet prosentteina (*$0-\%$*). Vertailtavien estimaattien Bayes-p-arvot ja Bayes-intervallit eri euromäärillä.*

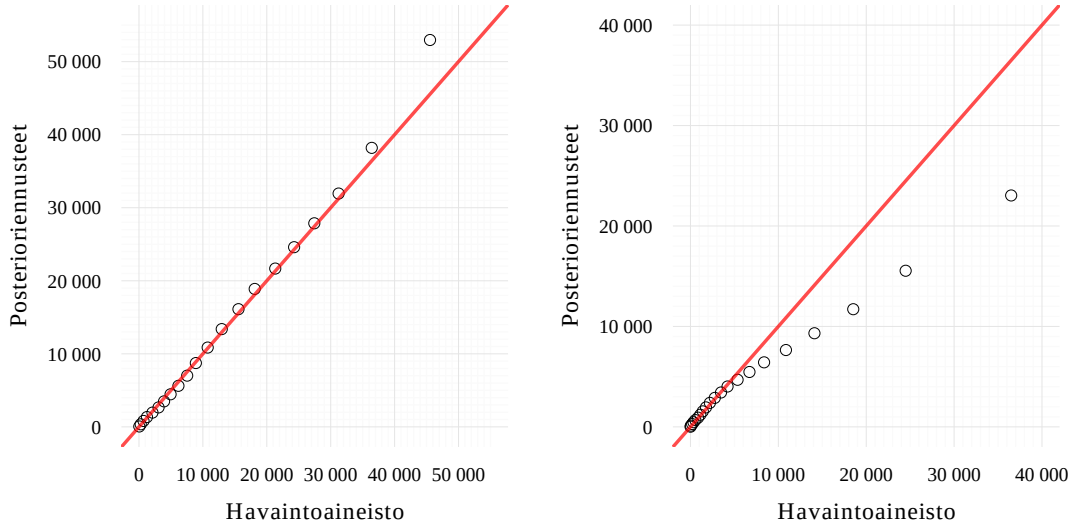
Kaikki havainnot ja ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %	$KA(yn)$	$KH(yn)$	$0-\%$
Posterioriennusteet	3 980	9 540	0	0	0	0	0	1 420	16 800	26 580	41 600	13 430	13 430	70,4
Aineiston havainnot	3 880	8 940	0	0	0	0	0	1 350	16 210	26 210	39 140	13 080	12 220	70,4
Bayes-p-arvo	0,91	1	0	0	0	0	0	0,71	0,88	0,79	1	0,91	1	0,48
Bayes-intervalli: 1 000 euroa	1	0,97	1	1	1	1	1	1	0,78	0,92	0,02	1	0,26	–
Bayes-intervalli: 500 euroa	1	0,30	1	1	1	1	1	1	0,43	0,60	0	0,69	0,01	–
Bayes-intervalli: 200 euroa	0,90	0,02	1	1	1	1	1	0,85	0,14	0,26	0	0,25	0	–

Henkilöiden keskiarvot	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %	$KA(yn)$	$KH(yn)$	$0-\%$
Posterioriennusteet	2 940	4 720	0	0	0	0	720	4 080	9 010	13 080	20 960	4 650	5 230	36,8
Aineiston henkilöt	2 720	6 290	0	0	0	0	0	1 900	9 050	17 230	30 190	6 310	8 330	57,0
Bayes-p-arvo	1	0	0	0	0	0	1	1	0,43	0	0	0	0	0
Bayes-intervalli: 1 000 euroa	1	0	1	1	1	1	1	0	1	0	0	0	0	–
Bayes-intervalli: 500 euroa	1	0	1	1	1	1	0	0	0,96	0	0	0	0	–
Bayes-intervalli: 200 euroa	0,36	0	1	1	1	1	0	0	0,59	0	0	0	0	–

Vertaan alla kuviossa 5.3 havaintoaineiston jakaumien prosenttipisteitä vastaavien posterioriennustejakaumien prosenttipisteisiin. Jakaumissa ovat mukana vain yli nollan olevan havainnot sekä ennusteet, ”ylinollat”. Koska nollien osuuksien vastaavuus havaintoaineiston ja posterioriennusteiden välillä on jo käsitelty yllä taulukossa 5.11, en käsittele nollaosuuksia tarkemmin enää. Ylinollien jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeamat vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).

Kaikkien havaintojen ja ennusteiden ylinollien jakaumat vastaavat toisiaan hyvin lukuun ottamatta posterioriennusteiden ylinollien jakauman oikeaa häntää, jossa ennusteet ovat suurempia kuin havaintoaineiston arvot (vrt. 99. prosenttipiste). Kuten jo edellä kirjoitin, henkilöiden keskiarvojen osalta posterioriennusteiden (ylinollien) jakauma vastaa huonosti havaintoaineiston jakaumaan (niiden keskiarvoa lukuun ottamatta). Tämä johtuu siitä, että Somaliassa syntyneiden lopullisessa mallissa ei ole toistaiseksi sellaisia satunnaisvaikutuksia, jotka ottaisivat mallissa sen huomioon, että korkeat (ja matalat) palkkasummat ryvästyvät henkilöittäin. Seurauksena tästä on, että elinkaariennusteiden ennustetuissa palkkasummissa on liian vähän henkilöiden välistä vaihtelua.



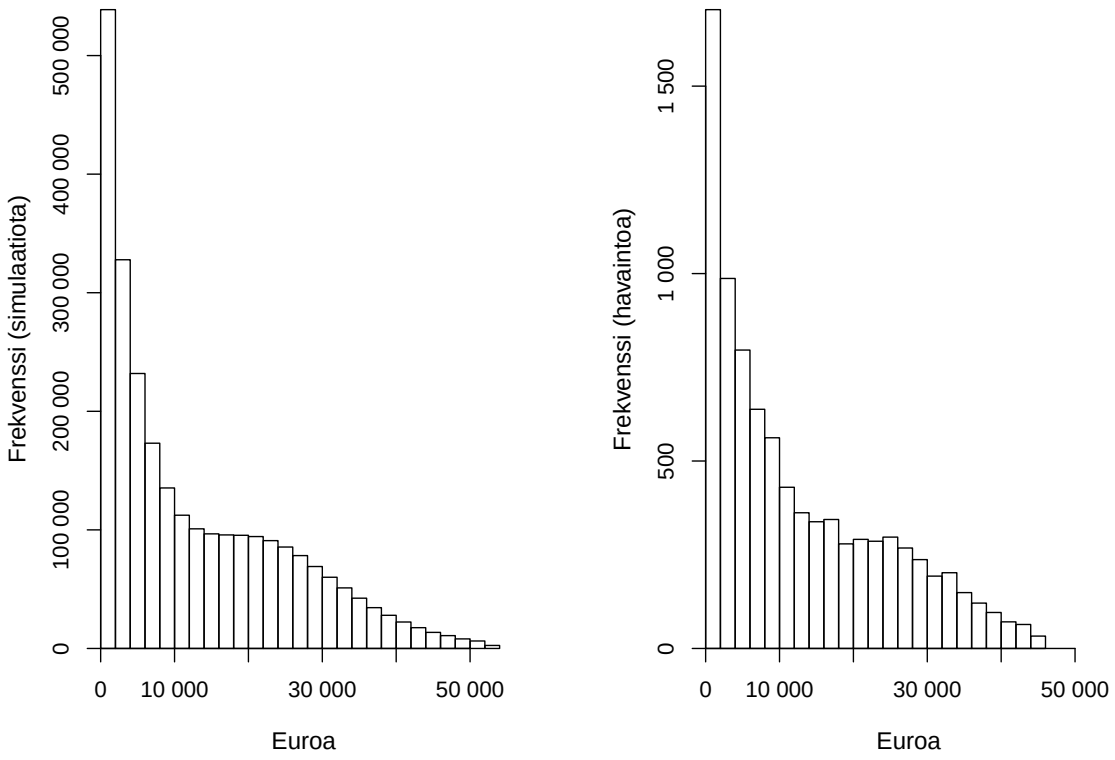
(a) *Kaikki havainnot ja ennusteet.*

(b) *Henkilöiden keskiarvot.*

Kuva 5.3: *Havaintoaineiston yli nollan olevien havaintojen jakauman ja posterioriennusteiden yli nollan olevien ennusteiden jakauman vertailu. "Ylinollien" jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

Verrataan vielä posterioriennusteiden ja havaintoaineiston ylinollien jakaumia prosenttipisteiden vertailukuvioiden (Q-Q plot) lisäksi histogrammien avulla. Nämä on esitetty kuvassa 5.4. Vasemmalla oleva histogrammi esittää palkkasummien yli nollan olevien posterioriennusteiden jakaumaa. Oikealla oleva histogrammi esittää puolestaan palkkasummien yli nollan olevien havaintojen jakaumaa. Kummastakin jakaumasta on katkaistu pois ylin prosentti yksityisyydensuojan varmistamiseksi.

Myös ylinollia koskevat histogrammit osoittavat, että kaikkien havaintojen ja posterioriennusteiden jakaumat muistuttavat toisiaan – lähinnä lukuun ottamatta posterioriennusteiden oikeaa häntää, joka histogrammikuviossakin ylettyy enemmän oikealle kuin havaintoaineistojen oikea häntä. Histogrammikuviosta näkyy myös, miten palkkasummien ylinollia koskevan mallin toinen gamma-jakauma ottaa huomioon palkkasummien paikallisen huipun noin 20 000–25 000 euron kohdalla.



Kuva 5.4: *Posterioriennustetarkistus histogrammeilla. Posterioriennusteiden yli nollan olevien ennusteiden jakauma ja havaintoaineiston yli nollan olevien havaintojen jakauma. Posterioriennusteiden ylinollien jakauma vasemmalla ja havaintoaineiston ylinollien jakauma oikealla. Kummastakin jakaumasta on katkaistu pois ylin prosentti yksityisyydensuojan varmistamiseksi.*

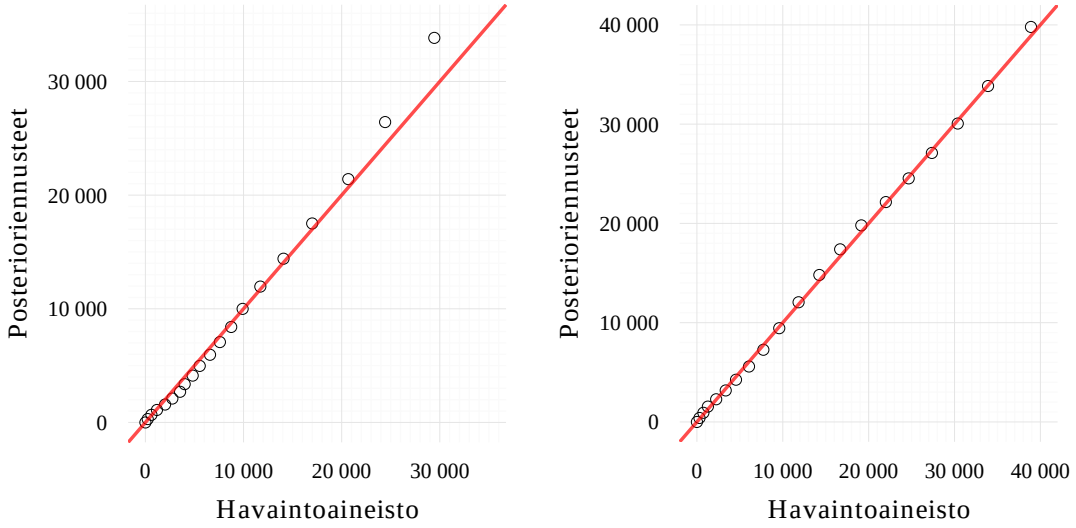
Posterioriennustarkistukset selittävien muuttujien luokkien mukaan

Vertaan nyt posterioriennusteita havaintoaineistoon tärkeimpien selittävien muuttujien luokkien mukaan. Tärkeimpiä selittäviä muuttujia ovat sukupuoli sekä aikaan liittyvät muuttujat: ikä, maahanmuuttovuosi, maassaoloaika ja kalenterivuosi. Viimeinen näistä muuttujista ei ollut lainkaan mallissa, koska maahanmuuttovuosi, maassaoloaika ja kalenterivuosi eivät voi olla samaan aikaan lineaarisessa mallissa. Esitän posterioriennustetarkistukset myös maahanmuuttoiän mukaan sekä iän mukaan niille, joilla on kulunut yli 15 vuotta ensimmäisestä muutosta Suomeen.

Vertaan posterioriennusteiden nollaosuuksia sekä ”ylinollien” keskiarvoja selittävien luokkien mukaan vastaaviin havaintoaineiston nollaosuuksiin ja ”ylinollien” keskiarvoihin. Koska palkkasummat koostu-

vat nollista sekä yli nollan olevasta osasta, esitän nollaosuudet ja ylinollien keskiarvot erikseen. Esimerkiksi iän kohdalla vertaan posterioriennusteiden nollaosuuksia havaintoaineiston nollaosuuksiin niin, että lasken sekä posterioriennusteista että havaintoaineistosta jokaiselle iälle erikseen nollien osuuden 18-vuotiaista 62-vuotiaisiin. Ylinollien osalta lasken puolestaan erikseen jokaiselle iälle havaintojen ja ennusteiden ylinollien keskiarvot.

Ennen kuin vertaan nollaosuuksia ja ylinollien keskiarvoja selittävien muuttujien mukaan, esitän vielä ylinollien jakaumien tarkistukset erikseen sekä naisille että miehille. Kuviossa 5.5 on esitetty ylinollien jakaumien prosenttipisteet sukupuolen mukaan. Kuviosta selviää, että miehillä (oikeanpuoleinen osakuva) posterioriennusteiden ylinollien jakauma vastaa hyvin havaintoaineiston ylinollien jakaumaa – myös jakauman oikean hännän osalta. Naisten kohdalla (vasemmanpuoleinen osakuva) posterioriennusteiden ylinollien jakauma poikkeaa havaintoaineiston ylinollien jakaumasta posterioriennusteiden ja havaintojen jakauman oikean hännän osalta – samoin kuin kaikkien posterioriennusteiden ja havaintojen jakaumissa (ks. edelllä kuva 5.3). Tämä viittaa siihen, että posterioriennusteiden ylinollien jakauman oikea häntä poikkeaa havaintoaineiston jakauman hännästä nimenomaan naisia koskevan kehnommin aineistoon sopivan posterioriennustejakauman vuoksi.

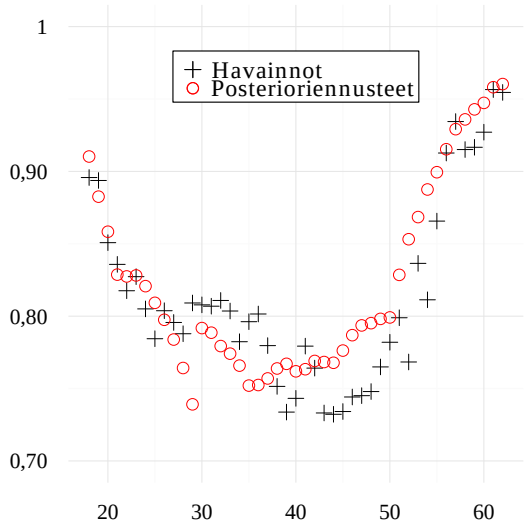


(a) *Naiset.*

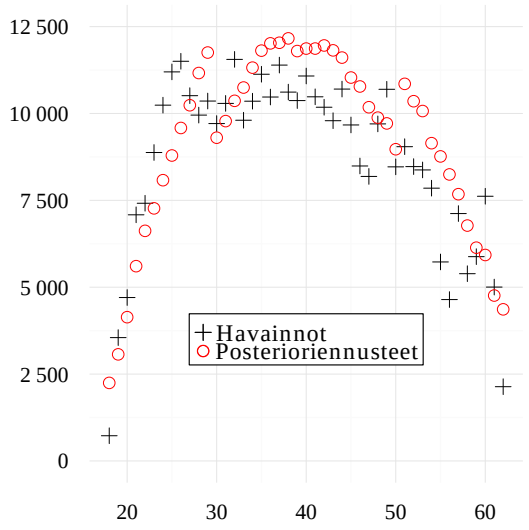
(b) *Miehet.*

Kuva 5.5: *Havaintoaineiston yli nollan olevien havaintojen jakauman ja posterioriennusteiden yli nollan olevien ennusteiden jakauman vertailu sukupuolen mukaan. "Ylinollien" jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

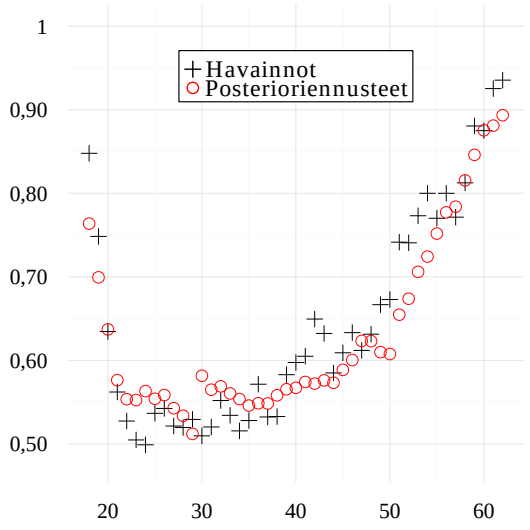
Posterioriennustetarkistus iän ja sukupuolen mukaan.



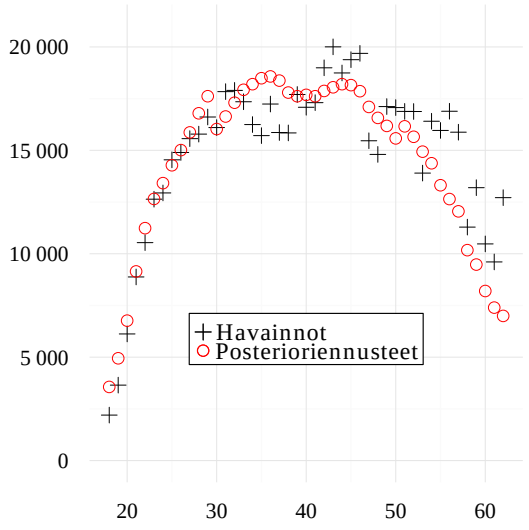
(a) *Nollien osuudet iän mukaan. Naiset.*



(b) *Ylinollien keskiarvot iän mukaan. Naiset.*



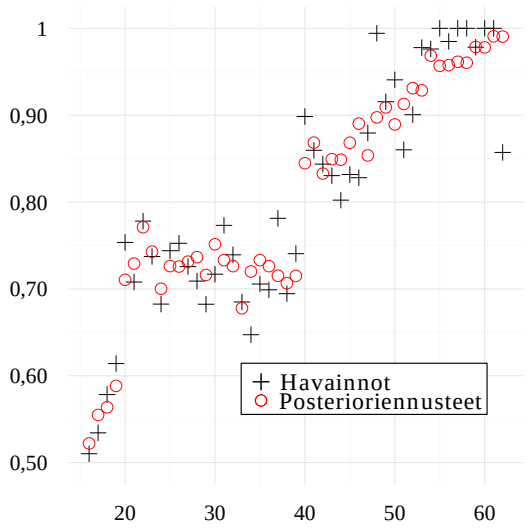
(c) *Nollien osuudet iän mukaan. Miehet.*



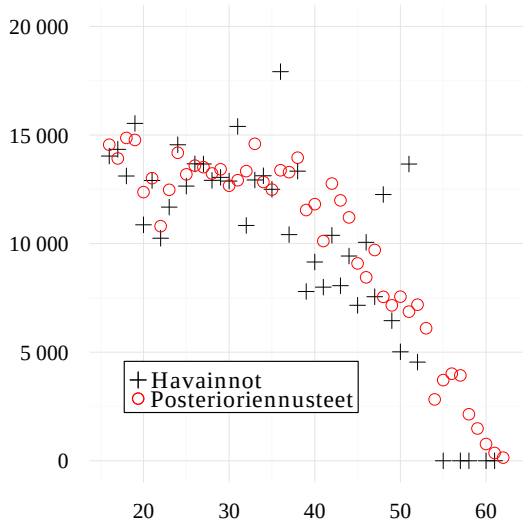
(d) *Ylinollien keskiarvot iän mukaan. Miehet.*

Kuva 5.6: Havainnot ja posterioriennusteet iän ja sukupuolen mukaan.

Posteriorienustetarkistus maahanmuuttoiän ja maassaoloajan mukaan.

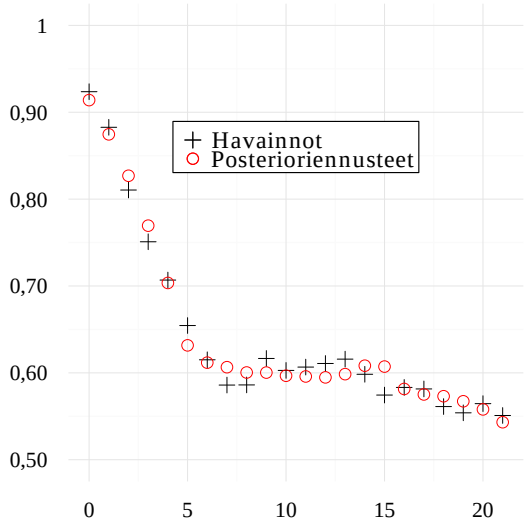


(a) *Nollien osuudet maahanmuuttoiän mukaan.*

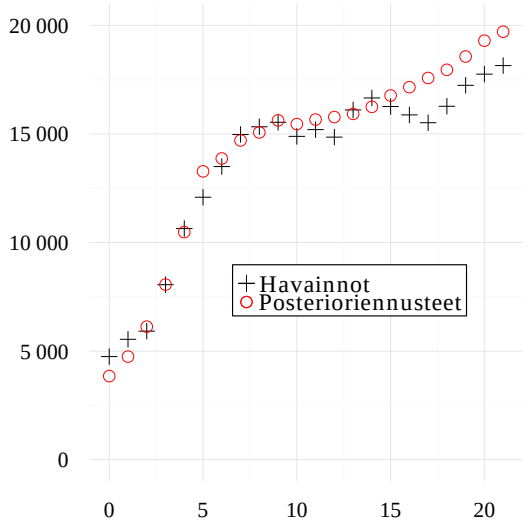


(b) *Ylinollien keskiarvot maahanmuuttoiän mukaan. Kuviosta on poistettu viisi havaintoa yksilönsuojan varmistamiseksi.*

Kuva 5.7: Havainnot ja posteriorienusteet maahanmuuttoiän mukaan.



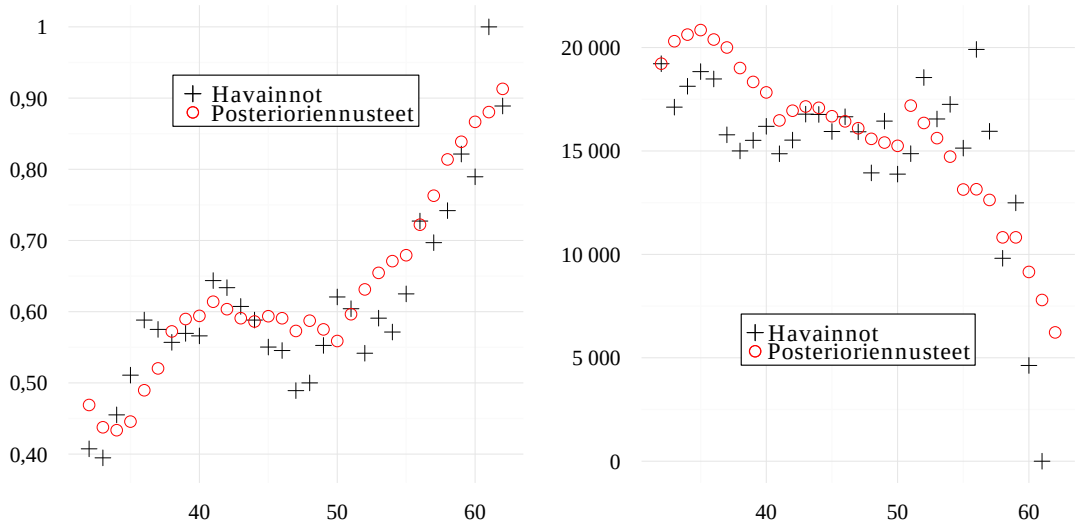
(a) *Nollien osuudet maassaoloajan mukaan.*



(b) *Ylinollien keskiarvot maassaoloajan mukaan.*

Kuva 5.8: Havainnot ja posteriorienusteet maassaoloajan mukaan.

Posterioriennustetarkistus iän mukaan. Yli 15 vuotta maassa olleet.

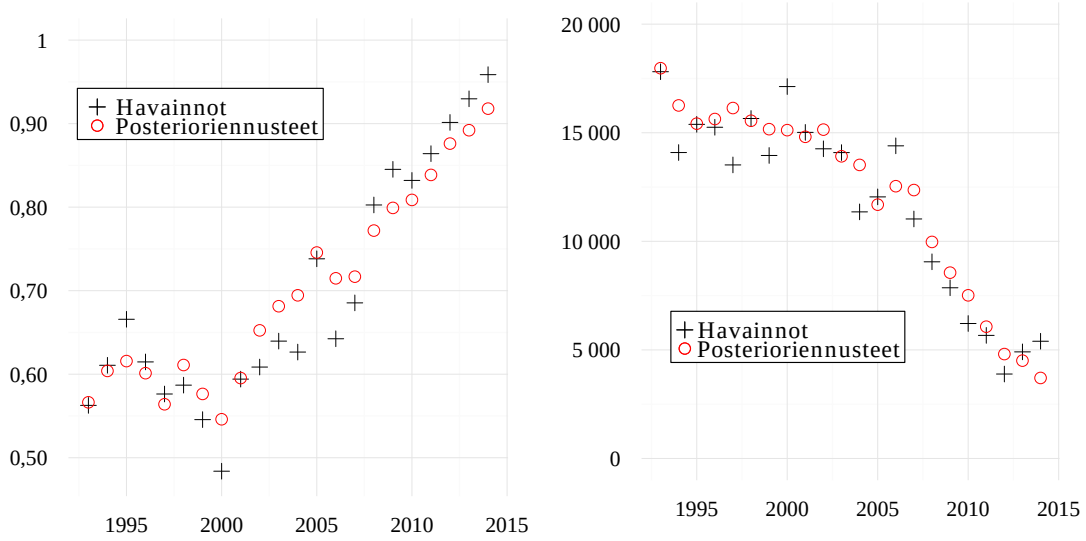


(a) *Nollien osuudet iän mukaan, yli 15 vuotta maassa olleet.*

(b) *Ylinollien keskiarvot iän mukaan, yli 15 vuotta maassa olleet.*

Kuva 5.9: Havainnot ja posterioriennusteet iän mukaan. Yli 15 maassa vuotta olleet.

Posterioriennustetarkistus maahanmuuttovuoden mukaan.

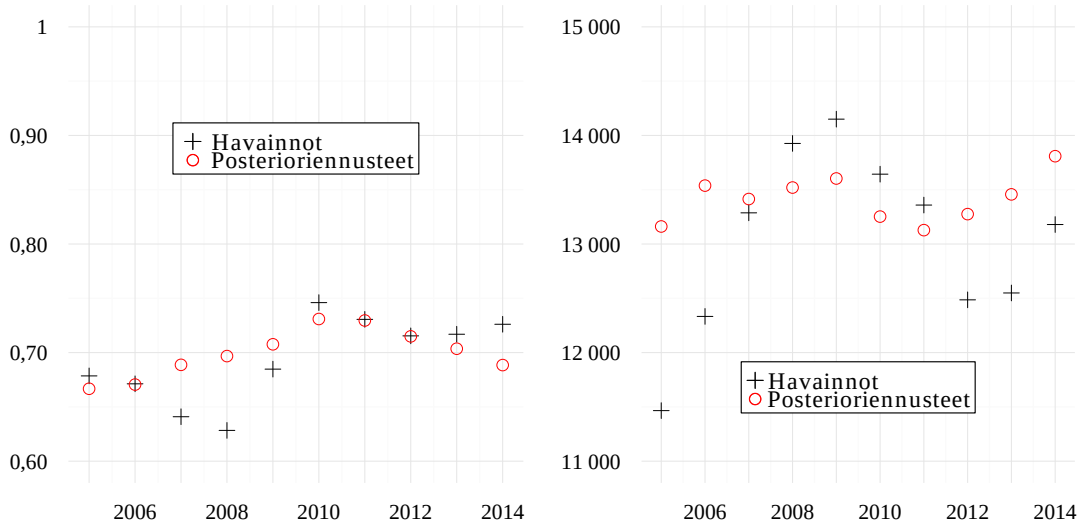


(a) *Nollien osuudet maahanmuuttovuoden mukaan.*

(b) *Ylinollien keskiarvot maahanmuuttovuoden mukaan.*

Kuva 5.10: Havainnot ja posterioriennusteet maahanmuuttovuoden mukaan.

Posterioriennustetarkistus kalenterivuoden mukaan.



(a) *Nollien osuudet kalenterivuoden mukaan.* (b) *Ylinollien keskiarvot kalenterivuoden mukaan.*

Kuva 5.11: Havainnot ja posterioriennusteet kalenterivuoden mukaan.

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan. Havaintoja kuvista 5.6–5.11.

Posterioriennusteiden nollaosuudet ja ylinollien keskiarvot selittävien muuttujien luokkien mukaan vastaavat hyvin havaintoaineiston vastaavia arvoja kuvissa 5.6–5.11. Ehkä suurin systemaattinen virhe posterioriennusteiden ja havaintoaineiston välillä on kuvion 5.8 oikeanpuoleisessa osakuvassa: Posterioriennusteiden ylinollien keskiarvot ovat 15 vuoden maassaoloajan jälkeen noin 2 000 euroa liian suuria verrattuna havaintoaineiston ylinollien keskiarvoihin.

5.4.2 Irak

5.4.2.1 Irak: havaintoaineisto

Taulukossa 5.7 on esitetty Irakissa syntyneiden 18–62-vuotiaiden palkkasummahavaintojen lukumäärä ja havaintojen jakauman keskeisimmät tunnusluvut. Havainnot koskevat vuosia 2005–2014. Taulukossa on esitetty samat tiedot myös havaintoaineiston henkilöiden nettovaikutushavaintojen keskiarvoille. Mallinnusaineistoon ei sisälly alle 16-vuotiaana muuttaneita eikä ennen vuotta 1992 muuttaneita. Ennen vuotta 1992 – ja varsinkin 1980-luvulla muuttaneiden Irakissa syntyneiden työllisyyden kehitys poikkeaa paljon myöhemmin muuttaneita. Tämä johtuu oletettavasti siitä, että suhteellisen harvat Irakissa syntyneet, jotka asuivat Suomessa ennen vuotta 1990, olivat muuttaneet muista kuin humanitaarisista syistä – toisin kuin vuoden 1991 jälkeen. Palkkasummahavainnot muodostavat selitettävän muuttujan palkkasummien

mallinnuksessa. Taulukossa 5.7 on esitetty kaikkien palkkasummahavaintojen jakaumatietojen lisäksi vielä erikseen niiden palkkasummahavaintojen, jotka ovat yli nollan, jakauman keskiarvo ja keskihajonta. Lisäksi taulukon alimmalla rivillä on esitetty nollien osuudet havainnoista.

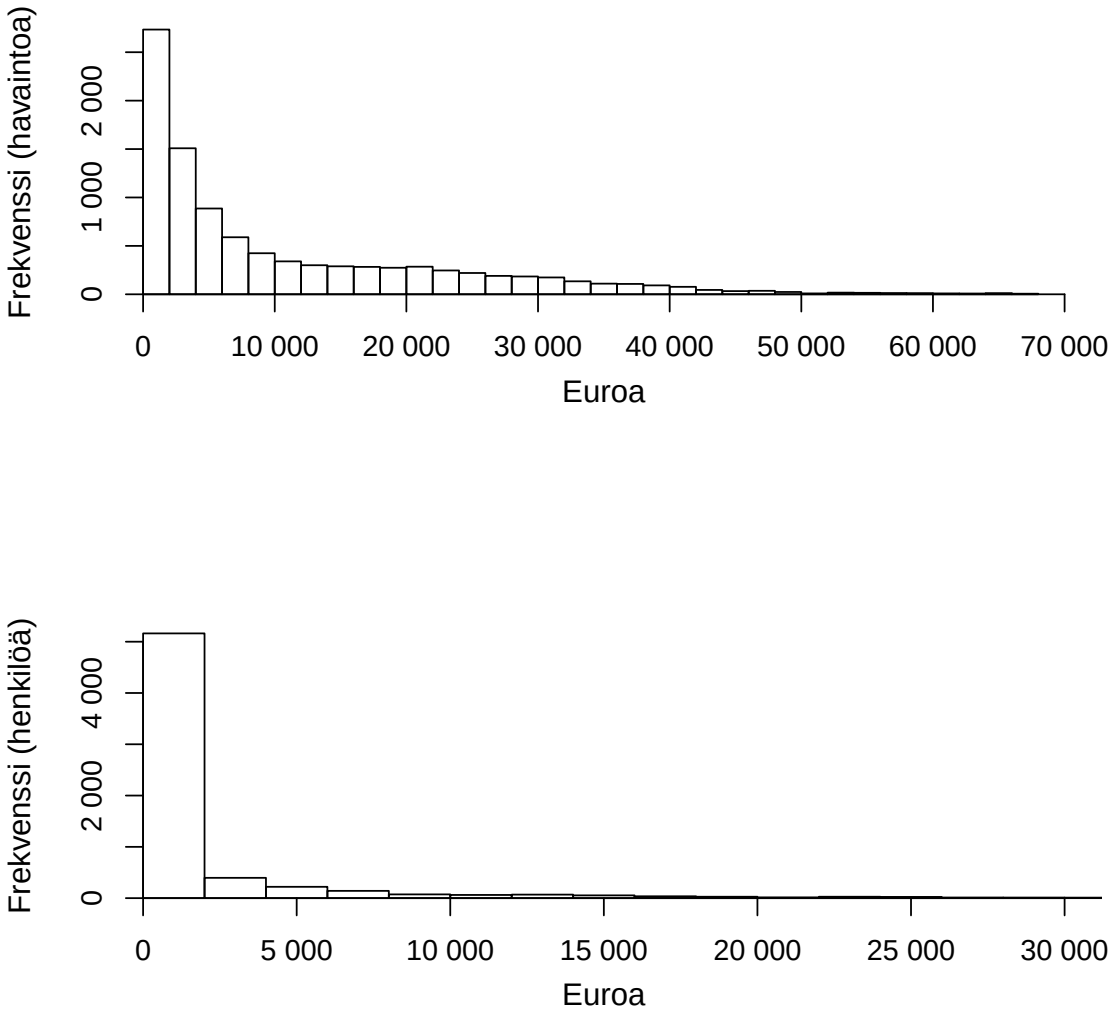
Irakissa syntyneiden palkkasummahavaintojen mallinnusaineisto koostuu 37 168 henkilövuositasoisesta palkkasummahavainnosta, jotka on poimittu paneeliaineistona/pitkittäisaineistona 6 406 henkilöltä. Taulukon 5.7 tietojen mukaan esimerkiksi kaikkien palkkasummahavaintojen keskiarvo on 3 070 euroa, keskihajonta on 9 190 euroa ja viisi prosenttia kaikista havainnoista on 22 100 euroa suurempia. Havaintoaineiston henkilöiden palkkasummahavaintojen keskiarvojen keskiarvo on 2 150 euroa ja keskihajonta 6 400 euroa. Yli nollien keskiarvo kaikkien havaintojen joukossa on 11 660 euroa, ja henkilöiden yli nollien keskiaro on puolestaan 5 290 euroa. Nollien osuus koko havaintoaineistossa on 74 prosenttia.

Taulukko 5.7: *Irakissa syntyneet 18–62-vuotiaat vuosina 2005–2014. Palkkasummahavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa sekä henkilöiden lukumäärä ja henkilötasoisten palkkasummahavaintojen keskiarvojen jakauman tunnusluvut mallinnusaineistossa. Palkkasummat on muunnettu vuoden 2011 ansiotasoon. Havaintojen osalta "ylinollat" käsittävät vain ne havainnot, joissa palkkasumma on yli nolla euroa; henkilöiden osalta ne henkilöt, joiden keskiarvo on yli nollan. Nollien osuudet kertovat nollien osuudet kaikkien havaintojen sekä kaikkien henkilöiden keskiarvojen joukoissa. Euromääräiset tunnusluvut pyöristetty lähimpään kymmeneen euroon.*

	Havainnot	Henkilöt
Kokonaislukumäärä (N)	37 168	6 406
Keskiarvo	3 070	2 150
Keskihajonta	9 190	6 400
1. prosenttipiste (1 %)	0	0
5. prosenttipiste (5 %)	0	0
10. prosenttipiste (10 %)	0	0
Alakvartiili (25 %)	0	0
Mediaani (50 %)	0	0
Yläkvartiili (75 %)	280	1 070
90. prosenttipiste (90 %)	9 530	5 840
95. prosenttipiste (95 %)	22 100	12 740
99. prosenttipiste (99 %)	40 870	31 810
Ylinollat: keskiarvo	11 660	5 290
Ylinollat: keskihajonta	15 860	9 180
Nollien osuus	0,74	0,59

Havaintoaineistossa noin 26 %:ia havainnoista on yli nollan. Kuvassa 5.12 on esitetty näiden yli nollien olevien palkkasummien jakauman histogrammi. Jakaumasta on katkaistu pois ylin prosentti. Histogrammista näkee, että havaintojen suurin tiheys (frekvenssi) on lähellä nollaa ja laskee oikealle päin mentäessä. Noin 20 000 euron kohdalla yli nollien jakaumassa on kuitenkin paikallinen huippu. Mallinnuksessa toi-

sen gamma-jakauman moodi sijoittuu suunnilleen tähän kohtaan. Tämä paikallinen huippu on helpommin nähtävissä Suomessa syntyneiden vastaavassa histogrammissa (ks. histogrammi, s. 314).



Kuva 5.12: *Irakissa syntyneet 18–62-vuotiaat. Yli nollan olevien palkkasummahavaintojen katkaistun jakauman histogrammi mallinnusaineistossa (yllä). Kaikkien aineiston henkilöiden palkkasummien keskiarvojen katkaistun jakauman histogrammi (alla). Jakaumista on poistettu ylin prosentti yksityisyydensuojan varmistamiseksi.*

Taulukossa 5.8 on esitetty Irakissa syntyneiden palkkasummamallien selittävien muuttujien jakaumien

keskiarvot, keskihajonnat ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden kaikkien havaintojen joukossa. Edellä parametrien selitetaulukossa (s. 262) on lyhyitä selityksiä selittävistä muuttujista.

Taulukko 5.8: *Irakissa syntyneet 18–62-vuotiaat. Palkkasummamallien selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo, ja KH on keskihajonta).*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,36	0,48	0	0	1
Ikä	35,7	10,4	27	34	43
Ikä^2	1383	798	729	1156	1849
Maassaolo (vuosia)	6,8	5,7	2	5	11
Maassaolo < 6	0,51	0,50	0	1	1
Maassaolo $\times$ (maassaolo < 6)	1,08	1,59	0	0	2
Maassaolo $\times$ (maassaolo < 6) $\times$ nainen	0,36	1,06	0	0	0
Maassaolo < 8	0,60	0,49	0	1	1
Maassaolo $\times$ (maassaolo < 8)	1,63	2,14	0	0	3
Maassaolo > 12	0,20	0,40	0	0	0
Maassaolo $\times$ (maassaolo > 12)	3,13	6,38	0	0	0
Ikä < 21	0,04	0,21	0	0	0
Ikä < 30	0,33	0,47	0	0	1
Ikä $< 30 \times$ ikä	8,17	11,72	0	0	22
$36 <$ ikä < 59	0,42	0,49	0	0	1
($36 <$ ikä < 59) $\times$ nainen	0,16	0,37	0	0	0
($36 <$ ikä < 59) $\times$ nainen $\times$ ikä	7,28	16,73	0	0	0
$39 <$ ikä < 56	0,31	0,46	0	0	1
($39 <$ ikä < 56) $\times$ nainen	0,12	0,32	0	0	0
($39 <$ ikä < 56) $\times$ nainen $\times$ ikä	5,49	15,00	0	0	0
Ikä > 55	0,04	0,20	0	0	0
(Ikä > 55) $\times$ ikä	2,52	11,88	0	0	0
Muuttovuosi v. 1992–1999	0,31	0,46	0	0	1
Muuttovuosi v. 2000–2005	0,23	0,42	0	0	0
Muuttovuosi v. 2006–2014	0,46	0,50	0	0	1
Korkein koulutus: 2. asteen koulutus	0,10	0,29	0	0	0
Korkein koulutus: 3. asteen koulutus	0,08	0,26	0	0	0
Korkein koulutus: ei 2.astetta/ei tietoa	0,83	0,38	1	1	1
Pääasiallinen toiminta: opiskelija	0,10	0,31	0	0	0
Pääasiallinen toiminta: työllinen	0,04	0,19	0	0	0
Pääasiallinen toiminta: muu	0,86	0,35	1	1	1
Kotikunta Uudellamaalla	0,34	0,47	0	0	1

Jatkuu seuraavalla sivulla

Taulukko 5.8 – jatkuu edelliseltä sivulta

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Muutto alle 20-v.	0,14	0,35	0	0	0

5.4.2.2 Irak: estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Taulukossa 5.9 on esitetty perättäisten palkkasummamallien parametrien odotusarvoestimaatit Irakissa syntyneille 18–62-vuotiaille. Olen käsitellyt perättäisten mallien rakentamista mallinnusaskel kerrallaan alaluvussa 5.3.1. Estimoitavien parametrien selitteet on puolestaan esitetty parametrien selitetaulukossa s. 262. Selitetaulukossa on myös selvennetty lyhyesti, mitä tarkoittaa, että parametri liittyy joko hurdleen (H), ensimmäiseen gamma-jakaumaan ($\mathcal{G}_1$), toiseen gamma-jakaumaan ($\mathcal{G}_2$) tai θ-indikaattoriin, joka kertoo, kuuluuko havainto ensimmäiseen gamma-jakaumaan (eikä toiseen) kahden gamma-jakauman sekoitusjakaumassa. Olen käsitellyt Somaliaa koskevien estimointitulosten yhteydessä (edellä luvussa 5.4.1.2) jo palkkasummamallien parametrien tulkintaa, joten en toista sitä tässä yhteydessä.

Taulukko 5.9: *Perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit, mallikohtaiset ennustetarkkuuden (Waic ja LOO) estimaatit ja keskihajonnat. Mallien valinta on esitetty alaluvussa 5.3.1. H:n ja θ:n vakioparametrit ovat oddseja, regressiokertoimet vedonlyöntisuhteita (engl. odds ratio). Gamma-jakaumien ($\mathcal{G}_1$ ja $\mathcal{G}_2$) vakioparametrit ovat euroja, regressiokertoimet odotusarvon lineaarisia kertoimia. Indikaattorimuuttujien osalta regressiokertoimet ovat muutoksia suhteessa vertailuryhmään, ja jatkuvien muuttujien osalta regressiokertoimet ovat muutoksia, kun muuttujan arvo kasvaa yhdellä keskihajonnalla.*

	Malli 1	Malli 2	Malli 3
Vakioparametrit:			
Vakio (H)	0,68	2,80	1,27
Vakio ($\mathcal{G}_1$)	11 660	10 490	12 210
Vakio ($\mathcal{G}_2$)		29 740	24 850
Vakio (θ)		15,71	3,91
Hurdle-regressiokertoimet (H):			
Nainen (H)			3,10
Ikä (H)			1,33
Ikä^2 (H)			1,09
Maassaolo (vuosia) (H)			0,85

Jatkuu seuraavalla sivulla

Taulukko 5.9 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3
Maassaolo $<$ 6 (H)			0,58
Maassaolo $\times$ (maassaolo $<$ 6) (H)			0,03
Maassaolo $\times$ (maassaolo $<$ 6) $\times$ nainen (H)			1,75
Ikä $<$ 21 (H)			1,41
Ikä $<$ 30 (H)			0,88
Ikä $<$ 30 $\times$ ikä (H)			0,92
(39 $<$ ikä $<$ 56) $\times$ nainen (H)			0,36
(39 $<$ ikä $<$ 56) $\times$ nainen $\times$ ikä (H)			1,18
Ikä $>$ 55 (H)			0,07
Ikä $>$ 55 $\times$ ikä (H)			6,83
Muuttovuosi v. 1992–1999 (H)			0,78
Muuttovuosi v. 2000–2005 (H)			0,69
Korkein koulutus: 2. asteen koulutus (H)			0,85
Korkein koulutus: 3. asteen koulutus (H)			0,55
Pääasiallinen toiminta: opiskelija (H)			0,72
Pääasiallinen toiminta: työllinen (H)			0,18
Kotikunta Uudellamaalla (H)			0,82
Muutto alle 20-v. (H)			0,77
1. gamma-jakauman regressiokertoimet ($\mathcal{G}_1$):			
Nainen ($\mathcal{G}_1$)			1,31
Ikä ($\mathcal{G}_1$)			1,59
Ikä^2 ($\mathcal{G}_1$)			0,73
Maassaolo (vuosia) ($\mathcal{G}_1$)			1,04
Maassaolo $<$ 6 ($\mathcal{G}_1$)			1,27
Maassaolo $\times$ maassaolo $<$ 6 ($\mathcal{G}_1$)			3,89
(39 $<$ ikä $<$ 56) $\times$ nainen ($\mathcal{G}_1$)			1,48
(39 $<$ ikä $<$ 56) $\times$ nainen $\times$ ikä ($\mathcal{G}_1$)			0,68
Ikä $>$ 55 ($\mathcal{G}_1$)			61,45
Ikä $>$ 55 $\times$ ikä ($\mathcal{G}_1$)			0,23
Muuttovuosi v. 1992–1999 ($\mathcal{G}_1$)			1,42
Muuttovuosi v. 2000–2005 ($\mathcal{G}_1$)			1,34
Korkein koulutus: 2. asteen koulutus ($\mathcal{G}_1$)			1,05
Korkein koulutus: 3. asteen koulutus ($\mathcal{G}_1$)			2,40
Pääasiallinen toiminta: opiskelija ($\mathcal{G}_1$)			1,71
Pääasiallinen toiminta: työllinen ($\mathcal{G}_1$)			2,65
Kotikunta Uudellamaalla ($\mathcal{G}_1$)			1,27
Muutto alle 20-v. ($\mathcal{G}_1$)			1,27
2. gamma-jakauman regressiokertoimet ($\mathcal{G}_2$):			
Nainen ($\mathcal{G}_2$)			0,71

Jatkuu seuraavalla sivulla

Taulukko 5.9 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3
Ikä ($\mathcal{G}_2$)			0,97
Ikä^2 ($\mathcal{G}_2$)			1,18
Maassaolo (vuosia) ($\mathcal{G}_2$)			1,29
Maassaolo $<$ 8 ($\mathcal{G}_2$)			0,96
Maassaolo $\times$ (maassaolo $<$ 8) ($\mathcal{G}_2$)			1,61
Maassaolo $>$ 12 ($\mathcal{G}_2$)			1,20
Maassaolo $\times$ (maassaolo $>$ 12) ($\mathcal{G}_2$)			0,98
Muuttovuosi v. 1992–1999 ($\mathcal{G}_2$)			1,02
Kuuluminen 1. gamma-jakaumaan, regressiokertoimet (θ):			
Nainen (θ)			0,88
Ikä^2 (θ)			1,79
Maassaolo (vuosia) (θ)			0,40
Maassaolo $<$ 8 (θ)			0,79
Muuttovuosi v. 1992–1999 (θ)			1,61
Muuttovuosi v. 2000–2005 (θ)			0,92
Muotoparametri($\mathcal{G}_1$)	0,68	0,67	0,80
Muotoparametri($\mathcal{G}_2$)		29,0	12,0
Gamma-jakaumien muotoparametrit:			
Muotoparametri($\mathcal{G}_1$)	0,68	0,67	0,80
Muotoparametri($\mathcal{G}_2$)		29,0	12,0
Mallien ennustetarkkuudet Waic ja LOO:			
Waic (estimaatti)	83 681	83 521	76 296
Waic (estimaatin keskivirhe)	(588)	(587)	(548)
LOO (estimaatti)	83 681	83 521	76 296
LOO (estimaatin keskivirhe)	(588)	(587)	(548)

Taulukon 5.9 ennustemittarien Waic ja LOO mukaan isommat mallit ovat parempia ennustajia kuin pienemmät mallit. Tosin ero mallin 1 ja mallin 2 välillä ei ole suuri, mikä johtuu siitä, että Irakissa syntyneiden yli nollien havaintoaineistoa voisi kuvata melko hyvin vain yhdellä gamma-jakaumalla kahden gamma-jakauman sekoitusjakauman sijaan, sillä ”koko vuoden töissä olleiden” määrä on suhteellisen pieni, mikä näkyy jakauman paikallisen huipun pienuutena noin 20 000 euron kohdalla (ks. jakauman histogrammi, kuva 5.12).

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” mallin on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän uusien maahanmuuttajien elinkaariennusteissa palkkasummien ennustamiseen Irakissa syntyneiden osalta. Irakin osalta lopullinen mallin on malli 3, joka sisältää kaikille havainnoille yhteiset selittävät muuttujat. Tutkimuksen tässä versiossa Irakissa syntyneille ei ole vielä olemassa mallia 4, joka sisältäisi palkkasummien mallin neljän vakiotermin satunnaisvaikutukset. Selittävien muuttujien vakiotermien ja regressiokertoimien tulkinnat on esitetty yllä, joten en toista niitä tässä.

Taulukossa 5.10 on esitetty Irakissa syntyneiden lopullisen palkkasummamallin (malli 3) posteriorieestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat Monte Carlo-keskivirheet (MC-keskivirheet) sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli ykköstä, en käsittele niitä tarkemmin.

Taulukko 5.10: *Irakissa syntyneiden 18–62-vuotiaiden palkkasummien lopullinen malli (malli 3). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakioparametrit:									
Vakio (H)	1,27	0,15	1,01	1,17	1,26	1,36	1,58	0,002	1,00
Vakio ($\mathcal{G}_1$)	12 210	1 290	9 850	11 300	12 110	13 070	14 850	20	1,00
Vakio ($\mathcal{G}_2$)	24 850	2 140	21 460	23 400	24 510	25 970	30 060	57	1,00
Vakio (θ)	3,91	0,38	3,24	3,65	3,88	4,13	4,70	0,0060	1,00
Regressiokertoimet (H):									
Nainen (H)	3,10	0,19	2,74	2,97	3,10	3,23	3,49	0,0015	1,00
Ikä (H)	1,33	0,19	0,99	1,19	1,31	1,45	1,75	0,0032	1,00
Ikä^2 (H)	1,09	0,16	0,82	0,98	1,09	1,19	1,44	0,0027	1,00
Maassaolo (vuosia) (H)	0,85	0,03	0,80	0,83	0,85	0,87	0,90	0,0004	1,00
Maassaolo <6 (H)	0,58	0,04	0,51	0,56	0,58	0,60	0,66	0,0003	1,00
Maassaolo $\times$ (maassaolo <6) (H)	0,03	0,01	0,02	0,03	0,03	0,04	0,04	0,0000	1,00
Maassaolo $\times$ (maassaolo <6) $\times$ nainen (H)	1,75	0,17	1,43	1,63	1,74	1,85	2,10	0,0006	1,00
Ikä <21 (H)	1,41	0,14	1,16	1,32	1,41	1,50	1,70	0,0005	1,00
Ikä <30 (H)	0,88	0,10	0,70	0,81	0,88	0,94	1,09	0,0007	1,00
(Ikä <30) $\times$ ikä (H)	0,92	0,14	0,68	0,82	0,91	1,00	1,23	0,0011	1,00
($39<$ ikä <56) $\times$ nainen (H)	0,36	0,03	0,30	0,34	0,36	0,38	0,42	0,0001	1,00
($39<$ ikä <56) $\times$ nainen $\times$ ikä (H)	1,18	0,13	0,94	1,09	1,18	1,26	1,45	0,0005	1,00
Ikä >55 (H)	0,07	0,09	0,01	0,02	0,04	0,09	0,32	0,0003	1,00
Ikä >55 $\times$ ikä (H)	6,83	3,53	2,34	4,32	6,05	8,47	15,93	0,0113	1,00
Muuttovuosi v. 1992–1999 (H)	0,78	0,05	0,70	0,75	0,78	0,81	0,88	0,0003	1,00
Muuttovuosi v. 2000–2005 (H)	0,69	0,03	0,63	0,67	0,69	0,71	0,75	0,0002	1,00
Korkein koulutus: 2. asteen koulutus (H)	0,85	0,04	0,77	0,82	0,85	0,87	0,92	0,0002	1,00
Korkein koulutus: 3. asteen koulutus (H)	0,55	0,03	0,51	0,54	0,55	0,57	0,61	0,0001	1,00
Pääasiallinen toiminta: opiskelija (H)	0,72	0,03	0,67	0,70	0,72	0,74	0,78	0,0001	1,00

Jatkuu seuraavalla sivulla

Taulukko 5.10 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Pääasiallinen toiminta: työllinen (H)	0,18	0,01	0,16	0,17	0,18	0,19	0,21	0,0000	1,00
Kotikunta Uudellamaalla (H)	0,82	0,02	0,78	0,81	0,82	0,84	0,87	0,0002	1,00
Muutto alle 20-v. (H)	0,77	0,04	0,70	0,75	0,77	0,80	0,85	0,0002	1,00
Regressiokertoimet ($\mathcal{G}_1$):									
Nainen ($\mathcal{G}_1$)	1,31	0,08	1,15	1,25	1,31	1,36	1,48	0,0006	1,00
Ikä ($\mathcal{G}_1$)	1,59	0,17	1,28	1,48	1,59	1,70	1,94	0,0028	1,00
Ikä^2 ($\mathcal{G}_1$)	0,73	0,07	0,61	0,69	0,73	0,78	0,88	0,0011	1,00
Maassaolo (vuosia) ($\mathcal{G}_1$)	1,04	0,05	0,95	1,00	1,03	1,07	1,13	0,0007	1,00
Maassaolo $<$ 6 ($\mathcal{G}_1$)	1,27	0,10	1,08	1,20	1,27	1,33	1,46	0,0008	1,00
Maassaolo $\times$ (maassaolo $<$ 6) ($\mathcal{G}_1$)	3,89	0,78	2,58	3,34	3,83	4,35	5,63	0,0061	1,00
(39 $<$ ikä $<$ 56) $\times$ nainen ($\mathcal{G}_1$)	1,48	0,14	1,23	1,39	1,48	1,57	1,78	0,0005	1,00
(39 $<$ ikä $<$ 56) $\times$ nainen $\times$ ikä ($\mathcal{G}_1$)	0,68	0,10	0,51	0,61	0,68	0,75	0,89	0,0003	1,00
Ikä $>$ 55 ($\mathcal{G}_1$)	61,45	108,76	2,94	13,57	29,34	63,83	325,29	0,3493	1,00
Ikä $>$ 55 $\times$ ikä ($\mathcal{G}_1$)	0,23	0,14	0,06	0,13	0,20	0,28	0,57	0,0004	1,00
Muuttovuosi v. 1992–1999 ($\mathcal{G}_1$)	1,42	0,12	1,19	1,33	1,41	1,50	1,67	0,0009	1,00
Muuttovuosi v. 2000–2005 ($\mathcal{G}_1$)	1,34	0,09	1,16	1,28	1,34	1,40	1,52	0,0006	1,00
Korkein koulutus: 2. asteen koulutus ($\mathcal{G}_1$)	1,05	0,06	0,94	1,01	1,05	1,09	1,18	0,0003	1,00
Korkein koulutus: 3. asteen koulutus ($\mathcal{G}_1$)	2,40	0,14	2,15	2,31	2,40	2,49	2,69	0,0006	1,00
Pääasiallinen toiminta: opiskelija ($\mathcal{G}_1$)	1,71	0,08	1,56	1,66	1,71	1,76	1,86	0,0004	1,00
Pääasiallinen toiminta: työllinen ($\mathcal{G}_1$)	2,65	0,15	2,37	2,55	2,65	2,75	2,96	0,0005	1,00
Kotikunta Uudellamaalla ($\mathcal{G}_1$)	1,27	0,04	1,19	1,24	1,27	1,30	1,35	0,0003	1,00
Muutto alle 20-v. ($\mathcal{G}_1$)	1,27	0,08	1,13	1,22	1,27	1,33	1,45	0,0005	1,00
Regressiokertoimet ($\mathcal{G}_2$):									
Nainen ($\mathcal{G}_2$)	0,71	0,02	0,67	0,70	0,71	0,73	0,75	0,0002	1,00
Ikä ($\mathcal{G}_2$)	0,97	0,08	0,83	0,91	0,96	1,01	1,15	0,0020	1,00
Ikä^2 ($\mathcal{G}_2$)	1,18	0,12	0,92	1,11	1,19	1,26	1,40	0,0034	1,00
Maassaolo (vuosia) ($\mathcal{G}_2$)	1,29	0,05	1,20	1,25	1,29	1,32	1,39	0,0008	1,00
Maassaolo $<$ 8 ($\mathcal{G}_2$)	0,96	0,05	0,86	0,93	0,96	0,99	1,07	0,0004	1,00
Maassaolo $\times$ (maassaolo $<$ 8) ($\mathcal{G}_2$)	1,61	0,15	1,34	1,51	1,60	1,70	1,93	0,0012	1,00

Jatkuu seuraavalla sivulla

Taulukko 5.10 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Maassaolo > 12 ($\mathcal{G}_2$)	1,20	0,11	1,00	1,12	1,19	1,26	1,41	0,0007	1,00
Maassaolo $\times$ (maassaolo > 12) ($\mathcal{G}_2$)	0,98	0,09	0,82	0,92	0,98	1,04	1,16	0,0006	1,00
Muuttovuosi v. 1992–1999 ($\mathcal{G}_2$)	1,02	0,04	0,95	0,99	1,02	1,04	1,09	0,0003	1,00
Regressiokertoimet ($\boldsymbol{\theta}$):									
Nainen (θ)	0,88	0,11	0,68	0,80	0,87	0,94	1,10	0,0008	1,00
Ikä^2 (θ)	1,79	0,15	1,53	1,69	1,78	1,88	2,10	0,0023	1,00
Maassaolo (vuosia) (θ)	0,40	0,04	0,32	0,37	0,40	0,43	0,49	0,0007	1,00
Maassaolo < 8 (θ)	0,79	0,13	0,57	0,69	0,77	0,86	1,07	0,0010	1,00
Muuttovuosi v. 1992–1999 (θ)	1,61	0,37	1,02	1,35	1,57	1,83	2,43	0,0028	1,00
Muuttovuosi v. 2000–2005 (θ)	0,92	0,15	0,66	0,81	0,91	1,00	1,24	0,0010	1,00
Muotoparametrit:									
Muotoparametri($\mathcal{G}_1$)	0,80	0,01	0,78	0,79	0,80	0,81	0,83	0	1,00
Muotoparametri($\mathcal{G}_2$)	12,0	1,3	9,7	11,2	11,9	12,7	14,9	0,0008	1,00

Huomioita estimointituloksista. Listaan lyhyesti joitakin huomioita keskeisimmistä vaikutuksista ja estimointituloksista taulukossa 5.10. Kun kirjoitan parametrin sijoittuvan jollekin tietylle intervallille, tarkoitan tällöin kyseisen parametrin posteriorijakauman 95%:n uskottavuusintervallia 2,5 %:sta 97,5 %:iin. Käytän lisäksi vedonlyöntisuhteesta sen englanninkieliseen nimitykseen, odds ratio, pohjautuvaa lyhennettä *OR*.

- **Sukupuoli:**
  - **Hurdle (H).** Naisindikaattorin OR on noin 2,75–3,50, eli naisilla nollapalkkasumman odds on noin 3-kertainen miehiin verrattuna. Yli ykkösen oleva OR tarkoittaa, että nollapalkkasumman todennäköisyys on suurempi kuin vertailuluokassa. Mikäli OR on pienempi kuin yksi, niin todennäköisyys on puolestaan pienempi kuin vertailuluokassa.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$).** Naisindikaattorin kerroin ensimmäisessä gamma-jakaumassa on noin 1,15–1,48 miehiin verrattuna. Ensimmäiseen gamma-jakaumaan kuuluvat palkkasummat syntyvät (luultavasti) henkilöiltä, jotka ovat vain osan vuotta töissä. Naisilla, jotka ovat vain osan vuotta töissä, palkkasummat ovat näin ollen noin 115–148 %:ia miesten palkkasummista.
  - **Toinen gamma-jakauma ($\mathcal{G}_1$).** Naisindikaattorin kerroin toisessa gamma-jakaumassa on noin 0,67–0,75 miehiin verrattuna. Toiseen gamma-jakaumaan kuuluvat palkkasummat syntyvät (luultavasti) henkilöiltä, jotka ovat koko vuoden töissä. Naisilla, jotka ovat koko vuoden töissä, palkkasummat ovat näin ollen noin 67–75 %:ia miesten palkkasummista, kun muut selittäjät huomioidaan.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ).** Naisindikaattorin OR noin 0,68–1,10 eli naisilla yli nollan olevan palkkasumman odds kuulua ensimmäiseen gamma-jakaumaan (eikä toiseen gamma-jakaumaan) on ainakin yli 75 %:n todennäköisyydellä pienempi kuin miehillä. Alle ykkösen oleva OR tarkoittaa, että ensimmäiseen gamma-jakaumaan kuulumisen todennäköisyys on pienempi kuin vertailuluokassa. Mikäli ensimmäinen gamma-jakauma vastaa vain osan vuotta työssä olevien palkkasummia (ja toinen gamma-jakauma koko vuoden töissä olevien), niin tämän tuloksen perusteella naisilla on miehiä pienempi todennäköisyys olla vain osan vuotta töissä verrattuna siihen, että olisi koko vuoden töissä, sillä naisindikaattorin odds on suurella todennäköisyydellä ykköstä pienempi.
- **Ikä:**
  - **Hurdle (H).** Alle 21-vuotiailla suurempi todennäköisyys saada nollapalkkasumma kuin muilla. Yli 55-vuotiailla todennäköisyys saada nolla nousee nopeasti iän kasvaessa.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$).** Iän päävaikutus nostaa palkkasummia odotusarvoisesti noin 1,6-kertaisiksi, kun ikä kasvaa yhden keskihajonnan. Toisaalta iän neliö on alle ykkösen, mikä tarkoittaa sitä, että iän neliön vaikutuksesta ikä aluksia nostaa palkkasummia noin 40-vuotiaaksi asti, mutta sen jälkeen palkkasummat kääntyvät laskuun iän neliön vaikutuksesta. Naisindikaattorin, ikäluokkaindikaattorin ja iän interaktio, jonka todennäköisyysjakauma sijaitsee suurimmaksi osaksi ykkösen alapuolella, tarkoittaa sitä, että 40–55-vuotiailla naisilla ikä nostaa palkkasummien odotusarvoa vähemmän kuin miehillä.

- **Toinen gamma-jakauma ($\mathcal{G}_2$)**. Ikään liittyvien vaikutusten jakaumat ovat keskittyneet ykkösen ympäristöön.

- **Maassaoloaika:**
  - **Hurdle (H)**. Maassaoloajan kasvu pienentää todennäköisyyttä nollapalkkasummalle. Maassaoloaika pienentää tätä todennäköisyyttä enemmän alle kuusi vuotta maassa olleilla kuin kauemmin maassa olleilla. Nollan todennäköisyys pienenee siten aluksi jyrkemmin, mutta tasoittuu sittemmin maassaoloajan kasvaessa.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$)**. Maassaoloajan kasvu nostaa ensimmäiseen gamma-jakaumaan kuuluvien palkkasummien odotusarvoa. Nousu on suurinta alle kuusi vuotta maassa olleilla.
  - **Toinen gamma-jakauma ($\mathcal{G}_2$)**. Maassaoloajan nousu yhdellä keskihajonnalla (5,7 vuotta) nostaa toiseen gamma-jakaumaan kuuluvien palkkasummien odotusarvoa noin 1,20–1,39-kertaisiksi. Maassaoloajan vaikutus ei vaikuttaisi riippuvan ainakaan paljoa maassaoloajasta.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ)**. Maassaoloaika vaikuttaa θ-parametriin samoin kuin hurdleen ja 1. ja 2. gamma-jakaumaan: maassaoloajan nousu pienentää yli nollan olevan palkkasumman todennäköisyyttä kuulua ensimmäiseen gamma-jakaumaan – ja nostaa todennäköisyyttä kuulua toiseen. Todennäköisyys kuulua ensimmäiseen gamma-jakaumaan pienenee nopeiten alle 6 vuotta maassa olleilla. Mikäli ensimmäinen gamma-jakauma vastaa vain osan vuotta työssä olevien palkkasummia (ja toinen gamma-jakauma koko vuoden töissä olevien), niin tämän tuloksen perusteella maassaoloaika kasvattaa todennäköisyyttä olla koko vuosi töissä – verrattuna siihen, että olisi vain osan vuodesta töissä.

- **Maahanmuuttovuosiluokka:**
  - **Hurdle (H)**. Vuosina 1992–1999 ja 2000–2005 muuttaneilla on pienempi todennäköisyys saada nollapalkkasumma kuin vertailuluokassa eli vuosina 2006–2014 muuttaneilla – kun muut vaikutukset, eli esimerkiksi maassaoloaika, otetaan huomioon.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$)**. Vuosina 1992–1999 ja 2000–2005 muuttaneilla on ensimmäisen gamma-jakauman euromääräinen palkkasumma on noin 1,16–1,67-kertainen kuin vertailuluokassa eli vuosina 2006–2014 muuttaneilla.
  - **Toinen gamma-jakauma ($\mathcal{G}_2$)**. Vuosina 1992–1999 muuttaneilla toisen gamma-jakauman euromääräinen palkkasumma on suunnilleen samansuuruinen kuin myöhemmin muuttaneilla.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ)**. Vuosina 1992–1999 muuttaneilla todennäköisyys kuulua ensimmäiseen gamma-jakaumaan on suurempi kuin vertailuluokassa eli vuosina 2006–2014 muuttaneilla. Vuosina 2000–2005 muuttaneilla todennäköisyys kuulua ensimmäiseen gamma-jakaumaan on noin 75 %:m todennäköisyydellä pienempi kuin vuosina 2006–2014 muuttaneilla.

5.4.2.3 Irak: posterioriennustetarkistukset

Palkkasummien posterioriennusteet on muodostettu yllä esitetyn mallin 3 (eli elinkaariennusteissa käytettävän) parametrien posteriorijakaumien estimaateista ja havaintoaineistosta niin, kuin on esitetty maassaoloaikojen mallinnuksen yhteydessä luvussa 3.2.6.

Posterioriennustetarkistus: Kaikkien palkkasummahavaintojen yhteisjakauma ja henkilöiden palkkasummien keskiarvojen yhteisjakauma.

Aloitan posterioriennustetarkistukset vertaamalla havaintoaineiston kaikkien palkkasummahavaintojen muodostamaa yhteisjakaumaa posterioriennustejakauman kaikkien ennusteiden muodostamaan yhteisjakaumaan. Ensimmäisen mallinnustavoitteen mukaan näiden jakaumien pitäisi vastata mahdollisimman hyvin toisiaan, erityisesti jakaumien tärkeimpien tunnuslukujen kuten odotusarvojen, keskihajontojen ja kvartiilien suhteen. Koska palkkasummat ovat jakautuneet nolliin ja yli nolliin, mallinnustavoitteiden tulisi täyttyä sekä nollien osuuden että ylinollien jakauman osalta.

Kaikkien havaintojen jakauman lisäksi vertaan kaikkien havaintoaineiston henkilöiden palkkasummien keskiarvojen jakaumaa ennustettuihin henkilöiden palkkasummien keskiarvojen jakaumaan. Nämä jakaumat muodostetaan havaintoaineistossa laskemalla palkkasummien keskiarvot kaikille aineiston henkilöille ja muodostamalla näistä jakauma. Posterioriennusteiden kohdalla lasketaan ennustettujen palkkasummien keskiarvot kaikille henkilöille ja muodostetaan näistä jakauma. Henkilötasoisten jakaumien vertailun tarkoituksena on osoittaa, että elinkaariennusteissa käytettävä nettovaikutusten malli tuottaa luotettavia ennusteita paitsi havaintotasolla niin myös henkilötasolla. Tämä tavoite ei kuitenkaan toteudu käytännössä niin kauan, kuin Irakissa syntyneiden lopullisena palkkasummamallina ei ole satunnaisvaikutuksia sisältävää mallia. Satunnaisvaikutukset ovat tässä tapauksessa henkilötason satunnaisvaikutuksia mallin neljässä vakiotermissä (vrt. tulokset Suomessa syntyneiden osalta, jossa nämä satunnaisvaikutukset ovat mukana lopullisessa mallissa).

Taulukossa 5.11 on esitetty yllä kuvatun posterioriennustetarkistuksen tulokset. Ylemmässä taulukon osassa verrataan kaikkien palkkasummahavaintojen jakauman tunnuslukujen estimaatteja vastaaviin posterioriennustejakauman estimaatteihin. Bayes-p-arvot ja Bayes-intervallit eri euromäärille ovat testisuureita estimaattien vastaavuuksille havaintoaineiston ja posterioriennusteiden välillä. Bayes-intervalli ilmaisee todennäköisyyden sille, että posterioriennustejakauman estimaatti poikkeaa korkeintaan ilmoitetun euromäärän verran havaintoaineiston estimaatista. Esimerkiksi 2 000 euron Bayes-intervallin sisään mahtuvat kaikkien havaintojen jakauman osalta varsin suurella todennäköisyydellä kaikki jakaumapisteet lukuun ottamatta nollan ympäristöön havaintoaineistossa sijoittuvaa 90. prosenttipistettä. Bayes-p-arvo ja Bayes-intervallit on määritetty tarkemmin työikäisten nettovaikutusmallinnuksen yhteydessä (luvussa 4.3.7).

Taulukon ylemmästä osasta käy ilmi, että kaikkien havaintojen ja posterioriennusteiden jakaumat vastaavat hyvin toisiaan niin jakaumien havaittujen keskiarvojen, keskihajontojen sekä nollaosuuksien osalta. Myös posterioriennusteiden ylinollien keskiarvo ja keskihajonta vastaavat hyvin havaintoaineiston vastaavia jakauman tunnuslukuja.

Taulukko 5.11: *Palkkasummien havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien estimaattien vertailu. Kaikkien ennusteiden jakauma ja ennustettujen henkilötasoisten keskiarvojen jakauma. Jakaumien odotusarvot (OA), keskihajonnat (KH), prosenttipisteet:* $1, 5, 10, 25, 50, 75, 90, 95, 99$*, ylinollien odotusarvot (OA(yn)), ylinollien keskihajonnat (KH(yn)) ja nollien osuudet prosentteina* $(0-\%)$*. Vertailtavien estimaattien Bayes-p-arvot ja Bayes-intervallit eri euromäärillä.*

Kaikki havainnot ja ennusteet	OA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %	$OA(yn)$	$KH(yn)$	$0-\%$
Posterioriennusteet	3 080	9 160	0	0	0	0	0	200	9 880	21 440	42 340	11 690	14 780	73,7
Aineiston havainnot	3 070	9 190	0	0	0	0	0	280	9 530	22 100	40 870	11 660	14 860	73,7
Bayes-p-arvo	0,57	0,44	0	0	0	0	0	0,12	0,78	0,21	0,95	0,56	0,41	0,50
Bayes-intervalli: 1 000 euroa	1	1	1	1	1	1	1	1	0,92	0,67	0,31	1	0,95	–
Bayes-intervalli: 500 euroa	1	0,90	1	1	1	1	1	1	0,56	0,39	0,15	0,86	0,68	–
Bayes-intervalli: 200 euroa	0,97	0,58	1	1	1	1	1	1	0,20	0,18	0,07	0,43	0,34	–

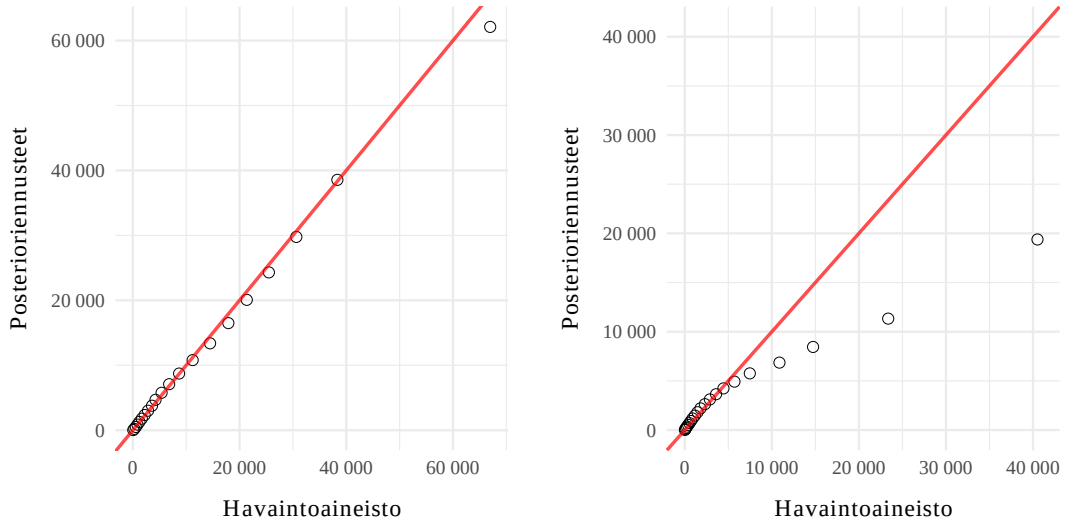
Henkilöiden keskiarvot	OA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %	$OA(yn)$	$KH(yn)$	$0-\%$
Posterioriennusteet	2 200	3 790	0	0	0	0	440	3 080	6 600	9 390	17 020	3 630	4 300	39,5
Aineiston henkilöt	2 150	6 400	0	0	0	0	0	1 070	5 840	12 740	31 810	5 290	9 180	59,4
Bayes-p-arvo	0,78	0	0	0	0	0	1	1	1	0	0	0	0	0
Bayes-intervalli: 1 000 euroa	1	0	1	1	1	1	1	0	0,89	0	0	0	0	–
Bayes-intervalli: 500 euroa	1	0	1	1	1	1	0,94	0	0,10	0	0	0	0	–
Bayes-intervalli: 200 euroa	0,99	0	1	1	1	1	0	0	0	0	0	0	0	–

Taulukon 5.11 alemmasta osasta selviää, miksi lopulliseen malliin tulisi vielä sisällyttää ainakin mallin vakiotermien osalta henkilötason satunnaisvaikutukset. Henkilöiden keskiarvojen osalta vain posterioriennusteiden keskiarvo on lähellä havaintoaineiston keskiarvoa. Nollaosuuksista huomaa selvästi, että havaintoaineistossa nollat keskittyvät selvästi tiettyihin henkilöihin syistä, joita mallissa olevat selittävät muuttujat eivät tavoita – toisin sanoen havaintoaineistossa olevasta *havaitsemattomasta heterogeenisuudesta* johtuen. Samasta syystä myös ylinollien keskihajonta on havaintoaineistossa huomattavasti suurempaa kuin henkilöiden keskiarvojen osalta posterioriennusteissa.

Vertaan alla kuviossa 5.13 havaintoaineiston jakaumien prosenttipisteitä vastaavien posterioriennustejakaumien prosenttipisteisiin. Jakaumissa ovat mukana vain yli nollan olevan havainnot sekä ennusteet, ”ylinollat”. Koska nollien osuuksien vastaavuus havaintoaineiston ja posterioriennusteiden välillä on jo käsitelty yllä taulukossa 5.11, en käsittele nollaosuuksia tarkemmin enää. Ylinollien jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeamat vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).

Kaikkien havaintojen ja ennusteiden ylinollien jakaumat vastaavat toisiaan hyvin lukuun ottamatta posterioriennusteiden ylinollien jakauman oikeaa häntää, jossa ennusteet ovat suurempia kuin havaintoaineiston arvot (vrt. 99. prosenttipiste). Kuten jo edellä kirjoitin, henkilöiden keskiarvojen osalta posterioriennusteiden (ylinollien) jakauma vastaa huonosti havaintoaineiston jakaumaa (niiden keskiarvoa lukuun ottamatta). Tämä johtuu siitä, että Irakissa syntyneiden lopullisessa mallissa ei ole toistaiseksi sellaisia satunnaisvaikutuksia, jotka ottaisivat mallissa sen huomioon, että korkeat (ja matalat) palkkasummat ryvästyvät henkilöittäin. Seurauksena tästä on, että elinkaariennusteiden ennustetuissa palkkasummissa on liian vähän henkilöiden välistä vaihtelua.



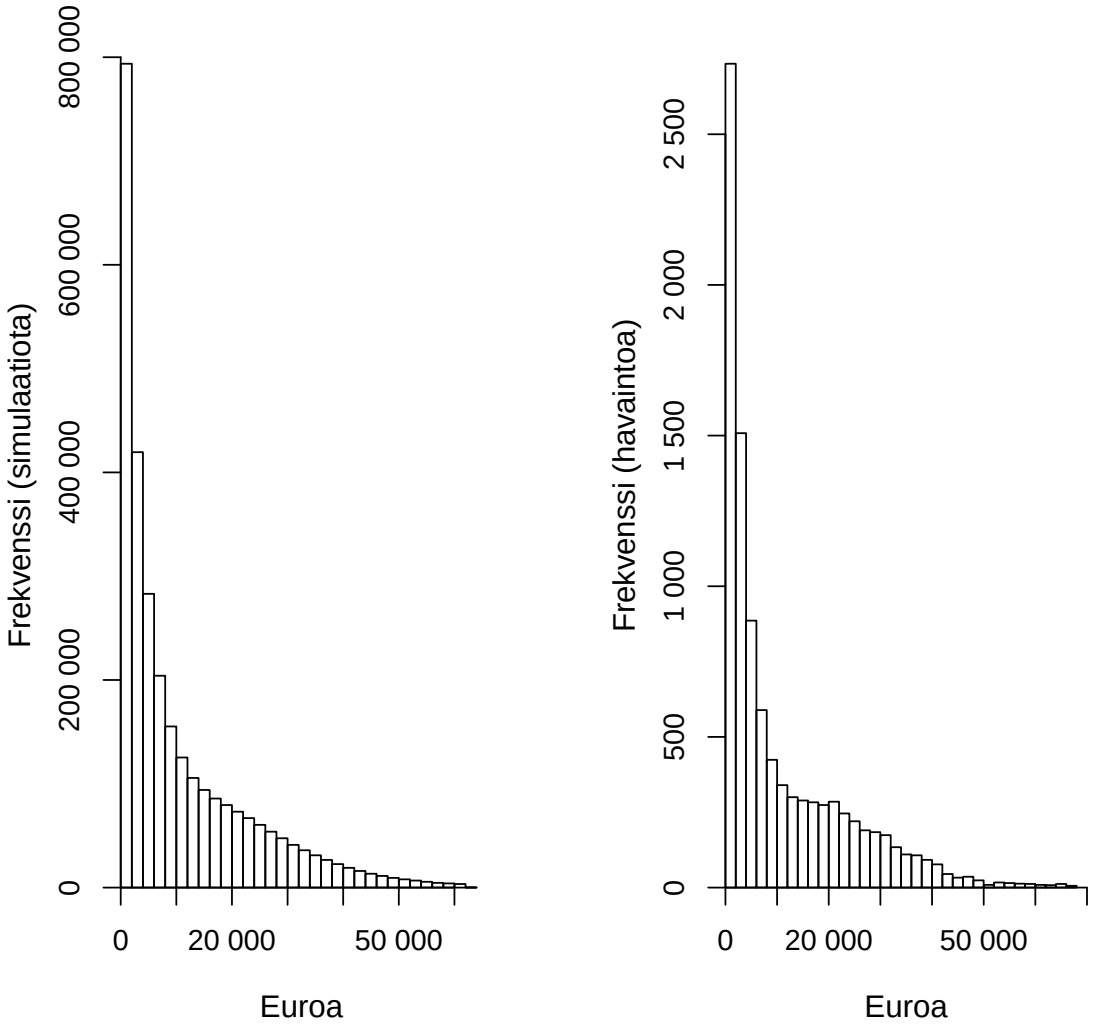
(a) *Kaikki havainnot ja ennusteet.*

(b) *Henkilöiden keskiarvot.*

Kuva 5.13: *Havaintoaineiston yli nollan olevien havaintojen jakauman ja posterioriennusteiden yli nollan olevien ennusteiden jakauman vertailu. "Ylinollien" jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

Verrataan vielä posterioriennusteiden ja havaintoaineiston ylinollien jakaumia prosenttipisteiden vertailukuvioiden (Q-Q plot) lisäksi histogrammien avulla. Nämä on esitetty kuvassa 5.14. Vasemmalla oleva histogrammi esittää palkkasummien yli nollan olevien posterioriennusteiden jakaumaa. Oikealla oleva histogrammi esittää puolestaan palkkasummien yli nollan olevien havaintojen jakaumaa. Kummastakin jakaumasta on katkaistu pois ylin prosentti yksityisyydensuojan varmistamiseksi.

Myös ylinollia koskevat histogrammit osoittavat, että kaikkien havaintojen ja posterioriennusteiden jakaumat muistuttavat toisiaan – lähinnä lukuun ottamatta posterioriennusteiden oikeaa häntää, joka histogrammikuviossakin ylettyy enemmän oikealle kuin havaintoaineistojen oikea häntä.



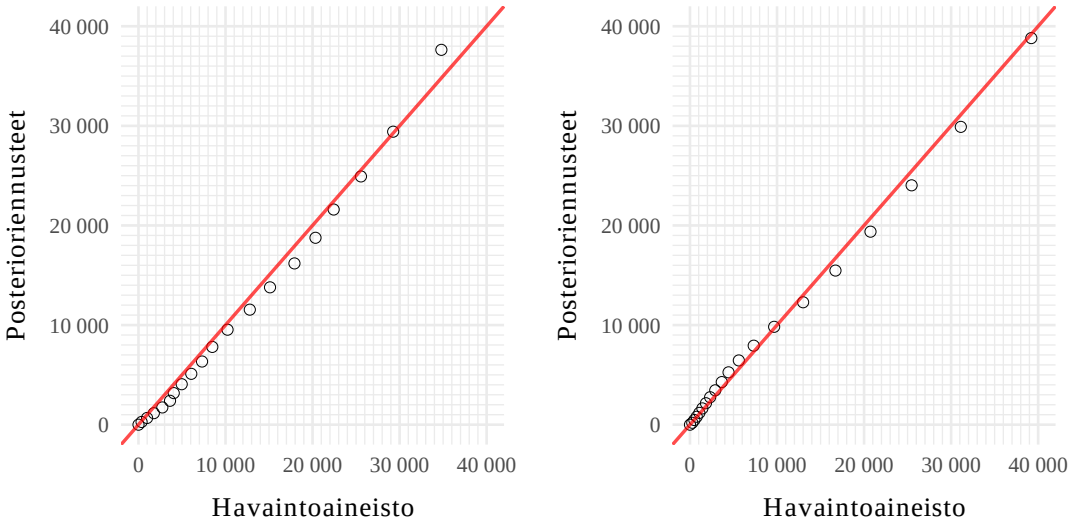
Kuva 5.14: *Posterioriennustetarkistus histogrammeilla. Posterioriennusteiden yli nollan olevien ennusteiden jakauma ja havaintoaineiston yli nollan olevien havaintojen jakauma. Posterioriennusteiden ylinollien jakauma vasemmalla ja havaintoaineiston ylinollien jakauma oikealla. Kummastakin jakaumasta on katkaistu pois ylin prosentti yksityisyydensuojan varmistamiseksi.*

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan

Vertaan nyt posterioriennusteita havaintoaineistoon tärkeimpien selittävien muuttujien luokkien mukaan. Tärkeimpiä selittäviä muuttujia ovat sukupuoli sekä aikaan liittyvät muuttujat: ikä, maahanmuuttovuosi, maassaoloaika ja kalenterivuosi. Viimeinen näistä muuttujista ei ollut lainkaan mallissa, koska maahanmuuttovuosi, maassaoloaika ja kalenterivuosi eivät voi olla samaan aikaan lineaarisessa mallissa. Esitän posterioriennustetarkistukset myös maahanmuuttoiän mukaan sekä iän mukaan niille, joilla on kulunut yli 15 vuotta ensimmäisestä muutosta Suomeen.

Vertaan posterioriennusteiden nollaosuuksia sekä ”ylinollien” keskiarvoja selittävien luokkien mukaan vastaaviin havaintoaineiston nollaosuuksiin ja ”ylinollien” keskiarvoihin. Koska palkkasummat koostuvat nollista sekä yli nollan olevasta osasta, esitän nollaosuudet ja ylinollien keskiarvot erikseen. Esimerkiksi iän kohdalla vertaan posterioriennusteiden nollaosuuksia havaintoaineiston nollaosuuksiin niin, että lasken sekä posterioriennusteista että havaintoaineistosta jokaiselle iälle erikseen nollien osuuden 18-vuotiaista 62-vuotiaisiin. Ylinollien osalta lasken puolestaan erikseen jokaiselle iälle havaintojen ja ennusteiden ylinollien keskiarvot.

Havainto- ja posterioriennustejakaumien vertailu sukupuolen mukaan. Ennen kuin vertaan nollaosuuksia ja ylinollien keskiarvoja selittävien muuttujien mukaan, esitän vielä ylinollien jakaumien tarkistukset erikseen sekä naisille että miehille. Kuviossa 5.15 on esitetty ylinollien jakaumien prosenttipisteet sukupuolen mukaan. Kuviosta selviää, että miehillä (oikeanpuoleinen osakuva) posterioriennusteiden ylinollien jakauma vastaa melko hyvin havaintoaineiston ylinollien jakaumaa – sopivuutta voisi vielä parantaa ensimmäiseen gamma-jakaumaan suurimaksi osaksi kuuluvien palkkasummien osalta (10 000 euroon asti). Naisten kohdalla (vasemmanpuoleinen osakuva) posterioriennusteiden ylinollien jakauma vastaa myös melko hyvin havaintojen jakaumaa. Erityisesti 10 000–20 000 euron välillä havaintoaineiston pisteet saavat suurempia arvoja kuin posterioriennusteiden vastaavat. Toisaalta jakauman oikeassa hännässä ennusteet ovat havaintoja suurempia.

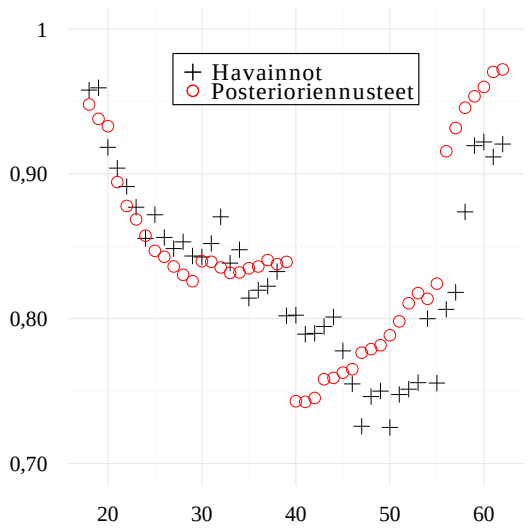


(a) *Naiset.*

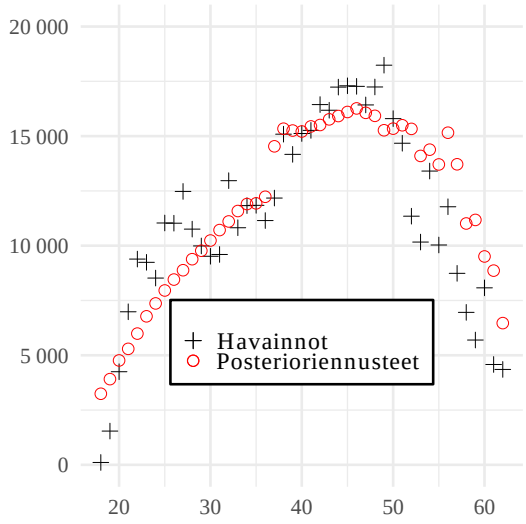
(b) *Miehet.*

Kuva 5.15: *Havaintoaineiston yli nollan olevien havaintojen jakauman ja posterioriennusteiden yli nollan olevien ennusteiden jakauman vertailu sukupuolen mukaan. "Ylinollien" jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

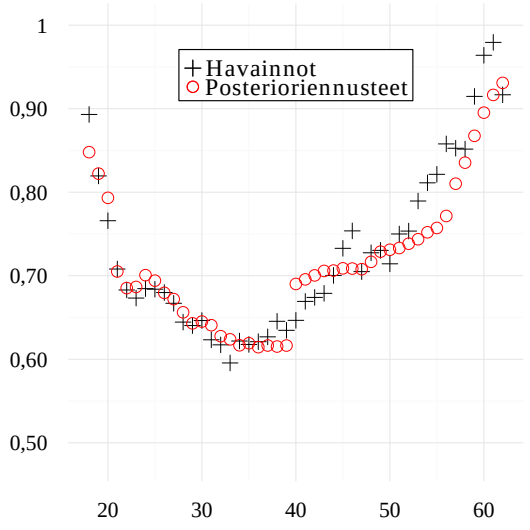
Posterioriennustetarkistus iän ja sukupuolen mukaan.



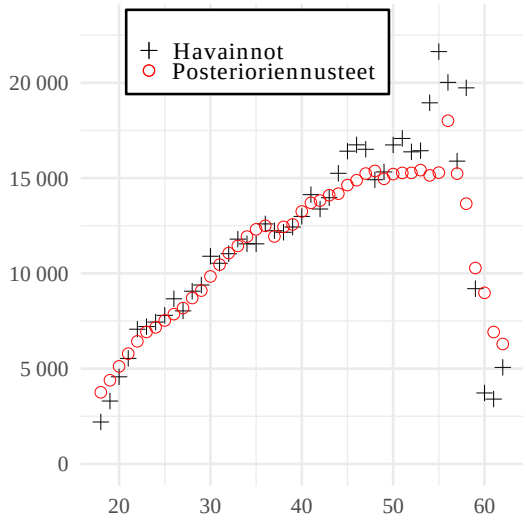
(a) *Nollien osuudet iän mukaan. Naiset.*



(b) *Ylinollien keskiarvot iän mukaan. Naiset.*



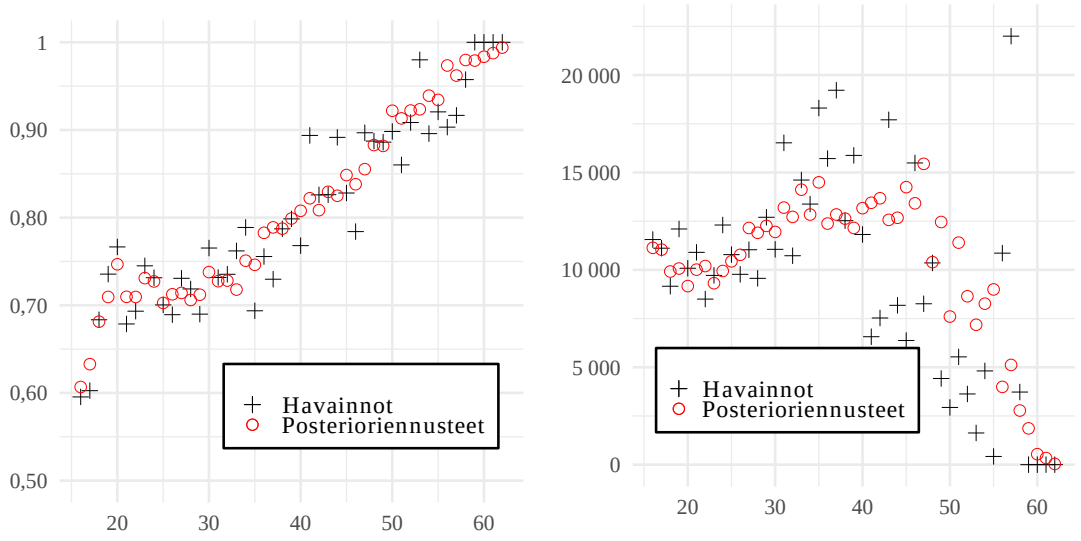
(c) *Nollien osuudet iän mukaan. Miehet.*



(d) *Ylinollien keskiarvot iän mukaan. Miehet.*

Kuva 5.16: Havainnot ja posterioriennusteet iän ja sukupuolen mukaan.

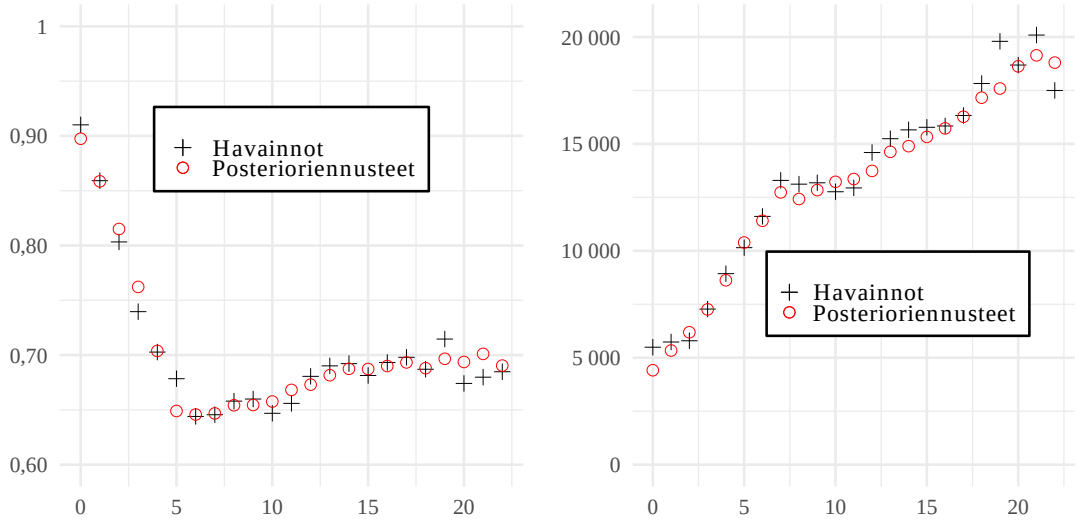
Posteriorienustetarkistus maahanmuuttoiän ja maassaoloajan mukaan.



(a) *Nollien osuudet maahanmuuttoiän mukaan.*

(b) *Ylinollien keskiarvot maahanmuuttoiän mukaan.*

Kuva 5.17: Havainnot ja posteriorienustteet maahanmuuttoiän mukaan.

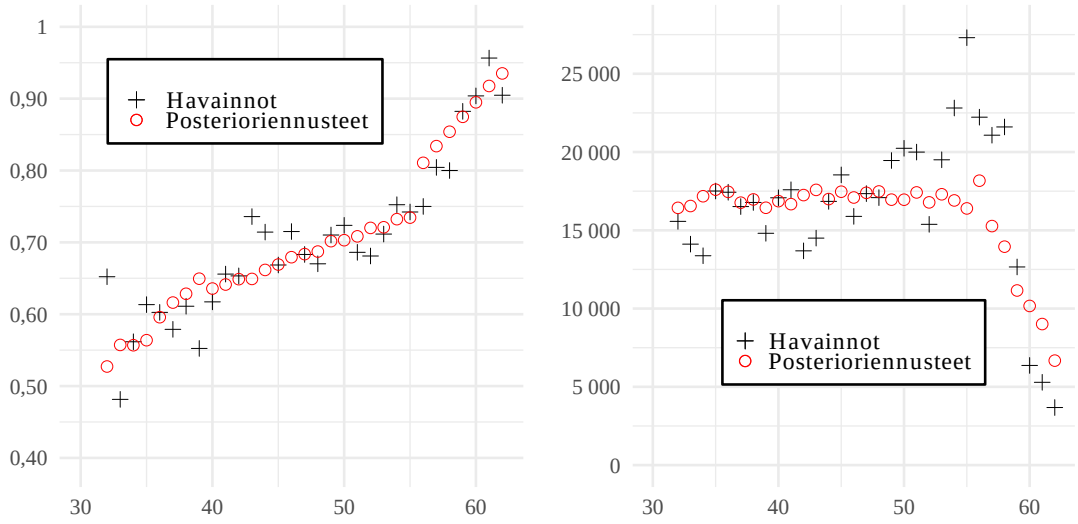


(a) *Nollien osuudet maassaoloajan mukaan.*

(b) *Ylinollien keskiarvot maassaoloajan mukaan.*

Kuva 5.18: Havainnot ja posteriorienustteet maassaoloajan mukaan.

Posterioriennustetarkistus iän mukaan. Yli 15 vuotta maassa olleet.

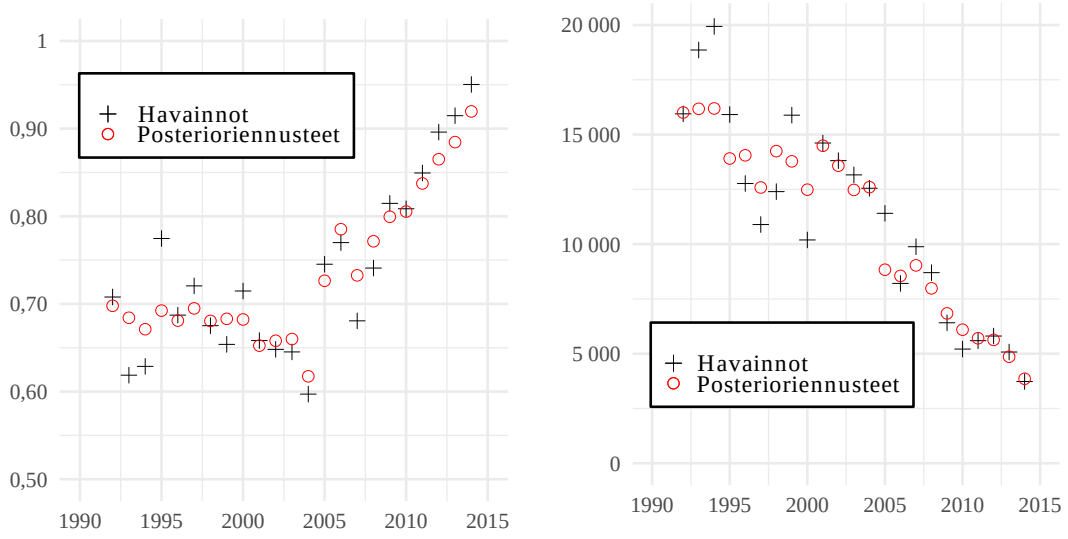


(a) *Nollien osuudet iän mukaan, yli 15 vuotta maassa olleet.*

(b) *Ylinollien keskiarvot iän mukaan, yli 15 vuotta maassa olleet.*

Kuva 5.19: Havainnot ja posterioriennusteet iän mukaan. Yli 15 vuotta maassa olleet.

Posterioriennustetarkistus maahanmuuttovuoden mukaan.

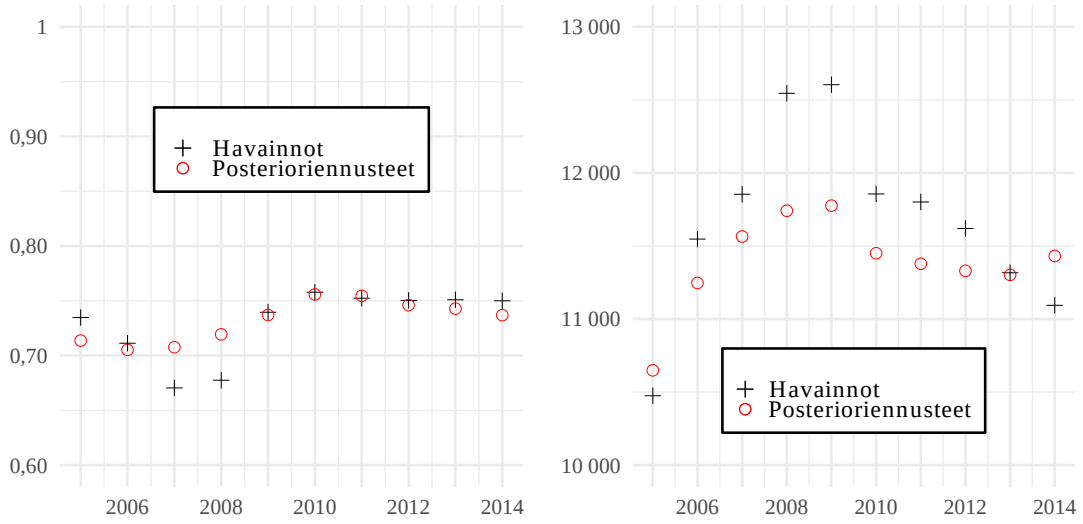


(a) *Nollien osuudet maahanmuuttovuoden mukaan.*

(b) *Ylinollien keskiarvot maahanmuuttovuoden mukaan.*

Kuva 5.20: Havainnot ja posterioriennusteet maahanmuuttovuoden mukaan.

Posterioriennustetarkistus kalenterivuoden mukaan.



(a) *Nollien osuudet kalenterivuoden mukaan.* (b) *Ylinollien keskiarvot kalenterivuoden mukaan.*

Kuva 5.21: Havainnot ja posterioriennusteet kalenterivuoden mukaan.

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan. Havaintoja kuvista 5.16–5.21.

Posterioriennusteiden nollaosuudet ja ylinollien keskiarvot selittävien muuttujien luokkien mukaan vastaavat hyvin havaintoaineiston vastaavia arvoja kuvissa 5.16–5.21. Ehkä suurimmat systemaattiseet virheet posterioriennusteiden ja havaintoaineiston välillä koskevat naisten palkkasummien nollaosuuksia ja ylinollia kuvassa 5.16. Mallia voisi vielä parantaa niin, että yli 40-vuotiaiden naisten posterioriennusteiden nollaosuudet vastaavaisivat paremmin havaintoaineiston arvoja. Sama koskee myös yli 50-vuotiaita naisia ylinollian osalta.

5.4.3 Suomi

5.4.3.1 Suomi: havaintoaineisto

Taulukossa 5.12 on esitetty Suomessa syntyneiden 18–62-vuotiaiden palkkasummahavaintojen lukumäärä ja havaintojen jakauman keskeisimmät tunnusluvut. Havainnot koskevat vuosia 2005–2014. Palkkasummahavainnot muodostavat selitettävän muuttujan palkkasummien mallinnuksessa. Taulukossa 5.12 on esitetty samat tiedot myös havaintoaineiston henkilöiden palkkasummahavaintojen keskiarvoille. Taulukossa 5.12 on esitetty kaikkien palkkasummahavaintojen jakaumatietojen lisäksi vielä erikseen niiden palkkasummahavaintojen, jotka ovat yli nollan, jakauman keskiarvo ja keskihajonta. Lisäksi taulukon alimmalla rivillä on esitetty nollien osuudet havainnoista.

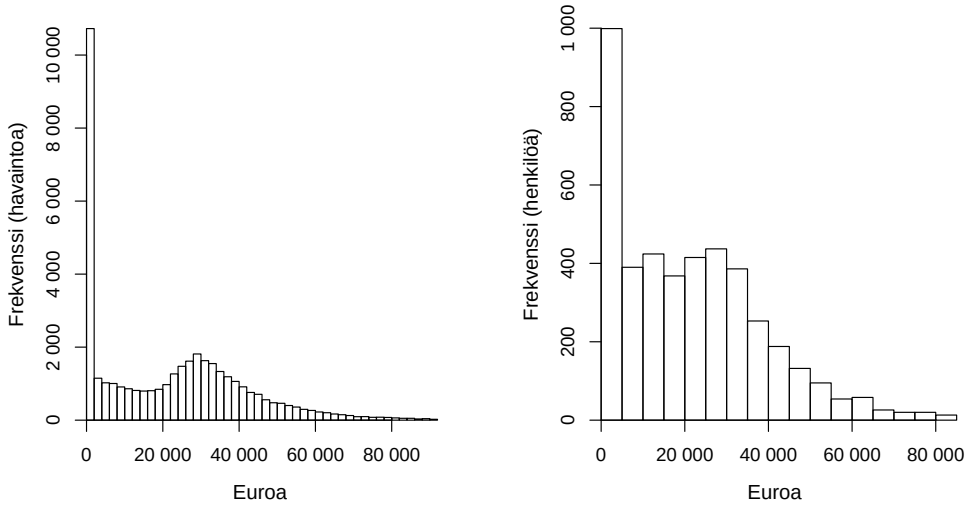
Suomessa syntyneiden palkkasummahavaintojen mallinnusaineisto koostuu 40 001 henkilövuositasoisesta palkkasummahavainnosta, jotka on poimittu paneeliaineistona/pitkittäisaineistona 4 324 henkilöltä. Mallinnusaineisto ei koostu kaikista koko tutkimusaineistoon kuuluvista Suomessa syntyneistä henkilöistä ja heitä koskevista palkkasummahavainnoista. Otin mallinnusaineistoa varten yksinkertaisen satunnaisotoksen (ilman palauttamista) kaikista tutkimusaineistoon sisältyvistä Suomessa syntyneistä henkilöitä niin, että näitä henkilöitä koskeva havaintoaineisto olisi noin 40 000 havaintoa. En käyttänyt koko tutkimusaineistoa, koska erityisesti satunnaisvaikutuksia sisältävän tilastollisen mallin ajamiseen olisi mennyt koko tutkimusaineistolla – jossa satoja tuhansia havaintoja – vähintään viikkoja. Sisällytin mukaan mallinnusaineistoon vain ne henkilöt, joita koskien tutkimusaineistossa on vähintään viisi havaintoa. Oletin, että tämä ratkaisu voisi parantaa satunnaisvaikutuksia sisältävän mallin estimoituvuutta. Tämän otosvalinnan seurauksena mallinnusaineistosta puuttuu erityisesti henkilöitä, joiden havainnot sijoittuvat ikäskaalan ala- tai yläpäähän (eli lähelle 18:aa ja 62:ä ikävuotta). Koska mallinnuksessa huomioidaan henkilöiden ikä, tällä otosvalinnalla ei pitäisi olla mitään merkitystä mallinnustulosten ja siten elinkaariennusteissa käytettävien palkkasummaennusteiden kannalta.

Taulukon 5.12 tietojen mukaan esimerkiksi kaikkien palkkasummahavaintojen keskiarvo on 22 790 euroa, keskihajonta on 22 480 euroa ja viisi prosenttia kaikista havainnoista on 59 660 euroa suurempia. Havaintoaineiston henkilöiden palkkasummahavaintojen keskiarvojen keskiarvo on 22 210 euroa ja keskihajonta 20 270 euroa. Yli nollien keskiarvo kaikkien havaintojen joukossa on 29 500 euroa. Nollien osuus koko havaintoaineistossa on 23 prosenttia.

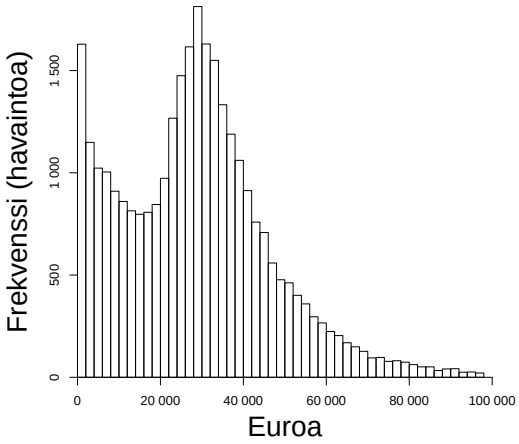
Taulukko 5.12: *Suomessa syntyneet 18–62-vuotiaat vuosina 2005–2014. Palkkasummahavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa sekä henkilöiden lukumäärä ja henkilötasoisten palkkasummahavaintojen keskiarvojen jakauman tunnusluvut mallinnusaineistossa. Havaintojen osalta "ylinollat" käsittävät vain ne havainnot, joissa palkkasumma on yli nolla euroa; henkilöiden osalta ne henkilöt, joiden keskiarvo on yli nollan. Nollien osuudet kertovat nollien osuudet kaikkien havaintojen sekä kaikkien henkilöiden keskiarvojen joukoissa. Euromääräiset tunnusluvut pyöristetty lähimpään kymmeneen euroon.*

	Havainnot	Henkilöt
Kokonaislukumäärä (N)	40 001	4 324
Keskiarvo	22 790	22 210
Keskihajonta	22 480	20 270
1. prosenttipiste (1 %)	0	0
5. prosenttipiste (5 %)	0	0
10. prosenttipiste (10 %)	0	0
Alakvartiili (25 %)	890	6 030
Mediaani (50 %)	22 170	19 740
Yläkvartiili (75 %)	35 080	32 640
90. prosenttipiste (90 %)	48 830	46 270
95. prosenttipiste (95 %)	59 660	56 050
99. prosenttipiste (99 %)	91 010	87 170
Ylinollat: keskiarvo	29 500	–
Ylinollat: keskihajonta	21 350	–
Nollien osuus	0,23	0,11

Havaintoaineistossa noin 26 %:ia havainnoista on yli nollan. Kuvassa 5.23 on esitetty näiden yli nollien olevien palkkasummien jakauman histogrammi. Jakaumasta on katkaistu pois ylin prosentti. Histogrammista näkee, että havaintojen suurin tiheys (frekvenssi) on lähellä nollaa ja laskee oikealle päin mentäessä. Noin 20 000 euron kohdalla yli nollien jakaumassa on kuitenkin paikallinen huippu. Mallinnuksessa toisen gamma-jakauman moodi sijoittuu suunnilleen tähän kohtaan. Tämä paikallinen huippu on helpommin nähtävissä Suomessa syntyneiden vastaavassa histogrammissa (ks. histogrammi, s. 314).



Kuva 5.22: *Suomessa syntyneet 18–62-vuotiaat. Palkkasummahavaintojen katkaistun jakauman histogrammi mallinnusaineistossa (vasemmalla) sekä henkilötasoisten palkkasummahavaintojen keskiarvojen katkaistun jakauman histogrammi (oikealla). Jakaumista on poistettu ylin prosentti yksityisyydensuojan varmistamiseksi.*



Kuva 5.23: *Suomessa syntyneet 18–62-vuotiaat. Yli nollan olevien palkkasummahavaintojen katkaistun jakauman histogrammi mallinnusaineistossa. Jakaumasta on poistettu ylin prosentti yksityisyydensuojan varmistamiseksi.*

Taulukossa 5.13 on esitetty Suomessa syntyneiden palkkasummamallien selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Indikaattorimuuttujien kohdalla keskiarvo kertoo kyseisen muuttujan osuuden kaikkien havaintojen joukossa. Edellä parametrien selitetaulukossa (s. 262) on lyhyitä selityksiä selittävistä muuttujista.

Taulukko 5.13: *Suomessa syntyneet 18–62-vuotiaat. Palkkasummamallien selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta).*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,49	0,50	0	0	1
Ikä	41,1	12,7	30	42	52
Ikä^2	1850	1035	900	1764	2704
Ikä < 21	0,05	0,21	0	0	0
Ikä < 30	0,24	0,43	0	0	0
(Ikä < 30) $\times$ ikä	5,74	10,33	0	0	0
Ikä > 55	0,16	0,37	0	0	0
(Ikä > 55) $\times$ ikä	9,67	21,79	0	0	0
Ikä > 60	0,04	0,19	0	0	0
Kalenterivuosi	2009	2,8	2007	2009	2012

5.4.3.2 Suomi: estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Taulukossa 5.14 on esitetty perättäisten palkkasummamallien parametrien odotusarvoestimaatit Suomessa syntyneille 18–62-vuotiaille. Olen käsitellyt perättäisten mallien rakentamista mallinnusaskel kerrallaan alaluvussa 5.3.1. Estimoitavien parametrien selitteet on puolestaan esitetty parametrien selitetaulukossa s. 262. Selitetaulukossa on myös selvennetty lyhyesti, mitä tarkoittaa, että parametri liittyy joko hurdleen (H), ensimmäiseen gamma-jakaumaan ($\mathcal{G}_1$), toiseen gamma-jakaumaan ($\mathcal{G}_2$) tai θ-indikaattoriin, joka kertoo, kuuluuko havainto ensimmäiseen gamma-jakaumaan (eikä toiseen) kahden gamma-jakauman sekoitusjakaumassa. Olen käsitellyt Somaliaa koskevien estimointitulosten yhteydessä (luvussa 5.4.1.2) jo palkkasummamallien parametrien tulkintaa, joten en toista sitä tässä yhteydessä.

Mallin 4 sisältämillä satunnaisvaikutuksilla on gammajakauma, $U \sim \mathcal{G}(1/\lambda, 1/\lambda)$, kuten olen esittänyt edellä mallin 4 määrittämisen yhteydessä. Taulukon 5.14 ”satunnaisvaikutusten parametrit” ovat siten parametreja, jotka löytyvät satunnaisvaikutusten gamma-jakaumista. Gamma-jakauman ominaisuuksista johtuen nämä parametrit ovat samalla gamma-jakautuneiden satunnaisvaikutusten variansseja.

Taulukko 5.14: *Perättäisten palkkasummamallien parametrien odotusarvoestimaatit, mallikohtaiset ennustetarkkuuden estimaatit ja keskihajonnat (Waic ja LOO). Mallien valinta on esitetty alaluvussa 5.3.1. H:n ja θ:n vakioparametrit ovat oddseja, regressiokertoimet vedonlyöntisuhteita (engl. odds ratio). Gamma-jakaumien ($\mathcal{G}_1$ ja $\mathcal{G}_2$) vakioparametrit ovat euroja, regressiokertoimet odotusarvon lineaarisia kertoimia. Indikaattorimuuttujien osalta regressiokertoimet ovat palkkasumman odotusarvon muutoksia suhteessa vertailuryhmään, ja jatkuvien muuttujien osalta regressiokertoimet ovat kertoimia, joilla odotusarvot kerrotaan, kun muuttujan arvo kasvaa yhdellä keskihajonnalla.*

	Malli 1	Malli 2	Malli 3	Malli 4
Vakioparametrit:				
Vakio (H)	1,43	0,29	0,24	0,04
Vakio ($\mathcal{G}_1$)	29 490	24 890	38 500	22 630
Vakio ($\mathcal{G}_2$)		34 250	40 690	37 070
Vakio (θ)		1,03	0,24	0,04
Hurdle-regressiokertoimet (H):				
Nainen (H)			0,71	0,78
Ikä (H)			1,41	0,99
Ikä^2 (H)			1,32	2,01
Ikä $<$ 21 (H)			1,59	1,97
Ikä $<$ 30 (H)			1,31	1,72
Ikä $<$ 30 $\times$ ikä (H)			1,63	2,55
Ikä $>$ 60 (H)			1,53	2,44
Kalenterivuosi (H)			1,01	1,36
1. gamma-jakauman regressiokertoimet ($\mathcal{G}_1$):				
Nainen ($\mathcal{G}_1$)			0,68	0,88
Ikä ($\mathcal{G}_1$)			1,63	1,31
Ikä^2 ($\mathcal{G}_1$)			0,77	0,88
Ikä $<$ 30 ($\mathcal{G}_1$)			2,46	4,30
Ikä $<$ 30 $\times$ ikä ($\mathcal{G}_1$)			2,01	3,22
Kalenterivuosi ($\mathcal{G}_1$)			1,01	0,99
2. gamma-jakauman regressiokertoimet ($\mathcal{G}_2$):				
Nainen ($\mathcal{G}_2$)			0,78	0,80
Ikä ($\mathcal{G}_2$)			1,20	1,11
Ikä^2 ($\mathcal{G}_2$)			0,96	0,95
Ikä $<$ 30 ($\mathcal{G}_2$)			1,62	1,11
Ikä $<$ 30 $\times$ ikä ($\mathcal{G}_2$)			1,72	1,13
Kalenterivuosi ($\mathcal{G}_2$)			1,00	1,00
Kuuluminen 1. gamma-jakaumaan, regr.kert. (θ):				
Nainen (θ)			1,08	2,15
Ikä (θ)			0,62	0,18

Jatkuu seuraavalla sivulla

Taulukko 5.14 – jatkuu edelliseltä sivulta

	Malli 1	Malli 2	Malli 3	Malli 4
Ikä^2 (θ)			1,39	2,32
Ikä < 30 (θ)			0,11	0,00
Ikä $< 30 \times$ ikä (θ)			0,14	0,00
Ikä > 55 (θ)			0,49	0,02
Ikä $> 55 \times$ ikä (θ)			2,30	52,26
Ikä > 60 (θ)			1,07	1,22
Kalenterivuosi (θ)			0,99	0,99
Gamma-jakaumien muotoparametrit:				
Muotoparametri($\mathcal{G}_1$)	1,43	0,85	0,94	1,31
Muotoparametri($\mathcal{G}_2$)		10,3	10,7	127,9
Satunnaisvaikutusten parametrit:				
λ (H)				5,55
λ ($\mathcal{G}_1$)				0,24
λ ($\mathcal{G}_2$)				0,11
λ (θ)				3,20
Mallien ennustetarkkuudet Waic ja LOO:				
Waic (estimaatti)	230 774	223 697	212 150	139 880
Waic (estimaatin keskivirhe)	(400)	(414)	(431)	(509)
LOO (estimaatti)	230 774	223 697	212 150	140 424
LOO (estimaatin keskivirhe)	(400)	(414)	(431)	(514)

Taulukon 5.14 ennustemittarien Waic ja LOO mukaan parametrimääriltään isommat mallit ovat parempia ennustajia kuin pienemmät mallit. Täten myös ennustemittarit tukevat sitä, että lopulliseen – elinkaariennusteissa käytettävään – malliin sisältyy ainakin vakiotermejä koskevat henkilötason satunnaisvaikutukset.

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” mallin on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän uusien maahanmuuttajien elinkaariennusteissa palkkasummien ennustamiseen Suomessa syntyneiden osalta. Suomin osalta lopullinen mallin on malli 3, joka sisältää kaikille havainnoille yhteiset selittävät muuttujat. Tutkimuksen tässä versiossa Suomessa syntyneille ei ole vielä olemassa mallia 4, joka sisältäisi palkkasummien mallin neljän vakiotermin satunnaisvaikutukset. Selittävien muuttujien vakiotermien ja regressiokertoimien tulkinnat on esitetty yllä, joten en toista niitä tässä.

Taulukossa 5.15 on esitetty Suomessa syntyneiden lopullisen palkkasummamallin (malli 4) posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat Monte Carlo-keskivirheet (MC-keskivirheet) sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli ykköstä, en käsittele niitä tarkemmin.

Vakioiden keskihajonnat ovat hyvin pieniä, sillä estimointisyistä käytin vakiotermeille hyvin informatiivisia jakaumia, kuten priorijakaumia koskevassa luvussa 5.3.2 esitettiin.

Taulukko 5.15: *Suomessa syntyneiden 18–62-vuotiaiden palkkasummien lopullinen malli (malli 4). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakioparametrit:									
Vakio (H)	0,04	0	0,04	0,04	0,04	0,05	0,05	0,0002	1,00
Vakio ($\mathcal{G}_1$)	22 630	0	22 630	22 630	22 630	22 630	22 640	0,04	1,00
Vakio ($\mathcal{G}_2$)	37 070	0	37 070	37 070	37 070	37 080	37 080	0,06	1,00
Vakio (θ)	0,04	0,01	0,02	0,03	0,04	0,04	0,06	0,0003	1,00
Regressiokertoimet (H):									
Nainen (H)	0,78	0,06	0,67	0,74	0,78	0,82	0,91	0,0012	1,00
Ikä (H)	0,99	0,06	0,87	0,94	0,98	1,03	1,12	0,0019	1,00
Ikä^2 (H)	2,01	0,14	1,75	1,91	2,00	2,10	2,30	0,0038	1,00
Ikä < 21 (H)	1,97	0,25	1,53	1,79	1,95	2,12	2,50	0,0012	1,00
Ikä < 30 (H)	1,72	0,63	0,80	1,26	1,62	2,05	3,20	0,0081	1,00
Ikä $< 30 \times$ ikä (H)	2,55	0,94	1,19	1,89	2,38	3,04	4,84	0,0124	1,00
Ikä > 60 (H)	2,44	0,29	1,91	2,23	2,42	2,62	3,05	0,0009	1,00
Kalenterivuosi (H)	1,36	0,03	1,30	1,34	1,36	1,38	1,42	0,0005	1,00
Regressiokertoimet ($\mathcal{G}_1$):									
Nainen ($\mathcal{G}_1$)	0,88	0,02	0,84	0,87	0,88	0,90	0,93	0,0003	1,00
Ikä ($\mathcal{G}_1$)	1,31	0,03	1,25	1,29	1,31	1,33	1,36	0,0007	1,00
Ikä^2 ($\mathcal{G}_1$)	0,88	0,03	0,83	0,86	0,88	0,90	0,93	0,0008	1,00
Ikä < 30 ($\mathcal{G}_1$)	4,30	0,43	3,51	3,99	4,28	4,57	5,20	0,0062	1,00
Ikä $< 30 \times$ ikä ($\mathcal{G}_1$)	3,22	0,37	2,55	2,96	3,21	3,47	4,01	0,0054	1,00
Kalenterivuosi ($\mathcal{G}_1$)	0,99	0,01	0,97	0,98	0,99	1,00	1,01	0,0004	1,00
Regressiokertoimet ($\mathcal{G}_2$):									
Nainen ($\mathcal{G}_2$)	0,80	0,01	0,78	0,79	0,80	0,81	0,82	0,0006	1,03
Ikä ($\mathcal{G}_2$)	1,11	0,01	1,09	1,10	1,11	1,11	1,13	0,0012	1,00
Ikä^2 ($\mathcal{G}_2$)	0,95	0,00	0,94	0,95	0,95	0,95	0,96	0,0001	1,00

Jatkuu seuraavalla sivulla

Taulukko 5.15 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Ikä < 30 ($\mathcal{G}_2$)	1,11	0,02	1,07	1,10	1,11	1,12	1,15	0,0007	1,00
Ikä $< 30 \times$ ikä ($\mathcal{G}_2$)	1,13	0,02	1,09	1,12	1,13	1,15	1,18	0,0007	1,00
Kalenterivuosi ($\mathcal{G}_2$)	1,00	0,00	1,00	1,00	1,00	1,00	1,01	0,0002	1,01
Regressiokertoimet (θ):									
Nainen (θ)	2,15	0,28	1,64	1,96	2,14	2,33	2,75	0,0055	1,00
Ikä (θ)	0,18	0,03	0,13	0,16	0,18	0,20	0,24	0,0010	1,00
Ikä^2 (θ)	2,32	0,29	1,80	2,12	2,30	2,51	2,95	0,0104	1,00
Ikä < 30 (θ)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,0000	1,00
Ikä $< 30 \times$ ikä (θ)	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,0000	1,00
Ikä > 55 (θ)	0,02	0,02	0,00	0,01	0,01	0,02	0,07	0,0003	1,00
Ikä $> 55 \times$ ikä (θ)	52,26	42,60	9,35	24,14	40,57	66,36	165,25	0,5666	1,00
Ikä > 60 (θ)	1,22	0,27	0,77	1,02	1,19	1,38	1,82	0,0013	1,00
Kalenterivuosi (θ)	0,99	0,03	0,92	0,96	0,99	1,01	1,05	0,0016	1,00
Muotoparametrit:									
Muotoparametri($\mathcal{G}_1$)	1,31	0,02	1,28	1,30	1,31	1,32	1,34	0,0000	1,00
Muotoparametri($\mathcal{G}_2$)	127,9	2,3	123,4	126,3	127,8	129,4	132,4	0,002	1,00
Satunnaisvaikutusten parametrit:									
λ (H)	5,55	0,23	5,12	5,40	5,53	5,69	6,03	0,0010	1,01
λ ($\mathcal{G}_1$)	0,24	0,01	0,22	0,23	0,24	0,25	0,27	0,0000	1,00
λ ($\mathcal{G}_2$)	0,11	0,00	0,10	0,11	0,11	0,11	0,12	0,0000	1,02
λ (θ)	3,20	0,09	3,01	3,13	3,19	3,26	3,39	0,0001	1,01

Listaan lyhyesti joitakin huomioita keskeisimmistä vaikutuksista ja estimointituloksista taulukossa 5.15. Kun kirjoitan parametrin sijoittuvan jollekin tietylle intervallille, tarkoitan tällöin kyseisen parametrin posteriorijakauman 95%:n uskottavuusintervallia 2,5 %:sta 97,5 %:iin. Käytän lisäksi vedonlyöntisuhteesta sen englanninkieliseen nimitykseen, odds ratio, pohjautuvaa lyhennettä *OR*.

- **Sukupuoli:**
  - **Hurdle (H).** Naisindikaattorin OR on noin 0,7–0,9 eli naisilla nollapalkkasumman odds on noin 0,7–0,9-kertainen miehiin verrattuna. Alle ykkösen oleva OR tarkoittaa, että nollapalkkasumman todennäköisyys on pienempi kuin vertailuluokassa. Mikäli OR on suurempi kuin yksi, niin todennäköisyys on puolestaan suurempi kuin vertailuluokassa.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$).** Naisindikaattorin kerroin ensimmäisessä gamma-jakaumassa on noin 0,84–0,93 miehiin verrattuna. Ensimmäiseen gamma-jakaumaan kuuluvat palkkasummat ovat (luultavasti) henkilöiltä, jotka ovat vain osan vuotta töissä. Naisilla, jotka ovat vain osan vuotta töissä, palkkasummat ovat näin ollen noin 84–93 %:ia miesten palkkasummista, kun mallissa olevat muut tekijät huomioidaan.
  - **Toinen gamma-jakauma ($\mathcal{G}_2$).** Naisindikaattorin kerroin toisessa gamma-jakaumassa on noin 0,8 miehiin verrattuna. Toiseen gamma-jakaumaan kuuluvat palkkasummat ovat (luultavasti) henkilöiltä, jotka ovat koko vuoden töissä. Naisilla, jotka ovat koko vuoden töissä, palkkasummat ovat näin ollen noin 80 %:ia miesten palkkasummista, kun mallissa olevat muut tekijät huomioidaan.
  - **Kuuluminen ensimmäiseen gamma-jakaumaan (θ).** Naisindikaattorin OR on noin 1,6–2,8, eli naisilla yli nollan olevan palkkasumman todennäköisyys kuulua ensimmäiseen gamma-jakaumaan (eikä toiseen gamma-jakaumaan) on suurempi kuin miehillä. Alle ykkösen oleva OR tarkoittaa, että ensimmäiseen gamma-jakaumaan kuulumisen todennäköisyys on pienempi kuin vertailuluokassa. Mikäli ensimmäinen gamma-jakauma vastaa vain osan vuotta työssä olevien palkkasummia (ja toinen gamma-jakauma koko vuoden töissä olevien), niin tämän tuloksen perusteella naisilla on miehiä suurempi todennäköisyys olla vain osan vuotta töissä verrattuna siihen, että olisi koko vuoden töissä, sillä naisindikaattorin odds on suurella todennäköisyydellä ykköstä suurempi.
- **Ikä:**
  - **Hurdle (H).** Iän päävaikutuksen jakauma on ykkösen ympärillä, joten iän päävaikutuksella ei ole ainakaan voimakasta yhteyttä hurdleen. Iän neliö on yli ykkösen, mikä tarkoittaa sitä, että iän neliön vaikutuksesta ikä aluksi laskee todennäköisyyttä saada palkkasumman arvoksi nolla noin 40-vuotiaaksi asti, mutta sen jälkeen nollan todennäköisyys kääntyy nousuun iän neliön vaikutuksesta. Alle 21-vuotiailla on suurempi todennäköisyys saada nollapalkkasumma kuin muilla, samoin alle 30-vuotiailla ja yli 60-vuotiailla.
  - **Ensimmäinen gamma-jakauma ($\mathcal{G}_1$).** Iän päävaikutus on n. 1,25–1,36, eli iän päävaikutus nostaa ensimmäisen gamma-jakauman odotusarvoa. Iän neliön vaikutus on alle ykkösen, mikä tarkoittaa sitä, että iän neliön vaikutuksesta ikä aluksi nostaa palkkasumman odotusarvoa ensimmäisessä gamma-jakaumassa noin 40-vuotiaaksi asti, mutta sen jälkeen odotusarvo kääntyy laskuun iän neliön vaikutuksesta.

- **Toinen gamma-jakauma ($\mathcal{G}_2$)**. Iän päävaikutus on n. 1,1, eli iän päävaikutus nostaa hieman toisen gamma-jakauman odotusarvoa. Iän neliön vaikutus on alle hieman alle ykkösen (n. 0,95), mikä tarkoittaa sitä, että iän neliön vaikutuksesta ikä aluksi nostaa palkkasumman odotusarvoa toisessa gamma-jakaumassa noin 40-vuotiaaksi asti, mutta sen jälkeen odotusarvo kääntyy laskuun iän neliön vaikutuksesta.
- **Toinen gamma-jakauma (θ)**. Iän päävaikutus on n. 0,13–0,24 eli iän päävaikutus laskee todennäköisyyttä kuulua ensimmäiseen gamma-jakaumaan – eli olla vain osan vuodesta töissä. Iän neliön vaikutus on yli ykkösen, mikä tarkoittaa sitä, että iän neliön vaikutuksesta ikä aluksi laskee todennäköisyyttä kuulua ensimmäiseen gamma-jakaumaan noin 40-vuotiaaksi asti, mutta sen jälkeen tämä todennäköisyys kääntyy nousuun iän neliön vaikutuksesta.

- **Kalenterivuosi:**

  Kalenterivuosien osalta tutkimusaineisto on varsin rajoittunut, sillä se koskee vain vuosia 2005–2014. Mallin 4 estimointitulosten perusteella nollahavainnon todennäköisyys on hieman kasvanut vuosina 2005–2014. Muihin palkkasummamallin komponentteihin ($\mathcal{G}_1, \mathcal{G}_2, \theta$) kalenterivuodella ei vaikuta estimointitulosten perusteella olevan ainakaan ”suurta” vaikutusta.

5.4.3.3 Suomi: posterioriennustetarkistukset

Palkkasummien posterioriennusteet on muodostettu yllä esitetyn mallin 4 – eli elinkaariennusteissa käytettävän mallin – parametrien posteriorijakaumien estimaateista ja havaintoaineistosta niin, kuin on esitetty maassaoloaikojen mallinnuksen yhteydessä luvussa 3.2.6. Olen laatinut posterioriennusteet niin, kuin kyse olisi uusien maahanmuuttajien aineistosta, joten simuloin posterioriennusteita varten myös satunnaisvaikutukset .

Posterioriennustetarkistus: Kaikkien palkkasummahavaintojen yhteisjakauma ja henkilöiden palkkasummien keskiarvojen yhteisjakauma.

Aloitan posterioriennustetarkistukset vertaamalla havaintoaineiston kaikkien palkkasummahavaintojen muodostamaa yhteisjakaumaa posterioriennustejakauman kaikkien ennusteiden muodostamaan yhteisjakaumaan. Ensimmäisen mallinnustavoitteen mukaan näiden jakaumien pitäisi vastata mahdollisimman hyvin toisiaan, erityisesti jakaumien tärkeimpien tunnuslukujen kuten keskiarvojen, keskihajontojen ja kvartiilien suhteen. Koska palkkasummat ovat jakautuneet nolliin ja yli nolliin, mallinnustavoitteiden tulisi täyttyä sekä nollien osuuden että ylinollien jakauman osalta.

Kaikkien havaintojen jakauman lisäksi vertaan kaikkien havaintoaineiston henkilöiden palkkasummien keskiarvojen jakaumaa ennustettuihin henkilöiden palkkasummien keskiarvojen jakaumaan. Nämä jakaumat muodostetaan havaintoaineistossa laskemalla palkkasummien keskiarvot kaikille aineiston henkilöille ja muodostamalla näistä jakauma. Posterioriennusteiden kohdalla lasketaan ennustettujen palkkasummien keskiarvot kaikille henkilöille, ja muodostetaan näistä jakauma. Henkilötasoisten jakaumien vertailun tarkoituksena on osoittaa, että elinkaariennusteissa käytettävä nettovaikutusten malli tuottaa

luotettavia ennusteita paitsi havaintotasolla niin myös henkilötasolla.

Taulukossa 5.16 on esitetty yllä kuvatun posterioriennustetarkistuksen tulokset. Ylemmässä taulukon osassa verrataan kaikkien palkkasummahavaintojen jakauman tunnuslukujen estimaatteja vastaaviin posterioriennustejakauman estimaatteihin. Bayes-p-arvot ja Bayes-intervallit eri euromäärille ovat testisuureita estimaattien vastaavuuksille havaintoaineiston ja posterioriennusteiden välillä. Bayes-intervalli ilmaisee todennäköisyyden sille, että posterioriennustejakauman estimaatti poikkeaa korkeintaan ilmoitetun euromäärän verran havaintoaineiston estimaatista. Esimerkiksi 1 000 euron Bayes-intervallin sisään mahtuvat kaikkien havaintojen jakauman osalta yli 50 prosentin todennäköisyydellä kaikki jakaumapisteet lukuun ottamatta 99. prosenttipistettä, jossa tämä todennäköisyys on 0,26. Bayes-p-arvo ja Bayes-intervallit on määritetty tarkemmin työikäisten nettovaikutusmallinnuksen yhteydessä (luvussa 4.3.7).

Taulukon ylemmästä osasta käy ilmi, että kaikkien havaintojen ja posterioriennusteiden jakaumat vastaavat hyvin toisiaan niin jakaumien havaittujen keskiarvojen, keskihajontojen sekä nollaosuuksien osalta. Myös posterioriennusteiden ylinollien keskiarvo ja keskihajonta vastaavat hyvin havaintoaineiston vastaavia jakauman tunnuslukuja, sillä ero posterioriennusteiden estimaattien ja havaintoestimaattien välillä on ylinollien keskiarvon osalta noin 200 euroa ja keskihajonnan osalta vain noin 60 euroa.

Taulukon 5.16 alemmassa osassa on verrattu henkilöittäisiä ennustettujen palkkasummien keskiarvoja havaintoaineiston henkilöittäisten keskiarvojen jakaumaan. Yleinen johtopäätelmä tästä vertailusta on, että ennustejakauma ja havaintoaineiston jakauma vastaavat toisiaan hyvin keskiarvon osalta, nollien osuuksien osalta sekä prosenttipisteiden osalta 75 %:iin asti. Jakauman oikean reunan osalta kuitenkin havaintojen prosenttipisteet ovat suurempia kuin posterioriennusteiden pisteet. Myös keskihajonta on henkilöiden keskiarvojen osalta suurempi kuin posterioriennusteiden henkilöiden keskiarvojen jakaumassa. Näiden tulosten perusteella arvioisin, että käytetyt henkilötasoiset gamma-jakautuneet satunnaisvaikutukset eivät vastaa kovin täsmällisesti havaintoaineistoa. Palkkasummien mallia voisi siten kehittää ainakin kokeilemalla vielä erilaisia satunnaisvaikutusten jakaumia.

Taulukko 5.16: *Palkkasummien havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Kaikkien ennusteiden jakauma ja ennustettujen henkilötasoisten keskiarvojen jakauma. Jakaumien keskiarvot (KA), keskihajonnat (KH), prosenttipisteet:* $1, 5, 10, 25, 50, 75, 90, 95, 99$*, ylinollien keskiarvot (KA(yn)), ylinollien keskihajonnat (KH(yn)) ja nollien osuudet prosentteina* ($0-\%$)*. Vertailtavien estimaattien Bayes-p-arvot ja Bayes-intervallit eri euromäärillä.*

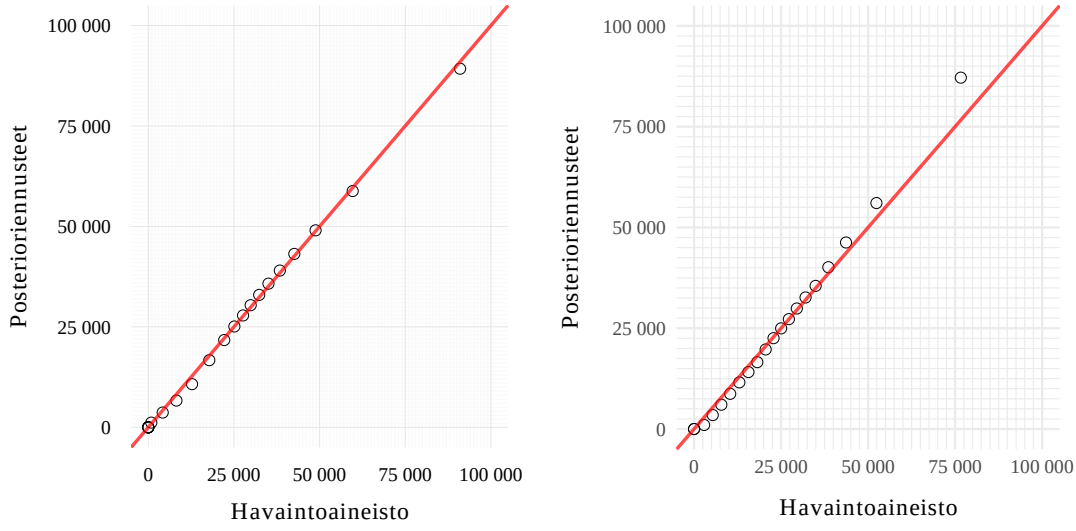
Kaikki havainnot ja ennusteet	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %	$KA(yn)$	$KH(yn)$	$0-\%$
Posterioriennusteet	22 560	22 380	0	0	0	1 180	21 730	35 760	49 040	58 810	89 260	29 310	21 290	23,0
Aineiston havainnot	22 790	22 480	0	0	0	890	22 170	35 080	48 830	59 660	91 010	29 500	21 350	22,7
Bayes-p-arvo	0,02	0,29	0	0	0	1	0	1	0,79	0,01	0,09	0,08	0,30	0,96
Bayes-intervalli: 1 000 euroa	1	0,96	1	1	1	1	1	0,98	1	0,65	0,26	1	0,93	–
Bayes-intervalli: 500 euroa	1	0,77	1	1	1	0,99	0,62	0,09	0,84	0,17	0,12	1	0,65	–
Bayes-intervalli: 200 euroa	0,37	0,31	1	1	1	0,20	0,09	0	0,42	0,05	0,06	0,55	0,25	–

Henkilöiden keskiarvot	KA	KH	1 %	5 %	10 %	25 %	50 %	75 %	90 %	95 %	99 %	$0-\%$
Posterioriennusteet	22 030	17 860	0	0	0	7 840	20 500	32 010	43 600	52 310	76 140	10,4
Aineiston henkilöt	22 210	20 270	0	0	0	6 030	19 740	32 640	46 270	56 050	87 170	10,6
Bayes-p-arvo	0,04	0	0	0	0,11	1	1	0	0	0	0	0,13
Bayes-intervalli: 1 000 euroa	1	0,01	1	1	1	0	0,90	0,96	0	0	0	–
Bayes-intervalli: 500 euroa	1	0	1	1	1	0	0,09	0,25	0	0	0	–
Bayes-intervalli: 200 euroa	0,54	0	1	1	0,99	0	0	0,01	0	0	0	–

Vertaan alla kuviossa 5.24 havaintoaineiston jakaumien prosenttipisteitä vastaavien posterioriennustejakaumien prosenttipisteisiin. Jakaumissa ovat mukana vain yli nollan olevan havainnot sekä ennusteet, ”ylinollat”. Koska nollien osuuksien vastaavuus havaintoaineiston ja posterioriennusteiden välillä on jo käsitelty yllä taulukossa 5.16, en käsittele nollaosuuksia tarkemmin enää. Ylinollien jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeamat vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).

Kaikkien havaintojen ja ennusteiden ylinollien jakaumat vastaavat toisiaan kuvan 5.24 vasemman osakuvan perusteella. Kuten jo edellä kirjoitin, henkilöiden keskiarvojen osalta posterioriennusteiden jakauma vastaa melko huonosti havaintoaineiston jakaumaa jakauman oikean hännän osalta, kuten viitatun kuvan oikea osakuva osoittaa.

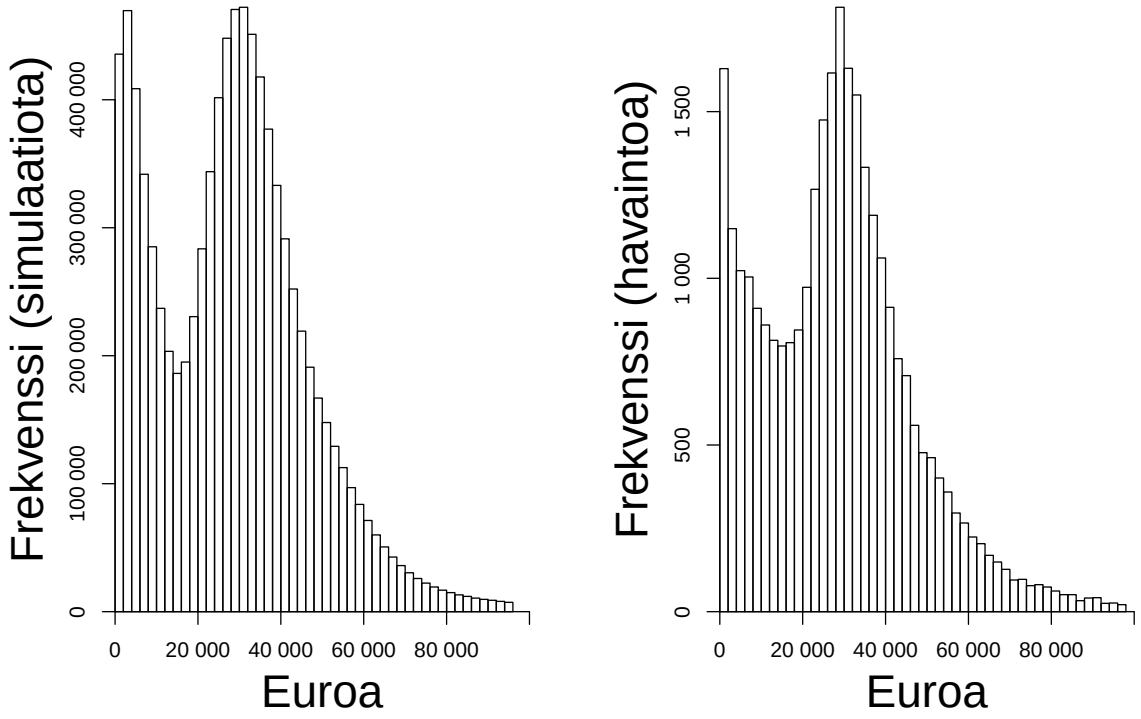


(a) *Kaikki havainnot ja ennusteet.*

(b) *Henkilöiden keskiarvot.*

Kuva 5.24: *Vasemmassa osakuvassa havaintoaineiston yli nollan olevien havaintojen jakauman ja posterioriennusteiden yli nollan olevien ennusteiden jakauman vertailu. ”Ylinollien” jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Oikeassa osakuvassa havaintoaineiston henkilöiden palkkasummahavaintojen keskiarvojen jakauman ja posterioriennusteiden vastaavan jakauman vertailu. Jakaumien prosenttipisteet:* $5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*

Verrataan vielä posterioriennusteiden ja havaintoaineiston ylinollien jakaumia prosenttipisteiden vertailukuvioiden (Q-Q plot) lisäksi histogrammien avulla. Nämä on esitetty kuvassa 5.25. Vasemmalla oleva histogrammi esittää palkkasummien yli nollan olevien posterioriennusteiden jakaumaa. Oikealla oleva histogrammi esittää puolestaan palkkasummien yli nollan olevien havaintojen jakaumaa. Kummastakin jakaumasta on katkaistu pois ylin prosentti yksityisyydensuojan varmistamiseksi. Myös ylinollia koskevat histogrammit osoittavat, että kaikkien havaintojen ja posterioriennusteiden jakaumat muistuttavat toisiaan.



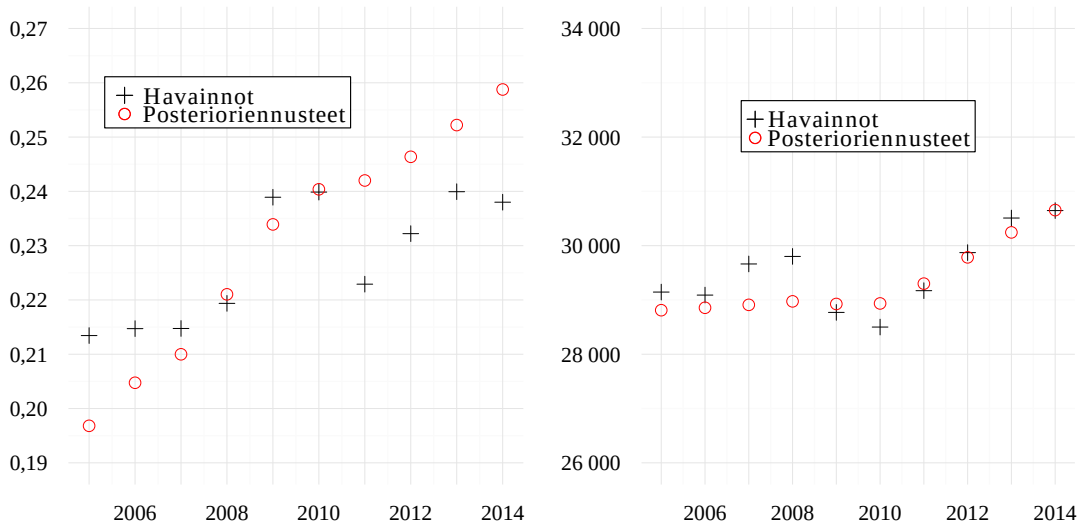
Kuva 5.25: *Suomessa syntyneet 18–62-vuotiaat. Nettovaikutusten lopullisen mallin posterioriennusteiden katkaistujen jakaumien histogrammit: kaikki ennustetut havainnot ja ennustetetut henkilötasoiset keskiarvot (ylärivi).*

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan

Vertaan nyt posterioriennusteita havaintoaineistoon selittävien muuttujien luokkien mukaan. Selittäviä muuttujia ovat sukupuoli, ikä ja kalenterivuosi.

Vertaan posterioriennusteiden nollaosuuksia sekä ”ylinollien” keskiarvoja selittävien luokkien mukaan vastaaviin havaintoaineiston nollaosuuksiin ja ”ylinollien” keskiarvoihin. Koska palkkasummat koostuvat nollista sekä yli nollan olevasta osasta, esitän nollaosuudet ja ylinollien keskiarvot erikseen. Esimerkiksi iän kohdalla vertaan posterioriennusteiden nollaosuuksia havaintoaineiston nollaosuuksiin niin, että lasken sekä posterioriennusteista että havaintoaineistosta jokaiselle iälle erikseen nollien osuuden 18-vuotiaista 62-vuotiaisiin. Ylinollien osalta lasken puolestaan erikseen jokaiselle iälle havaintojen ja ennusteiden ylinollien keskiarvot.

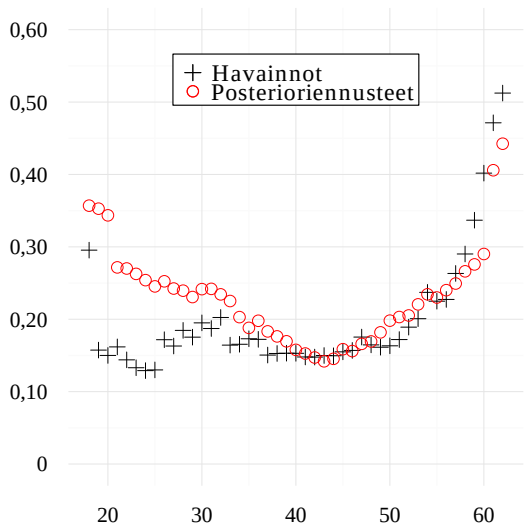
Posterioriennustetarkistus kalenterivuoden mukaan.



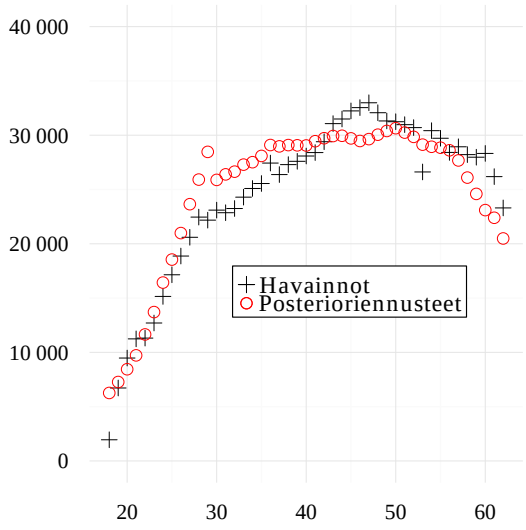
(a) *Nollien osuudet kalenterivuoden mukaan.* (b) *Ylinollien keskiarvot kalenterivuoden mukaan.*

Kuva 5.26: Havainnot ja posterioriennusteet kalenterivuoden mukaan.

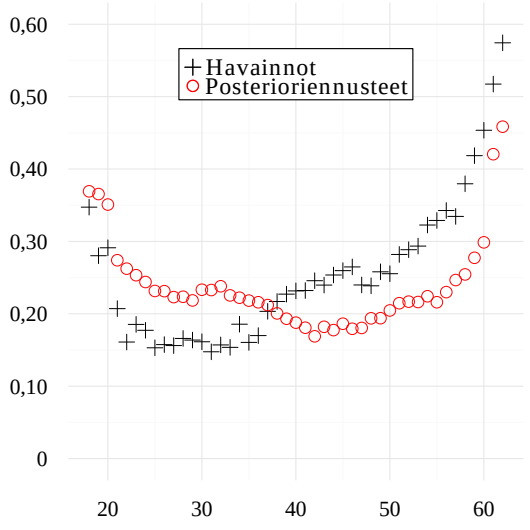
Posterioriennustetarkistus iän ja sukupuolen mukaan.



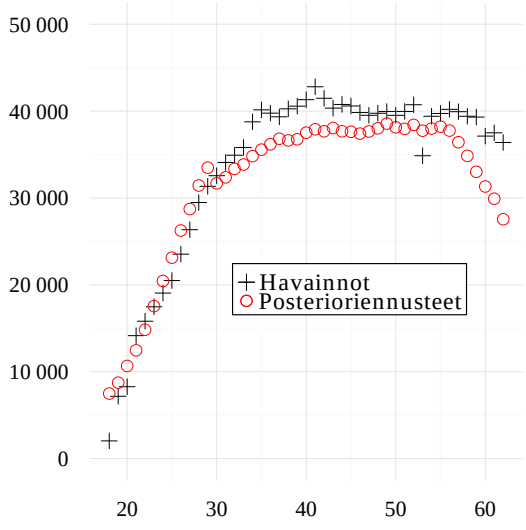
(a) *Nollien osuudet iän mukaan. Naiset.*



(b) *Ylinollien keskiarvot iän mukaan. Naiset.*



(c) *Nollien osuudet iän mukaan. Miehet.*



(d) *Ylinollien keskiarvot iän mukaan. Miehet.*

Kuva 5.27: Havainnot ja posterioriennusteet iän ja sukupuolen mukaan.

Posterioriennustarkistukset selittävien muuttujien luokkien mukaan. Havaintoja kuvista 5.4.3.3–5.26.

Kuvan naisia koskevan osakuvion perusteella naisten palkkasummien nollaosuuksien ennusteet vastaavat havaintoaineiston ikävuosittaisia havaintoja melko hyvin – lukuun ottamatta alle 30:n olevia ikävuosia. Ylinollien osalta posterioriennusteiden ikävuosittaiset keskiarvot poikkeavat eniten havaintoaineiston vastaavista keskiarvoista 30 ikävuoden ympärillä. Miesten osalta ennustetut nollaosuudet ovat liian suuria vajaaseen 40 ikävuoteen asti, tämän jälkeen ennustetut nollaosuudet ovat puolestaan liian pieniä suhteessa havaintoaineistoon. Palkkasummien mallia voisi parantaa jatkossa lisäämällä erityisesti hurdlea koskien malliin iän (tai ikäluokkaindikaattoreiden) ja sukupuolen välisiä interaktioita. Ylinollien osalta ikävuosittaiset posterioriennusteet vastaavat melko hyvin havaintoaineiston vastaavia keskiarvoja – joskin ennusteiden ikävuosittaiset keskiarvot ovat havaittuja pienempiä erityisesti 40 ikävuoden ja 60 ikävuoden kohdalla.

Kuvan 5.26 vasemmanpuoleisen osakuvan perusteella nollapalkkasummien osuus oli 21–22 %:ia vuosina 2005–2008, mutta vuosina 2009–2014 nollien osuuksien taso nousi 22–24 %:iin. Koska palkkasummien mallissa on mukana vain kalenterivuoden päävaikutus nollapalkkasummiin (hurdleen), tämä näkyy siinä, että kuvassa 5.26 on posterioriennusteiden nouseva trendi kalenterivuoden suhteen. Palkkasummien mallia saisi nollaosuuksien osalta parannettua helposti niin, että lisäisi malliin kalenterivuosiluokkaindikaattorin (onko kalenterivuosi suurempi tai yhtä suuri kuin vuosi 2009).

Kuvan 5.26 oikeanpuoleisen osakuvan perusteella ylinollissa voisi olla nouseva trendi vuodesta 2010 lähtien, mikä näkyy myös posterioriennusteiden ylinollien keskiarvoissa vuoden 2010 jälkeen. Edellä esitetyn lopullisen mallin estimaatteja koskevan taulukon 5.15 mukaan kalenterivuodella ei vaikututtaisi olevan (kalenterivuosia koskevien regressiokertoimien odotusarvoestimaattien mukaan) vaikutusta palkkasummien ylinollien keskiarvoon. Kuvan 5.26 oikeanpuoleisen osakuvan nouseva trendi johtuneekin lähinnä keski-iän noususta 18–62-vuotiaiden Suomessa syntyneen väestön keskuudessa. Koska keskimäärin kaikkein pienimmät palkkasummat sijoittuvat nuorempiin ikävuosiin, keski-iän nousu vähentää näiden nuorempien ikävuosien osuutta 18–62-vuotiaita koskevista havainnoista. Samalla 18–62-vuotiaiden ylinollien keskiarvot nousevat.

Luku 6

Muut ennusteparametrit ja eläkkeisiin liittyvien nettovaikutusten ennustaminen

6.1 Muut ennusteparametrit

Tämän tutkimustekstin johdannossa (luvussa 2.1) on määritetty julkisen talouden elinkaarivaikutusten komponentit sekä elinkaarivaikutusten funktio. Elinkaarivaikutusten komponentteja on seitsemän kappaletta, ja elinkaarivaikutusten funktion avulla näistä komponenteista muodostetaan julkisen talouden elinkaarivaikutukset. Näiden komponenttien *ennustejakaumista* saadaan vastaavasti muodostettua elinkaarivaikutusten *ennusteiden* todennäköisyysjakaumat.

Yksi seitsemästä elinkaarivaikutusten komponentista ovat ”muut parametrit”, joiden vektorista (listasta) käytän merkintää $\boldsymbol{\pi}$. Elinkaarivaikutusten ennusteiden yhteydessä käytän näistä parametreista nimitystä ”muut ennusteparametrit”. Kuten johdannossa todetaan, muut ennusteparametrit $\boldsymbol{\pi}$ ovat parametreja, joiden avulla elinkaarivaikutusten ennusteet muodostetaan muista elinkaarivaikutusten komponenteista elinkaarivaikutusten funktion avulla. Käytännössä tämä tarkoittaa elinkaariennusteiden (mahdollista) diskonttaamista diskonttauskertoimien (joka muodostaa yhden $\boldsymbol{\pi}$:n komponentin) avulla sekä eläkkeisiin liittyvien nettovaikutusten ennusteiden laskemista palkkasummien ennusteista.

Kuten johdannossa esitetään (luvussa 2.2), $\boldsymbol{\pi}$ on satunnaisvektori samoin kuten elinkaariennusteiden komponentit $\boldsymbol{y}_{uusi}$ (sis. esim. työiän ennustetut nettovaikutukset ja ennustetut palkkasummat) sekä näiden komponenttien malliparametrit $\boldsymbol{\theta}$. Satunnaisvektori $\boldsymbol{\pi}$ koostuu muiden ennusteparametrien listasta, jonka yksittäiset komponentit ovat (suurimmaksi osaksi) satunnaismuuttujia. Ne komponentit, jotka eivät ole satunnaismuuttujia, ovat vakioita. Jokaiseen ennusteparametriin, joka on satunnaismuuttuja, liittyy myös todennäköisyysjakauma, joka kuvaa tuohon parametriin liittyviä epävarmuuksia. Nämä epävarmuudet sisältyvät siten julkisen talouden nettovaikutusten elinkaariennusteisiin niin, kuin on esitetty johdannon luvussa 2.3.4.

Esitän tässä luvussa muiden ennusteparametrien elinkaariennusteissa käytetyt todennäköisyysjakaumat, sekä kuinka niihin on päädytty. Tavoitteena on, että nämä todennäköisyysjakaumat sisältävät ainakin

”realistisen” määrän näihin parametreihin liittyviä epävarmuuksia ennusteissa. Vakioparametrien osalta esitän perustelut niiden saamille vakioarvoille. Yleisesti ottaen elinkaariennusteita voisi vielä täydentää määrittelemällä vakioparametreille myös todennäköisyysjakaumat.

Muiden ennusteparametrien π posteriorijakauma

Lausutaan nyt Bayesin kaavaa (teoreemaa) seuraten muiden ennusteparametrien $\boldsymbol{\pi}$ niitä koskevalla havaintoaineistolla $\boldsymbol{d}$ ehdollinen todennäköisyysjakauma eli posteriorijakauma. Se on verrannollinen $\boldsymbol{\pi}$:n otosjakauman ($p(\boldsymbol{d}|\boldsymbol{\pi})$) ja priorijakauman väliselle tulolle:

$$p(\boldsymbol{\pi}|\boldsymbol{d}) \propto p(\boldsymbol{d}|\boldsymbol{\pi})p(\boldsymbol{\pi}). \tag{6.1}$$

Tämän luvun tavoitteena on määrittää $\boldsymbol{\pi}$:n posteriorijakauma. Osa $\boldsymbol{\pi}$:n sisältämistä yksittäisistä parametreista on joko muiden parametrien deterministisiä funktioita tai niiden todennäköisyysjakauma on ehdollinen muilla parametreilla. Muuten oletan parametrien olevan toisistaan riippumattomia, joten voin määrittää niiden yhteisen posteriorijakauman, $p(\boldsymbol{\pi}|\boldsymbol{d})$, komponentti (parametri) kerrallaan. Priorijakauma $p(\boldsymbol{\pi})$ summaa sen, mitä parametreista tiedetään havaintoaineiston $\boldsymbol{d}$ ulkopuolelta, ja otosjakauma $p(\boldsymbol{d}|\boldsymbol{\pi})$ puolestaan päivittää tämän tiedon havaintoaineiston tarjoamalla informaatiolla. Parametreja koskeva havaintoaineisto $\boldsymbol{d}$ koostuu lähinnä näiden parametrien $\boldsymbol{\pi}$ menneistä, havaituista arvoista. Uusien maahanmuuttajien elinkaariennusteissa määritetystä posteriorijakaumasta, $p(\boldsymbol{\pi}|\boldsymbol{d})$, poimitaan simuloituja arvoja.

Alla on lista $\boldsymbol{\pi}$:n komponenteista eli yksittäisistä ”muista ennusteparametreista” elinkaarivaikutusten ennusteissa. Olen jakanut yksittäiset parametrit listauksessa viiteen eri joukkoon. Jotkut listauksen parametrit sisältävät todellisuudessa useamman parametrin. Kirjoitan nämä auki näitä parametreja koskevan esittelyn yhteydessä. Olen myös listauksessa kirjoittanut niiden parametrien nimen jälkeen tekstin ”*kiinteitä*”, jotka ainakin vielä tässä tutkimuksen versiossa saavat vain kiinteitä arvoja, eli niille ei ole määritetty omia todennäköisyysjakaumia.

Muiden ennusteparametrien (π) lista:

1. Diskonttauskertoimet ja niillä ehdolliset parametrit:
   - Diskonttauskertoimet
   - Reaalikorko
   - Palkkojen reaalisen tason muutokset
2. Työnantajien maksujen osuudet palkoista:
   - Lakisääteisten työeläkemaksujen osuus palkoista
   - Sosiaalivakuutusmaksujen osuus palkoista
3. Eläkkeiden karttumisen osuus palkoista (*kiinteitä*)

4. Maksettavien kansaneläkkeiden määrä (eläkkeiden raja-arvot):
    - Maksettava täysi kansaneläke (*kiinteä*)
    - Maksettavan takuueläkkeen raja-arvo (*kiinteä*)
5. Ennustetuista eläkeajan tuloista maksettavat verot (veroprosentit):
    - Ansiotuloveroprosenttien tulorajat ja veroprosentit (*kiinteitä*)
    - Pääomatuloveroprosentti ja ennustetut pääomatulot eläkeiässä (*kiinteitä*)
    - Välillisten verojen tulorajat ja veroprosentit (tulorajat ja niitä vastaavat veroprosentit *kiinteitä*)

Esitän seuraavaksi muiden ennusteparametrien todennäköisyysjakaumien – tai vakioiden – määritelmät ja niiden perusteet. Noudatan tässä esitetyn listauksen järjestystä.

6.1.1 Diskonttauskertoimet ja niillä ehdolliset parametrit

6.1.1.1 Diskonttauskertoimet

Diskonttauskertoimia käytetään ennustettujen vuosittaisten elinkaarivaikutusten ennusteiden diskonttaamiseen nykyarvoon (tai ennusteen tekohetken arvoon). Elinkaariennusteiden diskonttausta tarvitaan erityisesti tämän tutkimustekstin III osassa, jossa uusia maahanmuuttajia koskevien elinkaariennusteiden tuloksia sovelletaan uusien maahanmuuttajien valintaan päätöksentekoteorian mukaisesti. Koska diskonttauskertoimia sovelletaan erityisesti tutkimuksen osassa III, olen myös sijoittanut diskonttauskertoimien mallinnusta käsittelevän luvun tutkimuksen III osaan (lukuun 11). Vuosittaiset diskonttauskertoimet ovat osamääriä: $\frac{1+inflaatio}{1+nimelliskorko}$, missä inflaatio ja nimelliskorko saavat diskonttauskertoimen vuotta koskevat arvot.

6.1.1.2 Reaalikoron vuosittaiset muutokset

Käsitän tässä tutkimuksessa reaalikoron nimelliskoron ja inflaation välisenä erotuksena eli $reaalikorko = nimelliskorko - inflaatio$. Yllä diskonttauskerroin määritettiin osamääränä $\frac{1+inflaatio}{1+nimelliskorko}$. Tästä seuraa, että reaalikorko on diskonttauskertoimen (lähes) deterministinen funktio, sillä kun diskonttauskertoimen (DK) määritelmästä otetaan puolittain logaritmi:

$log(diskonttauskerroin) = log(1 + inflaatio) - log(1 + nimelliskorko)$,

mistä saadaan likiarvo:

$diskonttauskerroin - 1 \approx inflaatio - nimelliskorko$,

mistä saadaan reaalikoron määritelmää noudattaen:

$reaalikorko \approx -diskonttauskerroin + 1$.

Tässä likiarvo pätee, kun $(diskonttauskerroin - 1)$, $inflaatio$ ja $nimelliskorko$ ovat kaikki ”pieniä” eli ”lähellä” nollaa. Diskonttauskertoimia koskevan aikasarjahavaintoaineiston (ks. luku 11.2.3) perusteella mainittu likiarvo pätee. Lisäksi samasta havaintoaineistosta voidaan todeta, että viimeisen 50 vuoden aikana (vuosina 1966–2015) likiarvo $reaalikorko \approx -diskonttauskerroin + 1$ oltaisiin voitu korvata yhtäsuuruusmerkillä ilman suurta virhettä. Havaitun reaalikoron ja $(-diskonttauskerroin + 1)$:n välinen korrelaatio oli tällä aikavälillä suurempi kuin +0,999, ja näiden suureiden välisen erotuksen suuruuden keskiarvo oli noin 0,0005 eli 0,05 prosenttiyksikköä.

Yllä esitetyn johdosta muodostan elinkaariennusteissa vuosittaiset reaalikorkojen ennusteet diskonttauskertoimien ennusteista siten että:

$$reaalikorkoennuste = -diskonttauskerroinennuste + 1. \tag{6.2}$$

Elinkaariennusteissa poistin reaalikorkoennusteiden kumulatiivisen kehityksen (potenssi)sarjoista ne ennustesarjat, jotka tuottivat 50 vuoden päähän kaikkein äärimmäiset ennusteet reaalikorolle. Poistin käytettävistä ennusteista ne ennustesarjat, joiden ennustamat reaalikorot 50 vuoden päässä olivat joko pienempiä kuin kaikkien ennusteiden 5. prosenttipiste tai suurempia kuin 95. prosenttipiste.

6.1.1.3 Palkkojen reaalisen tason vuosimuutokset

Palkkojen reaalisella tasolla tarkoitan vuosittaista palkkojen tasoa, joista on poistettu inflaation (kuluttajahintojen nousu). Palkkojen vuosittaiset nimelliset tasot poimin Tilastokeskuksen ansiotasoindeksistä (ATI). Vuosia 1948–2009 koskevat ansiotasoindeksin tiedot löysin ja poimin teoksesta (Turkkila (2011), Tilastoliite 6c., s. 259), jossa tietojen lähteenä on: Tilastokeskus, hinnat ja palkat. Vuosia 2010–2015 koskevat tiedot ovat Tilastokeskuksen sivuilta (SVT: *Ansiotasoindeksi*). Elinkustannushintojen vaikutuksen poistamiseksi käytin Tilastokeskuksen rahanarvonkerrointa (SVT: *Rahanarvonkerroin*). Tämä tarkoittaa sitä, että muunsin vuosittaiset ansiotasot vuoden 2011 euroiksi rahanarvonkertoimen avulla. Tämän jälkeen vielä laskin näin muodostetuista palkkojen reaalisista vuosittaisista tasoista palkkojen reaalisen tason vuosittaiset muutokset vuosina 1949–2015. Koska vuosina 1949–1951 on muusta havaintoaineistosta poikkeavien korkeita arvoja, jätin nämä pois palkkojen reaalisen tason vuosimuutosten aineistosta. Mikäli olisin sisällyttänyt myös nämä havainnot aineistoon, ennustetuista palkkojen tason reaalisista vuosimuutoksista olisi tullut hieman korkeampia.

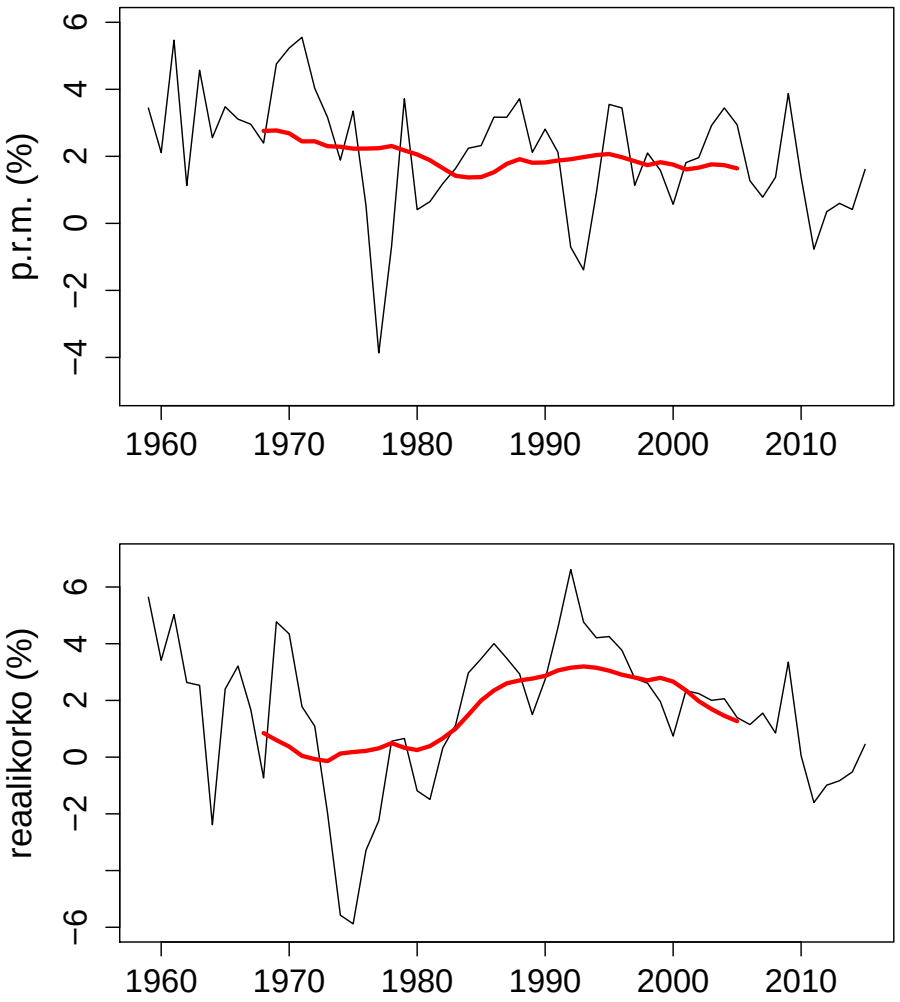
Taulukko 6.1: *Palkkojen reaalisen tason muutokset (%) vuosina 1952–2015. Muutosten jakauman keskiarvo (KA), keskihajonta (KH) ja prosenttipisteet.*

KA	KH	5 %	25 %	50 %	75 %	95 %
2,10	1,77	-0,72	1,13	2,12	3,35	4,85

Taulukossa 6.1 on esitetty palkkojen reaalisten vuosittaisten muutosten aikasarjan jakauman keskeiset tunnusluvut vuosille 1959–2011. Reaalipalkkojen keskimääräinen muutos vuosina 1959–2011 oli 2,10 prosenttia vuodessa, ja keskihajonta oli 1,77 prosenttiyksikköä. Reaalipalkkojen muutokset ovat olleet

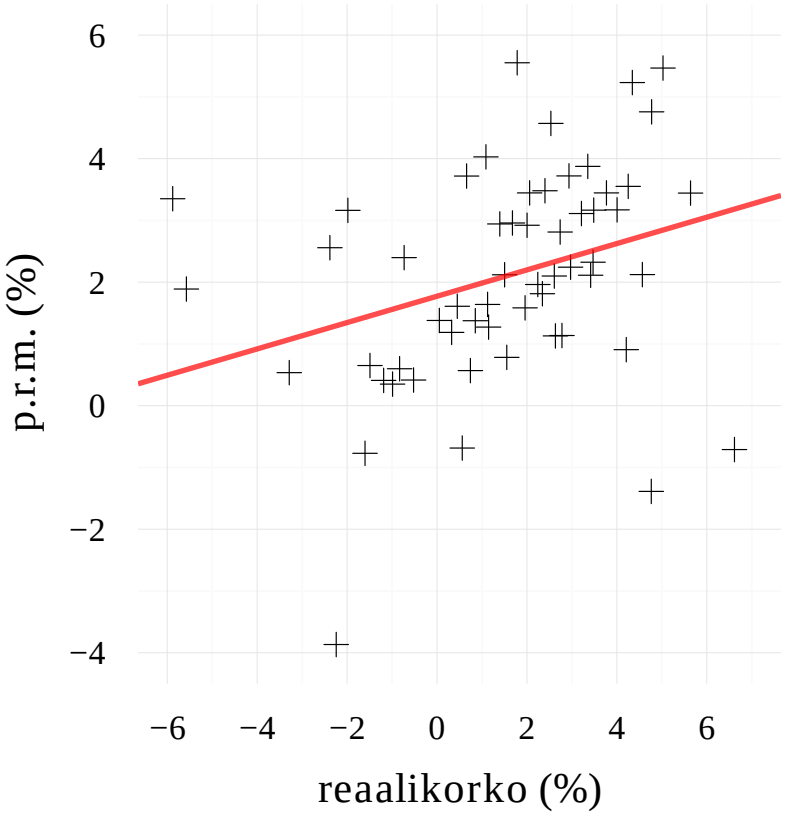
viimeisen 20 vuoden aikana mainitun ajanjakson keskiarvoa pienempiä, sillä vuosina 1996–2015 muutos oli keskimäärin 1,64 %:ia ja 2006–2015 muutos oli keskimäärin 1,09 %:ia. Tosin vuosien 1959–2011 aikana on ollut muitakin lyhyitä ajanjaksoja, jolloin palkkojen reaalinen kehitys on ollut useita vuosia keskiarvon alapuolella (esimerkiksi 1990-luvun alku), kuten selviää alla kuvan 6.1 ylimmästä osakuvasta. Kuvan perusteella reaalipalkkojen kehityksessä vaikuttaa olevan joka tapauksessa ”pieni” laskeva trendi ajassa.

Kuvassa 6.1 on esitetty vuosittaiset palkkojen reaalisen tason muutokset (p.r.m.) ja reaalikoron vuosittaiset arvot vuosina 1959–2015.



Kuva 6.1: *Palkkojen reaalisen tason vuosittaiset muutokset (p.r.m.) vuosina 1959–2015 (ylempi kuva). Vuosittaiset reaalikoron tasot vuosina 1959–2015 (alempi kuva). Punaiset viivat ovat liukuvia keskiarvoja 10 vuotta menneisyyteen ja tulevaisuuteen.*

Vuosittaisilla palkkojen reaalisilla muutoksilla (p.r.m.) ja vuosittaisilla reaalikoron arvoilla vaikuttaa olevan aikasarjan tietojen perusteella varsin vahva positiivinen lineaarinen yhteys. Tämä käy ilmi p.r.m:n ja reaalikoron välisestä hajontakuviosta (kuva 6.2), johon on myös lisätty reaalipalkkojen ja p.r.m:n välinen lineaarinen sovite. Havaintoaineistossa korrelaatio näiden välillä on +0,32, ja se pysyy +0,30:n ja +0,40:n välillä vaikka suurimmat poikkeavat havainnot otettaisiin pois tarkastelusta. Oletan tässä yhteydessä – ilman perusteluja – että p.r.m:n ja reaalipalkkojen välinen yhteys syntyy reaalikoron vaikutuksesta p.r.m:ään (eikä päinvastoin). Tämä tarkoittaa sitä, että edellä esitettyjä reaalikorkojen vuosittaisia ennusteita voidaan käyttää palkkojen reaalisten muutosten ennustamiseen. Palkkojen reaalisilla muutoksilla on täten ehdollinen todennäköisyysjakauma, joka on ehdollinen reaalikoron muutoksella.



Kuva 6.2: *Palkkojen reaalisen tason vuosittaisten muutosten (p.r.m.) ja vuosittaisten reaalikoron tasojen välinen hajontakuvio vuosina 1959–2015. Punainen viiva on lineaarinen sovite reaalikoron vuosittaisten tasojen ja p.r.m:n välillä.*

Koska palkkojen reaalisten muutosten (p.r.m.) jakauma vaikuttaa jakauman histogrammin perusteella (ei esitetty tässä) olevan likimain normaalisti jakautunut, laadin tälle p.r.m:lle yksinkertaisen lineaarisen mallin, jossa vakioparametrin lisäksi on kaksi selittävää muuttujaa, joiden yhteys prm:ään on esitetty yllä: kalenterivuosi ja vuosittainen reaalikorko. Malli on hyvin yksinkertainen myös siinä mielessä, että en sisällytä siihen p.r.m:n tai reaalikoron viivästyneiden arvojen vaikutuksia. Saatan kehittää tätä

mallia tutkimuksen myöhemmässä versiossa, jossa voin esimerkiksi tutkia p.r.m:n ja reaalikoron välisen mahdollisen syy–seuraus-suhteen suuntaa.

Tämän yksinkertaisen mallin estimaattien mukaan palkkojen reaalinen muutos (p.r.m.) on yhtenä vuotena:

$$p.r.m.(vuosi) = 2{,}83 + 0{,}204 \times reaalikorko - 0{,}037 \times (vuosi - 1959) + 0{,}4 + N(0; 1{,}59). \tag{6.3}$$

missä viimeinen termi on normaalijakautunut virhetermi, jonka keskihajonta on 1,59 %-yksikköä. Vuosi on kalenterivuosi. Vakioparametrin estimaatin (2,83) tulkinta on, että mallin mukaan p.r.m:n odotusarvo oli 2,83 %:ia vuonna 1959, kun reaalikorko kiinnitetään nollaan. Reaalikoron regressiokerroin (0,204) tarkoittaa puolestaan sitä, että p.r.m:n odotusarvo kasvaa 0,204:llä prosenttiyksiköllä, kun reaalikorko kasvaa yhden prosenttiyksikön. Kalenterivuoteen liittyvä regressiokerroin kertoo, että p.r.m:n vuosittainen odotusarvo on laskenut vuodesta 1959 vuoteen 2015 0,037:llä prosenttiyksiköllä vuosittain eli p.r.m:n odotusarvo on laskenut yhdellä prosenttiyksiköllä noin 27 vuodessa. Tämä näkyy edellä kuvassa 6.1 p.r.m:n laskevana trendinä.

Vaikka p.r.m:llä on havaintoaineistoon perustuva estimointitulosten perusteella laskeva trendi, niin p.r.m:n ennusteissa oletan, että tämä trendi ei jatku vuoden 2015 jälkeen. Tämä johtuu siitä, että esitellyillä estimointituloksilla – ja käytettävillä reaalikoron ennusteilla – ennustettujen p.r.m:ien odotusarvo on noin 1,0 %:ia. Tämä on myös pienempi luku, kuin mitä eläketurvakeskus (ETK) käyttää pitkän aikavälin lakisääteisten eläkkeiden ennusteissaan. ETK:n ennusteessa reaalikasvuoletus pitkällä aikavälillä on 1,4 prosenttia vuodessa (Tikanmäki et al. 2016, s. 28). Tästä syystä en oleta p.r.m:n odotusarvon laskevan enää 1,0 %:n tasosta. Sen sijaan lisäsin elinkaariennusteissa p.r.m:n odotusarvoon 0,4 %-yksikköä, jotta p.r.m:n ennusteiden keskiarvo vastaisi ETK:n käyttämää oletusta. Yllä olevassa p.r.m:n kaavassa oleva 0,4 %-yksikköä on mukana tästä syystä.

Käytin tässä yhteydessä p.r.m:lle frekventistä mallia eli estimoidut parametrit (vakio, regressiokertoimet ja virhetermin keskihajonta) ovat piste-estimaatteja. Esitettyä mallia voisi vielä kehittää siirtymällä bayesilaisen päättelyn piiriin, jolloin mainittuihin parametreihin vielä estimoinnin jälkeen sisältyvä epävarmuus siirtyisi myös elinkaariennusteisiin. Samassa yhteydessä olisi mahdollisuus lisätä malliin aineiston ulkopuolista informaatiota esimerkiksi sitä koskien, onko palkkojen reaalisella muutoksella laskevaa trendiä tulevaisuudessa – eli mikä on palkkojen reaalisen muutoksen (ajassa muuttuvan) odotusarvon todennäköisyysjakauma.

6.1.2 Työnantajien maksujen osuudet palkoista

Lakisääteisten työeläkemaksujen osuus palkoista

Vuosittaiset työnantajien eläkemaksut laskin palkansaajien eläkemaksuista käyttämällä vuosittain määräytyvien eläkemaksujen palkansaajien keskimääräisiä osuuksia (huomioiden palkansaajan iän) sekä vuosittain määräytyviä työnantajien keskimääräisiä osuuksia palkasta. Koska työnantajien maksun keskimääräinen osuus palkasta vaihtelee eläkelain mukaan, laskin työnantajien maksuosuudelle painotetun

keskiarvon yli eri eläkelakien. Painoina käytin eläkelakikohtaisia maksutuloja eri vuosina. Koska suurin osa eläkemaksuista karttuu TyEL:n ($\sim$ yksityinen sektori) VaEL:n ($\sim$ valtio) ja KuEL:n ($\sim$ kunnat) piiristä, käytin työnantajien keskimääräisen osuuksien laskemiseen yli eri eläkelakien, näiden kolmen eläkelain maksutietoja. Vuosille 2005—2007 käytin vuoden 2008 kerrointa. (ETK 2009–2012).

Lakisääteisten työeläkemaksujen osuus palkoista koostuu työnantajien ja palkansaajien maksuosuuksista. Käsittelen aluksi näiden osuuksien yhteenlaskettua osuutta palkoista, ennen kuin määritän erikseen työnantajien osuuden. Palkansaajien ennustetut työeläkemaksut sisältyvät jo työikäisten nettovaikutusennusteisiin, joten en käsittele palkansaajien osuuden ennustamista tarkemmin. Tavoitteena tässä yhteydessä on määrittää työnantajien keskimääräisille vuosittaisille maksuosuuksille uskottavat ennusteet.

Lakisääteisten työeläkemaksujen osuuksien aikasarjat on esittänyt esimerkiksi ETK. Aikasarjat on ilmoitettu eläkelakien mukaan, sillä maksuosuudet vaihtelevat niiden mukaan. Koska minulla ei ole rekisteriaineistossa tietoa, minkä eläkelain mukaan rekisteriaineistoista löytyvät tiedot eläkemaksuista on maksettu, määritin vuosittaiset keskimääräiset maksujen osuudet eri eläkelakien yli. Koska suurin osa eläkemaksuista karttuu TyEL:n, VaEL:n ja KuEL:n piiristä, käytin keskimääräisen osuuksien laskemiseen näiden kolmen eläkelain maksuosuuksia ja maksutuloja (karttuneet eläkemaksut). Laskin keskimääräisen maksosuuden painotettuna keskiarvona näiden kolmen eläkelain maksuosuuksien yli, missä painoina toimivat näiden kolmen eläkelain vuosittaiset maksutulot.

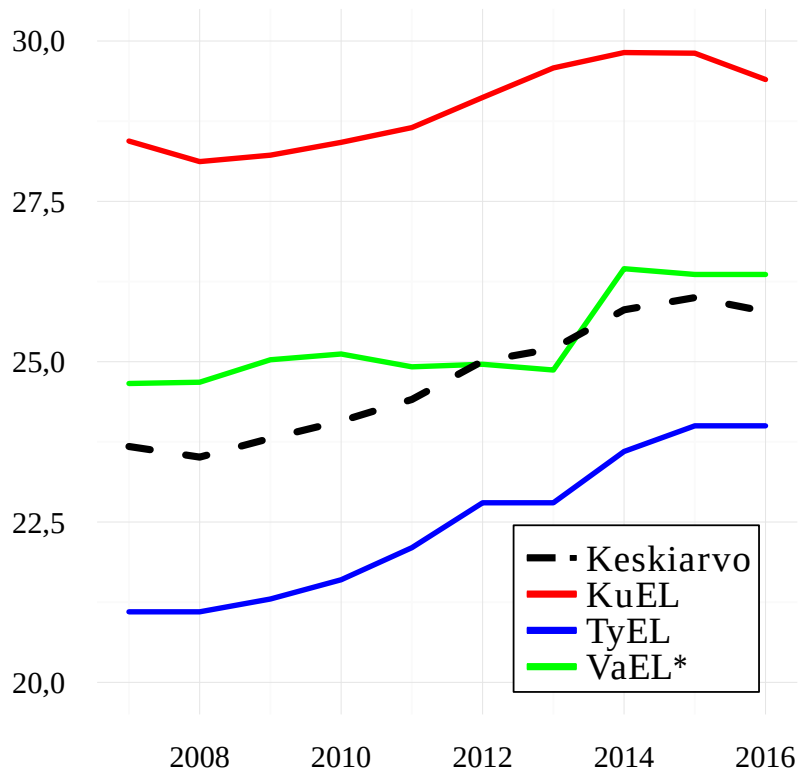
Keskimääräinen vuosittainen eläkemaksujen osuus palkoista vuosittain oli näin:

$$Keskiarvo = \frac{(TyEL_{os} \times TyEL_m) + (VaEL_{os} \times VaEL_m) + (KuEL_{os} \times KuEL_m)}{TyEL_m + VaEL_m + KuEL_m}, \tag{6.4}$$

missä $Keskiarvo$ on keskimääräinen vuosittainen maksuosuus palkoista, $TyEL_{os}$ on maksun keskimäärinen vuosittainen osuus TyEL:n piirissä ja $TyEL_m$ on TyEL:n piiristä karttunut vuosittainen eläkemaksutulojen summa. Merkinnät ovat samat muille eläkelaeille.

Kuvassa 6.3 on esitetty keskimääräiset lakisääteisten työeläkemaksujen osuudet (%) palkoista eläkelain mukaan vuosina 2007–2016. Näiden lisäksi kuvassa on esitetty eläkelakien osuuksien maksutuloilla painotettu keskiarvo samoille vuosille. Kuvan 6.3 tässä yhteydessä tärkeimmän viestin voisi tiivistää niin, että eläkemaksuosuuksien keskiarvo on noussut vuosina 2007–2016 noin 24 %:sta 26 %:iin ja että tämä nousu johtuu pääsiassa TyEL-maksuosuuden noususta.

Korkman et al. (2007) ennustivat oikein vuonna 2007, että eläkemaksujen osalta korotuspaineita on nimenomaan yksityisen sektorin (TyEL) maksuosuudessa, sillä ”julkisen sektorin eläkemaksut ovat jo nyt [2007] huomattavan paljon koreammat kuin yksityisen sektorin, eikä niissä ole sen vuoksi korotuspaineita, jos ennusteet toteutuvat.” (s. 29). Korkman et al. (2007) (s. 30) vuonna 2007 luoma TyEL-maksun ennustejakauma vastaa vuoden 2016 todellisuutta hyvin, sillä Korkmanin et al. noin 24 %:n mediaanienuste vuodelle 2016 vastaa hyvin todellista vuoden 2016 maksuosuutta, kuten kuvasta 6.3 selviää.



Kuva 6.3: *Keskimääräiset lakisääteisten työeläkemaksujen osuudet (%) palkoista eläkelain mukaan vuosina 2007–2016. Keskiarvo on painotettu keskiarvo TyEL-, KuEL- ja VaEL-maksuosuuksien välillä. Painoina ovat eläkelaittaiset vuosittaiset maksutulot. * VaEL:n vuoden 2016 osalta olen käyttänyt vuoden 2015 arvoa, koska todellinen vuoden 2016 luku paikkaa VaEL-maksuosuuksissa tehtyjä aiempia virheitä. Lähteet: ETK ja ETK.*

Koska Korkman et al. (2007) onnistuivat vuonna 2007 ennustamaan hyvin eläkemaksuosuuksien kehittymisen vuonna 2016 (TyEL-osuus nousee n. 24 %:iin, julkisella puolella ei samanlaisia korotuspaineita), päätin käyttää keskimääräisen työeläkemaksuosuuksien ennusteissa ennustejakaumaa, joka vastaa karkeasti Korkman et al. 2007 esittämää ennustejakaumaa TyEL-maksun kehittymiselle (s. 30). Viitatussa ennustejakaumassa TyEL-maksun osuuden mediaaniennuste nousee vuoteen 2030 mennessä noin 27–28 %:iin, millä tasolla se myös pysyy tämän jälkeen. Vuoteen 2030 ennusteiden mennessä ennusteen 90 %:n yläraja nousee noin 32 %:iin ja ennusteen 10 %:n alaraja laskee noin 20 %:iin. Tämän jälkeen ennusteiden näin määritetty 80 %:n uskottavuusväli pysyy suunnilleen samansuuruisena.

Käyttämäni ennuste eläkemaksuosuuksista on yksinkertaistettu versio Korkman et al. (2007):n esittämästä ennusteesta. Ennusteessani oletan, että TyEL-maksuosuuksien nousu jatkuu niin, että TyEL-maksu nousee vastaamaan keskimääräistä eläkemaksua eri eläkelakien yli. Käytän keskimääräisen maksusosuuden odotusarvona 27 %:ia, mikä vastaa suunnilleen yllä esitetyn Korkman et al. 2007:n mediaaniennustetta

vuodesta 2030 lähtien. Elinkaariennusteissa käytän 27 %:n odotusarvoa heti vuodesta 2015 (tai muusta elinkaariennusteiden aloitusvuodesta lähtien). Oletan lisäksi maksuosuuksien todennäköisyysjakauman olevan normaalijakautunut ja sen keskihajonnan olevan neljä prosenttiyksikköä. Tällä keskihajontavalinnalla maksuosuusennusteista suunnilleen 80 %:ia osuu samalle 80 %:n uskottavuusvälille viitatussa esimerkkiennusteessa eli noin 20 %:n ja 33 %:n välille. Normaalijakauman, jonka odotusarvo on 27 ja keskihajonta 4, kertymäfunktion arvo on 32:n kohdalla 0,89 ja 22:n kohdalla noin 0,11. En ottanut huomioon ajassa perättäisten maksusosuuksien tilastollista riippuvuutta toisistaan. Ennustetta voisi parantaa vielä esimeriksi tältä osin.

Valitsemani odotusarvo eläkemaksuosuuksien ennusteiden odotusarvolle on lähellä myös ETK:n määrittämää ”riittävää” maksuosuutta eläkemaksuosuuksille ETK:n pitkän aikavälin ennusteissa lakisääteisiä eläkkeitä koskien. ETK:n ennusteraportin mukaan: ”Kaikkien työeläkelakien eläkemenojen rahoittamiseen riittävä maksutaso suhteessa koko talouden työtulosumaan olisi 28,1 [prosenttia].”. (Risku et al. 2013: tiivistelmä, s. 9). Mikäli käyttäisin elinkaariennusteissa maksuosuuksien odotusarvona 28 %:ia (27:n sijaan), tämä tekisi niistä elinkaariennusteista keskimääräisemmin positiivisempia, joissa ennustetut palkkasummat ovat korkeampia, kuin niistä, joissa ennustetut palkkasummat ovat matalampia. Tällöin Suomessa syntyneiden elinkaariennusteista muodostuisi keskimäärin positiivisempia kuin Irakissa ja Somaliassa syntyneillä maahanmuuttajilla.

Palkansaajien maksuosuus. Koska palkansaajien maksuosuus on ollut viime vuosina noin 7 %:ia palkasta, käytin työantajien osuuden ennusteen odotusarvona 20 %:ia (ETK). Lisäksi, koska palkansaajien osuus on ollut noin yksi neljäsosa kokonaiseläkemaksuista, oletin, että maksuosuuksien suuruuksissa myös noin yksi neljäsosa maksuosuuksiin sisältyvästä epävarmuudesta (ennusteiden keskihajonnasta) sisältyy palkansaajien – ennusteissa työnantajista riippumattomaan – osuuteen. Tästä syystä oletin, että työantajien normaalijakautuneen keskimääräisen maksuosuuden ennusteen keskihajonta on 3 prosenttiyksikköä.

Työnantajien maksuosuuksien ennusteet. Kaiken kaikkiaan elinkaariennusteissa työnantajien keskimääräisillä vuosittaisilla maksuosuuksilla (%) on normaalijakautunut ennustejakauma:

$$\textit{Työantajien keskimääräisen maksuosuuden vuosittainen ennuste} \sim N(20, 3). \tag{6.5}$$

Tämän normaalijakauman parametreille olisi voinut määrittää vielä omat priorijakaumat sen sijaan, että ne ovat nyt vakioita. Joka tapauksessa tälläkin määritelmällä työnantajien maksujen keskiarvoihin – eläkemaksuja kartuttavan ikäintervallin yli (tässä tutkimuksessa ikävuodet 18–62 eli 45 (ikä)vuotta) – sisältyy epävarmuuksia siten, että keskiarvo on 25 %:n todennäköisyydellä suurempi kuin 19,7 %:ia ja 75 %:n todennäköisyydellä pienempi kuin 20,3 %:ia. Samat keskiarvon uskottavuusintervallin rajat ovat 5 %:lle ja 95 %:lle noin 19,27 ja 20,74 %:ia.

Sosiaalivakuutusmaksujen osuus palkoista. Työnantajien sosiaalivakuutusmaksut koostuvat työnantajien maksamista sairaus- ja työttömyysvakuutusmaksuista (työeläkevakuutus jo edellä). Työnantajien näiden sosiaalivakuutusmaksujen maksuosuudet palkoista vuosina 2011–2017 on esittänyt esimerkiksi Sosiaali- ja terveysministeriö (STM). Nämä maksuosuudet on esitetty taulukossa 6.2. *"Maksut yhteensä"* on sairausvakuutusmaksuosuuden ja työttömyysvakuutusmaksuosuuden summa.

Taulukko 6.2: *Työnantajien sosiaalivakuutusmaksut osuuksina (%) palkoista vuosina 2011–2015. * Kun palkkasumma vuonna 2017 alle 2 059 500 euroa. Lähde: Sosiaali- ja terveysministeriö (STM).*

	2011	2012	2013	2014	2015	2016	2017
Sairausvakuutus	2,12	2,12	2,04	2,14	2,08	2,12	1,08
Työttömyysvakuutus*	0,8	0,8	0,8	0,75	0,8	1,0	0,8
Maksut yhteensä	2,92	2,92	2,84	2,89	2,88	3,12	1,88

Koska vuotta 2017 lukuun ottamatta maksujen keskiarvo on noin 2,9 %:ia palkasta ja koska maksuosuus näyttää vaihtelevan joitakin prosenttiyksikön desimaaleja vuosittain (ennen vuotta 2017), käytin elinkaariennusteissa seuraavaa vuosittaista ennustetta työnantajien sosiaalivakuutusmaksujen osuudelle palkoista:

$$\textit{Työantajien keskimääräisen maksuosuuden vuosittainen ennuste} \sim N(2{,}9; 0{,}15). \tag{6.6}$$

Olen antanut tässä keskihajonnan olla tässä suuri verrattuna vuosien 2011–2015 vaihteluun nähden, koska vuonna 2016 osuus poikkesi noin 0,2 %-yksikköä aiempien vuosien keskiarvosta.

6.1.3 Eläkkeiden karttumisen osuus palkoista ja palkattomilta ajoilta

Työeläkkeiden karttuminen palkkasummasta. Eläkkeiden karttumisen osuudella palkoista viittaan siihen prosenttiosuuteen, jolla työeläkkeitä karttuu henkilölle hänen vuosittaisesta palkkasummastaan. Käytän elinkaariennusteissa vuoden 2011 karttumisprosentteja, jotka ovat:

- 1,5 %:ia palkkasummasta alle 53-vuotiaille.
- 1,8 %:ia palkkasummasta 53–62-vuotiaille.
- Yli 62-vuotiaat ovat kaikki eläkkeellä elinkaariennusteissa.

Vuoden 2017 eläkeuudistuksen seurauksena työeläkkeiden karttuma muuttuu yllä olevaan verrattuna niin että:

- Vanhuuseläkeikä nousee asteittain.

- Sama eläkekarttuma, 1,5 %:ia palkkasummasta, 17-vuotiaasta maksimieläkeikään. Koskee viimeistään vuodesta 2026 alkaen kaikkia.
- Lisäksi eläkekarttuma koko palkkasummasta, ennen vuotta 2017 karttuma laskettiin palkoista, joista oli poistettu palkansaajan eläkemaksu.

Pitäydyn elinkaariennusteissa kuitenkin vuoden 2011 luvuissa, koska VATT:n (2014) esimerkkiä seuraten annan kaikkien henkilöiden siirtyä eläkkeelle 62 ikävuoden jälkeen. Koska vanhuuseläkeikä kuitenkin todellisuudessa nousee vuoden 2017 uudistuksen jälkeen, 53–62-vuotiaille käyttämäni korkeampi karttumisprosentti kompensoi karttuneen eläkkeen määrässä sitä, että elinkaariennusteissa ei ole mukana yli 62-vuotiailta karttuvia eläkkeitä. Lisäksi noudatan elinkaariennusteissa vuoden 2017 eläkeuudistusta siinä, että lasken ennustetut eläkekarttumat koko ennustetuista palkkasummista – enkä sellaisista palkkasummista, joista on aluksi poistettu ennustettu palkansaajan työeläkemaksu.

VEKL ja palkattomat ajat. Noudatan ETK:n pitkän aikavälin laskelmien esimerkkiä siinä, että eläkettä karttuu palkkasumman lisäksi myös alle kolmivuotiaan lapsen hoidon tai opiskelun ajalta (VEKL) sekä palkattomilta ajoilta eli käytännössä saaduista sosiaalietuuksista kuten ansiosidonnaisesta työttömyyskorvauksesta. Näiden kahden karttuman osalta päädyin sellaiseen yksinkertaiseen ratkaisuun, että henkilön ennustettu VEKL-karttuma on 2,9 %:ia ennustetuista palkkasummista karttuneesta työeläkkeestä ja palkattomien aikojen karttuma on puolestaan 5,3 %:ia ennustetuista palkkasummista karttuneesta työeläkkeestä. Nämä osuudet on saatu ETK:n pitkän aikavälin ennustelaskelmien eritellymmistä tuloksista, jotka pyysin ETK:lta (ETK 2018).

6.1.4 Maksettavien kansaneläkkeiden määrä

Kansaneläkkeen määrä riippuu ansioeläketuloista, perhesuhteista (yksin asuva vai avo- tai avioliitossa) sekä asumisajasta Suomessa. Käsittelen jäljempänä (luvussa 6.2.2.2), kuinka osana elinkaariennusteita muodostetaan saatujen kansaneläkkeiden ennusteet. Käsittelen tässä yhteydessä kansaneläkkeen määrän muodostavia parametreja eli täyden kansaneläkkeen määrää perhesuhteen mukaan.

Maksettava täysi kansaneläke. Täysi kansaneläke oli vuonna 2011 yksin asuvalle 586,46 e/kk ja avo- tai avioliitossa olevalle 520,19 e/kk (Mielonen 2011, s. 16). Vuonna 2016 samat rajat olivat – rahanarvonkertoimella vuoden 2011 euroihin muunnettuna – 600,83 e/kk ja 532,94 e/kk (ETK 2016, s. 17). Täyden kansaneläkkeen rajat nousivat siten reaalisesti keskimäärin vain vähän yli 2 euroa vuodessa. Tämä johtuu siitä, että kansaneläkettä tarkistetaan vuosittain kansaneläkeindeksillä, joka puolestaan määräytyy elinkustannusindeksi mukaan (ETK 2016, s. 21). Näin ollen kansaneläkerajat nousevat lähinnä vain hintojen – eivätkä esimerkiksi palkkojen – nousun mukaan. ETK:n pitkän aikavälin ennusteissa maksettavaa kansaneläkettä ja takuueläkettä korotetaan vuodesta 2022 lähtien kuitenkin indeksillä, joka seuraa puoliksi ansiotason ennustettua kasvua (ETK 2016, s. 32). Noudatan vastaavaa korotusta maksettavan kansaneläkkeen määrän ennusteissa.

Tilastokeskuksen julkaisemien väestörakennetietojen mukaan yli 64-vuotiaista oli avioliitossa vuonna 2011 noin 52 %:ia ja vuonna 2016 puolestaan noin 53 %:ia. Siviilisäätyjen mukaan jaoteltu tilasto ei sisällä luokitusta avoliitossa olevista. Viitatun tilaston perusteella oletan kuitenkin elinkaariennusteissa puolen eläkkeellä olevasta väestöstä olevan avo- tai avioliitossa. Tästä syystä käytän elinkaariennusteissa täyden kansaneläkkeen rajana keskiarvoa yksin asuvien rajan ja avo- tai avoliitossa olevien rajan välillä. Tämä keskiarvo, jota käytän, on 553,33 e/kk. Korvaan tässä käytännössä yksin asumiseen (tai avo- tai avioliitossa asumiseen) liittyvän epävarmuuden kiinteällä keskiarvolla. Elinkaariennusteista tulisi tältä osin realistisempia, jos keskiarvolla korvaaminen korvattaisiin elinkaariennusteissa arpomalla kukin eläkkeellä oleva henkilö yksin asuvaksi suunnilleen (ehdollistamattomalla) todennäköisyydellä 0,50.

Maksettavan takuueläkkeen raja-arvo. Takuueläkkeen määräämä vähimmäiseläkkeen määrä vuonna 2011 oli 687,74 e/kk (Mielonen 2011, s. 16). Vuonna 2016 tämä summa oli – rahanarvonkertoimella vuoden 2011 euroihin muunnettuna – 726,39 e/kk. Myös takuueläkkeen määrää tarkistetaan vuosittain kansaneläkeindeksillä, mutta takuueläkkeen määräämä vähimmäiseläkkeen määrä nousi suhteellisesti täyttä kansaneläkettä nopeammin vuosina 2011–2016, koska takuuläkkeeseen tehtiin vuonna 2016 tasokorotus (Mielonen 2011, s. 20–21). Ilman vuoden 2016 tasokorotusta, mikä merkitsi takuueläkkeen vähimmäiseläkkeen nousua 20 eurolla, myös takuueläkkeen vähimmäiseläke olisi noussut keskimäärin vuosittain vain noin neljällä eurolla vuosina 2011–2016. (ETK 2016, s. 21).

ETK:n pitkän aikavälin ennusteissa maksettavaa kansaneläkettä ja takuueläkettä korotetaan vuodesta 2022 lähtien indeksillä, joka seuraa puoliksi ansiotason ennustettua kasvua (ETK 2016, s. 32). Noudatan vastaavaa korotusta maksettavan takuueläkkeen määrän ennusteissa.

6.1.5 Ennustetuista eläkeajan tuloista maksettavat verot (veroprosentit)

Eläkeajan tuloista maksettavia veroja varten on määritettävä seuraavat parametrit: tuloveroprosentit eläketuloluokkien mukaan, (eläkeajan) pääomatulojen veroprosentti sekä välillisten verojen prosentit kaikista eläkeajan bruttotuloista. Eläkeajan pääomatulojen osalta on lisäksi määritettävä parametrit, joiden avulla ennustetuista työeläketuloista ennustetaan eläkeajan pääomatulojen määrä.

Ansiotuloveroprosenttien tulorajat ja veroprosentit. Välittömien verojen ja veronluonteisten maksujen veroprosentteina bruttoeläketuloluokittain käytän Mielonen 2011:n esittämiä tietoja (s. 26). Mielonen 2011 (s. 26) esittää siis kuukausittaisille bruttoeläketuloille välittömien verojen ja veronluonteisten maksujen prosentit. Näihin veroihin ja maksuihin sisältyvät kunnallisvero, valtionvero, sairausvakuutusmaksu ja kirkollisvero. En poistanut veronmaksuprosenteista kirkollisveron (pientä) osuutta, sillä veroprosentteihin sisältyy muutenkin epävarmuuksia sekä niistä puuttuu joitakin verolajeja kuten maksetut kiinteistöverot. Mielonen 2011 esittää samat tuloverorajat vuonna 2016 (s. 30). Vuosien 2011–2015 välillä matalien eläketulojen verotus on selvästi keventynyt. Käytän silti elinkaariennusteissa vuoden 2011 lukuja, sillä tällä ratkaisulla keskimäärin matalien ennustettujen eläketulojen maahanmuuttajaryhmille tulee suhteessa Suomessa syntyneisiin hieman enemmän maksettuja veroja eläkeiässä, kuin jos olisin käyttänyt vuoden 2016 lukuja.

Reaalisten bruttoeläketulojen tuloluokittaisiin veroprosentteihin sisältyy tietenkin epävarmuuksia, mutta en mallinna tätä epävarmuutta, sillä en koe omaavani tarpeeksi tietoa asiasta. Joka tapauksessa olisin ainakin varovainen reaalituloluokittaisten veroprosenttien nostamisen suhteen, sillä Turkkila 2011:n tutkimuksen mukaan reaalinen tuloverotus *keveni* 1990-luvun alusta ainakin vuoteen 2008 asti, johon tutkimus yltää (ks. esim. kuviot: s.116 ja s.153).

Pääomatuloveroprosentti ja ennustetut pääomatulot eläkeiässä. Käytin elinkaariennusteissa pääomatulojen veroprosenttina 30:tä prosenttia. Todellisuudessa (vuonna 2016) yli 30 000 euron pääomatulojen osalta veroprosentti on hieman korkeampi eli 34. Tällä oletuksella korkeammista pääomatuloista maksetaan hieman vähemmän veroja, kuin jos niiden osalta noudatettaisiin korkeampaa pääomatulojen veroprosenttia. Myös tässä yhteydessä pääomatuloverojen prosentin ennusteelle voisi muodostaa todennäköisyysjakauman, joka kuvaisi sen suuruuteen tulevaisuudessa liittyviä epävarmuuksia. Koska pääomatuloveroprosentin taso lienee yhteydessä Suomen-kaltaisten maiden pääomatuloveroprosenttien kehitykseen, tässä voisi pääomatuloveroprosentin ennusteen todennäköisyysjakauman perusteena voisi käyttää myös kansainvälistä kirjallisuutta aiheesta.

Tutkimuskäytössä olevan FLEED-rekisteriaineiston perusteella Suomessa syntyneillä eläkeläisillä pääomatulot ovat keskimäärin noin 10 %:ia bruttotyöeläketuloista. Tästä syystä käytän eläkeiän pääomatulojen mahdollisimman yksinkertaisena – ja deterministisenä – ennusteena 10 prosenttia ennustetuista työeläketuloista. Todellisuudessa pääomatulojen jakauma on hyvin vino, ja pääomatulot ovat kasaantuneet vain osalle väestöä. Pääomatulojen realistisissa ennusteissa pitäisi huomioida tämä vinous sekä pääomatulojen määrään liittyvä epävarmuus. Tällä ratkaisulla keskimäärin matalien ennustettujen eläketulojen – ja siten oletettavasti myös matalien pääomatulojen – maahanmuuttajaryhmille tulee suhteessa Suomessa syntyneisiin enemmän maksettuja eläkeiässä, kuin jos olisin käyttänyt realistisempaa ratkaisua, eli antanut pääomatulojen olla vinosti jakautuneita.

Välillisten verojen tulorajat ja veroprosentit. Eläkeiän maksettujen välillisten verojen suhteen käytän samanlaista ratkaisua, joka on esitetty tutkimuksen I osassa (luku 6.2.4). Käsittelen tätä ratkaisua hieman tarkemmin jäljempänä eläkeiän välillisten verojen ennustamisen yhteydessä. Joka tapauksessa tavoitteena on selvittää eläkeiän bruttotuloista (elinkaariennusteissa: työeläketulot + pääomatulot eläkeiässä) maksettujen välillisten verojen prosentit bruttotuloluokittain. Käytän tässä Jäntti (2009):n esittämiä välillisten verojen prosentteja kotitalouksien bruttorahatuloille per kotitalouden kulutusyksikkö. Koska viitatut veroprosentit on ilmaistu vain kulutusyksiköittäin, ennustin eläkeikäisille (yli 62-vuotiaille) kulutusyksiköiden määrän näiden kotitalouksissa. Henkilön itsensä lisäksi ennustin kotitaloudessa asuvan keskimäärin 0,45 henkilöä tulonjakotilastoon viitaten (SVT: *Tulonjakotilasto*). Tämä toisen henkilön asumisen todennäköisyys – 0,45 – on parametri, jota käytin ennustamaan toisen henkilön asumista samassa kotitaloudessa eläkeikäisenä.

6.2 Eläkkeisiin liittyvien vaikutusten ennustaminen palkkasummaennusteista ja muista ennusteparametreista

Palkkasummien ennusteet vaikuttavat elinkaariennusteisiin pääasiassa eläkeiän osalta eli yli 62-vuotiaana. Käytän elinkaariennusteissa yksinkertaista eläkejärjestelmää, joka on hyvin pitkälti VATT:n tutkimuksen (2014a) kaltainen. VATT:n tutkimuksen skenaarioissa oletetaan, että kaikki henkilöt jäävät eläkkeelle 63-vuotiaina (VATT 2014b). Seuraten tätä esimerkkiä oletan, että kaikki henkilöt jäävät eläkkeelle sinä vuonna, kun he täyttävät 63 vuotta. Lisäksi VATT:n tutkimuksen tapaan eläkkeet muodostuvat vain vanhuuseläkkeistä, eli esimerkiksi työkyvyttömyyseläkkeitä ei ole mukana. Työeläkkeen lisäksi mukana on ennustettu saatava kansaneläke ja takuueläke.

Vanhuuseläkkeitä ja niiden karttumista koskevat ennusteet, jotka perustuvat palkkasummaennusteisiin ja edellä tässä luvussa esitettyihin ennusteparametreihin, vaikuttavat seuraavasti nettovaikutusten elinkaariennusteisiin:

1. Positiivinen nettovaikutus:

– Työnantajien maksamat lakisääteiset työeläkevakuutusmaksut ja muut sosiaalivakuutusmaksut (18–62-vuotiaita koskien).

– Välittömät verot eläketuloista.

– Pääomatuloverot eläkeikäisenä.

– Välilliset verot eläkeikäisenä.

2. Negatiivinen nettovaikutus:

– Saatavat kokonaiseläkkeet (työeläke + kansaneläke + takuueläke).

– Maksettavat eläkeläisten asumistuet

Nämä nettovaikutukset määräytyvät simuloitujen palkkasummaennusteiden, edellä esitettyjen ”muiden ennusteparametrien” ennusteiden ja olotilaennusteiden mukaan. Esitän seuraavaksi, miten vaikutukset lasketaan näistä ennusteista elinkaariennusteiden yhteydessä.

Ennustetut palkkojen reaalisen tason muutokset. Käytin tulevaisuuteen sijoittuvista vuosittaisista palkkasummista ennusteita, joissa on huomioitu palkkojen reaalisen tason vuosimuutosten ennusteet. Palkkasummia käsittelevässä luvussa 5 laadin palkkasummien ennustemallit, joissa ennustetut palkkasummat ovat kuitenkin vain vuoden 2011 reaalisessa tasossa. Edellä esitettyjen ennustettujen palkkojen reaalisen tason vuosimuutosten avulla saan huomioitua palkkasummaennusteissa palkkojen reaalisen ta-

son muutoksen. Koska palkkojen reaalisen tason vuosittaisten muutosten ennusteiden odotusarvot ova positiivisia, palkat keskimäärin nousevat.

6.2.1 Ennustetut työnantajien maksamat lakisääteiset työeläkevakuutusmaksut ja muut sosiaalivakuutusmaksut

Koska työnantajien maksamat pakolliset eläkevakuutusmaksut ja muut sosiaalivakuutusmaksut (työttömyys- ja sairausvakuutusmaksut) eivät sisälly työikäisten nettovaikutusmallien aineistoihin – eivätkä siten myöskään ennusteisiin – laskin näiden maksujen ennusteet palkkasummien ennusteista, joissa on huomioitu palkkojen reaalisen tason muutoksen ennusteet. Työeläkevakuutusmaksut ja muut sosiaalivakuutusmaksut saatiin ennustettuina osuuksina ennustetuista palkkasummista. Näiden osuuksien ennusteet on esitetty edellä tässä luvussa.

6.2.2 Ennustetut saatavat kokonaiseläkkeet (työeläke + kansaneläke + takuueläke)

6.2.2.1 Ennustettu työeläkkeen määrä

Maksettavien kokonaiseläkkeiden ennusteet riippuvat karttuneiden työeläkkeiden ennustetuista yhteenlasketuista määristä 18–62-vuotiaana sekä ennustetuista kansaneläkkeen ja takuueläkkeen määristä sekä olotilaennusteista. Olotilaennusteet vaikuttavat maksettavien eläkkeiden määrään siten, että kansaneläkettä eikä takuueläkettä makseta, jos henkilö on ulkomailla. Mikäli henkilö on kuollut, niin henkilölle ei makseta enää mitään eläkkeitä. Olotilaennusteet vaikuttavat myös työeläkkeen karttumiseen siten, että eläkettä ei kartu, jos henkilö on ennusteen mukaan ulkomailla.

Maksettava työeläke karttuu 18–62-vuotiaana ennusteissa. Koska todellisuudessa karttumista tapahtuu joidenkin henkilöiden osalta myös yli 62-vuotiaana, karttuvat työeläkkeet ovat todellista hieman pienempiä. Vuosittaiset ennustetut työeläkekarttumat saadaan ennustetuista palkkasummista karttumisprosenteilla, jotka on esitetty aiemmin tässä luvussa.

Vuoden 2015 elinaikakerroin. Koska elinkaariennusteissa käytetään vuoden 2015 kuolleisuuslukuja, pienensin ennustettuja alkavia työeläkkeitä vuoden 2015 elinaikakertoimella. Elinaikakertoimen arvo vuodelle 2015 on 0,972.

Palkkakertoimella korotettu ennustettu alkavan työeläkkeen määrä. Eläkkeen saaminen alkaa kaikilla (elossa olevilla) henkilöillä elinkaariennusteissa sinä vuonna, kun he täyttävät 63 vuotta. Korotan ennustetut alkavat eläkkeet ennustetulla *palkkakertoimella*. Ennustettu palkkakerroin seuraa 80 %:sti ansiotason kasvua 20 %:sti hintojen nousua. Reaalisen ansiotason muutoksen ennusteet on esitetty aiemmin luvussa 6.1.1.3.

Työeläkeindeksillä korotettu ennustettu saatavan työeläkkeen määrä. Elinkaariennusteissa korotan saatavia ennustettuja eläkkeitä vuosittain ennustetulla työeläkeindeksillä. Ennustettu työeläkeindeksi seuraa 20 %:sti ansiotason kasvua 80 %:sti hintojen nousua.

Ennustettujen saatavien työeläkkeiden määrät elinkaariennusteiden laskelmissa. Käytän ennustettuja saatavia työeläkkeitä kolmeen tarkoitukseen elinkaariennusteissa:

1. Saatavan kansaneläkkeen ja takuueläkkeen määrän ennustamiseen.
2. Eläkeiässä maksettujen verojen ennustamiseen.
3. Negatiivisena julkisen talouden nettovaikutuksena.

Kansaneläkkeen ja takuueläkkeen määrän ennusteissa sekä eläkeiässä maksettujen verojen ennustamisessa käytän saatavista ennustetuista työeläkkeistä (ennustetulla) palkkakertoimella ja työeläkeindeksillä korotettuja summia. Negatiivisen nettovaikutuksen osalta käytän ennustettuja työeläkkeitä, joita *ei ole korotettu* palkkakertoimella eikä työeläkeindeksillä. Ne ovat siis vuoden 2015 ansiotasossa. Jälkimmäinen ratkaisu johtuu siitä, että mikäli eläkejärjestelmä on nk. *"jakojärjestelmä"*, karttuneen eläkkeen tulevaisuudessa maksavien henkilöiden palkat nousevat samaan tahtiin kuin palkkakerroin. Mikäli kyse on taas nk. *"rahastoivasta"* eläkejärjestelmästä, oletan, että rahastoitavilla eläkemaksuilla ja niiden sijoitustuotoilla saadaan ennustetut saatavat eläkkeet maksettua. Suomen työeläkejärjestelmä on sekoitus näistä kahdesta järjestelmästä (ks. yksityisen puolen osalta esimerkiksi: Risku 2015), mutta en ota sen tarkemmin elinkaariennusteissa kantaa siihen, että kuka maksaa ennustetut työeläkkeet. Eli maksetaanko henkilön ennustettu työeläke tulevaisuudessa reaalisesti korkeammista palkoista vai henkilön työuran aikana maksamista rahastoiduista työeläkemaksuista.

6.2.2.2 Ennustettu kansaneläkkeen määrä

Ennustettu kansaneläkkeen määrä riippuu maksettavan työeläkkeen ennustetusta määrästä, asumisajasta Suomessa sekä siitä, onko eläkkeensaaja avo- tai avioliitossa vai ei. Oletan, kuten myös VATT (2014b) implisiittisesti, että asumisajan vaatimus Suomessa täyttyy kaikkien ennustehenkilöiden kohdalla. Noudatan muutenkin pitkälti VATT:n (2014b) tapaa laskea ennustettava kansaneläkkeen määrä.

Olen edellä esittänyt perusteet kansaneläkkeen määrälle kansaneläkkeen määrän ennusteissa. Vuonna 2015 täyden kansaneläkkeen määrä ennusteissa on 553,33 e/kk. ETK:n pitkän aikavälin ennusteissa maksettavaa kansaneläkettä ja takuueläkettä korotetaan vuodesta 2022 lähtien indeksillä, joka seuraa puoliksi ansiotason ennustettua kasvua (ETK 2016, s. 32). Noudatan vastaavaa korotusta maksettavan kansaneläkkeen määrän ennusteissa.

En huomioinut (VATT:n (2014b) tavoin) ennusteissa sitä, että kansaneläkkeen määrä pienentyy vasta, kun työeläketulot ovat suuremmat kuin 51,79 e kk (Mielonen 2011, 15). En myöskään huomioinut sitä, että pienin maksettava kansaneläke vuonna 2011 oli 6,23 e/kk (mt.).

Ennuste eläkeiässä maksettavalle vuosittaiselle kansaneläkkeelle henkilön kohdalla oli täten:

$$\begin{aligned} \textit{kansanel_enn} &= 12 \times I((\textit{kansaneläkeraja} - (\textit{työel_kartt_enn})/12) > 0) \\ &\quad \times (\textit{kansaneläkeraja} - 0{,}5 \times (\textit{työel_kartt_enn})/12), \end{aligned} \tag{6.7}$$

missä *kansanel_enn* on ennuste eläkeiässä vuosittain maksettavalle kansaneläkkeen määrälle, ja *työel_kartt_enn* on vuosittain maksettavan työeläkkeen määrän ennuste, joka on summa työikäisenä vuosittain karttuneiden työeläkkeiden ennusteista (työeläkkeen määrä yllä). Kansaneläkeraja on mainittu 553,33 e/kk.

6.2.2.3 Ennustettu takuueläkkeen määrä

Takuueläke varmistaa, että henkilö saa eläkettä vähintään tietyn rajasumman verran kuukaudessa. Vuodelle 2015 käytän takuueläkerajana 726,39 e/kk, kuten edellä esitin. ETK:n pitkän aikavälin ennusteissa maksettavaa kansaneläkettä ja takuueläkettä korotetaan vuodesta 2022 lähtien indeksillä, joka seuraa puoliksi ansiotason ennustettua kasvua (ETK 2016, s. 32). Noudatan vastaavaa korotusta takuueläkerajan ennusteissa.

Ennustetun vuosittaisen takuueläkkeen määrän laskentakaava on täten yksinkertaisesti:

$$\textit{takuuel_enn} = 12 \times (\textit{takuueläkeraja} - (\textit{työel_kartt_enn}/12 + \textit{kansanel_enn}/12)), \tag{6.8}$$

missä $\textit{takuuel_enn}_{is}$ on ennustettu eläkeiässä vuosittain maksettava takuueläkkeen määrä, ja *työel_kartt_enn* sekä *kansanel_enn* ovat karttuneen työeläkkeen ja maksettavan kansaneläkkeen ennusteet.

6.2.3 Ennustetut maksettavat verot eläkeikäisenä

Ennustetut välittömät verot eläketuloista. Laskin välittömät verot ennustetuista maksettavien eläkkeiden kokonaismääristä (ennustettu työeläke + kansaneläke + takuueläke) eli bruttoeläketuloista. Käytetyt veroprosentit bruttoeläketulojen mukaan on esitetty aiemmin tässä luvussa.

Mikäli henkilö on simulaatiokohtaisen olotilaennusteen mukaan Suomessa, eläketulot, joista välittömät verot maksetaan, käsittävät sekä työeläkkeen, kansaneläkkeen että takuueläkkeen. Ulkomailla olevan välittömät verot maksetaan vain työeläkkeistä. Tässä oletan, että ennustetuista työeläkkeistä maksetaan kaikki välittömät verot Suomeen, vaikka henkilö olisikin ulkomailla.

Ennustetut pääomatuloverot eläkeikäisenä. Maksetut pääomatuloverot työikäisenä sisältyvät työikäisten nettovaikutusennusteisiin, mutta eläkeiän osalta ne eivät sisälly muihin eläkeikään liittyviin ennusteisiin.

Ennustin pääomatulojen olevan 10%:ia ennustetuista bruttoeläketuloista, ja pääomatuloveroprosentin olevan 30. Tämä ennustus perustuu siihen, että pääomatulot ovat Suomessa syntyneillä eläkeläisillä

FLEED-aineistossa keskimäärin 10 %:ia eläketuloista. Olotilaennusteiden mukaan ulkomailla olevien pääomatulot lasketaan vain työeläkkeistä.

Eläkkeistä maksettavien välittömien verojen ja eläkeiän pääomatuloverojen lisäksi en ennusta muita eläkeiän maksettuja välittömiä veroja, sillä yritystulot ym. tulot ovat keskimäärin hyvin pieniä verrattuna eläke- ja pääomatuloihin eläkeiässä FLEED-aineiston mukaan.

Ennustetut välilliset verot eläkeikäisenä. Välilliset verot eläkeiässä ennustetaan eläkeiän bruttotuloista niille henkilöille, jotka ovat olotilaennusteen mukaan Suomessa. Bruttotulot koostuvat ennustetuista bruttoeläketuloista ja pääomatuloista, sillä kuten edellä esitin FLEED-aineiston tietoihin perustuen, suurin osa eläkeiän tuloista koostuu näistä tulonlähteistä.

Eläkeiän maksetut välillisten verojen henkilövuositasoiset simulaatiokohtaiset ennusteet laskin hyvin pitkälti samalla tavoin kuin toteutuneet maksetut välilliset verot tutkimuksen I osassa (luku 6.2.4). Henkilövuositasoiset välilliset verot muodostuivat tutkimuksen I osassa henkilöiden vuosittaisista brutto(raha)tuloista ja henkilöiden kotitalouksien vuosittaisista kulutusyksiköiden määristä.

Koska välillisten verojen estimaatit olivat saatavilla niiden lähteestä (Jäntti 2009) vain kotitalouksien bruttorahatuloille per kotitalouden kulutusyksikkö, jouduin tekemään seuraavat kaksi oletusta koskien välillisten verojen ennustamista:

1. Kaikilla kulutusyksiköillä on yhtä suuret bruttotulot kotitaloudessa.

2. Kaikki henkilöt eläkeikäisten kotitalouksissa ovat yli 62-vuotiaita eläkeläisiä.

Tilastokeskuksen tilastojen mukaan 65-vuotiailla ja sitä vanhemmilla henkilöillä oli vuonna 2011 yhdessä kotitaloudessa keskimäärin 1,45 henkilöä, kun kotitalouksissa olleiden henkilöiden kokonaislukumäärä jaettiin kotitalouksien kokonaislukumäärällä (SVT: *Tulonjakotilasto*).

Keskiarvo 1,45 vastaa yhtä montaa kulutusyksikköä, mikäli 2. oletus eläkeikäisten kotitalouksien jäsenten rakenteesta yllä pitää paikkansa. Jotta ottaisin simulaatiokohtaisissa ennusteissa huomioon epävarmuuden kulutusyksikköjen määrästä kunkin henkilön i kotitaloudessa, simuloin kulutusyksikköjen määrän jakaumaa poimimalla arvoja Bernoulli-jakaumasta, jonka parametri sai arvoksi 0,45:

$$\textit{Kulutusyksikkömäärä}(i) = 1 + Ber(0{,}5). \tag{6.9}$$

Kun kaikilla kotitalouden kulutusyksiköillä on 1. oletuksen mukaisesti yhtä suuret bruttotulot jokaisessa kotitaloudessa, välillisten verojen prosentti henkilön tuloille saadaan kokonaiseläketuloista per kulutusyksikkö samalla tavoin kuin tutkimuksen I osassa.

6.2.4 Ennustetut maksettavat eläkeläisten asumistuet

Ennustetut maksettavat eläkkeensaajien asumistuet lasketaan samasta kaavasta, jota Kansaneläkelaitos (Kela) käyttää niiden laskemiseen:

$$0{,}85 \times (\textit{huomioonotettavat asumismenot} - (\textit{perusomavastuu} + \textit{mahdollinen lisäomavastuu})). \quad (6.10)$$

Eläkkeensaajien asumistuen ennusteita varten on siten laadittava ennusteet henkilöiden huomioonotettaville asumismenoille, perusomavastuulle sekä mahdolliselle lisäomavastuulle. Lisäomavastuu on 40 %:ia lisävastuutulorajan ylittävistä henkilön (ja mahdollisen puolison yhteenlasketuista) eläketuloista. (Kansaneläkelaitos: *eläkkeensaajien asumistuki*).

6.2.4.1 Huomioonotettavien asumismenojen ennusteet

Asumismenot vuonna 2017. Ennustetut eläkeläisten asumistuet lasketaan vain vuokralla asuville. Tämä johtuu siitä, että vuonna 2017 noin 93 %:ia kaikista eläkeläisten asumistuista meni vuokramenoihin (Kansaneläkelaitos 2018b, s. 1.). Yksin vuokralla asuvien eläkeläisten asumistuen saajien asumismenojen jakauma vuonna 2017 on puolestaan esitetty Kelan raportissa (Kansaneläkelaitos 2018a, s. 4). Yksin vuokralla asuvien eläkeläisten asumistuen saajien asumismenojen keskiarvo vuonna 2017 oli 518,09 euroa. Kelan viitatun raportin asumismenojen jakauma on esitetty kuviossa. Tämän kuvion perusteella arvioin asumismenojen olevan suunnilleen normaalijakautunut. Arvioin myös kuvion perusteella, että asumismenojen keskihajonta olisi noin 1 000 euroa vuodessa. Käytin täten eläkkeensaajien asumistuen saajien vuosittaisten asumismenojen ennusteena pohjana normaalijakaumaa, jonka odotusarvo oli $12 \times 518,09e$ ja keskihajonta puolestaan 1 000 euroa. (Kansaneläkelaitos 2018a, s. 4).

Asumismenojen reaaliset muutokset. Elinkaariennusteissa käytettävissä eläkkeensaajien asumistuen saajien asumismenojen ennusteissa oli otettava huomioon myös näiden asumismenojen reaalinen kasvu. Oletin reaalisten asumismenojen nousevan *odotusarvoisesti* yhtä paljon kuin ennustetut palkkatasot. Tämä siitäkin huolimatta, että vuosina 2000–2017 kaikkien vuokra-asuntoja vuokrat ovat nousseet noin 80 %:ia samaan aikaan, kun ansiotaso on noussut Tilastokeskuksen ansiotasoindeksin mukaan "vain"64 %:ia (SVT: *Asuntojen vuokrat*, SVT: *Ansiotasoindeksi*). Mikäli antaisin näiden asumismenojen nousta keskimäärin nopeammin kuin palkkojen, tämä tekisi Irakissa ja Somaliassa syntyneiden elinkaariennusteista keskimäärin vielä negatiivisempia verrattuna kantaväestön elinkaariennusteisiin, sillä Somaliassa ja Irakissa syntyneiden ennustetut pienet eläketulot merkitsevät keskimäärin suuria ennustettuja maksettavia eläkeläisten asumistukia.

Vaikka asumismenot nousevat odotusarvoisesti palkkojen tason kanssa (ks. palkkojen reaalisen tason muutosten ennusteet luvussa 6.1.1.3) ja vaikka käytän asumismenojen muutosennusteiden pohjalla palkkojen muutoksen ennusteita, niin huomioin asumismenojen muutoksen ja palkkojen muutoksen välisen eron ennusteissa lisäämällä palkkojen muutoksen ennusteeseen satunnaisen termin. Tämä satunnainen termi on normaalijakautunut odotusarvolla nolla ja keskihajonnalla 0,01 (1 %). Toisin sanoen asumismenojen muutos-% poikkeaa palkkojen muutoksen ennusteprosentista normaalijakautuneella satunnaisella

termillä, jonka keskihajonta on 1 %. Tämä on melko suuri poikkeama, sillä palkkojen muutoksen ennusteiden odotusarvo on noin 1,4 %:ia.

Asumismenojen ennusteet. Näin ollen eläkeläisten asumistuen saajan *i* asumismenojen ennuste kalenterivuonna *k* on:

$$\textit{Asumismeno(i,k)} \sim N(12 \times 518{,}09; 1000^2) \times \textit{Asumismenotaso(k)}. \tag{6.11}$$

Tässä kalenterivuoden *k* asumismenotason ennuste, *Asumismenotaso(k)*, muodostuu kumulatiivisesti yksittäisistä, kalenterivuotta *k* edeltävistä asumistason vuosittaisten muutosten ennusteista. Nämä yksittäiset kalenterivuosittaiset asumistason muutoksen ennusteet noudattavat palkkojen muutoksen ennusteita lukuun ottamatta yllä esitettyä satunnaista termiä. Huomautan, että asumismenojen odotusarvo jäi elinkaariennusteissa vuodelle 2015 vuoden 2017 euroissa, minkä vuoksi asumismenoennusteet ovat odotusarvoisesti noin 1 %:n suurempia, kuin jos olisin tehnyt tämän muunnoksen.

Enimmäisasumismenojen ennusteet. Eläkkeensaajien asumistuessa on määritelty enimmäissummat huomioonotettaville asumismenoille. Tämä enimmäissumma vaihtelee kolmen eri kuntaryhmän mukaan siten, että summa on korkein pääkaupunkiseudun kunnissa. Enimmäisasumismenoille täytyi laatia myös omat ennusteensa. Yksinkertaisuuden vuoksi valitsin näiden ennusteiden odotusarvoksi II kuntaryhmän vuoden 2017 enimmäisasumismenon: 7 356,25 euroa (muunnettuna vuoden 2015 euroihin, huomioimatta asumismenojen kasvua). Jotta huomioisin jollain tavalla enimmäisasumismenojen ennusteissa epävarmuuden esimerkiksi kuntaryhmästä, käytin enimmäisasumismenojen ennusteen normaalijakaumaa, jonka keskihajonta oli 500 euroa (tässä ei vielä huomioitu asumismenojen kasvua). (Kansaneläkelaitos).

Oletin enimmäisasumismenojen vuosittaisten muutosten ennusteiden noudattavan asumismenojen vuosittaisten muutosten ennusteita. Näin ollen eläkeläisten asumistuen saajan *i* enimmäisasumismenojen ennuste kalenterivuonna *k* on:

$$\textit{Enimmäisasumismeno(i,k)} \sim N(7356{,}25; 500^2) \times \textit{Asumismenotaso(k)}. \tag{6.12}$$

Huomioonotettavien asumismenojen ennusteet. Edellisten perusteella eläkeläisten asumistuen saajan *i* enimmäisasumismenojen ennuste kalenterivuonna *k* on:

$$\textit{Huomioonotettavat asumismenot(i,k)} = min(\textit{Asumismeno(i,k)}, \textit{Enimmäisasumismeno(i,k)}). \tag{6.13}$$

6.2.4.2 Perusomavastuun rajan ennusteet

Perusomavastuun raja vuonna 2018 oli 610,42e/kk eli noin 600e vuoden 2015 euroissa. Kelan mukaan perusomavastuun raja on sidottu kansaneläkeindeksiin, joka nousee hintojen mukaan. Koska ETK:n pitkän aikavälin ennusteissa kansaneläkkeet eivät nouse kansaneläkeindeksin mukaan, vaan puoliksi palkkojen ja puoliksi hintojen mukaan, päätin sitoa myös perusomavastuun rajan nousun tähän (ETK 2016, s. 32). Näin ollen perusomavastuun raja nousee samojen ennusteiden mukaan kuin täysi kansaneläke (ks. edellä luku 6.2.2.2). (Kansaneläkelaitos).

6.2.4.3 Lisäomavastuun rajan ennusteet

Lisäomavastuun raja vuonna 2018 oli yksin asuvalle 8 676 euroa eli noin 8 580 euroa vuoden 2015 euroissa. Raja oli vuonna 2018 henkilölle, jonka puoliso saa asumislisää, 13 936 euroa eli noin 13 780 euroa vuoden 2015 euroissa. Yksinkertaisuuden vuoksi sovelsin ennusteissa samaa lisäomavastuun rajaa myös henkilöille, joiden puoliso ei saa asumislisää. Kelan mukaan lisäomavastuun raja on sidottu kansaneläkeindeksiin, joka nousee hintojen mukaan. Koska ETK:n pitkän aikavälin ennusteissa kansaneläkkeet eivät nouse kansaneläkeindeksin mukaan, vaan puoliksi palkkojen ja puoliksi hintojen mukaan, päätin sitoa lisäomavastuun rajan nousun puoliksi palkkojen ennustettuun nousuun, kuten tein myös perusomavastuun rajan nousun kanssa.

6.2.4.4 Ennustetut eläkkeensaajien asumistuet elinkaariennusteissa

Vuokralla asumisen todennäköisyys alimmissa eläketuloluokissa. Käytännössä eläkeläisten asumistukea saavat vain alimpiin eläketuloluokkiin kuuluvat. Lisäksi edellä esitin, että ennustetut eläketulot lasketaan vain vuokralla asuville eläkeläisille. Tästä syystä ennustettujen eläkkeensaajien asumistukien laskemiseksi on määritettävä ennusteet vuokralla asumiseen eläkeiällä alimmille eläketuloluokille. Näiden todennäköisyyksien määrittämiseen käytin tutkimusaineiston tarjoamaa tietoa. Tutkimusaineiston mukaan 65–69-vuotiaista Suomessa syntyneen kantaväestön henkilöistä, jotka kuuluivat kahteen alimpaan eläketulodesiiliin, noin 23 %:ia asui vuokralla. *Kaikkien* ulkomailla syntyneiden (ei vain Irakissa ja Somaliassa syntyneet) joukossa tämä osuus oli noin 50 %:ia. Tästä syystä käytin vuokralla asumisen todennäköisyytenä eläkkeensaajien asumistuen ennusteissa 0,25:ttä kantaväestön kohdalla ja 0,50:ttä Irakissa ja Somaliassa syntyneiden osalta.

Eläkeiän tulojen ennusteet. Ennustetun lisäomavastuun laskemiseen tarvitaan ennusteet eläkeiän tuloista, sillä lisäomavastuuvähennys on 40 %:ia lisävastuutulorajan ylittävistä henkilön (ja mahdollisen puolison yhteenlasketuista) eläketuloista. Edellä määritetyt eläkeiän ennustetut tulot koostuvat ennustetuista työeläkkeistä, kansaneläkkeistä ja takuueläkkeistä sekä ennustetuista pääomatuloista, joka on 10 %:ia muista eläkeiän tuloista. Yksinkertaisuuden vuoksi annoin mahdollisen puolison ennustetut eläkeiän tulot olla yhtä suuret lukuun ottamatta tulojen määrään lisättävää normaalijakautunutta satunnaista termiä, jonka odotusarvo on nolla ja keskihajonta tuhat euroa.

Ennusteet puolisolle eläkeiässä. Käytän samoja ennusteita sille, että henkilöllä on puoliso eläkeikäisenä, kuin käytin maksettujen välillisten verojen ennusteiden laatimisen yhteydessä. Mikäli henkilöllä on puoliso eläkeiässä, niin hänelle maksettava ennustettu eläkkeensaajan asumistuki puolittuu (verrattuna siihen, että puolisoa ei ole). Käytin asumismenojen ennusteena myös puolison kanssa asuville henkilöille edellä määritettyjä asumismenojen ennusteita, jotka pohjautuvat *yksin asuvien* asumismenoihin. Koska oletettavasti puolison kanssa asuvilla asumismenot per henkilö ovat keskimäärin pienemmät kuin yksin asuvilla, ennustettavia eläkkeensaajien asumistukia tulee elinkaariennusteisiin nyt hieman enemmän kuin, jos asumismenot olisi ennustettu erikseen niille, joilla on puoliso.

Ennustetut eläkkeensaajien asumistuet elinkaariennusteissa. Kaikki edellä esitetty huomioon ottaen, eläkkeensaajan asumistuen ennuste henkilölle *i*, joka asuu ennustetusti yksin, kalenterivuonna *k* laskettiin seuraavasti elinkaariennusteissa (vrt. edellä lauseke 6.10):

$$
\begin{aligned}
&I(\textit{Asuu vuokralla(i,k)}) \times 0{,}85 \times (\textit{huomioonotettavat asumismenot(i,k)} \\
&- (\textit{perusomavastuu(k)} + \textit{mahdollinen lisäomavastuu(i,k)})), \text{missä:} \\
&\textit{lisäomavastuu(i,k)} = max(0{,}4 \times (\textit{eläkeiän tulot(i,k)} - \textit{lisäomavastuuraja(k)}), 0).
\end{aligned}
\tag{6.14}
$$

Tässä ennustettu *lisäomavastuuraja(k)* koskee yksin asuvaa. Ennustetusti puolison kanssa asuvalle ennustettu eläkkeensaajan asumistuen määrä laskettiin elinkaariennusteissa muuten samalla lukuun ottamatta seuraavia seikkoja:

- Ennustettu *lisäomavastuuraja(k)* koskee puolison kanssa asuvaa.
- Puolison ennustetut eläkeiän tulot lasketaan mukaan henkilön eläkeiän tuloihin lisäomavastuuta laskettaessa.
- Ennustettu maksettava eläkeläisen asumistuki kerrotaan lopuksi 0,5:llä, sillä puolet asumistuesta menee puolisolle.

Luku 7

Alle 15-vuotiaiden ja yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutukset

7.1 Alle 15-vuotiaiden nettovaikutukset

Käsittelen alle 15-vuotiaiden julkisen talouden nettovaikutuksia kahdessa osassa: ensiksi peruskouluikäisiä eli 7–14-vuotiaita koskien ja sitten alle peruskouluikäisiä eli 0–6-vuotiaita koskien.

7.1.1 7–14-vuotiaiden nettovaikutukset

Peruskouluikäisten julkisen talouden nettovaikutukset koostuvat elinkaariennusteissa lähes täysin perusopetuksesta aiheutuvista menoista. Tämä selviää vertaamalla perusopetuksen vuosittaisia menoja opiskelijaa kohti (SVT: *koulutuksen talous* (SVT)) VATT:n (2014) tutkimuksen käyttämiin ”yksilöllisiin julkisiin kulutusmenoihin” peruskouluikäisille (s. 34), jotka ovat peräisin ETK:n tutkimuksesta (Riihelä et al. 2011).

Suomen virallisen tilaston (SVT) mukaan vuosittaiset oppilaskohtaiset oppilaitosten käyttömenot perusopetuksessa ovat olleet 8 500–8 600 euroa vuosina 2010–2014 (vuoden 2015 euroissa). Summa on Tilastokeskuksen rahanarvoimenkertoimen perusteella vuoden 2011 euroissa noin 8 100 euroa. VATT:n viitatussa tutkimuksessa peruskouluikäisten keskimääräinen julkisen talouden nettovaikutus on (ilman ”kollektiivisesti kulutettuja menoja”) keskimäärin noin -9 000 euroa vuodessa kuvion 13 perusteella (s. 34) vuonna 2011, joten peruskouluikäisten julkisen talouden nettovaikutukset koostuvat suurimmaksi osaksi perusopetuksen menoista.

Stakesin tutkimuksen (Hujanen et al. (2008)) mukaan 7–17-vuotiaiden keskimääräiset terveysmenot olivat 1 070 euroa vuonna 2006 eli noin 1 200 euroa vuoden 2011 euroissa (s. 18). Peruskouluvuoden menojen

(8 100 euroa) ja 7–17-vuotiaiden terveysmenojen keskiarvon (1 200 euroa) summa on 9 300 euroa eli noin 9 000 euroa. Oletan, että peruskouluikäisillä ei ole keskimäärin juuri lainkaan julkisen talouden tuloja. Täten peruskouluikäisen nettovaikutuksen keskiarvo on noin -9 000 euroa vuodessa. Viitatun kuvion perusteella peruskouluikäisten keskimääräinen julkisen talouden vaikutus vaihtelee jonkin verran ikäluokittain, mutta en huomioi tätä elinkaariennusteissa.

Koska perusopetuksen menojen reaalinen taso on pysynyt samansuuruisena 2010-luvulla – ja koska koko koulutussektorin menot ovat pysyneet reaalisesti suunnilleen samalla tasolla vuodesta 2000 vuoteen 2015 – oletan elinkaariennusteissa perusopetuksen menon reaalisen tason odotusarvon pysyvän esitetyssä noin 8 100 eurossa per oppilas (vuoden 2011 euroissa) ja siten peruskouluikäisten julkisen talouden vuosittaisten ennusteiden odotusarvon olevan -9 000 euroa. Koska arvioin peruskouluikäisten julkisen talouden keskiarvon VATT:n (2014) tutkimuksen kuviosta – enkä tarkoista euromääräisistä luvuista – mainittuun -9 000 euroon sisältyy (vähän) mittausvirhettä. Lisäksi perusopetuksen menot vaihtelevat varmasti ”jonkin” verran oppilaskohtaisesti – esimerkiksi saadun erityisopetuksen menojen myötä. Sisällyttääkseni mukaan elinkaariennusteisiin mainitun mittausvirheen ja oppilaskohtaisen vaihtelun, käytän elinkaariennusteissa todennäköisyysjakaumaa kuvaamaan tätä peruskouluikäisten nettovaikutukseen sisältyvää epävarmuutta.

7–14-vuotiaiden nettovaikutusennusteiden jakauma. Täten oletan elinkaariennusteissa, että peruskouluikäisen 7–14-vuotiaan yhden vuoden julkisen talouden vaikutuksen ennuste on:

$$\textit{7–14-vuotiaan nettovaikutusennuste} \sim N(-9000, 300). \tag{7.1}$$

Vuosittainen ennuste noudattaa siis normaalijakaumaa, jonka odotusarvo on -9 000 euroa ja keskihajonta 300 euroa. Peruskouluikäisten ennusteisiin saisi sisällytettyä vielä lisää epävarmuutta kasvattamalla ennustejakauman keskihajontaa.

Maahanmuuttajien lapset. Maahanmuuttajien lasten vaikutusten osalta oletan (luvussa 8), että perusopetuksen menon vuosittainen odotusarvo on 1 485 euroa suurempi kuin esitetty 8 100 euroa, joten maahanmuuttajien lasten kohdalla peruskouluikäisen lapsen julkisen taloden nettovaikutuksen vuosittainen odotusarvo on -10 485 euroa. Tämä maahanmuuttajien lasten korkeampi perusopetuksen keskimääräinen taso on seurausta maahanmuuttajien lapsille annettavasta – maahanmuuttajille suunnatusta – erityisopetuksesta. Olen esittänyt tutkimuksen I osan luvussa 7, miten olen päätynyt mainittuun 1 485 euroon keskimääräisenä menona maahanmuuttajille suunnatusta erityisopetuksesta.

7.1.2 0–6-vuotiaiden nettovaikutukset

Kuten edellä mainitsin, VATT:n (2014) tutkimuksen käyttämien ”yksilöllisten julkisten kulutusmenojen” peruskouluikäisten tiedot ovat peräisin ETK:n tutkimuksesta (Riihelä et al. 2011). VATT:n tutkimuksen ikäluokittaiset menotiedot on esitetty ETK:n tutkimuksen kuviossa 7b. (s. 41). Tämä kuvio osoittaa, että 0–6-vuotiaiden vuosittaiset julkisen talouden menot ovat keskimäärin noin 2/3-osaa peruskouluikäisten keskimääräisistä menoista. Koska peruskouluikäisillä keskimääräinen meno on vuoden 2011 euroissa noin

9 000 euroa, 0–6-vuotiaan keskimääräinen meno on noin 6 000 euroa. Koska on realistista olettaa, että 0–6-vuotiaita koskien ei ole julkisen talouden tuloja, -6 000 euroa on julkisen talouden nettovaikutusten odotusarvo 0–6-vuotiaille elinkaariennusteissa.

Koska arvioin 0–6-vuotiaiden keskimääräisen julkisen talouden nettovaikutuksen keskiarvon kuviosta (Riihelä et al. 2011, s. 41) – enkä tarkoista euromääräisistä luvuista – mainittuun -6 000 euroon sisältyy (vähän) mittausvirhettä. Lisäksi 0–6-vuotiaiden menot vaihtelevat varmasti ”jonkin” verran henkilöiden välillä – esimerkiksi saatujen terveyspalveluiden menojen osalta. Sisällyttääkseni mukaan elinkaariennusteisiin mainitun mittausvirheen ja henkilöiden välisen vaihtelun, käytän elinkaariennusteissa todennäköisyysjakaumaa kuvaamaan tätä 0–6-vuotiaiden nettovaikutukseen sisältyvää epävarmuutta.

0–6-vuotiaiden nettovaikutusennusteiden jakauma. Täten oletan elinkaariennusteissa, että alle peruskouluikäisen 0–6-vuotiaan yhden vuoden julkisen talouden vaikutuksen ennuste on:

$$\textit{0–6-vuotiaan nettovaikutusennuste} \sim N(-6000, 300). \tag{7.2}$$

Vuosittainen ennuste noudattaa siis normaalijakaumaa, jonka odotusarvo on -6 000 euroa ja keskihajonta 300 euroa. Alle peruskouluikäisten ennusteisiin saisi sisällytettyä vielä lisää epävarmuutta kasvattamalla ennustejakauman keskihajontaa. Olettaisin, että erityisesti syntymän jälkeen henkilöiden välillä voi olla ”paljon” vaihtelua saaduissa terveyspalveluiden menoissa.

7.2 Yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutukset

Yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutukset (eläkkeisiin liittyvien nettovaikutusten ennustamisesta ks. luku 6) muodostuvat pääasiassa terveys- ja sosiaalipalveluiden menoista, kuten seuraavaksi selviää. Oletan elinkaariennusteissa VATT:n (2014) esimerkkiä seuraten, että kaikki jäävät eläkkeelle sen vuoden lopussa, jolloin täyttävät 62 vuotta. Täten yli 62-vuotiaiden nettovaikutukset muodostuvat elinkaariennusteissa eläkkeisiin liittyvistä vaikutuksista sekä tässä käsiteltävistä terveys- ja sosiaalipalveluiden menoista sekä muista julkisen talouden menoista.

Puutteet yli 62-vuotiaita koskevissa menoissa. Yli 62-vuotiaiden terveys- ja sosiaalipalveluiden menoja koskeviin tietoihin liittyy kolme puutetta, joiden vuoksi elinkaariennusteissa kaikilla maahanmuuttajaryhmillä sekä Suomessa syntyneillä näiden palveluiden odotusarvoiset menot ovat yhtä suuret (tätä puutetta elinkaariennusteissa käsitellään myös puuttuvia tietoja koskevassa luvussa 10):

1. Käytössä ei ole rekisteritietoja yli 70-vuotiaiden terveys- ja sosiaalipalveluiden käytöstä. Myös 63–70-vuotiaiden osalta henkilölukumäärät ovat joidenkin maahanmuuttajaryhmien kohdalla pienet, joten en käytä niitä koskevia rekisteritietoja.

2. Käytössä ei ole myöskään tietoa siitä, että kuinka suuren osan vanhuusiän palveluista asiakkaat maksavat itse. Elinkaariennusteita voidaan myöhemmin vielä muuttaa realistisemmiksi laatimalla tältä osin uskottavia ennusteita. Asiakasmaksujen ennusteet perustuisivat tässä tapauksessa henkilöiden ennustettuihin eläketuloihin.

3. Käytössä ei ole tietoa siitä, kuinka paljon terveys- ja sosiaalipalvelujen keskimääräiset yksikkömenot nousevat esimerkiksi Somaliassa ja Irakissa syntyneillä, jos mukaan laskettaisiin esimerkiksi mahdollinen tulkkaus ja siihen kulunut aika.

Kun viittaan jatkossa tässä luvussa yli 62-vuotiaiden julkisen talouden menoihin ja nettovaikutuksiin, viittaan nimenomaan niihin menoihin ja vaikutuksiin, jotka aiheutuvat terveys- ja sosiaalipalveluiden menoista. Olisin voinut poimia yli 62-vuotiaiden ikäluokittaisten menojen keskiarvojen arviot edellä tässä luvussa viitatuista VATT:n (2014, s. 34) ja ETK:n (Riihelä et al. 2011, s. 41) tutkimusten kuvioista. Halusin kuitenkin määrittää eri terveys- ja sosiaalipalveluiden luokille omat ennusteensa.

7.2.1 Yli 62-vuotiaiden terveys- ja sosiaalipalveluiden menot

Poimin sukupuolen ja ikäluokat mukaan eritellyt yli 62-vuotiaiden terveys- ja sosiaalipalveluiden keskiarvot Stakesin aihetta käsittelevästä raportista (Hujanen et al. 2008). Erittelin raportissa esitettävät sosiaali- ja terveyspalvelut ja niiden menot viiteen eri luokkaan:

1. Erikoissairaanhoito.
2. Perusterveydenhuolto: vuodehoito, muu perusterveydenhuolto (ilman kotihoitoa) ja hammashoito.
3. Palveluasuminen (ml. tehostettu palveluasuminen) ja laitoshoito (mm. vanhainkodit).
4. Kotihoito.
5. Muut sosiaali- ja terveyspalvelut.

Olisin voinut käsitellä sosiaali- ja terveyspalveluiden menoja yhtenä kokonaisuutena, mutta käsittelemällä menoja eritellysti näiden viiden luokan kautta voin:

- Ottaa menoennusteissa huomioon keskiarvojen lisäksi menojen henkilövuosittaisen vaihtelun sosiaali- ja terveyspalvelun luokan mukaan.
- Laskea tämän tutkimuksen mahdollisessa päivitetyssä versiossa uskottavat ennusteet palveluista maksetuille asiakasmaksuille.

Stakesin (2008) raportissa esitetään vain sosiaali- ja terveyspalveluiden menojen keskiarvoja. Koska sosiaali- ja terveyspalveluiden menoissa on kuitenkin suurta vaihtelua henkilövuosittain, menoennusteista muodostuu realistisempia, jos tämä vaihtelu voidaan sisällyttää jotenkin ennusteisiin mukaan. Tämän

vaihtelun suuruutta on helpompi arvioida palveluiden luokkien mukaan. Lisäksi mahdolliset asiakasmaksuennusteet on käytännössä laadittava palveluiden luokkien mukaan, sillä esimerkiksi palveluasumisesta ja perusterveydenhuollosta maksettavat asiakasmaksut määräytyvät eri tavoin.

Alla taulukossa 7.1 on esitetty henkilöiden sosiaali- ja terveyspalveluiden keskimääräiset menot palveluiden luokan, sukupuolen ja ikäluokan mukaan vuonna 2006. Luvut ovat taulukossa vuoden 2006 euroissa, mutta varsinaisia elinkaariennusteita varten ne muunnetaan ennustevuoden (v. 2015) euroiksi Tilastokeskuksen rahanarvonkertoimella. Taulukosta käy ilmi esimerkiksi, että 65–74-vuotiaiden miesten keskimääräinen meno sairaanhoidosta oli 1 707 euroa vuonna 2006. Perusterveydenhuoltoon ja palveluasumiseen kuuluvat palvelut on esitetty yllä listassa.

63–64-vuotiaiden erikoissairaanhoidon keskiarvoluvut ovat 41–62-vuotiaiden ja 65–74-vuotiaiden luokkien keskiarvojen keskiarvo. Perusterveydenhuoltoa, palveluasumista ja vanhainkotiasumista ja kotihoitoa koskien en käyttänyt lainkaan näitä palveluiden luokkia koskevia keskiarvoja 63–64-vuotiaille, mutta suhteellisen pienet 63–64-vuotiaiden menot näistä palveluista sisältyvät 63–64-vuotiaiden muihin sosiaali- ja terveyspalveluiden menoihin.

Taulukko 7.1: *Yli 62-vuotiaiden terveys- ja sosiaalipalveluiden euromääräiset menot vuonna 2006 palveluiden luokan, ikäluokan ja sukupuolen mukaan. Luvut ovat näiden palveluiden keskimääräisiä menoja ikäluokan ja sukupuolen mukaan määritetyissä väestöissä. Lähde: Hujanen et al. (2008), s. 18, 51, 53–55, 57, 62.*

Palvelun luokka	**Sukupuoli**	**63–64-v.**	**65–74-v.**	**75–84-v.**	**yli 84-v.**
Erikoissairaanhoito	Miehet	1 226,5	1 707	2 588	3 114
	Naiset	1 024,5	1 340	2 010	2 158
Perusterveydenhuolto	Miehet	–	844	1 701	3 310
	Naiset	–	879	1 859	4 306
Palveluasuminen ja vanhainkodit	Miehet	–	359	1 531	5 530
	Naiset	–	354	2 197	8 248
Kotihoito	Miehet	–	219	705	2 040
	Naiset	–	226	1 035	2 830
Muut sosiaali- ja terveyspalvelut	Miehet	982,5	1 238	1 877	2 619
	Naiset	1 417,5	1 297	1 647	1 715
Sosiaali- ja terveyspalvelut yhteensä	Miehet	2 209	4 362	8 402	16 613
	Naiset	2 442	4 101	8 748	19 257

Sairaanhoidon ja perusterveydenhuollon menojen ennustejakaumat. Käytän elinkaariennusteissa sairaanhoidon ja perusterveydenhuollon ikäluokka- ja sukupuolikohtaisten ennusteiden keskiarvojen pohjana taulukossa 7.1 esitettyjä lukuja. Koska keskiarvot oletettavasti vaihtelevat jonkin verran vuosittain, lisäsin elinkaariennusteissa vuosittaisiin ikäluokittaisiin erikoissairaanhoidon menojen keskiarvoihin "kohinaa"normaalijakautuneesta satunnaistermistä, jonka odotusarvo oli nolla ja jonka keskihajonta oli 1/20-osa ikäluokittaisen menon odotusarvosta. Keskiarvoja (odotusarvoja) lukuun ottamatta erikois-

sairaanhoidon ja perusterveydenhuollon menojen ennusteiden jakaumat määräytyvät käytössä olevasta tutkimusaineistosta.

Olen laskenut erikoissairaanhoidon ja perusterveydenhuollon laitoshoidon ja avohoidon menot tutkimuksen I osassa (luvussa 6.3) THL:ltä saamieni rekisteritietojen sekä näitä palveluita koskevien yksikkökustannusten pohjalta. Tutkimusaineisto käsittää henkilöiden tietoja 70-vuotiaiksi asti. Näin ollen minulla on käytössäni henkilövuosittaiset arviot henkilöiden saamien erikoissairaanhoidon ja perusterveydenhuollon palveluiden menoista 70-vuotiaiksi asti.

Koska tarkoitus on määrittää yli 62-vuotiaille erikoissairaanhoidon ja perusterveydenhuollon menojen jakaumat, poimin lasketuista henkilövuositasoisista erikoissairaanhoidon ja perusterveydenhuollon menoista menot 66–70-vuotiaille. Käytin näistä 66–70-vuotiaita koskevia jakaumista saatuja estimaatteja ennustejakaumien määrittämiseksi kaikille yli 62-vuotiaiden ikäluokille. Varsinaisissa elinkaariennusteissa skaalasin 66–70-vuotiaita koskevan jakauman estimaatit niin, että ikäluokittaiset ennustejakaumat saivat odotusarvoikseen taulukossa 7.1 esitetyt ikäluokittaiset keskiarvot.

Koska rekisteriaineiston 66–70-vuotiaita koskevat sairaanhoidon ja perusterveydenhuollon menojen havaitut jakaumat eivät noudata mitään tunnettuja (todennäköisyys)jakaumia, estimoin ne käyttäen nk. *ei-parametrisia (kernel-)menetelmiä* käyttäen. Käytin tähän tarkoitukseen R-ohjelman valmista *density*-funktiota. Näin ollen – toisin kuin muuten tässä tutkimuksessa – nämä estimaatit ovat frekventistiä, ei-bayesilaisia estimaatteja. Voin korvata nämä ei-bayesilaiset estimaatit jossakin tutkimuksen mahdollisessa tulevassa versiossa bayesilaisilla ei-parametrisilla estimaateilla, jotka perustuvat myös mainittuihin normaalijakautuneisiin kerneleihin, joita kutsutaan myös *normaalijakaumaprosesseiksi* (*engl. Gaussian Process*). Käytännön menetelmän tällaiseen bayesilaiseen normaalijakaumaprosesseihin perustuvaan todennäköisyysjakauman tiheysfunktion estimointiin antavat esimerkiksi Kundu and Dunson (2011).

Palveluasumisen, vanhainkotiasumisen ja kotihoidon menojen ennustejakaumat. Palveluasumisen ja kotihoidon osalta määritin hyvin yksinkertaiset ennustejakaumat. Määritin THL:n ylläpitämän tietopankin (sotkanetin) ja Tilastokeskuksen ikäluokkien henkilömäärätilastojen perusteella ikäluokittaiset osuudet sille, että henkilö on joko tavallisessa tai tehostetussa palveluasumisessa tai vanhainkodissa. Samoin määritin myös ikäluokittaiset osuudet sille, että henkilö saa kotihoitoa. (Tilasto-ja indikaattoripankki Sotkanet.fi, SVT: *väestörakenne*).

Käytin näitä osuuksia elinkaariennusteissa ikäluokittaisina todennäköisyyksinä saada näitä palveluita. Kotihoidon palveluiden saamisen todennäköisyys oli myös ehdollinen palveluasumisen (ml. vanhainkotiasuminen) saamisella siten, että elinkaariennusteissa henkilö ei voi saada kotihoidon palveluita, mikäli hän on jo palveluasumisen tai vanhainkodin asiakkaana. Oletin näissä ennusteissa epärealistisesti, että henkilövuosittaiset todennäköisyydet saada näitä palveluita ovat toisistaan riippumattomia samoilla henkilöillä. Tämä oletus on epärealistinen, koska henkilö, joka on jonain vuonna palveluasumisessa, lienee todennäköisemmin edelleen palveluasumisessa seuraavana vuonna kuin henkilö, joka ei ole palveluasumisessa.

Oletin lisäksi yksinkertaistaen, että mikäli henkilö saa ennusteen mukaan palveluasumispalveluita (tai on vanhainkodissa) tai kotihoidon palveluita jonain vuonna, hän saa aina ikäluokkansa ja sukupuolensa määräämän keskimääräisen euromääräisen menon verran näitä palveluita. Laskin nämä keskimääräiset

palvelunsaajien menot koko väestöä koskevista keskiarvoista, jotka on esitetty edellä taulukossa 7.1 sekä näiden palvelunsaajien osuuksista eri väestöissä.

Muiden sosiaali- ja terveyspalveluiden menojen ennusteet. Muiden terveyspalveluiden ennusteina käytin yksinkertaisesti taulukossa 7.1 esitettyjä ikäluokittaisia keskiarvoja sukupuolen mukaan.

7.2.2 Yli 62-vuotiaiden muiden julkisten palveluiden menot

Yli 62-vuotiaita koskien on olemassa muitakin julkisen talouden menoja kuin sosiaali- ja terveyspalveluihin ja eläkkeisiin (ja muihin tulonsiirtoihin) liittyviä menoja. Arvioin tämän summan suuruuden vertaamalla Stakesin (2008) raportissa esitettyä 90-vuotiaiden keskiarvoista sosiaali- ja terveysmenojen suuruutta VATT:n (2014) raportissa esitettyyn summaan koskien keskiarvoista 90-vuotiaiden ”yksilöllisten julkisten kulutusmenojen summaa”. VATT:n raportissa esitetyn kuvion mukaan kaikkien julkisten palveluiden menojen summa 90-vuotiailla oli noin 23 000 euroa per henkilö vuonna 2010 (s. 24). Stakesin (2008) raportin kuvion mukaan 90-vuotiaiden sosiaali- ja terveysmenot olivat vuonna 2006 keskimäärin noin 19 000 euroa, mikä on vuoden 2010 euroissa noin 20 500 euroa (s. 17). Tästä voidaan päätellä, että muiden kuin terveys- ja sosiaalipalveluiden menot muodostavat keskimäärin noin 11%:ia kaikista julkisten palveluiden menoista 90-vuotiaana, sillä: $(23000 - 20500)/23000 \approx 11\%$. Oletin yksinkertaisuuden vuoksi, että tämä sama suhde pätee kaikkiin yli 62-vuotiaiden ikäluokkiin, eli että muut kuin terveys- ja sosiaalimenot muodostavat noin 11 %:ia kaikista yli 62-vuotiaiden julkisten palveluiden menoista keskimäärin.

Taulukossa 7.2 on esitetty näin muodostetut ikäluokittaiset elinkaariennusteissa käytetyt keskiarvot yli 62-vuotiaiden julkisen talouden menoille.

Taulukko 7.2: *Yli 62-vuotiaiden yksilöllisten julkisten palveluiden vuosittaiset keskimääräiset menot palveluiden luokan ja ikäluokan mukaan vuoden 2011 euroissa. Luvut ovat näiden palveluiden keskimääräisiä menoja ikäluokan mukaan määritetyissä väestöissä. Lähteet: Hujanen et al. (2008), 18; VATT (2014), 24.*

Palvelun luokka	**63–64-v.**	**65–74-v.**	**75–84-v.**	**yli 84-v.**
Sosiaali- ja terveyspalvelut yhteensä	2 599	4 715	9 630	20 789
Muut julkiset palvelut yhteensä	312	566	1 156	2 495
Kaikki julkiset palvelut yhteensä	2 911	5 281	10 785	23 284

Luku 8

Maahanmuuttajien lasten elinkaarivaikutukset

8.1 Yleistä: maahanmuuttajien lasten yhteenlasketut elinkaarivaikutukset

Suomeen ensimmäisen kerran yli 15-vuotiaina muuttaneilla maahanmuuttajilla voi olla lapsia siten, että lapset ovat muuttaneet joko vanhempiensa mukana Suomeen tai sitten maahanmuuttajille on syntynyt lapsia sen jälkeen, kun he ovat muuttaneet Suomeen. Käsittelen tässä tutkimuksen luvussa sitä, kuinka yli 15-vuotiaana Suomeen ensimmäisen kerran muuttaneiden elinkaarivaikutusten ennusteisiin otetaan mukaan myös heidän lastensa elinkaarivaikutukset.

Koska maahanmuuttajien lasten osalta voisi olla epärealistista olettaa, että maahanmuuttajien lasten nettovaikutukset olisivat Suomessa syntyneisiin verrattuna samanlaiset kuin heidän vanhemmillansa, käsittelen lasten elinkaarivaikutuksia erikseen. Maahanmuuttajien lasten elinkaarivaikutusten ennustamisessa on nettovaikutusten tason lisäksi myös toinen komponentti: Kuinka monta lasta Suomeen yli 15-vuotiaina muuttaneilla on yhteensä sekä heidän mukanaan muuttaneina lapsina että Suomessa syntyneinä. Toisin sanoen maahanmuuttajien lasten elinkaarivaikutusten ennusteissa on ennustettava sekä lasten tulevat nettovaikutukset sekä heidän ennustettu lukumääränsä.

Käsittelen aluksi maahanmuuttajien lasten tulevien nettovaikutusten ennusteita maahanmuuttajaryhmittäin. Tämän jälkeen käsittelen yli 15-vuotiaana Suomeen ensimmäisen kerran muuttaneiden lasten määrien ennustuksia maahanmuuttajaryhmittäin. Tavoitteena erityisesti nettovaikutusten ennusteiden kohdalla on ottaa huomioon se epävarmuus, joka maahanmuuttajien lasten tuleviin nettovaikutuksiin sisältyy siitä syystä, että usean maahanmuuttajaryhmän kohdalla maahanmuuttajien alle 8-vuotiaina Suomeen muuttaneet tai Suomessa syntyneet lapset ovat aineistossa korkeintaan 25-vuotiaita. Toisaalta, koska Suomessa syntyneiden henkilöiden osalta 22-vuotiaiden henkilöiden osalta on heitä koskevien tietojen perusteella mahdollista ennustaa – kuten jäljempänä osoitan – heidän julkisen talouden nettovaikutuksiaan 32–35-vuotiaina, myös maahanmuuttajien lasten kohdalla hyödynnän heistä 22-vuotiaina saatavilla

olevia tietoja heidän myöhempien nettovaikutustensa ennustamiseen.

8.2 Maahanmuuttajien lasten nettovaikutusten ennustaminen

Maahanmuuttajien lasten elinkaarivaikutuksen ennusteissa on kyse ehdollisista todennäköisyysjakaumista. Maahanmuuttajien lasten nettovaikutusten ennustaminen perustuu siihen oletukseen, että lapsilla nettovaikutusten ehdolliset ennustejakaumat ovat samanlaisia kuin Suomessa syntyneellä kantaväestöllä. Ehdollisuus tarkoittaa tässä sellaisia tietoja henkilöistä, jotka vaikuttavat siihen, millaisia henkilön nettovaikutukset ovat tulevaisuudessa. Henkilöitä koskevia tietoja, joihin viittaan tässä, ovat tutkimusaineistosta löytyvät tiedot kuten pääasiallinen toiminta ja toimeentulotuen saaminen. Käsittelen näitä tietoja tarkemmin jäljempänä. Se, että maahanmuuttajien lapsilla on samanlainen elinkaarivaikutusten ehdollinen ennustejakauma kuin kantaväestöllä, tarkoittaa tutkimuksen ennusteasetelmassa sitä, että oletan maahanmuuttajien lasten elinkaarivaikutusten ehdollisten ennustejakaumien olevan 35-vuotiaana samanlaiset kuin kantaväestöllä, kun ehdollistajina käytetään tietoja henkilöistä, kun he olivat 22-vuotiaita. Perustelen seuraavaksi alla näiden ikävuosien valinnan.

8.2.1 Alle 8-vuotiaina muuttaneet ja Suomessa syntyneet lapset

Maahanmuuttajien lasten elinkaariennusteiden tulisi olla uskottavia sekä Suomessa syntyneille maahanmuuttajien lapsille sekä niille lapsille, jotka ovat muuttaneet Suomeen alle 16-vuotiaina vanhempiensa mukana. Ennusteiden uskottavuuden kannalta on tärkeää huomioida, että maahanmuuttoikä saattaa vaikuttaa siihen, millaiset maahanmuuttajan lapsen tiedot ovat 22-vuotiaana esimerkiksi koulutuksen suhteen. Mikäli tällainen vaikutus on olemassa, oletettavasti vanhempana Suomeen muuttaneilla lapsilla on esimerkiksi keskimäärin heikompi koulutus 22-vuotiaana kuin nuorempana Suomeen muuttaneilla tai Suomessa syntyneillä. Koska useassa maahanmuuttajaryhmässä Suomessa syntyneitä 22-vuotiaita maahanmuuttajien lapsia oli vain ”vähän” vuoteen 2014 ulottuvassa tutkimusaineistossa, päätin sisällyttää mukaan maahanmuuttajien lasten ennusteaineistoon myös sellaiset maahanmuuttajien lapset, jotka ovat muuttaneet Suomeen 0–7-vuotiaina. Oletin, että heillä 22-vuotiaana tiedot ovat samanlaiset kuin Suomessa syntyneillä, koska kummankin ryhmän henkilöt ovat käyneet koko suomalaisen peruskoulun.

Käytin täten maahanmuuttajien lasten elinkaariennusteiden aineistona maahanmuuttajien lapsia, jotka olivat joko syntyneet Suomessa tai muuttaneet Suomeen ensimmäisen kerran 0–7-vuotiaina. Käytän tätä ennusteaineistoa myös niiden maahanmuuttajien lasten ennusteisiin, jotka muuttavat Suomeen ensimmäisen kerran 8–15-vuotiaina. Oletettavasti ennusteista näitä vanhempina muuttaneita kohtaan tulee näin ollen positiivisempia, kuin jos ennusteaineistona olisi käytetty 8–15-vuotiaina Suomeen muuttaneiden tietoja 22-vuotiaina.

8.2.2 Tiedot maahanmuuttajien lapsista 22-vuotiaina

Poimin tiedot, jonka pohjalta laadin maahanmuuttajien lasten nettovaikutusennusteet, 22-vuotiailta henkilöiltä. Nämä maahanmuuttajien lapset olivat joko syntyneet Suomessa tai muuttaneet Suomeen alle 8-vuotiaina, kuten kirjoitin yllä. Tämän valinnan ratkaisivat aineiston asettamat rajoitteet.

Ikäryhmän ylärajan valinta. Tutkimusaineistoon, joka ulottuu vuoteen 2014 asti, kuului joidenkin maahanmuuttajaryhmien osalta hyvin vähän sellaisia maahanmuuttajien lapsia, jotka olivat vähintään 25-vuotiaita ja olivat joko muuttaneet Suomeen alle 8-vuotiana tai syntyneet Suomessa. Maahanmuuttajien lasten alle 25-vuotiaiden ikäryhmissä oli yleisesti yli 15-vuoitiaita koskien sitä vähemmän henkilöitä, mitä vanhemmasta ikäryhmästä oli kyse. Tästä syystä halusin valita mahdollisimman pienen (nuoren) iän siksi ikävuodeksi, jonka pohjalta tiedot ennusteita varten poimittaisiin. Näin valittavalle iälle muodostui yläraja.

Ikäryhmän alarajan valinta. Alarajan ikävalinnalle muodosti ensiksi se, että alle 20-vuotiaita henkilöitä koskien tutkimusaineistosta ei löydy sellaisia tietoja, jotka ennustaisivat henkilöiden tulevia nettovaikutuksia (tai eroja tulevien nettovaikutusten odotusarvossa)yli 30-vuotiaina. En valinnut ennusteaineistoiksi myöskään 20- enkä 21-vuotiaita. Tämä johtuu siitä, että mallinnusaineistona käytettävässä Suomessa syntyneiden väestössä miehillä vuoden lopun pääasiallinen toiminta oli asevelvollisuuden suorittaminen 20-vuotiailla 24 %:lla ja 21-vuotiailla 5 %:lla näistä väestöistä vuosina 2005–2011. Koska uskoin henkilön pääasiallisen toiminnan vaikuttavan nettovaikutuksiin yli 30-vuotiaana, en valinnut mallinnusaineistojen iäksi 20–21-vuotiaita, koska asevelvollisuuden suorittaminen pääasiallisena toiminta kohdistui juuri näihin ikävuosiin, enkä pitänyt asevelvollisuuden suorittamista sellaisena mittarina, joka maahanmuuttajaryhmillä ennustaisi samalla tavalla tulevia nettovaikutuksia kuin Suomessa syntyneiden kantaväestössä. Koska 22-vuotiailla kantaväestön miehillä asevelvollisuus oli pääsiallisena toimintana vain 1,3 %:lla väestöstä, pidin 22-vuotiaita sopivana alarajana ikävuodelle, jonka pohjalta tulevia nettovaikutuksia mallinnettaisiin.

Edellä mainituista syistä johtuen käytin 22-vuotiaiden tietoja mallintamaan ja ennustamaan nettovaikutuksia 32–35-vuotiaana ja palkkasummia 35–39-vuotiaana.

8.2.3 Julkisen talouden ennustetut nettovaikutukset 32–35-vuotiaana

Tietoja 22-vuotiaista käytetään mallintamaan ja ennustamaan julkisen talouden nettovaikutuksia 32–35-vuotiaana. Suomessa syntyneen kantaväestön 22-vuotiaita koskevien tietojen avulla mallinnetaan samojen henkilöiden nettovaikutuksia, kun he ovat 32–35-vuotiaita. Maahanmuuttajien lasten kohdalla puolestaan käytän tietoja heistä 22-vuotiaina ennustamaan heidän nettovaikutuksiaan 32–35-vuotiaina kantaväestön mallinnustulosten pohjalta.

Valitsin 32–35-vuotiaat mallinnuksen – ja siten ennustamisen – kohteiksi kahdesta syystä. Ensiksi halusin mallintaa 22-vuotiaiden nettovaikutuksia mahdollisimman pitkälle tulevaisuuteen, vähintään yli 30-vuotiaille. Tämä johtuu siitä, että alle 30-vuotiaista vielä merkittävä osa väestöstä on opiskelijoita,

joiden julkisen talouden nettovaikutukset ovat varsin negatiiviset. Näiden henkilöiden osalta täten alle 30-vuotiaana havaitut julkisen talouden nettovaikutukset eivät kerro siitä, mitkä ovat heidän nettovaikutuksensa parhaassa työiässä. Tässä yhteydessä tarkoitan parhaalla työiällä yli 30-vuotiaita (suurin osa väestöstä ei opiskelijoita) ja alle 60-vuotiaita (suurin osa väestöstä ei vielä eläkkeellä), eli ikäryhmiä, joissa työllisten osuus väestöstä on kaikkein korkein.

Oletus: 32–35-vuotiaiden nettovaikutukset ennustavat nettovaikutuksia parhaassa työiässä. Keskeinen oletus on täten, että henkilöiden julkisen talouden nettovaikutukset 32–35-vuotiaina ennustavat heidän nettovaikutuksiaan koko työiän – ja siten elinkaaren – yli. ”Vakioin” mallinnuksessa miesten ja naisten välisen eron vaikutuksen pois, sillä naisten julkisen talouden nettovaikutuksiin 32–35-vuotiaina verrattuna miehiin vaikuttaa lastensaanti ja perhe-etuuksien kasaantuminen naisille (ks. ikäryhmittäisistä saaduista tulonsiirroista sukupuolen mukaan tutkimuksen I osa Salminen 2015a, luku 6.2.1.1 kuviot 5 ja 6).

Julkisen talouden nettovaikutukset ovat aineistossa saatavilla vuosille 1995–2011. Tämä asettaa rajoitteet sille, kuinka suuren ikäintervallin yli 22-vuotiaita koskevien tietojen avulla voidaan mallintaa samojen henkilöiden tulevia nettovaikutuksia. Henkilöt, jotka olivat 22-vuotiaita vuonna 1995, olivat 38-vuotiaita vuonna 2011, joten 38-vuotiaat olisivat mallinnuksen yhteydessä korkein valittava ikävuosi nettovaikutusten mallintamiselle 22-vuotiaiden tietojen perusteella. Päädyin kuitenkin valitsemaan mallinnusta koskevaksi iäksi 32–35-vuotiaat, sillä tällöin sain kasvatettua mallinnusaineiston lukumäärää, koska mukaan tulivat kaikki Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosina 1995–2001. Henkilöittäiset nettovaikutukset 32–35-vuotiaina voisi summata eri tavoin. Käytin tähän summaamiseen yksinkertaisia (aritmeettisia) keskiarvoja.

8.2.4 32–62-vuotiaiden nettovaikutusennusteet 32–35-vuotiaiden ennustetuista nettovaikutuksista

Maahanmuuttajien lasten elinkaariennusteet ennustan lopullisesti yllä esitettyjen 32–35-vuotiaita koskevien ennusteiden pohjalta. Keskeisenä oletuksena tässä on ensiksi – niin kuin edellä todettiin – että maahanmuuttajien lasten nettovaikutukset ovat 32–35-vuotiaina ehdollisesti (22-vuotiaiden tiedoilla) samanlaiset kuin Suomessa syntyneellä kantaväestöllä. Toiseksi oletan, että 32–35-vuotiaille laadittuja nettovaikutusennusteita voidaan käyttää ennusteina henkilön nettovaikutuksille yli koko työiän 32-vuotiaasta 62-vuotiaaksi.

Tässä tapauksessa ennusteet ovat käytännössä ennusteita henkilön vuosittaisten nettovaikutusennusten erotukselle Suomessa syntyneen kantaväestön työikäisten ikävuosittaisten nettovaikutusennusteiden odotusarvoista. Tämä tarkoittaa sitä, että jos henkilön nettovaikutusennuste 32–35-vuotiaana on esimerkiksi 2 000 euroa pienempi kuin Suomessa syntynyttä kantaväestöä koskevan ennusteen odotusarvo, niin silloin tämän henkilön nettovaikutusennusteen odotusarvo on 32–62-vuotiaana joka vuosi 2 000 euroa *pienempi*, kuin mitä on Suomessa syntyneellä henkilöllä (kun ikä ja sukupuoli huomioidaan). Mikäli henkilön nettovaikutusennuste 32–35-vuotiaana on puolestaan esimerkiksi 3 000 euroa *suurempi* kuin Suomessa syntynyttä kantaväestöä koskevan ennusteen odotusarvo, niin silloin tämän henkilön nettovaikutusennusteen odotusarvo on 32–62-vuotiaana joka vuosi 3 000 euroa suurempi, kuin mitä on Suomessa syntyneellä

henkilöllä (kun ikä ja sukupuoli huomioidaan).

Koska 32–35-vuotiaita koskevat nettovaikutusennusteet saadaan 22-vuotiaita koskevien tietojen pohjalta, tällä tavalla 22-vuotiaita koskevien tietojen pohjalta saadaan nettovaikutusten ennusteet 32–62-vuotiaille työikäisille.

8.2.5 Maahanmuuttajien lasten 32–62-vuotiaiden nettovaikutusten ennusteiden todennäköisyysjakauma

Yllä esitetyn perusteella esitän seuraavaksi maahanmuuttajien lasten elinkaarivaikutusten ennusteiden todennäköisyysjakauman. Kuten olen johdannossa esittänyt (luvuissa 2.2 ja 2.3), elinkaarivaikutusten ennusteet ovat todennäköisyysjakaumia. Todennäköisyysjakaumat kuvaavat elinkaariennnusteisiin sisältyviä epävarmuuksia. Maahanmuuttajien lasten elinkaariennusteisiin sisältyvä epävarmuus voidaan jakaa kahteen osaan. Ensiksikin siihen epävarmuuten, joka sisältyy siihen, että 22-vuotiaita koskevien tietojen perusteella ennustetaan heidän julkisen talouden keskimääräisiä nettovaikutuksia 32–35-vuotiaina. Toinen epävarmuuden osa on sama kuin Suomessa syntyneiden työikäisten nettovaikutusennusteisiin – ja siten elinkaariennusteisiin – sisältyvä epävarmuus.

8.2.5.1 22-vuotiaita koskevien tietojen perusteella nettovaikutusten ennustejakauma 32–35-vuotiaana

Käytetään nyt Suomessa syntyneitä 22-vuotiaita koskevista havaituista tiedoista merkintää $\boldsymbol{X}_{22v}^{hav}$. Käytetään näiden samojen henkilöiden havaituista 32–35-vuotiaiden nettovaikutushavainnoista merkintää $\boldsymbol{y}_{32_35v}^{hav}$. Olkoon samalla $\boldsymbol{\phi}$ niiden parametrien satunnaisvektori (satunnaismuuttujien lista), joilla 32–35-vuotiaiden henkilöiden nettovaikutuksia mallinnetaan näiden henkilöiden 22-vuotiaana havaittujen tietojen perusteella. Tarkoituksena on määrittää 32–35-vuotiaille nettovaikutusten ennusteet, $y_{32_35v}^{enn}$, heitä koskevien 22-vuotiaana havaittujen tietojen perusteella.

$y_{32_35v}^{enn}$:n posterioriennustejakauma on määritettyjä merkintöjä noudattaen (ks. tarkemmin posterioriennustejakaumien muodostamisesta luvut 2.2.4 ja 2.2.5):

$$p(y_{32_35v}^{enn}|\boldsymbol{y}_{32_35v}^{hav}, \boldsymbol{X}_{22v}^{hav}) \propto \int_{\boldsymbol{\phi}} p(y_{32_35v}^{enn}|\boldsymbol{y}_{32_35v}^{hav}, \boldsymbol{X}_{22v}^{hav}, \boldsymbol{\phi}) p(\boldsymbol{y}_{32_35v}^{hav}|\boldsymbol{X}_{22v}^{hav}, \boldsymbol{\phi}) p(\boldsymbol{\phi}) d\boldsymbol{\phi}. \tag{8.1}$$

Näin määritetty posterioriennustejakauma perustuu $\boldsymbol{\phi}$:n posteriorijakaumaan, joka voidaan esittää Bayesin kaavan (teoreeman) mukaisesti verrannollisuutena:

$$p(\boldsymbol{\phi}|\boldsymbol{y}_{32_35v}^{hav}, \boldsymbol{X}_{22v}^{hav}) \propto p(\boldsymbol{y}_{32_35v}^{hav}|\boldsymbol{X}_{22v}^{hav}, \boldsymbol{\phi}) p(\boldsymbol{\phi}). \tag{8.2}$$

Kuten olotilamallinnuksen, nuorten ja työikäisten nettovaikutusmallinnuksen sekä palkkasummamallinnuksen yhteydessä, myös tässä luvussa tavoitteena on selvittää $\boldsymbol{\phi}$:n posteriorijakauma, ja lauseketta 8.2 hyödyntämällä laatia posterioriennustejakauma 32–35-vuotiaiden nettovaikutuksille, jota voidaan käyttää osana elinkaariennusteiden luomista.

8.2.5.2 32–35-vuotiaita koskevien ennusteiden perusteella 32–62-vuotiaiden nettovaikutusten ennusteiden todennäköisyysjakauma

Keskistetään nyt kaikki 32–35-vuotiaita koskevat ennusteet niin, että yksittäisistä ennusteista, $y^{enn}_{32_35v}$, vähennetään 32–35-vuotiaiden havaittujen nettovaikutusten, $\boldsymbol{y}^{hav}_{32_35v}$, euromääräinen keskiarvo. Kutsutaan näitä ennusteita erotusennusteiksi y^{enn}_{erotus}:

$$y^{enn}_{erotus} = y^{enn}_{32_35v} - \text{keskiarvo}(\boldsymbol{y}^{hav}_{32_35v}). \tag{8.3}$$

Tässä keskiarvo$(\cdot)$ on "keskiarvofunktio", jossa sen argumentista otetaan keskiarvo. Nyt erotusennusteella y^{enn}_{erotus} on muuten sama posterioriennustejakauma kuin $y^{enn}_{32_35v}$:lla, mutta nyt $y^{enn}_{32_35v}$:n posterioriennustejakaumaa on siirretty Suomessa syntyneiden 32–35-vuotiaiden keskimääräisen nettovaikutuksen verran. Näin 32–35-vuotiaita koskevista ennusteista, $y^{enn}_{32_35v}$, saadaan ennustejakauma sille, kuinka paljon maahanmuuttajien lasten työikäisten nettovaikutusennusteet poikkeavat vuosittain Suomessa syntyneen kantaväestön ennusteista 32–62-vuotiaana.

Mikäli tätä poikkeamaa tai erotusta ei huomioitaisi, maahanmuuttajien lasten työikäisten nettovaikutusennusteet olisivat täsmälleen samanlaiset kuin Suomessa syntyneellä kantaväestöllä. Mikäli Suomessa syntyneen kantaväestön yksittäistä, yhtä vuotta koskevaa työiän nettovaikutusennustetta merkitään $y^{enn}_{\text{työikä}}$, niin maahanmuuttajien lasten vastaavat nettovaikutusennusteet 32–62-vuotiaina ovat muotoa: $y^{enn}_{\text{työikä}} + y^{enn}_{erotus}$. Kuten edellä kirjoitin, mikäli maahanmuuttajan lapsen erotusennuste, y^{enn}_{erotus}, on nollaa pienempi, tällä maahanmuuttajan lapsella työiän nettovaikutusennusteet ovat odotusarvoisesti negatiivisempia 32–62-vuotiaana kuin Suomessa syntyneellä. Mikäli erotusennuste on puolestaan nollaa suurempi, maahanmuuttajan lapsella työiän nettovaikutusennusteet ovat odotusarvoisesti positiivisempia 32–62-vuotiaana kuin Suomessa syntyneellä.

Puuttuvat epävarmuudet maahanmuuttajien lasten 32–62-vuotiaiden nettovaikutusennusteissa. On tärkeää huomioida, että nyt maahanmuuttajien lasten nettovaikutusten ennusteista jää 32–62-vuotiaiden osaltakin puuttumaan yllä kuvatun ratkaisun myötä (ainakin) kahden tekijän tuottama lisäepävarmuus ennusteisiin. Ensiksi oletan, että maahanmuuttajien lasten elinkaariennusteet ovat samanlaiset kuin kantaväestöllä, kun henkilön nettovaikutusten keskiarvo tiedetään 32–35-vuotiaana. Jopa siinäkin tapauksessa, että tämä oletus pitäisi *keskimäärin* paikkansa (mitä epäilen), tälle vaikutukselle voisi määrittää todennäköisyysjakauman, joka kuvaisi esimerkiksi sitä, että tämä vaikutus vaihtelee henkilöittäin.

Toiseksi oletan, että maahanmuuttajien lasten 22-vuotiaiden tiedoista voidaan ennustaa henkilöiden nettovaikutuksia 32–35-vuotiaana samoin parametrien ϕ (posteriori)jakaumin kuin Suomessa syntyneen kantaväestön osalta. Näiden parametrien posteriorijakaumat on estimoitu havaintoaineistosta, joka perustuu Suomessa syntyneeseen väestöön, jotka olivat 22-vuotiaita vuosina 1995–2001 ja 32–35-vuotiaita vuosina 2005–2011. Ymmärrettävästi näiden parametrien jakaumat saattavat olla erilaisia maahanmuuttajien lapsia koskevissa ennusteissa. Tästä aiheutuvat lisäepävarmuus ϕ:n todennäköisyysjakaumassa jää nyt kuitenkin pois maahanmuuttajien lasten elinkaariennusteista.

8.2.6 32–62-vuotiaiden palkkasummaennusteiden ero verrattuna kantaväestöön 35–39-vuotiaiden ennustetuista palkkasummista

Edellisissä luvuissa määritettiin, kuinka 22-vuotiaita koskevien tietojen avulla voidaan mallintaa nettovaikutuksia 32–35-vuotiaana ja kuinka näitä tuloksia voidaan käyttää ennustamaan (maahanmuuttajien lasten) nettovaikutuseroja verrattuna kantaväestöön 32–62-vuotiaana. Laadin vastaavalla tavalla 22-vuotiaiden tietojen pohjalta palkkasummien ennusteet 35–39-vuotiaille ja käytän näitä tuloksia ennustamaan (maahanmuuttajien lasten) palkkasummaennusteiden eroja verrattuna kantaväestöön 32–62-vuotiaana. Näin saatuja maahanmuuttajien lasten

Palkkasummaennusteiden erot 32–62-v: Hurdle-ero, Gamma1-tn-ero, Gamma1-ero, Gamma2-ero. Palkkasummaennusteet koostuvat neljästä todennäköisyysjakaumasta: nk. hurdlen jakaumasta, gamma1-jakaumasta, gamma2-jakaumasta sekä ennusteen todennäköisyydestä kuulua gamma1-jakaumaan (ks. tarkemmin palkkasummamallinnusta käsittelevä luku 5). Maahanmuuttajien 22-vuotiaiden tietoja käyttäen näille neljälle palkkasummien komponentille laaditaan ennusteet 35–39v-vuotiaaksi. Ja näiden komponenttien ennustettuja eroja 35–39-vuotiaana verrattuna kantaväestön 35–39-vuotiaiden keskiarvoihin käytetään ennustamaan eroja näissä komponenteissa 32–62-vuotiaana.

Hurdle kertoo yksinkertaisesti todennäköisyyden sille, että palkkasummaennuste saa arvokseen nollan. Ennustetut hurdle-erot (35–39-vuotiaana) kertovat toisin sanoen, mikä on Irakissa ja Somaliassa syntyneiden maahanmuuttajien lasten todennäköisyyksien ero verrattuna kantaväestöön 32–62-vuotiaana, että palkkasumma on vuoden aikana tasan nolla.

Nk. gamma1-jakauma koskee karkeasti ottaen vain osan vuotta töissä olevien palkkasummien ennustejakaumaa, ja gamma2-jakauma puolestaan koskee koko vuoden täysaikaisissa töissä olevien palkkasummien ennustejakaumaa. ”Gamma1-tn” on todennäköisyys sille, että palkkasummaennuste kuuluu gamma1-jakaumaan, eli että henkilö ei ole koko vuotta täysiaikaisessa työssä. ”Gamma1-tn-ero” on näin ollen tämän todennäköisyyden keskimääräinen ero verrattuna kantaväestöön. Ennustettuja erot (35–39-vuotiaana) kertovat toisin sanoen, mikä on Irakissa ja Somaliassa syntyneiden maahanmuuttajien lasten todennäköisyyksien ero verrattuna kantaväestöön 32–62-vuotiaana, että henkilö on vain osan vuotta töissä (mikäli hän on töissä).

Gamma1-jakauma on vain osan vuotta töissä olevien palkkasummaennusteiden jakauma. Ennustettuja euromääräiset erot (35–39-vuotiaana) kertovat, mikä on Irakissa ja Somaliassa syntyneiden maahanmuuttajien lasten keskimääräinen euromääräinen ero palkkasummissa verrattuna kantaväestöön 32–62-vuotiaan, kun henkilö on vain osan vuotta töissä. Gamma2-jakaumaa koskien laaditaan samanlaiset ennusteet, mutta nyt nämä ennusteet koskevat koko vuoden kokoaikatöissä olevien palkkasummaennusteiden euromääräisiä eroja verrattuna kantaväestöön.

Alla luvussa 8.5 laaditaan kuvattu palkkasummaennusteiden eromalli 35–39-vuotialle 22-vuotiaiden tietoja perusteella. Elinkaarivaikutusten ennusteiden yhteydessä luvussa 9.4.5 esitetään maahanmuuttajien lasten palkkasummien keskimääräiset eroennusteet verrattuna kantaväestöön palkkasummien esitetyille neljälle komponentille.

Maahanmuuttajien lasten eläkkeisiin liittyvien vaikutusten ennusteet. Maahanmuuttajien lapsille lasketaan ennustetuista palkkasummaennusteista ennustetut eläkkeisiin liittyvien vaikutusten ennusteet, kuten on esitetty luvussa 6.2. Samalla tavalla eläkkeisiin liittyvien vaikutusten ennusteet laaditaan muillekin ennustepopulaatioille ennustetuista palkkasummista. Koska maahanmuuttajien lasten ennustettuihin palkkasummaennusteisiin sisältyvät palkkasummaennusteiden erot verrattuna kantaväestöön, näin ollen myös ennustettuihin eläkevaikutuksiin sisältyy eroja verrattuna kantaväestöön.

8.3 Mallinnus- ja ennusteaineistot

8.3.1 Suomessa syntyneiden 22-vuotiaiden mallinnusaineisto

Käytin Suomessa syntyneiden havaintoaineistoa mallintamaan henkilöiden nettovaikutuksia 32–35-vuotiaina perustuen tietoihin, joita heistä on 22-vuotiaina, kuten edellä selostin. Havaintoaineisto muodostuu Suomessa syntyneistä henkilöistä, jotka olivat 22-vuotiaita vuosina 1995–2001 ja elossa Suomessa 32–35-vuotiaina vuosina 2005–2011. Tämä havaintoaineisto muodostaa osajoukon tutkimuksen I osassa esitellystä tutkimusaineistosta (ks. Salminen 2015a, luku 5). Tutkimuksen I osassa esitellyn aineiston Suomessa syntyneet muodostavat noin yhden prosentin yksinkertaisen satunnaisotoksen henkilöistä, joilla on ollut Suomi kotikuntana 15–70-vuotiaina vähintään yhtenä kalenterivuonna vuosina 1988–2011.

8.3.2 Maahanmuuttajien lasten määritelmästä ja ennusteissa käytettävästä osajoukosta

Maahanmuuttajan lapsi on laajempi käsite kuin tutkimuksen I osassa käytetty käsite *toisen sukupolven maahanmuuttaja* (ks. Salminen 2015a, luku 3). Toisen sukupolven maahanmuuttaja vastaa määritelmänä Tilastokeskuksen käyttämää *taustamaan* käsitettä (ks. SVT: *väestörakenne: käsitteet*). Ero toisen sukupolven maahanmuuttajan ja maahanmuuttajan lapsen välillä on, että maahanmuuttajien lapsiin kuuluvat myös ulkomailla syntyneet henkilöt, kun toisen sukupolven maahanmuuttajiin kuuluvat vain Suomessa syntyneet maahanmuuttajien lapset. Toisen sukupolven maahanmuuttajat ovat toisin sanoen osajoukko kaikista maahanmuuttajien lapsista. Lisäksi tässä tutkimuksessa sisällytän maahanmuuttajien lapsiin kaikki ne henkilöt, joilla on ainakin yksi tiedossa oleva vanhempi, joka on syntynyt ulkomailla. Toisen sukupolven maahanmuuttajat määritin tutkimuksen I osassa niin, että joukkoon kuuluvat vain ne henkilöt, joiden kaikki tiedossa vanhemmat ovat syntyneet ulkomailla.

Mikäli tutkimusaineistossa olisi ollut kaikkien 10 suurimman maahanmuuttajaryhmän kohdalla enemmän kuin joitakin kymmeniä 22-vuotiaita toisen sukupolven maahanmuuttajia, olisin valinnut maahanmuuttajien lasten elinkaarivaikutusten ennusteaineistoihin vain toisen sukupolven maahanmuuttajia – eli vain Suomessa syntyneitä. Valitsin kuitenkin maahanmuuttajien lapsia koskevan laveamman määritelmän, joten ennusteaineistoon kuuluu myös henkilöitä, jotka ovat syntyneet ulkomailla. Näin maahanmuuttajaryhmäkohtaisiin ennusteaineistoihin kuuluu kymmenien henkilöiden sijasta satoja henkilöitä. Valitsin ulkomailla syntyneiden lasten osalta ennusteaineistoihin kuitenkin vain alle 8-vuotiaina Suomeen muut-

taneet, jotta heitä koskevien ennusteiden tulosten pohjalta voisin ennustaa Suomessa syntyneiden maahanmuuttajien lasten elinkaarivaikutuksia. Uskottavuus perustuu tässä oletukseen, että Suomessa koko peruskoulun käyneet maahanmuuttajien lapset eivät eroa 22-vuotiaina alempana esiteltävien havaittujen tietojen pohjalta Suomessa syntyneistä maahanmuuttajien lapsista.

Rajasin maahanmuuttajien lasten 22-vuotiaita koskevan aineiston koskemaan vain vuosia 2005–2014. Maahanmuuttajaryhmäkohtaisiin ennusteaineistoihin kuuluvat siten vain ne maahanmuuttajien Suomessa syntyneet tai alle 8-vuotiaina Suomeen ensimmäisen kerran muuttaneet lapset, jotka olivat 22-vuotiaita vuosina 2005–2014. Joidenkin maahanmuuttajaryhmien kohdalla Suomessa syntyneet maahanmuuttajien lapset olivat 22-vuotiaita jo 1990-luvulla. Koska kalenterivuodella saattaisi olla vaikutusta 22-vuotiaiden maahanmuuttajaryhmäkohtaisiin tietoihin, halusin välttää tämän mahdollisen kalenterivuosivaikutuksen maahanmuuttajaryhmien välillä rajoittamalla aineiston koskevaan vain suhteellisen lyhyttä ja historiallisesti lähellä olevaa ajanjaksoa.

Sisällytin mukaan ennusteaineistoihin ne henkilöt, joiden vanhemmista vähintään yksi on syntynyt ulkomailla. Maahanmuuttajaryhmäkohtaiset lapsimäärät kasvoivat yleisesti ottaen alle 10 prosenttia verrattuna lapsimääriin, joihin olisin ottanut mukaan vain ne, joiden kaikki tiedossa olevat vanhemmat ovat syntyneet ulkomailla. Päädyin tähän ratkaisuun siitä syystä, että lapsimääräennusteet lasketaan hedelmällisyysluvuista, jotka perustuvat puolestaan naisten saamien lasten määriin naisten syntymävaltion mukaan. Tällöin osa lapsista voi olla sellaisia, joiden kohdalla isä onkin Suomesta, ja tästä syystä sisällytin maahanmuuttajien lasten mallinnusaineistoihin myös lapset, joissa toinen vanhemmista voi olla Suomessa syntynyt.

Muuttopäivämäärä Suomeen sisältyy tutkimusaineiston rekisteritietoihin, ja tämän tiedon avulla laskin ensimmäisen Suomeen muuton iän jokaiselle maahanmuuttajan lapselle. Tutkimusaineisto sisältää kaikki ulkomailla syntyneet maahanmuuttajat (myös maahanmuuttajien lapset) ja Suomessa syntyneet toisen sukupolven maahanmuuttajat, joilla on ollut kotikunta Suomessa 15–70-vuotiaina vähintään yhtenä kalenterivuonna vuosina 1988–2014.

8.3.3 Havaitut tiedot 22-vuotiaista maahanmuuttajien lapsista

Maahanmuuttajien lasten maahanmuuttajaryhmäkohtaiset nettovaikutusennusteet perustuvat tietoihin maahanmuuttajien Suomessa syntyneistä tai alle 8-vuotiaana Suomeen ensimmäisen kerran muuttaneista lapsista 22-vuotiaina. Esitän alla maahanmuuttajaryhmäkohtaisesti nämä tiedot. Suomessa syntyneiden ja alle 8-vuotiaana muuttaneiden lisäksi esitän samat tiedot koskien myös niitä maahanmuuttajien lapsia, jotka ovat muuttaneet Suomeen ensimmäisen kerran 8–14-vuotiaina sekä 15–20-vuotiaina. Kymmenen suurimman maahanmuuttajaryhmän lisäksi esitän samat 22-vuotiaita koskevat tiedot Suomessa syntyneille.

Maahanmuuttajien 22-vuotiaista lapsia koskien esitän tietoja koskien pääsiallista toimintaa, koulutusta, syrjäytyneisyyttä, toimeentulotuen saamista, menoja rikosasioista, ulosottomenoista ja asumisesta Uudellamaalla. Nämä rekisteritiedot ovat sen vuoden lopusta, jolloin henkilöt täyttivät 22 vuotta.

Pääasiallinen toiminta 22-vuotiaana. Olen määrittänyt pääasiallisen toiminnan tutkimuksen I osassa viisiluokkaiseksi (ks. Salminen 2015a, luku 3), mikä perustuu Tilastokeskuksen luokitukseen. Nämä viisi luokkaa ovat: työllinen, opiskelija, työtön, eläkeläinen ja muu työvoiman ulkopuolella. Viimeiseen luokkaan kuuluvat määritelmässäni asevelvollisuuttaan suorittavat. Maahanmuuttajien 22-vuotiaiden lasten tiedoissa olen yhdistänyt luokat eläkeläinen ja muu työvoiman ulkopuolella yhdeksi luokaksi, sillä eläkeläisiä on vain hyvin pieni osuus 22-vuotiaista. Lisäksi joidenkin maahanmuuttajaryhmien kohdalla yhdistin tähän luokkaan myös työttömät, mikäli alle 5 henkilöä olisi muuten muodostanut muuten yksittäisen osuustiedon. Valitsin pääasiallisen toiminnan ja mainitun luokituksen 32–35-vuotiaiden nettovaikutuksen mallintamiseen 22-vuotiaiden tiedoilla, koska oletin, että 22-vuotiaina opiskelijoilla tai työllisillä olleilla olisi odotusarvoisesti positiivisemmat nettovaikutukset 32–35-vuotiaina kuin luokkaan työttömät, eläkeläiset ja muu työvoiman ulkopuolella kuuluneilla, kun muut vaikuttavat tekijät pidetään vakioina.

Suoritettu 2. asteen tutkinto 22-vuotiaana. Tutkimusaineisto sisältää tiedon korkeimmasta suoritetusta tutkinnosta. Peruskoulun tai ensimmäisen asteen koulutuksen suorittamistietoa ei tosin löydy. Koska useimpien 22-vuotiaiden henkilöiden kohdalla korkein suoritettu tutkinto on (olemassa olevilla tiedoilla) toisen asteen tutkinto (tai sen puute), tieto, joka kertoo korkeimmasta koulutuksesta, on: henkilöllä on vähintään 2. asteen koulutus. Mikäli henkilöllä on ollut 22-vuotiaana myös korkeakoulututkinto, niin hän kuuluu myös tähän luokkaan. Valitsin kaksiluokkaisen koulutusluokituksen (on 2. asteen koulutus tai ei ole 2. asteen koulutusta) mukaan 32–35-vuotiaiden nettovaikutuksen mallintamiseen 22-vuotiaiden tiedoilla, koska oletin, että 22-vuotiaina 2. asteen tutkinnon suorittaneilla olisi odotusarvoisesti positiivisemmat julkisen talouden nettovaikutukset kuin niillä henkilöillä, joilla 2. asteen tutkintoa ei ollut, kun muut vaikuttavat tekijät pidetään vakioina. Mukana alla olevissa taulukoissa on myös tieto ylioppilastutkinnon suorittamisesta. En kuitenkaan käyttänyt tietoa ylioppilastutkinnon suorittamisesta nettovaikutusten mallintamisessa, sillä tietoa ei löydy 1990-lukua koskien tutkimusaineistosta.

Syrjäytyneet 22-vuotiaat. Summatakseni yhteen yllä esitettyjä tietoja 22-vuotiaiden pääasiallisesta toiminnasta ja koulutuksesta esitän ryhmittäin syrjäytyneiden osuudet maahanmuuttajaryhmittäin sekä Suomessa syntyneille. Esitän syrjäytyneiden osuudet eriteltynä sukupuolen mukaan sekä koko väestöä koskien yhteensä. Mahdolliset erot sukupuolten välillä syrjäytyneiden osuuksissa maahanmuuttoiän mukaan ovat mielestäni kiinnostavia tuloksia yleisen maahanmuuttokeskustelun kannalta, sillä erot sukupuolten välisissä nettovaikutuksissa ovat joidenkin maahanmuuttajaryhmien kohdalla suuria ainakin ensimmäisen sukupolven maahanmuuttajien osalta, kuten tämän tutkimuksen I osassa olen osoittanut.

Syrjäytyneeksi määritän sellaisen 22-vuotiaan, jolla ei ole vähintään 2. asteen koulutusta eikä hän ole työllinen eikä opiskelemassa. Tämä määritelmä on lähellä Pekka Myrskylän käyttämää määritelmää (ks. YLE: 15.9.2012) sekä OECD:n käyttämää määritelmää (ks. YLE: 15.9.2016) syrjäytyneestä. Myrskylän käyttämästä määritelmästä oma määritelmäni eroaa siinä, että Myrskylän määritelmässä syrjäytyneitä *eivät* ole henkilöt, jotka ovat kotona hoitamassa lasta, vaikka he muuten täyttäisivät käyttämäni syrjäytyneen määritelmän. Joka tapauksessa 22-vuotiaille Suomessa syntyneille alla esittämäni luvut syrjäytyneiden 22-vuotiaiden osuudesta ovat lähelllä Myrskylän lukua 15–29-vuotiaille suomen- ja ruotsinkielisille. OECD:n määritelmässä ei huomioida suoritettua 2. asteen tutkintoa, joten OECD:n määritelmän mukaan syrjäytyneiden osuus on korkeampi kuin Myrskylän ja itse käyttämäni määritelmän mukaan mitattuna.

Toimeentulotuen saaminen 22-vuotiaana. Tieto koskien toimeentulon saamista vuoden aikana kertoo, onko henkilö saanut yli nolla euroa toimeentulotukea sen vuoden aikana, kun täytti 22 vuotta. Valitsin tiedon 32–35-vuotiaiden nettovaikutuksen mallintamiseen 22-vuotiaiden tiedoilla, koska oletin, että toimeentulotuen saajilla olisi odotusarvoisesti negatiivisemmat nettovaikutukset 32–35-vuotiaana kuin niillä henkilöillä, jotka eivät olleet saanut toimeentulotukea, kun muut vaikuttavat tekijät pidetään vakioina. Kokeilin mallinnusta myös niin, että jaoin toimeentulotuen saadun määrän muutamaan luokkaan, mutta erot vaikutuksissa olivat pieniä lukuun ottamatta luokkaa, joka koski henkilöitä, jotka eivät olleet saanet lainkaan toimeentulotukea.

Rikosasioita ja ulosottoja koskevat julkisen talouden menot 22-vuotiaana. Tutkimuksen I osassa rikosasioihin, ulosottoihin ja julkiseen oikeusapuun liittyviä oikeusmenoja käsitellään luvussa 6.5.1 (Salminen 2015a). Tieto menoista rikosasioista vuoden aikana kertoo, onko henkilöön liittyen rikosasioihin liittyviä menoja (yli nolla euroa) sen vuoden ajalta, jolloin henkilö täytti 22 vuotta. Nämä menot koostuvat mm. esitutkintamenoista, syyttämismenoista, oikeudenkäyntimenoista ja rangaistusten täytäntöönpanojen menoista, ja käytetyt rekisteritiedot ovat peräisin *Tilastokeskuksen Syytetyt, tuomitut ja rangaistukset* -tilastosta, eli henkilö ei aiheuta menoja tämän määritelmän mukaan, jos hän on ollut vain rikoksen uhrina. Tieto ulosottomenoista kertoo, onko henkilöön liittyen yli nolla euroa ulosottomenoja sinä vuonna, kun henkilö täytti 22 vuotta.

Oikeusmenot 22-vuotiaana. Tieto oikeusmenoista vuoden aikana puolestaan kertoo, onko henkilöön liittyen ollut yli nolla euroja menoja joko rikosasioihin, ulosottoihin tai saatuun julkiseen oikeusapuun liittyen. Lisäksi, mikäli henkilö oli saanut rikosasioihin liittyen sakkoja, tässä tapauksessa määrittelin henkilön kuulumaan samaan luokkaan kuin ne, joihin liittyen oli yli nollan euron verran menoja edellä mainituista syistä johtuen. En voinut kuitenkaan käyttää ulosottomenotietoa enkä oikeusmenotietoa mallinnuksessa, sillä ulosottomenoja ja oikeusapumenotietoja ei löydy tutkimusaineistosta 1990-luvulle eikä siten myöskään oikeusmenotietoa. Rikosasioihin liittyvä menotieto kuitenkin löytyy, ja valitsin tämän tiedon 35-vuotiaiden nettovaikutuksen mallintamiseen 22-vuotiaiden tiedoilla, koska oletin, että niillä henkilöillä, joihin liittyen on 22-vuotiaina rikosasioihin liittyviä menoja, olisi odotusarvoisesti negatiivisemmat nettovaikutukset 32–35-vuotiaina kuin niillä henkilöillä, joihin liittyen rikosasioihin liittyviä menoja ei ollut.

Kotikunta Uudellamaalla 22-vuotiaana. Uusimaa-indikaattori kertoo, asuiko henkilö Uudellamaalla (vai jossain toisessa maakunnassa) sen vuoden lopussa, kun täytti 22 vuotta. Valitsin tiedon mukaan, koska oletin, että Uudellamaalla asuneilla olisi odotusarvoisesti positiivisemmat nettovaikutukset 32–35-vuotiaina kuin muualla asuneilla.

8.3.4 Maahanmuuttajaryhmäkohtaiset tiedot 22-vuotiaista maahanmuuttajien lapsista vuosina 2005–2014

Alla taulukoissa 8.1–8.10 on esitetty maahanmuuttajaryhmittäin maahanmuuttajien lapsia koskevia tietoja. Tiedot koskevat tietoja sen vuoden lopussa, kun maahanmuuttajien lapset täyttivät 22 vuotta. Tiedot ovat vuosilta 2005–2014. Tiedot on jaettu taulukoissa sarakkeittain maahanmuuttoiän mukaan: Luokka ”Alle 8” sisältää alle 8-vuotiaana Suomeen muuttaneet maahanmuuttajien lapset sekä Suomessa syntyneet maahanmuuttajien lapset. Sarake ”Yht.” koskee kaikkia alle 21-vuotiaina Suomeen muuttaneita maahanmuuttajien lapsia yhteensä. Taulukoiden oikeanpuoleisimmassa sarakkeessa on esitetty vertailun vuoksi koko Suomessa syntyneen väestön 22-vuotiaita koskevan otoksen tiedot vuosina 2005–2011.

Henkilölukumääriä lukuun ottamatta kaikki muut tiedot taulukoissa ovat prosenttilukuja. Poimitaan esimerkiksi taulukossa 8.1 riviä ”Ulosottomenoja vuoden aikana” ja saraketta ”Alle 8” vastaavan solun tieto: 8 (%). Luku tarkoittaa sitä, että alle 8-vuotiaana Suomeen muuttaneista (tai Suomessa syntyneistä) kahdeksalla prosentilla oli rikosasioista johtuvia julkisen talouden menoja sinä vuonna, kun he täyttivät 22 vuotta vuosien 2005–2014 aikana. Tässä tapauksessa on kyse maahanmuuttajien lapsista, joiden vanhemmista ainakin yksi on syntynyt Irakissa. Vertailun vuoksi koko Suomessa syntynyttä väestöä koskevassa otoksessa rikosasioista johtuvia julkisen talouden menoja oli kolmella prosentilla sinä vuonna, kun he täyttivät 22 vuotta vuosien 2005–2011 aikana.

Taulukoissa ei ole esitetty prosenttitietoja niitä soluja koskien, joissa esitetty prosenttiluku koskisi alle viittä henkilöä. Näissä tapauksissa taulukoiden solut on korvattu ajatusviivalla: –.

Taulukko 8.1: *Henkilölukumäärät (N) ja tiedot (%). Irakissa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Irak	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**398**	432	244	1 074	*4 271*
Pääasiallinen toiminta: työllinen	**39**	37	25	35	*60*
Pääasiallinen toiminta: opiskelija	**22**	24	36	26	*28*
Pääasiallinen toiminta: työtön	**14**	21	18	18	*5*
Pääasiallinen toiminta: muu	**25**	19	20	21	*7*
On ylioppilastutkinto	**23**	15	4	16	*63*
On vähintään 2. asteen koulutus	**53**	40	17	40	*84*
Syrjäytyneet, miehet	**21**	23	29	24	*5*
Syrjäytyneet, naiset	**26**	28	36	29	*4*
Syrjäytyneet, kaikki yhteensä	**24**	25	32	26	*4*
Saaneet toimeentulotukea vuoden aikana	**50**	50	52	50	*11*
Menoja rikosasioista vuoden aikana	**8**	10	7	9	*3*
Ulosottomenoja vuoden aikana	**31**	27	15	25	*3*
Kotikunta Uudellamaalla	**54**	44	40	47	*28*

Taulukko 8.2: *Henkilölukumäärät (N) ja tiedot (%). Somaliassa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Somalia	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**567**	562	419	1 548	*4 271*
Pääasiallinen toiminta: työllinen	**36**	43	32	37	*60*
Pääasiallinen toiminta: opiskelija	**24**	22	20	22	*28*
Pääasiallinen toiminta: työtön	**10**	11	19	12	*5*
Pääasiallinen toiminta: muu	**31**	24	29	28	*7*
On ylioppilastutkinto	**20**	9	2	11	*63*
On vähintään 2. asteen koulutus	**35**	29	10	26	*84*
Syrjäytyneet, miehet	**37**	26	33	32	*5*
Syrjäytyneet, naiset	**29**	28	56	36	*4*
Syrjäytyneet, kaikki yhteensä	**34**	27	43	34	*4*
Saaneet toimeentulotukea vuoden aikana	**49**	45	58	50	*11*
Menoja rikosasioista vuoden aikana	**9**	10	6	8	*3*
Ulosottomenoja vuoden aikana	**49**	34	22	36	*3*
Kotikunta Uudellamaalla	**92**	86	67	83	*28*

Taulukko 8.3: *Henkilölukumäärät (N) ja tiedot (%). Saksassa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Saksa	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**39**	29	15	83	*4 271*
Pääasiallinen toiminta: työllinen	**44**	55	–	–	*60*
Pääasiallinen toiminta: opiskelija	**41**	21	–	–	*28*
Pääasiallinen toiminta: työtön	**0**	–	–	–	*5*
Pääasiallinen toiminta: muu	**15**	–	–	–	*7*
On ylioppilastutkinto	**67**	59	–	–	*63*
On vähintään 2. asteen koulutus	**77**	72	53	71	*84*
Syrjäytyneet, miehet	**–**	–	–	–	*5*
Syrjäytyneet, naiset	**–**	–	–	–	*4*
Syrjäytyneet, kaikki yhteensä	**–**	–	–	–	*4*
Saaneet toimeentulotukea vuoden aikana	**15**	–	–	–	*11*
Menoja rikosasioista vuoden aikana	**–**	–	–	–	*3*
Ulosottomenoja vuoden aikana	**–**	–	–	–	*3*
Kotikunta Uudellamaalla	**56**	52	47	53	*28*

Taulukko 8.4: *Henkilölukumäärät (N) ja tiedot (%). Ruotsissa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Ruotsi	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**158**	15	12	185	*4 271*
Pääasiallinen toiminta: työllinen	**52**	47	67	52	*60*
Pääasiallinen toiminta: opiskelija	**25**	–	–	–	*28*
Pääasiallinen toiminta: työtön	**10**	–	–	–	*5*
Pääasiallinen toiminta: muu	**14**	–	–	–	*7*
On ylioppilastutkinto	**36**	–	–	–	*63*
On vähintään 2. asteen koulutus	**66**	60	–	–	*84*
Syrjäytyneet, miehet	**14**	–	–	–	*5*
Syrjäytyneet, naiset	**12**	–	–	–	*4*
Syrjäytyneet, kaikki yhteensä	**13**	–	–	–	*4*
Saaneet toimeentulotukea vuoden aikana	**25**	–	–	–	*11*
Menoja rikosasioista vuoden aikana	**6**	–	–	–	*3*
Ulosottomenoja vuoden aikana	**23**	–	–	–	*3*
Kotikunta Uudellamaalla	**35**	47	50	37	*28*

Taulukko 8.5: *Henkilölukumäärät (N) ja tiedot (%). Kiinassa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Kiina	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**139**	109	59	307	*4 271*
Pääasiallinen toiminta: työllinen	**55**	38	36	45	*60*
Pääasiallinen toiminta: opiskelija	**30**	40	42	36	*28*
Pääasiallinen toiminta: työtön tai muu	**15**	22	22	19	*12*
On ylioppilastutkinto	**75**	59	25	60	*63*
On vähintään 2. asteen koulutus	**83**	72	42	71	*84*
Syrjäytyneet, miehet	**16**	12	24	16	*5*
Syrjäytyneet, naiset	**9**	17	19	14	*4*
Syrjäytyneet, kaikki yhteensä	**12**	15	22	15	*4*
Saaneet toimeentulotukea vuoden aikana	**–**	10	10	–	*11*
Menoja rikosasioista vuoden aikana	**–**	–	–	–	*3*
Ulosottomenoja vuoden aikana	**–**	–	–	–	*3*
Kotikunta Uudellamaalla	**76**	64	59	68	*28*

Taulukko 8.6: *Henkilölukumäärät (N) ja tiedot (%). Thaimaassa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Thaimaa	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**58**	185	94	337	*4 271*
Pääasiallinen toiminta: työllinen	**59**	56	42	52	*60*
Pääasiallinen toiminta: opiskelija	**16**	16	26	18	*28*
Pääasiallinen toiminta: työtön tai muu	**26**	29	33	29	*12*
On ylioppilastutkinto	**24**	9	–	–	*63*
On vähintään 2. asteen koulutus	**47**	44	20	38	*84*
Syrjäytyneet, miehet	**22**	18	35	23	*5*
Syrjäytyneet, naiset	**14**	16	22	18	*4*
Syrjäytyneet, kaikki yhteensä	**17**	17	28	20	*4*
Saaneet toimeentulotukea vuoden aikana	**28**	20	26	23	*11*
Menoja rikosasioista vuoden aikana	**–**	5	7	–	*3*
Ulosottomenoja vuoden aikana	**33**	22	24	25	*3*
Kotikunta Uudellamaalla	**55**	56	47	53	*28*

Taulukko 8.7: *Henkilölukumäärät (N) ja tiedot (%). Turkissa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Turkki	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**130**	158	103	391	*4 271*
Pääasiallinen toiminta: työllinen	**55**	53	50	53	*60*
Pääasiallinen toiminta: opiskelija	**18**	20	21	20	*28*
Pääasiallinen toiminta: työtön	**9**	9	11	10	*5*
Pääasiallinen toiminta: muu	**17**	18	18	17	*7*
On ylioppilastutkinto	**20**	15	–	–	*63*
On vähintään 2. asteen koulutus	**49**	39	18	37	*84*
Syrjäytyneet, miehet	**12**	21	20	18	*5*
Syrjäytyneet, naiset	**20**	18	26	21	*4*
Syrjäytyneet, kaikki yhteensä	**15**	20	22	19	*4*
Saaneet toimeentulotukea vuoden aikana	**23**	25	24	24	*11*
Menoja rikosasioista vuoden aikana	**6**	8	–	–	*3*
Ulosottomenoja vuoden aikana	**26**	20	13	20	*3*
Kotikunta Uudellamaalla	**68**	59	65	64	*28*

Taulukko 8.8: *Henkilölukumäärät (N) ja tiedot (%). Ent. Jugoslaviassa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Entinen Jugoslavia	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**626**	438	110	1 174	*4 271*
Pääasiallinen toiminta: työllinen	**63**	58	49	60	*60*
Pääasiallinen toiminta: opiskelija	**14**	16	27	16	*28*
Pääasiallinen toiminta: työtön	**10**	10	13	11	*5*
Pääasiallinen toiminta: muu	**13**	16	11	14	*7*
On ylioppilastutkinto	**20**	17	12	18	*63*
On vähintään 2. asteen koulutus	**64**	60	47	61	*84*
Syrjäytyneet, miehet	**10**	12	12	11	*5*
Syrjäytyneet, naiset	**11**	13	18	12	*4*
Syrjäytyneet, kaikki yhteensä	**11**	13	14	12	*4*
Saaneet toimeentulotukea vuoden aikana	**26**	22	36	26	*11*
Menoja rikosasioista vuoden aikana	**6**	6	6	6	*3*
Ulosottomenoja vuoden aikana	**27**	11	8	19	*3*
Kotikunta Uudellamaalla	**55**	49	27	50	*28*

Taulukko 8.9: *Henkilölukumäärät (N) ja tiedot (%). Virossa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Viro	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**875**	730	344	1 949	*4 271*
Pääasiallinen toiminta: työllinen	**63**	61	51	60	*60*
Pääasiallinen toiminta: opiskelija	**21**	18	19	20	*28*
Pääasiallinen toiminta: työtön	**6**	7	10	7	*5*
Pääasiallinen toiminta: muu	**11**	13	21	13	*7*
On ylioppilastutkinto	**40**	36	19	35	*63*
On vähintään 2. asteen koulutus	**69**	67	46	64	*84*
Syrjäytyneet, miehet	**9**	12	20	12	*5*
Syrjäytyneet, naiset	**7**	10	25	12	*4*
Syrjäytyneet, kaikki yhteensä	**8**	11	23	12	*4*
Saaneet toimeentulotukea vuoden aikana	**18**	18	26	19	*11*
Menoja rikosasioista vuoden aikana	**4**	7	4	5	*3*
Ulosottomenoja vuoden aikana	**17**	16	23	18	*3*
Kotikunta Uudellamaalla	**69**	63	58	65	*28*

Taulukko 8.10: *Henkilölukumäärät (N) ja tiedot (%). Ent. NL:ssa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Oikeanpuoleisessa sarakkeessa Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

Entinen Neuvostoliitto	**Alle 8**	8–14	15–20	Yht.	*Suomi*
Henkilölukumäärät (N)	**2 373**	3 049	1 268	6 690	*4 271*
Pääasiallinen toiminta: työllinen	**60**	54	42	54	*60*
Pääasiallinen toiminta: opiskelija	**23**	26	32	26	*28*
Pääasiallinen toiminta: työtön	**6**	8	12	8	*5*
Pääasiallinen toiminta: muu	**11**	12	14	12	*7*
On ylioppilastutkinto	**47**	38	20	37	*63*
On vähintään 2. asteen koulutus	**73**	68	55	67	*84*
Syrjäytyneet, miehet	**9**	10	15	11	*5*
Syrjäytyneet, naiset	**7**	9	12	9	*4*
Syrjäytyneet, kaikki yhteensä	**8**	10	13	10	*4*
Saaneet toimeentulotukea vuoden aikana	**20**	25	34	25	*11*
Menoja rikosasioista vuoden aikana	**4**	6	6	5	*3*
Ulosottomenoja vuoden aikana	**15**	14	11	14	*3*
Kotikunta Uudellamaalla	**56**	47	41	49	*28*

8.4 Mallinnus: julkisen talouden nettovaikutukset 32–35-vuotiaana 22-vuotiaita koskevien tietojen perusteella

8.4.1 Estimoitavien parametrien posteriorijakauman määritys ja simulointi

Esitän tässä luvussa, kuinka mallinnan Suomessa syntyneiden henkilöiden keskimääräisiä nettovaikutuksia 32–35-vuotiaana tiedoilla, joita henkilöistä on, kun he olivat 22-vuotiaita. Selitettävänä muuttujana mallinnuksessa on henkilöiden vuosittaiset keskimääräiset julkisen talouden nettovaikutukset, kun he olivat 32–35-vuotiaita. Selittävinä muuttujina ovat puolestaan edellisessä luvussa yllä esitetyt tiedot henkilöistä, kun he olivat 22-vuotiaita. Koska myös työikäisten (20–62-vuotiaiden) nettovaikutusten mallinnuksen yhteydessä (luvussa 4) selitettävänä muuttujana oli työiän nettovaikutus, mallinnus koskien 32–35-vuotiaiden nettovaikutuksia muistuttaa hyvin paljon työikäisten nettovaikutusten mallinnusta.

Tässä luvussa esitettävä 32–35-vuotiaiden nettovaikutusten mallinnus on kuitenkin yksinkertaisempi tehdä kuin työikäisten nettovaikutusmallinnuksen yhteydessä, sillä tässä tapauksessa jokaista henkilöä kohti on vain yksi selitettävän muuttujan havainto. Lisäksi mallinnuksessa ei tarvitse käytännössä huomioida maassaoloaikaan eikä kalenterivuoteen liittyviä vaikutuksia, koska havainnot ovat vain muutamalta kalenterivuodelta. Tämä tarkoittaa yksinkertaisempaa mallinnusta siinä mielessä, että 32–35-vuotiaiden nettovaikutushavainnot eivät ole hierarkkisesti rakentuneet, kuten työikäisten nettovaikutusmallinnuksen kohdalla oli. Lisäksi, koska jokaista henkilöä koskien on vain yksi havainto, perättäisten havaintojen välistä riippuvuutta ei ole – niin kuin on työikäisten nettovaikutushavaintoja koskevassa havaintoaineistossa. Koska mallinnus on näin ollen samankaltainen kuin työikäisten nettovaikutusmallinnuksen yhteydessä (luvussa 4.3.3), esitän mallinnusaskeleet tässä yhteydessä hyvin lyhyesti.

Koska elinkaarivaikutusten ennustaminen tapahtuu bayesilaisen tilastollisen päättelyn piirissä, noudatan myös 32–35-vuotiaiden nettovaikutusmallinnuksen yhteydessä bayesilaista mallinnusta. Tämä tarkoittaa sitä, että tavoitteena on selvittää estimoitavien mallinnusparametrien posteriorijakauma (ks. posteriorijakauman muodostamisesta esimerkiksi luku 4.3.5) . Tämä posteriorijakauma on verrannollinen aineistomallin määrittämän otosjakauman ja estimoitavien parametrien priorijakaumien tulolle. Kuten työikäisten nettovaikutusmallinnuksen yhteydessä, nyt on määritettävä sekä 32–35-vuotiaiden nettovaikutusten aineistomalli sekä estimoitavien parametrien priorijakaumat.

Otosjakauman määrittäminen: Vino Studentin t-jakauma. Koska 32–35-vuotiaita koskevassa nettovaikutushavaintojen aineistossa ei ole ainakaan sellaista hierarkkisuutta, joka minun mielestäni pitäisi huomioida 32–35-vuotiaiden nettovaikutusten aineistomallin määrittämisessä, oikeastaan ainoaksi tehtäväksi jää määrittää sopiva otosjakauma. Lisäksi edellä on jo määritetty käytettävät selitettävät muuttujat. Koska 32–35-vuotiaiden nettovaikutusmallinnuksen yhteydessä havainnot koskevat myös työiäkäisiä, oli mielestäni sopivaa olettaa, että 32–35-vuotiaiden nettovaikutusten jakauma noudattaisi likimain samaa todennäköisyysjakaumaa kuin kaikkien työikäisten mallinnuksessa käytetty otosjakauma. Tämä otosjakauma – josta havainnot on oletetusti generoitu – oli työikäisten nettovaikutusten lopullisessa mallissa vino Studentin t-jakauma. Kokeilin kuitenkin työikäisten nettovaikutusmallinnuksen yhteydessä otosjakaumaksi myös normaalijakaumaa sekä (ei-vinoa) Studentin t-jakaumaa. Noudatan tästä syystä samaa aineistomallien – ja otosjakaumien – hierarkkista valintaa 32–35-vuotiaiden nettovaikutusten mallinta-

miseen kuin noudatin työikäisten nettovaikutusten mallinnuksen yhteydessä.

Aineistomallit 1–4 vastaavat tästä syystä otosjakaumiltaan työikäisten nettovaikutusmallinnuksen yhteydessä esitettyjä mallinnusaskeleita (ks. luku 4.3.3). Malli 1 on tyhjä normaalimalli eli käytän otosjakaumana normaalijakaumaa, ja malliin ei sisälly lainkaan selittäviä muuttujia. Malli 2 on myös ”tyhjä”, ja siinä otosjakaumana on Studentin t-jakauma. Mallissa 3 ovat mukana kaikki selittävät muuttujat, eli tässä tapauksessa tiedot henkilöistä, kun he olivat 22-vuotiaita. Viimeisessä mallinnusaskeleessa (malli 4) otosjakaumana on ei-vinon Studentin t-jakauman sijaan vino Studentin t-jakauma.

Parametrien priorijakaumista ja muut mallinnukseen liittyvät seikat. Estimoitavien parametrien priorijakaumat ovat tässä yhteydessä myös samanlaisia – eli heikosti informatiivisia – kuin työikäisten nettovaikutusmallinnuksen yhteydessä (ks. luku 4.3.4). Tästä syystä en esitä niitä erikseen tässä yhteydessä. Samaten posteriorijakauman simulointi, simulaatioiden määriä koskevat valinnat, simulaatioketjujen konvergoitumista koskeva päättely sekä estimointitulosten perusteella tehty posterioriennustejakauman simulointi vastaavat työikäisten nettovaikutusmallinnuksen yhteydessä esitettyä (ks. luvut 4.3.5–4.3.7).

8.4.2 Havainnot, estimointitulokset ja posterioriennustetarkistus

Esitän tässä alaluvussa tulokset koskien Suomessa syntyneiden 32–35-vuotiaiden henkilöiden keskimääräisten nettovaikutusten mallinnusta niillä tiedoilla, joita henkilöistä on koskien sitä vuotta, jonka lopussa he olivat 22-vuotiaita. Esitän ensiksi mallinnuksessa käytetyn havaintoaineiston eli selitettävän muuttujan jakauman tunnusluvut sekä selittävien muuttujien keskeisimmät jakaumatiedot. Tämän jälkeen esitän mallinnuksen estimointitulokset. Posterioriennustetarkistuksessa varmistan, että käytetyn estimoidun mallin tuottamat ennusteet vastaavat jakaumaltaan alkuperäistä havaintoaineiston selitettävän muuttujan jakaumaa. Tulosten esittäminen tässä muodossa vastaa työikäisten ja nuorten nettovaikutusten mallinnuksen tulosten esittelyä luvussa 4.4. Koska käsittelen tuloksia – erityisesti estimointitulosten tulkintaa – tarkemmin mainitussa luvussa, käyn estimointitulokset vain lyhyesti läpi tässä luvussa. Alla esitettävät estimointitulokset ovat tulkinnallisesti samanlaisia työikäisten ja nuorten nettovaikutusmallinnuksen yhteydessä, sillä kummassakin tapauksessa mallinnus perustuu samoihin otosjakaumiin.

8.4.2.1 Havaintoaineisto

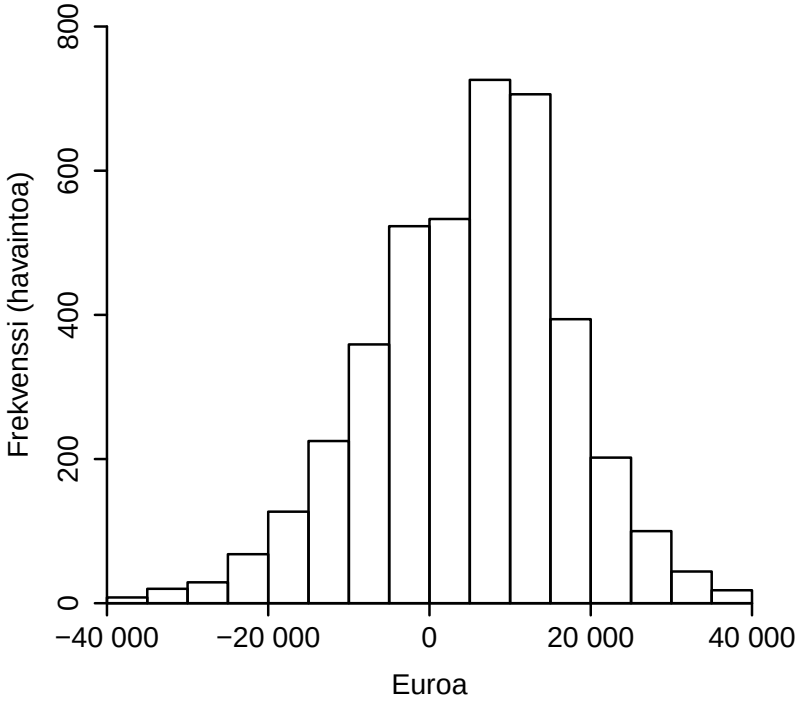
Alla taulukossa 8.11 on esitetty selitettävän muuttujan havaintolukumäärä ja havaintojen jakauman keskeisimpiä tunnuslukuja. Selitettävä muuttuja muodostuu niiden Suomessa syntyneiden henkilöiden nettovaikutusten keskiarvoista 32–35-vuotiaana, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana. Täten nettovaikutusten henkilöittäiset keskiarvot 32–35-vuotiaana koskevat vuosia 2005–2011. Aineistossa on 4 166 henkilöä, joiden nettovaikutusten keskiarvoista selitettävän muuttujan havaitut arvot koostuvat. Havaintojen keskiarvo on 4 700 euroa, ja keskihajonta on 15 900 euroa. Keskiarvo on täten suunnilleen yhtä suuri kuin kaikkien Suomessa syntyneiden työikäisten 20–62-vuotiaiden nettovaikutusten keskiarvo

(vrt. taulukko 4.12 s. 221). Havaintojen mediaani on 6 200 euroa. Koska havaintoaineiston mediaani on keskiarvoa suurempi, tämä viittaa selitettävän jakauman olevan vasemmalle vino.

Taulukko 8.11: *Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana ja 32–35-vuotiaita vuosien 2005–2011 aikana. Havainnot henkilöiden keskimääräisistä nettovaikutuksista 32–35-vuotiaana. Nettovaikutushavaintojen lukumäärä ja jakauman tunnusluvat mallinnusaineistossa. Euromääräiset tunnusluvut on pyöristetty lähimpään sataan euroon.*

	Havainnot
Kokonaislukumäärä (N)	4 166
Keskiarvo	4 700
Keskihajonta	15 900
1. prosenttipiste (1 %)	-38 900
5. prosenttipiste (5 %)	-17 500
10. prosenttipiste (10 %)	-11 800
Alakvartiili (25 %)	-3 200
Mediaani (50 %)	6 200
Yläkvartiili (75 %)	12 900
90. prosenttipiste (90 %)	19 700
95. prosenttipiste (95 %)	24 900
99. prosenttipiste (99 %)	38 800

Kuvassa 8.1 on esitetty selitettävän muuttujan jakauman histogrammi. Jakaumasta on katkaistu pois niiden ylin ja alin prosentti pois. Tämän 32–35-vuotiaita koskevan histogrammin perusteella selitettävällä muuttujalla voi olla likimain vino Studentin t-jakauma. Kaikkien Suomessa syntyneiden työikäisten 20–62-vuotiaiden osalta nettovaikutushavainnoilla on likimain vino Studentin t-jakauma, kuten luvussa 4 osoitettiin (vrt. myös kaikkien 20–62-vuotiaiden työikäisten nettovaikutusten histogrammi s. 222).



Kuva 8.1: *Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana ja 32–35-vuotiaita vuosien 2005–2011 aikana. Nettovaikutushavaintojen katkaistun jakauman histogrammi mallinnusaineistossa (yllä) sekä henkilöiden nettovaikutushavaintojen keskiarvojen katkaistun jakauman histogrammi (alla). Jakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Taulukossa 8.12 on esitetty selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Selittävät muuttujat koskevat tietoja henkilöistä sen vuoden lopussa, kun he olivat 22-vuotiaita. Kaikki selittävät muuttujat ovat tässä tapauksessa indikaattorimuuttujia, joten keskiarvot ovat indikaattorien osuuksia kaikkien havaintojen joukossa. Esimerkiksi muuttujan ”Pääasiallinen toiminta: työllinen” keskiarvo kertoo työllisten osuuden havaintoaineistossa.

Taulukko 8.12: *Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana ja 32–35-vuotiaita vuosien 2005–2011 aikana. Nettovaikutusmallien selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta). Tiedot ovat sen vuoden lopulta, kun henkilöt olivat 22-vuotiaita.*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,49	0,50	0	0	1
Pääasiallinen toiminta: työllinen	0,47	0,50	0	0	1
Pääasiallinen toiminta: opiskelija	0,34	0,47	0	0	1
Pääasiallinen toiminta: työtön	0,11	0,32	0	0	0
Pääasiallinen toiminta: muu	0,07	0,26	0	0	0
On vähintään 2. asteen koulutus	0,84	0,37	1	1	1
Saaneet toimeentulotukea vuoden aikana	0,21	0,40	0	0	0
Menoja rikosasioista vuoden aikana	0,03	0,16	0	0	0
Kotikunta Uudellamaalla	0,26	0,44	0	0	1

8.4.2.2 Estimointitulokset

Perättäisten mallien parametrien odotusarvoestimaatit

Esitin jo edellä lyhyesti käytetyt neljä mallinnusaskelta. Näitä mallinnusaskeleita vastaavat neljä perättäistä tai hierarkkista aineistomallia, jotka ovat:

- **Malli 1**: Tyhjä normaalimalli, eli otosjakaumana on normaalijakauma ilman selittäviä muuttujia.
- **Malli 2**: Tyhjä Studentin t-malli, eli otosjakaumana on Studentin t-jakauma ilman selittäviä muuttujia.
- **Malli 3**: Otosjakaumana on Studentin t-jakauma. Selittävät muuttujat ja regressiokertoimet ovat mukana mallissa.
- **Malli 4**: Otosjakaumana on vino Studentin t-jakauma.

Mallien perättäisyys tai hierarkkisuus tarkoittaa sitä, että pienemmän mallin parametrit sisältyvät aina suurempaan malliin (normaalijakauma on Studentin t-jakauma, jonka vapausasteparametrin arvo on ääretön). Taulukossa 8.13 on esitetty perättäisten mallien parametrien estimoitujen posteriorijakaumien odotusarvoestimaatit. Vakiotermin ja selittäviä muuttujia vastaavien regressiokertoimien lisäksi taulukon parametreja ovat hajontaparametri σ, t-jakauman vapausasteparametri ν sekä vinon t-jakauman vinousparametri γ. Parametrit on esitetty ja selitetty tarkemmin nuorten ja työikäisten nettovaikutusmallinnuksen yhteydessä (ks. parametrien selitetaulukko 4.1 s. 170 sekä parametrien tulosten tulkinnasta esimerkiksi Irakia koskevat tulokset luvussa 4.4.1.2).

Taulukko 8.13: *Perättäisten nettovaikutusmallien parametrien odotusarvoestimaatit ja mallikohtaiset ennustetarkkuuden (Waic ja LOO) estimaatit sekä niiden keskihajonnat.*

	Malli 1	Malli 2	Malli 3	Malli 4
Vakio	4 670	5 430	5 450	6 670
Nainen			-10 440	-10 380
Pääasiallinen toiminta: työllinen			5 320	5 220
Pääasiallinen toiminta: opiskelija			5 910	5 910
Pääasiallinen toiminta: työtön			1 600	1 530
On vähintään 2. asteen koulutus			4 530	4 490
Saaneet toimeentulotukea vuoden aika			-2 510	-2 440
Menoja rikosasioista vuoden aika			-7 320	-6 900
Kotikunta Uudellamaalla			1 700	1 810
σ	15 960	10 520	8 550	8 530
ν		3,99	3,36	3,47
γ				-840
WAIC (estimaatti)	11 837	10 559	9 278	8 711
WAIC (estimaatin keskivirhe)	(334)	(132)	(138)	(109)
LOO (estimaatti)	11 838	10 559	9 278	8 945
LOO (estimaatin keskivirhe)	(335)	(132)	(138)	(118)

Vakioparametri taulukossa 8.13 tulkitaan malleissa 1–3 niin, että se on euromääräinen odotusarvo henkilön nettovaikutusten keskiarvolle 32–35-vuotiaana, kun selittävien muuttujat saavat arvoksi keskiarvonsa (mallissa 4 pitää ottaa huomioon myös vinouden vaikutus). Olen skaalannut regressiokertoimien (taulukon 8.13 rivit ”Nainen”–”Kotikunta Uudellamaalla”) estimaatit niin, että estimaatit kertovat, kuinka monta euroa nettovaikutusten odotusarvo muuttuu, kun kunkin indikaattorimuuttujan arvo muuttuu vertailuluokasta indikaattorimuuttujan nimen ilmaisemaan arvoon (kun muut selittävät muuttujat pidetään vakioina). Pääasiallisen toiminnan vertailuluokkana on luokka ”muu”. Tämä tarkoittaa esimerkiksi sitä, että henkilön nettovaikutusten odotusarvo 32–35-vuotiaana muuttuu mallin 4 odotusarvoestimaatin mukaan 5 220 euroa positiivisemmaksi, kun henkilön pääasiallinen toiminta 22-vuotiaana on ”työllinen” vertailuluokan ”muu” sijaan. Samoin odotusarvo muuttuu mallin 4 odotusarvoestimaatin mukaan 4 490 positiivisemmaksi, kun henkilöllä on 22-vuotiaana vähintään 2. asteen koulutus – verrattuna siihen, että tätä koulutusta ei olisi ja muut selittävät muuttujat pidettäisiin samoissa arvoissaan.

Hajonta-, vapausaste- ja vinousparametrin tulkintaa olen käsitellyt tarkemmin työikäisten ja nuorten nettovaikutusmallien tulosten yhteydessä, joten en toista tulkintaa tässä. Samaten käsittelin tarkemmin ennustetarkkuuden mittareita WAIC:ia ja LOO:ta työikäisten ja nuorten nettovaikutusmallinnuksen yhteydessä. Joka tapauksessa taulukon8.13 WAIC- ja LOO-estimaattien mukaan parametrimäärältään suuremmilla malleilla on parempi ennustarkkuus kuin pienemmillä malleilla, sillä pienemmät estimaatit viittaavat parempaan ennustetarkkuuteen.

Lopullisen mallin parametrien posteriorijakaumat

”Lopullinen” malli on se malli edellä esitetyistä perättäin rakennetuista malleista, jota käytän henkilöiden nettovaikutusten ennustamiseen 32–35-vuotiaana niiden tietojan pohjalta, joita henkilöistä on 22-vuotiaana. Valitsin lopulliseksi malliksi mallin 4.

Taulukossa 8.14 on esitetty lopullisen mallin (malli 4) parametrien posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty näiden posteriorijakaumien odotusarvot (OA), keskihajonnat (KH), kvartiilit sekä jakaumien hännistä prosenttipisteet 2,5 (%) ja 97,5 (%). Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli yhtä, en käsittele niitä tarkemmin.

Taulukko 8.14: *Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana ja 32–35-vuotiaita vuosien 2005–2011 aikana. Nettovaikutusten lopullinen malli (malli 4). Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakio	6 670	280	6 140	6 470	6 670	6 850	7 210	19	1,01
Nainen	-10 380	320	-11 010	-10 590	-10 370	-10 170	-9 760	9	1,00
Pääasiallinen toiminta: työllinen	5 220	700	3 900	4 720	5 240	5 670	6 600	26	1,00
Pääasiallinen toiminta: opiskelija	5 910	710	4 500	5 440	5 930	6 380	7 300	27	1,00
Pääasiallinen toiminta: työtön	1 530	800	-80	1 010	1 520	2 040	3 200	27	1,00
On vähintään 2. asteen koulutus	4 490	480	3 580	4 170	4 480	4 820	5 440	14	1,00
Saaneet toimeentulotukea vuoden aikana	-2 440	430	-3 240	-2 740	-2 440	-2 150	-1 570	12	1,00
Menoja rikosasioista vuoden aikana	-6 900	1 210	-9 390	-7 680	-6 870	-6 110	-4 550	35	1,00
Kotikunta Uudellamaalla	1 810	370	1 070	1 560	1 820	2 050	2 550	11	1,00
σ	8 530	150	8 240	8 430	8 530	8 640	8 850	11	1,00
ν_y	3,47	0,19	3,12	3,33	3,46	3,59	3,86	0,018	1,01
γ	-840	170	-1 180	-950	-840	-720	-520	12	1,01

Esimerkkinä taulukon 8.14 estimaattien tulkinnasta voidaan käsitellä selittävää muuttujaa, ”Pääasiallinen toiminta: työtön”, koskevan regressiokertoimen posteriorijakaman estimointituloksia. Tämän indikaattorimuuttujan vertailuluokka on ”Pääasiallinen toiminta: muu”. Taulukon estimointitulosten mukaan on noin 97%:n todennäköisyys sille, että 22-vuotiaana työttömänä olleen nettovaikutusten odotusarvo on 32–35-vuotiaana suurempi kuin sellaisen henkilön, jolla pääasiallinen toiminta oli ”muu” 22-vuotiaana, kun muiden selittävien muuttujien arvot pidetään vakioina. Tämä johtuu siitä, että 97,5%:ia regressiokertoimen, ”Pääasiallinen toiminta: työtön”, posteriorijakauman todennäköisyysmassasta on -80 euron oikealla (eli positiivisella) puolella.

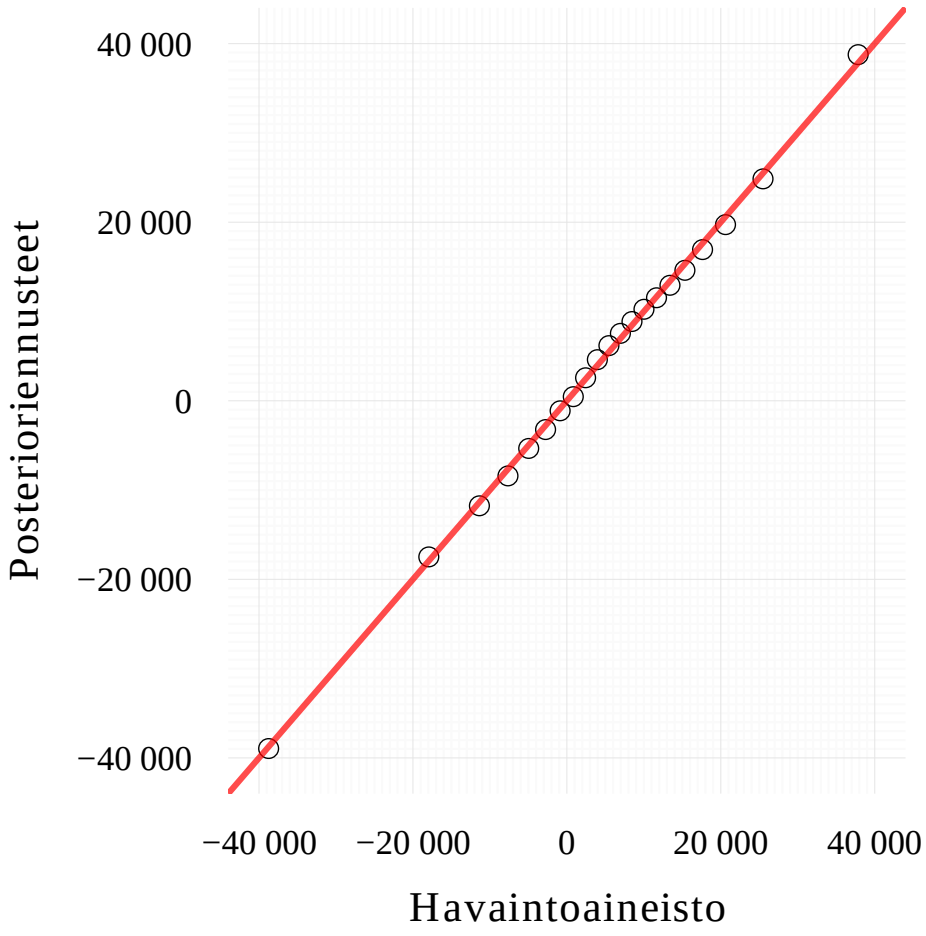
Huomautan vielä yleisesti ottaen estimointituloksia koskien, että tämän tutkimuksen I osan tulosten mukaan (kuviot 98 ja 99) Suomessa syntyneiden naisten ja miesten välinen ero havaituissa julkisen talouden nettovaikutuksissa on työiässä suurimmillaan 30–39-vuotiaiden ikäluokassa. Tämä näkyy taulukon 8.14 estimointituloksissa siinä, että naisten nettovaikutusten odotusarvo 32–35-vuotiaana on noin 10 000 euroa negatiivisempi kuin miesten. Se, että ero on suurimmillaan 30–39-vuotiaiden ikäluokassa, johtuu nähdäkseni lastensaannin vaikutuksesta. Tämä käy epäsuorasti ilmi esimerkiksi tutkimuksen I osan kuvioista 11 ja 12, jotka osoittavat, että naisten saamien tulonsiirtojen keskiarvo kasvaa 30–39-vuotiaana keskimäärin noin 4 000 euroa suuremmaksi kuin miesten, kun nuoremmassa ja vanhemmassa ikäluokassa ero on noin 2 000 euroa pienempi. En huomioi sukupuolen vaikutusta 32–35-vuotiaiden nettovaikutusennusteissa, jotka teen henkilöiden 22:n ikävuoden tietojen pohjalta.

8.4.2.3 Posterioriennustetarkistus

Posterioriennustetarkistuksessa vertaan lopullisen mallin tuottamien simuloitujen posterioriennusteiden jakaumaa alkuperäisen havaitun selitettävän muuttujan jakaumaan. Tavoitteena on tarkistaa, että lopullisen mallin tuottamien ennusteiden jakauma vastaa ”hyvin” havaintoaineiston selitettävän muuttujan jakaumaa. Olen käsitellyt tarkemmin posterioriennustejakaumien muodostamista olotilamallinnuksen yhteydessä luvussa 3.2.6, joten en toista selostusta tässä.

Vertaan alla kuviossa 8.2 havaintoaineiston selitettävän muuttujana jakauman prosenttipisteitä vastaaviin posterioriennustejakauman prosenttipisteisiin. Jakaumien prosenttipisteet, joille teen vertailut, ovat $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. Vasemmanpuoleisin prosenttipiste on 1. prosenttipiste ja oikeanpuoleisin on 99. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeama vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).



Kuva 8.2: *Posterioriennustetarkistus: havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Jakaumien prosenttipisteet:* $1, 5, 10, 15, \ldots, 80, 85, 95, 99$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä*

Koska prosenttipisteitä vastaavat ympyrät sijoittuvat kuvassa 8.2 hyvin diagonaalisille punaiselle viivalle, totean tämän perusteella, että edellä esitettyä vinoon t-jakaumaan perustuvaa mallia voidaan käyttää 32–35-vuotiaiden nettovaikutusten ennustamiseen 22-vuotiaita koskevien tietojen pohjalta.

8.5 Mallinnus: palkkasummat 35–39-vuotiaana 22-vuotiaita koskevien tietojen perusteella

8.5.1 Estimoitavien parametrien posteriorijakauman määritys ja simulointi

Mallinnusidea ja palkkasummat 35–39-vuotiaana selitettävänä muuttujana. Esitän tässä luvussa, kuinka mallinnan Suomessa syntyneiden henkilöiden palkkasummia 35–39-vuotiaana tiedoilla, joita henkilöistä on, kun he olivat 22-vuotiaita. Selitettävänä muuttujana mallinnuksessa on henkilöiden palkkasumma 35–39-vuotiaana. Selittävänä muuttujana käytetään vain yhtä havaintoa per aineiston henkilö. Henkilön yksittäinen palkkasummahavainto on siltä vuodelta, kun hän oli vanhin aineistossa. Eli jos henkilöltä löytyy aineistosta palkkasummahavainnot esimerkiksi 35-vuotiaasta 38-vuotiaaksi, niin tältä henkilöltä poimitaan mukaan vain palkkasummahavainto siltä vuodelta, kun hän oli 38-vuotias. En voinut käyttää näiden ikävuosien keskiarvoja henkilöittäin, sillä tämä olisi vääristänyt nollapalkkasummien todennäköisyyttä mallinnettaessa vuosittaisia palkkasummia. Useamman havainnon käyttäminen samoilta henkilöiltä olisi puolestaan luonut mallinnustilanteen, jossa tämä havaintoaineiston ryvästyneisyys henkilöihin olisi pitänyt ottaa huomioon, mikä puolestaan olisi lisännyt huomattavasti tähän mallinnukseen käytettyä aikaa.

Selittävät muuttujat: tiedot 22-vuotiaina. Selittävinä muuttujina mallinnuksessa ovat puolestaan edellisessä luvussa yllä esitetyt tiedot henkilöistä, kun he olivat 22-vuotiaita. Koska selitettävänä muuttujana ovat vuosittaiset palkkasummat, niiden mallinnus muistuttaa hyvin pitkälti luvussa 5 esitettyä mallinnusta 18–62-vuotiaiden vuosittaisille palkkasummille.

Palkkasummien mallinnus. Tässä luvussa esitettävä 35–39-vuotiaiden palkkasummien mallinnus on kuitenkin yksinkertaisempi tehdä kuin 18–62-vuotiaiden palkkasummien mallinnuksen yhteydessä, sillä tässä tapauksessa jokaista henkilöä kohti on vain yksi selitettävän muuttujan havainto. Lisäksi mallinnuksessa ei tarvitse käytännössä huomioida maassaoloaikaan eikä kalenterivuoteen liittyviä vaikutuksia, koska havainnot ovat vain muutamalta kalenterivuodelta. Tämä tarkoittaa yksinkertaisempaa mallinnusta siinä mielessä, että 35–39-vuotiaiden palkkasummahavainnot eivät ole hierarkkisesti rakentuneet, kuten työikäisten palkkasummamallinnuksen kohdalla oli. Lisäksi, koska jokaista henkilöä koskien on vain yksi havainto, perättäisten havaintojen välistä riippuvuutta ei ole – niin kuin on palkkasummahavaintoja koskevassa havaintoaineistossa. Koska mallinnus on näin ollen samankaltainen kuin palkkasummamallinnuksen yhteydessä (luvussa 5.3.1), en esitä mallinnusaskeleita enää uudelleen.

Koska elinkaarivaikutusten ennustaminen tapahtuu bayesilaisen tilastollisen päättelyn piirissä, noudatan myös 35–39-vuotiaiden palkkasummamallinnuksen yhteydessä bayesilaista mallinnusta. Tämä tarkoittaa sitä, että tavoitteena on selvittää estimoitavien mallinnusparametrien posteriorijakauma (ks. posteriorijakauman muodostamisesta esimerkiksi luku 5.3.3) . Tämä posteriorijakauma on verrannollinen aineistomallin määrittämän otosjakauman ja estimoitavien parametrien priorijakaumien tulolle. Kuten työikäisten palkkasummamallinnuksen yhteydessä, nyt on määritettävä sekä 35–39-vuotiaiden palkkasummien aineistomalli sekä estimoitavien parametrien priorijakaumat.

Otosjakauman määrittäminen: Hurdle-gamma-sekoitusjakauma. Koska 35–39-vuotiaita koskevassa palkkasummahavaintojen aineistossa ei ole ainakaan sellaista hierarkkisuutta, joka minun mielestäni pitäisi huomioida 35–39-vuotiaiden nettovaikutusten aineistomallin määrittämisessä, oikeastaan ainoaksi tehtäväksi jää määrittää sopiva otosjakauma. Lisäksi edellä on jo määritetty käytettävät selitettävät muuttujat. Koska 35–39-vuotiaiden palkkasummamallinnuksen yhteydessä havainnot koskevat myös 35–39-vuotiaita (ja muita työikäisiä), oli mielestäni sopivaa olettaa, että 35–39-vuotiaiden palkkasummien jakauma noudattaisi likimain samaa todennäköisyysjakaumaa kuin 18–62-vuotiaiden mallinnuksessa käytetty otosjakauma. Tämä otosjakauma – josta havainnot on oletetusti generoitu – oli palkkasummien lopullisessa mallissa hurdle-gamma-sekoitusjakauma. Hurdle on nollapalkkasummien osuus jakaumasta, ja kahden gamma-jakauman sekoitusjakauma muodostaa yli nollan olevien palkkasummien jakauman (ks. tarkemmin luku 5.3.1).

Parametrien priorijakaumista ja muut mallinnukseen liittyvät seikat. Estimoitavien parametrien priorijakaumat ovat tässä yhteydessä myös samanlaisia – eli heikosti informatiivisia – kuin palkkasummamallien yhteydessä (ks. luku 5.3.2). Tästä syystä en esitä niitä erikseen tässä yhteydessä. Samaten posteriorijakauman simulointi, simulaatioiden määriä koskevat valinnat, simulaatioketjujen konvergoitumista koskeva päättely sekä estimointitulosten perusteella tehty posterioriennustejakauman simulointi vastaavat palkkasummamallinnuksena yhteydessä esitettyä (ks. luku 5.3.4).

8.5.2 Havainnot, estimointitulokset ja posterioriennustetarkistus

Esitän tässä alaluvussa tulokset koskien Suomessa syntyneiden 35–39-vuotiaiden henkilöiden palkkasummien mallinnusta niillä tiedoilla, joita henkilöistä on koskien sitä vuotta, jonka lopussa he olivat 22-vuotiaita. Esitän ensiksi mallinnuksessa käytetyn havaintoaineiston eli selitettävän muuttujan jakauman tunnusluvut sekä selittävien muuttujien keskeisimmät jakaumatiedot. Tämän jälkeen esitän mallinnuksen estimointitulokset. Posterioriennustetarkistuksessa varmistan, että käytetyn estimoidun mallin tuottamat ennusteet vastaavat jakaumaltaan alkuperäistä havaintoaineiston selitettävän muuttujan jakaumaa. Tulosten esittäminen tässä muodossa vastaa palkkasummien mallinnuksen tulosten esittelyä luvussa 5.4. Koska käsittelen tuloksia – erityisesti estimointitulosten tulkintaa – tarkemmin mainitussa luvussa, käyn estimointitulokset vain lyhyesti läpi tässä luvussa. Alla esitettävät estimointitulokset ovat tulkinnallisesti samanlaisia kuin palkkasummamallien yhteydessä, sillä kummassakin tapauksessa mallinnus perustuu samoihin otosjakaumiin.

8.5.2.1 Havaintoaineisto

Selitettävän muuttujan havaintoaineisto. Alla taulukossa 8.15 on esitetty selitettävän muuttujan havaintolukumäärä ja havaintojen jakauman keskeisimpiä tunnuslukuja. Selitettävä muuttuja muodostuu niiden Suomessa syntyneiden henkilöiden palkkasummista 35–39-vuotiaana, jotka olivat 22-vuotiaita

vuosien 1995–2001 aikana. Täten palkkasummahavainnot 35–39-vuotiaana koskevat vuosia 2005–2014. Jokaiselta henkilöltä on vain yksi havainto. Havainto on siltä vuodelta, jolloin henkilö oli vanhin.

Aineistossa on 916 henkilöä, joiden yksittäisistä palkkasummahavainnoista selitettävän muuttujan havaitut arvot koostuvat. Palkkasummahavaintojen keskiarvo on 28 100 euroa (vuoden 2011 euroissa), ja keskihajonta on 26 100 euroa. Havaintojen mediaani on 28 400 euroa. Mediaani on tässä tapauksessa keskiarvoa suurempi, koska mukana havaintoaineistossa ovat myös nollapalkkasummat.

Taulukko 8.15: *Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana ja 35–39-vuotiaita vuosien 2005–2014 aikana. Havainnot henkilöiden palkkasummista 35–39-vuotiaana. Vain yksi havainto yhdeltä henkilöltä. Havainto siltä vuodelta, kun henkilö oli vanhin. Henkilöiden (ja samalla havaintojen) lukumäärä ja jakauman tunnusluvut mallinnusaineistossa. Euromääräiset tunnusluvut on pyöristetty lähimpään sataan euroon. Nollien osuus on nollapalkkasummien osuus havaintoaineistossa.*

	Havainnot
Henkilölukumäärä (N)	916
Keskiarvo	28 100
Keskihajonta	26 100
5. prosenttipiste (5 %)	0
10. prosenttipiste (10 %)	0
Alakvartiili (25 %)	7 300
Mediaani (50 %)	28 400
Yläkvartiili (75 %)	40 100
90. prosenttipiste (90 %)	53 400
95. prosenttipiste (95 %)	64 600
Nollien osuus	0,19

Selittävien muuttujien havaintoaineisto. Taulukossa 8.16 on esitetty selittävien muuttujien jakaumien keskiarvot, keskihajonnat ja kvartiilit. Selittävät muuttujat koskevat tietoja henkilöistä sen vuoden lopussa, kun he olivat 22-vuotiaita. Kaikki selittävät muuttujat kalenterivuosia lukuun ottamatta ovat tässä tapauksessa indikaattorimuuttujia, joten keskiarvot ovat indikaattorien osuuksia kaikkien havaintojen joukossa. Esimerkiksi muuttujan ”Pääasiallinen toiminta: työllinen” keskiarvo kertoo työllisten osuuden havaintoaineistossa.

Taulukko 8.16: *Suomessa syntyneet henkilöt, jotka olivat 22-vuotiaita vuosien 1995–2001 aikana ja 35–39-vuotiaita vuosien 2005–2014 aikana. Palkkasummamallin selittävien muuttujien standardoimattomien jakaumien tunnusluvut (KA on keski-arvo ja KH on keskihajonta). Tiedot ovat sen vuoden lopulta, kun henkilöt olivat 22-vuotiaita.*

	KA	KH	Alakvartiili	Mediaani	Yläkvartiili
Nainen	0,51	0,50	0	1	1
Pääasiallinen toiminta: työllinen	0,46	0,50	0	0	1
Pääasiallinen toiminta: opiskelija	0,36	0,48	0	0	1
Pääasiallinen toiminta: työtön	0,08	0,26	0	0	0
Pääasiallinen toiminta: muu	0,11	0,31	0	0	0
On vähintään 2. asteen koulutus	0,84	0,37	1	1	1
Saanut toimeentulotukea vuoden aikana	0,21	0,40	0	0	0
Menoja rikosasioista vuoden aikana	0,02	0,14	0	0	0
Kotikunta Uudellamaalla	0,26	0,44	0	0	1
Kalenterivuosi	1998	2	1996	1998	2000

8.5.2.2 Estimointitulokset: palkkasummamallin parametrien posteriorijakaumat

Taulukossa 8.17 on esitetty palkkasummamallin parametrien posterioriestimaatit. Parametrien posterioriestimaatit ovat peräisin niiden estimoiduista posteriorijakaumista. Taulukossa on esitetty parametrien posteriorijakaumien odotusarvojen (OA), keskihajontojen (KH) ja kvartiilien estimaatit sekä jakaumien hännistä prosenttipisteiden 2,5 (%) ja 97,5 (%) estimaatit. Lisäksi taulukossa on esitetty parametrien posteriorijakaumien odotusarvoestimaatteja koskevat MC-keskivirheet sekä simulaatioketjujen konvergenssia mittaava testisuure $\hat{R}$. Koska MC-keskivirheet ovat pieniä, ja R-hatut lähellä tavoitearvoaan eli ykköstä, en käsittele niitä tarkemmin.

Taulukko 8.17: *Palkkasummamalli: 35–39-vuotiaiden palkkasummat sen perusteella, mitä heistä tiedettiin 22-vuotiaina. Parametrien posteriorijakaumien odotusarvon (OA), keskihajonnan (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaatin MC-keskivirhe (MC-kv) ja konvergenssin testisuure ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Vakioparametrit:									
Vakio (H)	-1,62	0,10	-1,81	-1,69	-1,62	-1,56	-1,43	0,00	1,00
Vakio ($\mathcal{G}_1$)	1,95	0,11	1,72	1,88	1,95	2,02	2,15	0,00	1,00
Vakio ($\mathcal{G}_2$)	2,23	0,02	2,20	2,22	2,23	2,25	2,27	0,00	1,00
Vakio (θ)	-0,83	0,18	-1,17	-0,95	-0,83	-0,70	-0,50	0,01	1,00
Regressiokertoimet (H):									
Nainen (H)	0,09	0,09	-0,09	0,03	0,09	0,15	0,26	0,00	1,00
Pääasiallinen toiminta: työllinen (H)	-0,67	0,15	-0,95	-0,77	-0,67	-0,58	-0,38	0,00	1,00
Pääasiallinen toiminta: opiskelija (H)	-0,53	0,14	-0,82	-0,62	-0,53	-0,44	-0,24	0,00	1,00
Pääasiallinen toiminta: työton (H)	-0,09	0,10	-0,29	-0,16	-0,09	-0,02	0,13	0,00	1,00
Korkein koulutus: 2. asteen koulutus (H)	-0,37	0,09	-0,54	-0,42	-0,37	-0,31	-0,21	0,00	1,00
Saanut toimeentulotukea vuoden aikana (H)	-0,04	0,09	-0,21	-0,10	-0,04	0,02	0,15	0,00	1,00
Menoja rikosasioista vuoden aikana (H)	0,14	0,08	-0,01	0,09	0,14	0,19	0,30	0,00	1,00
Kotikunta Uudellamaalla (H)	-0,07	0,10	-0,26	-0,13	-0,07	-0,01	0,12	0,00	1,00
Kalenterivuosi (H)	0,10	0,10	-0,08	0,04	0,10	0,17	0,30	0,00	1,00
Regressiokertoimet ($\mathcal{G}_1$):									
Nainen ($\mathcal{G}_1$)	-0,27	0,08	-0,43	-0,33	-0,27	-0,22	-0,12	0,00	1,00
Pääasiallinen toiminta: työllinen ($\mathcal{G}_1$)	-0,08	0,23	-0,51	-0,23	-0,08	0,05	0,44	0,01	1,00
Pääasiallinen toiminta: opiskelija ($\mathcal{G}_1$)	0,12	0,23	-0,30	-0,02	0,10	0,25	0,65	0,01	1,00
Pääasiallinen toiminta: työton ($\mathcal{G}_1$)	-0,10	0,16	-0,41	-0,21	-0,10	-0,01	0,26	0,01	1,00
Korkein koulutus: 2. asteen koulutus ($\mathcal{G}_1$)	0,20	0,12	-0,04	0,13	0,20	0,27	0,44	0,00	1,00
Saanut toimeentulotukea vuoden aikana ($\mathcal{G}_1$)	-0,25	0,11	-0,47	-0,31	-0,24	-0,18	-0,05	0,00	1,00
Menoja rikosasioista vuoden aikana ($\mathcal{G}_1$)	-0,36	0,29	-0,79	-0,61	-0,42	-0,10	0,17	0,02	1,00
Kotikunta Uudellamaalla ($\mathcal{G}_1$)	0,06	0,08	-0,11	-0,00	0,06	0,11	0,22	0,00	1,00
Kalenterivuosi ($\mathcal{G}_1$)	-0,21	0,09	-0,40	-0,27	-0,20	-0,15	-0,04	0,00	1,00

Jatkuu seuraavalla sivulla

Taulukko 8.17 – jatkuu edelliseltä sivulta

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
Regressiokertoimet ($\mathcal{G}_2$):									
Nainen ($\mathcal{G}_2$)	-0,16	0,02	-0,19	-0,17	-0,16	-0,15	-0,13	0,00	1,00
Pääasiallinen toiminta: työllinen ($\mathcal{G}_2$)	0,12	0,05	0,03	0,09	0,12	0,15	0,22	0,00	1,00
Pääasiallinen toiminta: opiskelija ($\mathcal{G}_2$)	0,15	0,05	0,06	0,12	0,15	0,18	0,24	0,00	1,00
Pääasiallinen toiminta: työton ($\mathcal{G}_2$)	0,03	0,04	-0,04	0,01	0,03	0,05	0,10	0,00	1,00
Korkein koulutus: 2. asteen koulutus ($\mathcal{G}_2$)	0,06	0,02	0,02	0,05	0,06	0,07	0,10	0,00	1,00
Saanut toimeentulotukea vuoden aikana ($\mathcal{G}_2$)	-0,01	0,02	-0,04	-0,02	-0,01	0,00	0,02	0,00	1,00
Menoja rikosasioista vuoden aikana ($\mathcal{G}_2$)	-0,02	0,02	-0,05	-0,03	-0,02	-0,01	0,01	0,00	1,00
Kotikunta Uudellamaalla ($\mathcal{G}_2$)	0,05	0,02	0,02	0,04	0,05	0,06	0,08	0,00	1,00
Kalenterivuosi ($\mathcal{G}_2$)	-0,01	0,02	-0,04	-0,03	-0,01	-0,00	0,02	0,00	1,00
Regressiokertoimet (θ):									
Nainen (θ)	0,15	0,12	-0,09	0,07	0,15	0,23	0,38	0,00	1,00
Pääasiallinen toiminta: työllinen (θ)	-0,12	0,33	-0,75	-0,35	-0,13	0,10	0,55	0,01	1,00
Pääasiallinen toiminta: opiskelija (θ)	-0,07	0,32	-0,68	-0,28	-0,09	0,15	0,59	0,01	1,00
Pääasiallinen toiminta: työton (θ)	0,17	0,24	-0,29	0,00	0,16	0,33	0,64	0,01	1,00
Korkein koulutus: 2. asteen koulutus (θ)	0,04	0,14	-0,23	-0,06	0,04	0,13	0,32	0,00	1,00
Saanut toimeentulotukea vuoden aikana (θ)	0,02	0,14	-0,25	-0,07	0,03	0,12	0,29	0,00	1,00
Kotikunta Uudellamaalla (θ)	0,00	0,12	-0,24	-0,08	0,00	0,09	0,25	0,00	1,00
Kalenterivuosi (θ)	0,06	0,13	-0,19	-0,03	0,06	0,15	0,33	0,00	1,00
Muotoparametrit:									
Muotoparametri($\mathcal{G}_1$)	0,91	0,09	0,73	0,85	0,91	0,97	1,09	0,00	1,00
Muotoparametri($\mathcal{G}_2$)	11,96	1,43	9,42	10,98	11,91	12,83	15,01	0,04	1,00

Parametrien tulkinnasta. Taulukossa 8.17 parametrit on jaettu neljään luokkaan. Luokka on ilmoitettu suluissa kunkin parametrin nimen jälkeen:

- jos parametrin nimen jälkeen (H), parametri liittyy ”hurdleen”, joka kertoo, saako palkkasumma-havainto arvoksi nollan (eikä yli nollaa).
- jos parametrin nimen jälkeen ($\mathcal{G}_1$), parametri liittyy gamma1-jakaumaan eli vain osan vuotta töissä olevien palkkasummien jakaumaan.
- jos parametrin nimen jälkeen ($\mathcal{G}_2$), parametri liittyy gamma2-jakaumaan eli koko vuoden kokoaikaisissa töissä olevien palkkasummien jakaumaan.
- jos parametrin nimen jälkeen (θ), parametri liittyy θ:aan eli indikaattorimuuttujaan, joka kertoo, kuuluuko havainto gamma1-jakaumaan (eikä gamma2-jakaumaan).

Katso tarkemmin parametrien selitteistä palkkasummien mallinnuksessa esitetty taulukko 5.1 (s. 262) sekä esimerkiksi palkkasumma-aineistojen jakaumia käsittelevä kuva 5.1 (s. 246) sekä palkkasummien mallinnusta käsittelevä luku 5.3.1.

Hurdlea (H) ja θ:aa koskevat regressiokertoimet ovat taulukossa 8.17 log-oddseja. Gamma-jakaumia ($\mathcal{G}_1$, $\mathcal{G}_2$) koskevat regressiokertoimet ovat puolestaan vaikutusten logaritmeja. Mikäli hurdlea koskevat regressiokertoimet ovat nollaa suurempia, ne lisäävät nollapalkkasumman todennäköisyyttä (verrattuna vertailuluokkaan). Nollaa pienemmät regressiokertoimet puolestaan vähentävät nollapalkkasumman todennäköisyyttä (verrattuna vertailuluokkaan). Pääasiallisen toiminnan vertailuluokka on pääasiallisen toiminnan luokka: muu.

Mikäli θ:aa koskevat regressiokertoimet ovat nollaa suurempia, ne lisäävät palkkasumman todennäköisyyttä kuulua gamma1-jakaumaan (gamma2-jakauman sijasta), eli todennäköisyyttä siihen, että on vain osan vuodesta töissä tai osa-aikaisissa töissä (kuin koko vuoden kokoaikaisissa töissä). Mikäli regressiokertoimet ovat puolestaan nollaa pienempiä, palkkasummahavainnon todennäköisyys kuulua gamma1-jakaumaan pienenee ja siten todennäköisyys kuulua gamma2-jakaumaan kasvaa.

Vaikutusten logaritmeja koskevat gamma-jakaumien regressiokertoimet tulkitaan siten, että jos regressiokerroin on suurempi kuin kuin nolla, ne nostavat palkkasumman euromääräistä odotusarvoa. Jos ne ovat puolestaan nollaa pienempiä, ne laskevat odotusarvoja. Siten esimerkiksi niillä, jotka saivat toimeentulotukea 22-vuotiaina, on odotusarvoisesti pienemmät palkkasummat osan vuotta kestävistä tai osa-aikaisista töistä (gamma1-jakauma, $\mathcal{G}_1$), kuin niillä, jotka eivät saaneet toimeentulotukea 22-vuotiaina.

8.5.2.3 Posterioriennustetarkistus

Posterioriennustetarkistuksessa vertaan palkkasummamallin tuottamien simuloitujen posterioriennusteiden jakaumaa alkuperäisen havaitun selitettävän muuttujan jakaumaan. Tavoitteena on tarkistaa, että

lopullisen mallin tuottamien palkkasummien ennusteiden jakauma vastaa ”hyvin” havaintoaineiston selitettävän muuttujan eli palkkasummien jakaumaa. Olen käsitellyt tarkemmin posterioriennustejakaumien muodostamista olotilamallinnuksen yhteydessä luvussa 3.2.6, joten en toista selostusta tässä.

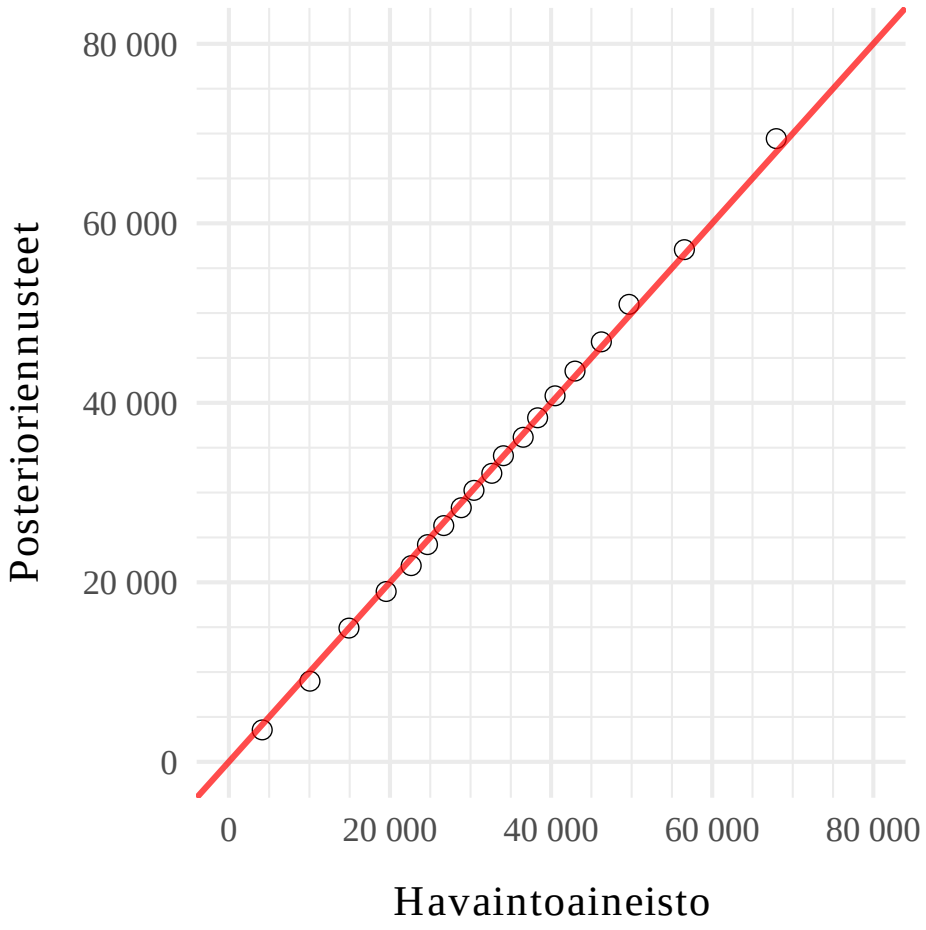
Alla taulukossa 8.18 on verrattu palkkasummamallista simuloitujen posterioriennusteiden tuottamaa palkkasummien jakaumaa alkuperäiseen havaintoaineiston palkkasummien jakaumaan. Taulukossa verrataan posterioriennusteiden jakauman keskiarvoa, keskihajontaa, mediaania ja nollien osuutta vastaaviin jakauman tunnuslukuihin alkuperäisessä havaintoaineistossa. Kuten taulukosta 8.18 käy ilmi, erityisesti mallin tuottamien palkkasummien keskiarvot ja nollien osuudet vastaavat "hyvin"alkuperäisen havaintoaineiston palkkasummien keskiarvoa ja nollien osuutta.

Taulukko 8.18: *Posterioriennustetarkistus: havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Posterioriennusteiden ja havaintoaineiston palkkasummien keskiarvot, keskihajonnat, mediaanit sekä nollapalkkasummien osuudet. Euromääräiset luvut on pyöristetty lähimpään sataan euroon.*

	Keskiarvo	Keskihajonta	Mediaani	Nollaosuus
Posterioriennusteet	28 500	27 400	27 700	0,19
Aineiston havainnot	28 100	26 100	28 400	0,19

Yli nollan olevien palkkasummien jakauma. Vertaan vielä tarkemmin mallin tuottamien yli nollan olevien palkkasummien jakaumaa alkuperäisen havaintoaineiston yli nollan olevien palkkasummien jakaumaan. Teen tämän vertailun alla kuvassa 8.3, jossa vertaan havaintoaineiston yli nollan olevien palkkasummien jakauman prosenttipisteitä vastaaviin posterioriennustejakauman prosenttipisteisiin. Jakaumien prosenttipisteet, joille teen vertailut, ovat $5, 10, 15, \ldots, 80, 85, 95$. Vasemmanpuoleisin prosenttipiste on 5. prosenttipiste ja oikeanpuoleisin on 95. prosenttipiste.

Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen viiva. Posterioriennusteiden prosenttipisteiden poikkeama vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*).



Kuva 8.3: *Posterioriennustetarkistus: havaintoaineiston ja posterioriennusteiden euromääräisten jakaumien vertailu. Yli nollan olevien palkkasummien jakaumien prosenttipisteet:* $5, 10, 15, \ldots, 80, 85, 95$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä*

Koska prosenttipisteitä vastaavat ympyrät sijoittuvat kuvassa 8.3 hyvin diagonaalisille punaiselle viivalle, totean tämän perusteella, että esitettyä hurdle-gamma-sekoitusjakaumaan perustuvaa mallia voidaan käyttää 35–39-vuotiaiden palkkasummien ennustamiseen 22-vuotiaita koskevien tietojen pohjalta.

8.6 Lasten määrien ennustaminen

Maahanmuuttajien lasten yhteenlasketut määrät koostuvat kahden komponentin summasta. Ensimmäinen summattava lasten määrä muodostuu niistä lapsista, jotka muuttavat Suomeen ulkomailla syntyneiden

vanhempiensa mukana. Tässä tapauksessa myös lapset ovat ulkomailla syntyneitä. Toinen summattavat lasten määrä muodostuu niistä lapsista, jotka syntyvät Suomessa vanhemmille, jotka ovat itse syntyneet ulkomailla.

Ennustan lasten määrät uusia maahanmuuttajia koskien vain naisille (äideille). Tällä ratkaisulla varmistan, että yhtä lasta ei lasketa kahteen kertaan maahanmuuttajien lasten määrään, sillä muuten sama lapsi voitaisiin liittää sekä isää että äitiä koskeviin lapsimääriin. Lisäksi yleisesti käytössä olevat hedelmällisyysluvut lasketaan naisia koskien.

8.6.1 Äitien mukana Suomeen muuttaneiden lasten lukumäärät

Uusien maahanmuuttajien mukana Suomeen muuttavien lasten lukumäärää ei tarvitse ennustaa. Tämä johtuu siitä, että lapset muuttavat yhtaikaa Suomeen äitiensä kanssa. Koska elinkaariennusteissa on käytettävissä tiedot, joita uusista maahanmuuttajista on muuttohetkellä, äitien mukana muuttavat lapset on tällainen tieto, joka tiedetään. Tieto on varma, eli siihen ei liity mittausvirhettä, kun käytettävissä olevat rekisteritiedot ovat luotettavia.

Käytetyllä määritelmällä äitien mukana muuttavien lasten lukumäärästä jäävät puuttumaan pelkästään lasten isien mukana muuttavat lapset sekä ne lapset, jotka muuttavat Suomeen jonain muuna vuonna, kuin heidän äitinsä muuttaa Suomeen ensimmäisen kerran. Tästä huolimatta suurin osa maahanmuuttajien lapsista muuttaa Suomeen samana vuonna, kuin heidän äitinsä muuttaa Suomeen – ainakin Irakissa ja Somaliassa syntyneiden maahanmuuttajien osalta.

Tämä selviää Tilastokeskuksen tutkimuskäyttööni laatimasta rekisteritiedosta, joka kertoo jokaisen maahanmuuttajan kohdalla kalenterivuosittain, kuinka monta heidän ulkomailla syntynyttä lastaan on muuttanut Suomeen. Samasta rekisteritiedosta saadaan haluttu tieto sille, kuinka moni maahanmuuttajanaisen lapsi muutti Suomeen samana vuonna kuin hänen äitinsä. Ainakin Irakissa ja Somaliassa syntyneiden osalta mukana muuttaneiden lasten määrä on paljon pienempi kuin seuraavaksi esitettävä ennuste Suomessa saatujen lasten määrälle. Tämä saattaa johtua siitä, että Irakissa ja Somaliassa syntyneet muuttavat yleensä Suomeen ensimmäisen kerran niin nuoressa iässä, että heille ei ole vielä ”ehtinyt” syntyä kuin pieni osa koko hedelmällisyysiän yli lasketusta ennustetusta kokonaislapsimäärästä.

8.6.2 Naisten Suomessa saamien lasten lukumäärien ennustaminen

Ennusteet maahanmuuttajanaisten Suomessa saamien lasten määristä perustuvat ennusteisiin ikäryhmittäisistä hedelmällisyysluvuista maahanmuuttajaryhmittäin. Hedelmällisyyslukujen ennusteet puolestaan perustuvat havaittuihin ikäryhmittäisiin hedelmällisyyslukuihin vuosina 2000–2015.

Ikäryhmittäiset hedelmällisyysluvut. Maahanmuuttajaryhmäkohtaiset (havaitut) vuosittaiset ikäryhmittäiset hedelmällisyysluvut laskin Tilastokeskuksen julkisesti saatavilla olevista tiedoista. Nämä tiedot koskivat naisten määriä taustamaan mukaan sekä ikäryhmittäisiä tietoja naisten synnyttämistä lapsista

taustamaan mukaan (henkilölukumäärät: SVT: *Väestörakenne: Taustamaa ja syntyperä iän ja sukupuolen mukaan maakunnittain 1990 - 2015*; syntyneet: SVT: *Syntyneet: Elävänä syntyneet sukupuolen, äidin iän, taustamaan ja kielen mukaan 2000 - 2015*). Maahanmuuttajaryhmäkohtaiset eli tässä tapauksessa taustamaakohtaiset ikäryhmittäiset hedelmällisyysluvut laskin näistä tiedoista osamäärinä ikäryhmittäisten hedelmällisyyslukujen yleisen määritelmän mukaisesti: ”*Ikäryhmittäisellä hedelmällisyysluvulla tarkoitetaan tietyn ikäryhmän naisten synnyttämien elävien lasten määrää saman ikäryhmän keskiväkiluvun tuhatta naista kohden.*” (Tilastokeskus 2017a: *Käsitteet: Ikäryhmittäinen hedelmällisyysluku*).

Tilastokeskus määrittelee taustamaan seuraavasti: ”*Mikäli henkilön molemmat vanhemmat ovat syntyneet ulkomailla, on taustamaa ensisijaisesti biologisen äidin syntymävaltio. Jos henkilöllä on vain tieto ulkomailla syntyneestä isästä, on taustamaa isän syntymävaltio. Jos kummankaan vanhemman syntymävaltiosta ei ole tietoa, on taustamaa ulkomailla syntyneiden henkilöiden osalta henkilön oma syntymävaltio. Suomessa syntyneiden henkilöiden, joiden vanhemmista ei ole tietoa, ja joiden on päätelty olevan ulkomaalaistaustaisia, taustamaa on tuntematon.*” (Tilastokeskus 2017a: *Käsitteet: Syntyperä ja taustamaa*).

Maahanmuuttajaryhmäkohtaiset kokonaishedelmällisyysluvut. Laskin maahanmuuttajaryhmittäin ikäryhmittäisten hedelmällisyyslukujen pohjalta jokaiselle ryhmälle vuosittaiset kokonaishedelmällisyysluvut vuosina 2000–2015. Tilastokeskuksen määritelmän mukaan: ”*Kokonaishedelmällisyysluku saadaan laskemalla yhteen yhdelle ikävuodelle lasketut hedelmällisyysluvut. Näin saatu luku tarkoittaa laskennallista lasten määrää, jonka nainen synnyttää kyseisen vuoden hedelmällisyyden pysyessä voimassa naisen koko hedelmällisen kauden edellyttäen, ettei hän kuole ennen tämän kauden päättymistä.*” (Tilastokeskus 2017a: *Käsitteet: Kokonaishedelmällisyys*).

Koska vuosien 2000–2015 hedelmällisyyslukuja käytetään aineistona tulevien hedelmällisyyslukujen ennusteille, halusin varmistaa kokonaishedelmällisyyslukujen aikasarjoja tarkastelemalla, kuinka realistista on, että hedelmällisyyslukujen ennustamiseen käytetään vuosien 2000–2015 tietoja. Erityisesti halusin selvittää aikasarjakuvioita tarkastelemalla, onko ryhmäkohtaisissa kokonaishedelmällisyysluvuissa (laskevia) trendejä. Niiden maiden kohdalla, joiden osalta aikasarjakuviot sisälsivät kuvioista havaittavissa olevia laskevia trendejä, käytin havaintoaineistona mallinnuksessa vain joitakin viimeisiä havaintoaineiston sisältämiä kalenterivuosia niin, että aikasarjan sisältämä laskeva trendi vaikutti kuvioiden perusteella olevan ”taittunut” ja tasaantunut johonkin kokonaishedelmällisyysluvun tasoon niissä.

Vaikka tässä tutkimuksen versiossa julkistetaan elinkaariennusteiden tulokset vain Somaliassa ja Irakissa syntyneitä koskien, laskin kokonaishedelmällisyysluvut ja ennustetut lasten lukumäärät myös muille 10 suurimpaan maahanmuuttajaryhmään kuuluville syntymävaltioille. Ne syntymävaltiot, joiden osalta käytin havaintoaineistona vuosien 2000–2015 koko aikasarjaa, olivat: Turkki, Viro, Saksa, Ruotsi ja Entinen Jugoslavia. Näiden maiden kohdalla kokonaishedelmällisyysluvulla ei vaikuttanut kuvioiden perusteella olevan trendiä kalenterivuosien suhteen. Somalian kohdalla käytin vuosia 2005–2015, sillä viimeisen n. 10 vuoden osalta aineistossa ei vaikuttanut olevan laskevaa kokonaishedelmällisyysluvun trendiä. Muiden maiden kohdalla käytin aineiston viittä viimeistä vuotta eli vuosia 2011–2015. Näitä maita olivat Irak, Entinen Neuvostoliitto, Thaimaa sekä Kiina, jonka osalta kokonaishedelmällisyysluvun trendi on ollut nouseva.

Trendien olemassaolo tai niiden puute ryhmäkohtaisissa havaintoaineistoissa saattaa johtua maahanmuuttajaväestöjen koostumusten muutoksista vuosina 2000–2015. Esimerkiksi korkeiden hedelmällisyyslukujen lähtömaista tulleilla maahanmuuttajilla hedelmällisyysluvut saattavat laskea maassaoloajan kasvaessa, kun muut hedelmällisyyslukuihin vaikuttavat tekijät pidetään vakioina. Tällöin ryhmäkohtainen kokonaishedelmällisyysluku riippuu osaltaan siitä, millainen on väestön suomessaoloaikojen jakauma kunakin vuonna. En mallinna ainakaan vielä tutkimuksen tässä versiossa maassaoloajan tai maahanmuuttoiän vaikutusta hedelmällisyyslukuihin. Aihetta käsitellään kuitenkin Suomen Perustan suunnitteilla olevassa tutkimuksessa, jossa laaditaan erilaisia väestöennusteita eri maahanmuuttajaryhmille Suomessa.

Lapsen saannin vuosittaiset ikäryhmittäiset todennäköisyydet. Lapsen saanti tiettynä kalenterivuonna 15–49-vuotiailla naisilla on joko–tai-tyyppinen tapahtuma: joko lapsi saadaan, tai lasta ei saada. Tällaisten joko–tai-tapahtumien (tai 0–1-tapahtumien) todennäköisyyksien mallintamiseen soveltuu yksiparametrinen Bernoulli-jakauma, jonka ainoa parametri kertoo, mikä on kiinnostuksen kohteena olevan tapahtuman toteutumisen tai onnistumisen todennäköisyys. Tässä tapauksessa parametri on ikäryhmittäinen todennäköisyys saada lapsi yhden kalenterivuoden aikana.

Koska lasten määriä mallinnetaan samalla koko väestölle, ollaan kiinnostuneita 0–1-tapahtumien toteutumisten yhteenlasketusta määrästä. Havaintoaineistossa tämä tarkoittaa syntyneiden lasten määrää tietyn ikäryhmän naisille. Tällaisten lukumäärien todennäköisyyksien mallintamiseen soveltuu binomijakauma, jolla on kaksi parametria: (riski)väestön henkilölukumäärä ja tapahtuman toteutumisen tai onnistumisen todennäköisyys, samalla lailla kuin Bernoulli-jakaumassa. Havaintoaineistossa (riski)väestön henkilölukumäärä on tietyn ikäryhmän naisten yhteenlaskettu lukumäärä vuosina 2000–2015 – tai maakohtaisesti myös lyhyemmällä aikavälillä, kuten edellä esitettiin.

Käytän tässä tutkimuksessa tilastolliseen mallintamiseen ja ennustamiseen bayesilaista päättelyä. Lapsimäärien ennustamisen osalta bayesilaiset ennusteet tarjoavat sen edun frekventistiseen päättelyyn verrattuna, että posteriorienusteet sisältävät estimoitaviin parametreihin sisältyvät epävarmuudet ennusteissa, sillä havaintoaineistot ovat melko pieniä varsinkin nuorimmissa ja vanhimmissa ikäluokissa. Havaintoaineistojen pienuus johtaa käytännössä ”suureen” epävarmuuteen parametrien arvoissa. Bayesilaisessa päättelyssä parametreille määritetään posteriorijakauma, joka on verrannollinen otosjakauman ja priorijakaumien tulolle.

Tässä yhteydessä posteriorijakauma on määritettävä ikäryhmittäisille todennäköisyyksille, $\psi_{15-19}, \ldots, \psi_{45-49}$, saada lapsi kalenterivuoden aikana. Ikäryhmittäisten todennäköisyyksien (ψ) alaindeksit kertovat, mitä ikäluokkaa ne koskevat. Nämä todennäköisyydet määritetään erikseen 5-vuotisikäryhmille: 15–19,20–24,. . . ,40–44,45–49. Naisen kuulumista johonkin näistä ikäluokista kuvaavat alla ikäluokkaindikaattori, $I_{ikäluokka}$, missä kunkin 5-vuotisikäluokan indikaattorin arvo on 1, mikäli nainen kuuluu tiettynä kalenterivuonna ikäluokkaan. Muuten indikaattorin arvo on 0.

Lasten määrän posteriorijakauma on beta-binomi-jakauma. Otosjakauma noudattaa mallissa binomijakaumaa yllä mainituista syistä. Otosjakauman parametreille on määritettävä vielä priorijakaumat. Kun otosjakauma on binomi-jakautunut, ”helppo” ja ”luonnollinen” valinta parametrien priorijakaumalle on $Beta(\alpha, \beta)$-jakauma, missä α ja β ovat kiinteitä nk. *hyperprioriparametreja*. Valinta on ”helppo” siinä

mielessä, että tällä valinnalla ikäryhmittäisillä lasten saannin todennäköisyyksien, ψ, posteriorijakaumilla on tunnetun tuloksen mukaan beta-binomi-jakauma, joka on samalla beta-jakauma. Koska beta-jakauma on hyvin tunnettu todennäköisyysjakauma, posteriorijakaumia ei tarvitse simuloida, ja ratkaisu on tässä mielessä helppo.

Priorijakauman valinta. Priorijakauman valinta on valitun posteriorijakauman tapauksessa siinä mielessä ”luonnollinen", että hyperprioriparametreilla on konkreettinen selitys: ne vastaavat tapahtumien/ykkösten ja ei-tapahtumien/nollien määriä havaintoaineiston ulkopuolella. Käytän hyvin epäinformatiivisia priorijakaumia, joissa sekä α:n että β:n arvot ovat ykkösiä. Tämä tarkoittaa siis sitä, että havaintoaineiston ulkopuolelta tulee tieto, että tietyssä ikäryhmässä on tapahtunut vuoden aikana yksi syntymä ja yksi ei-syntymä. Koska havaintoaineistossa henkilömäärät ovat sadoissa, posteriorijakauma määrittyy melkein täysin havaintoaineiston tarjoaman informaation kautta.

Ikäryhmittäiset lapsen saannin posteriorijakaumat. Ikäryhmittäisten lastensaantien todennäköisyyksien, $\boldsymbol{\psi}$, posteriorijakaumat ovat täten muotoa:

$$
\begin{aligned}
&p(\boldsymbol{\psi}|\boldsymbol{I}_{\textit{ikäluokka}}, \boldsymbol{N^{hav}}, \boldsymbol{H}, \alpha, \beta) \qquad (8.4)\\
&= (I_{\textit{ikäluokka}}(14 < \textit{ikä} < 20)Beta(H + \alpha\beta, H - N^{hav} + \alpha(1-\beta)),\\
&I_{\textit{ikäluokka}}(19 < \textit{ikä} < 25)Beta(H + \alpha\beta, H - N^{hav} + \alpha(1-\beta)), \ldots\\
&\ldots, I_{\textit{ikäluokka}}(44 < \textit{ikä} < 50)Beta(H + \alpha\beta, H - N^{hav} + \alpha(1-\beta)))
\end{aligned}
$$

missä $\boldsymbol{\psi}$ on vektori (lista), joka sisältää ikäryhmittäiset todennäköisyydet (ψ) saada lapsi vuoden aikana; $\boldsymbol{I}_{\textit{ikäluokka}}$ on ikäluokkaindikaattoreiden vektori; *ikä* on naisen ikä väestössä tiettynä kalenterivuotena; $\beta(\cdot,\cdot)$ ovat beta-jakaumia mainituin parametrein; N^{hav} ovat havaitut syntyneiden lasten lukumäärät tietyn ikäryhmän naisille; H on naisten lukumäärä tietyssä ikäryhmässä ja α sekä β ovat beta-binomi-mallin hyperprioriparametrit.

Saatavien lasten lukumäärän ennuste koko hedelmällisyysiän yli. Saatavien lasten lukumäärän ennuste koko hedelmällisyysiän yli, $\boldsymbol{N^{enn}}$, on vektori (lista), jonka komponentteja ovat ikäluokittaiset saatavien lasten lukumäärien ennusteet N^{enn}. Koko hedelmällisyysiän yli määritetty saatavien lasten lukumäärän posterioriennustejakauma on puolestaan yllä olevaan perustuen:

$$
p(\boldsymbol{N^{enn}}|\boldsymbol{N^{hav}}) = \int_{\boldsymbol{\psi}} \sum_{\boldsymbol{I}} p(N^{enn}|\boldsymbol{\psi}, \boldsymbol{I}_{\textit{ikäluokka}}) p(\boldsymbol{\psi}|\boldsymbol{I}_{\textit{ikäluokka}}, N^{hav}, H, \alpha, \beta) d\boldsymbol{\psi}, \qquad (8.5)
$$

Koska posterioriennustejakauma on Beta-jakautunut, posterioriennusteet voidaan poimia suoraan hyvin tunnetusta Beta-jakaumasta yllä mainituin parametrein.

8.7 Yhteenveto: maahanmuuttajien lasten yhteenlaskettujen elinkaarivaikutusten ennusteet

Erot 32–62-vuotiaiden nettovaikutusennusteissa verrattuna kantaväestöön. Olen esittänyt edellä luvuissa 8.2–8.4, kuinka maahanmuuttajien 22-vuotiaita lapsia koskevien tietojen perusteella voidaan ennustaa näiden henkilöiden julkisen talouden nettovaikutuksia 32–35-vuotiaana. Ennustemalli perustuu Suomessa syntynyttä koko väestöä koskevaan otosaineistoon henkilöistä, kun he olivat 22-vuotiaita ja edelleen 32–35-vuotiaita. Keskeisenä oletuksena ennusteessa on, että maahanmuuttajien lasten osalta 22-vuotiaiden tiedot vaikuttavat samalla tavalla henkilöiden nettovaikutuksiin 32–35-vuotiaana kuin koko Suomessa syntyneellä väestöllä.

Toinen oletus on, että ennustetta nettovaikutuksille 32–35-vuotiaana voidaan käyttää ennusteena maahanmuuttajien lasten vuosittaisten nettovaikutusten poikkeamille Suomessa syntyneen koko väestön nettovaikutusten ennusteen odotusarvosta 32–62-vuotiaana. Toisin sanoen, mikäli maahanmuuttajan lapsen nettovaikutusennuste 32–35-vuotiaana on *suurempi* kuin Suomessa syntyneen koko väestön keskiarvoennuste, oletuksen mukaan tämän maahanmuuttajan lapsen nettovaikutusennuste on 32–62-vuotiaana aina odotusarvoisesti *suurempi* kuin Suomessa syntyneen väestön keskiarvoennuste. Mikäli puolestaan maahanmuuttajan lapsen nettovaikutusennuste 32–35-vuotiaana on *pienempi* kuin Suomessa syntyneen koko väestön keskiarvoennuste, oletuksen mukaan tämän maahanmuuttajan lapsen nettovaikutusennuste on 32–62-vuotiaana aina odotusarvoisesti *pienempi* kuin Suomessa syntyneen väestön keskiarvoennuste. Maahanmuuttajien lasten nettovaikutusennusteet eivät sisällä ennustetta eläkkeisiin liittyvien vaikutusten mahdollisista eroista Suomessa syntyneeseen koko väestöön nähden.

Erot 32–62-vuotiaiden palkkasummaennusteissa ja siten eläkkeisiin liittyvissä vaikutuksissa verrattuna kantaväestöön. Luvussa 8.2.6 mallinsin 35–39-vuotiaiden palkkasummia 22-vuotiaita koskevien tietojen perusteella. Tämän mallin ja maahanmuuttajien 22-vuotiaita lapsia koskevien tietojen pohjalta laaditaan ennusteet maahanmuuttajien lasten palkkasummaennusteiden eroille verrattuna kantaväestöön. Koska palkkasummamallit koostuvat neljästä komponentista, näille kaikille komponenteille estimoitiin eroennusteet verrattuna kantaväestöön. Palkkasummaennusteiden eroista 32–62-vuotiaana saadaan puolestaan laskettua eläkkeisiin liittyvien vaikutusten erot verrattuna kantaväestöön.

Maahanmuuttajien lasten ennustetut yhteenlasketut elinkaarivaikutusten ennusteet. Koska maahanmuuttajien lasten nettovaikutusennusteet, palkkasummaennusteet ja lasten määrien ennusteet ovat ainakin tässä tutkimuksen versiossa riippumattomia, maahanmuuttajan lasten yhteenlaskettujen elinkaariennusteiden todennäköisyysjakauma saadaan yksinkertaisesti maahanmuuttajien lasten elinkaarivaikutusten todennäköisyysjakauman ja maahanmuuttajien lasten lukumäärien ennusteiden (luku 8.6) todennäköisyysjakauman välisenä tulona. Toisin sanoen maahanmuuttajat lasten yhteenlasketut ennustetut elinkaarivaikutukset saadaan laskettua seuraavasti: ennustetaan aluksi maahanmuuttajalle lasten määrien ennuste. Tämän jälkeen ennustetaan jokaiselle ennustetulle lapselle oma elinkaariennusteensa, kuten luvussa 9 esitetään. Näissä elinkaariennusteissa käytetään yllä esitettyjä nettovaikutus- ja palkkasummaeroennusteita verrattuna kantaväestöön. Myös yksittäisten lasten elinkaariennusteet ovat ainakin tässä tutkimuksen versiossa toisistaan riippumattomia.

Luku 9

Elinkaariennusteet

9.1 Yleistä elinkaariennusteista

Elinkaariennusteet tarkoittavat julkisen talouden nettovaikutusten ennusteita henkilöiden kokonaisten elinkaarien yli. Uusien maahanmuuttajien osalta elinkaaret alkavat siitä vuodesta, kun he muuttavat ensimmäisen kerran Suomeen. Suomessa syntyneillä elinkaaret alkavat siitä vuodesta, kun he ovat syntyneet. Laadin elinkaariennusteet jokaiselle ennustepopulaatioon kuuluvalle henkilölle näin määritellystä elinkaaren aloitusvuodesta siihen vuoteen asti, kunnes henkilö täyttäisi 101 vuotta.

Elinkaariennusteet perustuvat johdannossa (luvussa 2.1) määritettyyn *elinkaarivaikutusten funktioon* (ks. myös luku 12.2.3 päätöksentekoteorian esittelyn yhteydessä). Kuten johdannossa esitetään, elinkaarivaikutusten funktion argumentteina ovat *elinkaarivaikutusten komponentit*. Elinkaarivaikutusten funktio kertoo, kuinka elinkaarivaikutusten komponenteista saadaan laskettua henkilön julkisen talouden elinkaarivaikutukset.

Henkilön elinkaarivaikutusten seitsemän komponenttia ovat:

1. Olotila-ajat (aika Suomessa, ulkomailla ja kuolleena).
2. Julkisen talouden nettovaikutukset alle 15-vuotiaana.
3. Julkisen talouden nettovaikutukset 15–62-vuotiaana (nuoret ja työikäiset).
4. Palkkasummat 18–62-vuotiaana.
5. Muut ennusteparametrit, joiden avulla lähinnä lasketaan palkkasummista eläkkeisiin liittyvät julkisen talouden nettovaikutukset. Lisäksi diskonttauskertoimet kuuluvat näihin parametreihin.
6. Muut kuin eläkkeisiin liittyvät nettovaikutukset yli 62-vuotiaana.
7. Henkilön lasten yhteenlasketut elinkaarivaikutukset.

Käytännössä elinkaarivaikutusten funktio summaa näistä luetelluista komponenteista kaikki julkisen talouden tulot ja vähentää niistä kaikki julkisen talouden menot elinkaaren aikana. Olotila-aika kertoo, onko henkilö Suomessa vai ulkomailla tai kuollut. Jos henkilö on ulkomailla tai kuollut tiettynä kalenterivuonna, niin tuon kalenterivuoden vaikutus henkilön koko elinkaari vaikutukseen on nolla euroa (lukuun ottamatta muutto- ja kuolinvuosia).

Olen esittänyt luvuissa 3–8, kuinka esitellyille seitsemälle elinkaarivaikutusten komponentille laaditaan ennusteet. Nämä komponenttien ennusteet ovat ehdollisia todennäköisyysjakaumia, joten myös niistä elinkaarivaikutusten funktion avulla muodostuvat uusien maahanmuuttajien elinkaariennusteet ovat todennäköisyysjakaumia. Elinkaariennusteiden todennäköisyysjakaumat kuvaavat ennusteisiin sisältyviä epävarmuuksia niin, kuin olen esittänyt johdannon luvussa 2.3. Lisäksi päätöksentekoteoriaa käsittelevän luvun 12 yhteydessä esitän, kuinka epävarmuuksia sisältävistä elinkaariennusteista voidaan tehdä johtopäätöksiä potentiaalisten uusien maahanmuuttajien valitsemisen suhteen – kun maahanmuuttopolitiikan valinnoilla halutaan maksimoida hyöty Suomen julkiselle taloudelle. Uusien maahanmuuttajien elinkaariennusteiden todennäköisyysjakaumat – ja siten elinkaariennusteiden todennäköisyysjakaumat – ovat ehdollisia niillä tiedoilla, joita uusista maahanmuuttajista on saatavilla, kun he muuttavat ensimmäisen kerran Suomeen. Näitä tietoja ovat esimerkiksi sukupuoli, ikä, pääasiallinen toiminta ensimmäisenä vuonna Suomessa, (puutteellinen) koulutustieto, lasten määrä muuttaessa sekä maahanmuuttovuosi.

Esitän seuraavaksi, mistä henkilöistä uusien maahanmuuttajien populaatiot koostuvat elinkaariennusteissa, joiden tulokset esitän tämän luvun lopussa. Ennen elinkaariennusteiden tuloksia esitän lisäksi vielä, kuinka simuloin elinkaariennusteiden todennäköisyysjakaumat sekä kuinka sovellan käytännössä luvuissa 3–8 esitettyjä elinkaarivaikutusten komponenttien mallinnustuloksia elinkaariennusteissa.

9.2 Ennustepopulaatioiden poiminta

Elinkaariennusteita varten poimin Somaliassa ja Irakissa syntyneiden osalta tutkimuskäytössä olevista kokonaispopulaatioista noin 500 henkilön yksinkertaisen satunnaisotoksen ilman palauttamista niistä henkilöistä, jotka olivat muuttaneet Suomeen yli 15-vuotiaina ensimmäisen kerran vuosina 1995–2011. En käyttänyt näitä henkilöitä, joille ennusteet tehdään, tilastollisten mallien estimoinnissa (olotila-, nettovaikutus- ja eläkemaksumallit). Suomessa syntyvien osalta käsittelen keinotekoista vuonna 2000 syntynyttä kohorttia, joka käsittää yhteensä 200 henkilöä, joista 97 on naisia ja 103 miehiä.

Erillisten ennustepopulaatioiden poiminta perustuu tutkimuksen II osan kantavaan ideaan, eli siihen että elinkaariennusteet voidaan laatia juuri Suomeen muuttaneille uusille maahanmuuttajille. Mikäli olisin laatinut ennusteet samoille henkilöille, joiden pohjalta mallien estimoinnit tehtiin, henkilöiden välistä vaihtelua (erilaisuutta) maassa jo olevien maahanmuuttajien ja juuri muuttaneiden välillä ei olisi huomioitu.

Maahanmuuttomuuttovuodeksi elinkaariennusteissa vuosi 2015. Laadin ennusteet sellaiseen skenaarioon perustuen, missä ennustepopulaatioiden henkilöt muuttavat Suomeen ensimmäisen kerran vuonna 2015 – sen sijaan, että he olisivat muuttaneet Suomeen ensimmäisen kerran vuosina 1995–2011. Skenaarion muuttovuosi olisi voinut olla myös jokin muu kalenterivuosi tai useamman kalenterivuoden yhdistelmä. Valitsin kuitenkin vuoden 2015, koska kalenterivuosi on lähellä tämän tutkimuksen II osan julkaisuvuotta. Lisäksi Suomeen kohdistui vuonna 2015 ennennäkemättömän laaja Irakissa ja Somaliassa syntyneiden muuttoliike

Keskeinen oletus näin laadittavien ennusteiden validiteetille on, että Suomeen vuonna 2015 muuttaneet henkilöt olivat toteutuma samanlaisesta dataa generoivasta mekanismista kuin Suomeen 1995–2011 todellisuudessa muuttaneet olivat. Eli 2015 muuttaneet olivat samankaltaisia kuin vuosina 1995–2011 muuttaneet.

Uusiin maahanmuuttajahenkilöihin liittyvän epävarmuuden huomiointi elinkaariennusteissa. Ennustepopulaatioiden henkilöistä tiedetään vain edellä tilastollisissa malleissa olleiden selittävien muuttujien arvot, jotka koskivat tietoja henkilöistä heidän muuttaessaan Suomeen ensimmäisen kerran. Ennustepopulaatioiden henkilöillä on kuitenkin sellaisia havaitsemattomia tekijöitä tai ominaisuuksia, jotka vaikuttavat ennusteisiin, ja joita kutsutaan havaitsemattomaksi heterogeenisuudeksi. Ennustemalleissa tämä havaitsematon heterogeenisuus huomioidaan henkilötason satunnaisvaikutusten avulla erityisesti työikäisten nettovaikutusmalleissa. Nettovaikutusmalleissa huomioidaan lisäksi maahanmuuttovuosiin ja maassaoloaikoihin sisältyvä havaitsematon heterogeenisuus, jonka avulla ennusteisiin saadaan sisällytettyä se epävarmuus, joka ennusteisiin sisältyy ”uusien”, eli ei mallinnusaineistoon sisältyvien, maahanmuuttovuosien ja maassaoloaikojen vuoksi.

Elinkaariennusteet voidaan laatia ”oikeille” uusille maahanmuuttajille samalla tavalla, kuin laadin tässä luvussa keinotekoisen ennustepopulaation henkilöille.

Puute: Otosepävarmuus jää huomioimatta, kun elinkaariennusteiden tulokset yleistetään tulevien uusien maahanmuuttajien väestöille, joita ei ole vielä ”havaittu”. Kun esitettävät elinkaariennustetulokset yleistetään sellaisille tuleville uusille maahanmuuttajille, joita ei ole vielä havaittu – eli jotka eivät ole vielä pyrkimässä muuttamaan Suomeen, joten heistä ei tiedetä sukupuolta, ikää, jne. – ideaalisesti tässä tapauksessa elinkaariennusteissa pitäisi ottaa vielä huomioon uusien maahanmuuttajien ”poimintaan” liittyvä otosepävarmuus.

Tätä otosepävarmuutta ei oteta nyt huomioon, kun elinkaariennusteet perustuvat vain yhden ennustepopulaation poimintaan kaikista Suomeen 1995–2011 muuttaneista. Mikäli uusiin maahanmuuttajiin liittyvät otosepävarmuus haluttaisiin huomioida, niin ennustepopulaatio pitäisi poimia ”tarpeeksi” monta kertaa Suomeen 1995–2011 muuttaneiden joukosta ja suorittaa kaikki mallinnukset ja elinkaariennusteet jokaiselle poimitulle (mallinnus-) ja ennustepopulaatiolle erikseen. Otosepävarmuuden sisältävät elinkaariennusteet laadittaisiin sitten näiden kaikkien poimittujen ennustepopulaatioiden elinkaariennusteiden yli. En kuitenkaan laatinut tällaisia otosepävarmuuden sisältäviä elinkaariennusteita, sillä työmäärä niiden laatimisessa olisi ollut huomattavan suuri.

Mikäli laadittuja elinkaariennusteita sovelletaan jo havaittuihin potentiaalisiin uusiin maahanmuuttajiin

– eli sellaisiin Suomeen muuttavassa oleviin henkilöihin, joista tiedettäisiin heidän sukupuolensa, ikänsä, koulutuksensa, jne. – niin tässä tapauksessa otosepävarmuutta ei tietenkään tarvitsisi huomioida. Näihin ”havaittuihin” uusiin maahanmuuttajiin liittyisi tietenkin myös yllä mainittua havaitsematonta heterogeenisuutta, mutta tämä otetaan elinkaariennusteissa huomioon erityisesti nuorten ja työikäisten nettovaikutusennusteissa.

9.3 Elinkaariennusteiden simulointi ennustepopulaatioille

Laadin elinkaariennusteet ennustepopulaatioiden henkilöiden jokaiselle vuodelle siitä vuodesta lähtien, kun henkilö oli muuttanut ensimmäisen kerran Suomeen (Suomessa syntyneille syntymävuodesta lähtien), siihen vuoteen asti, kunnes henkilö täyttäisi 101 vuotta. Jotta ennusteissa huomioitaisiin epävarmuus, joka sisältyy malliparametrien arvoihin niiden estimoinnin jälkeen, käytin 300 satunnaisesti valittua simuloitua arvoa kunkin parametrin posteriorijakaumasta, jotka oli saatu mallien posteriorijakaumien simulaatioiden yhteydessä. Elinkaariennusteiden odotusarvojen simuloinnista johtuvat keskivirheet (Monte Carlo -keskivirheet) pysyivät kohtalaisen pieninä eli muutamassa tuhannessa eurossa valitulla 300 simulaation määrällä, kuten elinkaariennusteiden yhteydessä esitetyt keskivirheet osoittavat.

Elinkaariennusteissa jokaisen ennustepopulaation jokaista ennustettavaa vuotta kohti tuli näin ollen 300 simuloitua arvoa. Maahanmuuttajaryhmäkohtainen elinkaariennustesimulaatioiden kokonaismäärä ennustepopulaation henkilöille $i = 1, \ldots, N$ oli näin:

$$\sum_{i=1}^{N}(101 - \textit{maahanmuuttoikä}_i) \times 300(\textit{simulaatiota}), \tag{9.1}$$

missä N on ryhmäkohtainen ennustepopulaation henkilöiden kokonaislukumäärä ja $\textit{maahanmuuttoikä}_i$ on ennustepopulaatioon kuuluvan henkilön i ikä, kun hän muuttaa ensimmäistä kertaa Suomeen.

Lopulliset elinkaariennusteet sain henkilövuositasoisista simuloiduista ennusteista niin, että jokaisen henkilön kohdalla ennustesimulaatioiden tuottamat vuosittaiset nettovaikutusten arvot laskettiin yhteen simulaatiokohtaisesti. Elinkaariennusteita syntyi näin 300 kappaletta – eli simulaatioiden lukumäärä – jokaiselle ennustepopulaatioiden sisältämälle henkilölle. Ennustepopulaatiokohtaisesti simuloitujen ennusteiden kokonaislukumäärä oli näin $N \times 100$ simulaatiota, missä N on edellä mainittu ennustepopulaation henkilöiden kokonaislukumäärä. Esimerkiksi Irakissa syntyneiden elinkaariennusteet koostuvat $677 \times 300 = 203\,100$ simuloidusta elinkaariennusteesta.

Elinkaariennusteiden ennustepopulaatiokohtaiset tulokset koostuvat ennustepopulaation kaikkien elinkaariennustesimulaatioiden yhteisjakaumasta ja sitä koskevasta bayesilaisesta tilastollisesta päättelystä eli käytännössä mainitun yhteisjakauman tunnusluvuista. Kokonaisten jakaumien lisäksi käsittelen tuloksia esimerkiksi sukupuolen ja maahanmuuttoiän mukaan eriteltyinä.

9.4 Elinkaarivaikutusten komponenttien ennusteet mallinnuksen pohjalta

Esitän tässä alaluvussa, kuinka elinkaarivaikutusten komponenteille saadaan ennusteet luvuissa 3–8 esitetyistä malleista. Elinkaarivaikutusten komponenttien mallinnuksessa päämääränä oli, että malleja voidaan käyttää uusien maahanmuuttajien elinkaariennusteissa eli tässä esitettävää käyttötarkoitusta varten.

9.4.1 Olotila-aikojen ennusteet

Henkilön olotilalla on maassaolovuosittaisissa ennustesimulaatioissa kolme toisensa poissulkevaa tilaa: henkilö on yhden vuoden lopussa joko Suomessa, ulkomailla tai kuollut. Tila voi ennusteissa vaihtua kerran vuoden aikana maahan- ja maastamuuttoja koskien. Jos henkilö on ollut edellisen vuoden lopussa Suomessa, niin seuraavan vuoden aikana henkilö on voinut muuttaa ulkomaille, muttei muuttaa enää takaisin Suomeen samana vuonna. Mikäli henkilö on puolestaan ollut edellisen vuoden lopussa ulkomailla (muutettuaan aikaisemmin Suomesta pois), niin seuraavan vuoden aikana henkilö on voinut muuttaa takaisin Suomeen, muttei muuttaa enää takaisin ulkomaille samana vuonna. Henkilö voi ennusteissa kuolla kaikkina niinä vuosina, kun hän ei vielä ole kuollut. Uusien maahanmuuttajien olotila-aikojen ennusteissa käytetään samoja selittäviä muuttujia kuin olotilamallinnuksissa luvussa3.2.3.4. Näiden arvot tulevat ennustepopulaatioiden tiedoista. Olotila-aikojen ennusteet uusille maahanmuuttajille vastaavat luvussa 3.2.6 esitettyjä posterioriennustetarkistusten ennusteita.

9.4.1.1 Ennusteet suomessaoloajalle (ja Suomesta poismuuttamiselle)

Mikäli henkilö on ennustesimulaatiossa tietyn vuoden alussa Suomessa (eli edeltävän vuoden lopussa Suomessa) eikä henkilö kuolinennusteen mukaan kuole tuona vuonna, niin maassaoloajan ennuste kertoo, muuttaako henkilö tuona vuonna ulkomaille, vai pysyykö hän edelleen Suomessa. Jokaisessa elinkaariennusteiden simulaatiossa henkilö on ensimmäisen Suomeen-muuttonsa vuoden alussa Suomessa. Maassaoloennusteissa käytetään samoja selittäviä muuttujia kuin varsinaisessa maassaoloajan mallinnuksessa. Näiden arvot tulevat ennustepopulaatioiden tiedoista.

Mikäli henkilö ennusteissa joko muuttaa tiettynä vuonna ulkomaille tai kuolee, tuona vuonna henkilön suomessaoloaika ennusteissa on 0,5 vuotta. Lisäksi kaikkina vuosina, jolloin henkilö on muuttanut ulkomailta Suomeen (ml. ensimmäinen muutto Suomeen) henkilön suomessaoloaika on 0,5 vuotta. Mikäli henkilö on koko vuoden Suomessa, hänen ennustettu suomessaoloaikansa tuona vuonna on 1 vuosi. Henkilön ennustettu suomessaoloaika on puolestaan 0 vuotta niiden vuosien osalta, jolloin henkilö on koko vuoden joko ulkomailla tai kuollut.

Sinä vuonna, kun henkilö täyttää 71 vuotta, tai sen jälkeisinä vuosina, muuttamista ulkomaille ei enää ennusteta. Vastaavasti myöskään ulkomailta takaisin Suomeen muuttamista ei enää ennusteta tuossa iässä. Tämä tarkoittaa toisin sanoen sitä, että henkilö jää loppuiäkseen Suomeen tietyssä elinkaarivaikutuksen simulaatiossa, mikäli hän on Suomessa, kun täyttää 71 vuotta. Vastaavasti ulkomailla olevat jäävät koko loppuiäkseen ulkomaille.

9.4.1.2 Ennusteet ulkomailllaoloajalle (ja Suomeen paluumuuttamiselle)

Ulkomaillaoloajan ennusteissa selittävät muuttujat ovat samat kuin luvussa 3.3.1 esitetyssä mallinnuksessa, mutta nyt selittävien muuttujien arvot ovat peräisin ennusteaineistoista. Lisäksi selittävänä muuttujana oleva suomessaoloaika tulee yllä esitetystä suomesssaoloaikojen ennusteesta. Ulkomaillaoloajan ennusteet laadin niille vuosille, joiden alussa henkilö oli ulkomailla ja ei vielä kuollut. Ulkomaillaoloajan ennusteet kertovat, onko henkilö vielä ulkomailla tällaisen vuoden jälkeen, vai muuttiko hän takaisin Suomeen. Ennustettu ulkomaillaoloaika on 1 vuosi niiden vuosien osalta, jotka henkilö on kokonaan ulkomailla. Ennustettu ulkomaillaoloaika on puolestaan 0,5 vuotta niiden vuosien osalta, jolloin henkilö joko muuttaa Suomesta ulkomaille, muuttaa takaisin Suomeen tai kuolee.

Eläkeikäisten eli yli 62-vuotiaiden nettovaikutusennusteisiin vaikuttaa paitsi ennustettu suomessaolo(aika) niin myös ulkomaillaolo(aika). Esitän tarkemmin olotilojen ennusteiden vaikutukset nettovaikutusennusteisiin eläkeikäisillä alla luvussa 9.4.3.

Koska sekä maassaoloennusteet että ulkomaillaoloennusteet riippuvat Irakissa ja Somaliassa syntyneillä merkittävästi Suomeen muuton muuttovuoden luokasta (ks. Irakia ja Somaliaa koskevat tulokset luvuissa 3.2.7 ja 3.3.2), esitän Irakin ja Somalian osalta uusien maahanmuuttajien elinkaariennusteet erikseen eri maahanmuuttovuoden luokkiin perustuviin estimointituloksiin. Toisin sanoen esitän ennusteet erikseen siten, että maassaoloaika (ja ulkomaillaoloaika) noudattaa vuosina ennen vuotta 1994 Suomeen muuttaneiden olotila-aikojen estimointituloksia, vuosina 1994–1999 Suomeen muuttaneiden olotila-aikojen estimointituloksia sekä vuosina 2000–2009 Suomeen muuttaneiden olotila-aikojen estimointituloksia.

9.4.1.3 Kuolemisennusteet (elinajan ennusteet)

Kuollut on olotiloissa nk. absorboiva tila, eli jos henkilö on kuollut tiettynä vuonna, hän on kuollut todennäköisyydellä 1 myös kaikkina tulevina vuosina.

Ikävuosittaiset kuolinennusteet ovat yksinkertaisia: Mikäli henkilö ei ole vielä kuollut edeltävinä vuosina, niin hänen kuolintodennäköisyytensä tiettynä vuonna riippuu vain henkilön sukupuolesta sekä hänen iästään. Sukupuoli- ja ikävuosikohtaiset kuolintodennäköisyydet eli kuolleisuusluvut ovat Tilastokeskuksen kuolleisuusennusteista Helsinkiä koskien vuodelle 2015 (SVT: *väestöennuste*).

Ikävuodelle 100 lukua ei ennusteista löydy, joten käytin sitä koskien sukupuolittain kuolleisuuslukuja, jotka on saatu pyöristämällä 99:nnen ikävuoden sukupuolittaiset kuolleisuusluvut ylöspäin lähimpään satalukuun (kuolleisuusluvut esitetään kuolleiden määränä 1 000 henkilöä kohti).

Vaikka kuolleisuuslukujen odotetaan laskevan tulevaisuudessa, käytin silti vuoden 2015 lukuja, sillä tutkimuksessa tehtävät elinkaariennusteet eivät ota huomioon eliniän pitenemiseen liittyviä vaikutuksia kuten esimerkiksi elinaikakerrointa karttuneiden eläkkeiden arvon määrittämisessä. Yksinkertaisuuden vuoksi oletan, että kuolleisuusluvut ovat samat ulkomailla olevilla henkilöille kuin Suomessa olevilla henkilöillä.

9.4.2 Nettovaikutusennusteet työikäisille ja nuorille

Nettovaikutusennusteet työikäisille (20–62-vuotiaat) ja nuorille (15–19-vuotiaat) seuraavat luvussa 4 estimoiduista nettovaikutusmalleista, sillä laadin nettovaikutusennusteet maahanmuuttoryhmäkohtaisille uusien maahanmuuttajien ennustepopulaatioille, niin kuin posterioriennustetarkistukset on laadittu luvussa 4.3.7. Elinkaariennusteiden yhteydessä ainoastaan selittävien muuttujien arvot tulevat ennustepopulaatioiden tiedoista. Koska nettovaikutusmalleissa maahanmuuttovuodet on jaettu luokkiin ja maahanmuuttovuosi on elinkaariennusteissa vuosi 2015, oletan elinkaariennusteissa maahanmuuttovuoden vaikutuksen vastaavan nettovaikutusmallien uusinta eli lähinnä vuotta 2015 olevan maahanmuuttovuoden luokan vaikutusta. Työikäisten nettovaikutusennusteisiin lisätään vielä 18:nnesta ikävuodesta lähtien ennusteet työnantajien maksamista työeläkemaksuista ja muista sosiaalimaksuista, joiden ennusteet laadin eläkemaksuennusteiden pohjalta. Suomessa syntyneiden elinkaariennusteiden osalta kaikki henkilöt täyttävät 15 vuotta vuonna 2015.

Olotila-aikaennusteiden vaikutukset. Yllä esitetyt olotila-aikojen ennusteet vaikuttavat suoraviivaisesti siihen, miten vuosittaiset työikäisten nettovaikutusennusteet vaikuttavat koko elinkaaren yli karttuviin ennustettuihin julkisen talouden nettovaikutuksiin. Nuorten ja työikäisten vuosittaiset nettovaikutusennusteet kerrotaan suomessaolon ennusteilla, jolloin saadaan olotilat huomioon ottavat nettovaikutusennusteet elinkaariennusteita varten 15–62-vuotiaille. Mikäli henkilö on ennustetusti koko tietyn vuoden Suomessa, tuon vuoden ennustettu nettovaikutus kerrotaan ykkösellä. Mikäli henkilö on ennustetusti koko tietyn vuoden ulkomailla tai kuollut, tuon vuoden ennustettu nettovaikutus kerrotaan nollalla. Mikäli henkilö ennustetusti muuttaa Suomeen tai Suomesta tietyn vuoden aikana, tai ennustetusti kuolee tietyn vuoden aikana, tuon vuoden ennustettu nettovaikutus kerrotaan 0,5:llä.

Uusien maahanmuuttajien nuorten ja työikäisten nettovaikutusennusteisiin liittyy joitakin oletuksia, jotka käyn seuraavaksi läpi.

Vastaanotto- ja kotouttamismenot. Oletan, että Suomeen turvapaikanhakijoina heidän perheenjäseninään muuttavien vastaanotto- ja kotouttamismenot ovat 50 000 euroa henkilöä kohti . Käytän yksinkertaisuuden vuoksi samaa 50 000 euron summaa myös kiintiöpakolaisina Suomeen muuttavien osalta. Lisäksi oletan, että tämä summa jakaantuu tasaisesti viiden ensimmäisen maassaolovuodelle, eli vastaanotto- ja kotouttamismenot ovat 10 000 euroa henkilöä kohti viiden ensimmäisen vuoden aikana. Tämän summan muodostuminen esitetään luvussa 10.2.1. Samassa luvussa käsitellään myös summasta mahdollisesti puuttuvia vastaanotto- ja kotouttamismenoja.

Maassaoloajan vaikutuksen rajoite/rajoittamaton vaikutus. Nuorten ja työikäisten nettovaikutusennusteiden kohdalla oli päätettävä, kuinka maassaoloaika vaikuttaisi elinkaariennusteissa, kun maassaoloaika on yli 21 vuotta. Tällöin mennään havaintoaineiston ulkopuolelle, joka ulottuu 21 maassaolovuoteen asti nuorten ja työikäisten nettovaikutusten osalta. Päädyin rajoittamaan maassaoloajan vaikutuksen Irakissa syntyneiden tuloksissa 21 maassaolovuoteen eli yli 21 vuoden jälkeen maassaoloajalla on sama

vaikutus kuin 21 maassaolovuodella. Käytännössä, koska Somaliassa syntyneiden elinkaariennusteet olivat muutenkin niin negatiivisia, annoin maassaoloajan vaikutuksen kasvaa 21 maassaolovuoden jälkeen lineaarisesti.

Joka tapauksessa maassaoloajan vaikutuksen rajoitus (Irak) tai rajoittamattomuus (Somalia) eivät vaikuta juurikaan elinkaariennusteisiin keskiarvotuloksiin, ainakaan koko ennustepopulaatioita koskien, kun tarkastellaan koko iän Suomessa olevia. Tämä on osoitettu Irakissa syntyneiden osalta taulukossa 9.26 (s. 444), jossa elinkaariennusteet on muodostettu niin, että maassaoloajan vaikutukselle ei ole asetettu mitään rajoitteita, vaan se kasvaa lineaarisesti 21 maassaolovuoden jälkeenkin. Tämän rajoittamattoman tapauksen elinkaariennusteiden odotusarvo on -654 000 euroa, mutta samaan aikaan maassaolovaikutuksen rajoitteen sisältävän elinkaariennusteen odotusarvo on -630 000 euroa (ks. taulukko 9.25, 443)

Somaliassa syntyneiden osalta tämä on puolestaan osoitettu taulukossa 9.5 (s. 426), jossa elinkaariennusteet on muodostettu niin, että maassaoloajan vaikutus on rajoitettu 21 maassaolovuoteen. Mikäli tuloksia verrataan rajoittamattoman maassaolovaikutuksen tuottamiin elinkaariennusteisiin taulukossa 9.4 (s. 425), niin huomataan, että elinkaariennusteiden odotusarvo muuttuu -777 000:sta (maassaoloajan rajoittamattoman vaikutus) -793 000 (maassaoloajan 21 maassaolovuoteen rajoitettu vaikutus).

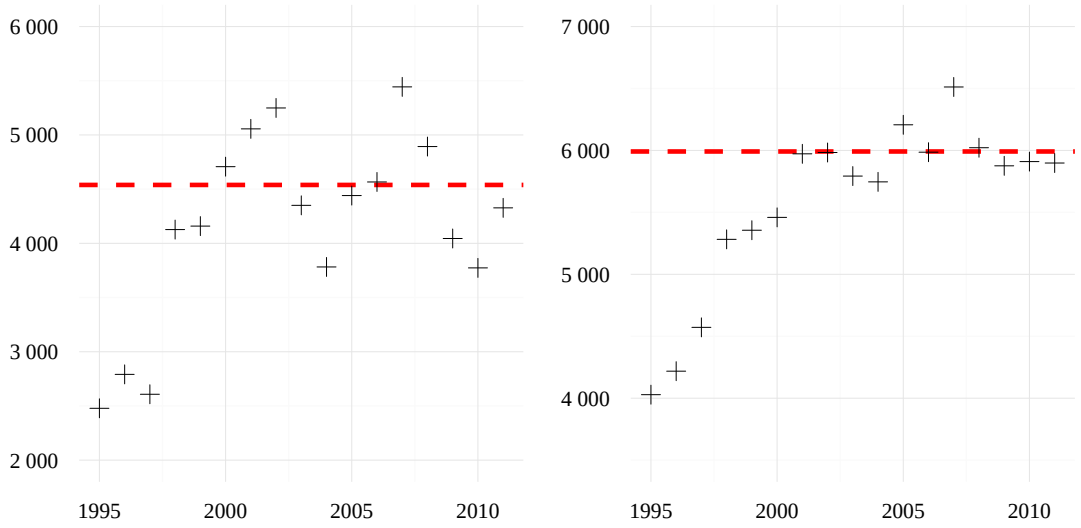
Maahanmuuttovuosiluokka. Koska nettovaikutusmalleissa maahanmuuttovuodet on jaettu luokkiin ja maahanmuuttovuosi on elinkaariennusteissa vuosi 2015, oletan elinkaariennusteissa maahanmuuttovuoden vaikutuksen vastaavan nettovaikutusmallien uusinta eli lähinnä vuotta 2015 olevan maahanmuuttovuoden luokan vaikutusta. Tämä maahanmuuttovuoden luokka on sekä Irakissa että Somaliassa syntyneiden nettovaikutusmalleissa vuosina 2007–2011 Suomeen muuttaneiden luokka.

Ei trendiä kalenterivuoden suhteen. Nuorten ja työikäisten nettovaikutusennusteissa ei ole trendiä kalenterivuoden suhteen. Tämä johtuu kahdesta asiasta. Ensiksi, havaintoaineiston perusteella työikäisten nettovaikutuksissa ei ole trendiä kalenterivuoden suhteen (vuosina 2001–2011). Toiseksi, minulla ei ole sellaista havaintoaineiston ulkopuolista informaatiota, joka kertoisi, että tulevaisuudessa työikäisten keskimääräisissä nettovaikutuksissa olisi trendi kalenterivuoden suhteen.

Kuvassa 9.1 alla on esitetty Suomessa syntyneiden työikäisten 20–62-vuotiaiden nettovaikutusten havaintojen vuosittaiset keskiarvot ja mediaanit vuosina 1995–2011. Luvut ovat euroja per Suomessa syntynyt työikäinen henkilö. Punainen katkoviiva vasemmassa osakuvassa on kaikkien havaintojen yhteinen keskiarvo vuosina 2001–2011, ja oikeanpuoleisessa osakuvassa punainen katkoviiva on vuosien 2001–2011 mediaanien keskiarvo.

Kuten kuviosta 9.1 ilmenee, sekä keskiarvoissa ja mediaaneissa ei ole havaittavissa mitään selvää – nousevaa tai laskevaa – trendiä vuosina 2001–2011. Myöskään formaalimmat tilastolliset testit eivät tue trendin olemassaoloa kalenterivuoden suhteen. Vuosina 1995–2000 julkisen talouden keskiarvot ja mediaanit ovat vuosia 2001–2011 pienemmät. Tämä johtuu erityisesti siitä, että saaduissa tulonsiirroissa on tapahtunut jokin muutos vuosien 1995–2000 aikana vuoteen 2001 mennessä (ks. tutkimuksen I osa, luku 6.2.1.1, aikasarjat kuvio 8). Tutkimusaineiston perusteella on vaikeaa sanoa, johtuuko muutos saaduissa tulonsiirroissa muutoksesta jossain yksittäisessä saadussa tulonsiirrossa, sillä tutkimusaineistossa ei ole käytössä yksittäisten tulonsiirtojen rekisteritietoja vaan kaikkien eri saatujen tulonsiirtojen summa

henkilölle vuosittain.



(a) *Keskiarvot. Punainen katkoviiva on vuosien 2001–2011 yhteinen keskiarvo.*

(b) *Mediaanit. Punainen katkoviiva on vuosien 2001–2011 mediaanien yhteinen keskiarvo.*

Kuva 9.1: *Suomessa syntyneiden työikäisten 20–62-vuotiaiden nettovaikutusten havaintojen vuosittaiset keskiarvot ja mediaanit vuosina 1995–2011 (euroa per työikäinen).*

Havaintoaineiston lisäksi minulla ei ole käytettävissä muuta sellaista havaintoaineiston ulkopuolista informaatiota, joka kertoisi, että työikäisten nettovaikutuksilla olisi havaintoaineiston eli vuoden 2011 jälkeisessä tulevaisuudessa trendi ajan suhteen. Trendin mahdollisen olemassaolon tuottaman lisäepävarmuuden voisi kuitenkin lisätä helposti työikäisten nettovaikutusennusteisiin siten, että nettovaikutusten ennustemallin odotusarvoon lisättäisiin satunnainen trendiä kuvaava parametri. Tälle parametrille määritettäisiin sitten todennäköisyysjakauma, joka kuvaisi siihen – eli trendin olemassaoloon – liittyviä epävarmuuksia ennusteissa.

Mikäli lisäisin tällaisen satunnaisen trendiparametrin työikäisten nettovaikutusten ennustemalliin, niin antaisin sen todennäköisyysjakauman odotusarvoksi nollan. Tämä ratkaisu kuvaisi sitä, ettei minulla ole käytössäni sellaista informaatiota, joka kertoisi, olisiko mahdollinen trendi todennäköisemmin positiivinen (kasvava) vai negatiivinen (laskeva) trendi kalenterivuoden suhteen. Käsittelen sekä johdannossa (luvussa 2.3.4) että päätöksentekoteorian yhteydessä (luvussa 12.3) nk. aineiston ulkopuolisten tekijöiden vaikutuksia elinkaarivaikutusten ennustetodennäköisyysjakaumiin. Päätöksentekoteoriaa koskevassa luvussa osoitan, että mikäli aineiston ulkopuolisten tekijöiden vaikutuksen odotusarvo on nolla nettovaikutusennusteissa, aineiston ulkopuoliset tekijät eivät vaikuta potentiaalisen uuden maahanmuuttajan valintaan, kun valinnalla maksimoidaan Suomen julkisen talouden hyöty.

9.4.3 Palkkasummaennusteet ja eläkkeisiin liittyvät nettovaikutusennusteet

9.4.3.1 Palkkasummaennusteet

Maassaoloajan vaikutuksen rajoite. Maassaoloaika on käytännössä pakko rajoittaa elinkaariennusteissa 21 maassaolovuoteen, eli 21 maassaolovuoden maassaoloaika palkkasummaennusteissa saman verran kuin 21 maassaolovuotta. Tämä johtuu siitä, että palkkasummamalleissa käytetään epälineaarisia linkkifunktioita (logaritmi- ja logit-linkkifunktiot), mikä tarkoittaa sitä, että rajoittamattoman maassaoloajan käyttö johtaisi joko epärealistisen suuriin tai pieniin palkkasummaennusteisiin.

Joka tapauksessa, mikäli palkkasummien nollaosuuksien ja yli nollien keskiarvojen kehitystä tarkastellaan Somaliassa syntyneillä maassaolovuosien mukaan (kuva 5.8, s. 286), niin huomataan, että nollien osuudet pysyvät suunnilleen samalla tasolla noin 7 maassaolovuoden jälkeen. Samoin yli nollien kasvava kehitys näyttää pysähtyvän noin 10 maassaolovuoden jälkeen – tai keskiarvo nousee ainakin hyvin hitaasti.

Irakissa syntyneiden palkkasummien nollaosuuksien ja yli nollien kehitys maassaoloajan suhteen on esitetty kuvassa 5.8 (s. 286). Kuvan perusteella myös Irakissa syntyneiden nollien osuuksissa ei tapahdu juurikaan kehitystä noin 5 maassaolovuoden jälkeen. Sen sijaan yli nollien keskiarvot vaikuttavat kasvavan aina 20 maassaolovuoteen asti. Viitatun kuvan perusteella on vaikeaa sanoa, että pysähtyykö yli nollien keskiarvojen kasvu noin 20 maassaolovuoden kohdalle. Viitatut kuviot tosin ovat hieman kehnoja maassaoloajan vaikutuksen marginaalivaikutuksen arvioimiseksi, sillä maassaoloaika kasvaa iän myötä ja keskimäärin nuorina muuttaneilla maahanmuuttajilla maassaoloajan kasvu tarkoittaa myös iän nousua parhaaseen työikään.

Maahanmuuttovuosiluokka. Koska palkkasummamalleissa maahanmuuttovuodet on jaettu luokkiin ja maahanmuuttovuosi on elinkaariennusteissa vuosi 2015, oletan elinkaariennusteissa maahanmuuttovuoden vaikutuksen vastaavan nettovaikutusmallien uusinta eli lähinnä vuotta 2015 olevan maahanmuuttovuoden luokan vaikutusta. Tämä maahanmuuttovuoden luokka on sekä Somaliassa syntyneiden palkkasummamalleissa vuosina 2002–2014 Suomeen muuttaneiden luokka, ja Irakissa syntyneiden palkkasummamalleissa tämä maahanmuuttovuoden luokka on 2006–2014 Suomeen muuttaneiden luokka.

9.4.3.2 Eläkkeisiin liittyvät nettovaikutusennusteet

Eläkkeisiin liittyvät nettovaikutusennusteet koostuvat seuraavista ennusteista. Positiiviset nettovaikutusennusteet ovat euromäärältään ei-negatiivisia. Negatiiviset nettovaikutusennusteet ovat euromäärältään ei-positiivisia.

1. Positiivinen nettovaikutus:

– Työnantajien maksamat lakisääteiset työeläkevakuutusmaksut ja sosiaalivakuutusmaksut (18–62-vuotiaita koskien).

– Välittömät verot eläketuloista.

– Pääomatuloverot eläkeikäisenä.

– Välilliset verot eläkeikäisenä.

2. Negatiivinen nettovaikutus:

– Maksettavat kokonaiseläkkeet (työeläke + kansaneläke + takuueläke).

– Maksettavat eläkeläisten asumistuet

Yksinkertainen eläkemalli, joka pitkälti VATT:n tutkimuksen mukainen. Palkkasummien – ja siten – maksettujen lakisääteisten eläkemaksujen ennusteet vaikuttavat elinkaariennusteisiin pääasiassa eläkeiän osalta eli yli 62-vuotiaana. Käytin elinkaariennusteissa yksinkertaista eläkejärjestelmää, joka on VATT:n tutkimuksen (2014a) kaltainen. VATT:n tutkimuksessa oletetaan, että kaikki henkilöt jäävät eläkkeelle 63-vuotiaina (VATT 2014b). Seuraten tätä esimerkkiä oletan elinkaariennusteissa, että kaikki henkilöt jäävät eläkkeelle sinä vuonna, kun he täyttävät 63 vuotta.

Ennustettu maksettava työeläke karttuu 18–62-vuotiaana. Koska todellisuudessa karttumista tapahtuu joidenkin henkilöiden osalta myös yli 62-vuotiaana, karttuvat työeläkkeet ovat todellista hieman pienempiä. Lisäksi VATT:n tutkimuksen tapaan elinkaariennusteiden eläkkeet muodostuvat vain vanhuuseläkkeistä, eli esimerkiksi työkyvyttömyyseläkkeitä ei ole mukana. Vanhuuseläkkeet karttuvat tässä tutkimuksessa vain lakisääteisistä eläkevakuutusmaksuista, eli mukana ei ole vapaaehtoisia eläkkeitä eikä myöskään eläkekarttumia muista lähteistä kuin työeläkemaksuista kuten esimerkiksi suoritetuista yliopistotutkinnoista tai joistakin sosiaalietuuksista, joista karttuu eläkettä. Työeläkkeen lisäksi mukana on ennustettu saatava kansaneläke ja takuueläke.

Suurin ero VATT:n tutkimukseen on, että oletan palkkojen vuosittaisen reaalisen kasvun olevan odotusarvoisesti nollaa suurempi myös tulevaisuudessa. Esitän palkkojen reaalisen kasvun ennusteet ja niiden perusteet tarkemmin luvussa 6 Esitän samassa luvussa myös tarkemmin, kuinka vuosittaisista palkkasummien ennusteista muodostetaan elinkaariennusteiden nk. *muiden ennusteparametrien* avulla maksettujen lakisääteisten eläkevakuutusten ja muiden eläkkeisiin liittyvien nettovaikutusten ennusteet.

Suomessaolo- ja ulkomaillaoloaikojen ennusteiden vaikutukset eläkkeisiin liittyviin nettovaikutusennusteisiin. Eläkeikäisten eli yli 62-vuotiaiden nettovaikutusennusteisiin vaikuttaa paitsi ennustettu suomessaolo(aika) niin myös ulkomaillaolo(aika):

- Olotilaennusteet vaikuttavat maksettavien eläkkeiden määrään siten, että kansaneläkettä eikä takuueläkettä makseta, jos henkilö on ulkomailla. Mikäli henkilö on kuollut, niin henkilölle ei makseta enää mitään eläkkeitä. Olotilaennusteet vaikuttavat myös työeläkkeen karttumiseen siten, että eläkettä ei kartu, jos henkilö on ennusteen mukaan ulkomailla.

- Mikäli henkilö on vuosittaisen olotilaennusteen mukaan Suomessa, eläketulot, joista välittömät verot maksetaan, käsittävät sekä työeläkkeen, kansaneläkkeen että takuueläkkeen. Ulkomailla olevan välittömät verot maksetaan vain työeläkkeistä. Tässä oletan, että ennustetuista työeläkkeistä maksetaan kaikki välittömät verot Suomeen, vaikka henkilö olisikin ulkomailla.
- Jos henkilö on olotilaennusteen mukaan ulkomailla, niin henkilön eläkeiän (yli 62-vuotiaana) pääomatulojen ennustamiseen käytetään eläketuloista vain työeläkettä (ks. tarkemmin eläkeiän pääomatulojen ennustaminen luvuissa 6.1.5 ja 6.2.3).
- Henkilön ennustettu vuosittainen maksettu välillisten verojen summa eläkeiässä (yli 62-vuotiaana) on aina tasan nolla euroa silloin, kun henkilö on olotilaennusteen mukaan ulkomailla.

9.4.3.3 Ennustetut työeläkkeet, kansaneläkkeet ja takuueläkkeet verrattuna ETK:n pitkän aikavälin vastaaviin ennusteisiin

Eläkkeisiin liittyvien ennusteiden uskottavuuden varmistamiseksi verrataan erityisesti Suomessa syntyneelle kantaväestölle ennustettuja eläkkeitä vastaaviin ETK:n (2016) laatimiin pitkän aikavälin eläkeennusteisiin (ETK 2016). Koska Suomessa syntyvän ennustepopulaation kaikki henkilöt ovat syntyneet vuonna 2000 ja koska he jäävät elinkaariennusteissa kaikki eläkkeelle sen vuoden lopussa, jolloin he täyttävät 62 vuotta, verrataan ennustettuja alkavia eläkkeitä vastaaviin ETK:n (2016) ennusteisiin. ETK:n raportin (2016) lisäksi pyysin tietoja suoraan ETK:lta sähköpostitse (ETK 2018).

Alla taulukossa 9.1 on esitetty sekä ETK:n pitkän aikavälin keskiennusteet alkaville eläkkeille vuonna 2062 että kaikille maksussa oleville eläkkeille samana vuonna. Taulukossa on esitetty vuonna 2062 alkavien eläkkeiden ennusteet Suomessa syntyneille sekä alkavien eläkkeiden ennusteet Irakissa ja Somaliassa syntyneille (eläkkeelle jääminen viimeistään vuonna 2062). Taulukossa on esitetty vuoden 2015 euroissa (ilman elinajankertoimen vaikutusta) saatavien työeläkkeiden, kansaneläkkeen, kansaneläkkeen ja takuueläkkeen summan sekä koko (”yhteensä”) eläkkeen keskiennusteet. Lisäksi taulukossa on ennusteissa käytetyt keskimääräiset iät jäädä eläkkeelle.

Taulukko 9.1: *ETK:n (2016) pitkän aikavälin keskiennusteet alkamassa sekä maksussa oleville eläkkeille vuonna 2062 taulukon kahdella ylimmällä tulosrivillä. Viidellä alemmalla tulosrivillä tämän tutkimuksen eläke-ennusteet Suomessa, Irakissa ja Somaliassa syntyneille ennustepopulaatioille. Irakissa ja Somaliassa syntyneitä koskien keskiennusteet sekä koko ennustepopulaatioille (kaikki) että 20–24-vuotiaan Suomeen muuttaville (20–24-v. muuttavat). Suomessa syntyvät jäävät kaikki eläkkeelle vuonna 2062. Keskiennusteet vuoden 2015 euroissa työeläkkeelle, pelkälle kansaneläkkeelle, kansaneläkkeen ja takuueläkkeen summalle sekä eläke yhteensä työeläkkeen ja kansaneläkkeiden summana. Eläkkeelle jäämisen keski-ikä ennusteissa. Lähteet: ETK (2016) ja ETK (2018).*

Ennuste	Työeläke	Kansanel.	Kansanel. + takuuel.	Yhteensä[a]	Eläkeikä
ETK, alkavat	3 023[b]	115,9	–	3 139	64,7
ETK, maksussa olevat	2 852[c]	–	156	3 008	64,7
Suomi	2 888	126	155	3 044	63
Irak (kaikki)	407	475	564	971	63
Irak (20–24-v. muuttavat)	500	464	523	1 023	63
Somalia (kaikki)	594	402	482	1 076	63
Somalia (20–24-v. muuttavat)	705	381	446	1 151	63

[a] sis. palkattoman karttuman ja VEKL:n. ETK:n ennusteissa työeläkkeen osuus jaettu vuoden 2062 ennustetulla elinajankertoimella.

[b] ETK:n keskiennuste vuodelle 2062 jaettu vuoden 2062 ennustetulla elinajankertoimella.

[c] ETK:n keskiennuste vuodelle 2062 jaettu vuoden 2062 ennustetulla elinajankertoimella. Koska elinaikakerrointa sovelletaan vain alkaviin eläkkeisiin, jakaminen v. 2062 elinajankertoimella tekee luvusta hieman liian suuren.

Taulukosta 9.1 voidaan päätellä, että Suomessa syntyneelle kantaväestölle tässä tutkimuksessa ennustetut:

- Keskityöeläkkeet vastaavat noin 96%:sti ETK:n ennustamia vuonna 2062 alkavia työeläkkeitä.
- Ennustetut pelkät kansaneläkkeet ovat noin 109 %:ia ETK:n ennustamista alkavien eläkkeiden pelkistä kansaneläkkeistä vuonna 2062.
- Ennustettujen alkavien kansaneläkkeiden ja takuueläkkeiden summa vastaa lähes täysin ETK:n ennustamista maksussa olevien kansaneläkkeiden ja takuueläkkeiden summasta vuonna 2062.
- Alkavat keskieläkkeet vastaavat noin 97%:sti ETK:n ennustamia vuonna 2062 alkavia keskieläkkeitä.

Kaiken kaikkiaan tässä tutkimuksessa laadittujen eläkkeiden keskiennusteet Suomessa syntyneelle kantaväestölle vastaavat ETK:n (2016) laatimia pitkän aikavälin eläke-ennusteiden keskiennusteita. Pieni verrattuna ETK:n alkavien eläkkeiden keskiennusteisiin saattaa johtua esimerkiksi siitä, että ETK:n ennusteissa henkilöt siirtyvät eläkkeelle keskimäärin 1,7 vuotta myöhemmin kuin tässä tutkimuksessa, kuten taulukosta 9.1 kävi yllä ilmi.

9.4.4 Alle 15-vuotiaiden ja yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutusennusteet

Alle 15-vuotiaiden ja yli 62-vuotiaiden muut kuin eläkkeisiin liittyvät nettovaikutusennusteet on esitetty luvussa 7.1. Sekä alle 15-vuotiaiden menojen että yli 62-vuotiaiden muihin kuin eläkkeisiin liittyvien menojen ennusteissa on oletuksena, että *reaalisilla* – eli vuoden 2011 euroisussa ilmaistuilla – menoilla ei ole trendiä kalenterivuoden suhteen. Olen perustellut tämän oletuksen mainitussa luvussa 7.1 vuosia 1995–2015 koskeviin havaintoihin perustuen. Oppilaskohtaiset reaaliset koulutusmenot ovat pysyneet suunnilleen samalla tasolla vuosina 2000–2015, ja reaalisissa keskimääräisissä vuosittaisissa terveys- ja sosiaalihuollon menoissa ei ole havaittavissa trendiä vuosina 1995–2015.

9.4.5 Maahanmuuttajien lasten yhteenlasketut nettovaikutusennusteet

Maahanmuuttajan elinkaariennusteessa maahanmuuttajan lapsien yhteenlasketut elinkaariennusteet muodostetaan, kuten on esitetty luvussa 8. Maahanmuuttajan lasten elinkaarivaikutusten ennusteet koostuvat kahdesta komponentista: yhden lapsen elinkaarivaikutuksen ennusteesta sekä lasten lukumäärän ennusteesta.

Yhden lapsen elinkaariennuste. Yhden lapsen elinkaarivaikutusten ennusteessa keskeistä on kolme oletusta:

1. Maahanmuuttajien 22-vuotiaita lapsia koskevilla tiedoilla voidaan ennustaa näiden lasten julkisen talouden keskimääräisiä nettovaikutuksia 32–35-vuotiaana ja palkkasummia 35–39-vuotiaana – samalla tavoin kuin Suomessa syntyneiden 22-vuotiaiden tiedoilla vuosina 1995–2001 mallinnettiin luvussa 8.2 samojen henkilöiden keskimääräisiä nettovaikutuksia 32–35-vuotiaana ja palkkasummia 35–39-vuotiaana.
2. Maahanmuuttajan lapsen poikkeama Suomessa syntyneiden keskimääräisestä nettovaikutuksesta 32–35-vuotiaana on samalla henkilön nettovaikutusten poikkeaman odotusarvo ikävuosittaisista Suomessa syntyneiden keskimääräisistä nettovaikutuksista 32–62-vuotiaana.
3. Maahanmuuttajan lapsen palkkasummaennusteiden eroja verrattuna Suomessa syntyneiden ennusteisiin 35–39-vuotiaana voidaan käyttää palkkasummien erojen ennusteina verrattuna kantaväestöön 32–62-vuotiaana.

Ainakin vielä tässä elinkaarivaikutuksia käsittelevän tutkimustekstin versiossa maahanmuuttajien lasten elinkaariennusteiden erot kantaväestöön nähden koskevat vain eroja 32–62-vuotiaana.

Maahanmuuttajien lasten määrien ennusteet. Maahanmuuttajien lasten määrät perustuvat rekisteritietoihin maahanmuuttajien mukana Suomeen muuttaneista lapsista sekä maahanmuuttaja- ja ikäryhmittäisiin hedelmällisyyslukuihin vuosina 2000–2015. Jälkimmäisten avulla ennustetaan maahanmuuttajien

Suomessa saamien lapsien määrät maahanmuuttajaryhmittäin, kuten on esitetty tarkemmin luvussa 8.6. Maahanmuuttajien mukana muuttavien lasten määrä on Irakissa ja Somaliassa syntyneiden osalta heidän keskimääräiseen kokonaislapsimäärään nähden alhainen, joten yksinkertaisuuden vuoksi oletan, että he kaikki muuttavat Suomeen vanhempiensa mukana 0-vuotiaina.

Maahanmuuttajan lasten yhteenlasketut nettovaikutusennusteet. Yhden maahanmuuttajan lasten yhteenlasketut nettovaikutusennusteet muodostuvat tämän henkilön lastenmääräennusteesta sekä näitä lapsia koskevista elinkaariennusteista. Allokoin kaikki lapsia koskevat yhteenlasketut elinkaariennusteet naisille, koska maahanmuuttajien lasten määrien ennusteet lasketaan vain naisille. Lasten määrät ennustetaan vain naisille, koska yleisesti ottaen (kokonais)hedelmällisyysluvut ilmoitetaan lapsimäärinä per nainen.

Olotila-aikojen ennusteiden vaikutus lasten nettovaikutusennusteisiin. Maahanmuuttajanaisten olotila-aikaennusteet vaikuttavat heidän lastensa yhteenlaskettuun nettovaikutusennusteeseen. Alle 16-vuotiaiden osalta oletan, että lapset ovat tiettynä kalenterivuonna Suomessa vain, jos olotilaennusteen mukaan heidän äitinsä on Suomessa. Mikäli äiti on olotilaennusteen mukaan ulkomailla, myös lapset ovat ennusteissa ulkomailla – ja siten heistä ei aiheudu julkisen talouden menoja tuona vuonna. Mikäli maahanmuuttajanainen on olotilaennusteen mukaan Suomessa, kun hänen lapsensa täyttää 16 vuotta, oletan tällöin, että lapsi jää koko loppuiäkseen Suomeen. Mikäli taas maahanmuuttajanainen on olotilaennusteen mukaan ulkomailla, kun hänen lapsensa täyttää 16 vuotta, oletan tällöin, että lapsi jää koko loppuiäkseen ulkomaille.

Elinkaariennusteiden tulokset uusien maahanmuuttajien elinkaariennusteissa. Taulukossa 9.2 alla esitetään maahanmuuttajien lasten ennusteita koskevia keskiarvotuloksia Somaliassa ja Irakissa syntyneitä ennustepopulaation naisia koskien, kun naiset ovat maahanmuutosta kuolemaan asti Suomessa. Mikäli naisten poismuutot huomioidaan elinkaariennusteissa, taulukossa 9.2 esitettävät tulokset olisivat erilaisia. Varsinaisissa esitetyissä elinkaariennustetuloksissa naisten poismuuttojen vaikutukset huomioidaan. Se, miten taulukossa 9.2 esitettäviin tuloksiin päädytään, on esitetty maahanmuuttajien lasten ennustettuja elinkaarivaikutuksia käsittelevässä luvussa 8.

Taulukko 9.2: *Maahanmuuttajien lasten ennusteita koskevia keskiarvotuloksia Somaliassa ja Irakissa syntyneitä ennustepopulaation naisia koskien, kun naiset ovat maahanmuutosta kuolemaan asti Suomessa. Keskiarvojen selitteet löytyvät alla tekstistä sekä maahanmuuttajien lasten vaikutusten mallintamista koskevassa luvussa 8.*

Keskiarvoennuste	Irak	Somalia
Lapsia per nainen	2,55	3,53
Nettovaikutusero: 32–62-v. (€)	-3 400	-3 800
Hurdle-ero (ero todennäköisyydessä)	0,12	0,17
Gamma1-tn-ero (ero todennäköisyydessä)	0,02	0,02
Gamma1-ero (€)	-9 600	-11 100
Gamma2-ero (€)	-4 100	-5 100
Yhden lapsen ennuste (€)	-170 000	-191 000
Lasten vaikutukset yhteensä per nainen (€)	-434 000	-674 000

Lapsia per nainen. Taulukon 9.2 ensimmäisen rivin luvut kertovat, kuinka monta lasta Irakissa ja Somaliassa syntyneillä yhteensä keskimäärin on. Lapset voivat muuttaa joko naisten (äitien) mukana Suomeen (vain äitien muuttovuonna mukana muuttaneet lasketaan mukaan), tai naiset voivat synnyttää lapset Suomessa. Näin lapsia karttuu Irakissa syntyneitä naisia koskien ennusteissa keskimäärin 2,55 lasta naista kohti ja Somaliassa syntyneitä naisia kohti 3,53 lasta.

On syytä huomata, että nämä luvut eivät ole Irakissa ja Somaliassa syntyneiden naisten *kokonaishedelmällisyyslukuja*, jotka kertovat, kuinka monta lasta he saisivat keskimäärin, mikäli he olisivat Suomessa koko lastensaanti-ikänsä. Taulukon 9.2 esitetyt luvut ovat kokonaishedelmällisyyslukuja pienempiä, sillä yli 15-vuotiaana Suomeen muuttavat naiset viettävät vain osan lastensaanti-iästään Suomessa ja esitettyjen lukujen perusteella ainakin osa heidän ulkomailla saamistaan lapsistaan ei muuta Suomeen ainakaan naisten maahanmuuttovuonna.

Nettovaikutusero: 32–62-v. Työikäisten nettovaikutusmallin keskimääräinen vuosittainen euromääräinen ero verrattuna Suomessa syntyneen kantaväestön nettovaikutusennusteisiin. Eli Irakissa syntyneillä lapsilla nettovaikutusennusteet 32–62-vuotiaana ovat keskimäärin -3 400 euroa pienemmät kuin Suomessa syntyneellä kantaväestöllä. Somaliassa syntyneillä tämä ero on keskimäärin -3 800 euroa. Luvussa 8.2.4 on esitetty tarkemmin, kuinka näihin tuloksiin päästään.

Palkkasummaennusteiden erot 32–62-v: Hurdle-ero, Gamma1-tn-ero, Gamma1-ero, Gamma2-ero. Palkkasummaennusteet koostuvat oikeastaan neljästä todennäköisyysjakaumasta: nk. hurdlen jakaumasta, gamma1-jakaumasta, gamma2-jakaumasta sekä ennusteen todennäköisyydestä kuulua gamma1-jakaumaan (ks. tarkemmin palkkasummamallinnusta käsittelevä luku 5). Hurdle kertoo yksinkertaisesti todennäköisyyden sille, että palkkasummaennuste saa arvokseen nollan. Taulukon 9.2 hurdle-erot kertovat toisin sanoen, mikä on Irakissa ja Somaliassa syntyneiden maahanmuuttajien lasten todennäköisyyksien ero verrattuna kantaväestöön 32–62-vuotiaana, että palkkasumma on vuoden aikana tasan nolla.

Irakissa syntyneiden lapsilla tämä todennäköisyys on 0,12 (12 %-yks.) korkeampi kuin kantaväestöllä, ja Somaliassa syntyneillä tämä ero on 0,17. Eli jos Suomessa syntyneen kantaväestön henkilöllä keskimääräinen todennäköisyys sille, että palkkasumma on nolla, on esimerkiksi 0,20, niin vastaavalla Irakissa ja Somaliassa syntyneiden lapsilla tämä todennäköisyys olisi 0,32 ja 0,37.

Nk. gamma1-jakauma koskee karkeasti ottaen vain osan vuotta töissä olevien palkkasummien ennustejakaumaa, ja gamma2-jakauma puolestaan koskee koko vuoden täysaikaisissa töissä olevien palkkasummien ennustejakaumaa. ”Gamma1-tn” on todennäköisyys sille, että palkkasummaennuste kuuluu gamma1-jakaumaan, eli että henkilö ei ole koko vuotta täysiaikaisessa työssä. ”Gamma1-tn-ero” on näin ollen tämän todennäköisyyden keskimääräinen ero verrattuna kantaväestöön. Sekä Irakissa että Somaliassa syntyneiden lapsilla palkkasummaennusteen todennäköisyys kuulua vain osan vuotta töissä olevien jakaumaan 32–62-vuotiaana on 0,02 korkeampi kuin kantaväestöllä.

Gamma1-jakauma on vain osan vuotta töissä olevien palkkasummaennusteiden jakauma. Irakissa syntyneiden lapsilla tämän jakauman tuottamien ennusteiden keskiarvot ovat 32–62-vuotiaana 9 600 euroa pienempiä vuodessa kuin kantaväestöllä, ja Somaliassa syntyneillä tämä ero on 11 100 euroa vuodessa. Koko vuoden täysaikaisissa töissä olevien ennustejakauman (gamma2-jakauma) osalta keskimääräiset palkkasummaennusteet ovat Irakissa syntyneiden maahanmuuttajien lasten osalta 32–62-vuotiaana 4 100 euroa vuodessa pienempiä kuin kantaväestöllä, ja Somaliassa syntyneiden maahanmuuttajien lasten osalta ero on 5 100 euroa.

Yhden lapsen ennuste. Yhden keskimääräisen Irakissa syntyneen maahanmuuttajan lapsen elinkaariennuste on 170 000 euroa negatiivisempi kuin Suomessa syntyvän kantaväestön lapsen elinkaariennuste. Somaliassa syntyneiden maahanmuuttajien lasten osalta tämä ennuste-ero on keskimäärin 191 000 euroa per lapsi. Nämä erot muodostuvat kolmesta eri komponentista:

1. Maahanmuuttajien suunnatusta erityisopetuksesta peruskoulussa. Tämän menon vuosittainen summa vuoden 2011 euroissa on 1 485 euroa per oppilas vuodessa, joten koko peruskoulun ajalta menoja syntyy näin yhteensä 13 365 euroa kantaväestöön verrattuna. Tämä on noin 14 000 euroa vuoden 2015 euroissa.

2. Keskimääräinen ero nettovaikutusennusteissa 32–62-vuotiaana: Keskimäärin eroa karttuu: $31 \times -3400\text{e} = -105400\text{e}$ Irakissa syntyneiden lapsilla, ja Somaliassa syntyneiden lapsilla puolestaan: $31 \times -3800\text{e} = -117800\text{e}$.

3. Loput keskimääräisistä eroista elinkaarivaikutuksissa aiheutuvat eläkkeisiin ja eläkemaksuihin liittyvistä vaikutuksista, jotka lasketaan, kuten on esitetty luvussa 6. Irakissa syntyneiden maahanmuuttajien lasten karttunut ero eläkevaikutuksista verrattuna kantaväestöön on noin 51 200 euroa, ja Somaliassa syntyneiden maahanmuuttajien lapsilla tämä ero on noin 60 000 euroa.

Lasten vaikutukset yhteensä per nainen. Keskimääräisistä lasten määristä naista kohti ja keskimääräisistä elinkaarivaikutuksista per yksi maahanmuuttajan lapsi saadaan tulona lasten keskimääräiset julkisen talouden nettovaikutusten ennusteet naista kohti. Irakissa syntyneellä naisella tämä summa on -434 000 euroa, ja Somaliassa syntyneellä naisella tämä summa on puolestaan -674 000 euroa.

9.4.5.1 Maahanmuuttajien lapsia koskevien elinkaariennusteiden vertailu THL:n tuloksiin koskien nuorten syrjäytymisen kustannusten minimielinkaarivaikutuksia: Käytetyt elinkaariennusteet ovat hyvin lähellä THL:n tuloksia.

THL:n minimiarvio vain peruskoulun käyneiden elinkaarivaikutuksista. THL julkaisi alkuvuodesta 2018 rekisteritietoihin perustuvan tutkimuksen syrjäytyneiden nuorten kustannuksista yhteiskunnalle (julkiselle taloudelle) koko heidän elinkaarensa aikana. Tutkimuksessa laskettiin rekisteritietoihin perustuen mm. kuinka paljon enemmän vain peruskoulun käyneet saavat tulonsiirtoja ja kuinka paljon he maksavat vähemmän veroja sellaisiin nuoriin verrattuna, joilla on peruskoulun jälkeinen tutkinto. Tutkimuksen tuloksen mukaan sellainen nuori, jolla ei ole peruskoulun jälkeistä tutkintoa, maksaa julkiselle taloudelle keskimäärin *vähintään* 370 000 euroa (enemmän kuin nuori, jolla tutkinto on). Edelleen tutkimuksen mukaan: ”*Summa [-370 000e] on tutkimuksen mukaan minimiarvio, sillä tarkastelun ulkopuolelle jää vielä lukuisia kustannuksia, kuten rikollisuuden, kulutusverojen ja rahan kierron vaikutukset, työttömien aktivointitoimenpiteet sekä erikoissairaanhoidon ja vanhustenhuollon kustannukset.*”. (THL 2018).

Maahanmuuttajien lasten elinkaariennusteet THL:n minimiarvion pohjalta. Sovelletaan nyt THL:n syrjäytymisen kustannuksia käsittelevän tutkimuksen tuloksia maahanmuuttajien lasten tietoihin 22-vuotiaana ja verrataan näin saatuja yksinkertaisia keskiarvoennusteita yllä taulukossa 9.2 esitettyihin keskimääräisiin Somaliassa ja Irakissa syntyneiden maahanmuuttajien lasten elinkaarivaikutusennusteisiin. Maahanmuuttajien lasten elinkaarivaikutuksia käsittelevän luvun 8 yhteydessä on esitetty tietoja maahanmuuttajien lapsista 22-vuotiaina vuosina 2005–2014. Niistä Irakissa syntyneiden maahanmuuttajien lapsista, jotka ovat joko syntyneet Suomessa tai muuttaneet Suomeen alle 8-vuotiaina, 53 %:lla on vähintään 2. asteen tutkinto suoritettuna (eli 47 %:lla on korkeintaan peruskoulututkinto) (ks. taulukko 8.1, s. 372). Somaliassa syntyneiden maahanmuuttajien 22-vuotiailla lapsilla, jotka ovat syntyneet joko Suomessa tai muuttaneet Suomeen alle 8-vuotiaina, 2. asteen tutkinto on suoritettuna 35 %:lla, eli 65 %:lla on vain korkeintaan peruskoulututkinto (ks. taulukko 8.2, s. 372). Sama osuus 22-vuotiailla Suomessa syntyneen kantaväestön henkilöillä on 84 %:ia vuosina 2005–2011.

THL:n viitatun tutkimuksen (THL (2018)) mukaan nuoren, jolla ei ole 2. asteen tutkintoa, kustannus elinkaaren aika on vähintään 362 000 euroa vuoden 2015 euroissa (370 000 v. 2018 rahassa). Ero Irakissa syntyneiden maahanmuuttajien lapsissa, joilla ei ole 2. asteen tutkintoa 22-vuotiaana, verrattuna kantaväestöön on $0{,}84 - 0{,}53 = 0{,}31$. Sama ero Somaliassa syntyneiden maahanmuuttajien lapsissa 22-vuotiaana on $0{,}84 - 0{,}25 = 0{,}49$.

Näin ollen Irakissa syntyneiden maahanmuuttajien lapsista (jotka ovat syntyneet Suomeen tai muuttaneet Suomeen alle 8-vuotiaina) aiheutuu kantaväestöön verrattuna keskimäärin vähintään $\mathbf{0{,}31 \times -362000e = -112200e}$ suuruinen vaikutus elinkaarien yli. Somaliassa syntyneiden maahanmuuttajien vastaavilla lapsilla tämä keskimääräinen summa on $\mathbf{0{,}49 \times -362000e = -177380e}$ elinkaarien yli.

THL:n minimiarvioista johdetut elinkaariennusteet ovat lähellä tämän tutkimuksen elinkaariennusteita Irakissa ja Somaliassa syntyneiden maahanmuuttajien lapsille. Taulukon 9.2 tulosten mu-

kaan yhden Irakissa syntyneen maahanmuuttajan lapsen keskimääräinen elinkaarivaikutusten ennuste on -170 000e verrattuna kantaväestöön, ja Somaliassa syntyneiden lapsilla vastaava luku on -191 000e. Mikäli näistä luvuista otetaan pois peruskouluvaikutukset, niin summat ovat -157 000 euroa ja -177 000 euroa. THL:n tutkimuksen pohjalta lasketut minimiarviot ovat puolestaan -112 200 euroa ja -177 380. Toisin sanoen elinkaariennusteissa käytetyt ennusteet maahanmuuttajien lasten elinkaarivaikutuksille tuottavat varsinkin Somaliassa syntyneiden osalta hyvin samanlaisia tuloksia, kuin jos olisi käytetty THL:n käyttämiä minimiarvioita syrjäytymisen kustannuksille elinkaarien yli.

9.4.6 Suomessa syntyneiden elinkaariennusteiden odotusarvoksi nolla euroa

Olen laatinut elinkaariennusteet niin, että Suomessa syntyvien elinkaariennusteiden odotusarvoksi tulee tasan nolla euroa. Tämä johtuu seuraavista syistä:

- Suomessa syntyneen kantaväestön elinkaariennusteiden odotusarvo riippuu siitä, mitä julkisen talouden vaikutuksia elinkaariennusteisiin sisältyy. Esimerkiksi luvussa 10 esitellään joitakin nyt esitettävistä elinkaariennusteista puuttuvia vaikutuksia, joihin lukeutuvat esimerkiksi jotkin karttumaltaan pienet verot. Mikäli ne olisivat mukana elinkaariennusteissa, Suomessa syntyneen kantaväestön elinkaariennusteet muuttuisivat joitakin kymmeniä tuhansia euroja. Täten yksi syy kantaväestön elinkaariennusteiden "nollaukselle"on, että tällöin ei kiinnitetä liikaa huomiota siihen, mikä sattuu olemaan kantaväestön elinkaariennusteiden odotusarvo.
- Mikäli julkisen talouden oletetaan olevan tasapainossa "pitkällä aikavälillä", tähän päädytään siten, että kantaväestön elinkaariennusteiden odotusarvo on nolla euroa – mikäli *kaikki* julkisen talouden vaikutukset jaetaan henkilöiden kesken.
- Kun kantaväestön elinkaariennusteiden odotusarvo on nolla euroa, maahanmuuttajaväestön elinkaarivaikutuksia on helpompi verrata kantaväestöön.

Suomessa syntyneiden diskonttaamattomien elinkaariennusteiden odotusarvo ilman mainittua nollausta on 84 247 euroa. Koska Suomessa syntyvien elinajanodote syntyessä on käytettyjen kuolleisuuslukujen perusteella 80,7 vuotta, odotusarvon nollaamiseksi jokaisen ikävuoden julkisen talouden nettovaikutusennusteesta vähennetään 1 044 euroa. Diskontattujen elinkaariennusteiden yhteydessä tämä luku on melkein sama, 1 041 euroa. Nämä vuosittaiset vähennykset tehdään kaikkien ennustepopulaatioiden kaikkien henkilöiden kaikille vuosittaisille ennusteille. Koska Suomessa syntyvä väestö on maassa keskimäärin pisimpään (syntymästä kuolemaan), näiden vähennysten summa on suurempi Suomessa syntyneillä kuin Irakissa ja Somaliassa syntyneillä maahanmuuttajilla.

9.4.7 Parannettavaa elinkaariennusteissa

Elinkaariennusteissa riittää (aina) parannettavaa. Olen listannut tähän keskeisimmät parannusasiat:

- Irakissa ja Somaliassa syntyneiden palkkasummamalleihin olisi lisättävä henkilötason satunnaiset vakiotermit. Nyt esitetyissä tuloksissa ei oteta huomioon henkilöiden välistä vaihtelua realistisen paljon. Lisäksi kaikkiin palkkasummamalleihin olisi lisättävä vielä autoregressiiviset termit.
- Nuorten ja työikäisten nettovaikutusmallit voisi ehdollistaa palkkasummilla. Eli antaa palkkasummien vaikuttaa nettovaikutuksiin. Nyt kummatkin niistä on mallinnettu erikseen.
- Kuolemisennusteet voisi ehdollistaa palkkasummilla ja työiän nettovaikutuksilla, jotta kuolemistodennäköisyydet olisivat ehdollisia palkoilla ja työiän nettovaikutuksilla.
- Sama koskee myös yli 62-vuotiaiden terveys- ja sosiaalipalveluiden käyttöä.

9.4.8 Elinkaariennusteiden komponenttien keskimääräiset kontribuutiot elinkaariennusteisiin: esimerkki

Ennen varsinaisia elinkaariennusteiden tuloksia esitän taulukossa 9.3 esimerkin siitä, kuinka elinkaariennusteiden odotusarvot muodostuvat eri elinkaarivaikutusten komponenteista. Esimerkkinä on Suomessa syntyvien elinkaariennusteiden odotusarvo (0 euroa), Somaliassa syntyneiden ja koko iän Suomessa olevien odotusarvo (-776 000 euroa) sekä Somaliassa syntyneiden, Suomeen 20–24-vuotiaana muuttavien ja koko iän Suomessa olevien odotusarvo (-951 300 euroa). ”Eläkevaikutusten” komponentti sisältää muun muassa työiässä maksettavat työnantajien sosiaalivakuutusmaksut (ks. ”eläkevaikutusten” osat edellä luvussa 9.4.3.2).

Taulukosta 9.3 huomataan, että Suomessa syntyville aiheutuu negatiivisia julkisen talouden nettovaikutuksia 0–15-vuotiaina, jotka eivät koske yli 15-vuotiaana Suomeen muuttavia Somaliassa syntyneitä. Myös 15–19-vuotiaan Suomessa syntyville karttuu negatiivisia julkisen talouden vaikutuksia, joita ei kartu myöskään 20–24-vuotiaana Suomeen muuttaville Somaliassa syntyneille.

Vaikka Suomessa syntyneille karttuu 0–19-vuotiaana sellaisia negatiivisia julkisen talouden nettovaikutuksia, joita Somaliassa syntyneillä ei ole, Somaliassa syntyneiden elinkaariennusteiden odotusarvot ovat silti yli -750 000 euroa verrattuna Suomessa syntyneisiin. Tämä johtuu lähinnä siitä, että Somaliassa syntyneiden ennustetut työiän nettovaikutukset ovat hyvin negatiiviset ja samalla ennustetut palkkasummat jäävät pieniksi, mikä tekee myös ”eläkevaikutuksista” hyvin negatiiviset verrattuna Suomessa syntyneeseen kantaväestöön.

Taulukko 9.3: *Elinkaariennusteiden odotusarvojen muodostuminen elinkaariennusteiden komponenttien euromääräisistä odotusarvoista. Esimerkkinä: Suomessa syntyvien elinkaariennusteiden odotusarvo (0 euroa), Somaliassa syntyneiden ja koko iän Suomessa olevien odotusarvo (-776 000 euroa) sekä Somaliassa syntyneiden, Suomeen 20–24-vuotiaana muuttavien ja koko iän Suomessa olevien odotusarvo (-951 300 euroa). ”Eläkevaikutusten” komponentti sisältää mm. työiässä maksettavat työnantajien sosiaalivakuutusmaksut (ks. ”eläkevaikutusten” osat edellä luvussa 9.4.3.2). ”Nollasiirto” on esitetty edellä luvussa 9.4.6, ja se tehdään, jotta Suomessa syntyneiden elinkaarivaikutusennusteiden odotusarvo olisi tasan nolla euroa.*

Komponentti	Suomi	Somalia (kaikki)	Somalia (20–24-v. muuttaneet)
Alle 15-vuotiaat	-116 400	0	0
Nuoret (15–19-vuotiaat)	-36 800	-6 100	0
Työikäiset (20–62-vuotiaat)	201 000	-446 600	-633 300
"Eläkevaikutukset"	238 000	-47 900	-33 900
Yli 62-vuotiaat	-201 100	-175 900	-176 600
Vastaanotto ja kotouttaminen	0	-46 700	-47 300
"Nollasiirto"	-84 200	-53 500	-60 200
Odotusarvo yhteensä	0	-776 600	-951 300

9.5 Uusien maahanmuuttajien (diskonttaamattomat) elinkaariennusteet

9.5.1 Somalia

Ennustetulokset eri suomessaoloaikojen mukaan. Esitän seuraavaksi varsinaiset elinkaariennustetulokset Somaliassa syntyneelle ennusteväestölle. Esitän samat ennustetulokset neljään kertaan siten, että vain taustalla oleva oletus ennustetusta suomesssaoloajasta vaihtuu niissä. Ensimmäiseksi esitän elinkaariennusteet alaluvussa 9.5.1.1 niin, että uusien maahanmuuttajien oletetaan jäävän koko loppuiäkseen Suomeen maahanmuutosta asti. Tämän jälkeen esitän elinkaariennusteet siten, että elinkaariennusteissa huomioidaan mahdolliset poismuutot sekä takaisinmuutot Suomeen (poismuuton jälkeen). Suomesta poismuuttojen ja takaisinmuuttojen ennustaminen on esitetty tämän tutkimuksen luvussa 3. Koska ennustetulosten rakenne on samanlainen kaikissa neljässä ennustetulosten luokassa, kirjoitan ennustetulosten tulkinnasta vain niiden ennustetulosten osalta, jotka koskevat koko loppuiäkseen Suomeen muuttavia.

2000-luvulla Pohjoismaihin muuttaneista pakolaisista suurin osa on jäänyt muuttomaahan. Koska Irakissa ja Somaliassa syntyneiden havaitut suomessaoloajat vaihtelevat paljon maahanmuuttovuoden luokan mukaan – siten, että myöhemmin muuttaneet jäävät suuremmalla todennäköisyydellä pitemmäksi aikaa ja myös pysyvästi Suomeen kuin aiemmin muuttaneet (ks. Irakia ja Somaliaa koskevat tulokset luvuissa 3.2.7 ja 3.3.2) – esitän elinkaariennustetulokset kolmen eri maahanmuuttovuoden luokan mukaan. Nämä maahanmuuttovuoden luokat ovat: vuosina 2000–2009 Suomeen ensimmäisen kerran muuttaneet, vuosina 1994–1999 Suomeen ensimmäisen kerran muuttaneet sekä ennen vuotta 1994 Suomeen ensimmäisen kerran muuttaneet. Näissä elinkaariennusteissa Suomesta poismuuton ja Suomeen takaisinmuuton ennusteet perustuvat mainittuja maahanmuuttoluokkia koskeviin estimaatteihin (ks. tarkemmin mallinnustulokset luvuissa 3.2.7 ja 3.3.2). Se, että varsinkin 1990-luvun jälkeen muuttaneet pakolaisista suurin osa on jäänyt maahan, on ilmiö, joka on havaittu myös Norjassa ja Tanskassa (ks. luku 13.4.1).

Elinkaariennusteet lasten vaikutusten kanssa. Esitän kaikki elinkaariennustetulokset sekä ilman maahanmuuttajien lasten ennustettuja vaikutuksia sekä niin, että lasten vaikutusten ennusteet ovat sisältyvät mukaan elinkaariennusteisiin. Käytän lasten ennustetuista elinkaarivaikutuksista lyhennettä ”l.v.”, joten ennusteet, joissa lasten vaikutukset on mukaan luettu saavat lyhenteen ”ml. LV” (mukaan luettu lasten vaikutukset).

Elinkaariennusteet koko väestölle yhteensä. Esitän elinkaariennusteet ensiksi erittelemättä koko ennustepopulaatiolle. Elinkaariennusteiden todennäköisyysjakauma muodostuu tällöin kaikkien ennustepopulaation henkilöiden kaikista elinkaariennustesimulaatioista. Elinkaariennusteiden erittelemätön yhteistodennäköisyysjakauma koostuu ennustesimulaatioista, joiden lukumäärä on 300 (simulaatiomäärä) $\times$ ennustepopulaation henkilöiden lukumäärä.

Elinkaariennusteet eritellysti maahanmuuttovuoden tietojen perusteella. Sen jälkeen, kun olen esittänyt elinkaariennusteet erittelemättä koko ennustepopulaatiolle, esitän elinkaariennusteet eritellysti

sen mukaan, mitä ennustepopulaation henkilöistä tiedetään, kun he muuttavat ensimmäisen kerran Suomeen. Nämä elinkaariennusteiden ”selittävät muuttujat”, joiden osalta esitän ennusteet eritellysti, ovat: sukupuoli, maahanmuuton ikäluokka, henkilöiden pääsiallinen toiminta Suomessa maahanmuuttovuonna sekä henkilöiden koulutustaso maahanmuuttovuonna.

Ensimmäisen maahanmuuton ikäluokat, joille esitän elinkaariennusteet, ovat: 16–19(-vuotiaana muuttavat), 20–24, 25–29, 30–34, 35–39, 40–49 sekä 50–62. Kaksi vanhinta ikäluokkaa käsittää enemmän ikävuosia, koska näissä maahanmuuton ikäluokissa on ennustepopulaatiossa vähemmän henkilöitä kuin nuoremmissa ikäluokissa. Maahanmuuttovuoden pääasiallisen toiminnan luokkia ovat puolestaan: työllinen, opiskelija, työtön ja muu. Luokka ”muu” käsittää kaikki henkilöt, jotka eivät olleet muuttovuonna työvoimassa ja jotka eivät olleet opiskelijoita. Muuttovuoden koulutustasossa on kaksi luokkaa: henkilöllä on muuttovuonna vähintään 2. asteen koulutus, henkilöllä ei ole vähintään 2. asteen koulutusta. Henkilön koulutustaso muuttovuonna sisältyy jälkimmäiseen luokkaan myös silloin, kun koulutustasotieto puuttuu.

Muuttovuoden tietojen perusteella eritellyt ennusteet eivät ole marginaalivaikutusten ennusteita. Nämä selittävien muuttujien mukaan eritellyt elinkaariennusteet eivät ole selittävien muuttujien marginaalivaikutusten ennusteita. Selittävien muuttujien marginaalivaikutusten elinkaariennusteet laadittaisiin siten, että pidettäisiin muut muuttovuoden tiedot – eli elinkaariennusteiden selittävät muuttujat – vakioarvoisina (esimerkiksi ennustepopulaation odotusarvoissa) ja annettaisiin vain sen selittävän muuttujan arvon vaihdella, jonka marginaalivaikutusta elinkaariennusteisiin haluttaisiin tutkia. Marginaalivaikutusten ennusteet eivät kuitenkaan ole selittävän muuttujan mukaan eriteltyjä ennusteita, koska niissä muut selittävät muuttujat on keinotekoisesti kiinnitetty vakioarvoiksi (odotusarvoihin ennustepopulaatiossa). Tämän seurauksena marginaaliennusteissa on pienempi keskihajonta kuin varsinaisissa selittävien muuttujien mukaan eritellyissä ennusteissa.

9.5.1.1 Koko ikä maahanmuutosta kuolemaan asti Suomessa

Esitän tässä luvussa elinkaariennusteet Somaliassa syntynyttä ennustepopulaatiota koskien niin, että oletan kaikkien ennustepopulaation henkilöiden jäävän Suomeen koko loppuiäkseen heidän ensimmäisestä Suomeen muutosta lähtien. Seuraavissa luvuissa esitän ennusteet samalla väestölle niin, että ennustetut Suomesta poismuutot ja paluumuutot otetaan huomioon elinkaariennusteissa.

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Alla taulukossa 9.4 on esitetty Somaliassa syntyneen ennustepopulaation elinkaariennusteiden yhteistodennäköisyysjakaumaa koskevia tuloksia. Elinkaariennustetulokset on esitetty ilman uusien maahanmuuttajien ennustettuja lasten elinkaarivaikutuksia (vasemmanpuoleinen sarake, ”Ilman LV”) sekä niin, että ennusteisiin sisältyvät ennustetut lasten elinkaarivaikutukset (oikeanpuoleinen sarake, ”ml. LV”).

Somaliassa syntyneiden ennustepopulaatio käsittää 490 henkilöä, niin kuin aiemmin on esitetty luvussa 9.2. Taulukossa 9.4 ennustepopulaation yhteistä ennustejakaumaa koskien on esitetty jakauman tunnus-

luvuista: odotusarvo, keskihajonta sekä esitetyt jakauman prosenttipisteet. MC-keskivirhe on jakauman odotusarvon Monte Carlo -keskivirhe, joka on ennustejakauman simuloinnista johtuvan virheen suuruus. Ennusteväestöä lukuun ottamatta kaikki muut ennustejakaumaa koskevat luvut ovat taulukossa 9.4 ilmaistu tuhansina euroina (nämä luvut on pyöristetty lähimpään tuhanteen euroon). Näin ollen esimerkiksi kaikkia ennustepopulaation henkilöitä koskevan ennustejakauman odotusarvo on ilman lasten vaikutuksia -777 000 euroa, ja keskihajonnan estimaatti on puolestaan 722 000 euroa. Odotusarvon simuloinnista aiheutuva MC-keskivirhe on puolestaan suuruudeltaan 3 000 euroa.

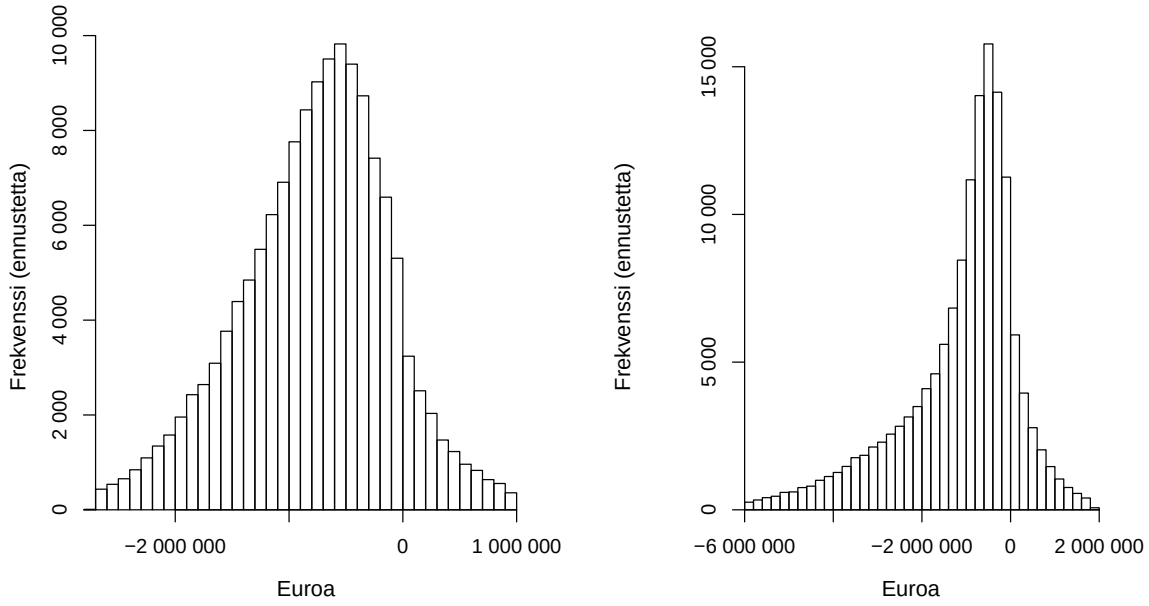
Somaliassa ja Irakissa syntyneiden elinkaariennusteiden jakaumilla ei ole keskihajontaa. Huomautan, että teknisesti ottaen Irakissa ja Somaliassa syntyneiden elinkaariennusteiden jakaumilla ei ole keskihajontaa – vaikka niille voidaan laskea tietysti aina esitetyt estimaatit elinkaariennusteiden simulaatioista. Tämä johtuu siitä, että Somaliassa ja Irakissa syntyvien nuorten ja työikäisten nettovaikutusten otosjakaumana on vino Studentin t-jakauma, jonka vapausasteet ovat alle arvoltaan alle 4. Tällä todennäköisyysjakaumalla ei ole tällöin keskihajontaa. Suomessa syntyneiden otosjakaumalla keskihajonta puolestaan on, koska käytän työikäisten nettovaikutusten ennustamisessa rajoitettua vinoa Studentin t-jakaumaa. Tämä keskihajontoja koskeva huomautus päätee myös tutkimuksen viimeisessä luvussa esitettäviin diskontattuihin elinkaariennusteisiin.

Lasten vaikutukset huomioivan elinkaariennusteiden yhteisjakauman odotusarvo on -1 118 000 euroa, ja keskihajonnan estimaatti on 1 663 000 euroa. Odotusarvo on negatiivisempi, kun lasten vaikutukset huomioidaan verrattuna ennusteisiin, joissa lasten vaikutuksia ei huomioida, koska Somaliassa syntyneiden maahanmuuttajien lasten elinkaariennusteet ovat keskimäärin negatiivsempia kuin Suomessa syntyneen kantaväestön henkilön keskimääräinen elinkaariennuste – 22-vuotiaita maahanmuuttajien lapsia ja kantaväestön samanikäisiä henkilöitä koskevien tietojen perusteella (ks. luku 8.2). Keskihajonta on lasten vaikutukset huomioivissa elinkaariennusteessa moninkertainen verrattuna ennusteisiin, joissa lasten vaikutuksia ei ole huomioitu. Tämä johtuu siitä, että lasten vaikutukset huomioivissa elinkaariennusteissa on kyse käytännössä useamman kuin yhden henkilön yhteenlasketuista elinkaariennusteista. Lasten vaikutukset sisältävissä ennusteissa lasketaan yhteen sekä henkilön oma elinkaariennuste sekä ennustettujen syntyvien lasten ennustetut elinkaarivaikutukset.

Elinkaariennusteiden mediaanit ovat -715 000 euroa (ilman lasten vaikutuksia) sekä -766 000 euroa (ml. lasten vaikutukset). Koska mediaanit ovat odotusarvoja positiivisemmat, tämä viittaa siihen, että Somaliassa syntyneen ennusteväestön kaikkien elinkaariennusteiden yhteisjakauma on vasemmalle vino. Taulukon 9.4 tulosten perusteella hieman yli 10 %:ia elinkaariennusteista on positiivisia, kun lasten vaikutuksia ei huomioida. Tämä osuus on suurempi – mutta kuitenkin alle 25 %:ia – kun lasten vaikutukset huomioidaan. Kun lasten vaikutuksia ei huomioida, noin 10 %:ia elinkaariennusteista on negatiivisempia kuin -1,7 miljoonaa euroa, ja kun lasten ennustetut elinkaarivaikutukset huomioidaan, noin 10 %:ia elinkaariennusteista on negatiivisempia kuin -3,1 miljoonaa euroa.

Taulukko 9.4: *Somaliassa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-777	-1 118
Keskihajonta	722	1 663
MC-keskivirhe	3	11
2,5:s prosenttipiste (2,5 %)	-2 329	-4 947
10. prosenttipiste (10 %)	-1 712	-3 146
Alakvartiili (25 %)	-1 197	-1 705
Mediaani (50 %)	-715	-766
Yläkvartiili (75 %)	-325	-278
90. prosenttipiste (90 %)	16	193
97,5:s prosenttipiste (97,5 %)	620	1 104



Kuva 9.2: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Taulukko 9.5: *Somaliassa syntyneen ennusteväestön elinkaariennusteet. Nuorten ja työikäisten nettovaikutusten ennustemallissa maassaoloajan vaikutus rajoitettu vain 21 vuoteen (toisin kuin muissa Somaliaa koskevissa tuloksissa). Eli 21 maassaolovuoden jälkeen maassaoloajalla sama vaikutus kuin 21 maassaolovuoden jälkeen. Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ennusteväestö (N)	490
Odotusarvo	-793
Keskihajonta	617
MC-keskivirhe	3
2,5:s prosenttipiste (2,5 %)	-2 137
10. prosenttipiste (10 %)	-1 608
Alakvartiili (25 %)	-1 168
Mediaani (50 %)	-733
Yläkvartiili (75 %)	-373
90. prosenttipiste (90 %)	-86
97,5:s prosenttipiste (97,5 %)	312

Elinkaariennusteet sukupuolen mukaan

Taulukossa 9.6 on esitetty Somaliassa syntynyttä ennusteväestöä koskevia elinkaariennusteiden tuloksia sukupuolen mukaan. Ennusteväestö muodostuu 248 naisesta ja 242 miehestä. Naisia koskevat tulokset ovat kaikkien naisten elinkaariennusteiden yhteistodennäköisyysjakaumaa koskevia jakauman tunnuslukuja. Samaa koskee miehiä. Taulukon kahdella ensimmäisellä rivillä tulokset on esitety, kun lasten vaikutuksia ei huomioida. Kahdella alemmalla rivillä lasten vaikutukset on huomioitu elinkaariennusteissa. Miehiä koskevat tulokset eivät kuitenkaan muutu, kun lasten vaikutukset huomioidaan. Tämä johtuu siitä, että olen liittänyt kaikki lasten ennustettujen vaikutusten tulokset vain heidän äiteihinsä. Päädyin tähän ratkaisuun, koska hedelmällisyysluvut ilmoitetaan yleensä vain naisia koskien.

Somaliassa syntyneen ennusteväestön naisten elinkaariennusteiden odotusarvo on -1 050 000 euroa, mediaani on -987 000 euroa ja keskihajonta on 706 000 euroa – kun lasten vaikutuksia ei huomioida. Kun lasten vaikutukset huomioidaan elinkaariennusteissa, vastaavat jakauman tunnusluvut ovat -1 725 000 euroa (odotusarvo), -1 552 000 euroa (mediaani) ja 2 083 000 euroa (keskihajonta). Miesten elinkaariennusteiden odotusarvo on puolestaan -496 000 euroa, mediaani -482 000 euroa ja keskihajonta 625 000 euroa. Somaliassa syntyneiden miesten elinkaariennusteet ovat näin ollen keskimäärin satoja tuhansia euroja positiivisempia kuin naisten, vaikka lasten vaikutuksia elinkaariennusteissa ei huomoitaisikaan.

Taulukko 9.6: *Somaliassa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-1 050	706	-1 971	-1 482	-987	-574	-251
Miehet	242	-496	625	-1 257	-851	-482	-146	216
Naiset (ml. LV)	248	-1 725	2 083	-4 065	-2 788	-1 552	-592	133
Miehet (ml. LV)	242	-496	625	-1 257	-851	-482	-146	216

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukossa 9.7 on esitetty Somaliassa syntynyttä ennusteväestöä koskevia elinkaariennusteiden keskeisimmät tulokset maahanmuuton ikäluokan mukaan. Koko ennusteväestö on jaettu seitsemään ikäluokkaan sen mukaan, minkä ikäisenä he muuttavat ensimmäisen kerran Suomeen (ennusteväestö koostuu vain 16–62-vuotiaana ensimmäisen kerran Suomeen muuttavista). Vanhimpiin ikäluokkiin sisältyy enemmän ikävuosia, koska Somaliassa syntyneistä harvempi muuttaa (on muuttanut) Suomeen ensimmäisen kerran näissä ikäluokissa. Koko ennusteväestön 490 Somaliassa syntyneestä henkilöstä 142 henkilöä muuttaa Suomeen, kun he ovat 16–19-vuotiaita. Toiseksi nuorimpaan ikäluokkaan, eli 20–24-vuotiaana ensimmäisen kerran Suomeen muuttaviin, kuuluu ennusteväestössä 81 henkilöä. Vanhimpaan ikäluokkaan, eli 50–62-vuotiaana muuttaviin, kuuluu 33 henkilöä.

Taulukon 9.7 ylemmässä puolikkaassa esitetään elinkaariennustetulokset aluksi ikäluokittain niin, ettei ennusteisiin sisälly lasten vaikutuksia. Taulukon alemmassa puolikkaassa maahanmuuttoikäluokittaiset elinkaariennusteet on esitetty puolestaan siten, että lasten vaikutukset ovat mukana ("ml. LV").

Negatiivisin elinkaariennusteiden odotusarvo on 20–24-vuotiaana Suomeen muuttavilla. Ilman lasten vaikutuksia tämän ikäluokan henkilöiden elinkaariennusteiden odotusarvo on -951 000 euroa, ja lasten vaikutukset mukaan lukien tämän ryhmän henkilöiden elinkaariennusteiden odotusarvo on -1 343 000 euroa. Alle 20-vuotiaana Suomeen muuttaneille elinkaarien pituudet Suomessa ovat odotusarvoisesti hieman pidempiä kuin 20–24-vuotiaana muuttavilla. Alle 20-vuotiaana muuttavilla elinkaariennusteiden odotusarvo on kuitenkin positiivisempi kuin 20–24-vuotiaana muuttavilla. Tämä johtuu siitä, että nettovaikutus- ja palkkasummamallien mukaan alle 20-vuotiaana muuttavien ennustetut julkisen talouden vaikutukset ovat sekä työikäisten nettovaikutusten että ennustetuista palkkasummista johdettavien vaikutusten mukaan positiivisempia kuin myöhemmin muuttaneilla. Ehkä tämä johtuu siitä, että alle 20-vuotiaana muuttavat sopeutuvat paremmin Suomeen kuin vanhemmissa ikäluokissa muuttavat. Tämän asian selittäminen ei kuitenkaan kuulun tämän tutkimuksen kysymyksenasettelun piiriin.

Vanhemmilla ikäluokilla elinkaariennusteiden odotusarvot ovat vähemmän negatiiviset kuin nuorempana muuttavilla. Tämä johtuu siitä, että vanhempana muuttavilla elinkaaret ovat odotusarvoisesti lyhyempiä kuin nuorempina Suomeen muuttavilla. Kaikkien vanhimpien muuttoikäluokkien elinkaariennusteiden

odotusarvot muuttavat suhteessa vähemmän kuin nuorempana muuttavien, kun lasten vaikutukset huomioidaan ennusteissa. Tämä johtuu siitä, että vanhempana muuttavien mukana muuttaa vähemmän alle 15-vuotiaita lapsia Suomeen kuin nuorempana muuttavien mukana. Lisäksi yli 40-vuotiaana muuttavien ennustettu Suomessa saatavien lapsien määrä on pieni.

Taulukon 9.7 mukaan elinkaariennusteiden keskihajonnat ovat sitä suurempia, mitä nuoremmalla iällä henkilö muuttaa Suomeen ensimmäisen kerran – kun lasten vaikutuksia ei huomioida. Kun 16–19-vuotiaana muuttavilla elinkaariennusteiden keskihajonta on 936 000 euroa, niin 50–62-vuotiaana muuttavilla ennusteiden keskihajonta on 286 000 euroa.

Taulukko 9.7: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-737	936	-1 947	-1 362	-708	-102	428
20–24	81	-951	801	-1 982	-1 487	-930	-393	40
25–29	85	-874	665	-1 742	-1 312	-850	-399	-53
30–34	64	-807	533	-1 492	-1 146	-784	-422	-141
35–39	34	-638	453	-1 220	-924	-611	-306	-84
40–49	51	-694	352	-1 132	-886	-664	-457	-274
50–62	33	-482	286	-824	-617	-445	-290	-161
16–19 (ml. LV)	142	-1 168	1 891	-3 521	-2 097	-874	-56	616
20–24 (ml. LV)	81	-1 343	1 769	-3 518	-2 148	-1 052	-324	221
25–29 (ml. LV)	85	-1 208	1 683	-3 187	-1 808	-896	-325	98
30–34 (ml. LV)	64	-1 118	1 560	-2 885	-1 625	-826	-343	2
35–39 (ml. LV)	34	-1 053	1 615	-3 110	-1 269	-618	-242	21
40–49 (ml. LV)	51	-926	1 292	-2 176	-1 101	-696	-424	-167
50–62 (ml. LV)	33	-482	286	-824	-617	-445	-290	-161

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukossa 13.5 on esitetty Somaliassa syntynyttä ennusteväestöä koskevien elinkaariennusteiden keskeisimmät tulokset maahanmuuttovuoden koulutustason mukaan. Koko ennusteväestö on jaettu kahteen luokaan sen mukaan, mikä henkilön koulutustaso on ensimmäisen maahanmuuton vuonna. Koulutustason kaksi luokkaan ovat: henkilöllä on muuttovuonna vähintään toiseen asteen koulutus, henkilöllä ei ole vähintään toisen asteen koulutusta muuttovuonna. Henkilön koulutustaso sijoittuu jälkimmäiseen luokkaan, mikäli henkilön koulutustaso ei ole tiedossa.

Somaliassa syntyneen koko ennusteväestön 490 henkilöstä 32:lla henkilöllä (7 %:lla) on vähintään toiseen asteen koulutus Suomeen muuttaessa. Lopuilla 458 henkilöllä (93 %:lla) ei ole toisen asteen koulutusta

muuttohetkellä, tai heidän koulutustasonsa ei ole tiedossa. Esitän koulutustason mukaisista elinkaariennusteiden jakaumatunnusluvuista taulukossa 13.5 vain odotusarvon, keskihajonnan ja mediaanin. Tulokset on jälleen ilmaistu sekä ilman lasten vaikutusten ennusteita (taulukon 13.5 ylemmät rivit) sekä lasten vaikutusten ennusteet mukaan lukien (taulukon 13.5 alemmat rivit, ”ml. LV”).

Taulukon13.5 tulosten perusteella elinkaariennusteiden odotusarvot ovat noin 70 000 euroa positiivisemmat niillä ennusteväestön henkilöillä, joilla on muuttaessa 2. asteen koulutus, kuin niillä ennusteväestön henkilöillä, joilla toisen asteen koulutusta ei ole. On kuitenkin muistettava, että ennustetulokset eivät ole marginaaliennusteita, eli ero elinkaariennusteiden odotusarvossa ei ole välttämättä seurausta koulutustason eroista. Voi olla, että ero selittyy esimerkiksi sillä, että 2. asteen koulutuksen omaavista suuremmalla osalla pääasiallinen toiminta oli muuttovuonna työllinen kuin vertailuluokassa (ei 2. asteen koulutusta). Koska odotusarvon ero kasvaa koulutustasoluokkien välillä, kun lasten vaikutukset huomioidaan, tämä viittaa siihen, että ne ennusteväestön henkilöt, joilla on vähintään 2. asteen koulutus muuttaessa, ovat keskimäärin vanhempia muuttovuonna kuin vertailuluokan (ei 2. asteen koulutusta) henkilöt. Tähän ikäeroon muuttaessa viittaavat myös ero elinkaariennusteiden keskihajonnoissa luokittain.

Elinkaariennusteiden mediaanit ovat jälleen kummassakin elinkaariennusteiden luokassa ennusteiden odotusarvoja positiivisempia. Tämä viittaa taas siihen, että elinkaariennusteiden jakaumat ovat vasemmalle vinoja.

Taulukko 9.8: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-711	588	-648
Ei 2. asteen koulutusta/ei tietoa	458	-781	730	-720
Vähintään 2. asteen koulutus (ml. LV)	32	-982	1 463	-662
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-1 128	1 675	-776

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukossa 13.4 on esitetty Somaliassa syntynyttä ennusteväestöä koskevien elinkaariennusteiden keskeisimmät tulokset maahanmuuttovuoden pääasiallisen toiminnan mukaan. Koko ennusteväestö on jaettu neljään luokkaan sen mukaan, mikä henkilön pääasiallinen toiminta on ensimmäisen maahanmuuton vuonna (tämä oletetaan tiedetyksi muuttohetkellä). Pääasiallisen toiminnan luokkia on neljä: työllinen, työtön, opiskelija sekä muu. Viimeiseen luokkaan kuuluvat kaikki ne henkilöt, jotka eivät olleet muuttovuonna työvoimassa ja jotka eivät olleet opiskelijoita. Somaliassa syntyneestä ennusteväestöstä suurimmalla osalla maahanmuuttovuoden pääsiallinen toiminta on ”muu”. Koko ennusteväestön 490 henkilöstä 406 kuuluu tähän luokkaan (83 %:ia). Opiskelijoita maahanmuuttovuonna ennusteväestössä on 48 hen-

kilöä eli noin 10 %:ia koko ennusteväestöstä. Työttömiä on 23 (5 %:ia) ja työllisiä 13 (vajaa 3 %:ia) henkilöä.

Koska pienimpiin luokkiin kuuluu vain vähän yli 10 henkilöä, esitän elinkaariennusteiden jakaumatunnusluvuista vain odotusarvon, keskihajonnan ja mediaanin. Tulokset on jälleen ilmaistu sekä ilman lasten vaikutusten ennusteita (taulukon 13.4 ylemmät rivit) sekä lasten vaikutusten ennusteet mukaan lukien (taulukon 13.4 alemmat rivit, ”ml. LV”).

Elinkaariennusteiden odotusarvot ovat negatiivisimmat niillä ennusteväestön henkilöillä, joiden pääasiallinen toiminta muuttovuonna on joko työtön tai ”muu”. Ilman lasten vaikutuksia näiden luokkien odotusarvot ovat -883 000 euroa ja -795 000. Lasten vaikutukset huomioiden näiden luokkien elinkaariennusteiden odotusarvot ovat noin -1,2 miljoonaa euroa. Niillä, joiden pääasiallinen toiminta muuttovuonna on opiskelijuus, elinkaariennusteiden odotusarvo ilman lasten vaikutuksia on -692 000 euroa, ja kun lasten vaikutukset huomioidaan, tämä odotusarvo on noin -881 000 euroa. Työllisillä elinkaariennusteiden odotusarvot ovat positiivisemmat kuin vertailuryhmissä: ilman lasten vaikutuksia odotusarvo on -314 000 euroa, ja lasten vaikutukset mukaan lukien odotusarvo on -367 000 euroa. Elinkaariennusteiden keskihajonta on pienintä työllisten luokassa. Mediaanit ovat kaikissa luokissa ennusteiden odotusarvoja pienempiä, mikä viittaa siihen, että kaikki ennustejakaumat ovat tässä yhteydessä vasemmalle vinoja.

Taulukko 9.9: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-314	561	-292
Opiskelija	48	-692	707	-622
Työtön	23	-883	621	-816
Muu	406	-795	728	-734
Työllinen (ml. LV)	13	-367	774	-294
Opiskelija (ml. LV)	48	-881	1 430	-626
Työtön (ml. LV)	23	-1 201	1 607	-893
Muu (ml. LV)	406	-1 165	1 704	-803

9.5.1.2 Suomessaoloaika vuosina 2000–2009 Suomeen muuttaneiden mukaan

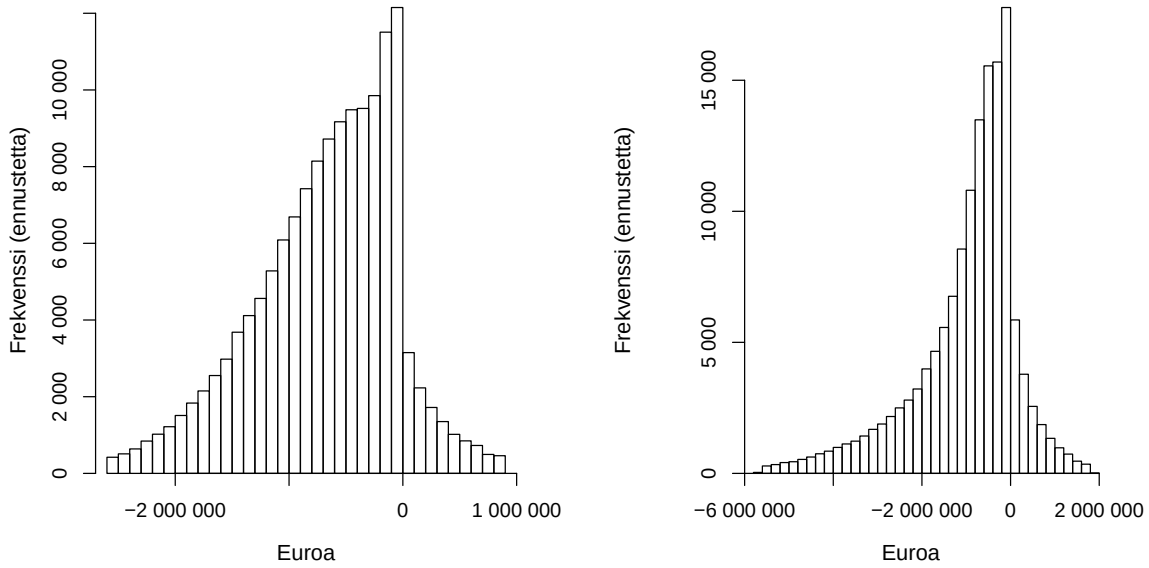
Esitän tässä luvussa elinkaariennusteet Somaliassa syntynyttä ennustepopulaatiota koskien niin, että elinkaariennusteissa huomioidaan ennustetut poismuutot sekä ennustetut paluumuutot Suomeen. Oletan tässä luvussa ennustetun suomessaoloajan (ja ulkomaillaoloajan) noudattavan vuosina 2000–2009 Suomeen muuttaneita koskevia estimaatteja, niin kuin ne on esitetty olotilamallinnuksen yhteydessä (luvuissa 3.2.7 ja 3.3.2). Esitän jäljempänä elinkaariennustetulokset myös niin, että oletan ennustetun suomessaoloajan noudattavan ennen vuotta 2000 Suomeen muuttaneita koskevia olotilamallinnuksen estimaatteja. Yleisesti ottaen tässä luvussa esitetyt elinkaariennusteiden estimaatit ovat negatiivisempia kuin jäljempänä esitettävät. Tämä johtuu siitä, että ennustetut suomessaoloajat ovat keskimäärin pitempiä, kun suomessaoloajan ennusteet perustuvat vuosina 2000–2009 Suomeen muuttaneita koskeviin estimaatteihin (kuin aiemmin muuttaneita koskeviin estimaatteihin).

En esitä tässä luvussa sitä, kuinka lukea elinkaariennusteiden tuloksia, sillä olen tehnyt jo niin yllä olevassa luvussa (9.5.1.1), missä kaikkien ennustepopulaation henkilöiden oletettiin jäävän loppuiäkseen Suomeen ensimmäisestä maahanmuutosta lähtien. Koska tulosten lukeminen on tässä yhteydessä samanlaista kuin yllä esitetyssä luvussa, en toista sitä enää tässä.

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.10: *Somaliassa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonan (KH) sekä prosenttipisteiden* $2,5; 10, 25, 50, 75, 90; 97,5 estimaatit$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-675	-975
Keskihajonta	685	1 527
MC-keskivirhe	4	9
2,5:s prosenttipiste (2,5 %)	-2 220	-4 551
10. prosenttipiste (10 %)	-1 581	-2 739
Alakvartiili (25 %)	-1 065	-1 475
Mediaani (50 %)	-581	-670
Yläkvartiili (75 %)	-196	-194
90. prosenttipiste (90 %)	-8	151
97,5:s prosenttipiste (97,5 %)	536	1 059



Kuva 9.3: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.11: *Somaliassa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-896	701	-1 842	-1 333	-810	-363	-100
Miehet	242	-448	586	-1 176	-779	-397	-100	151
Naiset (ml. LV)	248	-1 725	2 083	-4 065	-2 788	-1 552	-592	133
Miehet (ml. LV)	242	-496	625	-1 257	-851	-482	-146	216

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.12: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-613	856	-1 767	-1 157	-471	-65	320
20–24	81	-806	776	-1 870	-1 327	-714	-194	-1
25–29	85	-765	657	-1 642	-1 203	-692	-245	-37
30–34	64	-722	536	-1 439	-1 073	-678	-293	-83
35–39	34	-576	455	-1 177	-870	-525	-218	-58
40–49	51	-634	370	-1 103	-848	-607	-365	-167
50–62	33	-458	299	-818	-607	-425	-246	-115
16–19 (ml. LV)	142	-983	1 697	-3 040	-1 732	-693	-56	536
20–24 (ml. LV)	81	-1 150	1 634	-3 102	-1 837	-875	-201	193
25–29 (ml. LV)	85	-1 063	1 563	-2 793	-1 587	-797	-229	77
30–34 (ml. LV)	64	-1 005	1 468	-2 609	-1 454	-744	-257	-5
35–39 (ml. LV)	34	-947	1 516	-2 730	-1 166	-564	-186	-5
40–49 (ml. LV)	51	-841	1 232	-1 933	-1 042	-645	-344	-111
50–62 (ml. LV)	33	-458	299	-818	-607	-425	-246	-115

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.13: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-641	568	-574
Ei 2. asteen koulutusta/ei tietoa	458	-677	692	-582
Vähintään 2. asteen koulutus (ml. LV)	32	-885	1 378	-611
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-981	1 536	-676

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.14: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-292	530	-229
Opiskelija	48	-626	673	-540
Työtön	23	-773	605	-686
Muu	406	-687	690	-593
Työllinen (ml. LV)	13	-337	716	-236
Opiskelija (ml. LV)	48	-796	1 335	-568
Työtön (ml. LV)	23	-1 060	1 534	-792
Muu (ml. LV)	406	-1 011	1 560	-698

9.5.1.3 Suomessaoloaika vuosina 1994–1999 Suomeen muuttaneiden mukaan

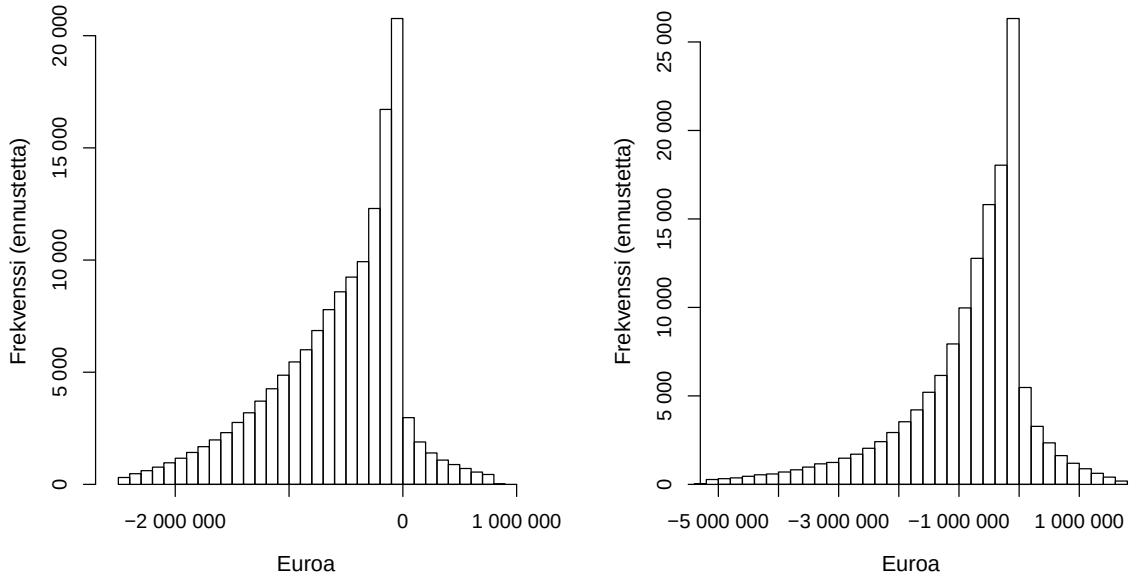
Esitän tässä luvussa elinkaariennusteet Somaliassa syntynyttä ennustepopulaatiota koskien niin, että elinkaariennusteissa huomioidaan ennustetut poismuutot sekä ennustetut paluumuutot Suomeen. Oletan tässä luvussa ennustetun suomessaoloajan (ja ulkomaillaoloajan) noudattavan vuosina 1994–1999 Suomeen muuttaneita koskevia estimaatteja, niin kuin ne on esitetty olotilamallinnuksen yhteydessä (luvuissa 3.2.7 ja 3.3.2). Olen yllä esittänyt elinkaariennustetulokset myös niin, että oletan ennustetun suomessaoloajan noudattavan vuosina 2000–2009 Suomeen muuttaneita koskevia olotilamallinnuksen estimaatteja sekä myös niin, että oletan kaikkien uusien maahanmuuttajien jäävän Suomeen loppuiäkseen ensimmäisen muuton jälkeen. Esitän vielä jäljempänä elinkaariennustetulokset myös niin, että oletan ennustetun suomessaoloajan noudattavan ennen vuotta 1994 Suomeen muuttaneita koskevia olotilamallinnuksen estimaatteja. Yleisesti ottaen tässä luvussa esitetyt elinkaariennusteiden estimaatit ovat negatiivisempia kuin jäljempänä esitettävät (ja positiivisemmat kuin edellä esitetyt). Tämä johtuu siitä, että ennustetut suomessaoloajat ovat keskimäärin pitempiä, kun suomessaoloajan ennusteet perustuvat vuosina 1994–1999 Suomeen muuttaneita koskeviin estimaatteihin kuin ennen vuotta 1994 muuttaneita koskeviin estimaatteihin (edellä esitetyissä ennustetuloksissa ennustetut suomessaoloajat olivat puolestaan keskimäärin pidempiä kuin tässä luvussa esitettävät).

En esitä tässä luvussa sitä, kuinka lukea elinkaariennusteiden tuloksia, sillä olen tehnyt jo niin yllä olevassa luvussa (9.5.1.1), missä kaikkien ennustepopulaation henkilöiden oletettiin jäävän loppuiäkseen Suomeen ensimmäisestä maahanmuutosta lähtien. Koska tulosten lukeminen on tässä yhteydessä samanlaista kuin yllä esitetyssä luvussa, en toista sitä enää tässä.

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.15: *Somaliassa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonan (KH) sekä prosenttipisteiden* $2,5; 10, 25, 50, 75, 90; 97,5 estimaatit$*. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Keskiarvo	-569	-832
Keskihajonta	638	1 393
MC-keskivirhe	4	9
2,5:s prosenttipiste (2,5 %)	-2 096	-4 129
10. prosenttipiste (10 %)	-1 442	-2 357
Alakvartiili (25 %)	-911	-1 259
Mediaani (50 %)	-425	-545
Yläkvartiili (75 %)	-124	-138
90. prosenttipiste (90 %)	-10	86
97,5:s prosenttipiste (97,5 %)	444	983



Kuva 9.4: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histograami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.16: *Somaliassa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-746	678	-1 700	-1 158	-598	-201	-58
Miehet	242	-388	538	-1 077	-671	-289	-75	78
Naiset (ml. LV)	248	-1 725	2 083	-4 065	-2 788	-1 552	-592	133
Miehet (ml. LV)	242	-496	625	-1 257	-851	-482	-146	216

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.17: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-497	774	-1 601	-919	-261	-53	207
20–24	81	-665	735	-1 723	-1 134	-456	-113	-7
25–29	85	-641	626	-1 526	-1 049	-491	-144	-24
30–34	64	-623	527	-1 341	-968	-525	-190	-55
35–39	34	-507	438	-1 106	-788	-425	-151	-41
40–49	51	-562	378	-1 049	-787	-530	-257	-107
50–62	33	-428	300	-794	-586	-390	-201	-86
16–19 (ml. LV)	142	-804	1 511	-2 555	-1 414	-490	-58	435
20–24 (ml. LV)	81	-971	1 505	-2 697	-1 566	-671	-137	129
25–29 (ml. LV)	85	-899	1 459	-2 416	-1 385	-627	-145	48
30–34 (ml. LV)	64	-884	1 359	-2 318	-1 294	-626	-189	-9
35–39 (ml. LV)	34	-843	1 380	-2 349	-1 063	-494	-145	-8
40–49 (ml. LV)	51	-762	1 160	-1 764	-965	-571	-250	-73
50–62 (ml. LV)	33	-428	300	-794	-586	-390	-201	-86

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.18: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-562	551	-460
Ei 2. asteen koulutusta/ei tietoa	458	-570	644	-422
Vähintään 2. asteen koulutus (ml. LV)	32	-792	1 304	-512
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-835	1 399	-547

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.19: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-264	488	-175
Opiskelija	48	-553	633	-427
Työtön	23	-657	583	-523
Muu	406	-576	643	-428
Työllinen (ml. LV)	13	-300	659	-180
Opiskelija (ml. LV)	48	-708	1 226	-476
Työtön (ml. LV)	23	-901	1 390	-642
Muu (ml. LV)	406	-860	1 425	-565

9.5.1.4 Suomessaoloaika ennen vuotta 1994 Suomeen muuttaneiden mukaan

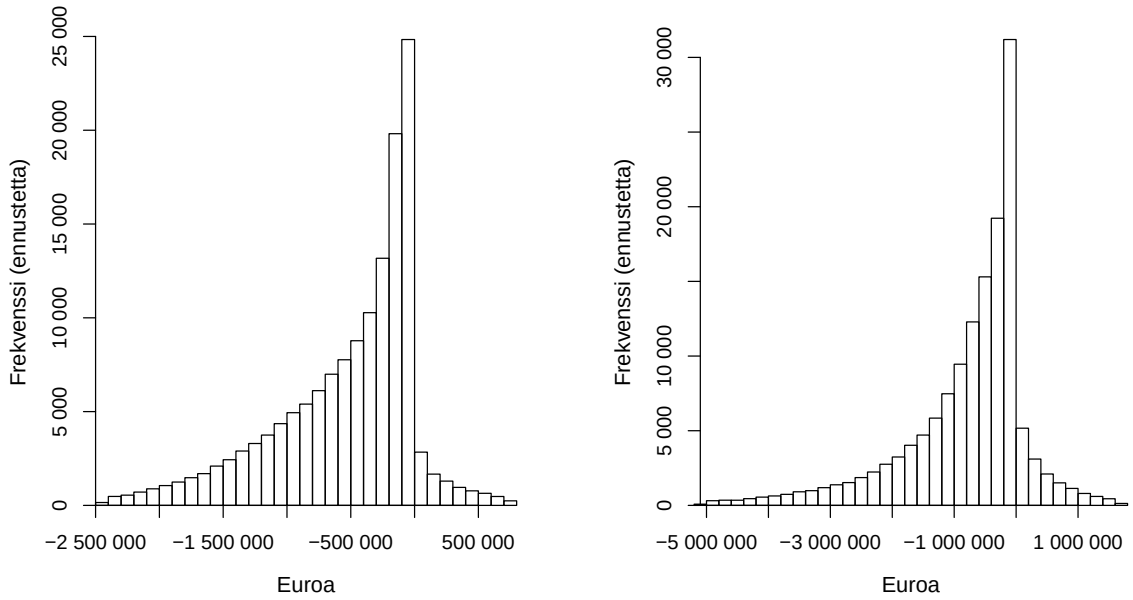
Esitän tässä luvussa elinkaariennusteet Somaliassa syntynyttä ennustepopulaatiota koskien niin, että elinkaariennusteissa huomioidaan ennustetut poismuutot sekä ennustetut paluumuutot Suomeen. Oletan tässä luvussa ennustetun suomessaoloajan (ja ulkomaillaoloajan) noudattavan ennen vuotta 1994 Suomeen muuttaneita koskevia estimaatteja, niin kuin ne on esitetty olotilamallinnuksen yhteydessä (luvuissa 3.2.7 ja 3.3.2). Olen yllä esittänyt elinkaariennustetulokset myös niin, että oletan ennustetun suomessaoloajan noudattavan sekä vuosina 2000–2009 Suomeen muuttaneita koskevia olotilamallinnuksen estimaatteja että vuosina 1994–1999 Suomeen muuttaneita koskevia olotilamallinnuksen estimaatteja. Esitin ensiksi tulokset myös niin, että oletan kaikkien uusien maahanmuuttajien jäävän Suomeen loppuiäkseen ensimmäisen muuton jälkeen. Yleisesti ottaen tässä luvussa esitetyt elinkaariennusteiden estimaatit ovat positiivisempia kuin kaikki edellä esitetyt elinkaariennustetulokset. Tämä johtuu siitä, että ennustetut suomessaoloajat ovat keskimäärin lyhyempiä tässä luvussa kuin edellä.

En esitä tässä luvussa sitä, kuinka lukea elinkaariennusteiden tuloksia, sillä olen tehnyt jo niin yllä olevassa alaluvussa (9.5.1.1), missä kaikkien ennustepopulaation henkilöiden oletettiin jäävän loppuiäkseen Suomeen ensimmäisestä maahanmuutosta lähtien. Koska tulosten lukeminen on tässä yhteydessä samanlaista kuin yllä esitetyssä luvussa, en toista sitä enää tässä.

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.20: *Somaliassa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonan (KH) sekä prosenttipisteiden* $2,5; 10, 25, 50, 75, 90; 97,5 estimaatit$*. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-526	-776
Keskihajonta	620	1 335
MC-keskivirhe	4	8
2,5:s prosenttipiste (2,5 %)	-2 062	-3 965
10. prosenttipiste (10 %)	-1 382	-2 230
Alakvartiili (25 %)	-838	-1 166
Mediaani (50 %)	-349	-481
Yläkvartiili (75 %)	-106	-119
90. prosenttipiste (90 %)	-11	50
97,5:s prosenttipiste (97,5 %)	392	951



Kuva 9.5: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histograami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.21: *Somaliassa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-687	667	-1 641	-1 076	-498	-163	-50
Miehet	242	-360	519	-1 035	-616	-237	-68	41
Naiset (ml. LV)	248	-1 725	2 083	-4 065	-2 788	-1 552	-592	133
Miehet (ml. LV)	242	-496	625	-1 257	-851	-482	-146	216

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.22: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-449	740	-1 509	-781	-205	-50	144
20–24	81	-611	713	-1 671	-1 036	-362	-98	-9
25–29	85	-594	618	-1 490	-985	-400	-120	-20
30–34	64	-585	528	-1 327	-923	-450	-155	-50
35–39	34	-472	432	-1 077	-739	-367	-130	-36
40–49	51	-531	380	-1 026	-766	-486	-214	-90
50–62	33	-404	303	-779	-560	-359	-169	-70
16–19 (ml. LV)	142	-741	1 445	-2 399	-1 273	-400	-62	376
20–24 (ml. LV)	81	-900	1 417	-2 540	-1 456	-588	-115	99
25–29 (ml. LV)	85	-846	1 407	-2 318	-1 310	-552	-124	19
30–34 (ml. LV)	64	-829	1 320	-2 219	-1 233	-569	-156	-10
35–39 (ml. LV)	34	-788	1 349	-2 192	-1 006	-445	-130	-9
40–49 (ml. LV)	51	-723	1 112	-1 671	-927	-536	-213	-63
50–62 (ml. LV)	33	-404	303	-779	-560	-359	-169	-70

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.23: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-530	541	-404
Ei 2. asteen koulutusta/ei tietoa	458	-526	626	-345
Vähintään 2. asteen koulutus (ml. LV)	32	-741	1 180	-471
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-778	1 345	-482

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.24: *Somaliassa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-256	478	-152
Opiskelija	48	-516	612	-368
Työtön	23	-614	582	-450
Muu	406	-531	625	-348
Työllinen (ml. LV)	13	-297	646	-160
Opiskelija (ml. LV)	48	-661	1 194	-415
Työtön (ml. LV)	23	-861	1 269	-590
Muu (ml. LV)	406	-800	1 367	-500

9.5.2 Irak

Esitän tässä luvussa elinkaariennusteet Irakissa syntynyttä ennustepopulaatiota koskien. Elinkaariennusteet vastaavat rakenteeltaan edellä Somaliassa syntyneitä koskevia vastaavia elinkaariennusteita, joten en toista ennusteita koskevia selitteitä tässä.

9.5.2.1 Koko ikä maahanmuutosta kuolemaan asti Suomessa

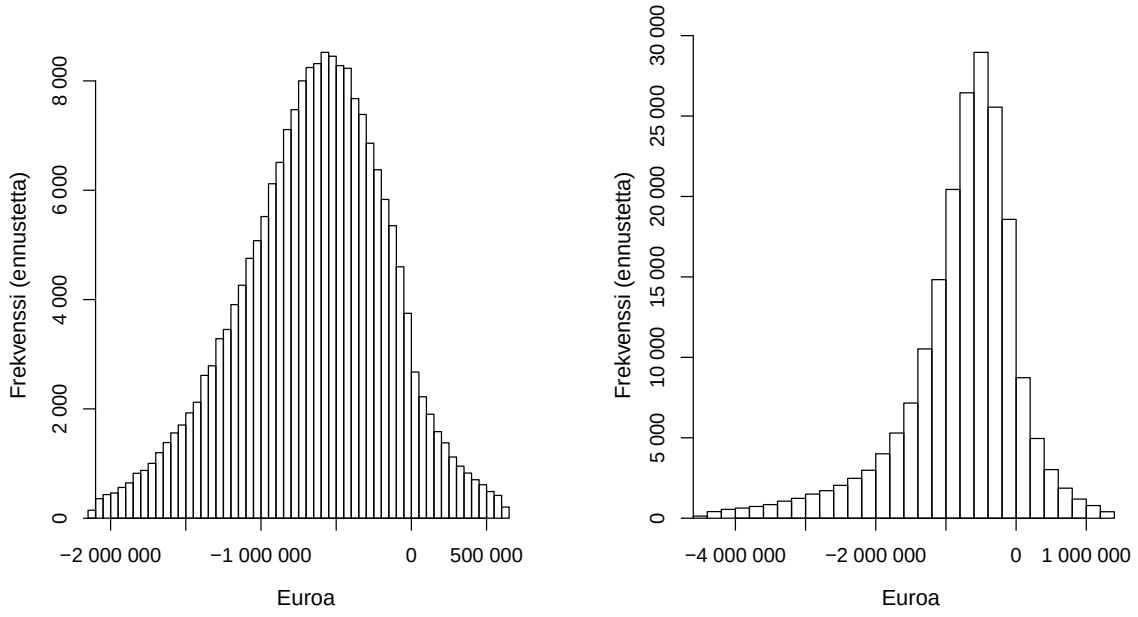
Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.25: *Irakissa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-654	-793
Keskihajonta	541	1 164
MC-keskivirhe	4	7
2,5:s prosenttipiste (2,5 %)	-1 821	-3 451
10. prosenttipiste (10 %)	-1 341	-1 893
Alakvartiili (25 %)	-968	-1 118
Mediaani (50 %)	-619	-639
Yläkvartiili (75 %)	-306	-277
90. prosenttipiste (90 %)	-43	50
97,5:s prosenttipiste (97,5 %)	361	713

Taulukko 9.26: *Irakissa syntyneen ennusteväestön elinkaariennusteet. Toisin kuin muissa Irakia koskevissa tuloksissa, tässä maassaoloajan vaikutuksen annetaan kasvaa 21 maassaolovuoden jälkeen maassaolovuosien mukaan. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ennusteväestö (N)	677
Odotusarvo	-630
Keskihajonta	633
MC-keskivirhe	4
2,5:s prosenttipiste (2,5 %)	-1 988
10. prosenttipiste (10 %)	-1 416
Alakvartiili (25 %)	-982
Mediaani (50 %)	-594
Yläkvartiili (75 %)	-257
90. prosenttipiste (90 %)	60
97,5:s prosenttipiste (97,5 %)	623



Kuva 9.6: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.27: *Irakissa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-808	542	-1 508	-1 126	-757	-449	-188
Miehet	460	-582	526	-1 242	-888	-554	-248	11
Naiset (ml. LV)	217	-1 241	1 829	-3 173	-2 059	-1 062	-404	251
Miehet (ml. LV)	460	-582	526	-1 242	-888	-554	-248	11

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.28: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-677	685	-1 543	-1 115	-662	-221	159
20–24	153	-690	623	-1 480	-1 088	-673	-272	57
25–29	147	-670	547	-1 366	-1 012	-652	-300	-16
30–34	108	-652	477	-1 262	-948	-629	-319	-85
35–39	62	-649	404	-1 161	-886	-624	-365	-159
40–49	70	-597	335	-1 009	-779	-569	-369	-198
50–62	34	-488	282	-830	-620	-451	-301	-167
16–19 (ml. LV)	103	-826	1 249	-2 086	-1 277	-694	-193	244
20–24 (ml. LV)	153	-844	1 227	-2 051	-1 246	-703	-242	155
25–29 (ml. LV)	147	-811	1 160	-1 908	-1 160	-673	-267	75
30–34 (ml. LV)	108	-812	1 268	-1 959	-1 111	-652	-283	9
35–39 (ml. LV)	62	-847	1 327	-2 154	-1 148	-669	-320	-36
40–49 (ml. LV)	70	-667	735	-1 272	-843	-579	-349	-154
50–62 (ml. LV)	34	-488	282	-830	-620	-451	-301	-167

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.29: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-624	467	-583
Ei 2. asteen koulutusta/ei tietoa	543	-661	558	-629
Vähintään 2. asteen koulutus (ml. LV)	134	-735	966	-598
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-808	1 207	-651

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.30: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

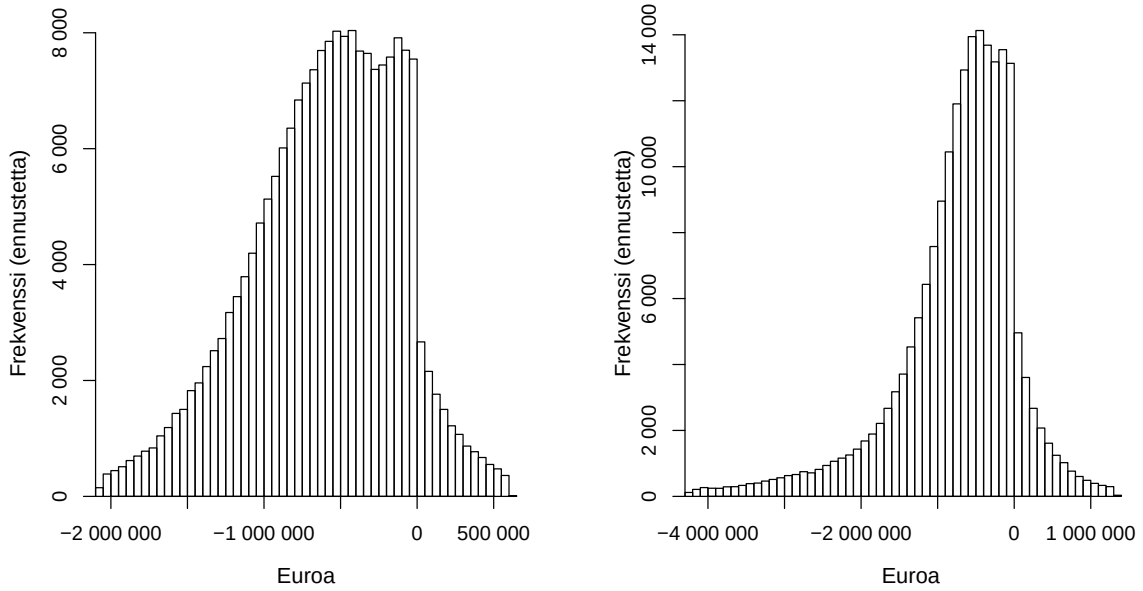
Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-257	480	-246
Opiskelija	77	-700	557	-663
Työtön	79	-649	509	-616
Muu	493	-670	538	-632
Työllinen (ml. LV)	28	-257	480	-246
Opiskelija (ml. LV)	77	-871	1 234	-694
Työtön (ml. LV)	79	-737	961	-629
Muu (ml. LV)	493	-820	1 201	-658

9.5.2.2 Suomessaoloaika vuosina 2000–2009 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.31: *Irakissa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonan (KH) sekä prosenttipisteiden* $2,5; 10, 25, 50, 75, 90; 97,5 estimaatit$*. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-603	-731
Keskihajonta	535	1 111
MC-keskivirhe	4	6
2,5:s prosenttipiste (2,5 %)	-1 784	-3 248
10. prosenttipiste (10 %)	-1 297	-1 757
Alakvartiili (25 %)	-919	-1 054
Mediaani (50 %)	-553	-584
Yläkvartiili (75 %)	-226	-215
90. prosenttipiste (90 %)	-27	31
97,5:s prosenttipiste (97,5 %)	339	685



Kuva 9.7: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.32: *Irakissa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-735	547	-1 457	-1 063	-682	-338	-99
Miehet	460	-541	517	-1 204	-847	-498	-186	-3
Naiset (ml. LV)	217	-1 135	1 745	-2 970	-1 862	-950	-350	228
Miehet (ml. LV)	460	-541	517	-1 204	-847	-498	-186	-3

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.33: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-624	665	-1 484	-1 050	-580	-153	126
20–24	153	-636	610	-1 431	-1 024	-594	-193	32
25–29	147	-620	543	-1 327	-965	-580	-217	-11
30–34	108	-598	480	-1 222	-899	-560	-235	-49
35–39	62	-597	410	-1 124	-851	-566	-285	-105
40–49	70	-548	345	-978	-751	-522	-291	-125
50–62	34	-453	292	-802	-601	-417	-249	-115
16–19 (ml. LV)	103	-761	1 181	-1 934	-1 203	-626	-148	217
20–24 (ml. LV)	153	-779	1 169	-1 909	-1 176	-637	-183	135
25–29 (ml. LV)	147	-749	1 123	-1 778	-1 098	-613	-204	57
30–34 (ml. LV)	108	-747	1 207	-1 802	-1 052	-592	-219	-2
35–39 (ml. LV)	62	-779	1 268	-1 970	-1 070	-622	-261	-23
40–49 (ml. LV)	70	-612	689	-1 195	-804	-532	-277	-94
50–62 (ml. LV)	34	-453	292	-802	-601	-417	-249	-115

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.34: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-576	462	-532
Ei 2. asteen koulutusta/ei tietoa	543	-609	551	-560
Vähintään 2. asteen koulutus (ml. LV)	134	-675	926	-553
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-745	1 152	-593

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.35: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

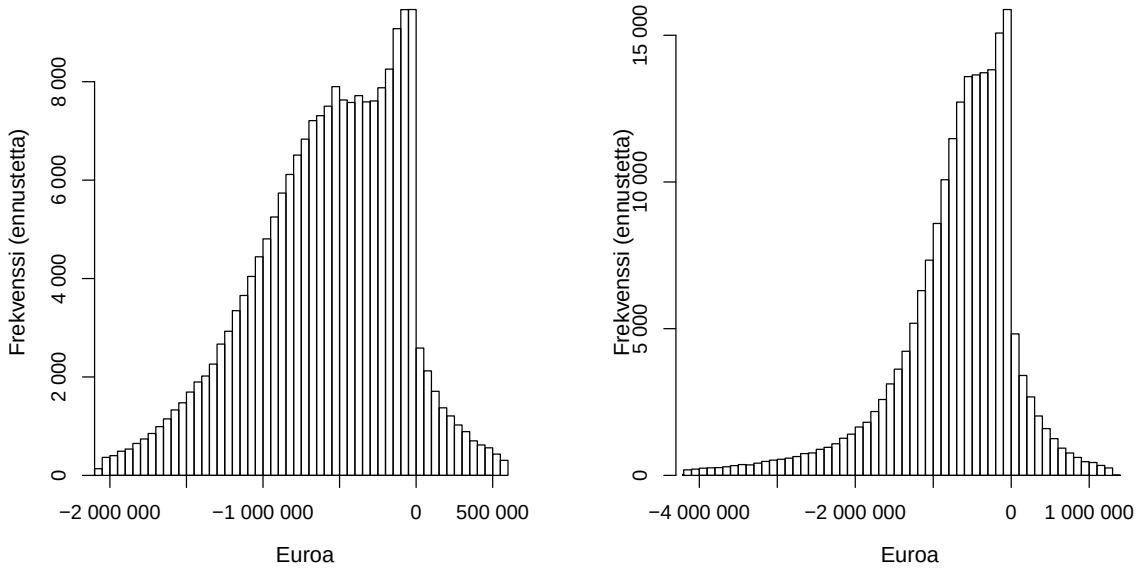
Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-244	472	-212
Opiskelija	77	-660	551	-610
Työtön	79	-597	500	-553
Muu	493	-615	533	-564
Työllinen (ml. LV)	28	-244	472	-212
Opiskelija (ml. LV)	77	-821	1 180	-655
Työtön (ml. LV)	79	-676	932	-573
Muu (ml. LV)	493	-753	1 145	-599

9.5.2.3 Suomessaoloaika vuosina 1994–1999 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.36: *Irakissa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonan (KH) sekä prosenttipisteiden* $2,5; 10, 25, 50, 75, 90; 97,5 estimaatit$*. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-580	-704
Keskihajonta	532	1 093
MC-keskivirhe	4	6
2,5:s prosenttipiste (2,5 %)	-1 767	-3 183
10. prosenttipiste (10 %)	-1 275	-1 709
Alakvartiili (25 %)	-894	-1 027
Mediaani (50 %)	-523	-556
Yläkvartiili (75 %)	-193	-187
90. prosenttipiste (90 %)	-22	23
97,5:s prosenttipiste (97,5 %)	324	681



Kuva 9.8: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.37: *Irakissa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-707	550	-1 439	-1 040	-650	-289	-75
Miehet	460	-521	513	-1 182	-824	-469	-159	-6
Naiset (ml. LV)	217	-1 093	1 717	-2 894	-1 789	-908	-317	230
Miehet (ml. LV)	460	-521	513	-1 182	-824	-469	-159	-6

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.38: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-605	663	-1 474	-1 023	-546	-129	120
20–24	153	-612	606	-1 414	-998	-554	-159	23
25–29	147	-597	540	-1 310	-944	-545	-185	-11
30–34	108	-577	475	-1 205	-880	-528	-206	-41
35–39	62	-573	410	-1 105	-830	-539	-248	-83
40–49	70	-524	350	-962	-729	-498	-254	-103
50–62	34	-437	294	-791	-591	-408	-222	-91
16–19 (ml. LV)	103	-735	1 170	-1 902	-1 167	-600	-132	215
20–24 (ml. LV)	153	-745	1 151	-1 840	-1 144	-602	-155	127
25–29 (ml. LV)	147	-723	1 089	-1 734	-1 076	-580	-176	49
30–34 (ml. LV)	108	-723	1 203	-1 757	-1 025	-567	-196	-5
35–39 (ml. LV)	62	-754	1 222	-1 896	-1 041	-596	-235	-24
40–49 (ml. LV)	70	-587	710	-1 171	-784	-507	-242	-77
50–62 (ml. LV)	34	-437	294	-791	-591	-408	-222	-91

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.39: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-551	463	-498
Ei 2. asteen koulutusta/ei tietoa	543	-588	548	-530
Vähintään 2. asteen koulutus (ml. LV)	134	-649	892	-525
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-718	1 137	-566

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.40: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

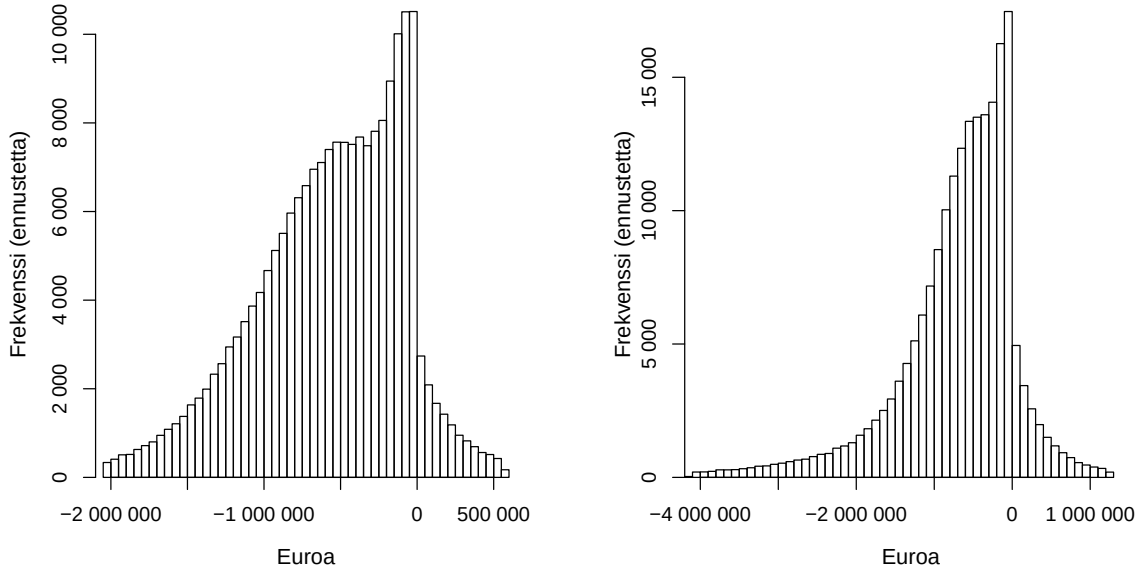
Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-235	463	-193
Opiskelija	77	-646	553	-595
Työtön	79	-571	499	-518
Muu	493	-591	530	-532
Työllinen (ml. LV)	28	-235	463	-193
Opiskelija (ml. LV)	77	-802	1 181	-636
Työtön (ml. LV)	79	-650	903	-542
Muu (ml. LV)	493	-724	1 125	-571

9.5.2.4 Suomessaoloaika ennen vuotta 1994 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.41: *Irakissa syntyneen ennusteväestön elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonan (KH) sekä prosenttipisteiden* $2,5; 10, 25, 50, 75, 90; 97,5 estimaatit$*. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-565	-688
Keskihajonta	528	1 066
MC-keskivirhe	4	6
2,5:s prosenttipiste (2,5 %)	-1 754	-3 102
10. prosenttipiste (10 %)	-1 261	-1 673
Alakvartiili (25 %)	-877	-1 008
Mediaani (50 %)	-501	-539
Yläkvartiili (75 %)	-174	-172
90. prosenttipiste (90 %)	-20	16
97,5:s prosenttipiste (97,5 %)	309	658



Kuva 9.9: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 9.42: *Irakissa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-684	549	-1 418	-1 013	-618	-256	-68
Miehet	460	-509	508	-1 175	-811	-450	-145	-6
Naiset (ml. LV)	217	-1 066	1 670	-2 821	-1 736	-879	-307	213
Miehet (ml. LV)	460	-509	508	-1 175	-811	-450	-145	-6

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 9.43: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden* $10, 25, 50, 75, 90$ *estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-587	652	-1 445	-999	-515	-119	106
20–24	153	-593	599	-1 390	-973	-523	-142	17
25–29	147	-584	536	-1 297	-927	-524	-170	-11
30–34	108	-563	476	-1 198	-865	-510	-184	-41
35–39	62	-558	413	-1 094	-818	-518	-223	-71
40–49	70	-514	351	-959	-724	-482	-237	-92
50–62	34	-427	299	-792	-581	-392	-207	-82
16–19 (ml. LV)	103	-716	1 129	-1 848	-1 152	-575	-124	196
20–24 (ml. LV)	153	-725	1 110	-1 803	-1 119	-575	-144	109
25–29 (ml. LV)	147	-710	1 064	-1 697	-1 058	-566	-164	38
30–34 (ml. LV)	108	-705	1 172	-1 701	-1 012	-555	-179	-6
35–39 (ml. LV)	62	-734	1 228	-1 866	-1 017	-581	-212	-15
40–49 (ml. LV)	70	-576	706	-1 157	-779	-494	-227	-68
50–62 (ml. LV)	34	-427	299	-792	-581	-392	-207	-82

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 9.44: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-539	461	-481
Ei 2. asteen koulutusta/ei tietoa	543	-572	543	-507
Vähintään 2. asteen koulutus (ml. LV)	134	-634	893	-509
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-701	1 104	-549

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 9.45: *Irakissa syntyneen ennusteväestön elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-234	457	-174
Opiskelija	77	-626	546	-569
Työtön	79	-561	497	-499
Muu	493	-575	527	-510
Työllinen (ml. LV)	28	-234	457	-174
Opiskelija (ml. LV)	77	-785	1 147	-622
Työtön (ml. LV)	79	-640	865	-528
Muu (ml. LV)	493	-706	1 100	-554

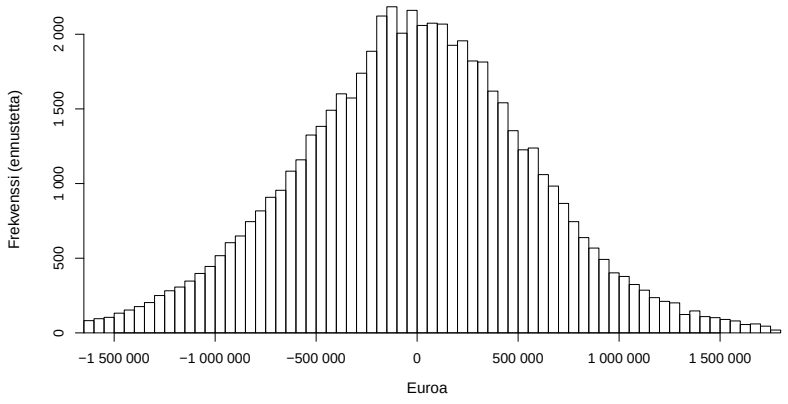
9.6 Suomessa syntyvien (diskonttaamattomat) elinkaariennusteet

Esitän tässä luvussa Suomessa syntyville kantaväestön henkilöille ekinkaariennusteet syntymästä kuolemaan asti.

9.6.1 Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 9.46: *Suomessa syntyneen ennusteväestön elinkaariennusteet. Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden keskiarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2, 5; 10, 25, 50, 75, 90; 97, 5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ennusteväestö (N)	200
Keskiarvo	0
Keskihajonta	898
MC-keskivirhe	6
2,5:s prosenttipiste (2,5 %)	-1 309
10. prosenttipiste (10 %)	-806
Alakvartiili (25 %)	-407
Mediaani (50 %)	-9
Yläkvartiili (75 %)	374
90. prosenttipiste (90 %)	744
97,5:s prosenttipiste (97,5 %)	1 282



Kuva 9.10: *Suomessa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi. Ennustejakaumasta on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

9.6.2 Elinkaariennusteet sukupuolen mukaan

Taulukko 9.47: *Suomessa syntyneen ennusteväestön elinkaariennusteet sukupuolen mukaan. Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden keskiarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	97	-201	859	-978	-586	-194	166	507
Miehet	103	189	893	-562	-193	165	535	890

Osa III

Elinkaariennusteet ja uusien maahanmuuttajien valinta

Luku 10

Elinkaariennusteista puuttuvat julkisen talouden vaikutukset

10.1 Elinkaariennusteista puuttuvat julkisen talouden vaikutukset

Olen esittänyt johdannossa (luvussa 1.2) tässä tutkimuksessa käytetyn määritelmän julkisen talouden nettovaikutuksille (ks. myös Salminen (2015a), luku 3). Julkisen talouden nettovaikutukset koostuvat julkisen talouden tulojen ja menojen erotuksesta. Määritelmän mukaan tutkimus keskittyy sellaisiin suoriin julkisen talouden tuloihin ja menoihin, jotka koskevat henkilöitä (eivätkä esimerkiksi yrityksiä).

On syytä huomioida, että tutkimuksessa ovat kyseessä nimenomaan *suorat* julkisen talouden vaikutukset. Suorat vaikutukset viittaavat sellaisiin julkisen talouden tuloihin ja menoihin, jotka voidaan liittää suoraan henkilöihin heidän maksaminaan veroinaan ja käyttäminään palveluinaan. Esimerkki epäsuorasta vaikutuksesta on henkilön väkivaltarikollisuudesta aiheutuneet epäsuorat julkisen talouden menot, jotka aiheutuvat lähinnä väkivallan uhrin hoitomenoista ja menetetyistä työtunneista. Tämän esimerkin tapauksessa epäsuorat vaikutukset ovat moninkertaiset suoriin vaikutuksiin nähden (ks. tutkimuksen I osa, alaluku 6.5.1), joten epäsuorien vaikutusten sisällyttäminen tutkimukseen voisi muuttaa tuloksia huomattavasti. En kuitenkaan sisällytä niitä mukaan, koska epäsuorille vaikutuksille olisi vaikeaa laatia kattavaa määritelmää sekä niiden suuruudesta olisi vaikeaa saada objektiivista (lähinnä rekisteriaineistoihin) perustuvaa tietoa.

Tutkimuksessa käytetty aineisto perustuu lähes täysin henkilövuositasoisiin rekisteritietoihin, kuten on esitetty johdannon luvussa 1.4. Tarkemmin tutkimusaineisto on kuvattu tutkimuksen I osassa, joka käsittelee toteutuneita julkisen talouden vaikutuksia ja jota voi pitää aineistonkuvauksena tätä tutkimuksen II osaa varten.

Käytettyihin rekisteriaineistoihin ei sisälly kuitenkaan kaikkia henkilöiden suoria julkisen talouden tuloja ja menoja. Olen käsitellyt näitä puuttuvia julkisen talouden tuloja ja menoja – ja siten puuttuvia julkisen talouden nettovaikutuksia – tutkimuksen I osan alaluvussa 5.5. Käsittelen tässä luvussa puuttuvia nettovaikutuksia kuitenkin vielä hieman eritellymmin kuin viitatussa tutkimuksen I osan alaluvussa.

Julkisen talouden menojen osalta puutteita on erityisesti maahanmuuttajien (erityis)palveluiden menoissa. Näitä palveluita ovat erityisesti turvapaikanhakijoina ja kiintiöpakolaisina sekä näiden perheenjäseninä maahan tulleiden vastaanotto- ja kotoutumismenot. Julkisen talouden tulopuolelta puuttuu joitakin verokertymältään pieniä veroja ja veronluonteisia maksuja. Lisäksi erityisesti vanhuusikää koskevien palveluiden osalta minulla ei ole käytössäni rekisteritietoja, joista selviäisi henkilöiden julkisten palveluiden käyttö sekä se, kuinka suuren osan palveluista palvelunsaajat ovat maksaneet itse.

Käsittelen vielä selvyyden vuoksi eritellysti yhteisöjen ja yhtymien toimintaan liittyviä julkisen talouden vaikutuksia, sillä tutkimuksen määritelmän mukaan vain henkilöiden maksamat verot ja veronluonteiset maksut kuuluvat tutkimukseen. Henkilö voi kuitenkin toimia yrittäjänä (esimerkiksi elinkeinonharjoittajana) siten, että henkilön yrityksestään saamien tulojen veronmaksajana on henkilö itse. Kun veronmaksajana on henkilö, nämä verot sisältyvät tutkimuksessa julkisen talouden tuloihin – ja siten myös julkisen talouden nettovaikutuksiin. Erittelen tämän vuoksi yhteisöjä ja yhtymiä koskevat julkisen talouden vaikutukset niihin, jotka sisältyvät tutkimukseen määritelmän mukaisesti, sekä niihin vaikutuksiin, jotka eivät puolestaan sisälly.

Arvioin puuttuvien nettovaikutusten kohdalla, kuinka paljon niiden huomiointi elinkaariennusteissa muuttaisi elinkaariennusteiden tuloksia. Esitettävät puuttuvat vaikutukset eivät ole kattava esitys kaikista puuttuvista vaikutuksista, mutta ne käsittävät mielestäni joitakin – ainakin euromääriltään suurimpia – keskeisimpiä puutteita.

10.2 Puuttuvat julkisen talouden menot

10.2.1 Puutteet maahanmuuttajien palveluiden menojen osalta

Maahanmuuttajien (erityis)palveluilla viittaan niihin julkisiin palveluihin, jotka on suunnattu vain maahanmuuttajille sekä heidän lapsilleen (nk. toisen sukupolven maahanmuuttajat). Näitä palveluita ovat erityisesti turvapaikanhakijoiden vastaanottopalvelut ja kotouttamispalvelut, kiintiöpakolaisten kotouttamispalvelut, maahanmuuttajille ja heidän lapsillensa tarjotut koulutuspalvelut sekä tulkkauspalvelut. Maahanmuuttajien palvelut ovat pääasiasiassa kuntien järjestämiä, ja erityisesti vastaanotto-, kotoutumis- ja tulkkipalveluiden osalta palveluiden järjestämisestä vastaa erillinen kunnan hallinnollinen yksikkö. Esimerkiksi Helsingissä tämä yksikkö on Maahanmuuttoyksikkö ja Espoossa kaupungin Maahanmuuttajapalvelut.

Vastaanotto- ja kotouttamismenot tutkimuksen I osassa. Elinkaariennusteiden osalta puutteita on erityisesti vastaanotto-, kotoutumis- ja tulkkipalveluiden menojen sekä kuntien maahanmuuttopalveluiden yksiköiden hallinnollisten menojen osalta. Maahanmuuttajia ja heidän lapsiaan koskevat koulutusmenot sen sijaan sisältyvät elinkaariennusteisiin – erityisesti työvoimakoulutusten osalta (ks. I osa, luku 6.6). Myös peruskoulutukseen sisältyvän maahanmuuttajille ja heidän lapsilleen suunnatun opetuksen menot sisältyvät ennusteisiin, jotka koskevat alle 16-vuotiaina Suomeen muuttaneita (ks. I osa, luku 7), jotka käytännössä ovat maahanmuuttajien lapsia. Perusopetuksen osalta ei kuitenkaan ole mahdollista

eritellä menojen jakautumista eri maahanmuuttajaryhmien kesken. Tämä koskee niin maahanmuuttajille suunnattua opetusta kuin kaikille oppilaille potentiaalisesti tarjottavaa erityisopetusta.

Rahoitus turvapaikanhakijoiden ja kiintiöpakolaisten vastaanotto- ja kotouttamispalveluihin (ml. tulkkauspalvelut) tulee valtiolta ELY-keskusten kautta. Koska tutkimusaineistossa ei ole tämän rahoituksen suuruuden osalta käytössä henkilövuositasoisia rekisteritietoja, minun oli tutkimuksen I osassa käytettävä toissijaisia lähteitä (lähinnä valtion talousarvioesityksiä) vastaanotto- ja kotoutumismenojen keskimääräisestä suuruudesta turvapaikanhakijaa, kiintiöpakolaista tai heidän perheenjäsentään kohti. Olen selostanut tutkimuksen I osan luvussa 6.7, kuinka määritin valtion talousarvioesitysten avulla keskimääräiset vastaanotto- ja kotouttamiskustannukset. Kuten myös selvennän mainitussa I osan luvussa, nämä keskimääräiset luvut jäävät todellisia keskimääräisiä menoja pienemmiksi.

Osoitan tämän vertaamalla käyttämääni arviota vastaanotto- ja kotoutumismenoista todellisiin toteutuneisiin menoihin vuonna 2011. Kun käytän keskimääräistä arviotani vastaanotto- ja kotoutumismenoista henkilöä kohti, saan vastaanotto- ja kotoutumismenojen kokonaissummaksi 40 miljoonaa euroa. Valtion tilinpäätöksen mukaan todellinen summa vuonna 2011 oli kuitenkin noin 150 miljoonaa euroa. Tämän perusteella elinkaariennusteisiin sisältyy ehkä vain noin kolmasosa todellisista vastaanotto- ja kotoutumismenoista. Ainakin yksi syy tähän eroon on se, että tutkimuksessani ei ole mukana vastaanotto- ja kotoutumispalveluihin sisältyviä menoja alaikäisten hoidosta, johon on varattu vuoden 2013 talousarvioesityksessä noin 20 miljoonaa euroa. Ylipäätään vastaanotto- ja kotoutumismenoista puuttuvat niihin perheenyhdistämisten kautta tulleisiin lapsiin liittyvät kotoutumispalveluiden menot, joiden osalta Maahanmuuttoviraston (osittain puutteellisesta) rekisteristä puuttuu linkki lapsen vanhempaan, joka toimii perheenyhdistäjänä. Niiden lapsien osalta, joilta linkki vanhempaan löytyy, sijoitan menot koskemaan lapsen vanhempia.

Lisäksi tutkimuksen I osassa korvaukset tulkkimenoista on määritetty sijoittuvan vain neljälle ensimmäiselle vuodelle Suomessa, mikä jättää huomiotta tulkkausmenot myöhemmiltä vuosilta. Oletan tutkimuksen I osassa myös, että vastaanottoaika turvapaikanhakijoiden osalta on yksi vuosi, mikä ainakin vuonna 2011 oli tutkimuksen I osassa esitettyjen tietojen perusteella todellista keskimääräistä aikaa lyhyempi aika rauenneiden hakemusten luokkaa lukuun ottamatta. Lisäksi – ja eritoten – en ole ottanut mukaan vastaanotto- ja kotouttamismenoja, joiden osalta ei ole täyttä varmuutta, sisältyvätkö ne jo mukaan tutkimukseen muiden julkisten talouden menojen myötä, eli sisältyvätkö ne jo esimerkiksi sosiaali- ja terveyspalveluiden menoihin. Laskelmani vuoden 2011 vastaanotto- ja kotoutumismenojen kokonaissummaksi, eli 40 miljoonaksi euroksi, perustuu lisäksi vain henkilöihin, joilla on kotikunta Suomessa, eli he ovat saaneet oleskeluluvan. Vastaanottomenoja aiheuttavat kuitenkin myös sellaiset henkilöt, jotka eivät saa oleskelulupaa Suomeen.

Arvioimani yhteenlasketut yhden turvapaikanhakijan keskimääräiset vastaanotto- ja kotoutumismenot olivat tutkimuksen I osassa noin 31 000 euroa vuonna 2011. Kiintiöpakolaisella vastaava luku oli noin 22 000 euroa vuonna 2011. Näissä luvuissa ei ole mukana turvapaikanhakijoiden ja kiintiöpakolaisten lasten perheenyhdistämisistä seuraavia kotouttamismenoja. Mikäli nämä luvut ovat edellä mainitun perusteella vain vajaa kolmasosa todellisista menoista, yhden turvapaikanhakijan keskimääräinen vastaanotto- ja kotoutumismeno olisi todellisuudessa keskimäärin yhteensä noin 100 000 euroa ja kiintiöpakolaisella noin 70 000 euroa, kun mukaan lasketaan myös turvapaikanhakijoiden ja kiintiöpakolaisten lasten perheenyhdistämisistä seuraavat kotouttamismenot. Turvapaikanhakijana Suomeen muuttaneen elinkaariennustees-

ta puuttuisi vastaanotto- ja kotouttamismenoihin liittyen tällöin noin 70 000 euroa, ja kiintiöpakolaisena tulleen osalta noin 48 000 euroa.

Vastaanotto- ja kotouttamismenot valtion v. 2016 tilinpäätöstietojen ja v. 2018 talousarvioesityksen perusteella. Arvioin vielä todellisten vastaanotto- ja kotouttamismenojen suuruutta valtion vuoden 2016 tilinpäätöstietojen sekä vuoden 2018 talousarvioesityksen sisältämien tietojen pohjalta. Vuoden 2016 tilinpäätöksessä sekä vuoden 2018 talousarvioesityksessä näkyy osa niistä vastaanotto- ja kotouttamismenoista, jotka aiheutuvat vuonna 2015 saapuneista turvapaikanhakijoista. Vuonna 2015 saapuneiden turvapaikanhakijoiden lukumäärä on toistaiseksi suurin Suomeen yhtenä vuonna saapuneiden turvapaikanhakijoiden lukumäärä.

Valtion vuoden 2016 tilinpäätöksen mukaan vastaanottomenot olivat vuonna 2016 noin 691 miljoonaa euroa, mikä on noin 655 miljoonaa vuoden 2011 euroissa (Valtioneuvoston kanslia, valtiovarainministeriö 2017, kohta 26.40., s. 22). Valtion vuoden 2018 talousarvioesityksen mukaan vuonna 2016 keskimäärin vastaanoton piirissä olevia henkilöitä oli 25 100 (kohta 26.40.). Tämä tarkoittaa sitä, että vuoden 2011 euroissa vastaanottomenot olivat noin 26 000 euroa vuodessa henkilöä kohti. Turvapaikkahakemusten keskimääräinen käsittelyaika – ilman muutoksenhakua – oli 261 vuorokautta vuonna 2016. Todellista nykyistä käsittelyaikaa – missä huomioitaisiin myös muutoksenhaut – on vaikea arvioida tarkasti, koska vuosi 2016 oli poikkeuksellisen ruuhkainen, mutta 261 vuorokautta ei toisaalta sisällä muutoksenhakuprosessia. Mikäli vastaanottomenot ovat tosiaan noin 26 000 per henkilö vuoden 2011 euroissa, niin tutkimuksen I osassa käyttämäni arvio yhteenlasketuista vastaanottomenoista *ja* kotouttamismenoista (31 000 euroa per henkilö) turvapaikanhakijoille on liian pieni. (Valtiovarainministeriö 2017).

Valtion vuoden 2018 talousarvioesityksen kotouttamista käsittelevän kohdan mukaan 15 516 pakolaista oli laskennallisten kuntakorvausten piirissä vuonna 2016. Vuoden 2016 tilinpäätöksen mukaan näiden pakolaisten kotouttamiseen käytettiin määrärahoja noin 131 miljoonaa euroa (noin 124 miljoonaa vuoden 2011 euroissa). Summa olisi vuoden 2011 euroissa noin 8 000 euroa per henkilö vuodessa. Kolmen vuoden aikana tämä olisi noin 24 000 euroa per henkilö. Vuonna 2017 pakolaisia oli laskennallisten kuntakorvausten piirissä 18 498 henkilöä, ja määrärahoja heidän kotouttamiselleen oli noin 232 miljoonaa euroa (noin 220 miljoonaa vuoden 2011 euroissa). Summa olisi vuoden 2011 euroissa noin 11 200 euroa per henkilö vuodessa. Kolmen vuoden aikana tämä olisi noin 33 600 euroa per henkilö. Tämän perusteella on vaikeaa sanoa, kuinka paljon kotouttamismenot ovat tarkkaan ottaen keskimäärin yhteensä per henkilö. Arvioimista hankaloittaa entisestään, että valtio korvaa turvapaikanhakijoina ja kiintiöpakolaisina tulleiden tulkkausmenot vielä kolmen/neljän vuoden jälkeen. (Valtiovarainministeriö 2017).

Mikäli noudatetaan vuoden 2017 kotouttamismenoja per henkilö sekä vuoden 2016 vastaanottomenoja per henkilö, niin vastaanotto- ja kotouttamismenot ovat vuoden 2011 euroissa ainakin 26 000 + 33 600 = 59 600 per turvapaikanhakijana Suomeen tullut. Tämä on noin kaksinkertainen summa tutkimuksen I osassa käyttämääni arvioon (31 000 euroa). Koska valtio korvaa kunnille kotouttamismenoja ainakin tulkkauksen osalta vielä kolmen vuoden jälkeen, todellinen vastaanoton ja kotouttamisen summa Suomeen jäävästä turvapaikanhakijasta on vielä mainittua 59 600 euroa korkeampi. Irakissa ja Somaliassa syntyneitä koskevissa tämän tutkimuksen elinkaariennusteissa vastaanotto- ja kotouttamismenojen oletetaan olevan 50 000e/henkilö, mikä vaikuttaa olevan edellä sanotun perusteella todellisia menoja pienempi luku.

Muut erityisesti maahanmuuttajiin liittyvät julkisen talouden menot. Muita erityisesti maahanmuuttajiin liittyviä menoja yllä esitettyjen lisäksi ovat esimerkiksi kuntien maahanmuuttoyksiköiden ja maahanmuuttajille suunnattujen palveluiden menot, joita valtio ei korvaa vastaanotto- ja kotoutumismenoista ja jotka eivät liity maahanmuuttajille suunnattuun perusopetukseen. Näiden menojen suuruuden arvioiminen edellyttäisi mahdollisuutta tutustua kuntien tilinpäätöksiin sillä tarkkuudella, että niistä selviäisi, kuinka suuren osan maahanmuuttajien palveluista valtio korvaa kunnille vastaanotto- ja kotoutumismenojen kautta.

Valtiolla itsellään on joitakin maahanmuuttajiin liittyviä menoja yllä esittyjen lisäksi. Esimerkiksi vähemmistövaltuutetun (nyk. yhdenvertaisuusvaltuutettu) ja syrjintälautakunnan toimintamenot olivat vuonna 2014 noin miljoona euroa (Valtioneuvoston kanslia, valtiovarainministeriö 2016, kohta 26.01.04, s. 23). Vähemmistövaltuutetun osalta ei ole nähdäkseni yksiselitteistä ratkaisua määrittää menojen jakautumista eri maahanmuuttajien ja Suomessa syntyneen kantaväestön kesken. Joka tapauksessa elinkaariennusteiden kokonaissummien kannalta vähemmistövaltuutetun menot ovat lähes olemattomia riippumatta menojen jakamisesta. Toinen esimerkki on opetus- ja kulttuuriministeriön koulutuksellisen tasa-arvon edistämiseen vuonna 2013 myöntämä noin 22,5 miljoonaa euroa. Tämän avustuksen myöntöpäätöksissä huomioitiin erityisesti koulunkäyntialueen väestön koulutustaso, työttömyysaste sekä maahanmuuttajien osuuden korkea määrä. Tässäkin tapauksessa on vaikeaa määrittää yksiselitteistä tapaa jakaa tämä meno eri maahanmuuttajien sekä Suomessa syntyneen kantaväestön kesken. (Opetus- ja kulttuuriministeriö 2013).

10.2.2 Muut puutteet julkisten palveluiden menojen osalta

Myös sellaisten julkisten palveluiden osalta, jotka eivät koske vain maahanmuuttajia, on todennäköisesti olemassa sellaisia maahanmuuttajia koskevia puuttuvia menoja, jotka eivät sisälly rekisteriperusteiseen tutkimusaineistoon eivätkä siten elinkaariennusteisiin. Tällaisia julkisten palveluiden menoja ovat esimerkiksi palvelut, joiden yksikkökustannus on todellisuudessa keskimäärin korkeampi maahanmuuttajilla (tai ainakin joillakin maahanmuuttajaryhmillä) kuin Suomessa syntyneellä kantaväestöllä. Tutkimuksen I osassa viittaan esimerkiksi erikoissairaanhoidon menoihin (luku 6.3.1), joissa yksikkökustannus on joillakin maahanmuuttajilla muuta väestöä korkeampi palveluun käytetyn pidemmän ajan vuoksi:

”Niin tässä [Salminen (2015)] kuin Stakesin ja VATT:n tutkimuksessa [Gissler et al. (2006)] ei voitu ottaa huomioon maahanmuuttajien mahdollisesti tarvitsemien erityispalveluiden menoja. Tutkimuksessa [Gissler et al. (2006)] asiasta kirjoitetaan näin: ‘*Kustannustietoja tarkastellessa on huomioitava, ettei maahanmuuttajien hoitokustannuksiin voitu sisällyttää kustannuspainolisää kieli- ja kulttuurikysymysten takia. Selvitysten mukaan maahanmuuttajien äitiyshuoltoon liittyvien kustannusten on laskettu olevan jopa kaksinkertaiset, mikäli mukana on tulkki. Ilman tulkkejakin käyntien kustannukset voivat olla keskimääräisiä suuremmat, sillä henkilökunnan työaikaa kuluu enemmän, vaikka muut kiinteät kustannukset pysyisivät samoina (Hujanen Timo, henkilökohtainen tiedonanto).*’ (Gissler et al. 2006, 22).” (Salminen 2015, luku 6.3.1).

Erikoissairaanhoidon lisäksi esimerkiksi erityisopetuksen saanti peruskoulussa on sellainen palvelu, jossa voi olla eroja kantaväestön ja maahanmuuttajaryhmien välillä.

Julkisten palveluiden yksikkökustannusten erojen lisäksi maahanmuuttajaryhmien ja kantaväestön kohdalla saattaa olla eroja sellaisten julkisten palveluiden käytössä, joista ei ole tietääkseni lainkaan valtakunnallisia rekisteritietoja. Esimerkki tällaisesta palvelusta on kuntien asumispalveluiden käyttö. Koska menot julkisista palveluista muodostuvat palveluiden käytön sekä palveluiden yksikkökustannusten tuloista, suurempi (pienempi) määrä tietyn palvelun käyttöä tarkoittaa suurempaa (pienempää) julkisen talouden menoa palvelunkäytöstä.

On myös joitakin – lähinnä sairausvakuutuksesta maksettavia – saatuja tulonsiirtoja, joista rekisteritiedot on tutkimusaineistossa vain vuotta 2011 koskien. En sisällyttänyt näitä tulonsiirtoja tutkimusaineistoon, jonka pohjalta laadin elinkaariennusteet. Nämä tulonsiirrot, joista on tiedot vain vuotta 2011 koskien, on lueteltu tutkimuksen I osan liitteessä 2.

Vanhuusiän osalta merkittävä puute tutkimuksessa on, että minulla ei ole käytössäni rekisteritietoja yli 70-vuotiaiden käyttämistä sosiaali- ja terveyspalveluista. Näin ollen en tiedä, onko eri syntymävaltioiden (ja henkilöiden) välillä eroja näiden palveluiden käytön jakaumissa. Käytän tämän vuoksi kaikille henkilöille koko Suomen väestöä koskevaa jakaumaa näiden palveluiden käytön osalta.

Lisäksi kaikenlaiset epäsuorat julkisen talouden vaikutukset jäävät pois. Näitä ovat esimerkiksi euromääräisesti suuret epäsuorat negatiiviset vaikutukset väkivaltarikoksista, jotka aiheutuvat mm. näiden rikosten uhrien mahdollisesta kuntouttamisesta ja tulonmenetyksistä.

En esitä arviota siitä, kuinka suuri vaikutus elinkaariennusteisiin on niillä julkisten palveluiden menoja koskevilla puutteilla, jotka koskevat muita kuin maahanmuuttajien palveluita. Tätä voisi olla vaikeaa uskottavasti arvioida toisin kuin puuttuvien vastaanotto- ja kotouttamismenojen edellä. Puuttuvien vastaanotto- ja kotouttamismenojen suuruutta pystyin arvioimaan suoraan valtion talousarvioesitysten ja tilinpäätösten tietojen perusteella.

10.3 Puuttuvat julkisen talouden tulot

Julkisen talouden menojen lisäksi on myös joitakin julkisen talouden tuloja, jotka puuttuvat rekisteriperusteisesta tutkimusaineistosta ja siten elinkaariennusteista. Nämä puutteet ovat lähinnä joitakin henkilöiden maksamia veroja ja veronluonteisia maksuja.

Verohallinnon Taskutilasto 2013 esittää verokertymät Suomessa vuonna 2011 eri veroille ja veronluonteisille maksuille (s. 18–19). Esitän seuraavaksi ne *henkilöiden* maksamat verot ja veronluonteiset maksut, jotka puuttuvat tutkimusaineistoista ja joiden osuus kokonaisverokertymästä on yli 0,2 %:ia (yli 165 miljoonaa euroa). Jätän pois tarkastelusta seurakuntien verot (kirkollisvero) sekä valtion osuuden RAY:n tuotoista, sillä ne eivät ole mielestäni tämän tutkimuksen määritelmän mukaisia henkilöiden maksamia suoria julkisen talouden tuloja.

Varainsiirtovero, autovero, ajoneuvovero ja kiinteistöverot. Puuttuvia veroja ja veronluonteisia maksuja ovat (suluissa niiden osuus kokonaisveromäärästä vuonna 2011) Verohallinnon Taskutilaston ”muiden verojen” luokassa: varainsiirtovero (0,7 %), autovero (1,3 %) ja ajoneuvovero (0,9 %). Kuntien veroista puuttuvat lähinnä (henkilöiden maksamat) kiinteistöverot (1,5 %). Euromääräisesti henkilöt maksoivat vuonna 2014 noin kolmasosan kaikista maksetuista kiinteistöveroista (eli 0,5 % kokonaisverokertymästä) (Veronmaksajain keskusliitto 2014). Mikäli koko varainsiirtoveron, autoveron ja ajoneuvoveron kertymä sekä kolmasosa kiinteistöveroista käsitetään henkilöiden maksamiksi, niin tämä tarkoittaa n. 2,7 miljardin verokertymän puuttumista koko väestön osalta vuodessa. On huomioitava, että tästäkin määrästä osa on muiden kuin henkilöiden maksamia. Koko Suomen väestöön suhteutettuna tämä tarkoittaa keskimäärin noin 530 euron vuosittaisen verokertymän puuttumista tutkimuksessa henkilöä kohti (vuoden 2011 havaintojen osalta). Koko elinkaaren yli tämä tarkoittaa keskimäärin noin 40 000 euron puuttumista maksetuista veroista vuoden 2011 euroissa.

Käytössäni olevassa tutkimusaineistossa on rekisteritiedot henkilöiden asuntokuntien autojen lukumääristä sekä henkilöiden maksamista kiinteistöveroista vuonna 2011. Ongelmana näiden käytössä on, että autoista maksettujen verojen jakaminen asuntokunnan jäsenten kesken ei olisi yksiselitteistä. Lisäksi minulla ei ole käytössäni rekisteritietoa autoista maksettavien ajoneuvoverojen suuruuksista. Henkilöiden maksamista kiinteistöveroista ei puolestaan ole tietoja kaikille vuosille 1995–2011. Näitä tietoja tutkimalla saisin kuitenkin käsityksen siitä, kuinka näiden verojen maksaminen jakautuu suunnilleen eri maahanmuuttajaryhmien ja Suomessa syntyneen kantaväestön osalta. Jätän tämän ainakin toistaiseksi tekemättä.

Puuttuvat asiakasmaksut vanhuusiän palveluista. Julkisen talouden tuloihin liittyvä vanhuusiän palveluja koskeva puute on, että en tiedä, minkä osuuden asiakkaat maksavat itse saamistaan palveluista. Ainakin kotihoidon osalta asiakkaan maksama osuus riippuu asiakkaan omista tuloista. Tämä osuus kuitenkin vaihtelee kunnittain, joten vaikka elinkaariennusteisiin sisältyy ennuste eläketuloista, niin minulla ei ole kovin tarkkaa tietoa siitä, kuinka suuren osuuden henkilö maksaa saamistaan vanhuusiän palveluista. Samat puutteet koskevat asiakasmaksuja myös palvelu- ja laitosasumisen osalta. Jatkossa näiden asiakasmaksujen suuruuden voisi mallintaa jollain ehdollisella todennäköisyysjakaumalla, joka kuvaisi asiakasmaksuun sisältyvää epävarmuutta. Tämä todennäköisyysjakauma olisi ehdollinen ennustetuilla eläketuloilla.

Koska kuukausittaisissa asiakasmaksuissa voi olla eri kunnissa asuvilla ja eri eläketulot omaavilla vähintään satojen eurojen eroja, puuttuvat vanhuusiän asiakasmaksut saattavat olla suurin yksittäinen puute julkisen talouden vaikutuksissa tässä tutkimuksessa.

10.4 Tutkimukseen sisältyvät ja siitä puuttuvat yhteisöjen ja yhtymien vaikutukset

Julkisen talouden tuloja ja menoja aiheutuu yhteisöjen ja yhtymien toiminnasta. Tämän tekstin johdannossa sekä tutkimuksen I osassa on määritetty, että tässä tutkimuksessa tarkastellaan vain sellaisia

julkisen talouden vaikutuksia, jotka ovat henkilöitä suoraan koskevia julkisen talouden tuloja tai menoja (ks. tarkemmin johdanto, luku 1.2 sekä tutkimuksen I osa, luku 3).

Osa yhteisöjen toiminnan aiheuttamista julkisen talouden vaikutuksista kuitenkin sisältyy tämän määritelmän puitteissa mukaan tutkimukseen. Lisäksi elinkaariennusteiden osalta käsitän määritelmän ulkopuolelta lisäksi työnantajien maksamat eläkevakuutus- ja sosiaaliturvamaksut myös henkilöiden maksamina julkisen talouden tuloina. Tämä johtuu yhtäältä siitä, että nämä maksut saadaan liitettyä rekisteritietojen avulla suoraan henkilöön, joita ne koskevat. Toisaalta tämä johtuu siitä, että näillä maksuilla kustannetaan muun muassa henkilöiden eläkkeitä, sairauksista aiheutuvia menoja ja ansiosidonnaisia työttömyyskorvauksia, jotka sisältyvät julkisen talouden menoihin tässä tutkimuksessa.

Käyn tässä alaluvussa lyhyesti läpi, mitkä muut yhteisöjen toimintaan liittyvät julkisen talouden vaikutukset sisältyvät – tai eivät sisälly – elinkaariennusteisiin. Aloitan julkisen talouden tuloista ja käyn tämän jälkeen läpi julkisen talouden menot. Lopuksi esitän arvion siitä, kuinka kaikkien yhteisövaikutusten sisällyttäminen tutkimukseen vaikuttaisi elinkaariennusteiden tuloksiin.

10.4.1 Henkilöverotuksen kautta tutkimukseen sisältyvät julkisen talouden vaikutukset

Julkisen talouden tulopuoli koostuu maksetuista veroista ja veronluonteisista maksuista. Se, kuuluvatko yrityksen tuloksesta maksetut verot ja veronluonteiset maksut tutkimuksen määritelmän mukaan elinkaarivaikutusten ennusteisiin, riippuu yrityksen yhtiömuodosta. Yksityisten elinkeinonharjoittajien ja yhtymien tuloksista maksetut verot ja veronluonteiset maksut sisältyvät elinkaariennusteisiin, sillä niissä veronmaksajina ovat henkilöt. Koska näiden yhtiömuotojen osalta veronmaksu tapahtuu käytännössä henkilöiden maksamina ansiotulo- ja pääomatuloveroina, nämä julkisen talouden tulot sisältyvät tutkimuksen määritelmän mukaan julkisen talouden nettovaikutuksiin. Lisäksi kaikista henkilöiden vuosittain maksamista ansiotulo- ja pääomatuloveroista on käytettävissä olevassa tutkimusaineistossa rekisteritiedot. Näin ollen tutkimusaineistossa – ja siten elinkaarivaikutusennusteissa – on huomioitu henkilöiden maksamat verot niiden yritysten tuloksesta, joissa he ovat yksityisinä elinkeinonharjoittajina, sekä henkilöiden maksamat verot yhtymien tuloksesta.

10.4.2 Yhteisöverot eivät sisälly julkisen talouden vaikutuksiin tässä tutkimuksessa

Osakeyhtiöiden ja muiden yhteisöjen maksamat yhteisöverot eivät puolestaan sisälly määritelmän mukaan henkilöiden suoriin julkisen talouden vaikutuksiin, sillä yhteisöveron maksajana on yhteisö eikä henkilö. Tutkimuksen määritelmää julkisen talouden vaikutuksista voisi laajentaa siten, että tarkasteltaviin julkisen talouden vaikutuksiin sisältyisivät myös yhteisöjä koskevat julkisen talouden tulot ja menot. Koska maahanmuuttajat ovat yksittäisiä henkilöitä (kuten myös Suomessa syntyneeseen kantaväestöön kuuluvat henkilöt), ei ole mielestäni selvää, kuinka yhteisöjen julkisen talouden vaikutukset tulisi määritellä jaettavaksi henkilöiden kesken.

Osakeyhtiöiden kohdalla julkisen talouden vaikutukset voisi teoriassa jakaa henkilöille osakeomistussuhteiden mukaan. Tämän määritelmän tekisi monimutkaiseksi toteuttaa ainakin se, että osakeyhtiön

osakkeenomistajina voi olla esimerkiksi muita osakeyhtiöitä. Lisäksi tutkimusaineistossa ei ole käytössä tietoja henkilöiden omistamista osakkeista, joten yhteisöverotuottoja ei voi käytännössä jakaa omistusten perusteella tutkimusaineistoon sisältyvien henkilöiden kesken. Koska henkilöiden osakkeista saamat osinkotulot ja osakkeista saadut myyntivoitot verotetaan pääomatuloina, pääomatulojen perusteella voisi epäsuorasti allokoida ainakin osan maksetuista yhteisöveroista henkilöille.

Yhteisöverojen osalta on huomioitava myös, että niiden maksujen jakauma on erittäin vino. Kauppalehden mukaan vuonna 2013 yhteisöveroverovelvollisia oli 311 000. Näistä kuitenkaan kaksi kolmasosaa ei maksanut lainkaan yhteisöveroja vuonna 2013. Yli miljoona euroa yhteisöveroa maksaneita yhtiöitä oli tuolloin 473 eli 0,2 % kaikista verovelvollisista yhtiöistä. Nämä vajaa 500 yhtiötä maksoivat yhteensä melkein 60 %:ia koko 4,6 miljardin euron yhteisöverokertymästä vuonna 2013. Vain 1,6 %:ia yhteisöverovelvollisista maksoi noin 80 %:ia kaikista yhteisöveroista. Tämän perusteella vaikuttaa selvältä, että mikäli yhteisöverot allokoitaisiin henkilöille julkisen talouden tuloina omistusten mukaan, näiden julkisen talouden tulojen suuruus riippuisi henkilötasolla siitä, kuinka paljon henkilöt omistavat suurimpien yhtiöiden osakkeita. (Kauppalehti 4.11.2014).

Yhteisöjen maksamat välilliset verot eivät sisälly julkisen talouden vaikutuksiin tässä tutkimuksessa

Maksettuja yhteisöveroja vielä tärkeämpi yhteisöjen ja yhtymien toimintaan liittyvä julkisen talouden tulojen lähde ovat maksetut välilliset verot. Välillisiä veroja ovat arvonlisävero (alv) ja valmisteverot.

Maksetut arvonlisäverot. Arvonlisäveroja kertyi vuonna 2011 noin 17 miljardia euroa (Verohallinto 2013). Vertaamalla tätä summaa työikäisten 20–62-vuotiaiden maksamiin välillisiin veroihin (tutkimuksen I osa 6.2.4.), on selvää, että joko tutkimuksen I osan menetelmällä ei saada selville kaikkia henkilöiden (kuluttajien) maksamia arvonlisäveroja, tai osa arvonlisäveroista on viime kädessä yhteisöjen maksamia (vähennykset huomioituna). Kun kuluttaja (henkilö) ostaa arvonlisäverollisen tuotteen tai palvelun, niin viime kädessä kuluttaja maksaa alv:n. Silloin kun toinen yritys ostaa viime kädessä tuotteen tai palvelun, useimmiten yritys voi vähentää alv:n ostoistaan (Verohallinto b). Arvonlisäveron kertymän perusteella kuitenkin ainakin osa arvonlisäveroista – jossa maksajana ei ole kuluttaja – jää yrityksiltä vähentämättä. Lisäksi julkisyhteisöjen osalta yleisesti ottaen kunnat ja valtion virastot ja laitokset maksavat arvonlisäveroa, kun ne myyvät tavaroita tai palveluita liiketoiminnan muodossa (Verohallinto a,Verohallinto c).

Maksetut valmisteverot. Arvonlisäverojen lisäksi yhteisöt maksavat osan valmisteverojen, erityisesti energiaverojen kertymästä. Polttoaineverojen kertymä vuonna 2011 oli noin 3,9 miljardia euroa, mikä oli noin 4,7 %:ia kokonaisverokertymästä (Verohallinto 2013). On vaikeaa arvioida saatavilla tiedoilla, kuinka suuri osuus tästä oli yhteisöjen ja yhtymien maksamia veroja. Tämän arvioimista vaikeuttaa myös se, että osa yhteisöistä on joko vapautettu joistakin energiaveroista tai ne saavat energiaveroista veronpalautuksia (Verohallinto d).

Kuten yhteisöverojen osalta yllä, yhteisöjen maksamien välillisten verojen allokoinnille henkilöille ei ole olemassa mitään objektiivista sääntöä. Tämä on yksi keskeisimmistä syistä, miksi jätin yhteisöjen vaikutukset pois tutkimuksesta, sillä yhteisövaikutusten allokointivalinta olisi mielestäni liian kiistanalai-

nen. Yhteisöjen maksamat välilliset verot voitaisiin jakaa tasan esimerkiksi koko väestön tai pelkästään työikäisten kesken. Henkilöiden todellista vaikutusta yhteisöjen maksamien välillisten verojen syntyyn kuvastaisi kuitenkin paremmin allokointitapa, jossa välilliset verot jaettaisiin tasan esimerkiksi kaikkien työllisten (ml. yrittäjät) kesken tai painotettuna työtulojen mukaan.

Maksetut välilliset verot sisältyvät julkisen talouden vaikutuksiin, kun niitä maksetaan (epäsuorasti) elinkeinotoiminnan tuloista.

Maksetut kiinteistöverot. Edellä mainittujen verojen ja veronluonteisten maksujen lisäksi yhteisöt maksoivat euromääräisesti vuonna 2014 noin 60 %:ia kaikista maksetuista kiinteistöveroista (eli noin 1 %:n kokonaisverokertymästä) (Veronmaksajain keskusliitto 2014).

Puuttuvat menot: erityisesti yritystuet. Yritysten toiminnan aiheuttamista julkisen talouden menoista tutkimuksen ulkopuolelle jäävät erityisesti yritystuet. Yrittäjiksi ryhtyvien henkilöiden starttirahat puolestaan sisältyvät tutkimuksessa julkisen talouden menopuoleen (ks. tutkimuksen I osa, liite 2: Tutkimukseen sisältyvät suorat tulonsiirrot, saadut tulonsiirrot).

Johtopäätös puuttuvista yhteisöjä koskevista julkisen talouden vaikutuksista. Mikäli tutkimuksen määritelmää julkisen talouden vaikutuksista laajennettaisiin koskemaan yllä mainittuja yhteisöjen suoria julkisen talouden vaikutuksia, tämän vaikutus elinkaariennusteisiin riippuisi siitä, miten nyt ”puuttuvat” vaikutukset allokoitaisiin tutkittavien väestöjen kesken. Mikäli vaikutukset jaettaisiin tasan esimerkiksi kaikkien työikäisten kesken, elinkaariennusteiden jakaumien erot syntymävaltioiden *välillä* pysyisivät suunnilleen samansuuruisina, kuin ne nyt ovat ilman mainittuja yhteisövaikutuksia. Mikäli taas yhteisövaikutukset jaettaisiin työllisyyden tai työtulojen mukaan, niin Suomessa syntyneiden elinkaariennusteet muuttuisivat keskimäärin positiivisemmiksi *verrattuna* sellaisiin syntymävaltioihin, joilla on matala työllisyysaste ja keskimäärin matalat työtulot.

10.5 Yhteenveto puuttuvista julkisen talouden vaikutuksista

Esitän vielä yhteenvedon yllä esitetyistä puuttuvista julkisen talouden vaikutuksista. Yllä esitetty ei ole kattava esitys kaikista puuttuvista vaikutuksista, mutta käsittää mielestäni joitakin – ainakin euromääriltään suurimpia – keskeisimpiä puutteita.

Elinkaariennusteita koskevat tulokset muuttuisivat huomattavasti, mikäli alla taulukossa 10.1 esitettävistä puutteista tutkimukseen sisällytettäisiin erityisesti yhteisövaikutuksen työllisyyden tai ansiotulojen mukaan jaettuina. Mikäli jako tehtäisiin työllisyyden tai ansiotulojen perusteella, niin Suomessa syntyneen kantaväestön elinkaariennusteet muuttuisivat positiivisemmiksi verrattuna sellaisiin maahanmuuttajaryhmiin, joissa työllisyys ja ansiotulot ovat keskimäärin matalia.

Elinkaariennusteet muuttuisivat oletettavasti myös melko paljon, mikäli vanhuusiän palveluiden asiakasmaksut sisällytettäisiin ennusteina, joissa asiakasmaksun määrä riippuisi henkilön ennustetuista eläkeiän tuloista. Koska Suomessa syntyneillä on eläkeiässä ennustetusti keskimäärin huomattavasti suuremmat tulot kuin Irakissa ja Somaliassa syntyneillä, Suomessa syntyneiden elinkaariennusteisiin sisältyisi keskimäärin huomattavasti enemmän ennustettuja vanhuusiän palveluiden asiakasmaksuja. Tämä puolestaan lisäisi Suomessa syntyneiden elinkaariennusteiden keskimääräistä eroa verrattuna Irakissa ja Somaliassa syntyneiden elinkaariennusteisiin.

Taulukko 10.1: *Lyhyt yhteenveto puuttuvista julkisen talouden vaikutuksista elinkaariennusteissa.*

Puuttuva vaikutus	**Lyhyt yhteenveto puuttuvasta vaikutuksesta**
1. Puuttuvat julkisen talouden menot	
Maahanmuuttajien (erityis)palvelut, erityisesti vastaanotto- ja kotoutumismenot.	Turvapaikanhakijoina tai kiintiöpakolaisina Suomeen tulleiden vastaanotto- ja kotouttamismenojen summana käytetään 50 000 euroa per henkilö. Vuoden 2016 valtion tilinpäätöstietojen ja vuoden 2018 valtion talousarvioesityksen perusteella yhden turvapaikanhakijan vastaanotto- ja kotouttamismenot ovat yhteensä ainakin noin 60 000 per henkilö. Todellinen luku voi olla tätä huomattavasti suurempi erityisesti, mikäli turvapaikkakäsittely kaikkine hakemuksineen venyy useiden vuosien mittaiseksi. Perusopetukseen sisältyvän maahanmuuttajien lasten koulutuksen keskimääräiset menot mukana, mutta lapsina Suomeen saapuvien vastaanotto- ja kotouttamismenot eivät ole mukana. Erityisopetuksen menojen mahdollisista eroista ryhmien välillä ei tietoja.

Jatkuu seuraavalla sivulla

Taulukko 10.1 – jatkuu edelliseltä sivulta

Puuttuva vaikutus	Lyhyt yhteenveto puuttuvasta vaikutuksesta
Muut maahanmuuttajiin liittyvät julkisen talouden menot.	Käsittää kaikki vain maahanmuuttajille tarkoitettujen palveluiden, avustuksien yms. menot, jotka eivät yllä olevassa luokassa. Esimerkiksi kuntien maahanmuuttajien palvelut, joita valtio ei korvaa. Tästä osasta kuntien maahanmuuttajiin liittyviä menoja on hankala saada tarpeeksi tarkkoja tilinpäätöstietoja. Valtion osalta ainakin osa yhdenvertaisuusvaltuutetun ja maahanmuuttajiin liittyvien avustusten menoista. Näitä valtion menoja vaikeaa jakaa yksiselitteisesti eri maahanmuuttajien ja muun väestön kesken.
Julkisten palveluiden yksikkökustannusten erot Suomessa syntyneen kantaväestön ja eri maahanmuuttajaryhmien välillä.	Suomessa syntyneen kantaväestön ja eri maahanmuuttajaryhmien käyttämien julkisten palveluiden keskimääräisissä yksikkökustannuksissa voi olla eroja, joiden suuruudesta ei ole laajaa tietoa. Esimerkiksi äitiyshuoltoon liittyvien palveluiden yksikkökustannukset voivat olla kaksinkertaiset verrattuna kantaväestöön, mikäli mukana on tulkki.
Puuttuvat rekisteritiedot julkisista palveluista.	Puuttuvia rekisteritietoja saaduista julkisista palveluista ovat esimerkiksi kuntien asumispalvelut, joten niitä koskevat menot puuttuvat elinkaariennusteista.
Puuttuvat saadut tulonsiirrot.	Puuttuvat rekisteritiedot saaduista tulonsiirroista koskevat lähinnä joitakin sairausvakuutuksesta maksettavia tulonsiirtoja. Nämä rekisteritiedot löytyvät vain vuodelle 2011 tutkimusaineistossa, ja olen luetellut ne tutkimuksen I osan liitteessä 2.
Vanhuusiän palvelut.	Vanhuusiän saaduista julkisista palveluista ei ole rekisteritietoja tutkimusaineistossa, joten en tiedä, kuinka ne jakautuvat henkilöittäin ja siten myös ryhmittäin.
2. Puuttuvat julkisen talouden tulot	

Jatkuu seuraavalla sivulla

Taulukko 10.1 – jatkuu edelliseltä sivulta

Puuttuva vaikutus	Lyhyt yhteenveto puuttuvasta vaikutuksesta
Puuttuvat henkilöiden maksamat verot ja veronluonteiset maksut.	Henkilöiden maksamista veroista puuttuvat (pienimpiä veroja huomioimatta) varainsiirtovero, autovero, ajoneuvovero ja kiinteistövero. Vuonna 2011 näiden verojen kertymä oli noin 2,7 miljardia euroa, kun kiinteistöveroista lasketaan mukaan vain henkilöiden maksama osuus. Tämä tarkoittaa koko Suomen väestön osalta keskimäärin noin 530 euron vuosittaisen verokertymän puuttumista henkilöä kohti, mikä tarkoittaa elinkaaren yli keskimäärin noin 40 000 euron puuttumista henkilöiden maksamista veroista. Todellisuudessa tämä luku on pienempi, sillä kiinteistöveron lisäksi myös muista mainituista veroista muut kuin henkilöt maksavat osan.
Puuttuvat tiedot vanhuusiän palveluiden asiakasmaksuista.	Ainakin osa vanhuusiän julkisista palveluista on palvelunsaajien itsensä maksamia. Esimerkiksi kotihoidonpalveluissa palvelunsaaja maksaa osan palvelun maksusta. Tämä osuus riippuu sekä palvelunsaajan tuloista että kunnasta. Sama koskee myös palvelu- ja laitosasumista. En ainakaan toistaiseksi sisällytä elinkaariennusteisiin ennustetta siitä, kuinka suuren osan vanhuusiän palveluista palvelunsaajat maksavat itse. Tämä ennuste riippuisi henkilön ennustetuista eläketuloista. **Näiden asiakasmaksujen puute on kenties suurin julkisen talouden vaikutusten puute elinkaariennusteissa.**
3. Puuttuvat yhteisöjen ja yhtymien vaikutukset	
Henkilöverotuksen kautta elinkaariennusteisiin *sisältyvät* vaikutukset.	Henkilöiden elinkeinonharjoittajina ja yhtymien kautta saadut tulot verotetaan henkilöiden oman ansio- ja pääomatuloverotuksen kautta. Nämä tiedot sisältyvät käytössä olevaan rekisteriperustaiseen tutkimusaineistoon – ja siten myös elinkaariennusteisiin.
Yhteisöverot eivät sisälly elinkaariennusteisiin tutkimuksen määritelmän vuoksi.	Mikäli yhteisöverot sisällytettäisiin mukaan elinkaariennusteisiin, niin niiden vaikutukset riippuisivat yhteisöverokertymän jakamisesta henkilöiden kesken. Ei ole yksiselitteistä tapaa tehdä tämä allokointi. Suurin osa yhteisöveroista on noin 500 suurimman yrityksen maksamia. Mikäli yhteisöverokertymä jaettaisiin omistusten mukaan, niin vaikutus elinkaariennusteisiin riippuisi siitä, miten henkilöt omistavat kaikkien suurimpia yrityksiä.

Jatkuu seuraavalla sivulla

Taulukko 10.1 – jatkuu edelliseltä sivulta

Puuttuva vaikutus	Lyhyt yhteenveto puuttuvasta vaikutuksesta
Yhteisöjen maksamat välilliset verot ja kiinteistöverot eivät sisälly elinkaariennusteisiin tutkimuksen määritelmän vuoksi.	Mikäli yhteisöjen maksamat muut verot kuin yhteisöverot sisällytettäisiin mukaan elinkaariennusteisiin, niin niiden vaikutukset riippuisivat näiden verojen kertymän jakamisesta henkilöiden kesken. Ei ole yksiselitteistä tapaa tehdä tämä allokointi. Mikäli jako tehtäisiin esimerkiksi työllisyyden tai ansiotulojen perusteella, niin Suomessa syntyneen kantaväestön elinkaariennusteet muuttuisivat positiivisemmiksi verrattuna sellaisiin maahanmuuttajaryhmiin, joissa työllisyys ja ansiotulot ovat keskimäärin matalia.
Yritystuet eivät sisälly elinkaariennusteisiin tutkimuksen määritelmän vuoksi. Henkilöiden saamat starttirahat sen sijaan ovat mukana.	Mikäli yritystuet sisällytettäisiin mukaan elinkaariennusteisiin, niin niiden vaikutukset riippuisivat yritystukien jakamisesta henkilöiden kesken. Ei ole yksiselitteistä tapaa tehdä tämä allokointi.

Luku 11

Elinkaariennusteiden diskonttaus nykyarvoon

11.1 Ennustettujen julkisen talouden nettovaikutusten nettonykyarvo

Tutkimuksen seuraavassa luvussa päätöksentekoteoriaa sovelletaan potentiaalisten uusien maahanmuuttajien valintaan Suomeen muuttaviksi. Tarkoituksena on valita – tai jättää valitsematta – Suomeen muuttaviksi potentiaaliset uudet maahanmuuttajat siten, että julkisen talouden odotettavissa oleva hyöty maksimoituu potentiaalisten uusien maahanmuuttajien valinnan tuloksena. Kuten esitän seuraavassa luvussa, määritän julkisen talouden hyödyn uudesta maahanmuuttajasta tämän julkisen talouden euromääräisenä elinkaarivaikutuksena. Päätös uuden potentiaalisen maahanmuuttajan valitsemisesta Suomeen muuttavaksi perustuu siihen, mikä on uuden maahanmuuttajan ennustetun julkisen talouden elinkaarivaikutusten (todennäköisyysjakauman) odotusarvo (ks. tarkemmin seuraava luku päätöksentekoteorian soveltamisesta uusien maahanmuuttajien valinnassa).

Päätöksenteko perustuu elinkaarivaikutusten odotusarvojen *päätöksentekohetken* euromääräiseen arvoon. Tätä varten tulevat (ennustetut) euromääräiset elinkaarivaikutukset on muunnettava päätöksentekohetken – jota kutsuttakoon *nykyhetkeksi* – arvoon. Koska elinkaariennusteet muodostuvat elinkaaren yli yhteenlasketuista julkisen talouden nettovaikutusten vuosittaisista ennusteista, käytännössä jokainen vuosittainen euromääräinen ennuste on muunnettava nykyhetken arvoon. Tulevien ennustettujen julkisen talouden nettovaikutusten nykyarvojen summaa kutsutaan nettovaikutusten *nettonykyarvoksi* (*engl. net present value* (Clemen and Reilly 2014, 34–35) tai *present discounted value* (Stiglitz 2000, 273)).

Tulevat nettovaikutukset diskontataan nykyarvoon *diskonttauskertoimilla* (*engl. discount factor*). Tulevien nettovaikutusten tapauksessa jokaiselle tulevalle vuodelle määritetään tai ennustetaan oma diskonttauskertoimensa. Vuosittaiset diskonttauskertoimet riippuvat kahdesta eri tekijästä, joiden funktioita ne ovat. Ensinnäkin inflaation vuoksi tulevien nettovaikutusten nimellisarvot ovat suurempia kuin niiden nykyarvot. Toiseksi nykyraha on (useimmiten) kalliimpaa kuin tuleva raha – toisin sanoen yksi euro tänään on arvokkaampi kuin yksi euro huomenna. Nimelliskorko määrittää (tässä yhteydessä), kuinka

paljon arvokkaampi raha on tiettynä vuonna verrattuna sitä seuraavaan vuoteen. Käytän tässä yhteydessä nimelliskorkoa reaalikoron sijaan, sillä näin sekä hintojen nousu että korot ilmaistaan nimellisarvoisina.

Jos kalenterivuoden t hintojen nousu on $inflaatio(t)$, ja tuon kalenterivuoden 12 kk:n nimelliskoron taso on $nimelliskorko(t)$, niin kyseisen kalenterivuoden t diskonttauskertoimen määritelmä on tällöin:

$$Diskonttauskerroin(t) := \frac{1 + inflaatio(t)}{1 + nimelliskorko(t)}. \tag{11.1}$$

Olkoon $V = \{0, 1, \ldots, T-1, T\}$ niiden kalenterivuosien joukko, joille nettovaikutusennusteet laaditaan ja joiden yhteenlaskettu summa halutaan ilmaista nykyarvossa. Yksittäinen kalenterivuosi, jolle ennuste laaditaan, on t siten että $t \in V$. Kalenterivuosi on nykyinen/ennustehetken kalenterivuosi silloin, kun $t = 0$. Kalenterivuoden t ennustetun julkisen talouden nettovaikutuksen, $nettovaikutus(t)$, nykyarvon määritelmä on tällöin:

$$NA(t) := nettovaikutus(t) \times \prod_{i=1}^{t} \frac{1 + inflaatio(i)}{1 + nimelliskorko(i)}, \tag{11.2}$$

missä $NA(t)$ on kalenterivuoden t ennustetun julkisen talouden nettovaikutuksen nykyarvo.

Kun kalenterivuosi $t = 0$, eli on kyse nykyisestä kalenterivuodesta, diskonttauskerrointa ei lasketa. Mikäli hinnat nousevat kalenterivuonna $t = 1$ edellisvuodesta esimerkiksi 2 %:ia ja nimelliskorko on tuona vuonna esimerkiksi 3 %:ia, tällöin $Diskonttauskerroin(1) = (1 + 0,02)/(1 + 0,03) \approx 0,99$. Tällöin tuon kalenterivuoden ennustetun nettovaikutuksen ($nettovaikutus(1)$) nykyarvo on: $NA(1) = nettovaikutus(1) \times (1 + 0,02)/(1 + 0,03) \approx nettovaikutus(1) \times 0,99$. Koska nimelliskorko on useimmiten inflaatiota suurempi, ennustettujen nettovaikutusten diskontatut *nykyarvot* ovat useimmiten pienempiä kuin niiden ennustetut diskonttaamattomat nimellisarvot.

Kaikkien tulevien kalenterivuosien $\{0, 1, \ldots, T-1, T\}$ nykyarvojen summan eli nettonykyarvon NNA (*engl. NPV*) määritelmä on edellä olevan perusteella:

$$NNA := \sum_{t=0}^{T} \left(nettovaikutus(t) \times \prod_{i=1}^{t} \frac{1 + inflaatio(i)}{1 + nimelliskorko(i)} \right). \tag{11.3}$$

Kun yksittäiset nettovaikutusennusteet koskevat yhtä henkilöä, ja kalenterivuodet $V = \{0, 1, \ldots, T-1, T\}$ käsittävät kaikki kalenterivuodet tämän henkilön ensimmäisen maahanmuuton kalenterivuodesta hänen kuolemaansa asti, tällöin ennustettujen nettovaikutusten nettonykyarvo NNA on kyseisen henkilön elinkaarivaikutuksen ennusteen nykyarvo.

Ennen kuin käsittelen diskonttauskertoimien mallinnusta ja mallinnuksen pohjalta laadittavia diskonttauskertoimien ennusteita, esittelen havaintoaineistot, joihin mallinnus perustuu.

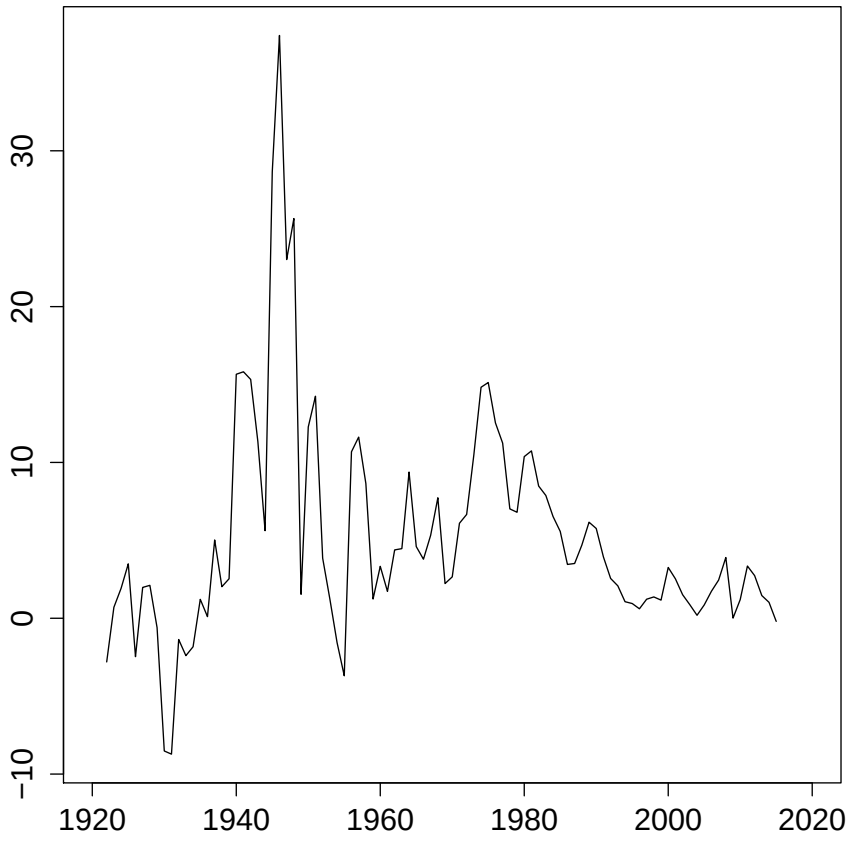
11.2 Diskonttauskertoimien havaintoaineisto

11.2.1 Rahanarvonkertoimen vuosimuutokset 1922–2015

Diskonttauskertoimen vuosittaisten arvojen muodostaman osamäärän osoittaja on $1 + inflaatio$, missä inflaatio on hintojen muutos edellisestä kalenterivuodesta. Inflaation yleisenä mittarina käytetään Tilastokeskuksen julkaisemaa *kuluttajahintaindeksiä*, joka *"[...] kuvaa Suomessa asuvien kotitalouksien Suomesta ostamien tavaroiden ja palveluiden hintakehitystä."*. Tilastokeskuksen mukaan kuluttajahintaindeksit sopivat kuitenkin vain lyhyen aikavälin tarkasteluihin hintojen muutoksissa. Edelleen Tilastokeskuksen mukaan *elinkustannusindeksi "[...] on uusimmasta kuluttajahintaindeksistä [...] ketjuttamalla laskettu pitkä aikasarja ja se kehittyy siten samalla tavalla kuin kuluttajahintaindeksi."*. Se on "käytännöllisin" mittari tutkia hintojen kehitystä, sillä *"[...] indeksiuudistukset eivät katkaise sarjaa ja elinkustannusindeksin pisteluvut julkaistaan kuukausittain samaan aikaan kuin kuluttajahintaindeksi."*. (Tilastokeskus 2017b).

Tilastokeskuksen julkaisema (SVT: *rahanarvonkerroin*) perustuu vuosien 1921–2015 osalta viralliseen elinkustannusindeksiin. Koska elinkustannusindeksi on "käytännöllisin" mittari hintojen kehityksen seuraamiseen pitkällä aikavälillä, valitsin vuosittaiset rahanarvonkertoimen muutokset havaintoaineistoksi kuvaamaan inflaatioon suuruutta vuosina 1922–2015.

Alla kuvassa 11.1 on esitetty rahanarvonkertoimien prosentuaaliset vuosimuutokset vuosina 1922–2015. Esimerkiksi vuoden 2015 arvo aikasarjassa kertoo, kuinka monta prosenttia on rahanarvonkertoimen muutos vuodesta 2014 vuoteen 2015 suhteessa vuoden 2015 rahanarvonkertoimeen. Vuoden 2014 arvo puolestaan kertoo, kuinka monta prosenttia on rahanarvonkertoimen muutos vuodesta 2013 vuoteen 2014 suhteessa vuoden 2014 rahanarvonkertoimeen, jne. Suurimmat rahanarvon vuosimuutokset ovat tapahtuneet sota-aikaan ja heti sotien jälkeen 1940-luvulla. 2000-luvulla kaikki vuosimuutokset ovat olleet alle 4 %:ia, ja useana vuonna lähellä nollaa. Vuonna 2015 suhteellinen muutos edellisestä vuodesta oli noin -0,2 %:ia.



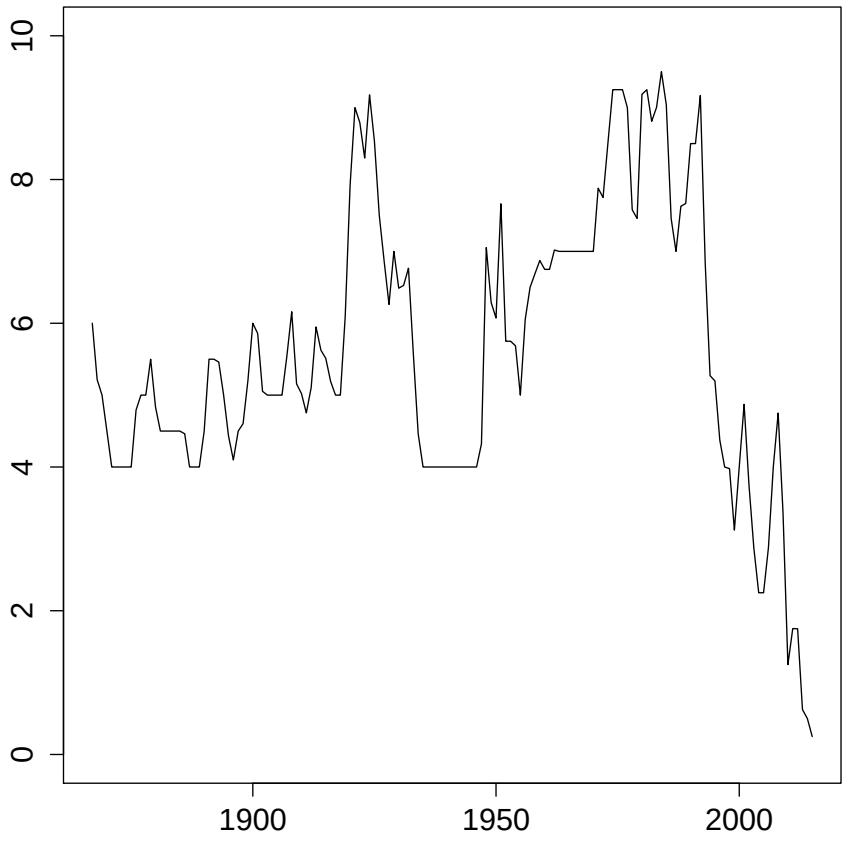
Kuva 11.1: *Rahanarvonkertoimen vuosittaiset muutokset 1922–2015 (%). Lähde: SVT: rahanarvonkerroin.*

11.2.2 Suomen Pankin peruskorko vuosina 1867–2015

Suomen Pankin peruskorko on 12 kuukauden markkinaperusteinen korko, jonka tiedot ovat saatavilla vuodesta 1867 lähtien. Suomen Pankin kuvaus peruskorosta on:

”Peruskorko on hallinnollinen korko, jonka suuruuden määräsi vuoden 1998 loppuun asti Suomen Pankki. Vuoden 1999 alusta alkaen peruskoron vahvistaa valtiovarainministeriö vuosittain kesäkuussa ja joulukuussa seuraavaksi puoleksi kalenterivuodeksi.”. (Suomen Pankki 2017).

”Peruskorko lasketaan sen vahvistamista edeltäneiden kolmen kalenterikuukauden aikana julkaistun 12 kuukauden markkinakoron perusteella. Peruskorko on mainittujen kolmen kuukauden ajalta laskettu markkinakoron keskiarvo pyöristettynä lähimpään prosenttiyksikön neljännekseen. Markkinakorko on toistaiseksi 12 kuukauden euriborkorko (365 päivää).”. (Suomen Pankki 2017).



Kuva 11.2: *Suomen Pankin peruskorko vuosina 1867–2015 (%). Lähde: Suomen Pankki 2017.*

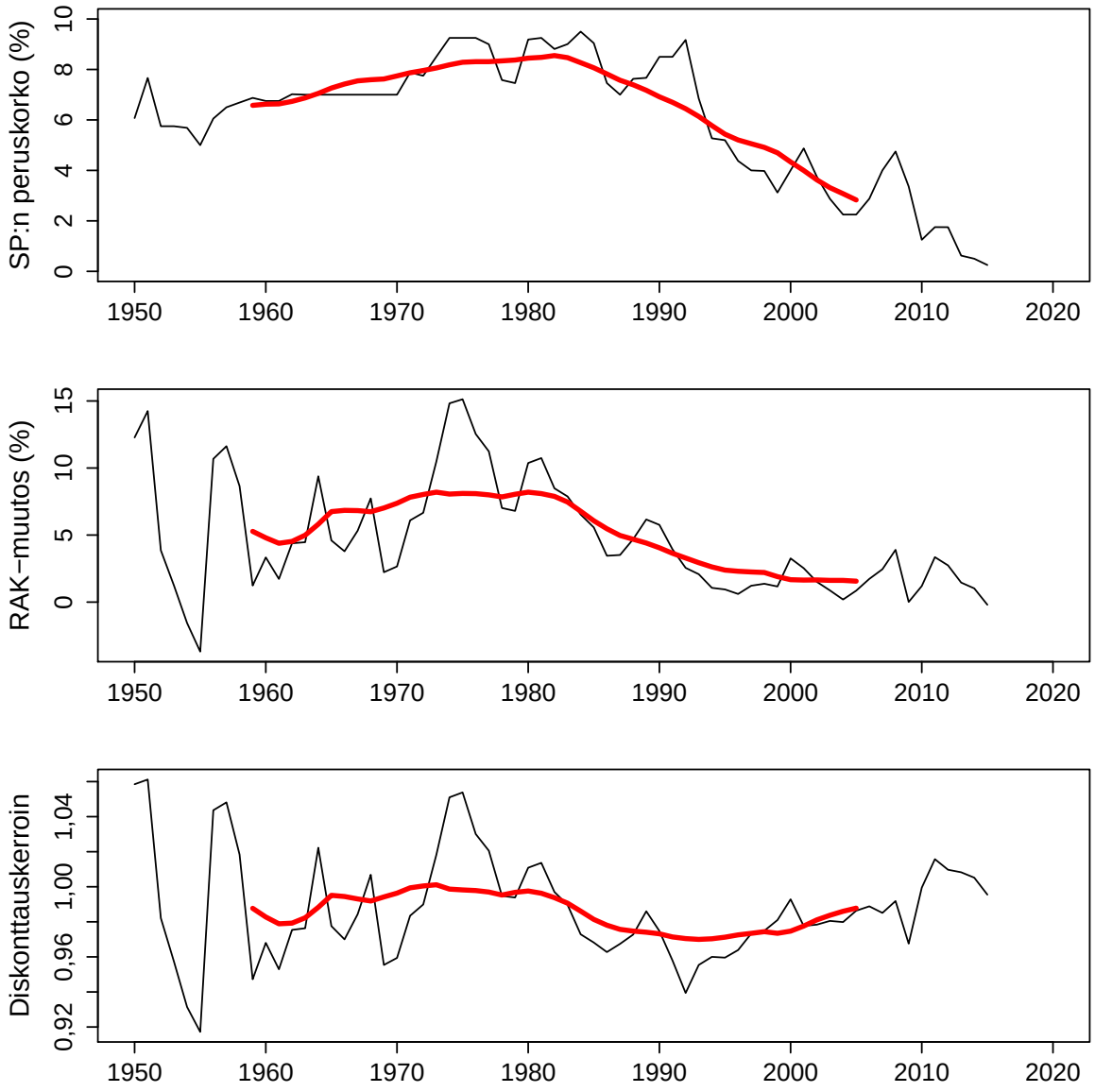
Kuvassa 11.2 on esitetty Suomen Pankin peruskoron keskimääräinen vuosittainen taso vuosina 1867–2015. Mikäli korkotaso on vaihdellut kalenterivuoden sisällä, olen laskenut keskimääräisen tason paino-

tettuna keskiarvona eri korkotasoista. Painoina käytin niiden päivien lukumäärää, jolloin kukin korkotaso on vallinnut yhden kalenterivuoden aikana. Peruskoron keskimääräinen vuositaso on vaihdellut vuoden 1867 ja 1900-luvun lopun muodostaman ajanjakson aikana noin 4 %:n ja 9,5 %:n muodostaman haarukan välillä. 2000-luvulla – ja erityisesti 2010-luvulla – peruskoron taso on ollut alhaisemmalla tasolla kuin milloinkaan vuodesta 1867 lähtien. Vuonna 2015 peruskoron taso oli 0,25 %:ia, ja vuonna 2016 sekä vuoden 2017 alussa (ei kuviossa) peruskoron taso oli tasan 0 %:ia.

11.2.3 Diskonttauskertoimet vuosina 1950–2015

Edellä vuosittaisten diskonttauskertoimien on määritetty olevan osamäärä: $\frac{1+inflaatio}{1+nimelliskorko}$. Käytän diskonttauskertoimien ennusteiden pohjalla olevan diskonttauskertoimien mallinnuksen aineistona havaittuja, menneitä vuosittaisia diskonttauskertoimia. Käytän mallinnukseen vain vuosien 1950–2015 diskonttauskertoimia, sillä 1940-luvun diskonttauskertoimet ovat poikkeuksellisen korkeita verrattuna vuosien 1950–2015 diskonttauskertoimiin. Mikäli olisin sisällyttänyt myös 1940-luvun diskonttauskertoimet mukaan mallinnukseen, ennustettavista diskonttauskertoimista olisi tullut suurempia, ja siten tulevien elinkaarivaikutusten (diskontattujen) ennusteiden nykarvosta kauempana nollasta olevia, kuin ne nyt ovat elinkaariennusteiden diskontatuissa tuloksissa. Jätin samalla pois vuodet ennen vuotta 1940. Myös nämä vuodet kuitenkin tarjoavat varsinaisen mallinnusaineiston ulkopuolelta sellaista informaatiota, jonka mukaan 1940-lukua lukuun ottamatta vuosittaiset diskonttauskertoimet ovat sijoittuneet 0,90:n ja 1,10:n väliin.

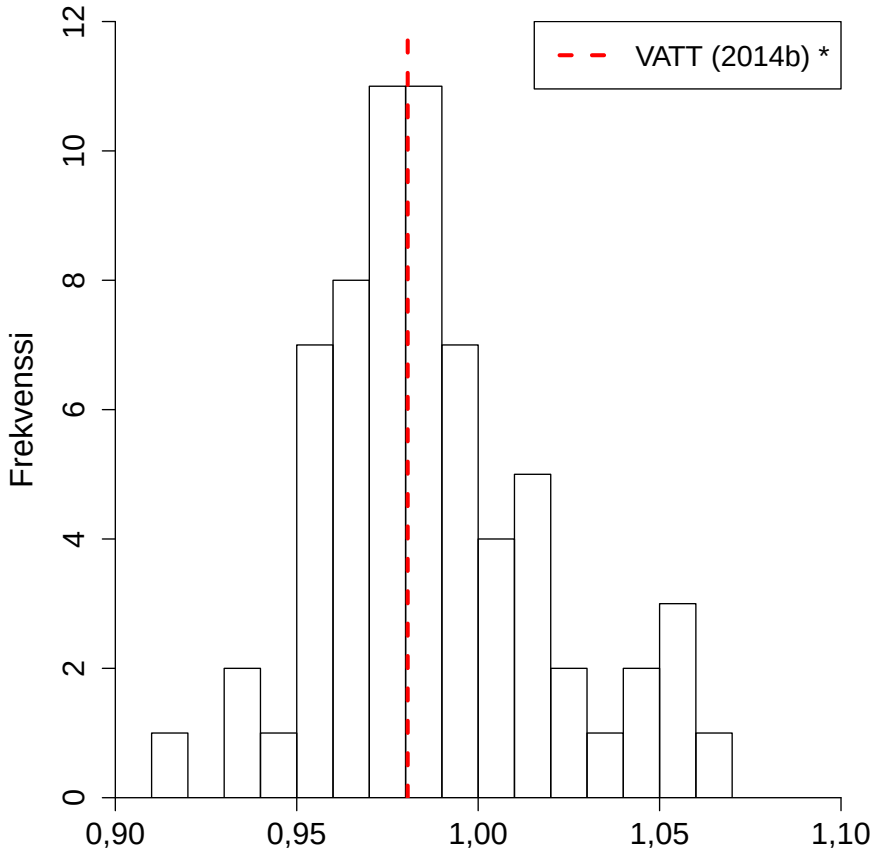
Alla kuviossa 11.3 on esitetty Suomen Pankin peruskoron ja rahanarvonkertoimen muutosten aikasarjat vuosille 1950–2015 sekä niiden pohjalta laskettujen diskonttauskertoimien aikasarja samalle ajanjaksolle. Diskonttauskertoimien aikasarja seuraa 1950-luvulta 1990-luvulle asti pitkälti rahanarvonkertoimen muutosten aikasarjaa. Tämä johtuu siitä, että peruskorko pysyi melko samanlaisena 1950-luvulta 1990-luvulle. Tämän jälkeen peruskoron vuositaso on kuitenkin laskenut nykyiseen nollaan. Tämä kehitys on nostanut diskonttauskertoimien arvot 2010-luvulla ykkösen ympäristöön.



Kuva 11.3: *Suomen Pankin (SP) peruskorko, rahanarvonkertoimen (RAK) vuosittaiset muutokset sekä diskonttauskertoimet vuosina 1950–2015. Punaiset viivat ovat liukuvia keskiarvoja 10 vuotta menneisyyteen ja tulevaisuuteen. Lähteet: Suomen Pankki 2017 ja SVT: rahanarvonkerroin.*

Alla kuviossa 11.4 on esitetty vuosien 1950–2015 diskonttauskertoimien jakauma histogrammina. Histogrammin perusteella diskonttauskertoimien moodi on ollut hieman alle yksi vuosina 1950–2015. Huomionarvoista mallintamisen kannalta histogrammissa on se, että sen perusteella diskonttauskertoimien havaittu jakauma muistuttaa normaalijakaumaa.

Kuvioon on myös merkitty VATT:n (2014b) tutkimuksen käyttämä kiinteä oletus tulevien nettovaikutusten diskonttauskertoimelle – mikäli oletetaan, että VATT:n tutkimuksen kiinteä hintojen nousun taso ja korkotaso vastaavat rahanarvonkertoimen muutoksia ja Suomen Pankin peruskorkoa. Olen tässä yhteydessä olettanut, että VATT:n (2014b) tutkimuksen liitteessä esitetty ”reaalikoron” taso on oikeasti nimelliskorko. Näin diskonttauskertoimessa on sekä osoittajassa että nimittäjässä nimellisarvoinen tekijä, ja näin myös oletettu kiinteä diskonttauskertoimen arvo on lähellä havaittujen kertoimien huippua, kuten kuvio 11.4 osoittaa.



Kuva 11.4: *Diskonttauskertoimet vuosina 1950–2015. Frekvenssi ilmaisee vuosien määrän. * VATT (2014b) on kyseisessä tutkimuksessa käytetty oletus tuleville diskonttauskertoimille, mikäli oletetaan, että VATT:n tutkimuksen kiinteä hintojen nousun taso ja korkotaso vastaavat rahanarvonkertoimen muutosta ja Suomen Pankin peruskorkoa. Lähteet: Suomen Pankki 2017 ja SVT: rahanarvonkerroin.*

Yllä olevan diskonttauskertoimien aikasarjakuvion perusteella näyttää siltä, että diskonttauskertoimien arvot ovat vaihdelleet suunnilleen saman keskiarvotason ympärillä koko aikasarjan yli. Tämä tarkoittaa sitä, että diskonttauskertoimella ei ole ollut ainakaan vuosina 1950–2015 mitään pitkän aikavälin trendiä.

Tähän viittaavat myös alla taulukossa 11.1 esitetyt periodikohtaiset keskiarvot. Taulukon toisella rivillä periodikohtaiset keskiarvot on esitetty suhteessa koko ajanjakson 1950–2015 keskiarvoon. Kuten taulukon luvut kertovat, periodikohtaiset keskiarvot eivät ole poikenneet pitkän ajan keskiarvosta, kun tarkastellaan noin 20 vuoden periodeita. Tällä vuosikymmenellä eli 2010-luvulla keskiarvo on ollut korkeampi, mutta en tee siitä mitään pitkän aikavälin johtopäätöksiä diskonttauskertoimen mallinnuksen ja ennustamisen yhteydessä.

Taulukossa 11.1 on esitetty myös periodikohtaiset diskonttauskertoimien havaitut keskihajonnat sekä näiden periodikohtaisten keskihajontojen suhde koko aikavälin 1950–2015 havaittuun keskihajontaan. Toisin kuin havaittujen periodikohtaisten keskiarvojen kohdalla, keskihajontojen osalta havaitut keskihajonnat ovat pienentyneet suuruudeltaan ajan myötä, eli näiden lukujen perusteella diskonttauskertoimien keskihajonnalla saattaa olla ollut pitkän aikavälin laskeva trendi. Tämä tarkoittaa sitä, että mitä kauemmaksi menneisyyteen vuodesta 2015 mennään kohti vuotta 1950, sitä suurempia ovat olleet vuosittaiset diskonttauskertoimien keskimääräiset poikkeamat diskonttauskertoimien pitkän aikavälin keskiarvosta.

Taulukko 11.1: *Diskonttauskertoimien periodikohtaiset keskiarvot ja keskihajonnat vuosina 1950–2015. Periodikohtaisten keskiarvojen ja keskihajontojen suhteet koko ajanjakson lukuihin. Lähteet: Suomen Pankki 2017 ja SVT: rahanarvonkerroin.*

	1950–2015	**1950–1971**	**1972–1993**
Keskiarvo periodin aikana	0,988	0,986	0,992
Suhteessa v. 1950–2015 keskiarvoon	1	0,998	1,004
Keskihajonta periodin aikana	0,030	0,041	0,030
Suhteessa v. 1950–2015 keskihajontaan	1	1,34	1,00
	1994–2005	**2006–2015**	**2011–2015**
Keskiarvo periodin aikana	0,985	0,997	1,007
Suhteessa v. 1950–2015 keskiarvoon	0,997	1,009	1,019
Keskihajonta periodin aikana	0,016	0,014	0,007
Suhteessa v. 1950–2015 keskihajontaan	0,53	0,47	0,24

11.3 Diskonttauskertoimien mallinnus ja ennustaminen

11.3.1 Diskonttauskertoimien mallinnus

Diskonttauskertoimien mallinnustavoitteena on laatia diskonttauskertoimille sellainen aikasarjamalli, jonka pohjalta voidaan tehdä uskottavia ennusteita tuleville diskonttauskertoimille. Koska diskonttauskertoimien aineisto ja tulevat arvot muodostavat aikasarjan, etsin sopivaa mallia aikasarja-analyysin piiristä. Koska aikasarja-aineisto on lyhyt (66 havaintoa), malliparametreihin – ja siten ennusteisiin – sisältyy vielä paljon epävarmuuksia estimoinnin jälkeen. Koska käytän mallinnuksessa bayesilaista päättelyä, kuten muutenkin tässä tutkimuksessa, voin kuitenkin sisällyttää mallin pohjalta tehtäviin ennusteisiin tämän epävarmuuden. Lisäksi bayesilaisten menetelmien käyttö mahdollistaisi mallinnusaineiston ulkopuolisen informaation sisällyttämisen malliin ja ennusteisiin, mutta jätän tämän mahdollisuuden käyttämättä. Koska haluan ennusteiden sisältävän kuitenkin mieluummin ”liikaa” epävarmuutta kuin liian vähän, käytän diskonttauskertoimien ennusteiden osalta suurempaa varianssia eli vaihtelua, kuin mallinnusaineisto osoittaisi koskien viimeistä noin 20–30 vuotta.

Käytin diskonttauskertoimien mallintamiseen ”tavanomaisia” yksiulotteisten aikasarjojen sekä ehdollisen varianssin (*engl. conditional heteroskedasticity, CH*) mallintamiseen soveltuvia menetelmiä. En esittele menetelmiä tarkemmin, sillä niistä löytyy paljon kirjallisuutta, esimerkiksi Lütkepohl and Krätzig (2004). Tavanomaisilla yksiulotteisten aikasarjojen menetelmillä viittaan nk. *ARIMA*-malleihin. Koska diskonttauskertoimilla ei havaintoaineiston perusteella ole trendiä ajassa, tiputin mainitusta kirjainyhdistelmästä ”I”:n pois jo ennen, kuin kokeilin erilaisia aikasarjamalleja. Mallinvalintakriteerinä käytin edellä tässä tutkimuksessa (luvussa 4.3.2.3) esiteltyjä ennustetarkkuuden mittareita WAIC:ia ja LOO:ta.

Aloitin mallinnuksen ”tyhjästä” mallista, joka ei sisältänyt lainkaan prosessin (aikasarjan) menneisiin havaintoihin viittaavia parametreja. Se oli siis tavallinen normaalijakaumaan perustuva regressiomalli, jossa oli mukana vain vakiotermi. Tämän jälkeen laajensin mallia AR- ja MA-termeillä kasvattaen näiden asteita aina yhden kerrallaan. Diskonttauskertoimien aikasarjojen autokorrelaatio- ja osittaisautokorrelaatiodiagrammit viittasivat aikasarjan taustalla olevan stokastisen prosessin olevan joko AR(1)-, MA(1)- tai ARMA(1,1)-prosessi. Estimoin näihin malleihin perustuvat malliparametrien posteriorijakaumat Stan-ohjelmalla, jonka esittely löytyy aiemmin tästä tutkimuksesta luvusta 4.3.6.1, ja Stan-ohjelman ohjeet aikasarjamallinnuksesta esittelee puolestaan: Stan Development Team 2015, osa II, luku 7.

Käytin yleisesti ottaen kaikkien mallien kaikkien parametrien kohdalla melko epäinformatiivisia priorijakaumia. Tässä tapauksessa tosin mallin ulkopuolista informaatiota esimerkiksi korkojen pitkän aikavälin kehityksestä olisi saatavilla (esimerkiksi: Council of Economic Advisers 2015). Kokeilin myös AR(2)-mallia. Waic- ja LOO-kriteereiden perusteella valitsin malliksi kuitenkin AR(1)-mallin. Höllensin myös prosessin normaalijakaumaoletusta käyttämällä normaalijakauman sijaan Studentin t-jakaumaa sekä vinoa normaalijakaumaa. Luovuin kuitenkin Studentin t-jakaumasta – eli käytännössä jakauman vapausasteparametrin lisäämisestä malliin – havaintojen vähyyden vuoksi. Vino normaalijakauma oli mallinvalintakriteereiden mukaan yhtä hyvä kuin tavallinen normaalijakauma. Päätin käyttää pienempää mallia, eli tavalliseen normaalijakaumaan perustuvaa AR(1)-mallia.

Ar(1)-mallissa nykyisten havaintojen odotusarvo riippuu prosessin edellisestä havainnosta. Tämä ei kui-

tenkaan huomioi vielä normaalijakautuneen prosessin varianssin (tai keskihajonnan) mahdollisia riippuvuussuhteita yli ajan. Tästä syystä laajensin mallin käsittämään vielä mallin varianssin (tai keskihajonnan) mahdollisen riippuvuuden ajassa. Kokeilin nk. GARCH(0,1)- ja GARCH(1,0)-malleja yhdessä AR(1)-mallin kanssa. Näissä malleissa normaalijakautuneen prosessin jokaisen havaintoajanhetken varianssi riippuu joko ajassa edeltävän havainnon poikkeamasta prosessin odotusarvosta tai edellisen havainnon varianssista. Käytin näiden mallien ennustetarkkuuden kriteerinä edelleen WAIC:ia ja LOO:ta. GARCH-parametrien lisääminen malliin ei näiden kriteereiden perusteella parantanut mallien ennustetarkkuutta, joten en sisällyttänyt lopulliseen diskonttauskertoimien malliin näitä parametreja.

Sen sijaan diskonttauskertoimien mallin ennustetarkkuutta paransi keskihajonnan trendin lisääminen malliin (ennustetarkkuus ei parantunut, vaikka lisäsin vielä mainitut GARCH-parametrit malliin). Keskihajonnan trendi tarkoittaa sitä, että prosessin keskihajonta muuttuu lineaarisesti ajassa. Havaintoaineiston tarkastelun perusteella diskonttauskertoimien keskihajonta on pienentynyt vuosien mittaan vuosien 1950–2015 aikana, joten tämä viittaisi siihen, että keskihajonnalla olisi negatiivinen trendi. Näin vaikuttaa olevan myös alla esitettävien estimointitulosten perusteella.

Olkoon mallinnusaineiston kalenterivuosien joukko $K = \{1950, 1951, \ldots, 2014, 2015\}$. Lopullisesta diskonttauskertoimien (aineisto)mallista muodostui yllä esitetyn testaamisen perusteella AR(1)-malli, jonka keskihajonnalla on laskeva trendi. Normaalijakautunut aineistomallin otosjakauma voidaan kirjoittaa näin ollen:

$$y_t \sim N(vakio + \rho y_{t-1}, \sigma(t)^2) \tag{11.4}$$

ja

$$\sigma(t) = \sigma(2015) + \sigma_{trendi} \times (t - 2015), \tag{11.5}$$

missä y_t on diskonttauskertoimien aikasarjan havainto vuonna t, $vakio$ on mallin vakio, ρ on autoregressiivinen kerroin, y_{t-1} on aikasarjan havainto vuonna $t-1$ ja keskihajonta vuonna t on puolestaan $\sigma(t)$. Kalenterivuoden t ja vuoden 2015 välinen erotus, $t - 2015$, saa arvoksi negatiivisia kokonaislukuja lukuun ottamatta vuotta 2015, jolloin sen arvon on nolla. Keskihajonnan vertailu- tai vakiotasona toimii keskihajonnan taso vuonna 2015, $\sigma(2015)$, ja σ_{trendi} on keskihajonnan trendi, joka ilmaisee, kuinka paljon keskihajonta keskimäärin muuttuu vuodessa.

Käytin estimoinnissa standardoitua versiota mallinnettavasta aikasarjasta. Aineistomallin estimoitavilla parametreilla oli seuraavat melko epäinformatiiviset priorijakaumat:

$vakio \sim N(0, 1)$.

$\rho \sim \mathcal{U}(-1, 1)$.

$sigma(2015) \sim Cauchy(0, 10)$.

$\sigma_{trendi} \sim \mathcal{U}(-1, 0)$.

Rajoitin σ_{trendi}:n saamaan vain ei-positiivisia arvoja, jotta $\sigma(t)$ olisi aina positiivinen. Mikäli σ_{trendi}:n posteriorijakauma olisi keskittynyt nollan lähelle, olisin vaihtanut kokonaan mallia.

Bayesilaisessa päättelyssä parametrien yhteisposteriorijakauma on verrannollinen (aineistomallin) otosjakauman ja estimoitavien parametrien priorijakauman väliselle tulolle. Alla taulukossa 11.2 on esitetty parametrikohtaiset posteriorijakaumien estimaatit diskonttauskertoimien mallille.

Taulukko 11.2: *Diskonttauskertoimien malli. Parametrien posteriorijakaumien odotusarvojen (OA), keskihajontojen (KH) ja prosenttipisteiden estimaatit sekä odotusarvoestimaattien MC-keskivirheet (MC-kv) ja parametrikohtaiset simulaatioketjujen konvergenssien testisuureet ($\hat{R}$).*

	OA	KH	2,5 %	25 %	50 %	75 %	97,5 %	*MC-kv*	$\hat{R}$
vakio	0,009	0,071	-0,138	-0,037	0,009	0,057	0,145	0,001	1,00
ρ	0,65	0,09	0,47	0,59	0,65	0,71	0,83	0,002	1,00
$\sigma(2015)$	0,28	0,06	0,18	0,24	0,27	0,32	0,43	0,002	1,00
σ_{trendi}	-0,013	0,003	-0,019	-0,015	-0,013	-0,012	-0,008	0,000	1,00

Käytin estimoinnissa standardoitua versiota aikasarjahavaintojen jakaumasta, eli muunsin havaintojen jakauman keskiarvon nollaan ja kokonaisvaihtelun ykkösen arvoiseksi. Tästä syystä taulukossa 11.2 $vakio$-parametrin posteriorijakauma on keskittynyt nollan ympärille. Autoregressiivisen kertoimen parametria ρ koskevat posteriorijakauman estimointitulokset tarkoittavat sitä, että perättäisillä diskonttauskertoimien arvoilla on vahva lineaarinen yhteys. Parametrin posteriorijakauma on suhteellisen leveä, mikä johtuu havaintojen vähyydestä sekä parametria koskevan priorijakauman epäinformatiivisuudesta.

Parametrin $\sigma(2015)$ posteriorijakauma kuvaa diskonttauskertoimien keskihajonnan tasoa vuonna 2015, kun autokorrelaation tuottama korrelaatio perättäisten havaintojen välille otetaan huomioon (ja kun aikasarja-aineiston kokonaisvaihtelu on standardoitu ykkösen suuruiseksi). Parametrin σ_{trendi} posteriorijakauma kuvaa sitä, kuinka paljon diskonttauskertoimien keskihajonta on pienentynyt keskimäärin vuodessa. Kyseisen parametrin estimoitu odotusarvon luku, $-0,013$, tarkoittaa sitä, että keskihajonta pienenee 1,3 %-yksikköä vuodessa suhteessa aikasarjan kokonaisvaihteluun. Aikasarjan kokonaisvaihtelu – jonka arvo on yksi standardoituna – muodostuu sekä $\sigma(t)$:n kuvaamasta satunnaisvaihtelusta että autokovarianssista havaintojen välillä. Autokovarianssin standardoitua arvoa edustaa ρ.

11.3.2 Diskonttauskertoimien ennustaminen

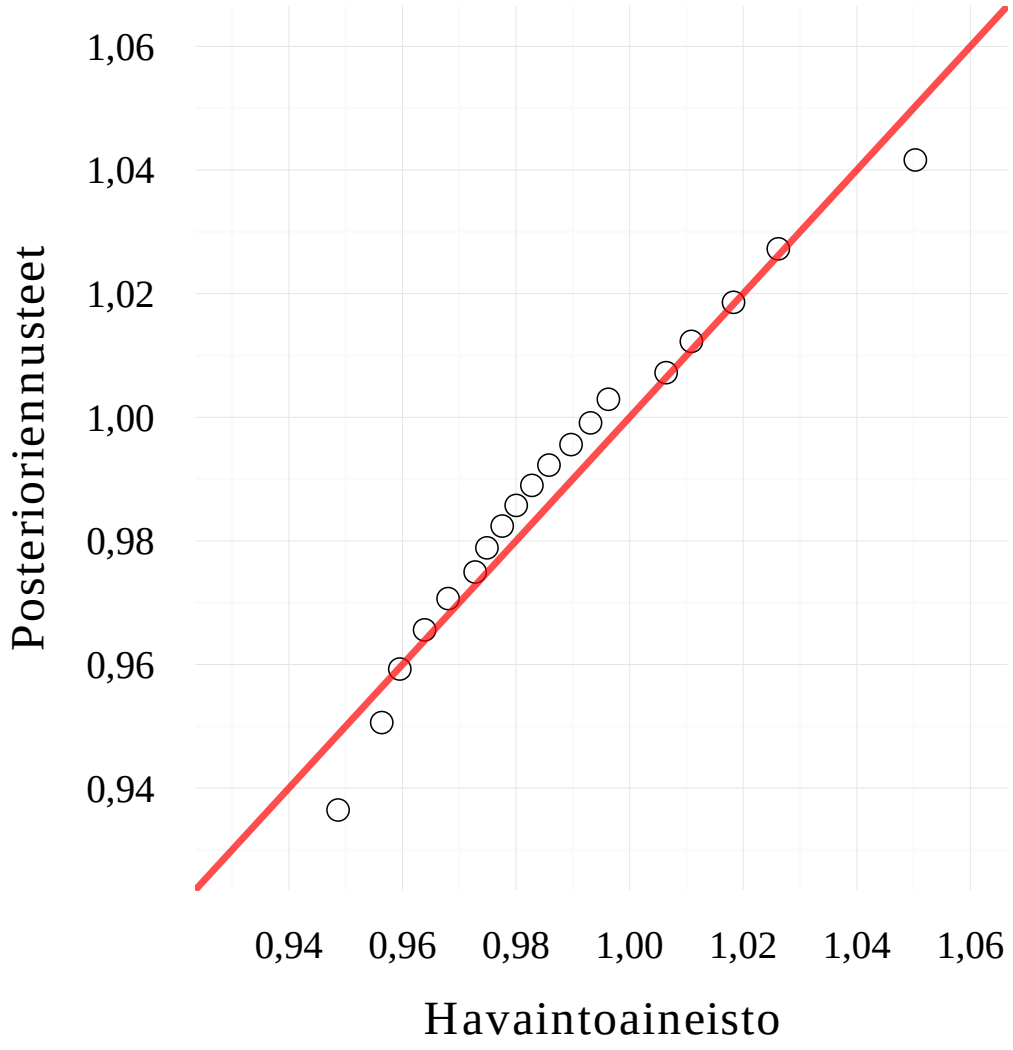
Tulevaisuuteen sijoittuvien elinkaariennusteiden diskonttauskertoimet poimitaan diskonttauskertoimien posterioriennustejakaumasta. En paneudu posterioriennustejakaumiin tässä kohtaa sen enempää, sillä olen käsitellyt aihetta aiemmin tässä tutkimuksessa (esimerkiksi olotila-aikojen mallinnuksen yhteydessä luvussa 3.2.5). Posterioriennustejakauman simulointi tapahtuu joka tapauksessa poimimalla simuloituja arvoja yllä määritetystä diskonttauskertoimien parametrien yhteisposteriorijakaumasta.

11.3.2.1 Posterioriennustetarkistus

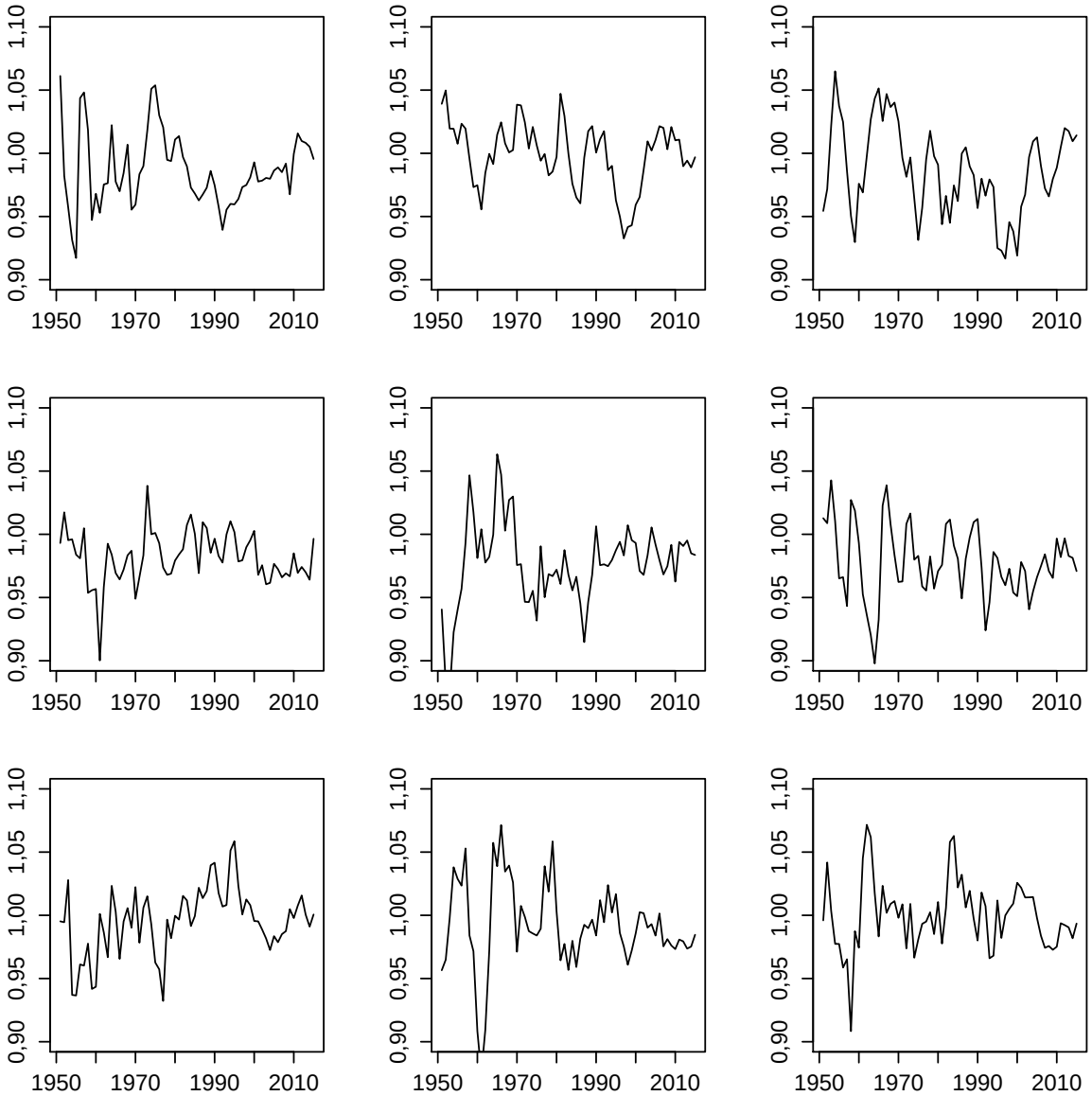
Posterioriennusteita voidaan käyttää mallin realistisuuden tarkistamiseen suhteessa havaintoaineistoon, tässä tapauksessa siis diskonttauskertoimien havaintoaineistoon vuosilta 1950–2015. Poimin posterioriennustejakaumasta 4 000 realisaatiota diskonttauskertoimien (vaihtoehtoisiksi) aikasarjoiksi vuosille 1950–2015. Vaikka käytin mallinnuksessa standardoitua versiota havaintoaineiston aikasarjasta, poimin posterioriennusteet alkuperäisen havaintoaineiston standardoimattomassa mittakaavassa. Mallia voi pitää realistisena siinä tapauksessa, että posterioriennusteiden jakauma muistuttaa havaintoaineiston jakaumaa. Posterioriennusteiden jakauman ja havaintoaineiston jakauman keskiarvo sekä keskihajonta ovat arvoiltaan lähellä toisiaan. Sekä posterioriennusteiden että havaintoaineiston keskiarvot ovat noin 0,988. Posterioriennusteiden keskihajonta on puolestaan noin 0,032, kun havaintoaineistossa se on noin 0,030.

Alla kuvassa 11.5 on esitetty hajontakuviossa posterioriennusteiden ja havaintoaineistojen jakaumien prosenttipisteiden vastaavuuden vertailu. Prosenttipisteet on esitetty kuvassa viiden prosenttipisteen välein. Vasemmanpuoleisin prosenttipiste on 5. prosenttipiste, ja oikeanpuoleisin on 95. prosenttipiste. Havaintoaineiston ja posterioriennusteiden jakaumat vastaavat sitä paremmin toisiaan, mitä lähempänä jakaumien samat prosenttipisteet ovat toisiaan. Täyttä vastaavuutta ilmaisee kuvioon piirretty punainen suora. Posterioriennusteiden prosenttipisteiden poikkeamat vastaavista havaintoaineiston prosenttipisteistä lasketaan pystyakselin suuntaisena erotuksena prosenttipisteen ja punaisen suoran pisteen (jolla on sama vaaka-akselin koordinaatti) välillä. Tämä jakaumien vertailutapa tunnetaan yleisesti myös englanninkielisen terminsä kautta: *Q-Q plot* (*Quantile-Quantile plot*). Kuvasta 11.5 voi päätellä, että posterioriennusteiden jakauma ja havaintoaineiston jakauma vastaavat ”suunnilleen” toisiaan. Havaintoaineiston pienuudesta johtuen mitään ennustettujen ja havaittujen prosenttipisteiden täydellistä vastaavuutta ei ollut odotettavissakaan.

Alla kuvassa 11.6 on esitetty vuosien 1951–2015 havaitut diskonttauskertoimet sekä kahdeksan satunnaisesti valittua aikasarjan realisaatiota diskonttauskertoimien posterioriennustejakaumasta. Pudotin vuotta 1950 koskevan havainnon pois, sillä posterioriennusteissa ensimmäinen ennuste on kiinnitetty havaittuun vuoden 1950 arvoon. Posterioriennustejakaumasta poimitut aikasarjat vaikuttavat kuvan perusteella samanlaisilta kuin havaintoaineiston aikasarja: Diskonttauskertoimet vaihtelevat ajassa suunnilleen saman tason ympärillä, ja vaihtelu on simuloiduissa aikasarjoissa suunnilleen yhtä suurta kuin havaitussa aikasarjassa. Lisäksi simuloitujen aikasarjojen vaihtelu pienenee ajassa, ja perättäiset diskonttauskertoimet vaikuttavat muistuttavan toisiaan arvoiltaan samoin kuin havaitussa aikasarjassa.



Kuva 11.5: *Posterioriennustetarkistus hajontakuviolla. Diskonttauskertoimien posterioriennusteiden jakauman vertailu havaintoaineiston jakaumaan. Posterioriennusteiden ja havaintoaineiston jakaumien prosenttipisteet* $5, 10, \ldots, 90, 95$. *Punainen suora ilmaisee täyttä vastaavuutta havaintoaineiston ja posterioriennusteiden jakaumien (prosenttipisteiden) välillä.*



Kuva 11.6: *Vuosien 1951–2015 havaitut diskonttauskertoimet (vasen yläkulma) sekä kahdeksan realisaatiota aikasarjoista, jotka on poimittu diskonttauskertoimien posterioriennustejakaumasta.*

11.3.2.2 Ennustettujen diskonttauskertoimien käyttö

Käytän poimintoja diskonttauskertoimien posterioriennustejakaumasta ennusteina tuleville diskonttauskertoimille niissä elinkaarivaikutusten ennusteissa, joissa tulevat nettovaikutusten ennusteet diskontataan nykyarvoonsa eli ennusteiden laatimisajankohdan arvoon.

Ennustettujen diskonttauskertoimien keskihajonnasta. Diskonttauskertoimien mallissa diskonttauskertoimien keskihajonnalla, $\sigma(t)$, on ajassa laskeva lineaarinen trendi. Käytännössä tämä trendi ei voi jatkua samansuuruisena vuoden 2015 jälkeen vuosikymmeniksi tulevaisuuteen, koska keskihajonnan on oltava positiivinen. Tältä osin eräs vaihtoehto olisi ollut korvata mallinnuksessa lineaarinen laskeva trendi sellaisella laskevalla funktiolla, joka säilyy kuitenkin positiivisena tulevaisuudessa.

Koska havaittujen diskonttauskertoimien keskihajonnalla on esitetty ajassa laskeva trendi, olisin voinut valita käytettäväksi ennustetuissa diskonttauskertoimissa keskihajonnan ennusteita, joka vastaa mallinnustulosten perusteella vuoden 2015 tasoa diskonttauskertoimien keskihajonnassa. Päätin kuitenkin käyttää diskonttauskertoimien ennusteissa keskihajonnan tasona vuosien 1950–2015 keskimääräistä keskihajonnan tasoa, mikä edustaa suurempaa vaihtelua, kuin mitä on todellisuudessa havaittu diskonttauskertoimissa viimeisen noin 30 vuoden aikana (ks. taulukko 11.1). Tämän valinnan seurauksena diskonttauskertoimien ennusteissa on enemmän vaihtelua, kuin todellisuudessa on havaittu noin viimeisen 30 vuoden aikana. Tällä valinnalla voin perustella sitä, että tulevien diskonttauskertoimien ennusteisiin sisältyy *vähintään* realistinen määrä tulevaisuuteen liittyvää epävarmuutta.

Diskonttauskertoimien ennusteiden keskiarvosta. Diskonttauskertoimien ennusteiden keskihajonnan (jakauman) lisäksi on syytä tarkastella myös ennusteiden keskiarvojen jakauman realistisuutta. Taulukossa 11.3 on esitetty diskonttauskertoimien 60 vuoden ennusteiden keskiarvojen jakauman tunnuslukuja. Tämä 60 vuoden diskonttauskertoimien keskiarvojen ennustejakauma on simuloitu niin, että jokaisesta elinkaariennusteessa käytetystä vuosittaisista diskonttauskertoimien ennusteiden sarjasta on laskettu keskiarvo ennustetuille diskonttauskertoimesta ensimmäisestä ennusteesta kuudenteenkymmenenteen ennusteeseen. Nämä 60 vuoden keskiarvot ennusteille on laskettu kaikille ennustesimulaatioille erikseen, ja näiden keskiarvojen jakauma on esitettynä taulukossa 11.3. On syytä huomata, että kyseessä ei ole yksittäisten ennustettujen diskonttauskertoimien jakauma, jossa on tietenkin enemmän vaihtelua, vaan diskonttauskertoimien pitkän aikavälin ennusteiden *keskiarvojen* jakaumasta.

Taulukko 11.3: *Elinkaariennusteissa käytettyjen ennustettujen diskonttauskertoimien 60 vuoden keskiarvojen jakauma. Jakauman odotusarvon (OA), keskihajonnan (KH) ja eräiden prosenttipisteiden estimaatit.*

OA	KH	2,5 %	10 %	25 %	50 %	75 %	90 %	97,5 %
0,989	0,009	0,971	0,977	0,983	0,988	0,944	1,000	1,006

Taulukon 11.3 perusteella neljäsosassa diskonttauskertoimien ennustesarjoja 60 vuoden diskonttauskertoimien keskiarvo on pienempi kuin 0,983, ja 10 %:ssa ennustesarjoja tämä pitkän aikavälin keskiarvo on alle 0,77. Toisaalta myös neljäsosassa diskonttauskertoimien ennustesarjoja 60 vuoden diskonttauskertoimien keskiarvo on suurempi kuin 0,944 ja 10 %:ssa ennustesarjoja tämä pitkän aikavälin keskiarvo on suurempi kuin 1,000. Jälkimmäinen ennusteiden keskiarvo tarkoittaa sitä, että 10 %:ssa ennustesarjoja diskontatut yksittäisten vuosien kertoimet keskimäärin kasvattavat diskontattujen ennusteiden itseisarvoja verrattuna diskonttaamattomiin ennusteisiin – kun yleensä diskontatut ennusteet ovat itseisarvoltaan pienempiä kuin diskonttaamattomat. Tämä osaltaan osoittaa, miten leveä diskonttauskertoimien elinkaariennusteissa käytetty jakauma on. Lisäksi diskonttauskertoimien ennustesarjojen 60 vuoden keskiarvojen jakauman keskihajonta on 0,9 %-yksikköä. Tämä *60 vuoden* ennusteiden keskiarvojen vaihtelu on lähes 2/3-osaa *yksittäisten* diskonttauskertoimien havaintojen vaihtelusta (ks. taulukko 11.1) viimeisen 20 vuoden aikana (v. 1994–2015).

Pysyvästi korkeammalle tasolle jäävät diskonttauskertoimet pienentäisivät diskontattujen ja diskonttaamattomien elinkaariennusteiden välisiä eroja. Edelliseen päättelyyn diskonttauskertoimien ennusteiden odotusarvon jakauman realistisuudesta tuo kuitenkin lisäepävarmuutta viimeisen 10 vuoden aikana (vuosina 2006–2015) havaittu diskonttauskertoimien keskiarvo. Diskonttauskertoimien keskiarvo on ollut vuosina 2006–2017 keskimäärin 0,997, eli se on ollut lähellä ykköstä. Diskonttauskertoimet ovat olleet vuosina 2006–2015 korkeita, koska Suomen Pankin peruskorko on ollut (keskimäärin) historiallisen alhaalla viimeisen 20 vuoden aikana samaan aikaan, kun rahanarvonkertoimen vuosimuutokset ovat olleet myös lähellä nollaa (ks. edellä kuviot 11.2 ja 11.3).

Mikäli vuoden 2006–2015 havainnot edustaisivat pidemmän aikavälin pysyvää siirtymää vuosien 1950–2015 keskimääräisestä tasosta, elinkaariennusteissa käytetyt diskonttauskertoimien (posteriori)ennusteet olisivat keskimäärin toteutuvia tulevia diskonttauskertoimia pienempiä. Tämä tarkoittaisi nettonykyarvoon diskontattujen elinkaariennusteiden olevan ennusteissa lähempänä nollaa, kuin mitä ne olisivat nykyistä korkeammalla diskonttauskertoimien keskimääräisellä tasolla.

Käytän kuitenkin diskonttauskertoimien ennusteissa ennusteiden odotusarvon jakaumana yllä esitettyä diskonttauskertoimien mallin vakiotermin posteriorijakaumaa. Tähän on kaksi syytä. Ensiksi, vuosien 2006–2015 korkeampi diskonttauskertoimien taso voi olla vain väliaikainen ilmiö. Koska perättäisten diskonttauskertoimien välillä on havaintoaineiston perusteella vahva lineaarinen yhteys, diskonttauskertoimet voivat pysyä jollakin (poikkeavalla) tasolla ilman, että kyse olisi diskonttauskertoimien prosessin odotusarvon jakauman siirtymisestä. Toiseksi, jos antaisin diskonttauskertoimien odotusarvon jakauman siirtyä oikealle/korkeammalle tasolle, niin tämä tarkoittaisi sitä, että elinkaariennusteiden diskontatut arvot keskimäärin lähenisivät diskonttaamattomia arvoja. Erityisesti Somaliassa ja Irakissa syntyneiden elinkaariennusteiden osalta tämä tarkoittaisi diskontattujen elinkaariennusteiden olevan keskimäärin negatiivisempia, kuin mitä ne nyt ovat. Vaihtoehtoinen ratkaisu odotusarvon jakauman suhteen voisi näin ollen vähentää tulosten uskottavuutta, mikäli näyttäisi siltä, että olen pelkän havaintoaineiston seuraamisen sijaan suosinut ratkaisua, joka tekee ennusteista keskimäärin negatiivisempia.

Mikäli kuitenkin päätyisin toisenlaiseen ratkaisuun diskonttauskertoimien odotusarvon jakauman suhteen – lähinnä vakuuttavan aineiston ulkopuolisen informaation perusteella – ottaisin tämän odotusarvoa

koskevan informaatioon huomioon diskonttauskertoimien mallin vakion priorijakauman valinnassa. Nyt priorijakauma on standardinormaalijakauma, mutta jos aineiston ulkopuolinen informaatio viittaisi siihen, että odotusarvon jakauma on tulevaisuudessa havaintoaineiston keskiarvoa suurempi, standardinormaalijakauman vakion priorijakaumana voisi korvata jollain sellaisella jakaumalla, jonka odotusarvo on nollaa suurempi.

Lopuksi totean vielä kertauksena, että bayesilaisen päättelyn käyttäminen (diskonttauskertoimien) ennusteissa tuo ennusteisiin ainakin kaksi etua verrattuna frekventistisiin menetelmiin. Ensinnäkin siten ennusteisiin sisältyy parametreihin niiden estimoinnin jälkeen edelleen sisältyvä epävarmuus. Diskonttauskertoimien ennusteiden kohdalla tämä tarkoittaa esimerkiksi sitä, että autoregressiiviseen kertoimeen (ρ) sisältyvä suuri epävarmuus tulee mukaan ennusteisiin. Toinen etu on se, että havaintoaineiston ulkopuolista informaatiota voidaan käyttää ennusteissa estimoitavien parametrien priorijakaumien kautta. En kuitenkaan käytä tätä etua diskonttauskertoimien ennusteissa.

Luku 12

Päätöksentekoteoria ja Suomeen muuttavien uusien maahanmuuttajien valinta

12.1 Yleistä päätöksentekoteoriasta

Päätöksentekoteoria tai päätöksentekoanalyysi (*engl. decision theory, decision analysis*) on tilastotieteen ja taloustieteen sovellusala (ks. klassisesta alan kirjallisuudesta erityisesti DeGroot (1970) ja Berger (1980), uudemmasta bayesilaisesta teoriapainotteisesta kirjallisuudesta Robert (2007) sekä uudemmasta teoriaa ja käytännön sovelluksia esittelevästä bayesialaisesta kirjallisuudesta Rios Insua et al. (2012) ja Gelman et al. (2014), luku 9. Käytännön soveltajille suunnattu uudempi teos on puolestaan Clemen and Reilly (2014)). Päätöksentekoteoriaa sovelletaan tilanteissa, joissa päätöksentekijän (*engl. decision-maker, DM*) on mahdollista tehdä vaihtoehtoisia valintoja tai tekoja (*engl. action*) ja jokainen tehty valinta johtaa – viitatussa kirjallisuudessa käytetyn terminologian mukaan – *tulokseen*, josta on laskettavissa *seuraus*, tai *palkkioon* (*engl. result (consequence), reward*), jolle on määritettävissä *hyödyn* tai *utiliteetin* arvo päätöksentekijän hyötyfunktiosta (*engl. utility function*).

12.1.1 Ennustetun hyödyn odotusarvo valinnasta

Päätöksentekoteorian soveltamista varten päätöksentekotilanteessa täytyy voida määrittää päätöksentekijän mahdolliset valinnat, valintojen tulosten tuottamat seuraukset sekä seurausten tuottama utiliteetti päätöksentekijälle. Mikäli määritettävissä olevia yksittäisiä ja erillisiä valintoja on k kappaletta, kaikkien valintojen joukko V koostuu näistä yksittäisistä päätöksentekijän valintavaihtoehdoista: $v_1, v_2, \ldots, v_{k-1}, v_k \in V$. Mikäli jokaisen valinnan $v_n, n = 1, \ldots, k$ seuraukselle on määritettävissä funktio, käytetään siitä Rios Insua et al. (2012) esimerkkiä seuraten merkintää: $c(v_n, \tau)$, missä τ on

valinnan tuottama tulos (tulevaisuudessa).

Bayesilaisessa viitekehyksessä tulokselle on määritetty havainnoilla (y) ehdollistettu posteriorijakauma: $p(\tau|y) \propto p(y|\tau)p(\tau)$, mikä on voimassa silloin, kun posteriorijakauma ei riipu tehdystä valinnasta (Rios Insua et al. 2012, 34). Jos tehdyn valinnan v_n seurauksen päätöksentekijälle tuottama hyöty voidaan kirjoittaa hyötyfunktiona, $U(c(v_n, \tau))$, ja posteriorijakaumalla $p(\tau|y)$ on olemassa odotusarvo, niin kirjallisuudessa esitetyn yleisen tuloksen mukaan päätöksentekijän utiliteetin odotusarvo voidaan lausua tällöin päätöstä v_n koskien:

$$E_\tau[U(c(v_n, \tau))|y] \propto \int U(c(v_n, \tau))p(y|\tau)p(\tau)d\tau. \tag{12.1}$$

Tässä yhteydessä on huomautettava, että useimmiten kirjallisuudessa hyötyfunktion sijaan käytetään nk. *loss*-funktiota, joka saadaan helposti utliliteettifunktiosta, sillä loss-funktio $L(v_n, \tau) = -U(v_n, \tau)$. Yksinkertaisuuden vuoksi oletan tässä ja alempana, että päätöksen utiliteetti saadaan suoraan päätöksen tuloksesta: $U(v_n, \tau) = U(c(v_n, \tau))$.

Odotusarvoa koskeva tulos on esitetty viitatussa kirjallisuudessa hieman eri muodoissa ja useimmiten loss-funktiota käyttäen. Odotusarvo teoksessa DeGroot (1970)(s. 137–138) sisältää myös havaintoaineiston satunnaisuuden ja siinä lossin odotusarvoa nimitetään *"riskiksi"* tai *"Bayes-riskiksi"*; odotusarvo on teoksissa Berger (1980)(s. 159) ja Robert (2007)(s. 62) *posteriorin lossin odotusarvo* (*engl. posterior expected loss*); Rios Insua et al. (2012)(s. 25) ja Gelman et al. (2014)(s. 238), jotka määrittävät tulokselle utiliteetin hyötyfunktion avulla, nimittävät odotusarvoa puolestaan *utiliteetin odotusarvoksi* tai *ennustetuksi utiliteetin odotusarvoksi* (*engl. (predictive) expected utility*).

Lisäksi Gelman et al. (2014)(s. 238) kehottavat käyttämään muusta kirjallisuudesta poikkeavaa merkintätapaa koskien tuloksen posteriorijakaumaa $p(\tau|y)$ ja valintojen hyötyfunktiota $U(v, \tau)$. Gelman et al. (2014) korostavat hyödyn laskemista erikseen jokaiselle valinnalle ja sen vuoksi he kehottavat ehdollistamaan valinnalla v posteriorin ja hyötyfunktion: $p(\tau|y, v)$ ja $U(\tau)|v$. Pitäydyn muun kirjallisuuden käyttämässä merkintätavassa, mutta pyydän lukijaa huomioimaan, että tuloksen posteriorijakauma lasketaan aina erikseen jokaiselle valinnalle, vaikka tätä ei ole erikseen merkitty hyödyn odotusarvon lausekkeeseen.

12.1.2 Ennustetun hyödyn odotusarvon maksimointi valinnasta

Päätöksentekoanalyysin varsinaisena päämääränä on löytää ne yksittäiset valinnat $v \in V$, jotka maksimoivat ennustetun utilliteetin odotusarvon (tai puolestaan minimoivat lossin odotusarvon) (Rios Insua et al. (2012), 34; Gelman et al. (2014), 238). Odotusarvon utiliteetin maksimointi voidaan esittää tällöin muodossa:

$$\max_{v \in V} E_\tau[U(v, \tau)|y]. \tag{12.2}$$

Berger (1980) käyttää tällaisten valintojen osajoukosta nimitystä Bayes-valinta (*Bayes action*) (s. 159). Bayes-valintoja on vain yksi silloin, kun yhden valinnan utiliteetin odotusarvo on suurempi (ja odotusarvoinen loss puolestaan pienempi) kuin kaikkien muiden valintojen. Bayes-valintoja on useampi kuin yksi

silloin, kun useammalla kuin yhdellä valinnalla on sama utiliteetin odotusarvo, joka on suurempi kuin kaikilla muilla valinnoilla. Huomautettakoon vielä, että DeGroot (1970) käyttää lossin *minimiin* johtavien valintojen löytämisen sijaan *infimumia* ilmeisesti sen vuoksi, että loss-odotusarvon minimoivaa valintaa ei ole aina löydettävissä valintojen joukosta (s. 138).

12.1.3 Tuloksen hyötyfunktio ja sen posteriorijakauman päivitys lisäinformaatiolla

12.1.3.1 Hyötyfunktio ja suurin rahallinen odotusarvo

Kuten edellä mainitsin kirjallisuuteen viitaten, utiliteetin odotusarvon määrittämiseen tarvitaan (ainakin) kolme komponenttia: valintojen määritys, valintojen tulosten posteriorijakauman määritys sekä päätöksentekijän utiliteettifunktion $U(v, \tau)$ määritys. Hyötyfunktiota ja hyötyä/utiliteettia koskevat päätöksentekoteoriassa samanlaiset määritelmät, mitä niillä on taloustieteessä (ks. esim. Robert (2007), 54–65 ja Berger (1980), 47–60). Koska hyötyfunktion ja loss-funktion perusteellinen määrittäminen voi olla haastavaa, usein käytännössä sovelletaan tiettyjä ”helppoja” loss-funktioita, joita ovat lossin neliövirhe, lineaarinen loss-funktio ja 0–1-loss-funktio (Berger (1980), 60–64). Lineaarinen loss-funktio (ja utiliteettifunktio) tarkoittaa sitä, että tietyn suuruisella valinnan tuloksella on aina samansuuruinen vaikutus lossiin (tai utiliteettiin) riippumatta lossin tai utiliteetin tasosta.

Mikäli valintojen tulokset ovat euromääräisiä, ja valintojen tulosten utiliteettifunktio on lineaarinen, utiliteetti voidaan ilmaista euroina, sillä utiliteetti on suoraan verrannollinen euromääräisiin valintojen tuloksiin nähden. Yllä määritetystä utiliteetin odotusarvosta voidaan tällaisessa tapauksessa käyttää nimitystä *rahallinen odotusarvo* (*engl. Expected Monetary Value, EMV*) (Clemen and Reilly (2014), 122). Odotusarvon maksimi on siten suurin rahallinen odotusarvo, (*engl. Highest EMV*). Mikäli valintojen euromääräiset tulokset ulottuvat (”pitkälle”) tulevaisuuteen, on otettava huomioon, että raha on päätöksentekohetkellä kalliimpaa kuin tulevaisuudessa. Sen vuoksi valintojen yhteenlasketut tulevat euromääräiset vaikutukset on syytä ilmaista nykyarvossaan eli nettonykyarvona (*engl. net present value, NPV*) (Clemen and Reilly (2014), 33–35). Käsittelen tulevien euromääräisten vaikutusten diskonttaamista nykyarvoonsa luvussa 11.

Valinnan hyödyn odotusarvo on verrannollinen valinnan tuloksen odotusarvoon, kun hyötyfunktio on lineaarinen:

$$E_\tau[U(v,\tau)|y] \propto \int U(v,\tau)p(y|\tau)p(\tau)d\tau \propto \int \tau p(y|\tau)p(\tau)d\tau \propto E_\tau[(v,\tau)|y]. \tag{12.3}$$

Valinnan tuottama hyöty saadaan näin ollen (vakiokerrointa vaille) tuloksen τ odotusarvona. Jälkimmäinen odotusarvon merkintätapa yllä korostaa sitä, että tuloksen τ odotusarvo lasketaan erikseen jokaiselle valinnalle v. Hyödyn maksimoi se valinta, jonka tuloksen τ odotusarvo on suurin. Mikäli lisäksi on kyse EMV:stä, valinnan tuloksen odotusarvo on ilmaistu euroissa.

12.1.3.2 Valinnan tuloksen posteriorijakauma ja posteriorijakauman päivitys lisäinformaatiolla

Edellä valinnan tuloksen posteriorijakauma määritettiin seuraavasti: $p(\tau|y) \propto p(y|\tau)p(\tau)$, missä y on havaintoaineisto. Koska kyse on bayesilaisesta päättelystä, valinnan tulokseen τ voidaan aina liittää sen priorijakauman $p(\tau)$ kautta informaatiota, jota ei ole saatavissa havaintoaineistosta y. Tämä ominaisuus mahdollistaa myös valinnan tuloksen τ päivittämisen uudella havaintoaineistolla: tällöin posteriorijakaumaa $p(\tau|y)$ käytetään uudella havaintoaineistolla päivitetyn valinnan tuloksen τ priorijakaumana.

Mikäli seurauksen hyötyfunktio on lineaarinen ja odotusarvo euromääräinen, valinnan tuloksen τ päivittäminen uudella aineistolla kertoo, kuinka monta euroa valinnan odotettavissa oleva hyöty muuttuu lisähavaintojen tuoman informaation johdosta. Tällaisessa tilanteessa lisähavaintojen tuottamalle informaatiolle voidaan laskea *informaation (rahallinen) odotusarvo* (*engl. expected value of information*). Se on jokaisen valinnan kohdalla erotuksen itseisarvo: $|EMV_{uusi}(\tau) - EMV_{vanha}(\tau)|$, missä $EMV_{uusi}(\tau)$ on uudella havaintoaineistolla päivitetyn valinnan tuloksen τ rahallinen odotusarvo ja $EMV_{vanha}(\tau)$ on rahallinen odotusarvo ilman lisähavaintojen tuomaa informaatiota (Clemen and Reilly (2014), 554).

12.1.3.3 Sekventiaalinen analyysi ja lisäinformaation arvo

Mikäli valintojen hyötyjen odotusarvot on mahdollista päivittää uudella havaintoaineistolla *ja* päätöksenteko on mahdollista siirtää seuraavaan ajanhetkeen, jolloin voi valita joko lisäinformaation hankkimisen tai päätöksenteon väliltä, kyse on nk. päätöksenteon *sekventiaalisesta analyysista* (*engl. sequential analysis*) (ks. Rios Insua et al. (2012), 35–37 ja Berger (1980), luku 7). Sekventiaalisessa analyysissa päätöksentekijä tekee käytännössä kahdenlaisia päätöksiä:

1. Päätöksen siitä, lykätäänkö ”varsinaisen” päätöksen tekemistä ja kerätään lisäinformaatiota varsinaista päätöstä varten.

2. Edellä esitetty varsinainen päätöksentilanne: valinta, josta seuraa tulos, jolle on määritetty posteriorijakauma ja hyötyfunktio.

Lisäinformaation hankintaan liittyy yleensä jonkinlainen kustannus. Sekventiaalisessa analyysissa päätöksentekijän odotettavissa oleva utiliteetti maksimoidaan koko ajassa jatkuvan sekvenssin yli: Mikäli päätöksentekijä uskoo lisäinformaation olevan ”tarpeeksi” arvokasta (vrt. informaation määritetty arvo yllä), päätöksentekijän kannattaa lykätä varsinaisen päätöksen tekemistä ja hankkia lisäinformaatiota. Muussa tapauksessa päätöksentekijä tekee varsinaisen päätöksen, josta seuraa tulos.

En paneudu sekventiaaliseen analyysiin – ja siihen, milloin lisäinformaatio on tarpeeksi arvokasta päätöksenteon lykkäämiseksi – tarkemmin, sillä alla esitettävässä päätöksenteon sovelluksessa maahanmuuttajien valintaa koskien päätöksentekoa ei voi käytännössä lykätä myöhemmäksi. Potentiaalisten uusien maahanmuuttajien valintaa ei voi käytännössä lykätä esimerkiksi vuosiksi eteenpäin. Sekventiaalinen analyysi kuitenkin tarjoaa esimerkin päätöksentekoanalyysin tilanteesta, jossa päätöksiä tehdään useilla eri ajanjaksoilla havaintojen tuoman lisäinformaation perusteella. Maahanmuuttajien valinnassa päätöksiä tehdään myös useina eri ajankohtina siten, että tiedot maahanmuuttajien vaikutuksista (Suomen julkiselle

taloudelle) päivittyvät. Tällöin hyödyn maksimointi tapahtuu myös yli ajan siten, että optimaalinen valinta (potentiaalisten uusien maahanmuuttajien suhteen) voi olla erilainen eri aikoina saatavilla olevan informaation perusteella.

Clemen and Reilly (2014) määrittävät sekventiaalisen päätöksenteon hieman eri tavalla kuin edellä viitattu teoriakirjallisuus. Heidän määritelmänsä mukaan sekventiaalisesta päätöksenteosta on kyse tapauksissa, joissa päätöksiä tehdään useita yli ajan siten, että jokaisella päätöksenteon hetkellä käytetään tuolloin saatavilla olevaa informaatiota päätöksentekijän odotettavissa olevan utiliteetin maksimoimiseksi. He antavat esimerkin maanviljelijästä, joka tekee säähän liittyvän ratkaisun joka päivä. Ratkaisu riippuu sekä nykyisestä säästä että sääennusteesta tulevalle päivälle siten, että saatavilla oleva informaatio jokaisen päivän päätöstä koskien päivittyy myös päivittäin. Tämä vastaa skemaattisesti hyvin pitkälti ajassa jatkuvaa päätöksentekoprosessia koskien potentiaalisten uusien maahanmuuttajien valintaa. (Clemen and Reilly (2014), 65–66 ja 78–79).

12.2 Päätöksentekoteorian sovellus: julkinen talous ja potentiaalisen uuden maahanmuuttajan valinta

Esitin edellä päätöksentekoteorian perusteet alan kirjallisuuteen viitaten. Keskeisin tavoite päätöksentekoteoriassa on löytää ne päätöksentekijän valinnat, joilla päätöksentekijän odotusarvoinen hyöty maksimoituu valintojen tuottamien tulosten seurauksena. Päätöksentekoteorian soveltamiseksi on kyettävä määrittämään kolme asiaa: päätöksentekijän valinnat, valintojen tuloksen posteriorijakauma sekä valintojen tulosten hyötyfunktio päätöksentekijälle. Valintojen tuloksen hyötyfunktion odotusarvon on lisäksi oltava olemassa.

Sovellan seuraavaksi päätöksentekoteoriaa potentiaalisten uusien maahanmuuttajien valintaan. Tätä varten on määritettävä potentiaalisen uuden maahanmuuttajan valintaan liittyvät valintamahdollisuudet, julkisen talouden hyötyfunktio potentiaalisen uuden maahanmuuttajan valinnasta sekä posteriorijakauma potentiaalisen uuden maahanmuuttajan valinnan tulokselle eli euromääräiselle vaikutukselle Suomen julkiseen talouteen.

12.2.1 Potentiaalisen uuden maahanmuuttajan valinta

”Päätöksentekijänä” tässä yhteydessä on julkinen talous, jonka odotusarvoista hyötyä potentiaalisten uusien maahanmuuttajien valinnasta maksimoidaan. Olen tutkimuksen I osassa määrittänyt julkisen talouden koostuvan valtiosta, kunnista, kuntayhtymistä ja sosiaaliturvarahastoista. Tässä yhteydessä, eli julkisen talouden päätöksentekijänä, viittaan kuitenkin lainsäätäjiin ja toimeenpanovaltaan eli Suomen eduskuntaan ja hallitukseen. Eduskunta ja hallitus ovat käytännössä ne instituutiot, jotka voivat päättää potentiaalisten uusien maahanmuuttajien valinnasta.

Olen tutkimuksen I osassa määrittänyt maahanmuuttajan tarkoittavan henkilöä, joka on syntynyt ulko-

mailla ja jolla on kotikunta Suomessa. Uudella maahanmuuttajalla viittaan tässä yhteydessä sellaiseen maahanmuuttajaan, joka on muuttamassa Suomeen ensimmäisen kerran (maahanmuuttoja Suomeen voi olla useita, jos välissä on poismuuttoja). Potentiaalisella uudella maahanmuuttajalla viittaan sellaiseen maahanmuuttajaan, jonka Suomeen muuttamiseen Suomen eduskunta ja hallitus voivat päätöksillään vaikuttaa. Esimerkiksi EU-alueelta tulevat uudet maahanmuuttajat eivät ole tässä yhteydessä ”potentiaalisia” maahanmuuttajia, sillä heidän maahanmuuttoonsa eduskunnan ja hallituksen on vaikeaa vaikuttaa, koska Suomi on EU:n jäsenvaltio.

EU:n ulkopuolelta tulevien potentiaalisten uusien maahanmuuttajien Suomeen muuttamiseen eduskunta ja hallitus voivat vaikuttaa joko suorasti tai epäsuorasti. Suora vaikutus on sellainen, että potentiaalisen uuden maahanmuuttajan maahantulo sallitaan tai estetään muuttamalla lakeja, jotka koskevat potentiaalisen uuden maahanmuuttajan Suomeen muuttoa. Päätös tässä tapauksessa koskee siis lainsäädäntöä, jolla määritetään, millaiset potentiaaliset uudet maahanmuuttajat voivat muuttaa Suomeen. Epäsuora vaikuttaminen tarkoittaa puolestaan sitä, että julkisen talouden päätöksentekijä tekee sellaisia ratkaisuja, jotka vähentävät *vetotekijöitä* Suomeen muuton suhteen. Tällaiset epäsuorat päätökset laskevat potentiaalisen uuden maahanmuuttajan *todennäköisyyttä* muuttaa Suomeen. Mikäli julkisen talouden päätöksentekijä sallii tai estää potentiaalisen uuden maahanmuuttajan muuton Suomeen suoralla lainsäädännöllisellä ratkaisulla, muutto toteutuu tai estyy täydellä varmuudella. Epäsuorat, eli vetovoimatekijöihin liittyvät ratkaisut, muuttavat potentiaalisten uusien maahanmuuttajien todennäköisyyttä muuttaa Suomeen.

Yksinkertaisuuden vuoksi julkisen talouden päätöksentekijän ratkaisut – koskien potentiaalisen uuden maahanmuuttajan muuttoa Suomeen – ovat tässä yhteydessä suoria. Voisin myös käsitellä ratkaisuja, joissa muutetaan Suomeen kohdistuvia vetotekijöitä. Tällöin potentiaalisen uuden maahanmuuttajan Suomeen muuttamiselle tulisi määrittää todennäköisyys, joka riippuu päätöksentekijän (poliittisista) valinnoista. Koska tilastollinen päättely tapahtuu tässä yhteydessä bayesilaisen päättelyn piirissä, olisi sinänsä varsin helppoa käsitellä maahanmuuton toteutumisen – päätöksellä ehdollista – todennäköisyyttä. Maahanmuuton toteutumiselle määritettäisiin (ehdollinen) todennäköisyysjakauma, kuten kaikille muillekin parametreille, jotka eivät ole vakioita.

Sovellan päätöksentekoteoriaa erityisesti uusiin potentiaalisiin Irakissa ja Somaliassa syntyneisiin maahanmuuttajiin, jotka muuttavat Suomeen pääasiassa humanitaarisen maahanmuuton seurauksena. Tästä syystä oletan, että julkisen talouden päätöksentekijä voi valita näiden potentiaalisten uusien maahanmuuttajien kohdalla joko Suomeen muuton tai ei muuttamista Suomeen. Perustelen tätä oletusta seuraavilla kiintiöpakolaisten ja turvapaikanhakijoiden valintaan liittyvillä tekijöillä.

Kiintiöpakolaisten kohdalla valinta on suoraan eduskunnan tehtävissä, sillä eduskunta päättää vuosittain kiintiöpakolaisten määrän (Maahanmuuttovirasto 2017). Myös turvapaikanhakijoiden kohdalla potentiaalisia uusia maahanmuuttajia on mahdollista valita suoraan, sillä useat eurooppalaiset ja länsimaiset maat, jotka ovat sitoutuneet samoihin kansainvälisiin sopimuksiin kuin Suomi, noudattavat turvapaikanhakijoiden maahantulon suhteen hyvin erilaisia käytäntöjä, kuten Riikka Purra (2016) kirjoittaa. Purra myös kirjoittaa siitä, miten eri länsimaat ovat tehneet omia ratkaisujaan turvapaikanhakijoiden mahdollisten vetotekijöihin suhteen, sekä pohtii sitä, miten Suomi voisi noudattaa näiden maiden esimerkkiä (ks. erit. s. 36–41).

Mikäli julkisen talouden päätöksentekijällä on mahdollisuus valita suoraan EU:n ulkopuolelta tuleva

potentiaalinen uusi maahanmuuttaja, niin mahdollisten valintojen joukko (kaksio) V muodostuu kahdesta mahdollisesta valinnasta:

$V = \{$"Potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen.",
"Potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen."$\}$.

Toisaalta mikäli julkisen talouden päätöksentekijällä on mahdollisuus vaikuttaa niihin vetovoimatekijöihin, jotka vaikuttavat EU:n ulkopuolelta tulevan potentiaalisen uuden maahanmuuttajan muuttoon Suomeen, niin valinta V muodostuu kahden mahdollisen valinnan joukosta (kaksiosta):

$V = \{$"Vähennetään potentiaalisen uuden maahanmuuttajan Suomeen muuton vetovoimatekijöitä.",
"Lisätään (tai pidetään samana) potentiaalisen uuden maahanmuuttajan Suomeen muuton vetovoimatekijöitä."$\}$.

Tässä jälkimmäisessä tapauksessa maahanmuuton toteutumiselle on määritettävä ehdollinen todennäköisyysjakauma, joka on ehdollinen tehdyllä päätöksellä. Lisäksi tässä tapauksessa valintoja on mahdollisesti enemmän kuin kaksi, sillä on päätettävä mahdollisten vetovoimatekijöiden tasojen väliltä.

Koska oletan, että kaikkien valintavaihtoehtojen kohdalla pelkän valinnan tekemisen kustannus on nolla (tästä enemmän jäljempänä), julkisen talouden odotettavissa oleva hyöty maksimoituu kummassakin yllä esitellyssä valintatilanteessa "'samansuuntaisella" valinnalla: eli joko maahanmuuton suoralla estämisellä ja vetovoimatekijöiden vähentämisellä tai maahanmuuton suoralla sallimisella tai vetovoimatekijöiden pitämisellä samassa tasossa tai lisäämällä niitä.

Vaikka käsittelen päätöksiä henkilötasolla, niin käytännössä maahanmuuttoon liittyvät valinnat koskettavat enempää kuin yhtä potentiaalista uutta maahanmuuttajaa. Joka tapauksessa optimaalinen päätös on julkisen talouden odotettavissa olevan hyödyn kannalta aina yhtä suuri, mikäli yksittäisiä potentiaalisia uusia maahanmuuttajia koskevat taustatiedot ovat samanlaiset.

Potentiaalisten uusien maahanmuuttajien valinta ei tietenkään rajoitu vain yhteen ajanhetkeen. Käsittelen aluksi julkisen talouden hyödyn odotusarvon maksimointia yksittäisellä valinnalla, joka tapahtuu tietyllä ajanhetkellä siten, että valinnassa on käytettävissä tuohon ajanhetkeen mennessä karttunut informaatio päätöksen tuloksesta. Jäljempänä (luvussa 12.4) käsittelen tilannetta, jossa julkisen talouden hyödyn odotusarvo halutaan maksimoida yli ajan siten, että potentiaalisten uusien maahanmuuttajien valintaa analysoidaan esimerkiksi kalenterivuosikohtaisesti, jolloin valinnan tulosta koskeva informaatio karttuu myös ajan myötä. Tämä analyysi perustuu edellä esitettyyn päätöksenteon sekventiaaliseen analyysiin, ja erityisesti siihen, miten Clemen and Reilly (2014) käsittävät sen.

12.2.2 Julkisen talouden hyöty maahanmuuttajaa koskevan valintapäätöksen tuloksesta

Sovellan tässä yhteydessä päätöksentekoteoriaa sellaiselle julkisen talouden päätöksentekijälle, jonka hyöty koskien potentiaalisen uuden maahanmuuttajan valintaa riippuu vain valinnan euromääräisestä tuloksesta. Valinnan euromääräinen tulos on ennustettu julkisen talouden elinkaarivaikutus potentiaalisesta uudesta maahanmuuttajasta silloin, kun potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen – ja jos maahanmuuttajaa ei valita, tällöin tulos on nolla euroa, kuten jäljempänä esitän.

Oletan lisäksi, että julkisen talouden kannalta hyödyn muutos yhden euron muutoksesta valinnan ennustetussa tuloksessa on aina yhtä suuri. Toisin sanoen julkisen talouden hyötyfunktio on tässä lineaarinen, ja jäljempänä esitettävä odotettavissa oleva hyöty potentiaalisen uuden maahanmuuttajan valinnasta on (euromääräinen) odotusarvoinen rahallinen hyöty (EMV), mikä käsitteenä esiteltiin edellä.

Julkisen talouden päätöksentekijän lineaarinen, euroissa mitattava hyöty toteutuu sellaisissa tilanteissa, joissa julkisen talouden päätöksentekijän tavoitteena on uusien maahanmuuttajien avulla kartuttaa julkisen talouden resursseja. Mikäli uuden maahanmuuttajan elinkaarivaikutukset ovat positiiviset, näin karttuneita resursseja voidaan käyttää julkisen talouden toimintojen kuten sosiaali- ja terveyspalveluiden tuottamiseen. Jos taas puolestaan elinkaarivaikutukset ovat negatiiviset, resursseja ei riitä yhtä paljon muihin julkisen talouden toimintoihin kuin ilman maahanmuuttajan valintaa.

Uusien maahanmuuttajien julkisen talouden hyödyn mittaaminen – enemmän tai vähemmän epäsuorasti – euroissa näkyi käytännössä ainakin syksyllä 2015 turvapaikanhakijoista käydyn julkisen keskustelun ympärillä. Monet merkittävät suomalaiset mielipidevaikuttajat ottivat myönteisen kannan turvapaikanhakijoiden vastaanotolle nimenomaan turvapaikanhakijoista koituvien *oletettujen* julkisen talouden hyötyjen vuoksi. Viittaan lukuisiin tällaisiin lausuntoihin sekä käsittelen kriittisesti oletettujen hyötyjen toteutumisen realistisuutta havaintoaineiston valossa kirjoituksessani *Turvapaikanhakijoiden työllistymisestä* (Salminen 2015b).

Moniulotteinen hyötyfunktio uuden maahanmuuttajan valinnasta. Yllä julkisen talouden päätöksentekijällä on maahanmuuttajien valinnan suhteen vain yksi hyödyke: potentiaalisten uusien maahanmuuttajien julkisen talouden elinkaarivaikutukset. Tällaisen hyötyfunktion tapauksessa julkisen talouden päätöksentekijän ei tarvitse tehdä valintoja eri hyödykkeiden välillä. Julkisen talouden päätöksentekijän hyötyfunktio voi olla kuitenkin myös moniulotteinen potentiaalisen uuden maahanmuuttajan valinnan tuottaman tuloksen suhteen. Tällöin valitun maahanmuuttajan julkisen talouden elinkaarivaikutus on vain yksi hyödyke muiden hyödykkeiden joukossa. Muita hyödykkeitä voivat olla – esimerkiksi humanitaarisen maahanmuuton kohdalla – auttamisen halun toteuttaminen ja poliittisen suosion kasvattaminen potentiaalisen uuden maahanmuuttajan valinnalla tai valitsematta jättämisellä. Työperäisen maahanmuuton osalta potentiaalisten uusien maahanmuuttajien valinta voi aiheuttaa julkisen talouden päätöksentekijöille poliittisia kustannuksia, jos uusien muuttajien katsotaan esimerkiksi alentavan palkkatasoa.

Kun hyödykkeitä on potentiaalisen uuden maahanmuuttajan elinkaarivaikutusten lisäksi myös muita, julkisen talouden päätöksentekijä maksimoi hyötynsä valitsemalla näiden hyödykkeiden väliltä. Ennustetun hyödyn odotusarvon maksimi voi tällöin pitää sisällään myös sellaisten potentiaalisten uusien maahanmuuttajien valinnan, joiden julkisen talouden elinkaarivaikutusten euromääräinen odotusarvo on negatiivinen. Julkisen talouden päätöksentekijä voi olla valmis maksamaan tämän kustannuksen, koska tämän kustannuksen kompensoi julkisen talouden päättäjän hyötyfunktion maksimoinnin näkökulmasta esimerkiksi hyöty maahanmuuttajien valinnan tuomasta poliittisesta arvostuksesta.

Koska tämän tutkimuksen piiriin kuuluvat vain suorat julkisen talouden vaikutukset, pitäydyn jäljempänä yksinkertaisessa lineaarisessa euromääräisessä hyötyfunktiossa, jonka arvo määräytyy suoraan potentiaalisen uuden maahanmuuttajan ennustetusta elinkaarivaikutuksesta. Tällä valinnalla myös vältän hankalat kysymykset julkisen talouden hyödyn ja julkisen talouden päätöksentekijän (eduskunta ja hallitus) hyödyn

mahdollisista eroavaisuuksista. Käytännön päätöksenteossa on otettava huomioon hyötyfunktion mahdollinen moniulotteisuus uusien maahanmuuttajien valinnassa. Moniuloitteisen hyötyfunktion tapauksessakin potentiaalisten uusien maahanmuuttajien ennustetut euromääräiset elinkaarivaikutukset vaikuttavat hyötyfunktion saamiin arvoihin, joten potentiaalisen uuden maahanmuuttajan elinkaarivaikutuksen posterioriennustejakauma on hyödyllistä määrittää myös tällaista moniulotteisempaa valintatilannetta varten.

12.2.3 Tulos maahanmuuttajan valintapäätöksestä: uuden maahanmuuttajan elinkaarivaikutuksen posterioriennustejakauma

Uuden maahanmuuttajan valinnan tuloksen ehdollinen todennäköisyysjakauma on verrannollinen seuraavasti (ks. lausekkeen johtaminen johdannon luvussa 2.2):

$$p(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \propto p(\boldsymbol{\pi}) \int_{\boldsymbol{\theta}} p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta}) p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{12.4}$$

missä $\boldsymbol{y}_{uusi}$ on vektori, joka sisältää niiden komponenttien ennusteet (esimerkiksi ennustetut palkkasummat on eräs komponentti), joista uuden maahanmuuttajan elinkaarivaikutukset muodostuvat. Uutta maahanmuuttajaa koskeva selittävien muuttujien matriisi on $\boldsymbol{x}_{uusi}$. Se koostuu kaikista niistä uutta maahanmuuttajaa koskevista, havaittavissa olevista tiedoista, joita käytetään selittävinä muuttujina elinkaarivaikutusten komponenttikohtaisissa ennusteissa (esimerkiksi ennustetut julkisen talouden nettovaikutukset työiässä). Elinkaarivaikutusten komponenttikohtaisten ennustemallien kaikki parametrit sisältyvät puolestaan vektoriin $\boldsymbol{\theta}$. Ennustekomponenttien vektori $\boldsymbol{y}_{uusi}$ on ehdollisesti riippumaton jo maassa olevien (”vanhojen”) maahanmuuttajien havaintoaineistosta, kun ehdollistajana on parametrivektori $\boldsymbol{\theta}$. Parametrivektori $\boldsymbol{\pi}$ koostuu ”muista” ennusteparametreista, mm. diskonttauskertoimista.

Parametrivektorin $\boldsymbol{\theta}$ posteriorijakauma on ehdollinen tiedoilla, jotka koskevat Suomessa olevia tai aiemmin olleita (”vanhoja”) maahanmuuttajia:

$$p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \propto p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}|\boldsymbol{\theta}) p(\boldsymbol{\theta}).$$

Elinkaarivaikutusten komponenttien mallinnuksissa käytetyt havaitut selitettävät muuttujat sisältyvät vektoriin $\boldsymbol{y}_{vanha}$, ja näissä malleissa käytetyt selittävät muuttujat puolestaan matriisiin $\boldsymbol{x}_{vanha}$. Nämä havaitut tiedot koskevat Suomessa olevia tai aiemmin olleita (”vanhoja”) maahanmuuttajia.

Kirjoitetaan kaiken havaintoaineiston sisältävä vektori muodossa: $\boldsymbol{h} = \begin{bmatrix} \boldsymbol{x}_{uusi} & \boldsymbol{y}_{vanha} & \boldsymbol{x}_{vanha} \end{bmatrix}$. Kaikki havaintoaineisto koostuu siis ”vanhoja” maahanmuuttajia koskevista selittävistä ja selitettävistä muuttujista sekä ”uusia” maahanmuuttajia koskevista selittävistä muuttujista.

Elinkaarivaikutusten funktio. Uuden maahanmuuttajan valinnan seuraus, josta julkisen talouden ennustettu valintakohtainen hyöty saadaan, on uuden maahanmuuttajan elinkaarivaikutus. Määritetään funk-

tio $f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})$ kuvauksena komponenttien ennusteista elinkaarivaikutuksen ennusteeseen:

$$f : \{\textbf{olotila-ajat}\} \times \{\text{nettovaikutukset: alle 15-v.}\} \times \{\text{nettovaikutukset: nuoret ja työikäiset}\} \times \{\text{palkkasummat: 18–62-v.}\} \times \{\text{muut nettovaikutukset: yli 62-v.}\} \times \{\boldsymbol{\pi}\} \times \{\text{elinkaarivaikutukset: lapset}\} \rightarrow \{\text{elinkaarivaikutus}\}. \tag{12.5}$$

Funktion määrittelyjoukko muodostuu osaennusteiden joukkojen – eli elinkaarivaikutusten komponenttien – sekä muiden ennusteparametrien ($\boldsymbol{\pi}$) karteesisesta tulosta. Funktion arvojoukko muodostuu elinkaarivaikutusten ennusteiden joukosta. Käytännössä funktio tarkoittaa elinkaarivaikutusennusteiden laskemista nettovaikutusennusteista (työikä, ei-työikä, lapset), palkkasummaennusteista ja olotilaennusteista. Mahdollisia olotiloja on kolme: henkilö on aina joko Suomessa, ulkomailla tai kuollut. Muut ennusteparametrit ($\boldsymbol{\pi}$) tarkoittavat tässä esimerkiksi työeläkemaksujen työntekijän ja työnantajan maksuprosentteja työtulosta. Elinkaarivaikutuksen funktiota sekä uusien maahanmuuttajien elinkaarivaikutusten ennustejakaumien muodostamista käsittelen tässä esitettyä tarkemmin johdannossa luvussa 2.1. Elinkaarivaikutusten laskeminen käytännössä on esitetty luvussa 9.

Mikäli julkisen talouden hyöty uuden maahanmuuttajan elinkaarivaikutuksesta on lineaarinen ja euromääräinen, julkisen talouden ennustettu hyöty saadaan suoraan uuden maahanmuuttajan euromääräisestä ehdollisesta elinkaarivaikutuksen ennusteesta:

$$u(f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})) = f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}).$$

Jotta ennustettu hyöty todella vastaisi määritelmäänsä, julkisen talouden nettovaikutusten havaintoaineistossa olisi oltava ideaalisesti kaikki tutkimuksen määritelmän mukaiset henkilöiden suorat julkisen talouden euromääräiset nettovaikutukset. Joitakin julkisen talouden suoria vaikutuksia kuitenkin puuttuu havaintoaineistosta, sillä niitä ei sisälly käytössäni olevaan rekistereihin perustuvaan tutkimusaineistoon. Käsittelen näitä puuttuvia tietoja luvussa 10. Kuten esitän mainitussa luvussa, suurimmat puutteet liittyvät erityisesti maahanmuuttajien erityispalveluiden menoihin, sillä niistä ei ole olemassa ainakaan sellaisia henkilövuositasoisia rekisteritietoja, jotka olisivat käytössäni. Mikäli nämä julkisen talouden menot olisivat mukana tutkimusaineistossa, erityisesti humanitaarisista syistä maahan tulleiden elinkaarivaikutusten odotusarvot muuttuisivat negatiivisemmiksi, kuin mitä ne nyt ovat.

Tulos julkiselle taloudelle on nolla euroa, kun uutta maahanmuuttajaa ei valita Suomeen. Yllä esitetty potentiaalisen uuden maahanmuuttajan valinnan tuloksen posteriorijakauma pätee sekä sen valinnan osalta, että potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen sekä myös sen valinnan osalta, että potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen. Kun maahanmuuttaja valitaan muuttavaksi Suomeen, julkisen talouden hyöty määräytyy komponenttikohtaisten ennusteiden avulla elinkaarivaikutusten edellä esitetyn funktion kautta. Käsittelen seuraavaksi lyhyesti julkisen talouden hyötyä, kun potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen.

Kun potentiaalista uutta maahanmuuttajaa ei valita Suomeen, tämän potentiaalisen maahanmuuttajan elinkaarivaikutus – ja tässä tapauksessa hyöty julkiselle taloudelle – on tasan nolla euroa. Tämä seuraa

esimerkiksi elinkaarivaikutusennusteiden – olotilaennusteita koskevien – komponenttien kautta siitä, että tällaisen henkilön olotila on koko elämänsä ajan ulkomailla eikä Suomessa. Se, että elinkaarivaikutus on tässä tapauksessa tasan nolla euroa, on järkevää myös siitä syystä, että henkilöllä, joka ei koskaan muuta Suomeen, ei ole Suomen julkisen talouden kannalta tuloja eikä menoja.

Todellisuudessa siitä, että potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen, aiheutunee julkiselle taloudelle joitakin suoria menoja eli negatiivisia julkisen talouden nettovaikutuksia. Nämä menot aiheutuvat esimerkiksi rajavalvonnasta tai turvapaikkahakemusten käsittelystä ja turvapaikanhakijoiden vastaanotosta aiheutuvista menoista. Koska näistä menoista ei ole saatavilla henkilötason rekisteritietoja, ne jäävät pois nettovaikutuslaskelmista. Joka tapauksessa, jos potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen, elinkaarivaikutuksista puuttuu tässä tapauksessa enemmän menoja, kuin jos maahanmuuttajaa ei valita. Erityisesti asia on näin humanitaarisen maahanmuuton kautta Suomeen muuttaneiden kohdalla, koska vastaanotto- ja kotouttamismenoista puuttuu ainakin osa (ks. luku 10.2.1).

Potentiaalisen uuden maahanmuuttajan valitsemisella ja valitsematta jättämisellä voi olla myös epäsuoria julkisen talouden vaikutuksia. Ne eivät kuitenkaan sisälly tämän tutkimuksen piiriin.

12.2.4 Julkisen talouden ennustetun hyödyn odotusarvon maksimointi valintapäätöksestä

Olen edellä määrittänyt päätöksentekoteorian soveltamiseen tarvittavat osatekijät, kun teoriaa sovelletaan potentiaalisten uusien maahanmuuttajien valintaan. Olen määrittänyt kaksi mahdollista valintaa, joiden väliltä julkisen talouden päätöksentekijä tekee valinnan: joko potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen, tai häntä ei valita muuttavaksi Suomeen. Olen lisäksi määrittänyt julkisen talouden hyötyfunktion, eli potentiaalisen uuden maahanmuuttajan tuottaman hyödyn määrän julkiselle taloudelle. Se on uuden maahanmuuttajan julkisen talouden elinkaarivaikutus. Määritin hyötyfunktion lineaariseksi ja euromääräiseksi, joten uuden maahanmuuttajan (euromääräinen) elinkaarivaikutus kertoo suoraan maahanmuuttajan tuottaman hyödyn määrän Suomen julkiselle taloudelle. Kolmanneksi määritin maahanmuuttajan valinnan (tai valitsematta jättämisen) tuottaman tuloksen ja seurauksen posteriorijakauman. Seuraus, josta julkisen talouden hyöty lasketaan, on tässä tapauksessa potentiaalisen uuden maahanmuuttajan elinkaarivaikutuksen ennuste. Mikäli potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen, elinkaarivaikutuksen ennuste on aina tällaisen henkilön kohdalla tasan nolla euroa.

Kaiken edellä määritetyn perusteella julkisen talouden hyödyn odotusarvo potentiaalisen uuden maahanmuuttaja valinnasta tai valitsematta jättämisestä maksimoituu, kun maksimoidaan valinnalla $v \in V$ lauseke, missä merkinnät ovat, kuten edellä on esitetty:

$$\max_{v \in V} \int_{\boldsymbol{\theta}} \int_{\boldsymbol{y}_{uusi}} u(v, f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})) p(\boldsymbol{y}_{uusi} | \boldsymbol{x}_{uusi}, \boldsymbol{\theta}) d\boldsymbol{y}_{uusi} p(\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha} | \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{12.6}$$

Tässä määritelmässä uusien maahanmuuttajien elinkaariennusteet ovat ehdollisesti riippumattomia vanhoista maahanmuuttajista (ehdollistajana malliparametrit $\boldsymbol{\theta}$). Mikäli hyötyfunktio on lineaarinen ja eu-

romääräinen, odotusarvon maksimointiongelman ratkaisu pelkistyy suurimman rahallisen odotusarvon (EMV) löytämiseen kahden valinnan väliltä:

$$\max\left(E[f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})|\boldsymbol{h}], 0\right). \tag{12.7}$$

Julkisen talouden hyödyn odotusarvo on $E[f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})|\boldsymbol{h}]$, kun potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen.Julkisen talouden hyödyn odotusarvo on puolestaan nolla (euroa), mikäli uutta potentiaalista maahanmuuttajaa ei valita muuttavaksi Suomeen.

Kun uuden maahanmuuttajan elinkaarivaikutuksen odotusarvo on suurempi kuin nolla, julkisen talouden hyödyn odotusarvo maksimoituu sillä valinnalla, jolla potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen. Mikäli puolestaan uuden maahanmuuttajan elinkaarivaikutuksen odotusarvo on pienempi kuin nolla, julkisen talouden hyöty maksimoituu sillä valinnalla, jolla potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen.

12.3 Aineiston ulkopuolisten tekijöiden vaikutukset ja potentiaalisen uuden maahanmuuttajan valinta

Malliparametreja koskeva informaatio muodostuu lähes täysin käytössä olevasta tutkimusaineistosta. Olen esittänyt tämän tutkimustekstin johdannossa (luvussa 2.3), kuinka elinkaarivaikutusten ennusteisiin liittyvät todennäköisyysjakaumat kuvaavat niitä epävarmuuksia, mitä ennusteisiin liittyy. Nämä todennäköisyysjakaumat ovat elinkaariennusteiden muodostavien satunnaismuuttujien (ehdollisia) todennäköisyysjakaumia, niin kuin olen johdannossa esittänyt. Elinkaarivaikutusten ennusteiden komponenttien, $\boldsymbol{y}_{uusi}$, malliparametreilla, $\boldsymbol{\theta}$, on ehdollinen todennäköisyysjakauma. Se on ehdollinen ”vanhoja” eli Suomessa jo olevia maahanmuuttajia koskevalla havaintoaineistolla. Näin ollen tämä ehdollinen todennäköisyysjakauma on posteriorijakauma: $p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$. Koska käytettävissä oleva tutkimusaineisto on havaintomäärältään suuri, ja koska käytän yleisesti ottaen tässä tutkimuksessa melko epäinformatiivisia priorijakaumia, malliparametreihin $\boldsymbol{\theta}$ liittyvä epävarmuus elinkaariennusteissa määräytyy lähes täysin tutkimusaineiston mukaan.

Käytettävissä oleva tutkimusaineisto kattaa vaihdellen vuodet 1988–2014 – riippuen elinkaarivaikutusten komponentista. Tärkeimpien komponenttien osalta ainoastaan palkkasummia koskeva aineiston yltää vuoteen 2014 asti, muiden osalta vuoteen 2011. Laadin kuitenkin elinkaariennusteet uusille maahanmuuttajille, jotka muuttavat Suomeen ensimmäisen kerran vuoden 2011 jälkeen. Kuten esitän johdantoluvussa tutkimuskirjallisuuteen viitaten, on mahdollista, että malliparametrien ehdolliset todennäköisyysjakaumat ovat erilaisia vuoden 2011 jälkeen, kuin ne ovat pelkästään vuosiin 1988/1995–2011 sijoittuvasta tutkimusaineistosta estimoitaessa.

Aineiston ulkopuolisten tekijöiden vaikutukset, ϵ. Kutsun sellaisia tekijöitä, jotka voivat vaikuttaa malliparametrien ehdolliseen jakaumaan vuoden 2011 jälkeen, (tutkimus)aineiston ulkopuolisiksi tekijöiksi. Koska näiden aineiston ulkopuolisten tekijöiden vaikutusten suuruuteen liittyy myös epävarmuus-

tekijöitä, ne ovat myös satunnaismuuttujia, joilla on omat todennäköisyysjakaumansa. Käytän näiden satunnaismuuttujien muodostamasta satunnaisvektorista merkintätapaa: $\boldsymbol{\epsilon}$. Kun aineiston ulkopuoliset tekijät huomioidaan elinkaarivaikutusten jakaumassa, malliparametrien ja aineiston ulkopuolisten tekijöiden yhteistodennäköisyysjakauma voidaan kirjoittaa "ketjuttamalla" muotoon: $p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})p(\boldsymbol{\epsilon})$.

Kuten esitin johdannossa tutkimuskirjallisuuteen viitaten, eräs aineiston ulkopuolisen tekijän vaikutus voi olla vuoden 2011 jälkeen Suomeen muuttavien vuosittaisten muuttokohorttien aikaansaama. Tämä tarkoittaa sitä, että voi olla, että vuoden 2011 jälkeen Suomeen muuttaa jostakin syntymävaltiosta sellaisia uusia maahanmuuttajia, joiden julkisen talouden vaikutukset poikkeavat tutkimusaineistoon sisältyvien maahanmuuttovuosikohorttien vaikutuksista, *kun muut vaikutukset huomioidaan*. Tässä tapauksessa aineiston ulkopuolisen tekijän vaikutus kohdistuu erityisesti maahanmuuttovuosiluokan indikaattorimuuttujien regressiokertoimien todennäköisyysjakaumiin vuoden 2011 jälkeen elinkaarivaikutusten komponenttien ennusteissa.

Maassaoloaikaan liittyviin malliparametreihin vaikuttavia aineiston ulkopuolisten tekijöiden vaikutuksia on kaksi kappaletta. Vuoden 2011 jälkeisen tulevaisuuden lisäksi maassaoloaikaan liittyvien malliparametrien jakaumiin voi vaikuttaa myös niiden osalta elinkaariennusteissa suoritettava ekstrapolointi. Koska tutkimusaineistossa maassaoloaikoja on 23 maassaolovuoteen asti, mutta työikäisten nettovaikutusmallinnuksessa maassaoloaikoja ennustetaan maksimissaan 46 maassaolovuoteen saakka, maassaoloajan vaikutus 23 maassaolovuoden jälkeen on ekstrapoloitava.

Aineiston ulkopuolisten tekijöiden vaikutukset koskevat uusien maahanmuuttajien elinkaariennusteiden ehdollisia todennäköisyysjakaumia. On syytä huomioida, että aineiston ulkopuolisten tekijöiden vaikutus kohdistuu nimenomaan elinkaarivaikutusten *ehdolliseen* ennustejakaumaan, eli jakaumaan, jossa on huomioitu uusia maahanmuuttajia koskevat tiedot $\boldsymbol{x}_{uusi}$. Nämä tiedot koskevat esimerkiksi uuden maahanmuuttajan sukupuolta, ikää, koulutusta (tieto tosin puutteellinen) ja pääasiallista toimintaa muuttovuonna (karkeasti ottaen muuton syy). Mikäli esimerkiksi tulevissa vuosittaisissa muuttokohorteissa olisi saatavilla olevan koulutustiedon perusteella enemmän koulutettuja kuin havaintoaineistossa, niin tämä vaikutus tulisi huomioiduksi ennusteissa ilman, että aineiston ulkopuolisten tekijöiden vaikutusta huomioitaisiin lainkaan. Toki olisi mahdollista luoda myös ennustemalli sille, millaisia maahanmuuttajia Suomeen muuttaa tulevaisuudessa, eli millainen $\boldsymbol{x}_{uusi}$ on tulevaisuudessa. Tämän ennustemallin muodostamiseen voisi jälleen käyttää sekä olemassa olevan havaintoaineiston sisältämiä tietoja sekä aineiston ulkopuolelta tulevia tietoja. En kuitenkaan ainakaan tässä tutkimuksen versiossa laadi $\boldsymbol{x}_{uusi}$:lle omaa ennustemallia.

Aineiston ulkopuolisten tekijöiden vaikutukset ovat kaiken kaikkiaan siis sellaisia elinkaariennusteisiin liittyviä vaikutuksia, jotka eivät sisälly havaintoaineistoon. Aineiston ulkopuolisten tekijöiden vaikutusten huomiointi elinkaariennusteissa – ainakin niin kuin ne esitän tässä tutkimustekstissä hyvin yksinkertaisessa muodossa – tarkoittaa käytännössä aineiston ulkopuolisten tekijöiden (satunnaisten) vaikutusten $\boldsymbol{\epsilon}$ lisäämistä ennusteisiin.

Aineiston ulkopuolisten tekijöiden mahdollinen vaikutus potentiaalisten uusien maahanmuuttajien valintaa. Määritän tässä luvussa seuraavaksi ehdon sille, miten aineiston ulkopuolisten tekijöiden vaikutusten *huomiointi* voi muuttaa julkisen talouden hyödyn maksimoivan päätöksen – koskien potentiaalisen uuden maahanmuuttajan valintaa – kun päätöstä verrataan valintaan, jossa aineiston ulkopuolisten tekijöiden vaikutuksia *ei huomioida* elinkaariennusteiden jakaumassa. Tämän jälkeen esitän kuvioin erilaisia mahdollisuuksia sille, miten aineiston ulkopuoliset tekijät voivat vaikuttaa yhden malliparametrin todennäköisyysjakaumaan. Käsittelen esimerkkinä maassaoloajan regressiokerrointa työikäisten nettovaikutusennusteissa. Lopuksi esitän kuvioiden avulla esimerkkejä siitä, kuinka aineiston ulkopuoliset tekijät voivat vaikuttaa uuden maahanmuuttajan elinkaarivaikutusten todennäköisyysjakaumaan.

12.3.1 Aineiston ulkopuolisten tekijöiden vaikutukset ja julkisen talouden hyödyn odotusarvon maksimointi

Edellä on useaan otteeseen esitetty uuden maahanmuuttajan kaikella havaintoaineistolla ($\boldsymbol{h}$) ehdollisen elinkaariennusteen odotusarvo: $E[f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})|\boldsymbol{h}]$. Määritetään nyt tämä odotusarvo erikseen elinkaarivaikutusten todennäköisyysjakaumalle, jossa ei ole huomioitu aineiston ulkopuolisten tekijöiden vaikutuksia, sekä sellaiselle jakaumalle, jossa nämä vaikutukset on huomioitu. Oletan tässä, kuten edelläkin, että kaikki odotusarvot ovat olemassa. Tämä voidaan käytännössä varmistaa valitsemalla sopivat todennäköisyysjakaumat kaikille (satunnaisille) malliparametreille.

Määritetään elinkaarivaikutusten ennustejakauman odotusarvo ilman aineiston ulkopuolisten tekijöiden vaikutuksia ja siten, että aineiston ulkopuolisten tekijöiden vaikutukset, $\boldsymbol{\epsilon}$, on huomioitu odotusarvossa.

Ylempi odotusarvo on (vakiokerrointa vaille) elinkaarivaikutusten ennustejakauman odotusarvo ilman aineiston ulkopuolisten tekijöiden vaikutuksia. Alempana (vakiokerrointa vaille) olevassa odotusarvossa aineiston ulkopuolisten tekijöiden vaikutukset on sen sijaan huomioitu:

$$E(\text{elinkaari, ilman } \boldsymbol{\epsilon}) \tag{12.8}$$
$$:= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{y}_{uusi}} f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}) p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta}) d\boldsymbol{y}_{uusi} p(\boldsymbol{\theta}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) d\boldsymbol{\theta}.$$

$$E(\text{elinkaari, mukana } \boldsymbol{\epsilon}) \tag{12.9}$$
$$:= \int_{\boldsymbol{\epsilon}} \int_{\boldsymbol{\theta}} \int_{\boldsymbol{y}_{uusi}} f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi}) p(\boldsymbol{y}_{uusi}|\boldsymbol{x}_{uusi}, \boldsymbol{\theta}, \boldsymbol{\epsilon}) d\boldsymbol{y}_{uusi} p(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) d\boldsymbol{\theta} d\boldsymbol{\epsilon}.$$

Mikäli elinkaariennusteen laatijan argumenttina on, että aineiston ulkopuoliset, tulevaisuuteen liittyvät, (epävarmuus)tekijät vaikuttavat potentiaalisen uuden maahanmuuttajan elinkaarivaikutuksen jakauman

odotusarvoon, ennusteen laatijan on osoitettava, että elinkaariennusteen odotusarvo muuttuu – kun ennusteessa huomioidaan myös aineiston ulkopuolisten tekijöiden vaikutukset.

Määritelmä. Aineiston ulkopuolisten tekijöiden vaikutusten (ϵ) huomiointi muuttaa elinkaariennusteen odotusarvon, kun:

$$\begin{aligned} &E(\text{elinkaari, ilman } \boldsymbol{\epsilon}) \\ &\neq \\ &E(\text{elinkaari, mukana } \boldsymbol{\epsilon}). \end{aligned} \tag{12.10}$$

Mikäli elinkaariennusteen jakauman odotusarvo muuttuu aineiston ulkopuolisten tekijöiden vaikutuksesta, se voi olla joko suurempi tai pienempi kuin odotusarvo, jossa ei ole huomioitu aineiston ulkopuolisten tekijöiden vaikutuksia. On myös syytä huomata, että elinkaariennusteen *jakauma* voi muuttua, kun aineiston ulkopuoliset tekijät huomioidaan, vaikka *jakauman odotusarvo* ei muuttuisikaan. Aineiston ulkopuolisten tekijöiden vaikutusten huomiointi saattaa esimerkiksi lisätä elinkaariennusteisiin sisältyvää epävarmuutta – mikä näkyy leveämpänä ennustejakaumana – mutta odotusarvo voi pysyä silti samansuuruisena (kuin ennustejakaumassa, jossa näitä vaikutuksia ei huomioida).

Ehto sille, että aineiston ulkopuolisten tekijöiden vaikutukset muuttavat hyödyn maksimoivan valinnan, on, että uuden maahanmuuttajan elinkaariennusteen etumerkki vaihtuu. Formaalisti tämä ilmaistaan elinkaariennusteen ehdollisten odotusarvojen kautta.

Määritelmä. Aineiston ulkopuolisten tekijöiden vaikutusten (ϵ) huomiointi muuttaa odotettavissa olevan hyödyn maksimoivan valinnan, kun:

$$\begin{aligned} &\text{sign}(E(\text{elinkaari, ilman } \boldsymbol{\epsilon})) \\ &\neq \\ &\text{sign}(E(\text{elinkaari, mukana } \boldsymbol{\epsilon}))\,. \end{aligned} \tag{12.11}$$

Toisin sanoen, jos ehdollisen elinkaariennusteen odotusarvon etumerkki (*sign*) vaihtuu, kun elinkaariennusteessa otetaan huomioon aineiston ulkopuolisten tekijöiden vaikutukset ($\boldsymbol{\epsilon}$), tämä tarkoittaa sitä, että julkisen talouden odotettavissa olevan hyödyn maksimoiva ratkaisu potentiaalisen uuden maahanmuuttajan valinnan suhteen vaihtuu.

Esimerkkejä odotettavissa olevan hyödyn maksimoivan valinnan muuttumisesta, kun aineiston ulkopuoliset tekijät huomioidaan. Jos esimerkiksi uuden maahanmuuttajan ehdollinen elinkaariennuste ilman aineiston ulkopuolisten tekijöiden vaikutusta on +100 000 euroa (eli odotusarvon etumerkki on positiivinen), mutta odotusarvo muuttuu negatiiviseksi (esimerkiksi –30 000 euroon), kun aineiston ulkopuoliset tekijät huomioidaan ennusteessa, julkisen talouden hyödyn odotusarvon maksimoiva valinta muuttuu. Eli sen sijaan, että potentiaalinen uusi maahanmuuttaja valitaan Suomeen muuttavaksi (kun $\boldsymbol{\epsilon}$:ia **ei** huomioida), potentiaalinen uusi maahanmuuttaja kannattaa jättää valitsematta, kun aineiston ulkopuolisten tekijöiden vaikutukset huomioidaan ennusteessa.

Aineiston ulkopuolisten tekijöiden vaikutukset voivat muuttaa optimaalisen valinnan myös toiseen suuntaan. Jos uuden maahanmuuttajan elinkaariennusteen odotusarvo on etumerkiltään negatiivinen (esimerkiksi –150 000 euroa) ilman aineiston ulkopuolisten tekijöiden vaikutusten huomiointia, mutta muuttuu etumerkiltään positiiviseksi (esimerkiksi +40 000 euroon), kun aineiston ulkopuoliset tekijät otetaan huomioon elinkaariennusteen odotusarvossa, julkisen talouden hyödyn odotusarvon maksimoiva valinta muuttuu. Tällöin sen sijaan, että uutta potentiaalista maahanmuuttajaa ei valita Suomeen muuttavaksi (kun ϵ:ia **ei** huomioida), potentiaalinen uusi maahanmuuttaja kannattaakin valita muuttavaksi Suomeen, kun aineiston ulkopuolisten tekijöiden vaikutukset otetaan huomioon.

Yllä esitetty ehto aineiston ulkopuolisten tekijöiden mahdollisesta vaikutuksesta julkisen talouden optimaaliseen valintaan antaa tutkijalle tai muulle uuden maahanmuuttajan (ehdollisen) elinkaariennusteen laatijalle formaalin määritelmän tutkia sitä, muuttuuko julkisen talouden optimaalinen valinta aineiston ulkopuolisten tekijöiden vaikutuksesta.

Jos uuden maahanmuuttajan elinkaariennusteen jakauman odotusarvo on esimerkiksi –500 000 euroa ilman aineiston ulkopuolisten tekijöiden vaikutuksia, tällöin elinkaariennusteen laatijalla on selkeä kriteeri sille, muuttavatko aineiston ulkopuolisten tekijöiden vaikutukset julkisen talouden kannalta optimaalista valintaa. Aineiston ulkopuolisten tekijöiden on muutettava uuden maahanmuuttajan elinkaariennusteen todennäköisyysjakauman odotusarvoa tässä tapauksessa vähintään +500 000 euroa (positiivisemmaksi), jotta julkisen talouden optimaalinen valinta uuden maahanmuuttajan suhteen muuttuisi. Tällöin elinkaariennusteen odotusarvon etumerkki muuttuisi negatiivisesta positiiviseksi. Eli esimerkin tapauksessa julkisen talouden hyödyn maksimoiva valinta muuttuisi potentiaalisen uuden maahanmuuttajan valitsematta jättämisestä toiseen mahdolliseen valintaan eli siihen, että potentiaalinen uusi maahanmuuttaja valitaan Suomeen muuttavaksi.

12.3.2 Aineiston ulkopuolisten tekijöiden vaikutus yhteen malliparametriin

Esitän tässä luvussa esimerkin kautta erilaisia mahdollisuuksia sille, miten aineiston ulkopuoliset tekijät voivat vaikuttaa yhden malliparametrin todennäköisyysjakaumaan. Osana tätä kokonaisuutta osoitan, että aineiston ulkopuoliset tekijät (ϵ) eivät vaikuta työikäisten nettovaikutusmallin ennustettuun odotusarvoon silloin, kun ei ole tietoa siitä, ovatko ϵ-satunnaisvektorin komponentit todennäköisemmin suurempia vai pienempiä kuin nolla. Käytän työikäisten nettovaikutusmallia esimerkkinä, sillä se on lineaarinen malli ja siten sen tuottamien ennusteiden odotusarvo muodostuu regressiokertoimien jakaumien odotusarvojen summasta. Tietyin edellytyksin tai muutoksin samaa ideaa – aineiston ulkopuolisten tekijöiden vaikutuksia koskien – voidaan soveltaa myös muihin elinkaarivaikutusennusteiden komponentteihin ja niiden malliparametreihin.

Keskityn tarkastelemaan aineiston ulkopuolisten tekijöiden vaikutusta maassaoloajan regressiokertoimeen – jota kutsun tästä eteenpäin pelkästään *maassaoloajaksi* (lyh. *maika*) – työikäisten nettovaikutusmallissa. Kuten edellä esitin, aineiston ulkopuoliset tekijät voivat vaikuttaa maassaoloajan ehdolliseen todennäköisyysjakaumaan kahdesta eri syystä: ensiksi, maassaoloajan vaikutuksen jakauma voi olla vuoden 2011 jälkeen erilainen kuin vuoteen 2011 ulottuvasta tutkimusaineistosta estimoitu jakauma. Olkoon tämä vuoden 2011 jälkeisistä ulkopuolisista tekijöistä aiheutuva vaikutus maassaoloaikaan sa-

tunnaismuuttuja ϵ_{j2011}. Toiseksi, tutkimusaineistossa maassaoloaikahavaintoja on 23 maassaolovuoteen asti, mutta työikäisten nettovaikutusennusteissa maassaoloaikoja voi olla 46 maassaolovuoteen asti. Tällöin maassaoloajan ehdollinen todennäköisyysjakauma voi olla 23 maassaolovuoden jälkeen erilainen, kuin mitä se on 23 maassaolovuoteen asti. Olkoon tämä 23 maassaolovuoden jälkeisistä ulkopuolisista tekijöistä – eli käytännössä ekstrapolaatiosta (lyhennettynä *ep*) – aiheutuva vaikutus maassaoloaikaan satunnaismuuttuja ϵ_{ep}.

Yksinkertaisuuden vuoksi oletan tämän esityksen puitteissa, että aineiston ulkopuolisten tekijöiden vaikutukset ϵ_{j2011} ja ϵ_{ep} ovat riippumattomia aineiston ulkopuolisisten tekijöiden vaikutuksista muihin malliparametreihin. Todetaan vielä, että $\epsilon_{maika} \in \boldsymbol{\epsilon}$ on se aineiston ulkopuolisten tekijöiden vaikutusten satunnaisvektorin komponentti, joka vaikuttaa maassaoloajan regressiokertoimeen. Lisäksi $\epsilon_{j2011} \in \epsilon_{maika}$ ja $\epsilon_{ep} \in \epsilon_{maika}$.

Mikäli aineiston ulkopuolisten tekijöiden vaikutus huomioidaan elinkaariennusteissa, niin niillä ja malliparametreilla on yhteinen posteriorijakauma, joka voidaan lausua ”ketjuttamalla”:

$$p(\boldsymbol{\theta}, \boldsymbol{\epsilon}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) = p(\boldsymbol{\theta}|\boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})p(\boldsymbol{\epsilon}).$$

Maassaoloajan regressiokertoimen ja sitä koskevien aineiston ulkopuolisten tekijöiden vaikutusten yhteinen posteriorijakauma kirjoitetaan siten muodossa:

$$p(\beta_{maika}, \epsilon_{j2011}, \epsilon_{ep}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) = p(\beta_{maika}|\epsilon_{j2011}, \epsilon_{ep}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})p(\epsilon_{j2011}, \epsilon_{ep}).$$

Täten maassaoloajan regressiokertoimen ja aineiston ulkopuolisten tekijöiden vaikutusten osalta on määritettävä seuraavat todennäköisyysjakaumat:

1. $p(\beta_{maika}|\epsilon_{j2011}, \epsilon_{ep}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$.
2. $p(\epsilon_{j2011}, \epsilon_{ep})$.

Ensimmäinen näistä jakaumista on maassaoloajan regressiokertoimen havaintoaineistolla sekä aineiston ulkopuolisten tekijöiden vaikutuksilla ehdollinen jakauma. Aineiston ulkopuolisten tekijöiden vaikutuksia ovat ϵ_{j2011}, eli vuoden 2011 jälkeiset aineiston ulkopuoliset vaikutukset, ja ϵ_{ep}, eli 23 maassaolovuoden ylittävät ekstrapolaatiovaikutukset elinkaariennusteissa.

Toinen jakauma on mainittujen aineiston ulkopuolisten tekijöiden vaikutusten reunajakauma. Oletan tässä, että $p(\epsilon_{j2011})$ ja $p(\epsilon_{ep})$ ovat toisistaan riippumattomia, joskin niille voitaisiin määrittää tilastollinen riippuvuuskin, jos sellainen vaikuttaa elinkaariennusteen laatijan mukaan realistiselta vaihtoehdolta.

Neljä esimerkkiä siitä, miten aineiston ulkopuoliset tekijät voivat vaikuttaa maassaoloajan regressiokertoimen posteriorijakaumaan

Esitän seuraavaksi esimerkinomaisesti 4 erilaista vaihtoehtoa määrittää yllä esitetyt 2 todennäköisyysjakaumaa. Elinkaariennusteen laatija määrittää nämä jakaumat sen mukaan, millaista informaatiota hänellä on on koskien aineiston ulkopuolisia tekijöitä sekä niiden vaikutuksia maassaoloajan regressiokertoimeen. Tämä informaatio tulee tutkimusaineiston ulkopuolelta, koska kyse on nimensä mukaisesti aineiston ulkopuolisten tekijöiden vaikutuksesta.

Neljä esimerkkiä aineiston ulkopuolisten tekijöiden vaikutuksista maassaoloajan regressiokertoimen ehdolliseen todennäköisyysjakaumaan on esitetty alla kuvassa 12.3.3. Kuvan neljässä osakuvassa on jokaisessa niissä pystyakselilla maassaoloajan regressiokerroin, ja vaaka-akselilla on maassaoloaika vuosina. Työikäisten nettovaikutusmallin maassaoloajan regressiokerroin kertoo, miten nettovaikutus muuttuu, kun maassaoloaika kasvaa – eli nettovaikutuksen kulmakertoimen maassaoloajan suhteen. Tästä syystä pystyakseli on nimetty: ΔNettovaikutus/ΔMaassaoloaika. Osakuvissa alueet ovat sitä tummempia, mitä suurempi on maassaoloajan regressiokertoimen todennäköisyystiheys kyseisessä kohdassa. Kuvan alla on selitetty osakuviin viittaavin roomalaisin numeroin otsikoituina, millaisiin elinkaariennusteen laatijan käsityksiin aineiston ulkopuolisten tekijöiden vaikutuksista kunkin osakuvan esittämä maassaoloajan regressiokertoimen ennustejakauma perustuu.

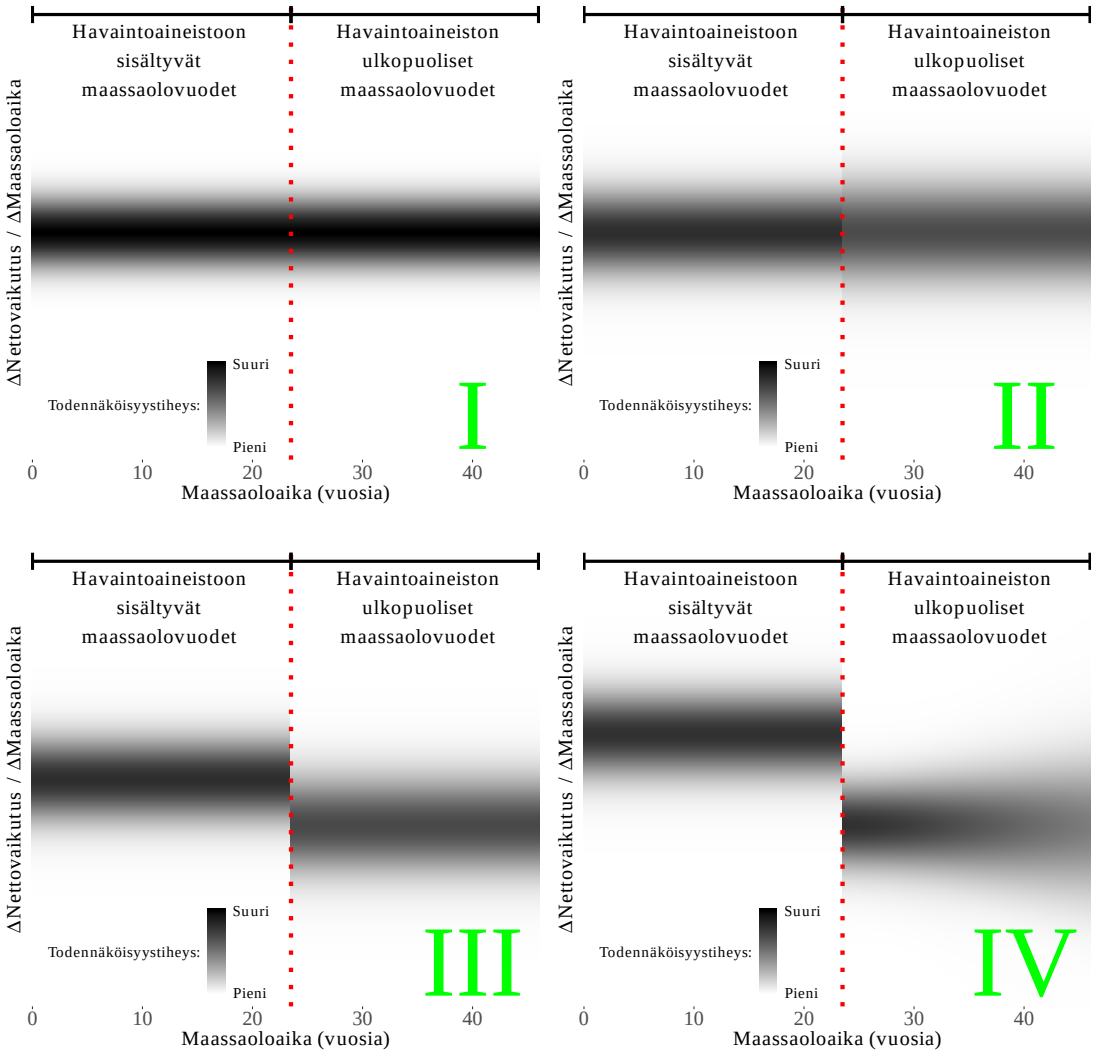
Osakuva I. Ei aineiston ulkopuolisten tekijöiden vaikutuksia.

Osakuvassa I maassaoloajan regressiokertoimeen ei liity aineiston ulkopuolisten tekijöiden vaikutuksia. Tämä voidaan esittää myös niin, että aineiston ulkopuolisten tekijöiden vaikutukset $p(\epsilon_{j2011})$ ja $p(\epsilon_{ep})$ saavat kummatkin vakioarvokseen nollan. Osakuvan I tapauksessa maassaoloajan regressiokertoimeen sisältyvä epävarmuus (tai sitä koskeva informaatio) määräytyy käytännössä täysin tutkimusaineiston pohjalta. Regressiokertoimen posteriorijakauma on likimain normaalijakautunut, koska havaintomäärä on suuri ja regressiokerroin on jatkuva sekä rajoittamaton satunnaismuuttuja.

Regressiokertoimen β_{maika} ja aineiston ulkopuolisten tekijöiden vaikutusten yhteisjakauma on, kun $\epsilon_{j2011} = 0$ ja $\epsilon_{ep} = 0$:

$$p(\beta_{maika}, \epsilon_{j2011}, \epsilon_{ep} | \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim \begin{cases} N(\mu_h, \sigma_h^2), & \text{kun } 0 \leq \textit{maassaoloaika} \leq 23{,}5 \\ N(\mu_h, \sigma_h^2), & \text{kun } 23{,}5 \leq \textit{maassaoloaika} \leq 46. \end{cases} \tag{12.12}$$

Tässä μ_h ja σ_h^2 ovat tutkimusaineistosta (h) estimoidun maassaoloajan posteriorijakauman odotusarvo ja varianssi. Vaikka tutkimusaineistossa maassaoloaika on diskreetti muuttuja, käsittelen sitä tässä esimerkissä jatkuvana muuttujana. Tällöin 23,5:ttä maassaolovuotta pienemmät arvot kuuluvat tutkimusaineistoon, ja siten 23,5:ttä maassaolovuotta suurempien arvojen vaikutus työikäisten nettovaikutusennusteisiin joudutaan ekstrapoloimaan maksimissaan 46 maassaolovuoteen asti.



Kuva 12.1: *Aineiston ulkopuolisten tekijöiden erilaisia vaikutuksia maassaoloajan regressiokertoimen todennäköisyysjakaumaan. Maassaoloajan regressiokertoimen todennäköisyysjakaumien (posteriorijakaumien) todennäköisyystiheydet maassaoloajan eri arvoilla.*

Osakuva II. Lisäepävarmuudet maassaoloajan vaikutukseen.

Ehkä yksinkertaisin tapa määrittää maassaoloajan ehdollinen jakauma, $p(\beta_{maika}|\epsilon_{j2011}, \epsilon_{ep}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, on lisätä aineiston ulkopuoliset vaikutukset pelkästä havaintoaineistosta estimoituun regressiokertoimen β_{maika} posteriorijakaumaan siten, että:

$$p(\beta_{maika}|\epsilon_{j2011}, \epsilon_{ep}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim N(\mu_h, \sigma_h^2) + \epsilon_{j2011} + \epsilon_{ep} = N(\mu_h + \epsilon_{j2011} + \epsilon_{ep}, \sigma_h^2). \tag{12.13}$$

Tällöin aineiston ulkopuolisten tekijöiden vaikutukset yksinkertaisesti siirtävät pelkän tutkimusaineiston

perusteella estimoitua β_{maika}:n jakaumaa. Koska tässä esimerkissä elinkaariennusteen laatija ei tiedä varmasti aineiston ulkopuolisten tekijöiden vaikutusten suuruutta, ϵ_{j2011} ja ϵ_{ep} ovat myös satunnaismuuttujia, joilla on omat todennäköisyysjakaumansa. Tämä tarkoittaa käytännössä sitä, että näiden vaikutusten lisääminen yllä esitetyllä tavalla β_{maika}:n pelkän aineiston avulla estimoituun jakaumaan, lisää maassaoloajan regressiokertoimeen liittyviä epävarmuuksia verrattuna tilanteeseen, jossa näitä vaikutuksia ei ole (vrt. osakuva I). Toisin sanoen maassaoloajan regressiokertoimen todennäköisyystiheys ei ole yhtä keskittynyt kuin osakuvassa I.

Aineiston ulkopuolisten tekijöiden vaikutuksille ϵ_{j2011} ja ϵ_{ep} on nyt määritettävä omat reunajakaumansa. Kuvaan näihin satunnaismuuttujiin liittyviä epävarmuuksia Studentin T-jakaumilla, joiden odotusarvot ovat nollia ja joiden vapausasteet ovat ”pieniä”, näissä esimerkeissä 4. Tässä tapauksessa elinkaariennusteen laatijalla ei ole aineiston ulkopuolisten tekijöiden vaikutusten suhteen sellaista informaatiota, joka kertoisi, onko maassaoloajan regressiokerroin tutkimusaineiston ”ulkopuolella” todennäköisemmin positiivisempi vai negatiivisempi kuin pelkän tutkimusaineiston perusteella estimoitu jakauma. Tästä syystä aineiston ulkopuolisten tekijöiden vaikutusten jakaumien odotusarvot ovat nollia, ja ne ovat symmetrisiä nollan suhteen. Elinkaariennusteen laatija haluaa kuitenkin pitää mahdollisina myös suuriakin poikkeamia, ja siksi käytetään T-jakaumia, joiden vapausasteet ovat pieniä.

Näin ollen regressiokertoimen β_{maika} ja aineiston ulkopuolisten tekijöiden vaikutusten yhteisjakauma on osakuvassa II:

$$p(\beta_{maika}, \epsilon_{j2011}, \epsilon_{ep} | \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim \tag{12.14}$$
$$\begin{cases} N(\mu_h + \epsilon_{j2011}, \sigma_h^2)\, T(4, 0, \sigma_{\epsilon_j2011}), & \text{kun } 0 \leq \textit{maassaoloaika} < 23{,}5 \\ N(\mu_h + \epsilon_{j2011} + \epsilon_{ep}, \sigma_h^2)\, T(4, 0, \sigma_{\epsilon_j2011})\, T(4, 0, \sigma_{\epsilon_ep}), & \text{kun } 23{,}5 \leq \textit{maassaoloaika} \leq 46. \end{cases}$$

Tässä $T(4, 0, \sigma_{\epsilon_j2011})$ on vuoden 2011 jälkeisten vuosien aineiston ulkopuolisten tekijöiden vaikutuksen reunajakauma, jolla on Studentin T-jakauma vapausasteilla 4, odotusarvolla 0 ja hajontaparametrilla σ_{ϵ_j2011}. $T(4, 0, \sigma_{\epsilon_ep})$ on puolestaan ekstrapolaatioon (maassaoloajat $\geq 23{,}5$) liittyvien aineiston ulkopuolisten tekijöiden vaikutuksen, ϵ_{ep}, T-jakautunut reunajakauma.

Nyt maassaoloajan regressiokertoimeen liittyy σ_{ϵ_j2011}:n tuoma lisäepävarmuus, kun maassaoloaika $< 23{,}5$. Kun maassaoloaika $\geq 23{,}5$, niin maassaoloajan regressiokertoimeen liittyy edelleen σ_{ϵ_j2011}:n tuoma lisäepävarmuus, mutta sen lisäksi myös ekstrapolaation tuoma lisäepävarmuus, ϵ_{ep}.

Lisäepävarmuudet näkyvät kuvassa 12.3.3, kun verrataan osakuvien I ja II esimerkkitilanteita toisiinsa. Maassaoloajan regressiokertoimen jakauma on II-osakuvassa leveämpi, mikä näkyy siinä, että todennäköisyystiheys on ”himmeämpi” osakuvassa II niissä kohdissa, joissa tiheys on kuitenkin suurimmillaan, verrattuna osakuvaan I. Kun ollaan yli 23,5:n maassaolovuoden, niin epävarmuus on vielä suurempaa osakuvassa II, mikä johtuu ekstrapolaation tuomasta lisäepävarmuudesta.

Sekä osakuvia I ja II vertaamalla että yllä esitetyistä jakaumamääritelmistä selviää, että sekä I- että II-tapauksessa maassaoloajan regressiokertoimen ehdollisen jakauman odotusarvo on sama eli aineistosta estimoitu μ_h. Tämä seuraa siitä, että kummassakin tapauksessa aineiston ulkopuolisten tekijöiden vaikutusten odotusarvot ovat nollia. Tämä tarkoittaa myös sitä, että aineiston ulkopuolisilla tekijöillä ei ole tässä tapauksessa vaikutusta myöskään työikäisten nettovaikutusennusteiden jakauman *odotusarvoon*, vaikka

ne lisäävätkin maassaoloajan regressiokertoimen lisäepävarmuuden kautta näihin ennusteisiin sisältyviä epävarmuuksia.

Ei vaikutusta potentiaalisen uuden maahanmuuttajan valintaan. Kaiken kaikkiaan aineiston ulkopuoliset tekijät eivät muuta työikäisten nettovaikutusennusteiden – ja siten elinkaarivaikutusten ennusteiden – odotusarvoa, kun aineiston ulkopuolisten tekijöiden vaikutusten reunajakaumien odotusarvot ovat nollia. Koska potentiaalisen uuden maahanmuuttajan valinnan hyödyn maksimoiva päätös riippuu juuri elinkaarivaikutusten odotusarvosta, tässä tapauksessa aineiston ulkopuolisten tekijöiden tuoma lisäepävarmuus ennusteisiin ei muuta hyödyn maksimoivaa päätöstä. Toisin sanoen sellaisissa ennustetilanteissa, joissa elinkaariennusteiden laatijalla ei ole informaatiota siitä, ovatko aineiston ulkopuolisten tekijöiden vaikutukset todennäköisemmin nollaa pienempiä vai suurempia, hyödyn maksimoiva päätös ei muutu, ja päätös perustuu tutkimusaineiston tarjoamaan informaatioon.

Osakuva III. Ekstrapoloitu maassaolovaikutus vain aineiston ulkopuolisesta informaatiosta.

Esimerkkitapauksessa III maassaoloajan regressiokertoimen ja aineiston ulkopuolisten tekijöiden vaikutusten yhteisjakauma on muuten samanlainen kuin esimerkissä II, mutta nyt elinkaariennusteen laatija tietää, että tutkimusaineisto ei kerro mitään maassaolovaikutuksen jakauman odotusarvosta, kun se koskee ekstrapoloitavia maassaolovuosia. Olen kirjoittanut tämän jakaumalausekkeisiin niin, että havaintoaineistosta estimoitu odotusarvo μ_h kerrotaan nollalla, kun on kyse maassaolovaikutuksen ekstrapoloitavasta osuudesta:

$$p(\beta_{maika}, \epsilon_{j2011}, \epsilon_{ep} | \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim \tag{12.15}$$
$$\begin{cases} N(\mu_h + \epsilon_{j2011}, \sigma_h^2)\, T(4, 0, \sigma_{\epsilon_j2011}), & \text{kun } 0 \leq \textit{maassaoloaika} < 23{,}5 \\ N(\mu_h \times 0 + \epsilon_{j2011} + \epsilon_{ep}, \sigma_h^2)\, T(4, 0, \sigma_{\epsilon_j2011})\, T(4, 0, \sigma_{\epsilon_ep}), & \text{kun } 23{,}5 \leq \textit{maassaoloaika} \leq 46. \end{cases}$$

Yllä osakuviossa III tämä näkyy siten, että maassaoloajan vaikutuksen jakauma sijaitsee ekstrapoloitavalla maassaoloaikojen osuudella (yli 23,5 vuotta) alemmalla tasolla kuin osakuvissa I ja II. Tämä johtuu siitä, että kun maassaoloajan vaikutuksen odotusarvo estimoidaan aineistosta, se on käytännössä aina positiivinen. Nyt, kun maassaoloajan jakauma ja sen odotusarvo määräytyvät vain aineiston ulkopuolisten tekijöiden vaikutuksesta, jakauma sijaitsee alemmalla tasolla, koska aineiston ulkopuolisten tekijöiden vaikutusten reunajakaumien odotusarvot ovat nollia – ja siten myös maassaoloajan vaikutuksen jakauman odotusarvo.

Osakuvassa III maassaoloajan vaikutuksen jakauma on ekstrapoloitavalla maassaoloaikojen osuudella (yli 23,5 vuotta) hieman kapeampi tai enemmän keskittynyt kuin osakuvassa II. Tämä johtuu siitä, että osakuvassa III maassaoloajan vaikutukseen ei sisälly ekstrapoloitavalla lainkaan havaintoaineiston tuomaa epävarmuutta, vaan kaikki epävarmuus siihen tulee havaintoaineiston ulkopuolisten tekijöiden vaikutuksesta (yhtä suurina kuin osakuvassa II).

Osakuvio IV. Maassaoloajan vaikutus positiivisempi kuin aineiston perusteella, ekstrapolaatioon liittyvä epävarmuus kasvaa maassaoloajan mukaan.

Esimerkkitapauksessa IV jakaumat ovat samanlaiset kuin esimerkissä III lukuun ottamatta kahta muu-

tosta. Ensiksi, nyt elinkaariennusteen laatijalla on sellaista aineiston ulkopuolista informaatiota koskien aikaa vuosien 2011 jälkeen, että hän tietää maassaoloajan regressiokertoimen odotusarvon olevan vuoden 2011 jälkeen μ_{ϵ_j2011}:n verran suurempi, kuin mitä se olisi pelkästään tutkimusaineistosta estimoituna. Tämän vuoksi ϵ_j2011:n reunajakauma saa nollan sijasta odotusarvoksi μ_{ϵ_j2011}:n, joka on tässä tapauksessa nollaa suurempi. Kuvion 12.3.3 osakuvassa IV tämä näkyy siinä, että maassaoloajan jakauma on siirtynyt korkeammalle tasolle kuin muissa osakuvioissa, kun tarkastellaan tutkimusaineistoon kuuluvia maassaoloaikoja.

Nyt elinkaariennusteen laatija haluaa myös, että ekstrapolaation vaikutukseen liittyvä epävarmuus kasvaa, mitä kauemmaksi tulevaisuuteen edetään tutkimusaineistoon kuuluvista maassaoloajan vuosista. Jotta tämä tulisi huomioitua elinkaariennusteissa, ekstrapolaatioon liittyvien aineiston ulkopuolisten tekijöiden vaikutusten jakauman hajontaparametrin arvo riippuu nyt maassaoloajasta: $\sigma_{\epsilon_ep}(\textit{maassaoloaika})$. Tässä esimerkkitapauksessa tämä hajontaparametri kasvaa lineaarisesti siten, että: $\sigma_{\epsilon_ep}(\textit{maassaoloaika}) = \sigma_{\epsilon_ep0} + k \times (\textit{maassaoloaika} - 23{,}5)$, missä σ_{ϵ_ep0} on hajontaparametrin vakiotaso osakuviossa III ja k on jokin positiivinen reaaliluku. Osakuvassa IV tämä näkyy siinä, että maassaoloajan regressiokertoimen jakauman leveys kasvaa maassaoloajan kasvaessa, kun tarkastellaan tutkimusaineiston ulkopuolelle sijoittuvia maassaoloaikoja.

Todennäköisyysjakaumat ovat näin ollen tässä esimerkkitapauksessa:

$$p(\beta_{maika}, \epsilon_{j2011}, \epsilon_{ep} | \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim \tag{12.16}$$
$$\begin{cases} N(\mu_h + \epsilon_{j2011}, \sigma_h^2)\, T(4, \mu_{\epsilon_j2011}, \sigma_{\epsilon_j2011}), \text{kun } 0 \leq \textit{maassaoloaika} < 23{,}5 \\ \\ N(\mu_h \times 0 + \epsilon_{j2011} + \epsilon_{ep}, \sigma_h^2)\, T(4, 0, \sigma_{\epsilon_j2011})\, T(4, 0, \sigma_{\epsilon_ep}(\textit{maassaoloaika})), \\ \text{kun } 23{,}5 \leq \textit{maassaoloaika} \leq 46. \end{cases}$$

12.3.3 Aineiston ulkopuolisten tekijöiden vaikutusten vektori ja elinkaarivaikutusten ennusteiden todennäköisyysjakauma

Edellä tutkittiin aineiston ulkopuolisten tekijöiden vaikutusta maassaoloajan regressiokertoimen todennäköisyysjakaumaan. Samalla tavoin kaikille elinkaarivaikutusten komponenttien malliparametreille $\boldsymbol{\theta}$ voidaan määrittää ehdolliset todennäköisyysjakaumat, jotka riippuvat aineiston ulkopuolisten tekijöiden vaikutuksista. Koska olen esittänyt elinkaarivaikutusten komponenttien posterioriennustejakauman jo aiemmin tässä luvussa sekä tämän tutkimuksen johdannossa luvussa 2.2.5, en perustele elinkaariennustetta koskevia jakaumavalintoja tässä enää sen tarkemmin. Joka tapauksessa aineiston ulkopuolisten tekijöiden vaikutusten jakaumavalintoja on olemassa erilaisia, niin kuin esitin yllä maassaoloajan kohdalla.

Otetaan tässä esimerkiksi työikäisten nettovaikutusmallin kaikki parametrit. Koska lähes kaikki malliparametrit ovat likimain normaalijakautuneita työikäisten nettovaikutusmallissa, kaikkien malliparametrien ehdollinen yhteisjakauma on lähellä d-ulotteista normaalijakaumaa, missä d on malliparametrien määrä:

$$p(\boldsymbol{\theta}_{nettovaikutus} | \boldsymbol{\epsilon}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha}) \sim N_d(\boldsymbol{\mu_h}, \boldsymbol{\Sigma}_h^2) + \boldsymbol{\epsilon} = N_d(\boldsymbol{\mu_h} + \boldsymbol{\epsilon}, \boldsymbol{\Sigma}_h^2). \tag{12.17}$$

Aineiston ulkopuolisten tekijöiden reunajakauma on puolestaan d-ulotteinen Studentin T-jakauma, jonka vapauasteet ovat 4:

$$\boldsymbol{\epsilon} \sim T_d(4, \boldsymbol{\mu}_\epsilon, \boldsymbol{\Sigma}_\epsilon), \tag{12.18}$$

Lisäksi tilanteessa, jossa elinkaariennusteen laatijalla ei ole sellaista informaatiota aineiston ulkopuolisten tekijöiden vaikutuksia koskien, joka kertoisi, ovatko vaikutukset todennäköisemmin nollaa suurempia vai nollaa pienempiä, aineiston ulkopuolisten tekijöiden todennäköisyysjakauman odotusarvovektori on d-ulotteinen 0-vektori:

$$E_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}] = \boldsymbol{\mu}_\epsilon = \mathbf{0}_{\mathrm{d}}. \tag{12.19}$$

Tämä tarkoittaa sitä, että elinkaariennusteen laatija ei tiedä, siirtyvätkö malliparametrien tutkimusaineistosta estimoidut jakaumat aineiston ulkopuolisten tekijöiden vaikutuksesta todennäköisemmin vasemmalle vai oikealle – eli saavatko ne pienempiä vai suurempia arvoja kuin tutkimusaineistosta estimoitu jakauma. Tällaista epävarmuutta kannattaa mielestäni kuvata todennäköisyysjakaumilla, jotka on keskistetty nollaan ja jotka ovat symmetrisiä nollan suhteen.

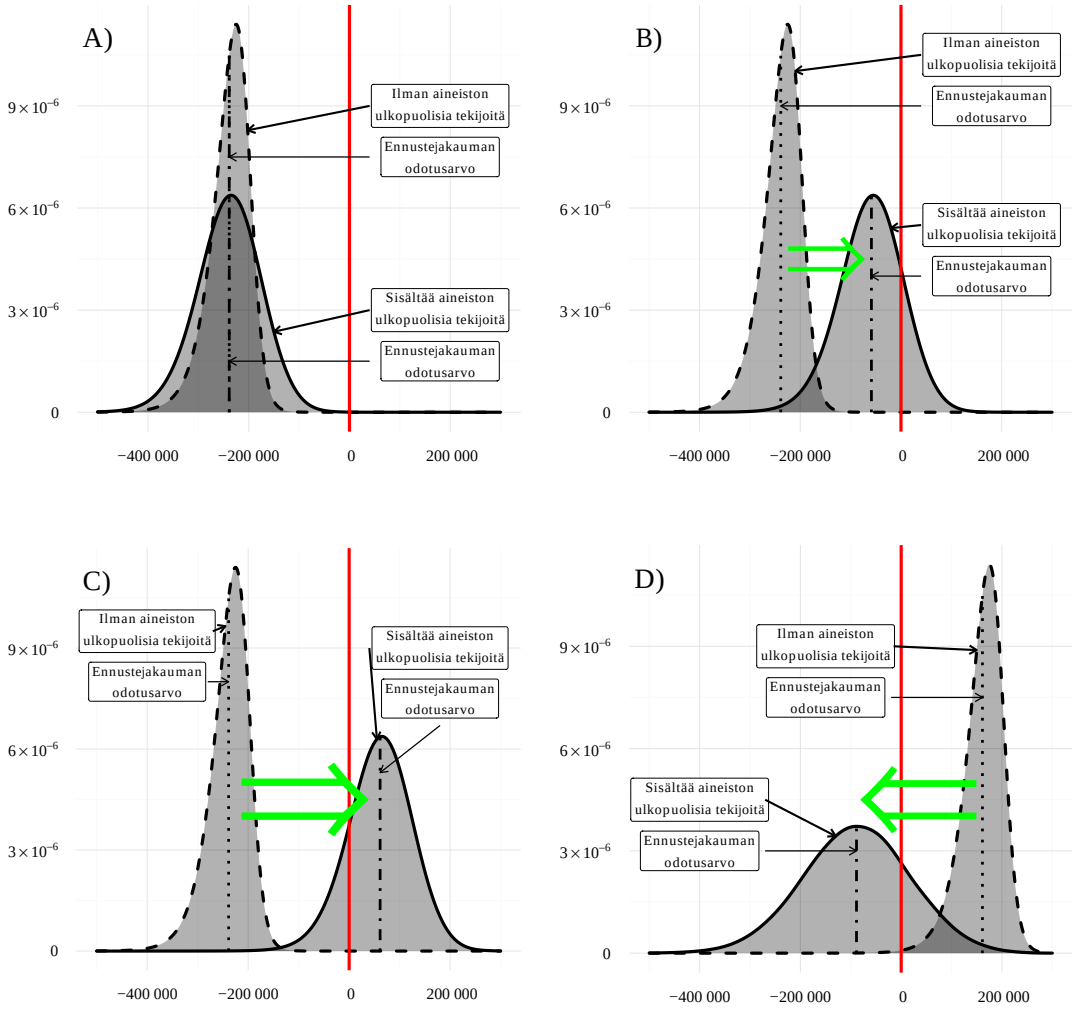
Neljä esimerkkiä siitä, miten aineiston ulkopuoliset tekijät voivat vaikuttaa uusien maahanmuuttajien elinkaariennusteiden todennäköisyysjakaumiin ja siten vaikuttaa uusien maahanmuuttajien valintaan.

Tarkastellaan nyt kuvioiden avulla esimerkinomaisia tilanteita, joissa aineiston ulkopuolisten tekijöiden vaikutusten huomiointi muuttaa elinkaarivaikutusten ennusteiden todennäköisyysjakaumia. Erityistä huomiota kiinnitän siihen, millainen vaikutus aineiston ulkopuolisten tekijöiden huomioinnilla on elinkaariennusteiden jakaumien odotusarvoihin. Mikäli jakauman odotusarvon – eli julkisen talouden odotetun hyödyn – etumerkki muuttuu aineiston ulkopuolisten tekijöiden vaikutuksesta, tällöin päätös koskien potentiaalisen uuden maahanmuuttajan valintaa vaihtuu. Kuten edellä määritettiin, potentiaalinen uusi maahanmuuttaja kannattaa valita muuttavaksi Suomeen, jos uuden maahanmuuttajan elinkaariennusteen odotusarvo on positiivinen. Jos puolestaan odotusarvo on negatiivinen, potentiaalista uutta maahanmuuttajaa ei kannata valita muuttavaksi Suomeen.

Kuvassa 12.2 alla on esitetty neljä erilaista tilannetta, joissa aineiston ulkopuoliset tekijät muuttavat elinkaarivaikutusten ennusteiden todennäköisyysjakaumia. Osakuvissa A),B),C) ja D) on esitetty katkoviivalla elinkaarivaikutusten ennustejakaumat silloin, kun aineiston ulkopuolisia tekijöitä ei huomioida. Elinkaarivaikutusten ennustejakaumat kiinteällä viivalla ovat puolestaan jakaumia, joissa aineiston ulkopuolisten tekijöiden vaikutukset on huomioitu. Kuvioista ilmenee siis, kuinka aineiston ulkopuolisten tekijöiden vaikutukset voivat muuttaa elinkaarivaikutusten jakaumia, kun näiden vaikutusten aiheuttamat muutokset huomioidaan. Muutokset riippuvat siitä, miten aineiston ulkopuolisten tekijöiden vaikutusten reunajakauma $p(\boldsymbol{\epsilon})$ ja sen vaikutukset malliparametrien jakaumiin on määritetty. Osakuvissa A),B),C) ja D) on esitetty pystyviivoilla jokaisen esimerkkijakauman odotusarvo.

Elinkaarivaikutusten jakaumat kuvassa 12.2 ovat keksittyjä esimerkkejä, mutta jakaumat on esitetty jotakuinkin realistisella skaalalla vaaka-akselin eurojen suhteen (todelliset jakaumat ovat vielä leveämpiä). Pystyakseli osakuvissa A),B),C) ja D) on todennäköisyystiheys. Punainen pystyviiva näissä kuvissa on nollan euron kohdalla. Se on päätöksentekoteorian perusteella kriittinen raja sille, kannattaako potentiaalista uutta maahanmuuttajaa valita Suomeen muuttavaksi. Mikäli ennustejakauman odotusarvo on nollan

vasemmalla puolella, potentiaalista uutta maahanmuuttajaa ei kannata valita Suomeen muuttavaksi. Mikäli taas ennustejakauman odotusarvo on nollan oikealla puolella, potentiaalinen uusi maahanmuuttaja kannattaa valita Suomeen muuttavaksi. Vihreät nuolet kuvioissa B),C) ja D) esittävät ennustejakaumien siirtymiä aineiston ulkopuolisten tekijöiden vaikutuksesta. Kuviossa A) nuolta ei ole, sillä aineiston ulkopuoliset tekijät huomioiva jakauma ei siirry verrattuna jakaumaan, jossa näitä vaikutuksia ei ole huomioitu.



Kuva 12.2: *Aineiston ulkopuolisten tekijöiden erilaisia vaikutuksia elinkaarivaikutusten ennustejakaumiin. Vihreä nuoli kuvaa vaikutusten aiheuttamia jakaumien siirtymiä. Pystyakselilla on todennäköisyystiheys, ja vaaka-akselilla ovat mahdolliset elinkaarivaikutukset euroina. Punainen pystyviiva on nollan euron kohdalla.*

Osakuva A). Elinkaarivaikutusten ennustejakauma ei siirry.

Osakuvassa A) elinkaariennusteen jakauma ei siirry, kun aineiston ulkopuolisten tekijöiden vaikutukset huomioidaan ennustejakaumassa. Tämä ilmenee osakuvassa A) siinä, että kummallakin esitetyllä ennustejakaumalla on sama odotusarvo. Aineiston ulkopuolisten tekijöiden vaikutusten huomiointi elinkaariennusteessa lisää kuitenkin ennusteeseen sisältyvää epävarmuutta, kun sitä verrataan ennustejakaumaan, jossa näitä tekijöitä ei ole huomioitu. Tämä näkyy osakuvassa A) siinä, että ennustejakauma, jossa aineiston ulkopuolisten tekijöiden vaikutukset on huomioitu, on leveämpi kuin ennustejakauma, jossa näitä vaikutuksia ei ole huomioitu. Kaiken kaikkiaan potentiaalisen uuden maahanmuuttajan valitsemisen kannalta tämä tarkoittaa sitä, että aineiston ulkopuoliset tekijät eivät vaikuta uuden maahanmuuttajan ehdollisen elinkaariennusteen jakauman odotusarvoon. Näin ollen ne eivät myöskään vaikuta julkisen talouden hyödyn odotusarvon maksimoivaan valintaan sitä päätöstä koskien, että valitaanko potentiaalinen uusi maahanmuuttaja Suomeen muuttavaksi.

Osakuvan A) ennustejakaumat syntyvät esimerkiksi sellaisessa tilanteessa, jonka olen esittänyt edellä työikäisten nettovaikutusmallin maassaoloajan vaikutusta koskevassa esimerkissä (kuvan osakuvassa II). Kun elinkaariennusteen laatijalla ei ole informaatiota sen suhteen, että ovatko aineiston ulkopuolisten tekijöiden vaikutukset malliparametreihin todennäköisemmin nollaa suurempia vai pienempiä, työikäisten nettovaikutusten ennustejakauman odotusarvo ei muutu aineiston ulkopuolisten tekijöiden vaikutuksesta. Toisin sanoen kun aineiston ulkopuolisten tekijöiden vaikutusten vektorin, $p(\boldsymbol{\epsilon})$, todennäköisyysjakauman odotusarvo on nollavektori, kaikkien nettovaikutusmallin regressiokertoimien odotusarvot määräytyvät *vain* havaintoaineistoon pohjautuviin estimaatteihin. Aineiston ulkopuoliset tekijät voivat lisätä regressiokertoimiin liittyviä epävarmuuksia, eli leventää niiden todennäköisyysjakaumia, mutta niiden odotusarvoihin aineiston ulkopuoliset tekijät eivät vaikuta silloin, kun aineiston ulkopuolisten tekijöiden jakauman odotusarvovektori on nollavektori.

Tällöin myöskään elinkaarivaikutusten ennustejakauman odotusarvo ei muutu, kun muiden elinkaariennusteiden komponenttien jakaumat (esimerkiksi palkkasummien ennustejakauma) pidetään samoina. Tällaisessa tapauksessa kuitenkin elinkaariennusteen jakauman varianssi kasvaa aineiston ulkopuolisten tekijöiden vaikutuksen tuoman lisäepävarmuuden myötä, joten päädytään osakuvan A) esittämään tilanteeseen.

Osakuva B). Elinkaarivaikutusten ennustejakauma siirtyy oikealle, hyödyn maksimoiva päätös ei kuitenkaan muutu.

Osakuvassa B) ennustejakauma siirtyy oikealle, kun aineiston ulkopuolisten tekijöiden vaikutus huomioidaan. Tällöin ennustejakauman, jossa nämä vaikutukset huomioidaan, odotusarvo on positiivisempi kuin ennustejakaumassa, jossa näitä vaikutuksia ei huomioida. Käyttäen aiemmin tässä luvussa määritettyjä odotusarvoja, tämä voidaan ilmaista:

$E(\text{elinkaari, ilman } \boldsymbol{\epsilon}) < E(\text{elinkaari, mukana } \boldsymbol{\epsilon})$.

Tällainen tilanne syntyy esimerkiksi silloin, kun elinkaariennusteen laatijalla on sellaista aineiston ulkopuolista informaatiota, että joidenkin (normaalijakautuneiden) regressiokertoimien jakaumien odotusarvot ovat tulevaisuudessa positiivisempia, kuin mitä ne ovat pelkästään tutkimusaineistosta estimoituina. Olen edellä esittänyt esimerkin tästä maassaoloajan osalta kuvan 12.3.3 osakuvassa IV.

Vaikka ennustejakauma osakuvassa B) siirtyy oikealle, kun aineiston ulkopuolisten tekijöiden vaikutukset huomioidaan, tämä ei kuitenkaan muuta julkisen talouden odotetun hyödyn maksimoivaa valintaa potentiaalisen uuden maahanmuuttajan osalta. Koska ennustejakauman odotusarvo on edelleen nollaa pienempi, potentiaalista uutta maahanmuuttajaa ei kannata valita muuttavaksi Suomeen. Se, ettei odotetun hyödyn maksimoiva valinta muutu aineiston ulkopuolisten tekijöiden vaikutuksesta, voidaan ilmaista hieman formaalimmin seuraavasti osakuvan B) osalta:

$\text{sign}(E(\text{elinkaari, ilman } \boldsymbol{\epsilon})) = \text{sign}(E(\text{elinkaari, mukana } \boldsymbol{\epsilon}))$.

Osakuvassa B) ennustejakauman odotusarvo saa edelleen saman etumerkin – eli miinuksen – kun aineiston ulkopuolisten tekijöiden vaikutus huomioidaan. Aiemmin tässä luvussa on määritetty, että aineiston ulkopuolisten tekijöiden vaikutukset muuttavat odotetun hyödyn maksimoivan valinnan vain silloin, kun ennustejakauman etumerkki vaihtuu niiden vaikutuksesta.

Osakuva C). Elinkaarivaikutusten ennustejakauma siirtyy oikealle, hyödyn maksimoiva päätös muuttuu.

Osakuvan C) esimerkkitilanne muistuttaa kuvan B) tilannetta siinä, että myös nyt ennustejakauman odotusarvo muuttuu positiivisemmaksi, kun aineiston ulkopuolisten tekijöiden vaikutukset huomioidaan ennusteissa:

$E(\text{elinkaari, ilman } \boldsymbol{\epsilon}) < E(\text{elinkaari, mukana } \boldsymbol{\epsilon})$.

Nyt, toisin kuin osakuvan B) esimerkissä, myös odotetun hyödyn maksimoiva valinta muuttuu, sillä:

$\text{sign}(E(\text{elinkaari, ilman } \boldsymbol{\epsilon})) \neq \text{sign}(E(\text{elinkaari, mukana } \boldsymbol{\epsilon}))$.

Toisin sanoen elinkaariennusteen jakauman odotusarvon etumerkki muuttuu miinusmerkkisestä plusmerkkiseksi, kun $\boldsymbol{\epsilon}$ huomioidaan ennusteessa. Tässä tapauksessa $\boldsymbol{\epsilon}$:n huomiointi ennusteessa vaikuttaa päätökseen niin, että potentiaalinen uusi maahanmuuttaja kannattaakin valita Suomeen muuttavaksi.

Osakuva D). Elinkaarivaikutusten ennustejakauma siirtyy vasemmalle, hyödyn maksimoiva päätös muuttuu.

Osakuvan D) esimerkkitilanteessa potentiaalisen uuden maahanmuuttajan elinkaarivaikutusten ennustejakauman odotusarvo on positiivinen, kun aineiston ulkopuolisten tekijöiden vaikutusta ei huomioida ennusteessa. Tässä esimerkkitilanteessa $\boldsymbol{\epsilon}$:n huomiointi ennustejakaumassa siirtää ennustejakaumaa vasemmalle, sillä:

$E(\text{elinkaari, ilman } \boldsymbol{\epsilon}) > E(\text{elinkaari, mukana } \boldsymbol{\epsilon})$.

Myös odotetun hyödyn maksimoiva valinta muuttuu, sillä:

$\text{sign}(E(\text{elinkaari, ilman } \boldsymbol{\epsilon})) \neq \text{sign}(E(\text{elinkaari, mukana } \boldsymbol{\epsilon}))$.

Toisin sanoen elinkaariennusteen jakauman odotusarvon etumerkki muuttuu plusmerkkisestä miinusmerkkiseksi, kun $\boldsymbol{\epsilon}$ huomioidaan ennusteessa. Tässä tapauksessa $\boldsymbol{\epsilon}$:n huomiointi ennusteessa vaikuttaa päätökseen niin, että potentiaalista uutta maahanmuuttajaa ei kannatakaan valita Suomeen muuttavaksi.

12.3.4 Aineiston ulkopuolisten tekijöiden vaikutukset ja elinkaarivaikutusten ennusteet tässä tutkimuksessa

Käyn tässä lyhyesti läpi, kuinka aineiston ulkopuoliset tekijät vaikuttaisivat tässä tutkimuksessa elinkaariennusteiden todennäköisyysjakaumiin, *mikäli* ne huomioitaisiin esitetyissä elinkaariennusteissa. Esitetyissä elinkaariennusteiden tuloksissa ei huomioida aineiston ulkopuolisten tekijöiden tuomia lisäepävarmuuksia elinkaariennusteisiin lukuun ottamatta lukua 13.2.4, jossa havainnollistan lisäepävarmuuksien tuomisen vaikutuksia elinkaariennusteisiin.

Epävarmuuksien lisäämisen vaikutus yksittäisiin malliparametreihin. Mikäli elinkaariennusteisiin sisällytettäisiin lisäepävarmuuksia (kuten luvussa 13.2.4 tehdään), tässä tutkimuksessa käytetty ratkaisu niiden suhteen perustuu siihen, että minulla ei ole sellaista tutkimusaineiston ulkopuolista informaatiota, joka kertoisi, ovatko malliparametrien $\boldsymbol{\theta}$ ehdolliset todennäköisyysjakaumat todennäköisemmin positiivisempia tai negatiivisempia kuin tutkimusaineistosta estimoidut jakaumat, kun huomioon otetaan aineiston ulkopuolisten tekijöiden vaikutus $\boldsymbol{\epsilon}$. Esimerkiksi työikäisten nettovaikutusmallin maassaoloajan regressiokertoimen kohdalla en tiedä, onko sen jakauma positiivisempi vai negatiivisempi kuin pelkästään tutkimusaineistosta estimoitu jakauma, kun tähän parametriin kohdistuvat aineiston ulkopuolisten tekijöiden vaikutukset ϵ_{j2011} ja ϵ_{ep} otetaan huomioon niin, kuin olen esittänyt alaluvussa 12.3.2. Näiden vaikutusten huomiointi maassaoloajan regressiokertoimen kohdalla olisi samanlainen, kuin mitä on esitetty edellä kuvan 12.3.3 osakuvassa II. Maassaoloajan regressiokertoimeen liittyvä epävarmuus kasvaisi näiden vaikutusten seurauksena, mutta jakauman odotusarvo ei muuttuisi.

Epävarmuuksien lisäämisen vaikutus elinkaariennusteiden jakaumiin. Kun samaa periaatetta sovelletaan myös muiden malliparametrien todennäköisyysjakaumiin, päädytään tilanteeseen, jossa malliparametreja koskeva epävarmuus kasvaa, mutta niiden odotusarvot eivät muutu. Elinkaariennusteiden jakaumien osalta tämä tarkoittaa yllä kuvan 12.2 osakuviossa A) esitettyä tilannetta. Elinkaariennusteiden jakaumista tulisi leveämpiä, eli ne sisältäisivät enemmän epävarmuuksia, kun aineiston ulkopuolisten tekijöiden vaikutukset otetaan huomioon. Silti näiden jakaumien odotusarvot eivät muuttuisi verrattuna ennustejakaumiin, joissa $\boldsymbol{\epsilon}$:ia ei ole huomioitu. Tällöin aineiston ulkopuolisten tekijöiden vaikutusten huomioon ottamisella ei myöskään ole vaikutusta siihen, kannattaako potentiaalista uutta maahanmuuttajaa valita Suomeen muuttavaksi vai ei. Tämä johtuu siitä, että julkisen talouden odotetun hyödyn maksimoiva päätös riippuu vain potentiaalisen uuden maahanmuuttajan elinkaarivaikutuksen ennustejakauman odotusarvon etumerkistä, kuten tässä luvussa on aiemmin esitetty.

Esimerkki epävarmuuden lisäämisestä elinkaariennusteisiin. Vuoteen 2015 diskontattujen elinkaariennusteiden tulosten yhteydessä esitän luvussa 13.2.4 Somaliassa ja Irakissa syntyneille uusille maahanmuuttajille diskontatut elinkaariennusteet, joihin on lisätty epävarmuutta yllä mainitun esimerkin mukaisesti. Lisään näissä diskontatuissa elinkaariennusteissa nettovaikutusennustemallin maassaoloajan regressiokertoimeen ”paljon” lisää epävarmuutta. Diskontattujen elinkaariennusteiden odotusarvot eivät kuitenkaan muutu, koska lisättävien epävarmuustermien odotusarvot ovat nollia. Tällöin myöskään Irakissa ja Somaliassa syntyneitä koskevia potentiaalisia uusia maahanmuuttajia koskevat valintaratkaisut

eivät muutu verrattuna elinkaariennusteisiin, joissa mainittua lisäepävarmuutta ei ole. Vaikka tässä tapauksessa lisäepävarmuudet eivät muuta diskontattujen elinkaariennusteiden jakaumien odotusarvoja, ne kuitenkin kasvattavat elinkaariennusteiden jakaumien keskihajontoja, eli lisäävät elinkaariennusteisiin sisältyviä epävarmuuksia.

12.4 Potentiaalisten uusien maahanmuuttajien valinta: ennustetun hyödyn odotusarvon maksimointi yli ajan

Malliparametrien posteriorijakaumien päivitys uuden havaintoaineiston informaatiolla. Elinkaariennusteen komponenttien mallien parametrit ($\boldsymbol{\theta}$) (niiden posteriorijakaumat) on estimoitu havaintoaineiston perusteella, joka koostuu vuosiin 1988–2011/2014 sijoittuvista henkilövuositason rekisteritiedoista, niin kuin tutkimuksessa on aiemmin esitetty. Havaintoaineistoa kuitenkin karttuu ajan kuluessa jokaisena uutena (kalenteri)vuotena: sekä niiden maahanmuuttajien osalta, jotka olivat jo Suomessa uuden (kalenteri)vuoden alkaessa, että niiden uusien maahanmuuttajien osalta, jotka muuttivat Suomeen uuden (kalenteri)vuoden aikana. Tätä karttunutta havaintoaineistoa voidaan käyttää malliparametrien ($\boldsymbol{\theta}$) posteriorijakaumien päivittämiseen. Uuden havaintoaineiston informaatiolla päivitetyt malliparametrit vaikuttavat uusien maahanmuuttajien elinkaariennusteiden jakaumiin ja näiden jakaumien odotusarvoihin. Tällä tavoin uusi havaintoaineisto voi myös muuttaa julkisen talouden hyödyn maksimoivan päätöksen uusien maahanmuuttajien valinnan suhteen.

Ennustemallien muuttaminen uuden havaintoaineiston informaation perusteella. Uuden havaintoaineiston pohjalta voidaan myös laajentaa tai muuttaa käytössä olevaan elinkaarivaikutusten ennustemallia. Esimerkiksi jos tulevaisuudessa maahanmuuttajia koskien on saatavilla tarkkoja koulutustietoja siltä vuodelta, kun he muuttavat ensimmäisen kerran maahan, tämä tieto kannattaa sisällyttää elinkaariennusteisiin. Uutta havaintoaineistoa voidaan käyttää myös posterioriennustetarkistuksiin, eli sen tarkistamiseen, että elinkaariennusteiden komponenttien ennustemallit tuottavat havaintoaineistoa vastaavia ennusteita (vrt. posterioriennustetarkistukset esimerkiksi palkkasummamallinnuksen yhteydessä). Mikäli ennustemallit eivät enää tuottaisi uutta havaintoaineistoa vastaavia ennusteita, niin ennustemalleja olisi syytä muuttaa niin, että vastaavuus paranisi.

Hyödyn maksimointi uusien maahanmuuttajien valinnoista yli ajan. Päätöksiä potentiaalisten uusien maahanmuuttajien Suomeen muuttamisesta ei tehdä myöskään vain yhtenä ajanhetkenä tai vain yhden ajanjakson yksikön aikana eli tässä yhteydessä yhden kalenterivuoden aikana. Päätöksiä potentiaalisten uusien maahanmuuttajien muuttamisesta Suomeen tehdään käytännössä koko ajan myös tulevaisuudessa. Koska päätöksentekoteorian tavoitteena on maksimoida päätösten hyöty päätöksentekijälle, tässä yhteydessä hyödyn maksimointi maahanmuuttajien valinnoista pitää tehdä yli tietyn ajan tai ajanjakson. Tämä tarkoittaa käytännössä valintapäätösten tekemistä useiden kalenterivuosien aikana tulevaisuudessa.

Päätöksiä seuraavien valintojen tulosten todennäköisyysjakaumat ovat uusien maahanmuuttajien elinkaarivaikutusten posterioriennustejakaumia (ks. alaluku 12.2.3). Näiden ennustejakaumien parametrien päivitys uudella tiedolla sekä valintojen hyödyn maksimointi yli ajan muistuttavat läheisesti aiemmin tässä

luvussa lyhyesti esitettyä sekventiaalista päätöksentekoanalyysia (ks. alaluku 12.1.3). Sekventiaalisessa analyysissa päätöksentekijä voi päättää lykätä varsinaisen päätöksen tekemistä ja näin kartuttaa informaatiota koskien varsinaisen päätöksen tulosta ajan yli siten, että päätöksentekijän hyöty maksimoituu koko tämän ajassa jatkuvan prosessin yli.

Potentiaalisten uusien maahanmuuttajien valinta yli ajan muistuttaa sekventiaalista analyysia ensiksi siinä mielessä, että ajan yli karttuvalla havaintoaineistolla voidaan päivittää valinnan tuloksen todennäköisyysjakauma. Toiseksi ajan yli tapahtuvat valinnat muistuttavat sekventiaalista analyysia siinä, että tavoitteena on maksimoida päätöksentekijän hyöty yli ajan. Erona sekventiaaliseen analyysin kuitenkin on, että päätöstä uuden maahanmuuttajan valinnasta Suomeen muuttavaksi (tai ei muuttavaksi) ei voi käytännössä lykätä esimerkiksi vuosikausiksi eteenpäin: potentiaalisia uusia maahanmuuttajia valitaan (tai ei valita) Suomeen muuttavaksi koko ajan vallitsevan maahanmuuttoa koskevan lainsäädännön ja sen toimeenpanon puitteissa.

Esitän seuraavaksi ensiksi, kuinka uutta havaintoaineistoa voidaan käyttää kahdella tavalla elinkaariennustemallien muokkaamiseen. Ensimmäinen tapa on, että uusien maahanmuuttajien elinkaariennusteiden komponenttien ($\boldsymbol{y}_{uusi}$) malliparametrien ($\boldsymbol{\theta}$) posteriorijakaumat päivitetään uudella havaintoaineistolla. Toinen tapa on, että elinkaariennustemalleja laajennetaan uusilla parametreilla tai käytössä olevia todennäköisyysjakaumien luokkia vaihdetaan. Kun olen esittänyt, kuinka uutta havaintoaineisto voidaan käyttää elinkaariennustemallien muokkaamiseen, esitän määritelmän sille, kuinka uudet maahanmuuttajat voidaan valita yli ajan siten, että julkisen talouden odotettavissa oleva hyöty maksimoituu potentiaalisten uusien maahanmuuttajien valinnoista yli ajan.

12.4.1 Malliparametrien vektorin posteriorijakauman päivitys uudella havaintoaineistolla

Olkoon elinkaarivaikutuksen komponenttien mallien kaikkien parametrien vektori $\boldsymbol{\theta}$, kuten edellä (alaluvussa 12.2.3) on esitetty. Olkoon ”vanhat havainnot” se havaintoaineisto, jota on alun perin käytetty $\boldsymbol{\theta}$-vektorin estimointiin. Parametrien vektorin $\boldsymbol{\theta}$ posteriorijakauma tällä vanhojen havaintojen aineistolla on:

$$p(\boldsymbol{\theta}|\text{vanhat havainnot}) \propto p(\text{vanhat havainnot}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

$p(\boldsymbol{\theta})$ on tässä yhteydessä parametrivektorin priorijakauma, joka summaa havaintoaineistojen ulkopuolisen informaation koskien parametrivektoria. Olkoon nyt ”uudet havainnot” se havaintoaineisto, joka on saatu kartutettua myöhemmin (kuin vanhat havainnot) ja jota käytetään parametrivektorin posteriorijakauman päivittämiseen. Kun oletetaan lisäksi, että uudet havainnot ovat ehdollisesti (ehtona $\boldsymbol{\theta}$) riippumattomia vanhoista havainnoista, niin parametrivektorin $\boldsymbol{\theta}$ uusilla havainnoilla päivitetty posteriorijakauma on:

$$p(\boldsymbol{\theta}|\text{uudet havainnot, vanhat havainnot}) \propto p(\text{uudet havainnot}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\text{vanhat havainnot}).$$

Priorijakaumana tässä on yllä esitetty vanhoihin havaintoihin pohjautuva $\boldsymbol{\theta}$:n posteriorijakauma. Uusi $\boldsymbol{\theta}$-vektorin posteriorijakauma on nyt ehdollinen sekä uusilla että vanhoilla havainnoilla. Parametrivektorin $\boldsymbol{\theta}$ ”iterointia” voidaan jatkaa vastaavalla tavalla aina, kun sitä koskien saadaan uutta havaintoaineistoa.

Tällöin sen uusimmalla aineistolla päivitetty posteriorijakauma saadaan esitetyllä tavalla, eli vanhoihin havaintoihin perustuva posteriorijakauma toimii aina parametrivektoria $\boldsymbol{\theta}$ koskevana priorijakaumana. Tämä on intuitiivisesti helppo ymmärtää, sillä priorijakauma summaa sen tiedon (todennäköisyysjakauman muodossa), joka on saatavilla parametrivektorista ennen (lisä)havaintoja. Näin parametrivektoria $\boldsymbol{\theta}$ voidaan päivittää esimerkiksi joka vuosi uusimmalla maahanmuuttajia koskevalla havaintoaineistolla.

Samaa satunnaismuuttujan tai -vektorin posteriorijakauman päivittämistä uudella havaintoaineistolla voidaan soveltaa myös muihin malliparametreihin $\boldsymbol{\pi}$.

12.4.2 Ennustemallien laajennus uusilla parametreilla ja ennustemalleissa käytettävien todennäköisyysjakaumien luokkien vaihtaminen

Olen käyttänyt tässä tutkimuksessa posterioriennustetarkistuksia sen varmistamiseen, että elinkaariennusteiden komponenttien ennustemallit tuottavat luotettavia ennusteita. Tämä tarkoittaa sitä, että ennustemallien tuottamisen ennusteiden avulla voidaan toistaa käytettävissä oleva havaintoaineisto. Kun tulevaisuudessa maahanmuuttajia (ja Suomessa syntyneitä koskien) karttuu uutta havaintoaineistoa, ennustemallien tuottamia ennusteita on syytä verrata myös uuteen havaintoaineistoon, jotta voidaan varmistua, että ennustemallit tuottavat edelleen luotettavia ennusteita.

Tulevaisuudessa ennustemalleja voidaan muokata seuraavasti:

1. Ennustemalleihin lisätään parametreja.
2. Ennustemalleissa käytettävien todennäköisyysjakaumien luokkia vaihdetaan.

Ennustemalleihin lisätään parametreja varsinkin silloin, jos uusia maahanmuuttajia koskien saadaan käyttöön uusia rekisteritietoja. Nämä voivat olla esimerkiksi tarkkoja tietoja heidän koulutustaustastaan tai maahanmuuton syystään. Näiden tietojen lisääminen ennustemalleihin lisäisi käytännössä malleissa olevien selittävien muuttujien ja niitä vastaavien regressiokertoimien määrää.

Ennustemalleihin voidaan lisätä parametreja myös posterioriennustetarkistusten pohjalta. Tällöin ennustemallien halutaan tuottavan paremmin havaintoaineistoa vastaavia ennusteita, kun malliin lisätään parametreja. Yksi esimerkki tästä on se, kun työikäisten nettovaikutusten mallinrakennuksessa normaalijakautuneeseen otosjakaumaan lisättiin vapausasteparametri. Tällöin tästä uudesta otosjakaumasta tuli Studentin t-jakauma, joka tuottaa lähemmin havaintoaineistoa vastaavia ennusteita kuin normaalijakauma.

Mikäli posterioriennustetarkistusten perusteella pelkkä parametrien lisäys ennustemalleihin ei tuota tarpeeksi hyvin uusia havaintoaineistoja vastaavia ennusteita, tällöin ennustemalleissa on kenties vaihdettava kokonaan käytössä olevien todennäköisyysjakaumien luokkia. Esimerkiksi palkkasummien ennustamisen osalta tämä voisi tarkoittaa sitä, että jos käytössä oleva hurdle-gamma-sekoitusjakauma-malli ei enää tulevaisuudessa tuota uusia havaintoaineistoja vastaavia ennusteita – eikä pelkkien uusien parametrien lisäys malliin korjaa tilannetta – kenties gamma-jakaumien sekoitusjakauma on vaihdettava esimerkiksi jonkin toisen todennäköisyysjakaumien luokan sekoitusjakaumaan.

12.4.3 Julkisen talouden ennustetun hyödyn odotusarvo yli ajan

Olkoon nyt $A = \{0, 1, 2, \ldots, N{-}1, N | N < \infty\}$ niiden (kalenteri)vuosien joukko, joiden yli julkisen talouden odotettavissa oleva hyöty halutaan maksimoida potentiaalisten uusien maahanmuuttajien valinnoista. Tässä vuosi 0 voi olla vaikka nykyinen kalenterivuosi, josta kalenterivuosien joukon A viimeinen vuosi on N:n vuoden päässä.

Olkoon lisäksi mahdollisten valintojen matriisi $\boldsymbol{V}$ yli ajan:

$$\boldsymbol{V}_{N\times 2} = \begin{bmatrix} V_0 \\ V_1 \\ \vdots \\ V_{N-1} \\ V_N \end{bmatrix}, \tag{12.20}$$

missä kalenterivuoden $a \in \{0, 1, 2, \ldots, N{-}1, N\}$ kaksi mahdollsta valintaa potentiaalisen uuden maahanmuuttajan suhteen ovat:

$V_a = \{$"Potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen vuonna a",
"Potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen vuonna a"$\}$.

Tällöin julkisen talouden hyödyn odotusarvo mahdollisilla valinnoilla $\boldsymbol{V}$ voidaan ilmaista matriisina:

$$E[\boldsymbol{V}, f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})|\boldsymbol{h}] = \begin{bmatrix} E[f(\boldsymbol{\pi}(0), \boldsymbol{y}_{uusi})|\boldsymbol{h}_0] & 0 \\ E[f(\boldsymbol{\pi}(1), \boldsymbol{y}_{uusi})|\boldsymbol{h}_1] & 0 \\ E[f(\boldsymbol{\pi}(2), \boldsymbol{y}_{uusi})|\boldsymbol{h}_2] & 0 \\ \vdots & \vdots \\ E[f(\boldsymbol{\pi}(N{-}1), \boldsymbol{y}_{uusi})|\boldsymbol{h}_{N-1}] & 0 \\ E[f(\boldsymbol{\pi}(N), \boldsymbol{y}_{uusi})|\boldsymbol{h}_N] & 0 \end{bmatrix}, \tag{12.21}$$

missä matriisin rivit vastaavat yksittäisiä (kalenteri)vuosia $a \in \{0, 1, 2, \ldots, N{-}1, N\}$ ja sarakkeet puolestaan (kalenteri)vuosikohtaisia päätöksiä. Oikeanpuoleinen sarake on N-pituinen nollavektori. Sen solut kertovat julkisen talouden euromääräisen hyödyn potentiaalisesta uudesta maahanmuuttajasta, mikäli tätä ei valita muuttavaksi Suomeen. Kuten aiemmin on esitetty, tällöin julkisen talouden nettovaikutus on nolla (euroa). Matriisin vasemmanpuoleinen sarakkeen solut puolestaan ilmaisevat julkisen talouden hyödyn odotusarvon, kun potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen. Kuten olen aiemmin esittänyt, odotusarvot ovat uuden maahanmuuttajan elinkaariennusteen todennäköisyysjakauman odotusarvoja (ks. alaluvut 12.2.3 (elinkaariennusteista) ja 12.2.4 (odotusarvoista ja odotetun hyödyn maksimoinnista)).

Jokaista vuotta $a \in A$ koskeva ehdollinen elinkaariennuste on päivitetty aina uusimmalla, yli ajan karttuneella kumulatiivisella havaintoaineistolla $\boldsymbol{h}_a$. Uusimman havaintoaineiston joukkoon $\boldsymbol{h}_a$ sisältyy siten sen osajoukkoina ennen vuotta a, eli vuosina $0, 1, 2, \ldots, a{-}1$, karttuneet havaintoaineistot:

$\boldsymbol{h}_0 \subset \boldsymbol{h}_1 \subset \boldsymbol{h}_2 \subset \cdots \subset \boldsymbol{h}_{a-1} \subset \boldsymbol{h}_a$. Kuten edellä, myös tässä yhteydessä havaintoaineisto koostuu ”vanhoja” (Suomessa jo asuneita) sekä ”uusia” (Suomeen ensimmäisen kerran muuttavia) maahanmuuttajia koskevasta havaintoaineistosta.

Muiden ennusteparametrien vektorin todennäköisyysjakauma vuonna a on $\boldsymbol{\pi}(a)$. Tällöin uuden maahanmuuttajan elinkaariennusteen odotusarvo on vuonna a: $E[f(\boldsymbol{\pi}(a), \boldsymbol{y}_{uusi})|\boldsymbol{h}_a]$. Samalla tämä on julkisen talouden hyödyn odotusarvo vuonna a muuttavasta uudesta maahanmuuttajasta. Kuten edellä, myös tässä yhteydessä oletan odotusarvon olevan olemassa uuden maahanmuuttajan elinkaariennusteen jakaumalle. Tässä tutkimuksessa huolehdin käytännössä elinkaariennusteiden odotusarvojen olemassaolosta sillä, että valitsin sopivat todennäköisyysjakaumat elinkaariennusteiden komponenttien (mm. olotilat, työikäisten nettovaikutukset) mallinnuksen yhteydessä.

12.4.4 Julkisen talouden ennustetun hyödyn odotusarvon maksimointi yli ajan

Kehitetään vielä yli ajan tapahtuvan potentiaalisten uusien maahanmuuttajien valinnan määritelmää. Määritetään kalenterivuosina $A = \{0, 1, 2, \ldots, N-1, N | N < \infty\}$ tapahtuvien valintojen *diskreetiksi N-ulotteiseksi avaruudeksi* $\mathcal{V} = V^N$, missä N on kalenterivuosien määrä ja V on joukko (kaksio), joka muodostuu kahden mahdollisen valinnan alkiosta, kuten edellä on esitetty. Nämä valinnat olivat: ”Potentiaalinen uusi maahanmuuttaja valitaan muuttavaksi Suomeen” sekä ”Potentiaalista uutta maahanmuuttajaa ei valita muuttavaksi Suomeen”. Diskreetin avaruuden $\mathcal{V}$ pisteet muodostuvat toisin sanoen vuosittaisista potentiaalisia uusia maahanmuuttajia koskevista valinnoista.

Olkoon nyt $(V_a)_{a \in A}$ kalenterivuosittaisten valintojen *jono* valintojen avaruudessa $\mathcal{V}$. Jonon (V_a) jäsen V_a on potentiaalista uutta maahanmuuttajaa koskeva valinta kalenterivuonna $a \in \{0, 1, 2, \ldots, N-1, N\}$, kuten edellisessä alaluvussa esitettiin.

Jotta päätöksentekijän odotettavissa olevan hyöty ajan yli maksimoituu potentiaalisten uusien maahanmuuttajien valinnoista, on löydettävä sellainen valintojen jono $(V_a)_{a \in A}$, joka maksimoi euromääräisen odotusarvon (EMV):

$$\max_{(V_a)} E[(V_a), f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})|\boldsymbol{h}]. \tag{12.22}$$

Koska summan odotusarvo on summan yhteenlaskettavien odotusarvojen summa ($E(\sum_{i=0}^{N} X_i) = \sum_{i=0}^{N} E(X_i)$), julkisen talouden hyödyn euromääräinen odotusarvo maksimoituu nyt uusien maahanmuuttajien valinnasta yli kalenterivuosien $0, 1, 2, \ldots, N-1, N$ seuraavasti:

$$\max_{(V_a)} E[(V_a), f(\boldsymbol{\pi}, \boldsymbol{y}_{uusi})|\boldsymbol{h}] = \sum_{a=0}^{N} \max\left(E[f(\boldsymbol{\pi}(a), \boldsymbol{y}_{uusi})|\boldsymbol{h}_a], 0\right). \tag{12.23}$$

Julkisen talouden odotusarvoinen hyöty maksimoituu sellaisella uusia maahanmuuttajia koskevalla valintojen jonolla (V_a), joka maksimoi esitetyn summan euromääräisen arvon. Elinkaariennusteiden odotusarvot, $E[f(\boldsymbol{\pi}(a), \boldsymbol{y}_{uusi})|\boldsymbol{h}_a]$, ovat summassa nettonykyarvoisia.

Summa maksimoituu, kun jokaisena (kalenteri)vuotena a potentiaalinen uusi maahanmuuttaja valitaan Suomeen, mikäli tämän elinkaariennusteen odotusarvo on kalenterivuoden a tiedoilla nollaa suurempi. Mikäli taas odotusarvo on nollaa pienempi, potentiaalista uutta maahanmuuttajaa ei valita Suomeen muuttavaksi.

Maksimointisääntö on siis samankaltainen kuin edellä yhdelle vuodelle esitetty. Nyt vain valinta perustuu uuden maahanmuuttajan *päivittyvän* elinkaariennusteen jakauman odotusarvoon, missä ennustejakauman parametrit on päivitetty jokaisena uutena kalenterivuonna uudella havaintoaineistolla. Jos $N = 0$, saadaan erikoistapauksena aiemmin tässä luvussa esitetty tulos julkisen talouden odotusarvoisen hyödyn maksimoinnista, kun päätöksiä tehdään vain yhdellä ajanhetkellä tai -ajanjaksolla (tässä yhtenä kalenterivuonna).

Koska uuden maahanmuuttajan elinkaariennusteen jakauma päivitetään joka vuosi uudella havaintoaineistolla, odotetun hyödyn maksimoivat päätökset voivat vaihdella vuosittain valittaessa suunnilleen samankaltaisia potentiaalisia uusia maahanmuuttajia. Esimerkiksi jos tällaista potentiaalista uutta maahanmuuttajaa koskevan elinkaariennusteen jakauman odotusarvo on ollut aiempina vuosina negatiivinen (päätös: ei valita muuttavaksi), mutta uuden havaintoaineiston tuoman informaation myötä elinkaariennusteen odotusarvo onkin positiivinen, tällöin julkisen talouden hyödyn maksimoiva päätös muuttuu ei-valinnasta samankaltaisen maahanmuuttajan valinnaksi.

12.5 Tiivistelmä

Esitän alla tiivistetyssä muodossa tämän luvun keskeisimmät asiat.

12.5.1 Yleistä päätöksentekoteoriasta

- Päätöksentekoteorian tai päätöksentekoanalyysin tavoitteena on maksimoida päätöksentekijän hyödyn (utiliteetin) odotusarvo päätöksentekotilanteessa. Päätöksentekijän kannattaa valita se vaihtoehto mahdollisten päätösten joukosta, jonka hyödyn odotusarvo on kaikkein suurin. Päätösten tuottamat tulokset sijoittuvat tulevaisuuteen, joten ne sisältävät epävarmuuksia, joita kuvataan todennäköisyysjakaumien avulla.

Päätöksentekoteorian soveltamisen edellytykset

- Jotta päätöksentekoteoriaa voidaan soveltaa valintatilanteessa, valintatilannetta koskien on pystyttävä määrittämään seuraavat kolme asiaa:
  - Päätöksentekijän kaikki valintavaihdot päätöksentekotilanteessa.
  - Valintojen (tulevaisuudessa) tuottamille tuloksille on määritettävä todennäköisyysjakaumat. Bayesilaisessa viitekehyksessä jakaumat ovat tulosten posteriorijakaumia.
  - Päätöksentekijän hyötyfunktio (utiliteettifunktio). Hyötyfunktion argumenttina ovat päätösten tuottamat tulokset. On kyettävä määrittämään, mikä on päätöksentekijän hyöty jokaisesta mahdollisesta päätöksen tuloksesta.
- Lisäksi hyödyn odotusarvon on oltava olemassa. Tämä riippuu hyödyn todennäköisyysjakauman jakaumaoletuksesta. Esimerkiksi Cauchy-jakaumalla ei ole odotusarvoa.

Euromääräinen hyöty

- Kun päätöksentekijän hyötyfunktio on lineaarinen ja hyödyn määrän yksikkö on 1 euro, hyödyn odotusarvo jokaisen valinnan kohdalla on valinnan rahallinen odotusarvo (*engl. Expected Monetary Value, EMV*). Tässä tapauksessa päätöksentekijän hyödyn odotusarvon maksimoiva valinta päätöksentekotilanteessa on se valinta, jonka euromääräinen odotusarvo on kaikkein suurin.

Lyhyesti sekventiaalisesta analyysista

- Sekventiaalisessa analyysissa varsinaista päätöksentekoa voidaan lykätä tulevaisuuteen. Päätöksen lykkäämisellä on usein kustannus, mutta toisaalta päätöksentekijälle karttuu tällöin enemmän informaatiota eri valintavaihtojen tuottamista tuloksista.

- Tämän tutkimuksen viitekehyksessä sekventiaalisesta analyysista opitaan, että päätöksentekijän odotettavissa oleva hyöty voidaan maksimoida myös yli ajan(jakson) ja siten useiden toisiaan seuraavien valintojen yli. Odotusarvon maksimoinnissa ajan yli hyödynnetään ajan yli karttuvan uuden havaintoaineiston tarjoamaa informaatiota, jolla voidaan päivittää varsinaisen valinnan tuloksen posteriorijakauma.

12.5.2 Päätöksentekoteorian soveltaminen potentiaalisen uuden maahanmuuttajan valintaan

- Tavoitteena on maksimoida julkisen talouden hyödyn odotusarvo potentiaalisen uuden maahanmuuttajan valinnasta.
- Päätöksentekijänä on käytännössä eduskunta ja hallitus, jotka laativat ja saattavat toimeen maahanmuuttoa koskevaa (uutta) lainsäädäntöä. Maahanmuuttoa koskevien lakien laatimisen ja toimeenpanon avulla potentiaalisia uusia maahanmuuttajia voidaan valita joko muuttavaksi Suomeen tai jättää valitsematta. Tämä koskee lähinnä EU:n ulkopuolelta tulevia maahanmuuttajia.
- Valitseminen voi olla joko suoraa tai epäsuoraa. Edellinen koskee suoraan maahanpääsyä ja oleskeluluvan saantia, jälkimmäinen puolestaan Suomeen muuton vetotekijöiden muuttamista.
- Potentiaalinen uusi maahanmuuttaja on sellainen henkilö, joka muuttaa ensimmäisen kerran Suomeen, mikäli hänet valitaan Suomeen muuttavaksi lainsäädännöllisin keinoin. Koska yhdellä henkilöllä voi olla elämänsä aikana useita muuttoja Suomeen, uusi maahanmuuttaja on määritelmän mukaan sellainen, joka on muuttamassa Suomeen ensimmäistä kertaa.

Päätöksentekoteorian soveltamisen edellytykset potentiaalisten uusien maahanmuuttajien valintaan

- Jotta päätöksentekoteoriaa voidaan soveltaa potentiaalisen uuden maahanmuuttajan valintaan, valintatilannetta koskien on määritettävä edellä esitetyt kolme asiaa: kaikki valintavaihtoehdot, valintojen tuottamien tulosten todennäköisyysjakaumat sekä julkisen talouden hyöty valintojen tulosten seurauksena.
  - Valintavaihtoehtoja on kaksi: Potentiaalinen uusi maahanmuuttaja joko valitaan Suomeen muuttavaksi, tai häntä ei valita Suomeen muuttavaksi.
  - Valintojen (tulevaisuudessa) tuottamien tulosten todennäköisyysjakaumat. Mikäli uusi maahanmuuttaja valitaan Suomeen muuttavaksi, tuloksen todennäköisyysjakauma on uudelle maahanmuuttajalle määritetty julkisen talouden elinkaarivaikutuksen posterioriennustejakauma. Mikäli uutta maahanmuuttajaa ei valita Suomeen muuttavaksi, tulos on aina tasan nolla euroa. Kummankin vaihtoehdon kohdalla tuloksista puuttuu tiettyjä julkisen talouden vaikutuksia. Suomeen muuttavien kohdalla puutteet on esitetty luvussa 10.

- Julkisen talouden hyötyfunktio on lineaarinen ja euromääräinen erityisesti silloin, kun uusia maahanmuuttajia halutaan valita Suomeen heidän tuottamiensa julkisen talouden nettovaikutusten perusteella. Hyötyfunktio voi olla erilainen, kun huomioidaan muitakin valinnan tuloksia kuin julkisen talouden nettovaikutukset. Kun hyötyfunktio on lineaarinen ja euromääräinen, julkisen talouden hyödyn odotusarvo on odotettavissa oleva rahallinen arvo (*Expected Monetary Value, EMV*) maahanmuuttajan valinnasta. Tämä on sama asia kuin uuden maahanmuuttajan elinkaarivaikutuksen odotusarvo.

Julkisen talouden hyödyn maksimoiva kriteeri

- Julkisen talouden hyöty mahdollisen uuden maahanmuuttajan valinnassa maksimoituu sillä valinnalla, joka maksimoi tuloksen odotettavissa olevan rahallisen arvon (EMV):
  - Kun potentiaalisen uuden maahanmuuttajan elinkaarivaikutuksen ennustejakauman euromääräinen odotusarvo on suurempi kuin nolla, niin tämä potentiaalinen uusi maahanmuuttaja kannattaa valita muuttavaksi Suomeen.
  - Kun potentiaalisen uuden maahanmuuttajan elinkaarivaikutuksen ennustejakauman euromääräinen odotusarvo on pienempi kuin nolla, niin tätä potentiaalista uutta maahanmuuttajaa ei kannata valita muuttavaksi Suomeen.

Aineiston ulkopuoliten tekijöiden vaikutukset: yleistä

- Niin kutsuttuja *aineiston ulkopuolisten tekijöiden vaikutuksia* elinkaariennusteisiin on käsitelty jo tarkemmin johdantoluvussa. Uusien maahanmuuttajien elinkaariennusteiden jakaumat on estimoitu käytännössä vuosia 1988/1995–2011/2014 koskevasta tutkimusaineistosta. Kuten johdantoluvussa on esitetty, tutkimuskirjallisuuden mukaan elinkaariennusteisiin voivat kuitenkin vaikuttaa erityisesti sellaiset tulevaisuuteen liittyvät tekijät, joita ei saa estimoitua tutkimusaineistosta, sillä ne sijoittuvat vuoden 2011 jälkeiseen tulevaisuuteen. Kutsun näitä vuoden 2011 jälkeiseen tulevaisuuteen liittyviä (epävarmuus)tekijöitä ”aineiston ulkopuolisten tekijöiden vaikutuksiksi”.
- Aineiston ulkopuolisten tekijöiden vaikutukset $\boldsymbol{\epsilon}$ välittyvät elinkaariennusteisiin malliparametrien $\boldsymbol{\theta}$ todennäköisyysjakaumien kautta. Mikäli $\boldsymbol{\epsilon}$-satunnaisvektoria ei huomioitaisi, $\boldsymbol{\theta}$:lla olisi vain tutkimusaineistoon perustuva posteriorijakauma elinkaariennusteissa.
- Jokaisen malliparametrin kohdalla voidaan määrittää, kuinka aineiston ulkopuoliset tekijät vaikuttavat niiden ehdollisiin todennäköisyysjakaumiin. Lisäksi aineiston ulkopuolisten tekijöiden vaikutuksille on määritettävä omat reunajakaumansa.
- Esimerkiksi työikäisten nettovaikutusmallin maassaoloajan regressiokertoimeen β_{maika} voi liittyä kaksi aineiston ulkopuolisten tekijöiden vaikutusta: ϵ_{j2011} ja ϵ_{ep}. Ensimmäinen näistä on se vuoden 2011 jälkeistä tulevaisuutta koskeva vaikutus regressiokertoimeen, mitä ei saada estimoitua havaintoaineistosta. Jälkimmäinen on puolestaan se ekstrapolaatioon liittyvä vaikutus, mitä

ei saada estimoitua havaintoaineistosta. Elinkaariennusteissa maassaoloajan vaikutus ekstrapoloidaan työikäisten nettovaikutuksia koskevissa ennusteissa maassaolovuosille 24–46 vuotta (tutkimusaineisto 23 maassaolovuoteen asti). Jotta aineiston ulkopuolisten tekijöiden vaikutukset saadaan mukaan ennusteisiin, maassaoloajan regressiokertoimelle määritetään ehdollinen jakauma $p(\beta_{maika}|, \epsilon_{j2011}, \epsilon_{ep}, \boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$, ja aineiston ulkopuolisille vaikutuksille määritetään puolestaan reunajakauma $p(\epsilon_{j2011}, \epsilon_{ep})$.

- Maassaoloajan regressiokertoimen aineistosta pelkästä aineistosta estimoitu posteriorijakauma $p(\beta_{maika}|\boldsymbol{y}_{vanha}, \boldsymbol{x}_{vanha})$ on likimain normaalijakautunut. Aineiston ulkopuolisten tekijöiden vaikutuksesta tämä normaalijakauma siirtyy ϵ_{j2011}:n ja ϵ_{ep}:n verran, kuten varsinaisessa tekstissä esitetään. Näillä aineiston ulkopuolisten tekijöiden vaikutuksilla on Studentin t-jakautuneet reunajakaumat: $T(4, \mu_\epsilon, \sigma_\epsilon)$, missä 4 on vapausaste, μ_ϵ odotusarvo ja σ_ϵ hajontaparametri.

- Mikäli elinkaariennusteen laatijalla ei ole sellaista aineiston ulkopuolista informaatiota, joka kertoisi, onko maassaoloajan regressiokerroin aineiston ulkopuolisten tekijöiden vaikutuksesta todennäköisemmin pienempi tai suurempi kuin pelkästä tutkimusaineistosta estimoitu, aineiston ulkopuolisten tekijöiden vaikutusten reunajakaumien odotusarvo $\mu_\epsilon = 0$, jolloin $T(4, 0, \sigma_\epsilon)$. Tässä tapauksessa maassaoloajan regressiokertoimen jakauman odotusarvo ei muutu, eikä siten myöskään työikäisten nettovaikutusennusteiden odotusarvo eikä elinkaariennusteiden odotusarvo.

- Työikäisten nettovaikutusmallin maassaoloajan regressiokerrointa koskenut tarkastelu voidaan toistaa kaikkien muiden malliparametrien kohdalla tietyin edellytyksin. Mikäli elinkaariennusteen laatija ei tiedä, ovatko malliparametrien todennäköisyysjakaumat aineiston ulkopuolisten tekijöiden vaikutuksesta todennäköisemmin positiivisempia vai negatiivisempia kuin tutkimusaineistosta estimoidut jakaumat, elinkaariennusteen jakauman odotusarvo ei muutu, vaikka jakauma levenisi ja sisältäisi siten enemmän epävarmuutta.

- Toisin sanoen, vaikka aineiston ulkopuoliset tekijät yleisesti ottaen lisäävät elinkaariennusteisiin sisältyvää epävarmuutta, tämä ei vielä tarkoita, että elinkaariennusteiden jakaumien odotusarvot muuttuisivat. Ennusteparametrien keskihajonnat voivat kasvaa epävarmuuden lisääntyessä, mutta elinkaariennusteiden odotusarvot voivat pysyä samoina.

- Aineiston ulkopuolisten tekijöiden vaikutusten huomiointi elinkaariennusteissa muuttaa julkisen talouden odotetun hyödyn maksimoivan ratkaisun maahanmuuttajien valinnassa vain, jos uuden maahanmuuttajan elinkaariennusteen jakauman *odotusarvon etumerkki* vaihtuu, kun aineiston ulkopuolisten tekijöiden vaikutukset otetaan huomioon elinkaariennusteissa.

- Jos esimerkiksi uuden maahanmuuttajan elinkaariennusteen jakauman odotusarvo on –500 000 euroa, kun aineiston ulkopuolisten tekijöiden vaikutuksia ei oteta huomioon, julkisen talouden hyödyn maksimoiva valinta muuttuu vain siinä tapauksessa, että jakauman odotusarvo siirtyy vähintään 500 000 euroa positiivisemmaksi, kun aineiston ulkopuolisten tekijöiden vaikutukset huomioidaan. Tämä edellyttää sitä, että elinkaariennusteen laatijalla on sellaista tietoa:

  - aineiston ulkopuolisten tekijöiden vaikutusten todennäköisyysjakaumasta (reunajakaumasta) ja

- siitä, miten aineiston ulkopuoliset tekijät vaikuttavat elinkaariennusteen odotusarvon siirtymiseen vähintään 500 000 euroa positiivisemmaksi.

Aineiston ulkopuolisten tekijöiden vaikutukset: ratkaisu tässä tutkimuksessa

- Käytän tämän tutkimuksen elinkaariennusteissa sellaista ratkaisua aineiston ulkopuolisten tekijöiden vaikutusten suhteen, että aineiston ulkopuolisten tekijöiden vaikutukset voivat lisätä uusien maahanmuuttajien elinkaariennusteisiin sisältyviä epävarmuuksia, mutta ne eivät muuta uusien maahanmuuttajien elinkaariennusteiden odotusarvoja. Tällöin ne eivät myöskään vaikuta uusien maahanmuuttajien julkisen talouden odotetun hyödyn maksimoivaan valintaan.
- Tämä ratkaisu perustuu siihen, että minulla ei ole sellaista tutkimusaineiston ulkopuolista informaatiota, joka kertoisi, ovatko malliparametrien jakaumat aineiston ulkopuolisten tekijöiden vaikutuksesta todennäköisemmin suurempia vai pienempiä kuin pelkästään tutkimusaineiston perusteella estimoidut jakaumat.
- Kun aineiston ulkopuolisten tekijöiden vaikutusten reunajakaumien odotusarvot ovat nollia, ne eivät muuta malliparametrien todennäköisyysjakaumien odotusarvoja. Tällöin ne eivät myöskään muuta elinkaariennusteiden komponenttien jakaumien odotusarvoja, vaikka ne leventäisivätkin näitä jakaumia. Tämä pätee erityisesti eniten elinkaariennusteeseen vaikuttavan komponentin mallin eli työikäisten nettovaikutusmallin kohdalla.
- Tässä tapauksessa myöskään potentiaalisten uusien maahanmuuttajien valinta ei muutu, vaikka aineiston ulkopuolisten tekijöiden vaikutukset ennusteissa huomioitaisiin. Tämä johtuu siitä, että julkisen talouden hyödyn odotusarvon maksimoiva valinta riippuu ainoastaan potentiaalisen uuden maahanmuuttajan elinkaariennusteen jakauman odotusarvosta. Koska odotusarvo ei muutu, hyödyn maksimoiva valinta on sama kuin pelkän tutkimusaineiston informaation pohjalta tehtävä valinta.

Ennustetun hyödyn maksimointi yli ajan

- Julkisen talouden hyödyn maksimointi yli ajan(jakson) tapahtuvista valinnoista muistuttaa sekventiaalista analyysia. Tosin toisin kuin sekventiaalisessa analyysissa, päätöstä mahdollisten uusien maahanmuuttajien valinnasta ei voi käytännössä lykätä tulevaisuuteen.
  - Uuden maahanmuuttajan elinkaariennusteen komponenttien parametrien posteriorijakaumat päivitetään ajassa karttuvan havaintoaineiston tarjoamalla uudella informaatiolla. Päivitys tapahtuu Bayesin kaavaa soveltamalla.
  - Ennustemalleja voidaan muokata lisäämällä niihin parametreja tai vaihtamalla kokonaan ennustemalleissa käytössä olevia todennäköisyysjakaumien luokkia. Mikäli uusia maahanmuuttajia koskien on tulevaisuudessa saatavilla uusia tietoja (esimerkiksi koulutuksesta), niin nämä voidaan lisätä ennustemalleihin uusina selittävinä muuttujina ja niitä vastaavina regressiokertoimina. Mikäli ennustemallit eivät onnistu toistamaan tulevaisuudessa havaintoaineistoja

”tarpeeksi” hyvin posterioriennustetarkistusten perusteella, ennustemalleja on parannettava lisäämällä niihin tarvittavia parametreja, tai jos tämäkään ei paranna ennusteita, vaihtamalla kokonaan ennustemalleissa käytettävien todennäköisyysjakaumien luokkia.

- Julkisen talouden hyöty maksimoituu yli ajan, kun jokaisella valinnan ajanhetkellä noudatetaan seuraavaa valintasääntöä: Valitaan uusi maahanmuuttaja muuttavaksi Suomeen, jos muuttohetkeen – kaikella karttuneella havaintoaineistolla – päivitetyn elinkaariennusteen odotusarvo on suurempi kuin nolla, ja jätetään uusi maahanmuuttaja valitsematta, jos ennustejakauman odotusarvo on nollaa pienempi.
- Kun tällä maksimointiperiaatteella ajan yli tehtyjen valintojen tulosten odotusarvot lasketaan yhteen, tämä summa on julkisen talouden hyödyn maksimoitu odotusarvo maahanmuuttajien valinnasta yli tietyn ajanjakson.
- Koska elinkaariennusteiden jakaumat muuttuvat uuden havaintoaineiston tarjoaman informaation myötä, on mahdollista, että samankaltainen maahanmuuttaja valitaan muuttavaksi Suomeen tiettynä ajanjaksona ja toisena ajanjaksona taas ei.

Luku 13

Johtopäätökset: Irakissa ja Somaliassa syntyneiden potentiaalisten uusien maahanmuuttajien valinta

13.1 Yhteenveto

Käyn yhteenvedossa lyhyesti läpi tämän elinkaariennusteita koskevan tutkimuksen sisällön tähän asti. Myöhemmin tässä luvussa esitän johtopäätökset, joita diskontatuista elinkaariennusteiden tuloksista voidaan tehdä koskien potentiaalisten uusien maahanmuuttajien valintaa päätöksentekoteoriaa soveltamalla.

13.1.1 Elinkaariennusteet ovat todennäköisyysjakaumia (I osa)

Johdannossa määritettiin tutkimuksen motivaatio, keskeisimmät määritelmät, tutkimuskysymykset ja esitettiin tutkimuksessa käytettävä (rekisteri)aineisto. Lisäksi johdannossa määritettiin elinkaariennusteille kaksi käyttötarkoitusta.

Elinkaariennusteiden kaksi käyttötarkoitusta. Ensiksi niiden avulla julkisen talouden nettovaikutukset voidaan esittää niin, että maahanmuuttajien maassaoloaika, sukupuoli, ikä, maahanmuuttovuosi ja muut tekijät otetaan samaan aikaan huomioon. Jotta diskonttaus ei vaikututa elinkaariennusteiden tulosten tulkintaan mainittua tulosten käyttötarkoitusta varten, elinkaariennusteiden tulokset on esitetty diskonttaamattomina luvussa 9.

Toinen elinkaariennusteiden käyttötarkoitus on soveltaa elinkaarivaikutusten ennustetuloksia uusien maahanmuuttajien optimaaliseen valintaan julkisen talouden kannalta. Optimaalisella valinnalla tarkoitetaan tässä yhteydessä uusien maahanmuuttajien valitsemista Suomeen muuttavaksi niin, että julkisen talouden hyödyn odotusarvo potentiaalisen uuden maahanmuuttajan valinnasta maksimoituu. Tässä maksimoin-

nissa käytetään nykyarvoon diskontattujen elinkaariennusteita. Jotta elinkaariennusteita voidaan käyttää uusien maahanmuuttajien valintaan, asetin johdannossa tavoitteeksi tehdä elinkaariennusteista ”uskottavia”.

Elinkaarivaikutusten funktio. Elinkaariennusteiden käyttötarkoitusten jälkeen johdannossa määritettiin ne komponentit (7 kappaletta), joista elinkaarivaikutukset saadaan laskettua elinkaarivaikutusten funktion avulla. Näitä komponentteja olivat mm. suomessaoloaika ja palkkasummat 18–62-vuotiaana vuosittain. Näiden komponenttien ennusteista saadaan muodostettua elinkaariennusteet. Näiden elinkaariennusteiden komponenttien ennusteisiin sisältyy epävarmuuksia.

Elinkaarivaikutusten ennusteet ovat todennäköisyysjakaumia. Epävarmuudet huomioidaan ennusteissa siten, että komponenttien ennusteet ovat todennäköisyysjakaumia, jotka kuvaavat ennusteisiin sisältyviä epävarmuuksia. Johdannossa on eritelty tarkemmin, mistä tekijöistä komponenttien ennustejakaumiin sisältyvät epävarmuudet muodostuvat. Ennusteisiin sisältyvät epävarmuudet perustuvat suurimmaksi osaksi käytössä olevaan havaintoaineistoon. Esitän kuitenkin myös, kuinka informaatio koskien havaintoaineiston ulkopuolisten tekijöiden vaikutuksia voidaan sisällyttää elinkaariennusteisiin.

13.1.2 Elinkaariennusteiden komponenttien uskottava mallinnus ja ennustaminen uusille maahanmuuttajille (II osa)

Tämän tutkimustekstin II osassa (luvuissa 3–9) esitettiin, kuinka uusien maahanmuuttajien elinkaariennusteet mudostetaan. Tämä tapahtui esittämällä, kuinka elinkaariennusteiden komponentit mallinnetaan tilastollisesti siten, että malleja voidaan käyttää uusien maahanmuuttajien elinkaariennusteissa. Komponenttien mallien tulokset olivat todennäköisyysjakaumia. Tavoitteena jokaisen komponentin mallinnuksen kohdalla oli, että elinkaarivaikutusten komponenttien mallit olisivat ”uskottavia” malleja käytettäviksi uusien maahanmuuttajien elinkaariennusteissa. Elinkaarivaikutusten komponenttien mallien uskottava käyttö elinkaariennusteissa perustuu seuraaviin – toteutuneisiin – mallinnustavoitteisiin:

1. Mallien pohjalta tehdyt ennusteet toistavat alkuperäisen havaintoaineiston.
2. Malleissa huomioidaan otanta-asetelman havaintoaineistoihin (paneeliaineistoja) tuottamat riippuvuudet. Paneeliaineistot mahdollistavat myös nk. havaitsemattoman heterogeenisuuden huomioon ottamisen ”uusia” tilastoyksiköitä (henkilöt, maassaolovuodet, maahanmuuttovuodet ja kalenterivuodet) koskevissa ennusteissa.
3. Malleissa huomioidaan erityisesti tärkeimmät aikaan liittyvät vaikutukset. Aikaan liittyviä vaikutuksia ovat ikään, maahanmuuttoikään, maassaoloaikaan, maassaolovuoteen ja kalenterivuoteen liittyvät vaikutukset.

1. uskottavuustekijä elinkaariennusteissa: Mallien pohjalta tehdyt ennusteet toistavat alkuperäisen havaintoaineiston

Ensimmäinen lueteltu elinkaarivaikutusten komponenttien ennustemallien ominaisuus tarkoittaa sitä, että komponenttimallien pohjalta simuloidut komponenttien ennusteiden jakaumat vastaavat/toistavat komponenttien alkuperäisen alkuperäisten havaintoaineistojen jakaumat. Esimerkiksi palkkasummien mallinnuksen kohdalla tämä tarkoitti sitä, että elinkaariennusteissa käytetyt palkkasummamallit tuottivat/-toistivat mallinnuksessa käytetyn palkkasummien havaintoaineiston jakauman, kun ”ennusteet” laadittiin mallinnuksessa käytettävälle selittävien muuttujien aineistolle.

Varmistin komponenttien mallinnusten yhteydessä nk. posterioriennustetarkistuksilla, että komponenttien ennustemallit tuottavat/toistavat todella komponenttien havaintoaineistojen jakaumat. Tämä tavoite toteutui erityisesti ”tärkeimpien” jakaumatunnuslukujen, kuten jakaumien keskiarvojen, keskihajontojen sekä kvartiilien, osalta ”hyvin tarkasti”. Tarkkuudella tarkoita tässä sitä, että ennustemallien tuottamien (posteriori)ennustejakaumien mainitut tärkeimmät tunnusluvut poikkesivat esimerkiksi palkkasummien mallinnuksen yhteydessä ”suurella” todennäköisyydellä korkeintaan joko 200, 500 tai 1 000 euroa havaintoaineiston jakauman vastaavista tunnusluvuista.

2. uskottavuustekijä elinkaariennusteissa: Malleissa huomioidaan otanta-asetelman havaintoaineistoihin (paneeliaineistoja) tuottamat riippuvuudet

Toinen lueteltu elinkaarivaikutusten komponenttien ennustemallien ominaisuus tarkoittaa siitä, että komponenttien mallinnuksessa otettiin huomioon komponenttien havaintoaineistoissa olevat otanta-asetelmasta johtuvat tilastolliset riippuvuudet havaintojen välillä. Nämä riippuvuudet aiheutuvat siitä, että yksittäiset havainnot ovat ryvästyneet havaintoaineistossa henkilöittäin, maahanmuuttovuosittain, maassaoloaikojen ja kalenterivuosien mukaan. Huomioin nämä kaikki riippuvuudet erityisesti nuorten ja työikäisten nettovaikutusten mallinnuksessa käyttämällä niissä tarkoitukseen sopivia hierarkkisia malleja.

Elinkaarivaikutusten komponenttien havaintoaineistossa on tilastoyksiköittäin (henkilöt, maassaolovuodet, maahanmuuttovuoden ja kalenterivuodet) vaihtelevaa nk. havaitsematonta heterogeenisuutta. Havaitsematon heterogeenisuus aiheutuu siitä, että komponentteja koskien ei ole havaintoaineistossa saatavilla kaikkea informaatiota, joka selittää komponenttien arvon vaihtelua *tilastoyksiköittäin*. Esimerkiksi nuorten ja työikäisten nettovaikutusten osalta nettovaikutukset voivat vaihdella henkilöittäin esimerkiksi tarkan (maahanmuuttovuoden) koulutustiedon tai maahanmuuton syyn perusteella. Tietoa näistä tekijöistä ei ole kuitenkaan havaintoaineistossa ja siten nuorten ja työikäisten nettovaikutusten mallissa selittävinä muuttujina, minkä johdosta havaintoaineistosta puuttuva tarkka koulutustieto ja tieto maahanmuuton syystä voivat aiheuttaa havaitsematon heterogeenisuutta henkilöiden välille.

Koska nuorten ja työikäisten nettovaikutusaineisto sisältää useita havaintoja jokaiselta tilastoyksiköltä, aineisto mahdollistaa havaitsemattoman heterogeenisuuden mallintamisen satunnaisvaikutuksien kautta. Nämä satunnaisvaikutukset ovat mukana mainituissa hierarkkisissa malleissa. Koska elinkaariennusteissa käytettävissä malleissa – erityisesti nuorten ja työikäisten nettovaikutusten osalta – on mukana satunnaisvaikutuksina tilastoyksiköittäin vaihteleva havaitsematon heterogeenisuus, tämä epävarmuuden lähde sisältyy mukaan myös elinkaariennusteisin, missä poimitaan uusia ”otoksia” tilastoyksiköistä. Esimerkiksi, koska elinkaariennusteet laaditaan uusille maahanmuuttajille, nämä henkilöt eivät sisälly mukaan havaintoaineistoon, jonka pohjalta elinkaariennusteiden mallit on laadittu. Nyt kuitenkin se vaihtelu, joka aiheutuu havaitsemattomasta heterogeenisuudesta henkilöiden välillä, sisältyy mukaan uusien maahan-

muuttajien elinkaariennusteisiin.

3. uskottavuustekijä elinkaariennusteissa: Malleissa huomioidaan erityisesti tärkeimmät aikaan liittyvät vaikutukset

Kolmas lueteltu elinkaarivaikutusten komponenttien ennustemallien ominaisuus tarkoittaa siitä, että komponenttien mallinnuksessa otettiin huomioon tärkeimmät aikaan liittyvät vaikutukset. Aikaan liittyviä vaikutuksia ovat ikään (ja maahanmuuttoikään), maassaoloaikaan, maassaolovuoteen ja kalenterivuoteen liittyvät vaikutukset. Käytännössä tämä tarkoittaa sitä, että aikaan liittyvät vaikutukset ovat komponenttien ennustemalleissa sekä selittävinä muuttujina (kun aikaan liittyvät vaikutukset ovat havaittuja) sekä aikaan liittyvinä satunnaisvaikutuksina (kun vaikutukset ovat havaitsemattomia). Jälkimmäisessä tapauksessa satunnaisvaikutukset kuvaavat havaitsematonta heterogeenisuutta aikaan liittyvien tilastoyksiköiden välillä. Esimerkiksi eri maahanmuuttovuosien välillä voi olla havaitsematonta heterogeenisuutta, jos maahanmuuttovuosiin liittyy sellaisia havaitsemattomia tekijöitä, jotka vaikuttavat elinkaarivaikutusten komponenttien arvojen vaihteluun.

Elinkaariennusteiden komponenttien mallien posterioriennustetarkistuksissa varmistin, että elinkaariennusteissa käytettävien komponenttien mallit toistavat alkuperäisen havaintoaineiston – aikaan liittyvien muuttujien mukaan. Tämä tarkoittaa sitä, että esimerkiksi palkkasummien malli toistaa palkkasummien havaintoaineiston, kun vertailu mallin tuottamien ennusteiden ja havaintoaineiston välillä tehdään aikaan liittyvien vaikutusten eli iän, maassaolovuosien ja maahanmuuttovuosien mukaan.

13.1.3 Elinkaariennusteet ja päätöksentekoteoria: Potentiaalisten uusien maahanmuuttajien valinta (III osa)

13.1.3.1 Puuttuvat julkisen talouden nettovaikutukset

Elinkaariennusteissa – ja erityisesti elinkaariennusteista tehtävissä johtopäätöksissä – olisi tavoiteltavaa että julkisen talouden nettovaikutusten elinkaariennusteisiin sisältyisivät ainakin kaikki henkilöihin suoraan liitettävät julkisen talouden tulot ja menot. Koska elinkaariennusteet perustuvat elinkaarivaikutusten komponenttien havaintoaineistoihin, käytännössä tavoite kaikkien yksilöllisten nettovaikutusten sisältymisestä elinkaariennusteisiin tarkoittaa sitä, että nämä kaikki vaikutukset sisältyvät elinkaariennusteiden komponenttien havaintoaineistoihin. Kaikki henkilöihin suoraan liitettävä julkisen talouden nettovaikutukset eivät kuitenkaan sisälly käytössä olevaan havaintoaineistoon, ja sen vuoksi esitin luvussa 10, mitä suoria vaikutuksia havaintoaineistoista – ja siten elinkaariennusteista – puuttuu. Esitin myös lyhyesti, miten elinkaariennusteiden tulokset muuttuisivat, jos nämä puuttuvat vaikutukset sisältyisivät ennusteisiin. Erityisesti kerroin, miten Irakissa ja Somaliassa syntyneiden uusien maahanmuuttajien elinkaariennusteet muuttuisivat *suhteessa* Suomessa syntyvien kantaväestön henkilöiden elinkaariennusteisiin, mikäli nämä puuttuvat suorat vaikutukset sisältyisivät elinkaariennusteisiin.

Henkilöihin suoraan liitettävissä vaikutuksissa on elinkaariennusteissa – kuten esitin luvussa 10 – joitakin puutteita erityisesti julkisen talouden menojen osalta. Henkilöiden maksamat verot ja veronluonteiset mak-

sut sisältyvät siis elinkaariennusteisiin lähes kokonaisuudessaan. Elinkaariennusteista puuttuu julkisen talouden menopuolelta erityisesti osa maahanmuuttajien erityispalveluiden menoista sekä vanhuusiässä henkilöiden itse maksama osuus julkisista vanhuusiän palveluista.

Maahanmuuttajien erityispalvelut. Maahanmuuttajien erityispalvelut liittyvät eritoten vastaanotto- ja kotoutumispalveluihin. Ne puuttuvat viranomaisrekistereihin perustuvasta tutkimusaineistosta, joten vastaanotto- ja kotoutumispalveluiden menoista puuttuu ainakin osa elinkaariennusteissa. Jos nämä puuttuvat menot sisältyisivät elinkaariennusteisiin, Suomessa syntyvien kantaväestön henkilöiden ennusteet eivät muuttuisi lainkaan, mutta Irakissa ja Somaliassa syntyneiden elinkaariennusteet olisivat nyt esitettyä negatiivisempia.

Vanhuusiän palveluiden asiakasmaksut. Toinen elinkaariennusteiden euromääräisesti suuri puute henkilöihin liitettävissä suorissa vaikutuksissa on henkilöiden itse – kokonaan tai osittain – maksamat vanhuusiän julkiset palvelut. Henkilöt maksavat osan saamistaan vanhuusiän palveluista itse. Tämä osuus vaihtelee kunnittain, mutta se on riippuvainen henkilön vanhuusiän tuloista. Koska Suomessa syntyneellä kantaväestöllä on ennustetusti suuremmat eläkeiän tulot kuin Somaliassa ja Irakissa syntyneillä maahanmuuttajilla, henkilöiden itse maksamien vanhuusiän palveluiden sisällyttäminen elinkaariennusteisiin muuttaisi Suomessa syntyneen kantaväestön elinkaariennusteita keskimäärin positiivisemmaksi verrattuna Somaliassa ja Irakissa syntyneiden elinkaariennusteisiin.

Elinkaariennusteisiin sisältyvät ja niistä puuttuvat yhteisöihin ja yhtymiin liittyvät vaikutukset. Henkilöihin suoraan liitettävien puuttuvien julkisen talouden nettovaikutusten lisäksi esitin luvussa 10.4, mitkä yhteisöjen ja yhtymien vaikutukset sisältyvät nyt elinkaariennusteisiin – ja miten elinkaariennusteet muuttuisivat, jos elinkaariennusteisiin sisällytettäisiin niistä nyt ”puuttuvat” yritysten ja yhteisöjen vaikutukset. On huomioitava, että puuttuneisuus on tässä kyseenalaista, sillä puuttuvat yhteisöihin liittyvät vaikutukset eivät käytetyn tulkinnan mukaan sisälly tutkimuksessa tarkasteltaviin henkilöiden suoriin julkisen talouden vaikutuksiin.

Elinkaariennusteisiin *sisältyvät* jo nyt sellaiset yhteisöjen ja yhtymien vaikutukset, jotka sisältyvät henkilöverotukseen (ansio- ja pääomatuloverotus). Tällaisia vaikutuksia ovat henkilöiden elinkeinonharjoittajina ja yhtymien toiminnan johdosta maksamat ansio- ja pääomatuloverot. Nämä sisältyvät käytössä olevaan käytössä olevaan rekisteriaineistoon. Sen sijaan yhteisöverot, yhteisöjen maksamat välilliset verot ja kiinteistöverot sekä yhteisöjen saamat yritystuet eivät sisälly elinkaariennusteisiin, sillä näissä verojen maksajina eivät ole henkilöt vaan yhteisöt.

Nämä yhteisöihin liittyvät vaikutukset voitaisiin kuitenkin allokoida henkilöiden kesken. Elinkaariennusteet muuttuisivat sen mukaan, miten tämä allokointi tehtäisiin. Käytännössä kuitenkin jokaisella ”realistisella” allokointitavalla joko Suomessa syntyvien kantaväestön henkilöiden elinkaariennusteet muuttuisivat keskimäärin positiivisemmiksi verrattuna Irakissa ja Somaliassa syntyneiden elinkaariennusteisiin, tai allokointi ei muuttaisi tätä suhdetta elinkaariennusteiden välillä.

13.1.3.2 Ennustetut diskonttauskertoimet

Jotta (diskonttaamattomien) elinkaariennusteiden tuloksia voidaan käyttää uusien maahanmuuttajien valinnassa päätöksentekoteorian puitteissa, elinkaariennusteet on diskontattava nykyarvoon. Nykyarvo on tässä yhteydessä elinkaariennusteiden arvo tässä tutkimuksessa esitettävien elinkaariennusteiden alkuvuonna 2015.

Diskontattavia elinkaariennusteita varten on laadittava ennusteet diskonttauskertoimille. Olen esittänyt luvussa 11, kuinka diskonttauskertoimien mallinnus ja ennusteet on laadittu. Diskonttauskertoimien ennusteet ovat todennäköisyysjakaumia. Tavoitteena mallinnuksessa oli, että diskonttauskertoimien mallin tuottamat ennusteet toistavat diskonttauskertoimien havaitun aikasarja-aineiston viimeisen 50 vuoden ajalta. Varmistin tämän tavoitteen toteutumisen posterioriennustetarkistuksilla.

Koska diskonttauskertoimien havaintoaineisto on pieni, ja bayesilaisessa mallinnuksessa käytettyjen parametrien priorijakaumat ovat leveitä (melko epäinformatiivisia), diskonttauskertoimien mallin tuottamien ennusteiden hajonta – eli niiden diskonttauskertoimien ennusteisiin liittyvä epävarmuus on ”suurta” (ks. tarkemmin diskonttauskertoimien ennusteisiin sisältyvän epävarmuuden suuruudesta luku 11.3.2).

Huomautin lisäksi diskonttauskertoimien mallinnuksen ja ennustamisen yhteydessä luvussa 11.3, että mikäli elinkaariennusteiden kattamassa tulevaisuudessa keskimääräinen reaalikorko jäisi 2010-luvun keskimääräiselle, historiallisen alhaiselle tasolle, tulevat toteutuvat diskonttauskertoimet olisivat keskimäärin käyttämiäni ennusteita korkeampia. Korkeammat keskimääräiset diskonttauskertoimet merkitsisivät sitä, että diskontatut elinkaariennusteet olisivat keskimäärin euromääräisesti lähempänä diskonttaamattomia elinkaariennusteita, kuin ne ovat matalammilla diskonttauskertoimilla. Toisin sanoen, jos muuttaisin käytettyjä diskonttauskertoimien ennusteita niin, että olettaisin diskonttauskertoimien olevan keskimäärin korkeampia, tällöin erityisesti Irakissa ja Somaliassa syntyneiden diskontatuista elinkaariennusteista tulisi keskimäärin negatiivisempia kuin, mitä ne alla esitettävissä diskontatuissa elinkaariennusteissa ovat.

13.1.3.3 Päätöksentekoteoria ja uusien maahanmuuttajien valinta

Esitin luvussa 12 päätöksentekoteorian yleiset perusteet sekä kuinka soveltaa päätöksentekoteoriaa potentiaalisten uusien maahanmuuttajien valintaan. Potentiaalinen uusi maahanmuuttaja on sellainen henkilö, joka muuttaa ensimmäisen kerran Suomeen, mikäli hänet valitaan Suomeen muuttavaksi lainsäädännöllisin keinoin. Kertaan seuraavaksi lyhyesti päätöksentekoteorian soveltamisen ehdot, sekä kuinka nämä toteutuvat potentiaalisten uusien maahanmuuttajien valinnassa (ks. yksityiskohtaisemmat perustelut luvussa 12).

Yleistä päätöksentekoteoriasta

Päätöksentekoteorian soveltamista on pystyttävä määrittämään seuraavat kolme asiaa:

1. Päätöksentekijän kaikki valintavaihdot päätöksentekotilanteessa.

2. Valintojen (tulevaisuudessa) tuottamille tuloksille on määritettävä todennäköisyysjakaumat. Bayesilaisessa viitekehyksessä jakaumat ovat tulosten posteriorijakaumia.

3. Päätöksentekijän hyötyfunktio (utiliteettifunktio). Hyötyfunktion argumenttina ovat päätösten tuottamat tulokset. On kyettävä määrittämään, mikä on päätöksentekijän hyöty jokaisesta mahdollisesta päätöksen tuloksesta.

Päätöksentekoteorian sovellus uusien maahanmuuttajien valintaan

Potentiaalisten uusien maahanmuuttajien osalta päätöksentekijänä on käytännössä eduskunta ja hallitus, jotka laativat ja saattavat toimeen maahanmuuttoa koskevaa (uutta) lainsäädäntöä. Maahanmuuttoa koskevien lakien laatimisen ja toimeenpanon avulla potentiaalisia uusia maahanmuuttajia voidaan valita joko muuttavaksi Suomeen tai jättää valitsematta. Tämä koskee lähinnä EU:n ulkopuolelta tulevia maahanmuuttajia.

Päätöksentekoteoriaa voidaan soveltaa potentiaalisten uusien maahanmuuttajien valintaan, sillä yllä mainitut päätöksentekoteoria soveltamisen kolme ehtoa toteutuvat potentiaalisten uusien maahanmuuttajien valinnassa seuraavasti:

1. Valintavaihtoehtoja on kaksi: Potentiaalinen uusi maahanmuuttaja joko valitaan Suomeen muuttavaksi, tai häntä ei valita Suomeen muuttavaksi.

2. Valintojen (tulevaisuudessa) tuottamien tulosten todennäköisyysjakaumat. Mikäli uusi maahanmuuttaja valitaan Suomeen muuttavaksi, tuloksen todennäköisyysjakauma on uudelle maahanmuuttajalle määritetty julkisen talouden elinkaarivaikutuksen (nykyarvoon diskontattu) posteriorienustejakauma. Mikäli uutta maahanmuuttajaa ei valita Suomeen muuttavaksi, tulos on aina tasan nolla euroa.

3. Julkisen talouden hyötyfunktio on lineaarinen ja euromääräinen erityisesti silloin, kun uusia maahanmuuttajia halutaan valita Suomeen heidän tuottamiensa julkisen talouden nettovaikutusten perusteella. Hyötyfunktio voi olla erilainen, kun huomioidaan muitakin valinnan tuloksia kuin julkisen talouden nettovaikutukset. Kun hyötyfunktio on lineaarinen ja euromääräinen, julkisen talouden hyödyn odotusarvo on odotettavissa oleva rahallinen arvo (Expected Monetary Value, EMV) maahanmuuttajan valinnasta. Tämä on sama asia kuin uuden maahanmuuttajan elinkaarivaikutuksen (nykyarvoon diskontattu) odotusarvo.

Julkisen talouden odotetun euromääräisen hyödyn maksimointi uusien maahanmuuttajien valinnasta

Kun julkisen talouden hyöty uudesta maahanmuuttajasta määräytyy lineaarisesti *vain* uuden maahanmuuttajan julkisen talouden elinkaariennusteesta, julkisen talouden odotettavissa oleva hyöty maksimoituu potentiaalisten uusien maahanmuuttajien valinnoista seuraavan yksinkertaisen säännön avulla:

- Jos potentiaalisen uuden maahanmuuttajan nykyarvoon diskontatun elinkaariennusteen (todennäköisyysjakauman) odotusarvo on negatiivinen – eli nollaa euroa pienempi – potentiaalista uutta maahanmuuttajaa ei kannata valita Suomeen muuttavaksi.
- Jos potentiaalisen uuden maahanmuuttajan nykyarvoon diskontatun elinkaariennusteen (todennäköisyysjakauman) odotusarvo on positiivinen – eli nollaa euroa suurempi – potentiaalinen uusi maahanmuuttaja kannattaa valita Suomeen muuttavaksi.

Irakissa ja Somaliassa syntyneiden potentiaalisten uusien maahanmuuttajien valinta päätöksentekoteorian perusteella

Irakissa ja Somaliassa syntyneiden uusien maahanmuuttajien sekä Suomessa syntyvien kantaväestön henkilöiden nykyarvoon diskontatut elinkaariennusteet esitetään seuraavaksi tässä luvussa. Näiden diskontattujen elinkaariennustetulosten yhteydessä esitetään myös diskontattujen elinkaariennustejakaumien odotusarvot. Odotusarvoja koskevien tulosten perusteella esitän tutkimusta koskevissa johtopäätöksissä myöhemmin tässä luvussa, että Irakissa ja Somaliassa syntyneitä uusia maahanmuuttajia ei pitäisi valita muuttavaksi Suomeen – mikäli valinnoilla halutaan maksimoida Suomen julkisen talouden odotettavissa oleva euromääräinen hyöty uusien maahanmuuttajien valinnoista. Kuten totean johtopäätösten yhteydessä, tämä odotetun hyödyn maksimoiva päätös koskee myös kaikkia Irakissa ja Somalaissa syntyneiden uusien maahanmuuttajien osajoukkoja, jotka on määritelty aineiston ”tärkeimpien” selittävien muuttujien mukaan: eli sukupuolen, maahanmuuttoiän, maahanmuuttovuoden pääasiallisen toiminnan sekä maahanmuuttovuoden koulutustason mukaan.

Aineiston ulkopuolisten tekijöiden vaikutukset potentiaalisten uusien maahanmuuttajien valintaan

Elinkaariennusteet perustuvat elinkaarivaikutusten komponentteja koskevaan havaintoaineistoon, joka koostuu pääosin eri viranomaisrekistereistä kootuista tiedoista. Rekisteriperustainen havaintoaineisto koskee komponenteittain vuosia 1988/1995/2005–2011/2014. Koska elinkaariennusteet sijoittuvat havaintoaineiston ulkopuolelle vuoden 2014 jälkeiseen aikaan, on mahdollista, että elinkaarivaikutusten komponenttien ennusteisiin vaikuttavat nk. aineiston ulkopuoliset tekijät. Näitä tekijöitä ei ole huomioitu havaintoaineistoon perustuvassa elinkaariennusteissa, sillä nimensä mukaisesti ne vaikuttavat elinkaarivaikutusten komponentteihin (tai oikeastaan komponenttien omiin parametreihin) havaintoaineiston ulkopuolelta.

Käsittelin päätöksentekoteoriaa koskevassa luvussa 12.3 (ks. myös johdannon luku 2.3.4) sitä, kuinka nämä aineiston ulkopuolisten tekijöiden vaikutukset voivat vaikuttaa uusien maahanmuuttajien elinkaariennusteiden jakaumiin (ks. erityisesti luvut 12.3.1–12.3.3). Esitin erilaisista vaikutuksista esimerkkejä sekä määritin kriteerin sille, milloin aineiston ulkopuolisten tekijöiden vaikutukset muuttavat julkisen talouden odotettavissa olevan hyödyn maksimoivan valintapäätökseen uusien maahanmuuttajien valinnasta. Odotettavissa olevan hyödyn maksimoiva päätös vaihtuu, mikäli aineiston ulkopuolisten tekijöiden vaikutuksesta uuden maahanmuuttajan diskontatun elinkaariennustejakauman odotusarvon etumerkki vaihtuu.

Jotta diskontattujen elinkaarivaikutusten jakaumien odotusarvot vaihtaisivat etumerkkiä aineiston ulkopuolisten tekijöiden vaikutuksesta, elinkaariennusteen laatijalla pitäisi olla käytettävissään jotain sellaista

havaintoaineiston ulkopuolista informaatiota, joka johtaa tällaiseen ennustejakauman siirtymiseen. Tällaista havaintoaineiston ulkopuolista informaatiota minulla ei ole.

On tärkeää huomata tässä yhteydessä, että jos aineiston ulkopuolisten tekijöiden vaikutukset tuovat pelkästään lisäepävarmuutta elinkaariennusteisiin, tällöin elinkaariennusteiden jakaumat levenevät, mutta niiden odotusarvot eivät muutu. Tämä tarkoittaa sitä, että myöskään julkisen talouden odotettavissa olevan hyödyn maksimoiva päätös ei muutu – verrattuna päätöstilanteeseen, jossa näitä aineiston ulkopuolisten tekijöiden vaikutuksia ei huomioida elinkaariennusteissa. Osoitin tämän teoriassa päätöksentekoteoriaan koskevan luvun luvussa 12.3.3. Osoitan saman jäljempänä luvussa 13.2.4 myös käytännössä: Kun lisään elinkaariennusteen komponentin parametrin ennusteeseen pelkkää kohinaa (lisäepävarmuutta), tämä vain lisää elinkaariennusteisiin sisältyvän epävarmuuden määrää, mutta koska diskontattujen elinkaariennusteiden odotusarvot eivät muutu, myöskään julkisen talouden hyödyn maksimoiva päätös ei puutu.

13.2 Nykyarvoon diskontatut elinkaariennusteet

Esitän seuraavissa luvuissa nykyarvoon – eli vuoteen 2015 – diskontatut elinkaariennusteet. En selosta nykyarvoon diskontattujen elinkaariennusteiden tulosten tulkintaa, koska elinkaariennusteiden tulosten tulkinta jo esitetty diskonttaamattomien elinkaariennusteiden tulosten esittämisen yhteydessä (luvussa 9.5).

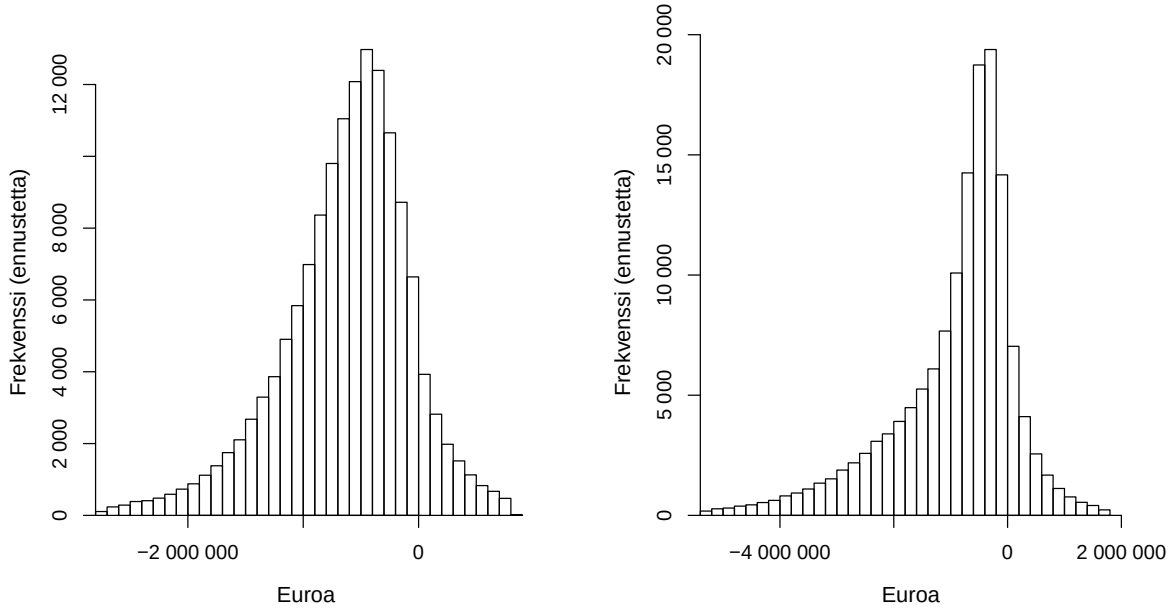
13.2.1 Somalia

13.2.1.1 Koko ikä maahanmuutosta kuolemaan asti Suomessa

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.1: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-641	-922
Keskihajonta	666	1 423
MC-keskivirhe	13	18
2,5:s prosenttipiste (2,5 %)	-2 147	-4 168
10. prosenttipiste (10 %)	-1 397	-2 543
Alakvartiili (25 %)	-937	-1 398
Mediaani (50 %)	-560	-615
Yläkvartiili (75 %)	-264	-231
90. prosenttipiste (90 %)	3	137
97,5:s prosenttipiste (97,5 %)	479	947



Kuva 13.1: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.2: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-863	685	-1 636	-1 156	-759	-450	-210
Miehet	242	-414	562	-1 026	-673	-383	-124	159
Naiset (ml. LV)	248	-1 419	1 787	-3 320	-2 228	-1 242	-498	82
Miehet (ml. LV)	242	-414	562	-1 026	-673	-383	-124	159

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.3: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-590	841	-1 552	-1 014	-517	-90	309
20–24	81	-772	762	-1 651	-1 143	-687	-298	23
25–29	85	-720	627	-1 460	-1 034	-650	-313	-47
30–34	64	-677	517	-1 277	-930	-616	-336	-122
35–39	34	-542	434	-1 045	-760	-490	-254	-74
40–49	51	-601	350	-1 002	-749	-548	-382	-240
50–62	33	-428	268	-739	-536	-384	-257	-152
16–19 (ml. LV)	142	-948	1 616	-2 841	-1 658	-663	-53	467
20–24 (ml. LV)	81	-1 096	1 524	-2 862	-1 744	-813	-250	166
25–29 (ml. LV)	85	-995	1 445	-2 581	-1 498	-705	-261	70
30–34 (ml. LV)	64	-932	1 327	-2 366	-1 357	-671	-284	-10
35–39 (ml. LV)	34	-876	1 380	-2 519	-1 088	-502	-207	6
40–49 (ml. LV)	51	-793	1 123	-1 825	-965	-584	-361	-166
50–62 (ml. LV)	33	-428	268	-739	-536	-384	-257	-152

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.4: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-246	504	-218
Opiskelija	48	-573	636	-491
Työtön	23	-736	596	-639
Muu	406	-656	673	-574
Työllinen (ml. LV)	13	-289	689	-222
Opiskelija (ml. LV)	48	-729	1 165	-499
Työtön (ml. LV)	23	-999	1 372	-723
Muu (ml. LV)	406	-961	1 464	-646

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.5: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

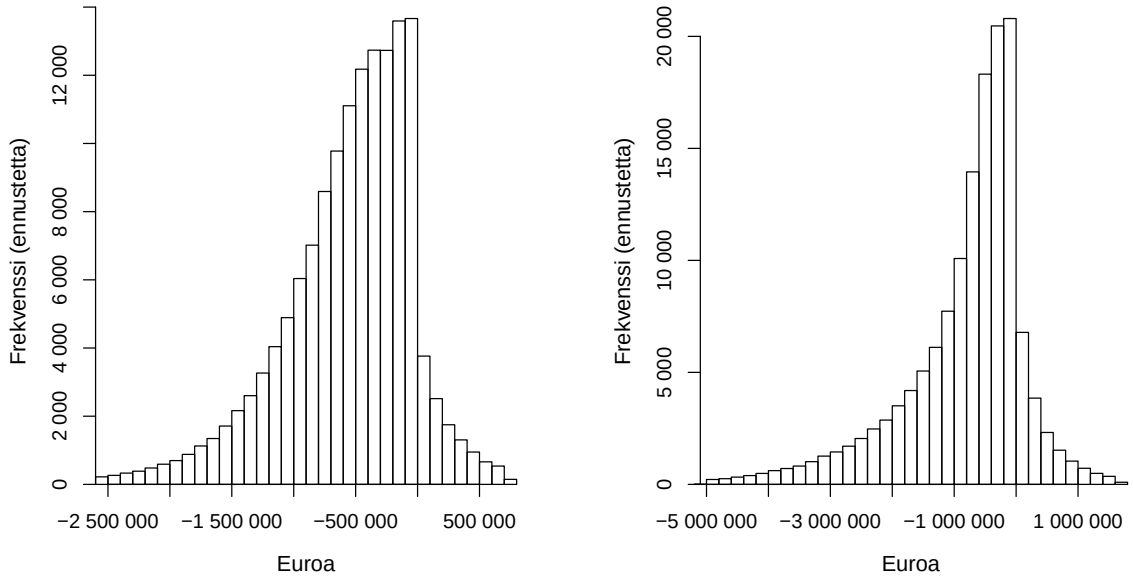
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-591	536	-515
Ei 2. asteen koulutusta/ei tietoa	458	-645	674	-563
Vähintään 2. asteen koulutus (ml. LV)	32	-809	1 267	-535
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-930	1 433	-622

13.2.1.2 Suomessaoloaika vuosina 2000–2009 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.6: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-562	-812
Keskihajonta	627	1 308
MC-keskivirhe	11	15
2,5:s prosenttipiste (2,5 %)	-2 011	-3 829
10. prosenttipiste (10 %)	-1 286	-2 241
Alakvartiili (25 %)	-838	-1 207
Mediaani (50 %)	-462	-544
Yläkvartiili (75 %)	-172	-173
90. prosenttipiste (90 %)	-10	101
97,5:s prosenttipiste (97,5 %)	408	885



Kuva 13.2: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.7: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-744	663	-1 508	-1 037	-632	-307	-96
Miehet	242	-375	526	-961	-616	-320	-90	110
Naiset (ml. LV)	248	-1 238	1 656	-2 997	-1 926	-1 035	-422	76
Miehet (ml. LV)	242	-375	526	-961	-616	-320	-90	110

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.8: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-497	767	-1 397	-863	-361	-64	231
20–24	81	-661	724	-1 521	-1 012	-534	-167	-4
25–29	85	-635	610	-1 374	-940	-540	-206	-35
30–34	64	-608	504	-1 217	-863	-533	-248	-77
35–39	34	-491	435	-1 006	-707	-422	-187	-51
40–49	51	-551	352	-968	-716	-505	-317	-158
50–62	33	-408	284	-728	-529	-366	-223	-110
16–19 (ml. LV)	142	-809	1 456	-2 465	-1 371	-540	-58	388
20–24 (ml. LV)	81	-948	1 408	-2 545	-1 490	-683	-171	138
25–29 (ml. LV)	85	-883	1 345	-2 310	-1 321	-629	-194	49
30–34 (ml. LV)	64	-842	1 238	-2 170	-1 225	-604	-226	-11
35–39 (ml. LV)	34	-792	1 286	-2 229	-1 007	-461	-168	-7
40–49 (ml. LV)	51	-723	1 064	-1 643	-905	-542	-305	-114
50–62 (ml. LV)	33	-408	284	-728	-529	-366	-223	-110

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.9: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-226	462	-182
Opiskelija	48	-524	604	-431
Työtön	23	-650	572	-542
Muu	406	-572	633	-471
Työllinen (ml. LV)	13	-264	625	-189
Opiskelija (ml. LV)	48	-665	1 094	-459
Työtön (ml. LV)	23	-891	1 305	-648
Muu (ml. LV)	406	-842	1 341	-567

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.10: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

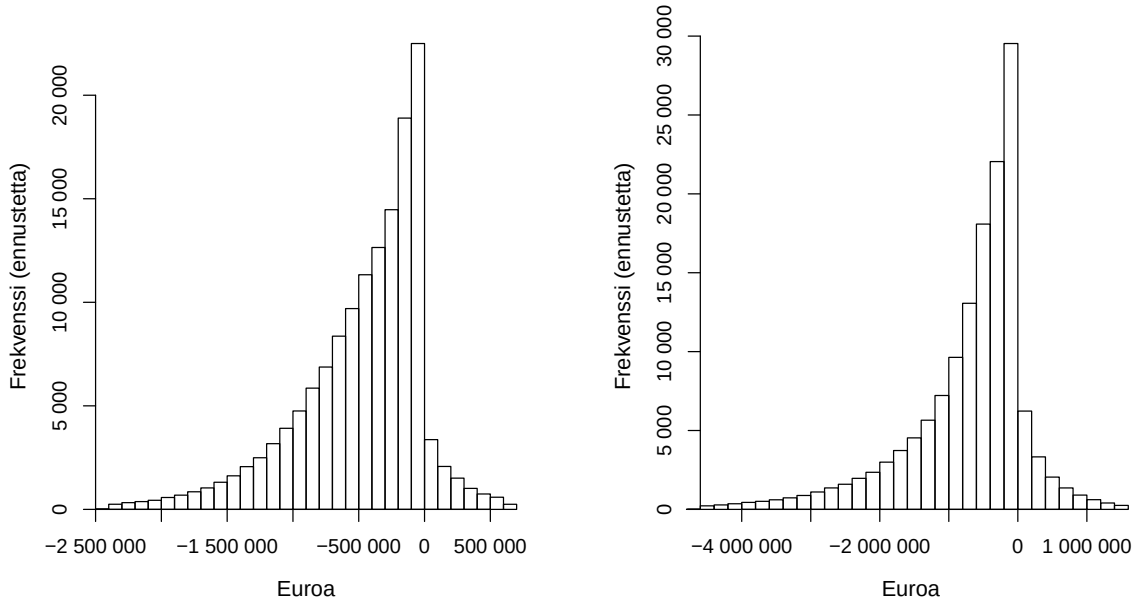
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-537	520	-460
Ei 2. asteen koulutusta/ei tietoa	458	-563	633	-463
Vähintään 2. asteen koulutus (ml. LV)	32	-736	1 198	-493
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-817	1 315	-549

13.2.1.3 Suomessaoloaika vuosina 1994–1999 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.11: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-478	-702
Keskihajonta	578	1 192
MC-keskivirhe	9	13
2,5:s prosenttipiste (2,5 %)	-1 865	-3 472
10. prosenttipiste (10 %)	-1 160	-1 941
Alakvartiili (25 %)	-718	-1 030
Mediaani (50 %)	-351	-453
Yläkvartiili (75 %)	-115	-129
90. prosenttipiste (90 %)	-11	46
97,5:s prosenttipiste (97,5 %)	332	796



Kuva 13.3: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.12: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-624	624	-1 370	-899	-484	-185	-57
Miehet	242	-328	482	-876	-535	-244	-70	56
Naiset (ml. LV)	248	-1 066	1 521	-2 642	-1 643	-847	-338	20
Miehet (ml. LV)	242	-328	482	-876	-535	-244	-70	56

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.13: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-407	684	-1 231	-680	-223	-52	142
20–24	81	-549	679	-1 372	-860	-368	-105	-8
25–29	85	-538	573	-1 259	-816	-399	-132	-25
30–34	64	-528	483	-1 143	-770	-425	-172	-54
35–39	34	-433	408	-936	-641	-351	-137	-37
40–49	51	-493	362	-912	-665	-442	-234	-103
50–62	33	-380	276	-701	-505	-340	-187	-83
16–19 (ml. LV)	142	-672	1 304	-2 075	-1 116	-408	-60	294
20–24 (ml. LV)	81	-808	1 288	-2 215	-1 264	-542	-127	83
25–29 (ml. LV)	85	-759	1 234	-2 033	-1 148	-506	-133	18
30–34 (ml. LV)	64	-748	1 151	-1 916	-1 083	-517	-173	-15
35–39 (ml. LV)	34	-712	1 181	-1 979	-907	-409	-132	-10
40–49 (ml. LV)	51	-663	1 011	-1 524	-835	-484	-231	-76
50–62 (ml. LV)	33	-380	276	-701	-505	-340	-187	-83

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.14: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-205	432	-137
Opiskelija	48	-466	566	-352
Työtön	23	-554	537	-429
Muu	406	-484	583	-354
Työllinen (ml. LV)	13	-239	593	-143
Opiskelija (ml. LV)	48	-599	1 006	-392
Työtön (ml. LV)	23	-763	1 207	-535
Muu (ml. LV)	406	-725	1 222	-470

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.15: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

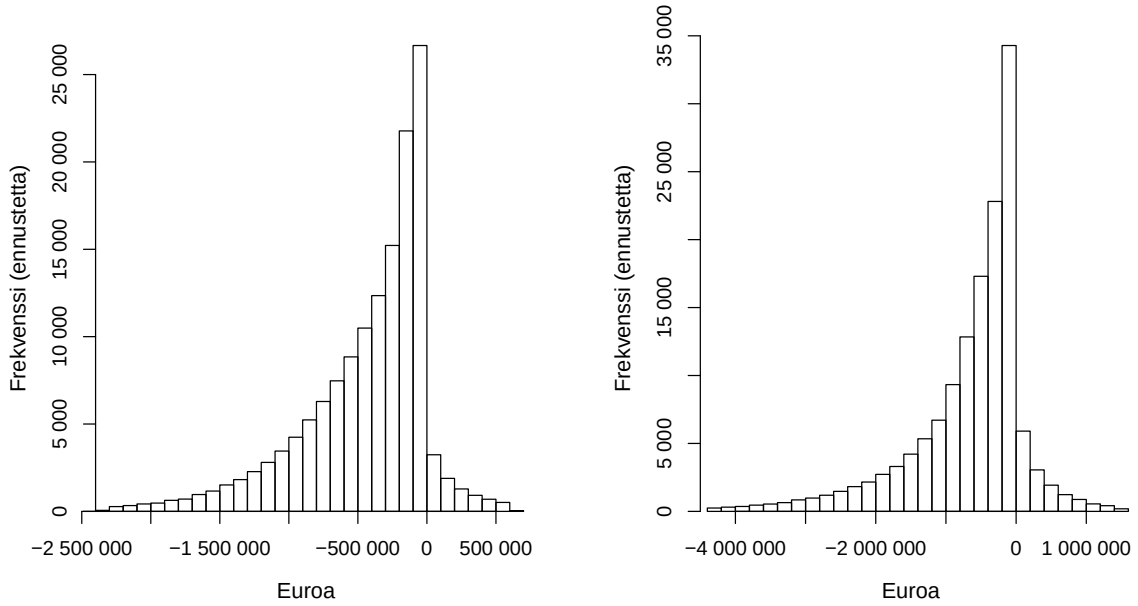
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-476	508	-376
Ei 2. asteen koulutusta/ei tietoa	458	-478	582	-349
Vähintään 2. asteen koulutus (ml. LV)	32	-665	1 117	-427
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-704	1 198	-455

13.2.1.4 Suomessaoloaika ennen vuotta 1994 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.16: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	490	490
Odotusarvo	-442	-654
Keskihajonta	556	1 144
MC-keskivirhe	8	12
2,5:s prosenttipiste (2,5 %)	-1 798	-3 306
10. prosenttipiste (10 %)	-1 106	-1 837
Alakvartiili (25 %)	-664	-954
Mediaani (50 %)	-299	-407
Yläkvartiili (75 %)	-100	-114
90. prosenttipiste (90 %)	-11	20
97,5:s prosenttipiste (97,5 %)	296	768



Kuva 13.4: *Vasemmalla Somaliassa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.17: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	248	-575	606	-1 312	-836	-415	-154	-50
Miehet	242	-306	461	-836	-492	-206	-64	28
Naiset (ml. LV)	248	-993	1 464	-2 513	-1 530	-765	-303	-1
Miehet (ml. LV)	242	-306	461	-836	-492	-206	-64	28

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.18: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	142	-368	646	-1 156	-583	-182	-50	95
20–24	81	-505	649	-1 310	-778	-304	-93	-9
25–29	85	-497	559	-1 208	-756	-335	-112	-21
30–34	64	-497	483	-1 106	-735	-375	-144	-48
35–39	34	-406	404	-902	-600	-310	-120	-34
40–49	51	-465	354	-891	-643	-412	-200	-87
50–62	33	-361	284	-689	-482	-315	-161	-68
16–19 (ml. LV)	142	-619	1 223	-1 933	-1 008	-345	-62	259
20–24 (ml. LV)	81	-750	1 222	-2 101	-1 161	-479	-110	60
25–29 (ml. LV)	85	-710	1 207	-1 936	-1 063	-453	-117	2
30–34 (ml. LV)	64	-700	1 117	-1 849	-1 023	-474	-147	-14
35–39 (ml. LV)	34	-674	1 169	-1 863	-861	-373	-120	-9
40–49 (ml. LV)	51	-625	996	-1 435	-801	-457	-203	-67
50–62 (ml. LV)	33	-361	284	-689	-482	-315	-161	-68

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.19: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	13	-201	425	-123
Opiskelija	48	-437	546	-309
Työtön	23	-523	535	-376
Muu	406	-446	560	-299
Työllinen (ml. LV)	13	-235	580	-127
Opiskelija (ml. LV)	48	-560	986	-350
Työtön (ml. LV)	23	-731	1 111	-500
Muu (ml. LV)	406	-674	1 173	-423

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.20: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	32	-446	478	-338
Ei 2. asteen koulutusta/ei tietoa	458	-442	561	-296
Vähintään 2. asteen koulutus (ml. LV)	32	-618	1 034	-391
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	458	-656	1 151	-409

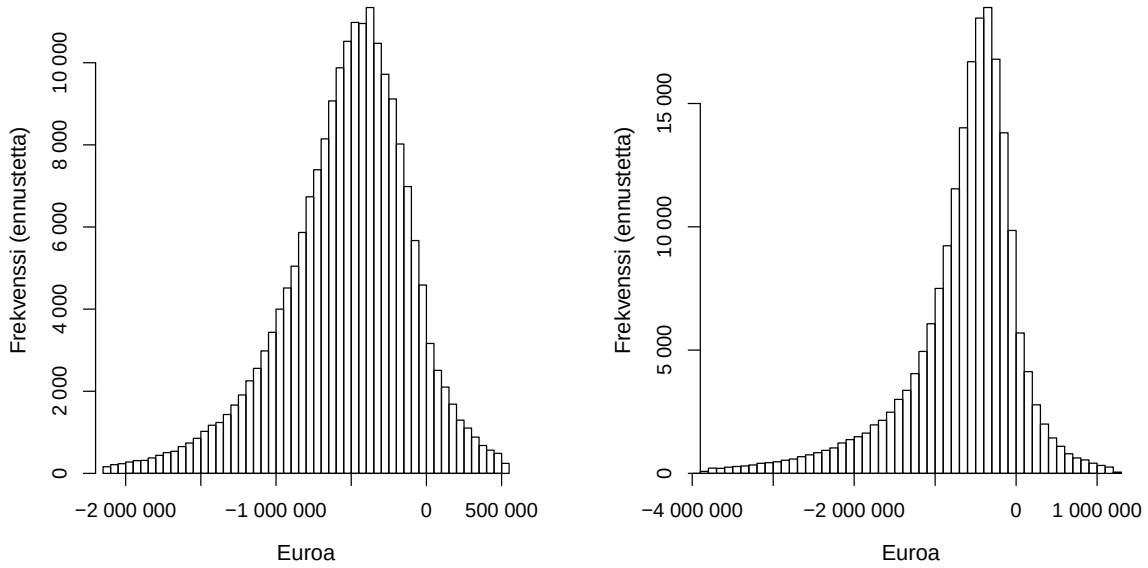
13.2.2 Irak

13.2.2.1 Koko ikä maahanmuutosta kuolemaan asti Suomessa

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.21: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-542	-656
Keskihajonta	504	987
MC-keskivirhe	10	12
2,5:s prosenttipiste (2,5 %)	-1 678	-2 889
10. prosenttipiste (10 %)	-1 111	-1 596
Alakvartiili (25 %)	-770	-910
Mediaani (50 %)	-486	-509
Yläkvartiili (75 %)	-248	-231
90. prosenttipiste (90 %)	-40	29
97,5:s prosenttipiste (97,5 %)	290	597



Kuva 13.5: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.22: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-665	525	-1 258	-889	-588	-355	-155
Miehet	460	-484	483	-1 029	-709	-438	-205	2
Naiset (ml. LV)	217	-1 020	1 533	-2 572	-1 649	-863	-346	168
Miehet (ml. LV)	460	-484	483	-1 029	-709	-438	-205	2

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.23: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-547	629	-1 261	-852	-491	-176	112
20–24	153	-562	572	-1 219	-838	-508	-214	41
25–29	147	-554	520	-1 143	-799	-501	-238	-16
30–34	108	-544	450	-1 063	-763	-494	-259	-73
35–39	62	-548	384	-992	-730	-502	-299	-139
40–49	70	-516	326	-883	-654	-469	-309	-177
50–62	34	-432	268	-731	-535	-389	-266	-157
16–19 (ml. LV)	103	-666	1 060	-1 751	-1 008	-523	-160	183
20–24 (ml. LV)	153	-684	1 049	-1 733	-1 001	-538	-195	120
25–29 (ml. LV)	147	-671	988	-1 637	-945	-526	-217	52
30–34 (ml. LV)	108	-677	1 057	-1 658	-924	-522	-236	-2
35–39 (ml. LV)	62	-710	1 104	-1 760	-969	-550	-275	-51
40–49 (ml. LV)	70	-574	660	-1 112	-723	-481	-298	-144
50–62 (ml. LV)	34	-432	268	-731	-535	-389	-266	-157

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.24: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-208	433	-187
Opiskelija	77	-580	515	-522
Työtön	79	-542	478	-488
Muu	493	-555	503	-496
Työllinen (ml. LV)	28	-208	433	-187
Opiskelija (ml. LV)	77	-720	1 034	-555
Työtön (ml. LV)	79	-615	814	-505
Muu (ml. LV)	493	-678	1 020	-523

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.25: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

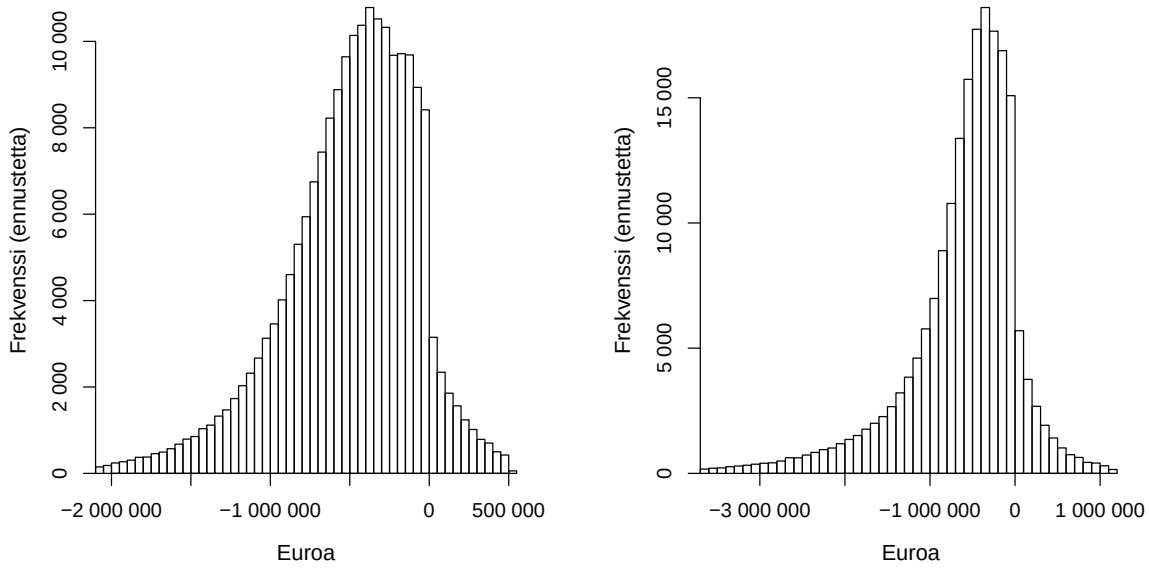
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-525	434	-469
Ei 2. asteen koulutusta/ei tietoa	543	-547	520	-491
Vähintään 2. asteen koulutus (ml. LV)	134	-570	783	-450
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-618	978	-472

13.2.2.2 Suomessaoloaika vuosina 2000–2009 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.26: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-502	-608
Keskihajonta	496	943
MC-keskivirhe	10	11
2,5:s prosenttipiste (2,5 %)	-1 630	-2 721
10. prosenttipiste (10 %)	-1 071	-1 490
Alakvartiili (25 %)	-728	-856
Mediaani (50 %)	-437	-467
Yläkvartiili (75 %)	-191	-186
90. prosenttipiste (90 %)	-28	13
97,5:s prosenttipiste (97,5 %)	271	561



Kuva 13.6: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.27: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-609	521	-1 201	-838	-530	-278	-91
Miehet	460	-452	475	-997	-673	-395	-159	-6
Naiset (ml. LV)	217	-939	1 461	-2 405	-1 504	-778	-306	142
Miehet (ml. LV)	460	-452	475	-997	-673	-395	-159	-6

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.28: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-505	605	-1 205	-802	-434	-130	89
20–24	153	-521	562	-1 173	-792	-452	-161	20
25–29	147	-514	507	-1 109	-759	-450	-182	-13
30–34	108	-503	456	-1 034	-719	-442	-199	-46
35–39	62	-507	388	-962	-698	-455	-242	-96
40–49	70	-476	332	-858	-629	-432	-255	-118
50–62	34	-402	270	-715	-515	-363	-227	-111
16–19 (ml. LV)	103	-616	999	-1 627	-942	-479	-129	162
20–24 (ml. LV)	153	-637	1 004	-1 615	-940	-492	-156	104
25–29 (ml. LV)	147	-624	946	-1 535	-892	-483	-175	34
30–34 (ml. LV)	108	-628	1 016	-1 550	-865	-476	-191	-8
35–39 (ml. LV)	62	-657	1 058	-1 634	-904	-509	-233	-36
40–49 (ml. LV)	70	-530	631	-1 059	-688	-444	-247	-94
50–62 (ml. LV)	34	-402	270	-715	-515	-363	-227	-111

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.29: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-200	434	-164
Opiskelija	77	-550	510	-481
Työtön	79	-501	471	-439
Muu	493	-512	495	-446
Työllinen (ml. LV)	28	-200	434	-164
Opiskelija (ml. LV)	77	-682	993	-525
Työtön (ml. LV)	79	-567	788	-458
Muu (ml. LV)	493	-627	973	-480

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.30: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

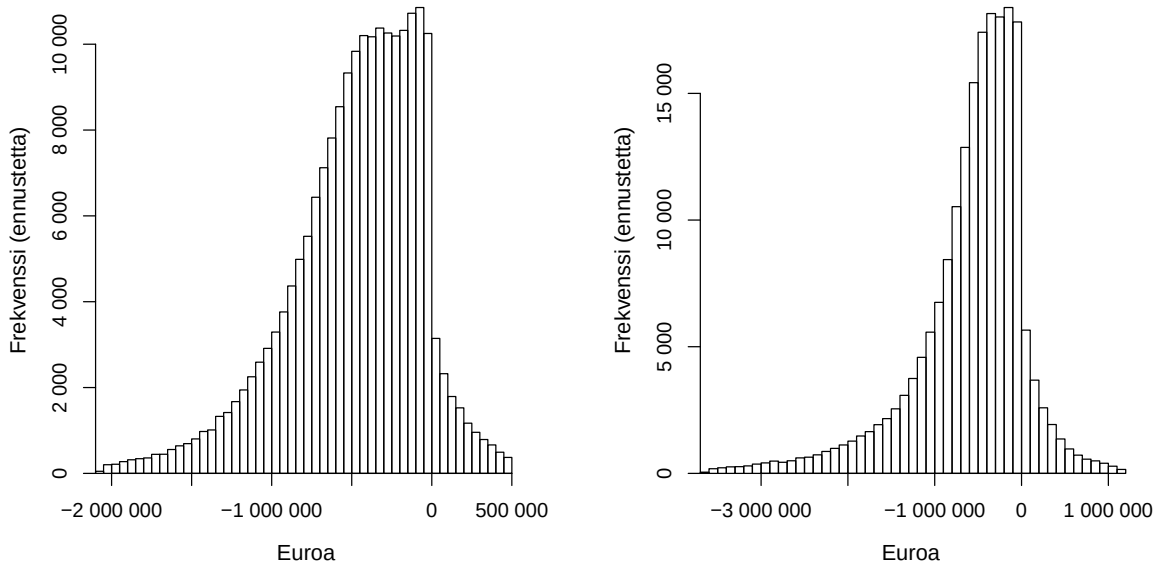
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-487	430	-428
Ei 2. asteen koulutusta/ei tietoa	543	-506	510	-440
Vähintään 2. asteen koulutus (ml. LV)	134	-570	783	-450
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-618	978	-472

13.2.2.3 Suomessaoloaika vuosina 1994–1999 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.31: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-484	-586
Keskihajonta	490	928
MC-keskivirhe	9	11
2,5:s prosenttipiste (2,5 %)	-1 612	-2 687
10. prosenttipiste (10 %)	-1 054	-1 442
Alakvartiili (25 %)	-708	-833
Mediaani (50 %)	-415	-447
Yläkvartiili (75 %)	-168	-165
90. prosenttipiste (90 %)	-22	7
97,5:s prosenttipiste (97,5 %)	261	552



Kuva 13.7: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histogrami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.32: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-586	518	-1 188	-817	-506	-243	-70
Miehet	460	-436	469	-981	-655	-374	-140	-7
Naiset (ml. LV)	217	-904	1 438	-2 344	-1 445	-741	-284	138
Miehet (ml. LV)	460	-436	469	-981	-655	-374	-140	-7

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.33: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-490	601	-1 189	-779	-412	-114	81
20–24	153	-502	554	-1 156	-768	-424	-138	14
25–29	147	-495	503	-1 092	-740	-424	-157	-11
30–34	108	-485	445	-1 010	-704	-422	-178	-40
35–39	62	-487	387	-942	-678	-433	-214	-78
40–49	70	-457	334	-841	-611	-413	-225	-98
50–62	34	-387	276	-698	-507	-353	-205	-87
16–19 (ml. LV)	103	-596	991	-1 577	-913	-458	-118	155
20–24 (ml. LV)	153	-609	983	-1 564	-910	-465	-135	93
25–29 (ml. LV)	147	-601	929	-1 485	-871	-458	-154	31
30–34 (ml. LV)	108	-608	1 006	-1 504	-849	-458	-171	-7
35–39 (ml. LV)	62	-637	1 016	-1 583	-881	-490	-210	-35
40–49 (ml. LV)	70	-509	640	-1 042	-669	-425	-219	-77
50–62 (ml. LV)	34	-387	276	-698	-507	-353	-205	-87

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.34: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-193	421	-150
Opiskelija	77	-538	510	-470
Työtön	79	-480	461	-411
Muu	493	-493	490	-423
Työllinen (ml. LV)	28	-193	421	-150
Opiskelija (ml. LV)	77	-666	985	-510
Työtön (ml. LV)	79	-546	766	-437
Muu (ml. LV)	493	-603	957	-458

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.35: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

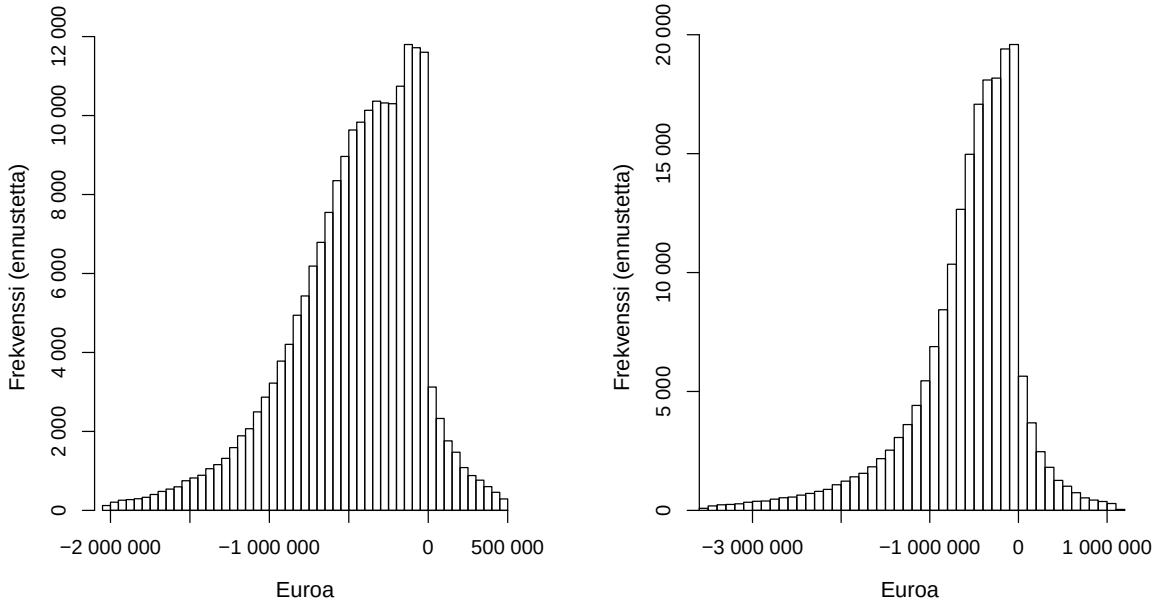
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-466	426	-405
Ei 2. asteen koulutusta/ei tietoa	543	-489	505	-418
Vähintään 2. asteen koulutus (ml. LV)	134	-547	760	-429
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-596	964	-452

13.2.2.4 Suomessaoloaika ennen vuotta 1994 Suomeen muuttaneiden mukaan

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.36: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman LV	ml. LV
Ennusteväestö (N)	677	677
Odotusarvo	-472	-573
Keskihajonta	484	902
MC-keskivirhe	9	10
2,5:s prosenttipiste (2,5 %)	-1 588	-2 619
10. prosenttipiste (10 %)	-1 036	-1 410
Alakvartiili (25 %)	-695	-818
Mediaani (50 %)	-399	-434
Yläkvartiili (75 %)	-154	-155
90. prosenttipiste (90 %)	-20	0
97,5:s prosenttipiste (97,5 %)	247	534



Kuva 13.8: *Vasemmalla Irakissa syntyneiden ennusteväestön elinkaariennusteiden yhteisjakauman histogrammi ilman lasten vaikutuksia, oikealla histograami, kun elinkaariennusteisiin sisältyy lasten yhteenlasketut ennustetut elinkaarivaikutukset. Ennustejakaumista on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.37: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	217	-568	514	-1 161	-801	-486	-222	-65
Miehet	460	-427	463	-967	-644	-361	-130	-7
Naiset (ml. LV)	217	-884	1 394	-2 294	-1 401	-723	-275	128
Miehet (ml. LV)	460	-427	463	-967	-644	-361	-130	-7

Elinkaariennusteet maahanmuuton ikäluokan mukaan

Taulukko 13.38: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuton ikäluokan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuton ikäluokan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ikäluokka	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
16–19	103	-477	594	-1 161	-762	-390	-106	70
20–24	153	-486	545	-1 137	-750	-403	-125	7
25–29	147	-484	495	-1 069	-727	-408	-148	-11
30–34	108	-474	442	-1 002	-696	-406	-162	-39
35–39	62	-476	388	-937	-668	-421	-197	-67
40–49	70	-447	328	-837	-605	-404	-213	-88
50–62	34	-381	284	-700	-499	-338	-192	-80
16–19 (ml. LV)	103	-584	958	-1 545	-902	-442	-111	138
20–24 (ml. LV)	153	-593	955	-1 514	-889	-447	-128	76
25–29 (ml. LV)	147	-590	895	-1 447	-850	-446	-145	22
30–34 (ml. LV)	108	-594	977	-1 455	-836	-447	-160	-9
35–39 (ml. LV)	62	-622	1 014	-1 546	-862	-478	-195	-29
40–49 (ml. LV)	70	-499	628	-1 020	-662	-419	-207	-69
50–62 (ml. LV)	34	-381	284	-700	-499	-338	-192	-80

Elinkaariennusteet muuttovuoden pääasiallisen toiminnan mukaan

Taulukko 13.39: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden pääasiallisen toiminnan mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden pääasiallisen toiminnan mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Pääasiallinen toiminta	Ennusteväestö (N)	OA	KH	50 %
Työllinen	28	-189	402	-140
Opiskelija	77	-522	498	-453
Työtön	79	-472	463	-403
Muu	493	-480	485	-405
Työllinen (ml. LV)	28	-189	402	-140
Opiskelija (ml. LV)	77	-652	969	-499
Työtön (ml. LV)	79	-539	738	-428
Muu (ml. LV)	493	-588	930	-445

Elinkaariennusteet muuttovuoden koulutustason mukaan

Taulukko 13.40: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet maahanmuuttovuoden koulutustason mukaan. Ennustetulokset ilman lasten vaikutuksia sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien (”ml. LV”). Maahanmuuttovuoden koulutustason mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden odotusarvojen (OA), keskihajontojen (KH) sekä mediaanien (50 %) estimaatit. Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

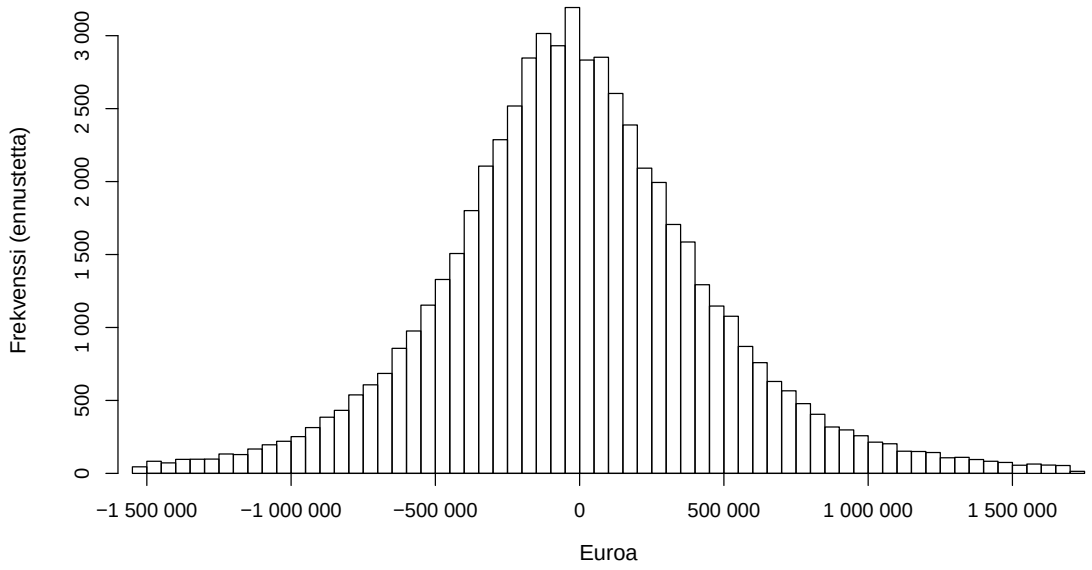
Koulutustaso	Ennusteväestö (N)	OA	KH	50 %
Vähintään 2. asteen koulutus	134	-457	425	-392
Ei 2. asteen koulutusta/ei tietoa	543	-476	498	-401
Vähintään 2. asteen koulutus (ml. LV)	134	-536	759	-416
Ei 2. asteen koulutusta/ei tietoa (ml. LV)	543	-582	934	-440

13.2.3 Suomi

Elinkaariennusteet koko ennustepopulaatiolle yhteensä

Taulukko 13.41: *Suomessa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Ennusteväestö (N)	200
Odotusarvo	0
Keskihajonta	754
MC-keskivirhe	8
2,5:s prosenttipiste (2,5 %)	-1 105
10. prosenttipiste (10 %)	-600
Alakvartiili (25 %)	-297
Mediaani (50 %)	-25
Yläkvartiili (75 %)	263
90. prosenttipiste (90 %)	593
97,5:s prosenttipiste (97,5 %)	1 184



Kuva 13.9: *Suomessa syntyneiden ennusteväestön diskontattujen elinkaariennusteiden yhteisjakauman histogrammi. Ennustejakaumasta on poistettu alin ja ylin prosentti yksityisyydensuojan varmistamiseksi.*

Elinkaariennusteet sukupuolen mukaan

Taulukko 13.42: *Suomessa syntyneen ennusteväestön diskontatut elinkaariennusteet sukupuolen mukaan. Sukupuolen mukaan eriteltyjen ennustejakaumien henkilömäärät ennusteväestössä (N), ennusteiden keskiarvon (OA), keskihajonnan (KH) sekä prosenttipisteiden estimaatit:* $10, 25, 50, 75, 90$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

Sukupuoli	Ennusteväestö (N)	OA	KH	10 %	25 %	50 %	75 %	90 %
Naiset	97	-154	741	-742	-419	-148	106	393
Miehet	103	145	737	-419	-161	93	386	722

13.2.4 Lisäepävarmuutta sisältävät nykyarvoon diskontatut elinkaariennusteet

Käsittelin luvussa 12 päätöksentekoteoriaa ja uusien maahanmuuttajien valintaa Suomeen, niin että valinnat maksimoivat Suomen julkisen talouden euromääräisen hyödyn valintojen seurauksista. Määritin päätöksentekoteoriaan perustuen, että potentiaalinen uusi maahanmuuttaja kannattaa valita Suomeen, mikäli hänen diskontatun elinkaariennusteen jakaumansa odotusarvo on positiivinen ja toisaalta jättää valitsematta Suomeen muuttavaksi, mikäli hänen diskontatut elinkaariennusteen jakaumansa on negatiivinen.

Olen yllä esittänyt diskontattujen elinkaariennusteiden jakaumat potentiaalisille uusille maahanmuuttajille, jotka ovat syntyneet joko Irakissa tai Somaliassa. Nämä elinkaariennustejakaumat perustuvat lähes täysin informaatioon, joka perustuu vuosiin 1988–2011/2014 sijoittuvaan tutkimusaineistoon. Kuten esitin luvussa 12.3, elinkaariennusteen tekijä voi sisällyttää havaintoaineiston ulkopuolisista tekijöistä johtuvia vaikutuksia. Nämä havaintoaineiston ulkopuoliset vaikutukset johtuvat siitä, että elinkaariennusteet ulottuvat havaintoaineiston ulkopuolisille vuosille vuoden 2014 jälkeiseen aikaan, ja elinkaariennusteen laatijalla voi olla tietoa siitä, miten ne vaikuttavat elinkaariennusteisiin – esimerkiksi niihin liittyviä epävarmuuksia.

Esitin edelleen luvussa 12.3, kuinka aineiston ulkopuoliset tekijät voivat vaikuttaa yhden malliparametrin posteriorijakaumaan (maasaoloajan regressiokertoimeen työikäisten nettovaikutusmallissa) sekä kuinka ne voivat mahdollisesti muuttaa elinkaariennusteiden jakaumia. Jälkimmäisessä tapauksessa määritin myös säännön sille, milloin aineiston ulkopuolisten tekijöiden vaikutukset voivat muuttaa potentiaalista uutta maahanmuuttajaa koskevan valintaratkaisun – verrattuna elinkaariennusteeseen, jossa näitä tekijöitä ei ole huomioitu.

Edelliseen liittyen esitin luvussa 12.3, että jos työikäisten nettovaikutusmallin yhteydessä elinkaarivaikutusten mallin laatijalla ei ole sellaista informaatiota, joka kertoisi, muuttaako vuoden 2014 (eli tutkimusaineiston jälkeinen) tulevaisuus regressiokertoimien jakaumia positiivisemmiksi vai negatiivisemmiksi aineistosta estimoituihin jakaumiin verrattuna, elinkaariennusteen laatija voi kuvata tätä epävarmuutta lisäämällä havaintoaineistosta estimoituihin regressiokertoimiin esimerkiksi Studentin t-jakautuneen satunnaisen termin, jonka odotusarvo on nolla. Odotusarvo on nolla, koska elinkaariennusteen laatija ei tiedä, vaikuttaako tulevaisuuteen liittyvä epävarmuus todennäköisemmin positiivisesti vai negatiivisesti regressiokertoimen arvoihin. Pienivapausasteinen Studentin t-jakauma mahdollistaa kuitenkin isojen poikkeamien liittämisen pelkästään aineistosta estimoituihin jakaumiin.

Tällaisessa tapauksessa, jossa regressiokertoimien jakaumiin lisätään odotusarvoltaan nolla oleva satunnainen termi, näiden regressiokertoimien odotusarvot eivät muutu eikä siten myöskään elinkaariennusteiden odotusarvo muutu. Tällöin myöskään elinkaariennusteen pohjalta tehtävä valintapäätös koskien uutta maahanmuuttajaa ei muutu tulevaisuuden tuoman epävarmuuden vuoksi. Toisaalta elinkaariennusteen jakaumaan tulee näin enemmän epävarmuutta, joten sen keskihajonta kasvaa.

Epävarmuuden lisääminen elinkaariennusteisiin. Esitän tästä syystä diskontatut elinkaariennusteiden tulokset esimerkinomaisesti Irakissa ja Somaliassa syntyneille siten, että työikäisten nettovaikutusennusteissa maassaoloajan regressiokertoimeen on lisätty ”paljon” epävarmuutta. Lisäsin työikäisten nettovaikutusennusteissa maassaoloajan lähes pelkästä havaintoaineistosta estimoituun posteriorijakaumaan Studentin t-jakautuneen termin, $T(4, 0, \sigma)$, jonka vapausaste on 4, ja sen odotusarvo on 0. Hajontaparametrin σ valitsin, niin että regressiokertoimeen lisättiin 10-kertainen määrä epävarmuutta (hajontaa). Näin näiden maassaoloajan regressiokertoimien hajonta 11-kertaistui pelkästä aineistosta estimoituun verrattuna. Tämä vastaa päätöksentekoteoriaa koskevan luvun kuvan 12.3.3 (s. 518) osakuvan II esittämää esimerkkiä.

Lisäepävarmuus ei muuta ennusteiden odotusarvoja, mutta niiden keskihajonnat kasvavat. Alla taulukoissa 13.43 ja 13.44 on esitetty elinkaariennusteiden jakaumatulosten vertailu Somalian ja Irakin osalta niin, että diskontatuissa elinkaariennusteissa ei ole lisäepävarmuuksia (vastaavat edellä esitettyjä tuloksia) sekä niin, että elinkaariennusteen osittain muodostavaan työikäisten nettovaikutusennusteen maassaoloajan regressiokertoimeen on lisätty yllä esitetyn tavoin ”paljon” lisäepävarmuutta. Kuten taulukoiden tulokset osoittavat, elinkaariennusteiden odotusarvot eivät muutu (pienet erot johtuvat simulaation tuottamasta virheestä), epävarmuuden lisäämisestä maassaoloajan regressiokertoimeen. Sen sijaan kummankin elinkaariennusteen osalla niiden keskihajonta kasvaa: Somalian osalta se kasvaa 665 tuhannesta 746 tuhanteen, ja Irakin osalta se kasvaa 503 tuhannesta 592 tuhanteen.

Tämä esimerkki siitä, että elinkaariennusteeseen sisältyvä epävarmuus kasvaa, mutta ennusteen odotusarvo ei silti muutu, vastaa päätöksentekoteoriaa koskevan luvun kuvan 12.2 (s. 517) osakuvan A antamaa esimerkkiä.

13.2.4.1 Somalia

Taulukko 13.43: *Somaliassa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman lisäepävarmuutta	Lisäepävarmuus mukana
Ennusteväestö (N)	490	490
Odotusarvo	-641	-640
Keskihajonta	666	715
MC-keskivirhe	13	18
2,5:s prosenttipiste (2,5 %)	-2 147	-2 271
10. prosenttipiste (10 %)	-1 397	-1 438
Alakvartiili (25 %)	-937	-947
Mediaani (50 %)	-560	-554
Yläkvartiili (75 %)	-264	-247
90. prosenttipiste (90 %)	3	44
97,5:s prosenttipiste (97,5 %)	479	582

13.2.4.2 Irak

Taulukko 13.44: *Irakissa syntyneen ennusteväestön diskontatut elinkaariennusteet. Ennustetulokset ilman lasten vaikutuksia ("Ilman LV") sekä lasten yhteenlasketut ennustetut elinkaarivaikutukset mukaan lukien ("ml. LV"). Ennustejakauman henkilömäärä ennusteväestössä (N), ennusteiden odotusarvon (OA) ja keskihajonnan (KH) estimaatit, ennusteiden odotusarvoa koskeva Monte Carlo -keskivirhe sekä prosenttipisteiden estimaatit:* $2,5; 10, 25, 50, 75, 90; 97,5$. *Euromääräiset tunnusluvut ovat tuhansia euroja per henkilö.*

	Ilman lisäepävarmuutta	Lisäepävarmuus mukana
Ennusteväestö (N)	677	677
Odotusarvo	-542	-538
Keskihajonta	504	592
MC-keskivirhe	10	19
2,5:s prosenttipiste (2,5 %)	-1 678	-1 889
10. prosenttipiste (10 %)	-1 111	-1 197
Alakvartiili (25 %)	-770	-794
Mediaani (50 %)	-486	-473
Yläkvartiili (75 %)	-248	-211
90. prosenttipiste (90 %)	-40	41
97,5:s prosenttipiste (97,5 %)	290	473

13.3 Johtopäätökset: Irakissa ja Somaliassa syntyneiden potentiaalisten uusien maahanmuuttajien valinta

13.3.1 Johtopäätökset diskontattujen elinkaariennusteiden odotusarvojen perusteella: Irakissa ja Somaliassa syntyneitä ei kannata valita muuttavaksi Suomeen

Julkisen talouden hyödyn maksimoiva kriteeri potentiaalisen uuden maahanmuuttajan valinnasta Suomeen. Päätöksentekoteoriaa ja potentiaalisten uusien maahanmuuttajien valintaa käsittelevässä luvussa **??** määritettiin, että potentiaalinen uusi maahanmuuttaja kannattaa valita muuttavaksi Suomeen, jos hänen nykyarvoon diskontatun elinkaariennusteensa jakauman odotusarvo on positiivinen. Edelleen määritettiin, että jos tämä euromääräinen odotusarvo on negatiivinen, niin potentiaalista uutta maahanmuuttajaa ei kannata valita muuttavaksi Suomeen. Tämä valintasääntö pätee silloin, kun valinnoilla halutaan maksimoida Suomen julkisen talouden odotettavissa oleva euromääräinen hyöty uusien maahanmuuttajien valinnoista.

Irakissa ja Somaliassa syntyneitä potentiaalisia uusia maahanmuuttajia ei kannata valita muuttavaksi Suomeen. Irakissa ja Somaliassa syntyneiden uusien maahanmuuttajien vuoteen 2015 diskontatut elinkaariennusteet on esitetty yllä luvussa 13.2. Irakissa syntyneeseen ennustepopulaation kuuluvien henkilöiden diskontattujen elinkaariennusteiden yhteisjakauman odotusarvo on euromäärältään negatiivinen kaikilla esitetyillä suomessaoloajan vaihtoehdoilla. Sama koskee myös Somaliassa syntyneiden ennustepopulaatiota. Tämä tarkoittaa sitä, että Irakissa ja Somaliassa syntyneitä potentiaalisia uusia maahanmuuttajia *ei kannata valita* Suomeen muuttaviksi – kun potentiaalisia uusia maahanmuuttajia tarkastellaan ilman erittelyjä esimerkiksi sukupuolen ja maahanmuuttoiän mukaan.

Valintoja koskeva johtopäätös eritellysti esimerkiksi sukupuolen mukaan: Irakissa ja Somaliassa syntyneitä potentiaalisia uusia maahanmuuttajia ei kannata valita muuttavaksi Suomeen. Irakissa ja Somaliassa syntyneiden uusien maahanmuuttajien vuoteen 2015 diskontatut elinkaariennusteet on esitetty yllä luvussa 13.2. Kun diskontattuja elinkaariennusteiden jakaumia tarkastellaan eritellysti sukupuolen, maahanmuuttoiän, maahanmuuttovuoden pääasiallisen toiminnan sekä maahanmuuttovuoden koulutustason mukaan, valintatulokset eivät muutu. Mainittujen uusia maahanmuuttajia koskevien ominaisuuksien mukaan eritellyistä diskontatuista elinkaariennusteista ei löydy yhtäkään sellaista ryhmää/osajoukkoa, jonka diskontattujen elinkaariennusteiden odotusarvot olisivat positiiviset. Tämä tarkoittaa sitä, että Suomeen muuttavaksi *ei kannata valita* niin miehiä kuin naisiakaan, nuorena tai vanhana muuttavia, maahanmuuttovuonna töissä tai opiskelemassa olevia, tai niitä, joilla on muuttaessa 2. asteen koulutus tai niitä, joilla 2. asteen koulutusta ei ole.

13.3.2 Johtopäätökset diskontattujen, lisäepävarmuutta sisältävien elinkaariennusteiden odotusarvojen perusteella: Irakissa ja Somaliassa syntyneitä ei kannata valita muuttavaksi Suomeen

Yllä luvussa 13.2.4 esitettiin esimerkit vuoteen 2015 diskontatuista elinkaariennusteista, joihin on lisätty epävarmuuksia (nettovaikutusennustemallin maassaoloajan regressiokertoimen kautta). Potentiaalisia uusia maahanmuuttajia koskevat valintatulokset näistä diskontatuista elinkaariennusteista, joihin on lisätty ennuste-epävarmuutta, ovat samanlaiset kuin yllä. Vaikka ennuste-epävarmuuksien lisääminen kasvattaa diskontattujen elinkaariennusteiden jakaumien hajontaa, jakaumien odotusarvot eivät kuitenkaan muutu.

13.3.3 Havaintoaineiston ulkopuolinen informaatio, mitä elinkaariennusteen tekijällä pitäisi olla, jotta valintapäätökset muuttuisivat.

Havaintoaineistoon perustuvien diskontattujen elinkaariennusteiden jakaumien odotusarvojen perusteella Somaliassa ja Irakissa syntyneitä potentiaalisia uusia maahanmuuttajia ei kannata valittavaksi Suomeen. Samaan valintapäätökseen päädytään, kun elinkaariennusteisiin lisätään (tulevaisuuden tuottamaa) lisäepävarmuutta siten, että elinkaariennusteen laatijalla ei ole sellaista informaatiota, muuttuvatko elinkaariennusteet Irakissa ja Somaliassa syntyneiden osalta todennäköisemmin positiivisemmiksi vai negatiivisemmiksi kuin havaintoaineiston perusteella estimoidut elinkaariennusteet.

Päätöksentekotekoteoriaa koskevassa luvussa (12.3.1) on esitetty määritelmä sille, milloin aineiston ulkopuolisten tekijöiden vaikutukset muuttavat valintapäätöksen koskien potentiaalisia uusia maahanmuuttajia. Tämän määritelmän mukaan julkisen talouden hyödyn maksimoiva valintapäätös muuttuu aineiston ulkopuolisten tekijöiden vaikutuksesta vain silloin, kun nämä aineiston ulkopuoliset tekijät muuttavat uuden maahanmuuttajan diskontatun elinkaariennusteen etumerkin (ks. lauseke 12.11, s. 508).

Toisin sanoen julkisen talouden hyödyn maksimoiva valintapäätös Irakissa ja Somaliassa syntyneitä koskien muuttuu vain silloin, jos elinkaariennusteen laatijalla on sellaista tutkimusaineiston ulkopuolista informaatiota, joka kertoo, että diskontattujen elinkaariennusteiden odotusarvot muuttuvat negatiivista positiivisiksi. Pelkkä epävarmuuden lisääminen elinkaariennusteisiin ei tee tätä.

Esimerkkinä Somalia: Havaintoaineiston ulkopuolinen informaatio, mitä elinkaariennusteen tekijällä pitäisi olla, jotta valintapäätökset muuttuisivat. Esimerkiksi Somaliassa syntyneillä uusilla maahanmuuttajille diskontatun elinkaariennusteen odotusarvo on -527 000e (ilman lasten vaikutusta) ja -774 000e (ml. lasten vaikutukset), jos suomessaoloaika noudattaa vuosina 2000–2009 Suomeen muuttaneiden maassaoloa (taulukon 13.6 tulosten mukaan). Tässä tapauksessa elinkaariennusteen laatijalla pitäisi olla sellaista havaintoaineiston ulkopuolista informaatiota, joka siirtäisi diskontattujen elinkaariennusteiden jakaumia positiivisemmiksi vähintään 527 000 euroa (ilman lasten vaikutuksia) tai vähintään 774 000 euroa (jos lasten vaikutukset otetaan mukaan). Havaintoaineiston ulkopuolisten tekijöiden vaikutusta elinkaariennusteiden jakaumiin ja niiden mahdollisia vaikutuksia uusia maahanmuuttajia koskeviin valintapäätöksiin käsitellään tarkemmin luvussa 12.3.3 (ks. erityisesti kuvan 12.2 esimerkit s. 517).

Tämän tutkimuksen tekijällä ei sellaista havaintoaineiston ulkopuolista informaatiota, joka siirtäisi diskontattujen elinkaariennusteiden odotusarvot negatiivisista positiivisiksi. Koska tämän tutkimuksen tekijällä ei ole mainitunlaista aineiston ulkopuolista informaatiota, joka siirtäisi diskontattujen elinkaariennusteiden jakaumia positiivisemmiksi niin, että niiden odotusarvot muuttuisivat negatiivisista positiivisiksi, Irakissa ja Somaliassa syntyneitä potentiaalisia uusia maahanmuuttajia koskevat johtopäätökset pysyvät esitettyinä. Irakissa ja Somaliassa syntyneitä potentiaalisia uusia maahanmuuttajia ei kannata valita muuttavaksi Suomeen, mikäli näillä valintapäätöksillä halutaan maksimoida julkisen talouden euromääräinen hyöty.

13.4 Katsaus pohjoismaisiin tutkimuksiin, joissa tulokset on eritelty pakolaistaustaisuuden mukaan

Esitän tämän elinkaarivaikutuksia käsittelevän tutkimuksen lopuksi vielä tutkimustuloksia muista pohjoismaalaisista tutkimuksista. Tutkimustulokset, joita esitän, käsittelevät pakolaistaustaisten maahanmuuttajien maassaoloaikoja, työllistymisten, työtulojen ja saatujen sosiaalietuuksien kehitystä maassaoloajan mukaan, maahanmuuttajien lasten koulutusta sekä julkisen talouden toteutuneita nettovaikutuksia että elinkaarivaikutuksia. Pakolaistaustaisuus tarkoittaa sitä, että tutkittavat maahanmuuttajat ovat muuttaneet Pohjoismaihin joko pakolaisina tai pakolaisina tulleiden perheenjäseninä. Pakolaisena muuttanut tarkoittaa puolestaan sitä, että on muuttanut Pohjoismaihin alun perin joko turvapaikanhakijana tai kiintiöpakolaisjärjestelmän kautta.

Vertaan maassaoloaikoja koskevia tuloksia tämän tutkimuksen luvussa 3.2.7 esitettyihin havainto- ja mallinnustuloksiin. Samaten vertaan muiden tutkimusten työllisyyttä, työtuloja ja saatuja sosiaalietuuksia koskevia tuloksia tutkimuksen luvussa 5.4 esitettyihin havainto- ja mallinnustuloksiin koskien palkkasummia. Maahanmuuttajien lasten koulutusta koskevia tuloksia vertaan tämän tutkimuksen luvussa 8.3.4 esitettyihin tietoihin koskien 22-vuotiaita maahanmuuttajien lapsia.

Julkisen talouden nettovaikutusten osalta ne tulokset muista tutkimuksista, jotka koskevat toteutuneita julkisen talouden vaikutuksia, vertautuvat tämän tutkimuksen I osan tuloksiin (ks. Salminen 2015a, luku 6.9) sekä tämän tutkimuksen luvussa 4.4 esitettyihin havainto- ja mallinnustuloksiin koskien nuorten ja työikäisten nettovaikutuksia. Ne tulokset muista tutkimuksista, jotka käsittelevät elinkaarivaikutuksia, vertautuvat puolestaan tämän tutkimuksen elinkaariennusteita koskeviin tuloksiin (diskonttaamattomat tulokset on esitetty luvuissa 9.6 ja 9.5 sekä nykyarvoon diskontatut tulokset puolestaan luvussa 13.2).

Pohjoismaiden ministerineuvoston laaja tutkimusartikkelikokoelma Pohjoismaista (2017) . Suurin osa alla esitettävistä muiden tutkimuksen tuloksista on peräisin Pohjoismaiden ministerineuvoston vuonna 2017 julkaisemasta tutkimusartikkelikokoelmasta: *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries* (Andersen et al. 2017). Artikkelikokoelma sisältää Norjan, Tanskan ja Suomen osalta tutkimusartikkeleita siitä, kuinka humanitaarisen maahanmuuton kautta (ml. perheenyhdistämiset) näihin maihin muuttaneet ovat työllistyneet ja miten heidän työtulonsa ja saamansa sosiaaliedut ovat kehittyneet maassaoloajan myötä verrattuna samanikäiseen kantaväestöön.

Artikkelikokoelma sisältäisi myös Ruotsia koskien tutkimusartikkelin (Åslund et al. 2017), jossa käsitellään maahanmuuttajien työllistymistä Ruotsiin muuton jälkeen – mutta toisin kuin Tanskaa, Norjaa ja Suomea koskevissa artikkeleissa, siinä ei esitetä suoraan eri maahanmuuttajaryhmille työllisyys-, työtulo-, ja sosiaalietuustietoja verrattuna samanikäiseen kantaväestöön. Vaikka näin ollen alla ei esitetä Ruotsia koskien pakolaisten työllisyys-, työtulo-, ja sosiaalietuuslukuja Ruotsia koskien, toisaalta Ruotsia koskien esitetään pakolaisia koskevia toteutuneita julkisen talouden vaikutuksia viitaten artikkeliin Ruist (2017).

Maahanmuuttajien työllisyyteen ja julkisen talouden vaikutuksiin liittyen on olemassa lukuisia muitakin tutkimuksia kuin artikkelikokoelmaan (Andersen et al. 2017) kuuluvat tutkimukset (ks. esim. tämän tutkimuksen I osan (Salminen 2015a) kirjallisuusluettelo). Valitsin kuitenkin Andersen et al. 2017:n

artikkelikokoelman tutkimusten tulokset vertailukohdaksi kolmesta syystä:

1. **Maahanmuuttajien erittely**. Humanitaarisen maahanmuuton (ml. perheenyhdistämiset) kautta Pohjoismaihin muuttaneet eritelty muista maahanmuuttajista artikkelin tutkimuksissa. Osassa tutkimuksista on lisäksi tehty erittely maahanmuuttajien syntymävaltioiden mukaan.
2. Erityisesti työllisyyttä ja työtuloja koskevat **tutkimusartikkelit perustuvat henkilömääriltään suuriin ja ajassa pitkäkestoisiin rekisteritietoihin**. Pitkäkestoisuudella viittaan siihen, että rekisteritiedot ylettyvät tutkimusartikkeleittain 1990-luvulta 2010-luvulle. Näin ollen tutkimusaineistot artikkeleissa ovat samanlaisia kuin tässä Suomen Perustan tutkimuksessa.
3. **Tutkimusartikkelit käsittelevät Suomea ja muita Pohjoismaita** (pl. Islanti). Pidän muita Pohjoismaita lähimpinä vertailumaina Suomelle.

13.4.1 Maassaoloaika

Yhteenveto Norjasta ja Tanskasta: Yli 90 %:ia pakolaisina tulleista maassa viisi vuotta maahanmuuton jälkeen.

13.4.1.1 Norja

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Norjaan sijoittuvan artikkelin: *Immigrant labor market integration across admission classes* (Bratsberg et al. 2017). Tässä tutkimusartikkelissa käsitellään pääasiallisesti Norjaan vuosina 1990–2013 saapuneiden maahanmuuttajien työllisyyttä, mutta se sisältää myös norjalaisiin viranomaisrekistereihin perustuvia tietoja maassaoloajoista. Tutkimuksessa maahanmuuttajat on jaettu kuuteen eri ryhmään, ja tutkimuksen tulokset on esitetty eritellysti näitä ryhmiä koskien. Näistä kuudesta ryhmästä kaksi koskee EU:sta ja OECD-maista tulleita maahanmuuttajia, ja neljä muuta luokkaa koskevat EU:n ja OECD-maiden ulkopuolelta tulleita maahanmuuttajia. EU- ja OECD-maiden ulkopuolelta tulevat on jaettu neljän ryhmään maahanmuuton syyn perusteella. Näiden luokkien nimet esiintyvät englanninkielisinä alla kuvassa 13.10, joka käsittelee maassaoloaikoja. Näiden kuuden englanninkielisen luokan nimet ovat suomeksi:

- EU- ja OECD-maiden ulkopuolelta muuttaneet:
  - *A. Refugee*. Humanitaarinen maahanmuutto/suojelutarpeen myöntäminen eli pakolaisina tai turvapaikanhakijoina Norjaan muuttaneet: *"[...] those admitted for protection (e.g., given refugee or asylum status)"* (Bratsberg et al. 2017, s. 19). Henkilölukumääriltään suurimmat lähtömaat aineistossa humanitaarisen maahanmuuton osalta ovat Bosnia, Irak ja Somalia (Bratsberg et al. 2017, s. 40), joten tilanne on samanlainen Suomen kanssa ainakin Irakin ja Somalian osalta (Bosniassa syntyneiden tarkkaa määrää on vaikeaa arvioida omassa tutkimusaineistossa, sillä ainakin osalla Bosniassa syntyneistä syntymävaltiona Tilastokeskuksen rekistereissä on Entinen Jugoslavia).

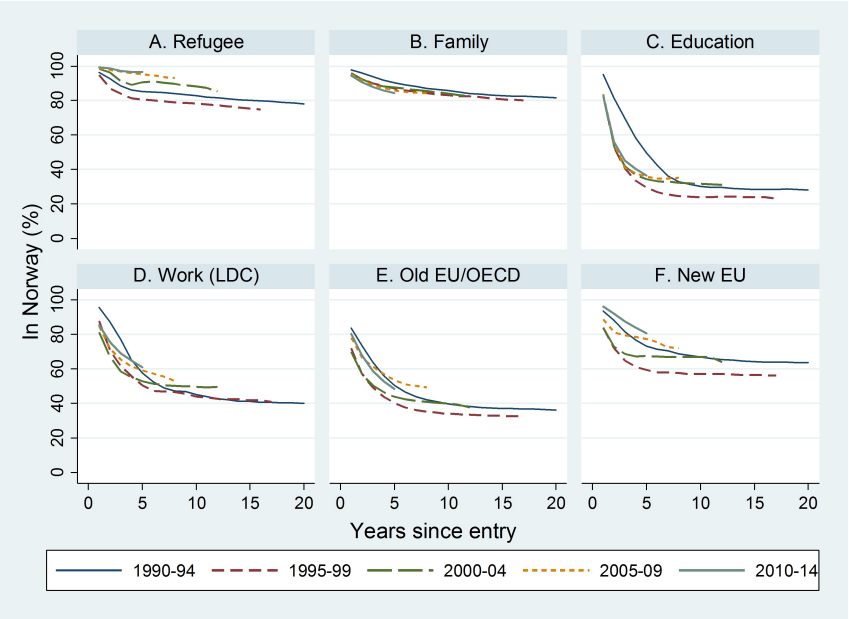
  - *B. Family*. Perheenyhdistämisten kautta Norjaan EU- ja OECD-maiden ulkopuolelta muuttaneet. Luokka sisältää mahdollisesti myös Norjassa syntyneelle kantaväestölle yhdistetyt perheenjäsenet (vrt. eri selitteet luokalle s. 19 ja 25).
  - *C. Education*. Opiskelu.
  - *D. Work (LDC)*. Työ. ”LDC” on ilmeisesti lyhennys sanoista ”Less Developed Countries” eli vähemmän kehittyneet maat.

- EU- ja OECD-maiden ulkopuolelta muuttaneet:
  - *E. Old EU/OECD*. Kaikki muut EU- ja OECD-maat paitsi ”uusi EU”.
  - *F. New EU*. Uusi EU eli EU-maat, jotka liittynet EU:hun vuosina 2004 ja 2007.

Tutkimus (Bratsberg et al. 2017) sisältää viranomaisrekistereistä peräisin olevia tietoja koskien kaikkia maahanmuuttajia, jotka olivat 18–47-vuotiaita, kun he muuttivat Norjaan vuosina 1990–2013. Tutkimuksessa esitetyt tulokset koskevat 25–62-vuotiaita, jotka eivät olleet opiskelijoita ja jotka olivat kunkin seurantavuoden lopussa Norjassa. Seuranta-aika tutkimuksessa käsitti vuodet 1993–2014. Maahanmuuttajien lisäksi tutkimukseen sisältyy 10 %:n satunnaisotos Norjassa syntyneestä kantaväestöstä. Bratsberg et al. 2017:n tutkimusaineisto on rakenteeltaan täten hyvin samankaltainen pitkittäis-/paneeliaineisto, jota käytän itse Suomea koskien tässä tutkimuksessa. (Bratsberg et al. 2017, s. 19, 25, 27).

Kuvassa 13.10 on esitetty Norjassa olevien maahanmuuttajien osuudet (%) maassaoloajan ja maahanmuuttovuoden luokan mukaan maahanmuuttajaryhmittäin. Maahanmuuttajaryhmiä on kuusi kappaletta, ja ryhmät on esitelty yllä. Kuvasta selviää, että humanitaarisina maahanmuuttajina Norjaan vuosina 1990–1994 muuttaneista (niistä, jotka sisältyvät tutkimusaineistoon) noin 80 %:ia oli vielä Norjassa 20 maassaolovuoden jälkeen. Vuosina 1995–1999 muuttaneista Norjaan on jäänyt kuvan 13.10 perusteella joitakin prosenttiyksikköjä vähemmän kuin aiemmin muuttaneista, kun maassaoloaika huomioidaan. Norjankin osalta – samoin kuin tässä tutkimuksessa on osoitettu Suomen osalta – 2000-luvulla muuttaneista humanitaarisista maahanmuuttajista suurempi osa on jäänyt Norjaan kuin aiemmin muuttaneista, kun maassaoloaika huomioidaan. Esimerkiksi 2000–2004 muuttaneista noin 90 %:ia oli Norjassa vielä 10 maassaolovuoden jälkeen.

Perheenyhdistämisten perusteella Norjaan muuttaneista osa on sellaisia henkilöitä, jotka on yhdistetty humanitaarisen maahanmuuton vuoksi Norjaan muuttaneille henkilöille. Perheisiin yhdistettyjen osalta kuvan 13.10 perusteella vaikuttaa siltä, että maahanmuuttovuodella ei ole juurikaan merkitystä Norjaan jäämiselle. Kuvan 13.10 perusteella vaikuttaa lisäksi siltä, että noin 80 %:ia perheisiin yhdistetyistä jää pysyvästi Norjaan.



Kuva 13.10: *Norjassa olevien osuudet (%) maassaoloajan ja maahanmuuttovuoden luokan mukaan maahanmuuttajaryhmittäin. Maahanmuuttajaryhmät on esitetty tekstissä yllä. Vaaka-akselilla on vuosien määrä maahanmuutosta. Eri kuvaajat vastaavat esitettyä viittä eri maahanmuuttovuoden luokkaa. Lähde: Bratsberg et al. 2017, s. 23.*

Kuvan 13.10 tuloksia tulkitessa on syytä pitää mielessä, että kuvassa ei ole esimerkiksi huomioitu humanitaaristen maahanmuuttajien lähtömaiden koostumuksessa tapahtuneita muutoksia (Balkanilta 1990-luvun alussa ja Irakista ja Somaliasta puolestaan 1990-luvun lopussa (Bratsberg et al. 2017, s. 27)). Muutokset maahan jäämisessä maahanmuuttovuoden luokan mukaan voivat siten johtua eroista eri lähtömaista tulevien jäämisessä Norjaan. Tämän tutkimuksen suomessaoloaikoja käsittelevässä luvussa olen osoittanut tulosten perusteella, että Suomeen jäämisen todennäköisyys vaihtelee kymmeniä prosenttiyksiköitä eri syntymävalitioista kotoisin olevien maahanmuuttajien välillä (ks. luvussa 3.2.7 esitetyt tulokset).

13.4.1.2 Tanska

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Tanskaan sijoittuvan artikkelin: *Labour market integration of refugees in Denmark* (Schultz-Nielsen 2017). Tässä tutkimusartikkelissa käsitellään pääasiallisesti Tanskaan vuosina 1997–2011 ensimmäistä kertaa muuttaneiden humanitaaristen maahanmuuttajien (artikkelissa ”pakolaiset”, *refugees*) työllisyyttä, mutta se sisältää myös tanskalaisiin viranomaisrekistereihin perustuvia tietoja maassaoloajoista. Tutkimuksessa humanitaarisia maahanmuuttajia koskevat tulokset eritellään sukupuolen mukaan sekä sen mukaan, tuliko maahanmuuttaja Tanskaan perheenyhdistämisen kautta vai ei. Humanitaarisiin maahanmuuttajiin lasketaan tässä tutkimuksessa siis myös humanitaaristen maahanmuuttajien Tanskaan yhdistämät perheenjäsenet (Andersen et al. 2017, s.

55).

Tutkimusaineisto Schultz-Nielsen 2017:n artikkelissa koostuu vuosina 1997–2011 Tanskaan muuttaneista humanitaarisista maahanmuuttajista (ml. perheenyhdistämiset). Tutkimuksen seuranta-aika kattaa vuodet 1997–2013. Tutkimusaineisto, joka on pitkittäis-/paneeliaineisto, muistuttaa muodoltaan oman tutkimukseni käytössä olevaan aineistoa. Schultz-Nielsen 2017:n tutkimusaineistoon sisältyvät vain ne henkilöt, jotka olivat muuttovuonna 17–36-vuotiaita. Lisärajoituksena on, että tulokset koskevat vain 25-vuotiaita ja vanhempia. Tutkimusaineisto sisältää vertailun vuoksi myös 10 %:n satunnaisotoksen Tanskassa vuosina 1961–1987 syntyneistä henkilöistä. (Schultz-Nielsen 2017, s. 66). Humanitaaristen maahanmuuttajien suurimpien lähtömaiden joukko tutkimusaineistossa sisältää mm. Irakin, Somalian, Afganistanin ja Iranin (ks. tarkemmin Schultz-Nielsen 2017, s. 86).

Taulukossa 13.45 on esitetty vähintään viisi vuotta Tanskassa olleiden osuudet (%) väestöittäin edellä esitetyssä tutkimusaineistossa. Tanskassa syntyneen kantaväestön osalta Tanskassa on pysynyt viiden vuoden seurantajakson jälkeen 98 %:ia väestöstä (ei taulukossa). Kaikista pakolaismiehistä (humanitaariset maahanmuuttajamiehet) Tanskassa on viiden vuoden jälkeen maahanmuutosta 95 %:ia. Pakolaisille perheenyhdistetyistä miehistä 91 %:ia oli Tanskassa vielä viisi vuotta maahanmuuton jälkeen. Naisten osalta vastaavat prosentit olivat 94 %:ia ja 96 %:ia. Tulokset muistuttavat täten yllä esitettyjä tuloksia Norjaa koskien sekä tässä tutkimuksessa esitettyjä tuloksia Suomeen 2000-luvulla muuttaneita Irakissa ja Somaliassa syntyneitä koskien. Näiden Norjaa, Tanskaa ja Suomea koskevien tulosten mukaan noin 90 %:ia 2000-luvulla muuttaneista humanitaarisista maahanmuuttajista jää Pohjoismaissa pysyvästi siihen maahan, johon he muuttavat.

Taulukko 13.45: *Vähintään viisi vuotta maassa olleiden osuudet (%) väestöittäin. "Pakolaiset" käsittävät kaikki humanitaarisen maahanmuuton kautta Tanskaan muuttaneet, ml. perheenyhdistämisten kautta tulleet. "Perheenyhdistetty muulle" käsittää perheenyhdistämiset, jotka eivät liity humanitaariseen maahanmuuttoon. Lähde: Schultz-Nielsen 2017, s. 67.*

	Pakolaiset	Perheenyhdistetty pakolaiselle	Perheenyhdistetty muulle
Miehet	95	91	85
Naiset	94	96	83

13.4.2 Työllisyys, työtulot ja saadut tulonsiirrot

Karkea yhteenveto Suomesta, Norjasta ja Tanskasta: Työllisyysasteet ja työtulot nousevat pakolaistaustaisilla verrattuna samanikäiseen kantaväestöön noin 8–10 maassaolovuoteen asti. Tämän jälkeen työllisyysteiden ja työtulojen erojen pieneneminen kantaväestöön joko pysähtyy (Suomi) tai alkaa kasvamaan uudestaan (Norja ja Tanska). Näin ollen työllisyys- ja työtuloerot kantaväestöön verrattuna jäävät "suuriksi".

Työllisyysastetta koskevia tuloksia maassaoloajan suhteen voi verrata tämän tutkimuksen palkkasummamallinnuksen yhteydessä esitettyihin kuviin nollapalkkasummien kehityksestä maassaolojana suhteen:

ks. Somalia (kuva 5.8a, s. 286) ja ks. Irak (kuva 5.18a, s. 309).

Saatujen tulonsiirtojen määrät verrattuna samanikäiseen kantaväestöön maassaoloajan suhteen ovat käänteisiä verrattuna työllisyysasteen ja työtulojen kehitykseen: Saatujen tulonsiirtojen määrä pienenee verrattuna samanikäiseen kantaväestöön noin 8–10 maassaolovuoteen asti. Tämän jälkeen saadut tulonsiirrot eivät enää pienene verrattuna samanikäiseen kantaväestöön, ja ne jäävät keskimäärin ”paljon” korkeammalle tasolle kuin samanikäisen kantaväestön saamat tulonsiirrot keskimäärin.

13.4.2.1 Suomi

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Suomeen sijoittuvan artikkelin: *Labor market integration of refugees in Finland* (Sarvimäki 2017a). Tässä tutkimusartikkelissa käsitellään humanitaaristen maahanmuuttajien (artikkelissa ”pakolaiset”, *refugees*) työllisyyttä, työtuloja ja saatuja etuisuuksia verrattuna Suomessa syntyneeseen kantaväestöön. Humanitaariset maahanmuuttajat on määritetty artikkelissa maahanmuuttajien syntymävaltion mukaan. Täten myös humanitaarisille maahanmuuttajille yhdistetyt perheenjäsenet ovat artikkelissa humanitaarisia maahanmuuttajia. Tutkimusaineisto koostuu vuosittaisista havainnoista, jotka koskevat kaikkia 25–60-vuotiaita maahanmuuttajia, jotka olivat muuttaneet Suomeen 18-vuotiaana tai vanhempana. Havainnot koskevat vuosia 1988–2013, eli tämä on tutkimuksen seurantajakso. Tutkimuksen Suomessa syntynyttä kantaväestöä koskevat tulokset perustuvat 10 %:n satunnaisotokseen koko väestöstä. Tutkimusaineisto on toisin sanoen paneeli-/pitkittäisaineisto, ja se muistuttaa tässä tutkimuksessa käyttämääni aineistoa. (Sarvimäki 2017a, s. 99).

Työllisyyserot verrattuna kantaväestöön

Taulukossa 13.46 on esitetty eri maahanmuuttajaryhmien työllisyysasteiden ero (%-yksikköä) verrattuna Suomessa syntyneeseen kantaväestöön. Työllisyyserot on esitetty sukupuolen sekä maassaoloajan mukaan. Maassaoloajan luokkia ovat: 1, 5, 10 ja 15 (vuotta Suomessa). Sarvimäki (2017a) on muodostanut työllisyyserot niin, että luvuissa huomioidaan (”vakioidaan”) iän ja sukupuolen erot maahanmuuttajaväestöissä verrattuna kantaväestöön. Työllisyyseroissa on huomioitu myös työllisyyshavaintojen kalenterivuosi. (ks. tarkempi selitys työllisyyserojen muodostamisesta: Sarvimäki 2017a, s. 104, taulukon 2 selite).

Taulukko 13.46: *Ennustetut erot (%-yksikköä) työllisyydessä maassaoloajan ja sukupuolen mukaan. Vaikutukset verrattuna Suomessa syntyneeseen kantaväestöön. Estimaattien keskivirheet suluissa. Lähde: Sarvimäki 2017a, s. 104.*

	Miehet				Naiset			
Vuosia maahanmuutosta:	1	5	10	15	1	5	10	15
Irak (est.)	-69	-53	-48	-50	-74	-68	-64	-63
Irak (kv)	(1)	(1)	(1)	(2)	(<1)	(1)	(1)	(2)
Somalia (est.)	-69	-55	-48	-48	-72	-68	-68	-66
Somalia (kv)	(1)	(1)	(2)	(2)	(<1)	(1)	(1)	(1)
Afganistan (est.)	-69	-44	-37	–	-77	-69	-58	–
Afganistan (kv)	(1)	(2)	(3)	–	(<1)	(1)	(2)	–
OECD (est.)	-19	-20	-14	-11	-31	-31	-26	-22
OECD (kv)	(<1)	(<1)	(1)	(1)	(<1)	(1)	(1)	(1)
Kaikki maahanmuuttajat (est.)	-26	-24	-22	-20	-45	-36	-28	-23
Kaikki maahanmuuttajat (kv)	(<1)	(<1)	(<1)	(<1)	(<1)	(<1)	(<1)	(<1)

Taulukon 13.46 tulosten mukaan Irakissa syntyneillä miehillä työllisyys on yhden vuoden maassaolon jälkeen 69 %-yksikköä heikompi kuin vastaavan ikäisillä Suomessa syntyneen kantaväestön miehillä. Tämä ero pienenee 10 vuoden maassaoloajan jälkeen 48 %-yksikköön, mutta on vielä 15 vuoden maassaoloajan jälkeen 50 %-yksikköä pienempi. Irakissa syntyneillä naisilla ero on ensimmäisen maassaolovuoden jälkeen 74 %-yksikköä, mistä se pienenee 10 ja 15 vuoden maassaoloajan jälkeen noin 63 %-yksikköön. Somaliassa syntyneitä koskevat tulokset ovat melko samanlaisia kuin Irakissa syntyneillä. Irakissa ja Somaliassa syntyneiden työllisyyden kehitys noudattaa suunnilleen samanlaista linjaa kuin Norjaa ja Tanskaa koskevat tulokset, jotka esitän myös tässä luvussa. Toisin sanoen työllisyysero kantaväestöön pienenee ensimmäisen noin 10 maassaolovuoden aikana, mutta sen jälkeen eron pieneneminen joko pysähtyy (taulukon 13.46 tulokset) tai ero alkaa kasvamaan uudestaan (Norjan ja Tanskan tulokset jäljempänä).

Työtulot verrattuna kantaväestöön

Työllisyyden jälkeen Sarvimäki (2017a) tarkastelee maahanmuuttajaryhmittäin työtulojen kehitystä maassaoloajan ja sukupuolen mukaan. Työtulot on määritetty artikkelissa palkka- ja yrittäjätulojen kokonaissummana (artikkelissa englanniksi: *[...]the sum of total wages, salary and entrepreneurial income[...]*). Työtulot sisältävät artikkelissa myös nollatulot. Työtulot on esitetty vuoden 2010 euroissa, johon työtulohavainnot on muunnettu Tilastokeskuksen rahanarvonkertoimella. Työtuloja koskevat estimaatit on muodostettu samalla tavalla kuin edellä työllisyyttä koskevat estimaatit, eli estimaateissa otetaan huomioon erot maahanmuuttajaväestön sukupuoli-, ikä- ja kalenterivuosijakaumissa verrattuna kantaväestöön. (s. 105).

Taulukossa 13.47 on esitetty eri maahanmuuttajaryhmien vuosittaisten työtulojen osuus (%) Suomessa

syntyneen kantaväestön työtuloista. Työtulojen osuudet on esitetty sukupuolen sekä maassaoloajan mukaan. Maassaoloajan luokkia ovat: 1, 5, 10 ja 15 (vuotta Suomessa). Sarvimäki 2017a on muodostanut työtulojen osuudet niin, että luvuissa huomioidaan (”vakioidaan”) iän ja sukupuolen erot maahanmuuttajaväestöissä verrattuna kantaväestöön. Työtulojen osuuksissa on huomioitu myös työllisyyshavaintojen kalenterivuosi. (ks. tarkempi selitys suhteellisten vuosittaisten tulojen muodostamisesta: Sarvimäki 2017a, s. 108, taulukon 3 selite).

Taulukko 13.47: *Suhteelliset vuosittaiset tulot (%) verrattuna Suomessa syntyneeseen kantaväestöön (kantaväestö = 100). Osuudet maassaoloajan ja sukupuolen mukaan. Osuuksien estimaattien keskivirheet suluissa. Lähde: Sarvimäki 2017a, s. 108.*

	Miehet				Naiset			
Vuosia maahanmuutosta:	1	5	10	15	1	5	10	15
Irak (est.)	4	15	22	21	1	7	13	15
Irak (kv)	(<1)	(1)	(1)	(2)	(<1)	(1)	(1)	(2)
Somalia (est.)	4	17	29	31	1	6	9	12
Somalia (kv)	(<1)	(1)	(1)	(2)	(<1)	(1)	(1)	(1)
Afganistan (est.)	4	25	38	–	1	6	23	–
Afganistan (kv)	(1)	(2)	(4)	–	(<1)	(1)	(3)	–
OECD (est.)	81	78	87	88	58	57	63	70
OECD (kv)	(1)	(1)	(1)	(2)	(<1)	(<1)	(1)	(1)
Kaikki maahanmuuttajat (est.)	52	57	60	60	31	41	52	59
Kaikki maahanmuuttajat (kv)	(<1)	(<1)	(<1)	(1)	(<1)	(<1)	(<1)	(1)

Taulukon 13.47 tulosten mukaan Irakissa syntyneillä miehillä työtulot ovat yhden vuoden maassaolon jälkeen 4 %:ia siitä, mitä vastaavan ikäisillä Suomessa syntyneen kantaväestön miehillä on keskimäärin. Irakissa syntyneiden miesten tulot kasvavat niin, että 10 vuoden maassaoloajan jälkeen heidän työtulonsa ovat 22 %:ia kantaväestön tuloista, mutta 15 vuoden maassaoloajan jälkeen työtulojen kasvu suhteessa kantaväestöön on pysähtynyt. Irakissa syntyneillä naisilla työtulot verrattuna kantaväestön naisiin on ensimmäisen maassaolovuoden jälkeen 1 %:n, mistä ne kasvavat 15 vuoden maassaoloajan jälkeen noin 15 %:iin. Somaliassa syntyneitä koskevat tulokset ovat melko samanlaisia kuin Irakissa syntyneillä. Sarvimäen mukaan kaikissa (tarkastelluissa) maahanmuuttajaryhmissä keskimääräiset työtulot ovat kantaväestöä pienemmät, ja erityisesti Afganistanista, Irakista ja Somaliasta kotoisin olevat pärjäävät erityisen huonosti: ”*All immigrant groups have lower average earnings than natives and those from Afghanistan, Iraq and Somalia fare particularly badly.*”. (Sarvimäki 2017a, s. 106).

Saadut etuudet verrattuna kantaväestöön

Työllisyyden ja työtulojen jälkeen Sarvimäki (2017a) tarkastelee maahanmuuttajaryhmittäin saatujen etuuksien kehitystä maassaoloajan ja sukupuolen mukaan. Etuudet on määritetty *ekvivalenssiskaalatusti*

(*engl. equivalence scaled*). Tämä tarkoittaa sitä, että jokainen kotitalouden jäsen saa painon siten, että ensimmäinen kotitalouden jäsen saa painon 1, muut yli 14-vuotiaat kotitalouden jäsenet saavat painot 0,7 ja alle 15-vuotiaat saavat painoksi 0,5. Ekvivalenssiskaalattu etuuksien määrä yhdessä kotitaloudessa on kotitalouden jäsenten saamien etuuksien summa jaettuna kotitalouden jäsenten painojen summalla. Jokainen kotitalouden jäsen saa siten etuuksiensa määräksi tämän ekvivalenssiskaalatun etuuksien summan. Sarvimäki (2017a) käyttää ekvivalenssiskaalattuja etuuksia, koska osa etuuksista on kohdistettu kotitalouksille yksittäisten henkilöiden sijaan. Etuuksia koskevat tulokset koskevat vuosia 1995–2013. (s. 108–109).

Taulukossa 13.48 on esitetty eri maahanmuuttajaryhmien vuosittaisten suhteellisten ekvivalenssiskaalattujen etuuksien osuudet (%). Saatuja etuuksia verrataan Suomessa syntyneen kantaväestön keskimääräisiin ekvivalenssiskaalattuihin etuuksiin. Saatujen etujen osuudet verrattuna kantaväestöön on esitetty sukupuolen sekä maassaoloajan mukaan. Maassaoloajan luokkia ovat: 1, 5, 10 ja 15 (vuotta Suomessa). Sarvimäki 2017a on muodostanut etujen osuudet niin, että luvuissa huomioidaan (”vakioidaan”) iän ja sukupuolen erot maahanmuuttajaväestöissä verrattuna kantaväestöön. Etujen osuuksissa on huomioitu myös etuushavaintojen kalenterivuosi. (ks. tarkempi selitys etuuksien osuuksien muodostamisesta: Sarvimäki 2017a, s. 111, taulukon 4 selite).

Taulukko 13.48: *Suhteelliset vuosittaiset ekvivalenssiskaalatut etuudet (%) verrattuna Suomessa syntyneeseen kantaväestöön (kantaväestö = 100). Osuudet maassaoloajan ja sukupuolen mukaan. Osuuksien estimaattien keskivirheet suluissa. Lähde: Sarvimäki 2017a, s. 111.*

	Miehet				Naiset			
Vuosia maahanmuutosta:	1	5	10	15	1	5	10	15
Irak (est.)	212	240	218	211	231	264	248	240
Irak (kv)	(2)	(3)	(4)	(4)	(2)	(3)	(3)	(6)
Somalia (est.)	210	196	193	172	215	238	228	203
Somalia (kv)	(2)	(2)	(3)	(3)	(3)	(3)	(3)	(4)
Afganistan (est.)	221	226	194	–	240	288	232	–
Afganistan (kv)	(4)	(5)	(7)	–	(3)	(4)	(7)	–
OECD (est.)	65	73	75	71	63	77	85	93
OECD (kv)	(1)	(1)	(1)	(2)	(1)	(1)	(2)	(4)
Kaikki maahanmuuttajat (est.)	95	110	116	114	125	139	139	132
Kaikki maahanmuuttajat (kv)	(<1)	(1)	(1)	(1)	(<1)	(1)	(1)	(1)

Taulukon 13.48 tulosten mukaan Irakissa syntyneillä miehillä suhteelliset etuudet ovat yhden vuoden maassaolon jälkeen 212 %:ia siitä, mitä vastaavan ikäisillä Suomessa syntyneen kantaväestön miehillä on keskimäärin. Irakissa syntyneiden miesten saamat etuudet vähenevät suhteessa kantaväestön miesten saamiin etuuksiin, että 10 vuoden maassaoloajan jälkeen heidän saamansa etuudet ovat 218 %:ia kantaväestön tuloista, mutta 15 vuoden maassaoloajan jälkeen saatujen etuuksien kasvu suhteessa kantaväestöön on lähestulkoon pysähtynyt. Irakissa syntyneillä naisilla saadut etuudet verrattuna kantaväestön

naisiin on ensimmäisen maassaolovuoden jälkeen 231 %:ia, mistä se laskee 15 vuoden maassaoloajan jälkeen noin 240 %:iin. Somaliassa syntyneitä koskevat tulokset ovat melko samanlaisia kuin Irakissa syntyneillä: tosin 15 vuoden maassaoloajan jälkeen Somaliassa syntyneet saavat vähemmän etuuksia kuin Irakissa syntyneet.

Sarvimäen johtopäätökset työllisyyttä, työtuloja ja saatuja etuuksia koskien

Sarvimäki (2017a) kirjoittaa artikkelinsa johtopäätöksissä edellä esitetyistä työllisyyttä, työtuloja ja saatuja etuuksia koskevista tuloksista: *"I find that earlier cohorts of immigrants from the main source countries of the 2015 asylum seekers – Iraq, Afghanistan and Somalia – had substantially lower employment rates and average earnings and collected more social benefits than natives. While the immigrant-native gap decreased over the first decade lived in Finland, it remained substantial. Furthermore, the differences in benefits remained roughly constant despite the increase in immigrants' employment rates and earnings."* (Sarvimäki 2017a, s. 112).

Tuloksia koskevien johtopäätösten mukaan Irakissa, Afganistanissa ja Somaliassa syntyneillä on tutkimusaineistossa merkittävästi alhaisempi työllisyysaste kuin kantaväestöllä. Työtulot olivat näillä maahanmuuttajilla puolestaan keskimäärin merkittävästi pienemmät kuin kantaväestöllä, ja saatuja työtuloja puolestaan enemmän kuin kantaväestöllä. Vaikka erot kantaväestöön pienenivät ensimmäisen 10 maassaolovuoden aikana, ne pysyivät vielä huomattavina. Saatujen etuuksien ero pysyi suunnilleen samansuuruisena maassaoloajan suhteen, vaikka työllisyys ja työtulot kasvoivatkin.

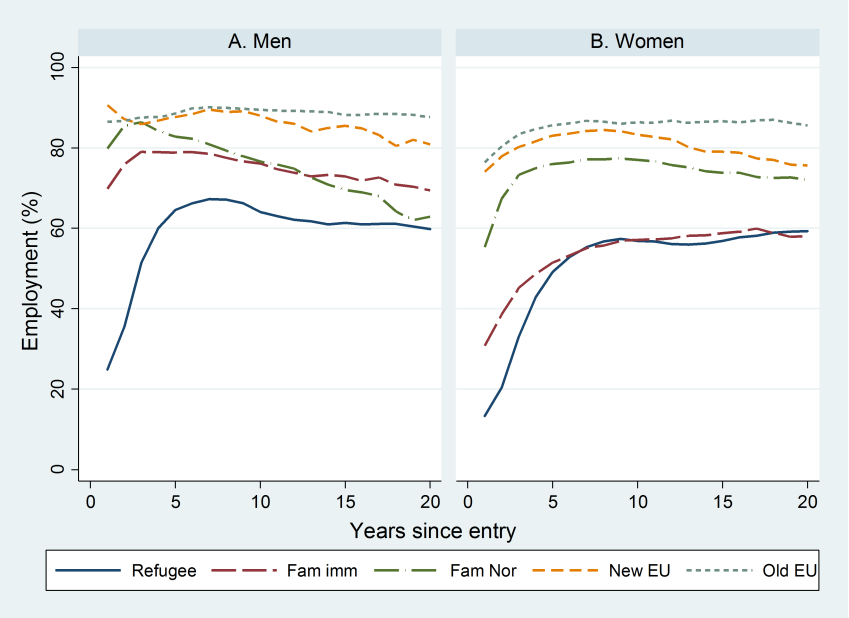
13.4.2.2 Norja

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Norjaan sijoittuvan artikkelin: *Immigrant labor market integration across admission classes* (Bratsberg et al. 2017). Tässä tutkimusartikkelissa käsitellään Norjaan vuosina 1990–2013 saapuneiden maahanmuuttajien työllisyyttä ja sosiaaliturvan käyttöä. Tutkimuksessa maahanmuuttajat on jaettu kuuteen eri ryhmään, ja tutkimuksen tulokset on esitetty eritellysti näitä ryhmiä koskien. Näistä kuudesta ryhmästä kaksi koskee EU:sta ja OECD-maista tulleita maahanmuuttajia, ja neljä muuta luokkaa koskevat EU:n ja OECD-maiden ulkopuolelta tulleita maahanmuuttajia. EU- ja OECD-maiden ulkopuolelta tulevat on jaettu neljän ryhmään maahanmuuton syyn perusteella. Olen esittänyt näiden luokkien nimet suomeksi maassaoloaikoja käsittelevässä alaluvussa (ks. 13.4.1.1).

Tutkimus (Bratsberg et al. 2017) sisältää viranomaisrekistereistä peräisin olevia tietoja koskien kaikkia maahanmuuttajia, jotka olivat 18–47-vuotiaita, kun he muuttivat Norjaan vuosina 1990–2013. Tutkimuksessa esitetyt tulokset koskevat 25–62-vuotiaita, jotka eivät olleet opiskelijoita ja jotka olivat kunkin seurantavuoden lopussa Norjassa. Seuranta-aika tutkimuksessa oli vuodet 1993–2014. Maahanmuuttajien lisäksi tutkimukseen sisältyy 10 %:n satunnaisotos Norjassa syntyneestä kantaväestöstä. Bratsberg et al. 2017:n tutkimusaineisto on rakenteeltaan täten hyvin samankaltainen pitkittäis-/paneeliaineisto, jota käytän itse Suomea koskien tässä tutkimuksessa. (Bratsberg et al. 2017, s. 19, 25, 27).

Työllisyysasteet

Kuvassa 13.11 on esitetty Norjassa olevien maahanmuuttajien työllisten osuudet (%) maassaoloajan ja sukupuolen mukaan maahanmuuttajaryhmittäin. Tulokset koskevat 25–62-vuotiaita vuosina 1993–2014. Henkilön ”työllisyys” on määritetty niin, että henkilö on työllinen, mikäli hänen pääasiallinen tulonlähteensä on työ (*”main source of income is work”*). Kuvasta selviää, että humanitaarisina maahanmuuttajina (”Refugee”) Norjaan muuttaneilla työllisyys nousee aluksi nopeasti noin 7 maassaolovuoteen asti sekä miehillä että naisilla. Tämän jälkeen miehillä työllisyysaste alkaa laskea hieman maassaoloajan suhteen niin, että se on 20 vuoden maassaoloajan jälkeen noin 60 %:ia. Naisilla työllisyysaste jää noin 10 vuoden jälkeen 60 %:n tasolle. Kuvan 13.11 tuloksia tulkittaessa on syytä muistaa, että tässä käsitellään kaikki humanitaarisia maahanmuuttajia yhtenä ryhmänä ja että näissä tuloksissa työllisyyteen mahdollisesti vaikututtavia muita tekijöitä kuin maassaoloaika, sukupuoli ja maahanmuuttoryhmä ei oteta huomioon. Nämä huomioidaan kuitenkin Bratsberg et al. 2017:n tutkimuksen taulukossa 6 (s. 41), jonka tuloksia käsittelen vielä alla.



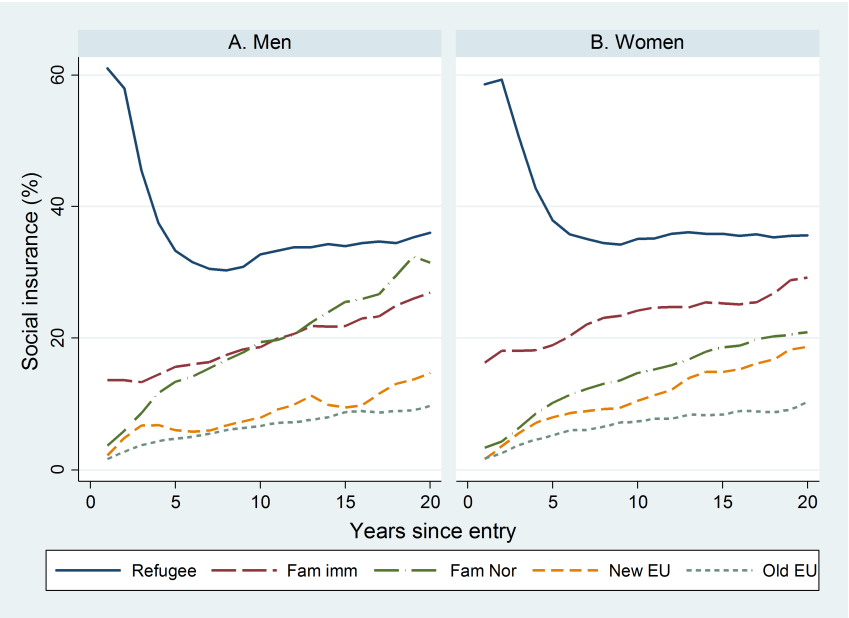
Kuva 13.11: *Henkilöiden osuudet, joiden pääasiallinen tulonlähde on työ. 25–62-vuotiaat maahanmuuttajaryhmän, sukupuolen (miehet vasemmalla, naiset oikealla) ja maassaoloajan mukaan vuosina 1993–2014. Lähde: Bratsberg et al. 2017, s. 28.*

Sosiaaliturvan tulonsiirrot

Kuvassa 13.12 on esitetty Norjassa olevien maahanmuuttajien niiden osuudet (%), joilla pääasiallinen tulonlähde ovat sosiaaliturvan tulonsiirrot (”social insurance transfers”). Kuvan tulokset on esitetty maassaoloajan ja sukupuolen mukaan maahanmuuttajaryhmittäin. Tulokset koskevat 25–62-vuotiaita vuosina 1993–2014. Kuvasta selviää, että humanitaarisina maahanmuuttajina (”Refugee”) Norjaan muuttaneilla

sosiaaliturva pääasiallisena tulonlähteenä laskee aluksi nopeasti noin 7 maassaolovuoteen asti sekä miehillä että naisilla. Tämän jälkeen miehillä sosiaaliturva pääasiallisena tulonlähteenä alkaa nousta hieman maassaoloajan suhteen niin, että se on 20 vuoden maassaoloajan jälkeen noin 35 %:ia. Naisilla vastaava osuus jää noin 7 vuoden jälkeen samaan noin 35 %:n tasolle. Kuvan 13.12 tulokset sosiaaliturvaa koskien humanitaarisilla maahanmuuttajilla ovat lähestulkoon käänteiset verrattuna työllisyysosuuksia koskeviin tuloksiin yllä kuvassa 13.11.

Kuvan 13.12 tuloksia tulkittaessa on syytä muistaa, että tässä käsitellään kaikkia humanitaarisia maahanmuuttajia yhtenä ryhmänä ja että näissä tuloksissa sosiaaliturvaan pääasiallisena tulonlähteenä mahdollisesti vaikututtavia muita tekijöitä kuin maassaoloaika, sukupuoli ja maahanmuuttoryhmä ei oteta huomioon. Nämä huomioidaan kuitenkin Bratsberg et al. 2017:n tutkimuksen taulukossa 6 (s. 41), jonka tuloksia käsittelen vielä alla.



Kuva 13.12: *Henkilöiden osuudet, joiden pääasiallinen tulonlähde ovat sosiaaliturvan tulonsiirrot ("social insurance transfers"). 25–62-vuotiaat maahanmuuttajaryhmän, sukupuolen (miehet vasemmalla, naiset oikealla) ja maassaoloajan mukaan vuosina 1993–2014. Lähde: Bratsberg et al. 2017, s. 29.*

Erot työllisyydessä ja sosiaaliturvan tulonsiirroissa verrattuna kantaväestöön

Bratsberg et al. 2017 laativat artikkelissaan lineaariset regressiomallit (OLS), joissa he "ennustavat" maahanmuuttajaryhmittäin eroja kantaväestöön työllisyydessä ja sosiaaliturvan käytössä. Näissä regressiomalleissa on otettu huomioon maahanmuuttajien: koulutustaso, onko koulutus saatu Norjassa, onko korkein koulutus Norjasta, puuttuuko koulutustieto, henkilön asuinpaikan työttömyysaste, maahanmuuttoikä sekä näiden interaktiot viiden maahanmuuttoluokan kanssa. Lisäksi malleissa on otettu huomioon

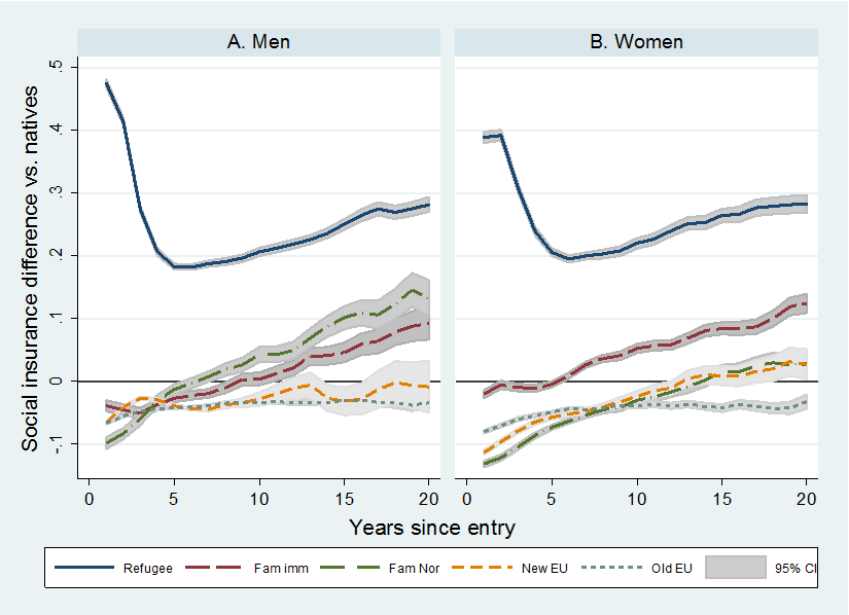
maahanmuuttajien: ikä, asuinlääni (”county”), havaintojen kalenterivuosi sekä syntymävaltion ja henkilön asuinpaikan työttömyysasteen interaktio Norjan kansalaisuuden kanssa. Ennustetut erot on esitetty alla olevissa kuvissa jokaiselle maahanmuuttajaryhmälle koulutustasolla, joka vastaa ryhmissä havaittua koulutustason painotettua keskiarvoa (”*Differentials are evaluated at the weighted average educational attainment in each immigrant sample*”). (Bratsberg et al. 2017, s. 32).

Ennustettu työllisyysero. Kuvassa 13.13 on esitetty ennustetut erot työllisten osuudessa norjalaiseen kantaväestöön verrattuna. Erot on esitetty maahanmuuttajaryhmän, sukupuolen ja maassaoloajan mukaan. Humanitaarisilla maahanmuuttajamiehillä (”Refugee”) ennustettu ero työllisyydessä verrattuna norjalaiseen kantaväestöön pienentyy aluksia nopeasti maassaoloajan kasvaessa siten, että noin 7 maassaolovuoden jälkeen ero on noin -20 %-yksikköä verrattuna kantaväestöön. Tämän jälkeen kuitenkin ero jälleen kasvaa siten, että 20 maassaolovuoden jälkeen ennustettu ero työllisyydessä on noin -30 %-yksikköä verrattuna kantaväestöön. Naisilla ennustettu ero verrattuna kantaväestöön on matalimmillaan noin -30 %-yksikköä verrattuna kantaväestöön (7–10 maassaolovuoden jälkeen). Tämän jälkeen ero jälleen kasvaa hieman niin, että se on noin -35 %-yksikköä 20 maassaolovuoden jälkeen.

Kuva 13.13: *Työ pääasiallisena tulonlähteenä: ennustettu ero työllisten osuudessa norjalaiseen kantaväestöön verrattuna. Erot maahanmuuttajaryhmän, sukupuolen (miehet vasemmalla, naiset oikealla) ja maassaoloajan mukaan. Harmaa alue on ennusteen 95 %:n luottamusväli. Lähde: Bratsberg et al. 2017, s. 32.*

Ennustettu ero sosiaaliturvan tulonsiirroissa. Kuva 13.14 käsittelee sosiaaliturvaa pääasiallisena tulonlähteenä. Kuvassa on esitetty ennustetut erot norjalaiseen kantaväestöön verrattuna niiden henkilöiden osuuksissa, joilla sosiaaliturva on pääasiallinen tulonlähde. Erot on esitetty maahanmuuttajaryhmän, su-

kupuolen ja maassaoloajan mukaan. Humanitaarisilla maahanmuuttajamiehillä (”Refugee”) ennustettu ero verrattuna norjalaiseen kantaväestöön pienentyy aluksi nopeasti maassaoloajan kasvaessa siten, että noin 7 maassaolovuoden jälkeen ero on noin 20 %-yksikköä verrattuna kantaväestöön. Tämän jälkeen kuitenkin ero jälleen kasvaa siten, että 20 maassaolovuoden jälkeen ennustettu ero työllisyydessä on noin 30 %-yksikköä verrattuna kantaväestöön. Naisilla ennustettu ero verrattuna kantaväestöön on matalimmillaan noin 20 %-yksikköä verrattuna kantaväestöön (7 maassaolovuoden jälkeen). Tämän jälkeen ero jälleen kasvaa hieman niin, että se on noin 30 %-yksikköä 20 maassaolovuoden jälkeen.



Kuva 13.14: *Sosiaaliturva pääasiallisena tulonlähteenä: ennustettu ero osuudessa norjalaiseen kantaväestöön verrattuna. Erot maahanmuuttajaryhmän, sukupuolen ja maassaoloajan mukaan. Harmaa alue on ennusteen 95 %:n luottamusväli. Lähde: Bratsberg et al. 2017, s. 33.*

Bratsberg et al. 2017 esittävät suurimpien humanitaaristen maahanmuuttajaryhmien osalta vielä erikseen maakohtaiset vaikutukset. Esitän niistä taulukossa 13.49 Irakia ja Somaliaa koskevat vaikutukset. Taulukon 13.49 mukaan Irakissa syntyneillä miehillä työllisyys on noin 6,6 %-yksikköä alhaisempi odotusarvoisesti kuin kaikilla humanitaarisilla maahanmuuttajilla yhteensä. Sosiaaliturva on pääasiallisena tulonlähteenä puolestaan odotusarvoisesti noin 6,4 %-yksikköä korkeampi kuin kaikilla humanitaarisilla maahanmuuttajilla yhteensä. Somaliassa syntyneillä miehillä vastaavat luvut ovat -11,1 %-yksikköä ja 8,2 %-yksikköä. Irakissa syntyneillä naisilla nämä luvut ovat -13,3 %-yksikköä ja 6,6 %-yksikköä, ja Somaliassa syntyneillä naisilla ne ovat puolestaan -21,7 %-yksikköä ja 23,2 %-yksikköä. (s. 41.).

Taulukko 13.49: *Työ ja sosiaaliturva pääasiallisena tulonlähteenä: Irakin ja Somalian syntymävaltiovaikutukset (regressiokertoimien estimaatit) sukupuolittain. Vaikutukset verrattuna humanitaaristen maahanmuuttajien ("Refugees") keskimääräiseen vaikutukseen (%-yksiköitä). Regressiokertoimien keskivirheet suluissa. Lähde: Bratsberg et al. 2017, s. 41.*

	Miehet		Naiset	
Syntymävaltio	Työllisyys	Sosiaaliturva	Työllisyys	Sosiaaliturva
Irak (estimaatti)	-6,6	6,4	-13,3	6,6
Irak (keskivirhe)	(0,8)	(0,7)	(1,2)	(1,2)
Somalia (estimaatti)	-11,1	8,2	-21,7	23,2
Somalia (keskivirhe)	(0,8)	(0,8)	(1,0)	(1,0)

Taulukon 13.49 tarkoittavat sitä, että kuviossa 13.13 ja 13.14 esitetyt ennustetut erot työllisyydessä ja sosiaaliturvan käytössä ovat Irakissa ja Somaliassa syntyneille vielä negatiivisempia (kuin humanitaarisille maahanmuuttajille yhteensä esitetyt ennusteet). Kaikkein suurin ero verrattuna humanitaarisiin maahanmuuttajiin yhteensä on Somaliassa syntyneillä naisilla. Kun kuvan 13.13 mukaan humanitaaristen maahanmuuttajanaisten työllisyys on 20 maassaolovuoden jälkeen noin 35 %-yksikköä negatiivisempi kuin Norjassa syntyneellä kantaväestöllä, niin taulukon 13.49 tulosten perusteella Somaliassa syntyneillä naisilla tämä ero on noin 57 %-yksikköä. Sosiaaliturva pääasiallisena lähteenä on Somaliassa syntyneillä naisilla puolestaan 20 maassaolovuoden jälkeen taulukon 13.49 ja kuvan 13.13 tulosten perusteella noin 50 %-yksikköä korkeampi kuin Norjassa syntyneellä kantaväestöllä (eikä vähän alle 30 %-yksikköä suurempi kuin kaikilla humanitaarisilla maahanmuuttajilla yhteensä).

Yhteenveto Norjaa koskevasta tutkimuksesta

Yhteenvetona Bratsberg et al. (2017):n tuloksista lainaan suoraan kyseisen artikkelin abstraktia: "*Some of our findings are unsettling: For immigrants admitted for protection or family reunification, we find that the initially encouraging labor-market integration process comes to a halt already after five to seven years, for most groups at employment levels still well below those of similar natives. More surprisingly, the integration process then runs into reverse, with consistently widening immigrant-native employment differentials over time.*" (s. 19).

Bratsberg et al. (2017):n mukaan jotkut tutkimuksen tuloksista ovat huolestuttavia. Tämä johtuu siitä, että humanitaarisilla ja perheisiin yhdistetyillä maahanmuuttajilla suotuisa työllisyyskehitys pysähtyy noin 5–7 vuoden maassaoloajan jälkeen – kun työllisyysasteet ovat useimmilla maahanmuuttajaryhmillä vielä huomattavasti kantaväestöä alhaisempia. Lisäksi tämän (5–7 maassaolovuoden) jälkeen ero kantaväestön työllisyyslukuihin alkaa jälleen kasvamaan. (s. 19).

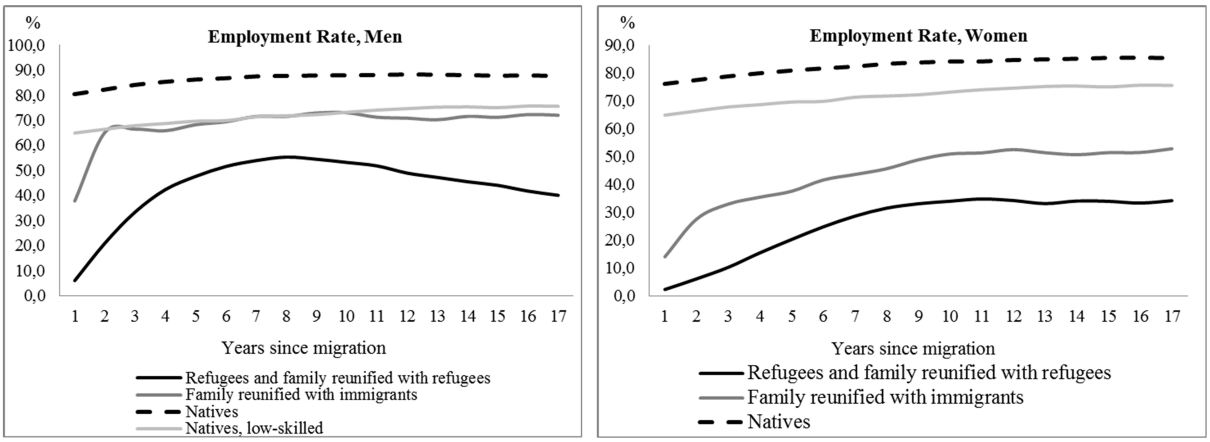
13.4.2.3 Tanska

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Tanskaan sijoittuvan artikkelin: *Labour market integration of refugees in Denmark* (Schultz-Nielsen 2017). Tässä tutkimusartikkelissa käsitellään pääasiallisesti Tanskaan vuosina 1997–2011 ensimmäistä kertaa muuttaneiden humanitaaristen maahanmuuttajien (artikkelissa ”pakolaiset”, *refugees*) työllisyyttä, tuloja ja sattuja tulonsiirtoja verrattuna Tanskassa syntyvään kantaväestöön. Tutkimuksessa humanitaarisia maahanmuuttajia koskevat tulokset eritellään sukupuolen mukaan sekä sen mukaan, tuliko maahanmuuttaja Tanskaan perheenyhdistämisen kautta vai ei. Lisäksi tulokset eritellään maahanmuuttajien osalta maahanmuuttovuosien kolmen luokan mukaan. Nämä maahanmuuttovuosien luokat ovat: vuosina 1997–2001, 2002–2006 ja 2007–2011 Tanskaan muuttaneet. Humanitaarisiin maahanmuuttajiin lasketaan tässä tutkimuksessa siis myös humanitaaristen maahanmuuttajien Tanskaan yhdistämät perheenjäsenet (Andersen et al. 2017, s. 55).

Tutkimusaineisto Schultz-Nielsen (2017):n artikkelissa koostuu vuosina 1997–2011 Tanskaan muuttaneista humanitaarisista maahanmuuttajista (ml. perheenyhdistämiset). Tutkimuksen seuranta-aika kattaa vuodet 1997–2013. Tutkimusaineisto, joka on pitkittäis-/paneeliaineisto ja on peräisin Tanskan viranomaisrekistereistä, muistuttaa muodoltaan oman tutkimukseni käytössä olevaan aineistoa. Schultz-Nielsen 2017:n tutkimusaineistoon sisältyvät vain ne henkilöt, jotka olivat muuttovuonna 17–36-vuotiaita. Lisärajoituksena on, että tulokset koskevat vain 25-vuotiaita ja vanhempia. Tutkimusaineisto sisältää vertailun vuoksi myös 10 %:n satunnaisotoksen Tanskassa vuosina 1961–1987 syntyneistä henkilöistä. (Schultz-Nielsen 2017, s. 66). Humanitaaristen maahanmuuttajien suurimpien lähtömaiden joukko tutkimusaineistossa sisältää mm. Irakin, Somalian, Afganistanin ja Iranin (ks. tarkemmin Schultz-Nielsen 2017, s. 86).

Työllisyysasteet vertailuväestöissä

Kuvassa 13.15 on esitetty työllisyysasteet (%) sukupuolen ja maassaoloajan mukaan maahanmuuttajaryhmittäin. Maahanmuuttajaryhmät ovat suomeksi: humanitaariset maahanmuuttajat ja heille yhdistetyt perheenjäsenet (*”Refugees and family reunified with refugees”*), perheenyhdistetyt muille kuin humanitaarisille maahanmuuttajille (”Family reunified with immigrants”), Tanskassa syntynyt kantaväestö (”Natives”) ja matalasti koulutetut kantaväestön henkilöt (”Natives, low-skilled”). Kuvan 13.15 tulosten mukaan humanitaarisilla maahanmuuttajamiehillä työllisyys nousee lähes nollasta (1 vuosi Tanskassa) noin 50 %:iin, kun muutosta Tanskaan kulunut noin kahdeksan vuotta. Tämän jälkeen kuitenkin työllisyysaste alkaa jälleen laskea, kun maassaoloaika kasvaa. Tämä tulos muistuttaa edellä esitettyä Norjaa koskevaa havaintoa. Humanitaarisilla maahanmuuttajanaisilla työllisyys nousee lähes nollasta (1 vuosi Tanskassa) noin 30 %:iin, kun muutosta Tanskaan kulunut noin 10 vuotta. Tämän jälkeen työllisyys ei naisilla enää nouse seuranta-ajan (17 maassaolovuotta) puitteissa. Tanskassa syntyneellä kantaväestöllä työllisyysaste on miehillä ja naisilla tutkimusaineiston seuranta-ajan puitteissa noin 80–90 %.



Kuva 13.15: *Työllisyysasteet (%) sukupuolen ("Men" on miehet, "Women" on naiset) ja maassaoloajan ("Years since migration") mukaan maahanmuuttajaryhmittäin. Maahanmuuttajaryhmät on esitetty suomeksi tekstissä. Lähde: Schultz-Nielsen 2017, s. 73.*

Työllisyyserot verrattuna kantaväestöön

Schultz-Nielsen 2017 laatii tutkimuksessaan "ennusteet" humanitaaristen maahanmuuttajien eroille verrattuna Tanskassa syntyneeseen kantaväestöön kolmessa eri tutkittavassa asiassa: työllisyysasteissa, palkkatuloissa ja saaduissa tulonsiirroissa. Tavoitteena näissä eroennusteissa on, että humanitaarisia maahanmuuttajia verrataan taustatekijöiltään "vastaaviin" kantaväestön henkilöihin. Tätä varten eroennusteissa otetaan huomioon ero mm. vertailtavien väestöjen sukupuolta, ikää ja lasten määrää koskevissa rakenteissa. Lisäksi ennusteissa otetaan huomioon mm. henkilöiden asuinpaikan työttömyysaste sekä henkilöiden lasten määrä ja lasten ikä. Laskentatapa on esitetty tarkemmin Schultz-Nielsen 2017:n liitteessä. (Schultz-Nielsen 2017, s. 73–74).

Taulukossa 13.50 on esitetty humanitaaristen maahanmuuttajien ennustetut erot (%-yksikköä) työllisyysasteessa verrattuna Tanskassa syntyneeseen kantaväestöön. Ennustetut erot esitetään maassaoloajan, sukupuolen ja maahanmuuttovuoden luokan mukaan. Tulokset on esitetty taulukossa neljän maassaoloajan mukaan: 2, 5, 10 ja 15 vuotta maahanmuutosta. Taulukon 13.50 estimaattien mukaan humanitaaristen maahanmuuttajamiesten ennustettu ero työllisyysasteessa verrattuna kantaväestöön kahden maassaolovuoden jälkeen vaihtelee -47 %-yksiköstä (v. 2002–2006 muuttaneet) -67,6 %-yksikköön (v. 1997–2001 muuttaneet). Ensimmäisen ja toisen muuttokohortin miehiä koskevien tulosten perusteella vaikuttaa siltä, että vaikka erinäiset taustatekijät huomioitaisiin, humanitaaristen maahanmuuttajien ero työllisyysasteessa verrattuna kantaväestöön pienenee noin 5–10 maassaolovuoteen asti, mutta sen jälkeen ero jälleen kasvaa. Naisilla puolestaan ero vaikuttaa pienenevän verrattuna kantaväestöön vielä 10–15 massaolovuoden jälkeenkin, mutta ero työllisyysasteessa on 10–15 maassaolovuoden jälkeen vielä 40 %-yksikköä verrattuna kantaväestöön 1. ja 2. muuttokohorttia koskevien tulosten perusteella.

Taulukko 13.50: *Humanitaaristen maahanmuuttajien ennustetut erot (%-yksikköä) työllisyysasteessa verrattuna Tanskassa syntyneeseen kantaväestöön. Ennustetut erot maassaoloajan, sukupuolen ja maahanmuuttovuoden luokan mukaan. Ennuste-erojen estimaattien (est.) keskivirheet (kv) suluissa. 1. muuttokohortti käsittää v. 1997–2001 muuttaneet, 2. muuttokohortti käsittää v. 2002–2006 muuttaneet ja 3. muuttokohortti käsittää v. 2007–2011 muuttaneet. Lähde: Schultz-Nielsen 2017, s. 75.*

	Miehet				Naiset			
Vuosia maahanmuutosta:	2	5	10	15	2	5	10	15
1. muuttokohortti (est.)	-67,6	-43,6	-30,9	-34,6	-77,6	-63,8	-46,1	-42,1
1. muuttokohortti (kv)	(0,7)	(0,7)	(0,6)	(0,9)	(0,5)	(0,5)	(0,6)	(0,9)
2. muuttokohortti (est.)	-47,0	-25,7	-30,8	–	-72,3	-51,6	-43,4	–
2. muuttokohortti (kv)	(1,5)	(1,2)	(1,4)	–	(1,0)	(1,0)	(1,2)	–
3. muuttokohdortti (est.)	-60,6	-38,3	–	–	-70,2	-54,3	–	–
3. muuttokohdortti (kv)	(1,7)	(1,9)	–	–	(1,3)	(1,5)	–	–

Erot palkkatuloissa verrattuna kantaväestöön

Taulukossa 13.51 on esitetty humanitaaristen maahanmuuttajien ennustetut erot (tuhatta Tanskan kruunua (DKK)) vuosittaisissa palkkatuloissa verrattuna Tanskassa syntyneeseen kantaväestöön. Artikkelissa vuosittaisista tuloista käytetään englanniksi nimitystä *annual earnings*, mutta koska nämä tulot eivät sisällä henkilöiden elinkeinonharjoittajana (artikkelissa engl.*self-employed*) ansaitsemia tuloja, tulkitsen tämän niin, että tulotiedoista puuttuvat henkilöiden yrittäjätulot (Schultz-Nielsen 2017, s. 76). Sen vuoksi suomennan ”earnings”-sanan tässä yhteydessä palkkatuloiksi. Nollatulot sisältyvät tuloksiin.

Ennustetut erot palkkatuloissa on esitetty taulukossa 13.51 maassaoloajan, sukupuolen ja maahanmuuttovuoden luokan mukaan. Tulokset on esitetty taulukossa neljän maassaoloajan mukaan: 2, 5, 10 ja 15 vuotta maahanmuutosta. Taulukon 13.51 estimaattien mukaan humanitaaristen maahanmuuttajamiesten ennustettu ero vuosittaisissa palkkatuloissa (tuhatta Tanskan kruunua) verrattuna kantaväestöön vaihtelee kahden maassaolovuoden jälkeen -229:sta (v. 2002–2006 muuttaneet) -259:ään (v. 1997–2001 muuttaneet). Maaliskuun 2018 valuuttakurssien perusteella nämä erot vuosittaisissa palkkatuloissa vastaavat noin -30 000 ja -34 000 euroa. Ensimmäisen ja toisen muuttokohortin miehiä koskevien tulosten perusteella vaikuttaa siltä, että vaikka erinäiset taustatekijät huomioitaisiin, humanitaaristen maahanmuuttajien ero vuosittaisissa palkkatuloissa verrattuna kantaväestöön pienenee noin 5–10 maassaolovuoteen asti, mutta sen jälkeen ero jälleen kasvaa. Naisilla puolestaan palkkatuloeroissa ei vaikuta juuri tapahtuvan kehitystä maassaoloajan suhteen. Ero kantaväestöön on kaikissa muuttokohorteissa ja maassaoloajan luokissa noin -170:n ja -190:n (tuhannen) Tanskan kruunun välillä.

Taulukko 13.51: *Humanitaaristen maahanmuuttajien ennustetut erot (tuhatta Tanskan kruunua (DKK)) vuosittaisissa palkkatuloissa verrattuna Tanskassa syntyneeseen kantaväestöön. Tuloerot vuoden 2015 kruunuissa. Ennustetut erot maassaoloajan, sukupuolen ja maahanmuuttovuoden luokan mukaan. Ennuste-erojen estimaattien (est.) keskivirheet (kv) suluissa. 1. muuttokohortti käsittää v. 1997–2001 muuttaneet, 2. muuttokohortti käsittää v. 2002–2006 muuttaneet ja 3. muuttokohortti käsittää v. 2007–2011 muuttaneet. Lähde: Schultz-Nielsen 2017, s. 77.*

	Miehet				Naiset			
Vuosia maahanmuutosta:	2	5	10	15	2	5	10	15
1. muuttokohortti (est.)	-259	-221	-203	-228	-169	-180	-174	-174
1. muuttokohortti (kv)	(2,2)	(2,2)	(2,7)	(4,0)	(1,4)	(0,5)	(0,6)	(0,9)
2. muuttokohortti (est.)	-229	-158	-211	–	-176	-168	-177	–
2. muuttokohortti (kv)	(4,3)	(4,3)	(5,3)	–	(2,2)	(2,3)	(3,2)	–
3. muuttokohdortti (est.)	-256	-231	–	–	-188	-193	–	–
3. muuttokohdortti (kv)	(4,0)	(5,2)	–	–	(2,3)	(2,9)	–	–

Erot saaduissa tulonsiirroissa verrattuna kantaväestöön

Taulukossa 13.52 on esitetty humanitaaristen maahanmuuttajien ennustetut erot (tuhatta Tanskan kruunua (DKK)) vuosittaisissa saaduissa tulonsiirroissa (artikkelissa englanniksi *transfer payments*) verrattuna Tanskassa syntyneeseen kantaväestöön. Saadut tulonsiirrot sisältävät mm. kuntoutustuet, työttömyystuet, opintotuet, asumistuet, lapsilisät ja työkyvyttömyystuet (Schultz-Nielsen 2017, s. 78.).

Ennustetut erot saaduissa tulonsiirroissa on esitetty taulukossa 13.52 maassaoloajan, sukupuolen ja maahanmuuttovuoden luokan mukaan. Tulokset on esitetty taulukossa neljän maassaoloajan mukaan: 2, 5, 10 ja 15 vuotta maahanmuutosta. Taulukon 13.52 estimaattien mukaan humanitaaristen maahanmuuttajamiesten ennustettu ero vuosittaisissa saaduissa tulonsiirroissa (tuhatta Tanskan kruunua) verrattuna kantaväestöön vaihtelee kahden maassaolovuoden jälkeen 69:sta (v. 2002–2006 muuttaneet) 106:een (v. 1997–2001 muuttaneet). Maaliskuun 2018 valuuttakurssien perusteella nämä erot vuosittaisissa saaduissa tulonsiirroissa vastaavat noin 9 000 ja 14 000 euroa. Ensimmäisen ja toisen muuttokohortin miehiä koskevien tulosten perusteella vaikuttaa siltä, että vaikka erinäiset taustatekijät huomioitaisiin, humanitaaristen maahanmuuttajien ero vuosittaisissa palkkatuloissa verrattuna kantaväestöön pienenee noin 5–10 maassaolovuoteen asti, mutta sen jälkeen ero jälleen kasvaa. Sama tulos koskee myös naisia taulukon 13.52 tulosten perusteella, eli erot saaduissa tulonsiirroissa verrattuna kantaväestöön pienenevät noin 5–10 ensimmäisen maassaoloajan aikana, mutta sen jälkeen ero saaduissa tulonsiirroissa verrattuna kantaväestöön jälleen kasvaa.

Taulukko 13.52: *Humanitaaristen maahanmuuttajien ennustetut erot (tuhatta Tanskan kruunua (DKK)) vuosittaisissa saaduissa tulonsiirroissa verrattuna Tanskassa syntyneeseen kantaväestöön. Saadut tulonsiirrot vuoden 2015 kruunuissa. Ennustetut erot maassaoloajan, sukupuolen ja maahanmuuttovuoden luokan mukaan. Ennuste-erojen estimaattien (est.) keskivirheet (kv) suluissa. 1. muuttokohortti käsittää v. 1997–2001 muuttaneet, 2. muuttokohortti käsittää v. 2002–2006 muuttaneet ja 3. muuttokohortti käsittää v. 2007–2011 muuttaneet. Lähde: Schultz-Nielsen 2017, s. 80.*

	Miehet				Naiset			
Vuosia maahanmuutosta:	2	5	10	15	2	5	10	15
1. muuttokohortti (est.)	106	80	54	61	92	90	71	70
1. muuttokohortti (kv)	(1,2)	(1,0)	(1,1)	(1,7)	(1,1)	(1,0)	(1,1)	(1,6)
2. muuttokohortti (est.)	69	40	56	–	38	34	56	–
2. muuttokohortti (kv)	(2,0)	(1,7)	(2,4)	–	(1,9)	(1,7)	(2,0)	–
3. muuttokohdortti (est.)	76	54	–	–	38	52	–	–
3. muuttokohdortti (kv)	(2,1)	(2,4)	–	–	(2,9)	(2,8)	–	–

Yhteenveto Tanskaa koskevasta tutkimuksesta

Yhteenvetona Schultz-Nielsen (2017):n tuloksista lainaan suoraan kaksi virkettä kyseisen artikkelin maahanmuuttajien integraatiota käsittelevästä luvusta (2.4.4): ”*As demonstrated in this study, a little more than 40 % of refugees are in employment after ten years in Denmark, their average earnings are lower and they receive more transfer income than natives. This situation has serious consequences, not only on the living conditions of the individual refugee, but also – if the flow of refugees reaches a considerable size – for the public finances.*” (s. 81).

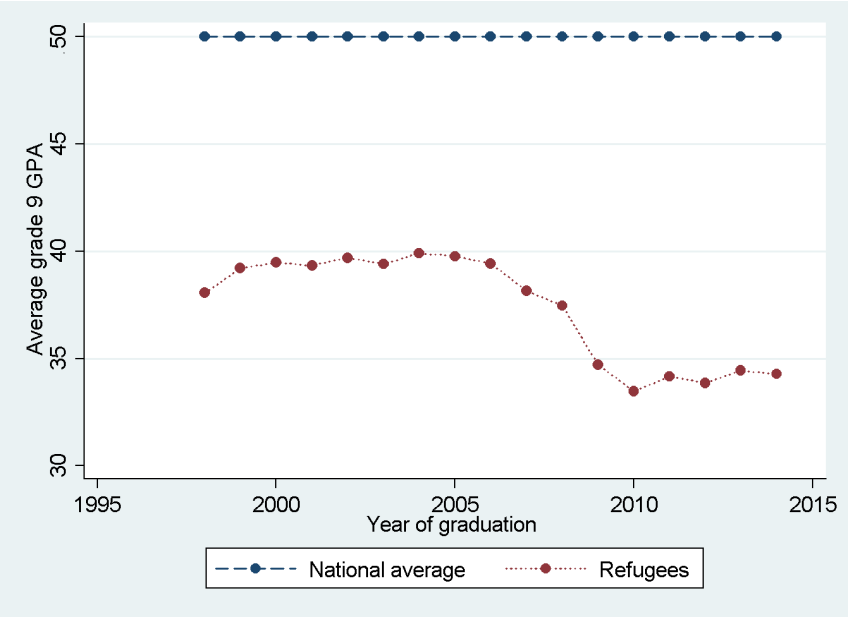
Schultz-Nielsen (2017):n mukaan hieman yli 40 %:ia humanitaarisista maahanmuuttajista on työllisiä 10 vuoden maassaolovuoden jälkeen, heidän palkkatulonsa ovat pienempiä ja saadut tulonsiirrot suurempia kuin kantaväestöllä. Tällä tilanteella on vakavat seuraukset sekä yksittäisten humanitaaristen maahanmuuttajien elinoloille että koko Tanskan julkiselle taloudelle – jos humanitaarinen maahanmuutto Tanskaan on lukumäärällisesti merkittävän suurta. (s. 81).

13.4.3 Humanitaaristen maahanmuuttajien lasten peruskoulutodistukset Ruotsissa

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Ruotsiin sijoittuvan artikkelin: "*The School Achievements of Refugee Children: Lessons from Sweden*"(Grönqvist and Niknami 2017). Artikkeli vertailee humanitaarisina maahanmuuttajina (artikkelissa: *”Refugee Children”*) Ruotsiin muuttaneiden lasten peruskoulun päättötodistuksia kaikkien oppilaiden vastaaviin todistuksiin Ruotsissa. Tutkimusaineisto koostuu rekisteritiedoista, jotka koskevat arvosanoja yhdeksännen luokan todistuksissa (jolloin oppilaat ovat yleensä 16-vuotiaita) vuosina 1998–2014. Ruotsin tilastokeskus (SCB) on lisäksi liittänyt tutkimusaineistoon tietoja koskien oppilaiden vanhempia, esimerkiksi vanhempien tulotasoa koskien. Tutki-

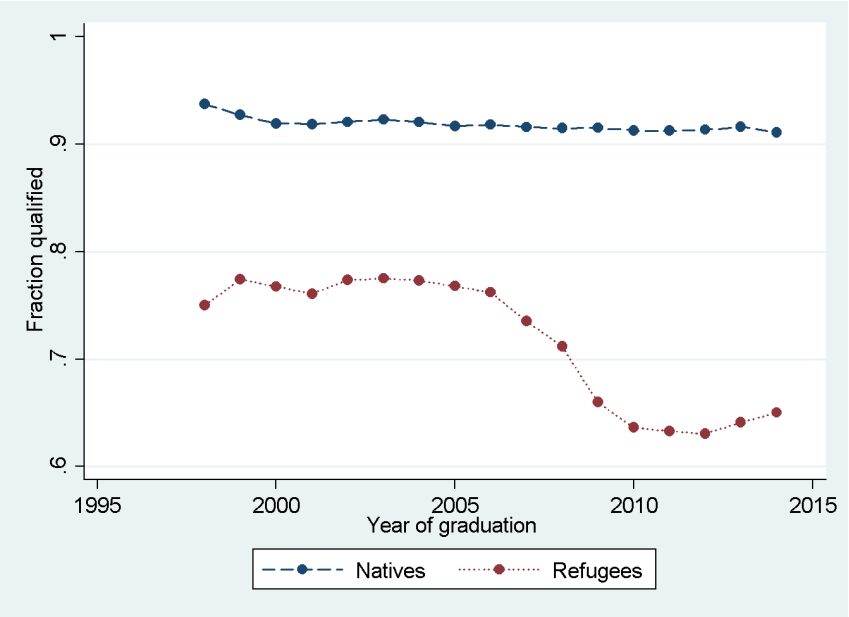
musaineisto sisältää yli 74 000 oppilasta, jotka muuttivat Ruotsiin humanitaarisina maahanmuuttajina (Andersen et al. 2017, s. 161).

Kuvassa 13.16 on esitetty peruskoulun päättötodistusten keskiarvot (GPA) Ruotsissa vuosina 1998–2014. Kuvassa on sekä Ruotsin kansallinen keskiarvo (*National average*) ja humanitaariset maahanmuuttajat (*Refugees*). Ruotsin kansallinen keskiarvo saa arvon 50. Humanitaarisina maahanmuuttajina Ruotsiin muuttaneiden keskiarvo on ollut vuosina 1998–2007 noin 40. Tämän jälkeen humanitaaristen maahanmuuttajien keskiarvo on laskenut 2010-luvulla noin 34:ään. Kuvaa tulkittaessa on huomioitava, että esimerkiksi muutoksia humanitaaristen maahanmuuttajalasten lähtömaissa ei oteta huomioon.



Kuva 13.16: *Peruskoulun päättötodistusten keskiarvot (GPA) Ruotsissa vuosina 1998–2014. Kansallinen keskiarvo (National average) ja humanitaariset maahanmuuttajat (Refugees). Lähde: Grönqvist and Niknami 2017, s. 170.*

Kuvassa 13.17 on esitetty niiden oppilaiden osuudet (%) Ruotsissa vuosina 1998–2014, jotka ovat oikeutettuja hakeutumaan toisen asteen opetukseen peruskoulun jälkeen (*Fraction qualified*). Kuvassa on sekä Ruotsin kantaväestön (*Natives*) ja humanitaaristen maahanmuuttajien (*Refugees*) lasten osuudet. Kantaväestön osuus on kuvan 13.17 perusteella ollut vuosina 1998–2014 noin 92 %:ia. Humanitaarisina maahanmuuttajina Ruotsiin muuttaneiden osuus on ollut vuosina 1998–2007 noin 77 %:ia. Tämän jälkeen humanitaaristen maahanmuuttajien keskiarvo on laskenut 2010-luvulla noin 65:een %:iin. Kuvaa tulkittaessa on huomioitava, että esimerkiksi muutoksia humanitaaristen maahanmuuttajalasten lähtömaissa ei oteta huomioon.



Kuva 13.17: *Toisen asteen koulutukseen hyväksyttyjen osuudet (fraction qualified) Ruotsissa vuosina 1998–2014. Kantaväestön (Natives) ja humanitaariset maahanmuuttajat (refugees). Lähde: Grönqvist and Niknami 2017, s. 172.*

Irakissa ja Somaliassa syntyneiden maahanmuuttajien lasten 2. asteen koulutus Suomessa. Tämän tutkimuksen tutkimusaineistoon ei sisälly rekisteritietoja peruskoulun päättötodistusten arvosanoista Suomessa. Olen kuitenkin esittänyt maahanmuuttajien lapsien elinkaarivaikutuksia käsittelevässä luvussa (ks. luku 8.3.4) maahanmuuttajien 22-vuotiaita lapsia koskevia tietoja maahanmuuttajaryhmittäin vanhempien syntymävaltion mukaan. Esitin samat tiedot myös Suomessa syntyneiden koko väestön 22-vuotiaita koskien. Osana näitä tuloksia esitin niiden henkilöiden osuudet väestöittäin, joilla on vähintään 2. asteen koulutus suoritettu 22-vuotiaana. Nämä maahanmuuttajien lapsien koulutusta koskevat tiedot on esitetty maahanmuuttoiän mukaan Irakissa ja Somaliassa syntyneille taulukossa 13.53. Taulukon alimmalla rivillä on sama tieto myös Suomessa syntyneiden koko väestön 22-vuotiaita koskien.

Taulukko 13.53: *Henkilöiden osuudet (%) väestöittäin syntymävaltion mukaan, joilla on vähintään 2. asteen koulutus 22-vuotiaana. Irakissa ja Somaliassa syntyneiden maahanmuuttajien 22-vuotiaat lapset maahanmuuttoiän mukaan vuosina 2005–2014. Luokka "Alle 8" sisältää myös Suomessa syntyneet maahanmuuttajien lapset. Lisäksi alimmalla rivillä on Suomessa syntyneiden 22-vuotiaiden tiedot vuosina 2005–2011. Lähde: Tilastokeskus (rekisteriaineistot).*

	Alle 8	8–14	15–20	Yht.
Irak	53	40	17	40
Somalia	35	29	10	26
Suomi	–	–	–	84

Suomessa syntyneiden koko väestössä 22-vuotiaista 84 %:lla oli vähintään 2. asteen koulutus suoritettuna vuosina 2005–2011. Irakissa syntyneiden maahanmuuttajien 22-vuotiailla lapsilla tämä osuus oli 40 %:ia, kun käsitellään koko 22-vuotiaiden väestöä. Mikäli asiaa tarkastellaan maahanmuuton ikäluokan mukaan, niin suurin osuus vähintään 2. asteen koulutuksen suorittaneista on alle 8-vuotiaana Suomeen muuttaneilla (sis. myös Suomessa syntyneet), 53 %:ia tästä väestöstä. Somaliassa syntyneiden maahanmuuttajien 22-vuotiailla lasta 26 %:lla oli vähintään 2. asteen tutkinto, kun käsitellään koko 22-vuotiaiden väestöä. Mikäli asiaa tarkastellaan maahanmuuton ikäluokan mukaan, niin suurin osuus vähintään 2. asteen koulutuksen suorittaneista on alle 8-vuotiaana Suomeen muuttaneilla (sis. myös Suomessa syntyneet), 35 %:ia tästä väestöstä.

Grönqvist and Niknami (2017):n tutkimuksen esittämät tulokset Ruotsia koskien (peruskoulutus) että tässä tutkimuksessa esitetyt tiedot 2. asteen koulutusta koskien eivät ole täysin vertailukelpoisia, sillä Ruotsin osalta tutkittava väestö koskee kaikkia humanitaarisia maahanmuuttajia ja Suomen osalta Irakissa ja Somaliassa syntyneitä. Irakia ja Somaliaa koskevat tulokset nimittäin osoittavat (taulukko 13.53), että 2. asteen koulutuksen suorittaminen vaihtelee näiden kahden humanitaarisen maahanmuuton ryhmän välillä yli 10 prosenttiyksikköä. Tulokset viittaavat joka tapauksessa siihen, että humanitaaristen maahanmuuttajien lapsilla suoritettu peruskoulutus ja 2. asteen koulutus jäävät keskimäärin kantaväestöä heikommaksi.

13.4.4 Julkisen talouden nettovaikutukset

13.4.4.1 Toteutuneet julkisen talouden nettovaikutukset Ruotsissa

Tutkimusartikkelien kokoelma (Andersen et al. 2017) sisältää Ruotsiin sijoittuvan artikkelin: ”*The fiscal impact of refugee immigration*” (Ruist 2017). Artikkeli vertailee humanitaarisina maahanmuuttajina (artikkelissa: *”refugee”*) Ruotsiin muuttaneiden julkisen talouden nettovaikutuksia verrattuna Ruotisin koko väestöön. Artikkelissaan Ruist (2017) käy aluksi läpi yleisempää keskustelua liittyen julkisen talouden nettovaikutusten laskemiseen. Tämän jälkeen Ruist esittää tuloksia kahdesta ruotsalaistutkimuksesta, joissa on tutkittu humanitaaristen maahanmuuttajien julkisen talouden nettovaikutuksia Ruotsissa. Käsittelen tässä yhteydessä näistä kahdesta vain Ruist (2015):n tuloksia, sillä toinen Ruist (2017):n esittelemä tutkimus ei vertaa nettovaikutustuloksia lainkaan kantaväestöön tai Ruotsin koko väestöön.

Ruist (2015) julkisen talouden toteutuneet nettovaikutukset humanitaarisilla maahanmuuttajilla vuonna 2007. Ruist (2015):ssä lasketaan julkisen talouden *toteutuneet* nettovaikutukset vuonna 2007 sekä kaikkien humanitaaristen maahanmuuttajien muodostamalle väestölle että Ruotsin koko väestölle. Humanitaaristen maahanmuuttajien väestöön lasketaan myös humanitaarisille maahanmuuttajille yhdistetyt perheenjäsenet. Ruist (2017) esittää lyhyesti, kuinka julkisen talouden nettovaikutukset on muodostettu Ruist (2015):ssä (s. 223). Laskentatapa muistuttaa hyvin paljon tämän tutkimuksen I osan Salminen (2015a) toteuttamistapaa.

Suurimmat erot oman tutkimukseni I osan ja Ruist (2015):n välillä ovat mielestäni seuraavat: Ensiksikin Ruist (2015) sisällyttää mukaan julkisen talouden tuloihin myös yritysten maksamat verot (tämän tutkimuksen määritelmän mukaan yhteisöverot eivät sisälly mukaan henkilöihin liitettäviin vaikutuksiin

(ks. tarkemmin yhteisöihin liittyvien vaikutusten sisältymisestä tähän tutkimukseen luku 10.4)). Toiseksi Ruist (2015):n tulokset koskevat vain vuotta 2007, kun tämän tutkimuksen I osassa esitettiin tulokset kaikille vuosille 1995–2011. Kolmanneksi Ruist (2015) esittää tulokset vain Ruotsin koko väestölle sekä kaikille humanitaarisille maahanmuuttajille yhteensä, kun tämän tutkimuksen I osassa esitän tulokset erikseen 10 suurimmalle (syntymävaltion mukaan määritellylle) maahanmuuttajaryhmälle sekä Suomessa syntyneelle kantaväestölle. Lisäksi esitän tulokset syntymäalueiden mukaan sekä sukupuolen, iän ja pääasiallisen toiminnan mukaan.

Ruist (2017) huomauttaa lisäksi, että Ruist (2015):een ei sisälly ”epäsuoria” (artikkelissa ”dynamic”) julkisen talouden vaikutuksia. Niitä ei sisälly myöskään tähän tutkimukseen. Epäsuoria julkisen talouden vaikutuksia ovat Ruist (2017):n mukaan kaikki sellaiset vaikutukset, jotka vaikuttavat julkiseen talouteen muiden (kokonais)talouden toimijoiden kautta. (s. 217). Vaikka näin ollen Ruist (2015):een (eikä myöskään omaan tutkimukseeni) sisälly ”todellista vaikutusta”, Ruist (2017):n mukaan suorien vaikutusten (artikkelissa ”redistribution”), Ruist uskoo, että suorat vaikutukset ovat ”kohtalainen sijaisarvio” (artikkelissa *”fair proxy”*) todellisille vaikutuksille: ”*[...] while the calculated redistribution is not equal to the true effect, my belief that it is a fair enough proxy for this effect is indeed the very reason why I am calculating it.*” (Ruist 2017, s. 217).

Taulukossa 13.54 on esitetty julkisen talouden nettovaikutuksia koskevat tulokset Ruistin (2015) tutkimuksesta. Taulukon ylemmillä rivillä ovat julkisen talouden tulot (yhteensä) sekä julkisen talouden tulojen komponentit. Taulukon alemmilla riveillä (viimeistä riviä lukuun ottamatta) ovat julkisen talouden menot (yhteensä) sekä julkisen talouden menojen komponentit. Taulukon viimeisellä rivillä ovat julkisen talouden nettovaikutukset. Luvut koskevat vuotta 2007 ja ovat miljardeja Ruotsin kruunuja (SEK). Taulukon ensimmäisellä sarakkeella ovat tulokset koskien koko Ruotsin väestöä, toisella sarakkeella humanitaariset maahanmuuttajat yhteensä sekä kolmannella sarakkeella on prosenttiosuus, jonka humanitaariset maahanmuuttajat muodostavat kustakin julkisen talouden vaikutuksesta.

Taulukon 13.54 esittämistä tuloksista selviää esimerkiksi, että humanitaariset maahanmuuttajat saivat yli puolet kaikista kruunumääräisistä sosiaaliturvaetuuksista (artikkelissa *social assistance*) vuonna 2007. Myös maahanmuuttajakohtaisissa menoissa humanitaaristen maahanmuuttajien osuus (28,7 %) on (ei yllättävästi) korkea. Julkisen talouden vaikutuksia, joissa humanitaaristen maahanmuuttajien osuus on pieni koko väestöön verrattuna, ovat taulukon 13.54 mukaan henkilöiden maksamat suorat verot (3 %:ia), yritysten maksamat suorat verot (0 %:ia, toisin kuin useimpien vaikutusten kohdalla, joiden tulokset perustuvat rekisteritietoihin, tämä perustuu Ruistin omaan oletukseen), sekä eläkkeet (1,2 %:ia).

Koko Ruotsin väestöllä julkisen talouden nettovaikutukset olivat taulukon 13.54 tulosten perusteella 0 kruunua, sillä julkisen talouden tulot ja menot ovat yhtä suuret. Humanitaaristen maahanmuuttajien yhteenlaskettu julkisen talouden nettovaikutus oli puolestaan -32,5 miljardia Ruotsin kruunua vuonna 2007. Maaliskuun 2018 valuuttakursseilla tämä summa on noin -3,2 miljardia euroa (on huomioitava, että Ruistin tulokset ovat ilmeisesti vuoden 2007 hinnoissa, joten nykyrahassa ilmoitetut summat olisivat nimellisesti suurempia).

Taulukko 13.54: *Toteutuneet julkisen talouden nettovaikutukset Ruotsissa vuonna 2007 (miljardia Ruotsin kruunua (SEK)). Koko väestön ja humanitaaristen maahanmuuttajien vaikutukset julkisen talouden sektoreittain sekä julkisen talouden tulot, menot ja nettovaikutukset yhteensä. Lähde: Ruist (2017), s. 228.*

	Koko väestö	Humanitaariset maahanmuuttajat	Humanit. osuus (%)
Julkisen talouden tulot	**1473,3**	**49,4**	**3,4**
Suorat verot, henkilöt	536,8	16,2	3,0
Sosiaalivakuutusmaksut	391,15	15,3	3,9
Kulutusverot	398,6	15,9	4,0
Suorat verot, yritykset	106,6	0	0
Muut tulot	39,8	2,0	5,1
Julkisen talouden menot	**1473,3**	**81,9**	**5,6**
Sosiaaliturva	7,8	4,3	55,4
Eläkkeet	286,7	3,3	1,2
Muut saadut tulonsiirrot	195,4	18,8	9,6
Sairaalahoito ja vanhustenhuolto	240,6	9,1	3,8
Vammaistuet	37,2	2,1	5,7
Koulutus ja lastenhoito	158,0	4,9	3,1
Rikokset ja oikeus	30,6	7,7	25,3
Maahanmuuttajakohtaiset menot	22,3	6,4	28,7
Muut menot	494,8	25,3	5,1
Julkisen talouden nettovaikutukset	**0**	**-32,5**	–

Kiinnostava huomio, jonka Ruist (2017) esittää tuloksia koskien, on, että humanitaaristen maahanmuuttajien negatiivinen julkisen talouden nettovaikutus johtuu 80 %:sti matalista julkisen talouden tuloista (eikä korkeista julkisen talouden menoista) (s. 229). Sama ilmiö on todennettavissa myös Suomen osalta ainakin suorien (netto)tulonsiirtojen osalta (ks. tutkimuksen I osa, luku 6.2), eli suurempi osa eroista nettotulonsiirroissa Suomessa syntyneen kantaväestön ja Somaliassa ja Irakissa syntyneiden välillä aiheutuu eroista maksetuissa tulonsiirroissa (verot, yms.) kuin eroista saaduissa tulonsiirroissa.

13.4.4.2 Elinkaarivaikutukset

Ruotsi

Edellä esitettiin pakolaistaustaisten maahanmuuttajien toteutuneita julkisen talouden nettovaikutuksia Ruotsissa vuonna 2007. Nämä luvut olivat peräisin Ruist (2015):n tutkimuksesta. Ruist tutkii pakolaistaustaisten maahanmuuttajien julkisen talouden vaikutuksia myös Ruotsin *Regeringskanslietille* (”Hallituksen kanslialle”) laatimassaan tutkimuksessa Ruist (2018). Tässä uudemmassa tutkimuksessa hän laskee pakolaistaustaisten toteutuneiden julkisen talouden nettovaikutusten summan vuodelle 2015. Kun ruotsalaisen kantaväestön julkisen talouden toteutuneen nettovaikutuksen summa vuonna 2015 oli tasan

0 (Ruotsin kruunua), niin samana vuonna pakolaistaustaisten yhteenlaskettu julkisen talouden nettovaikutus oli -41,5 miljardia Ruotsin kruunua eli noin 4 miljardia euroa (helmikuun 2019 vaihtokursseilla). (Ruist 2018, s. 66–69).

Elinkaariennusteiden laskentatavasta. Ruist (2018) käyttää vuodelle 2015 laskettuja toteutuneita julkisen talouden vaikutuksia sen laskemiseen, mitkä ovat pakolaistaustaisten maahanmuuttajien julkisen talouden elinkaarivaikutukset Ruotsissa. Elinkaarivaikutusten ennusteiden laskentatapa on esitetty tarkemmin raportin sivuilla 75–76. Joka tapauksessa elinkaariennusteet perustuvat laskettuihin toteutuneisiin julkisen talouden nettovaikutuksiin. Laskelmissa otetaan huomioon maahanmuuttajien muuttoiät Ruotsiin, maassaoloajan vaikutus nettovaikutuksiin sekä maahanmuuttajien alkuperämaa.

Elinkaarivaikutusennusteissa maassaoloajan annetaan vaikuttaa julkisen talouden nettovaikutuksiin (positiivisesti) 30 ensimmäisen maassaolovuoden aikana, minkä jälkeen ennusteissa julkisen talouden nettovaikutukset pysyvät maassaoloajan suhteen vakioina eläkeikään asti. Elinkaariennusteissa oletetaan, että pakolaistaustaiset maahanmuuttajat jäävät koko loppuiäkseen Ruotsiin. Ruistin mukaan tämä oletus ei poikkea paljoa todellisuudesta, sillä pakolaistaustaustaisten maahanmuuttajien poismuutto Ruotsista on pientä (vrt. aiemmin luvussa 13.4.1 esitetyt vastaavat tulokset Norjasta ja Tanskasta sekä olotilaaikojen mallinnuksen yhteydessä esitetyt vastaavat tulokset Suomeen 2000-luvulla muuttaneita koskien). Ruist (2018) olettaa, että ruotsalaisen kantaväestön ja pakolaistaustaisten maahanmuuttajien nettovaikutukset ovat samansuuruiset eläkeikäisenä. Tämä oletus on tämän tutkimuksen tietojen perusteella varsin epärealistinen, koska Somaliassa ja Irakissa syntyneiden työeläkkeet jäävät keskimäärin pieniksi – ja kansaneläkkeet puolestaan suuriksi – verrattuna Suomessa syntyneeseen kantaväestöön. Koska toteutuneet julkisen talouden vaikutukset riippuvat pakolaisilla voimakkaasti lähtömaasta, Ruist (2018) jakaa pakolaiset kahteen lähtömaaluokkaan: korkeamman integraatiopotentiaalin ("*hög potential*") omaaviin lähtömaihin (Entinen Jugoslavia, Etiopia/Eritrea ja Iran) sekä matalamman integraatiopotentiaalin ("*låg potential*") omaaviin lähtömaihin (Irak ja Somalia). (Ruist 2018, s. 71, 74–76).

Elinkaariennusteiden tulokset. Ruist (2018):n elinkaariennusteiden mukaan matalamman integraatiopotentiaalin lähtömaista tulevien (keskimääräiset) nettovaikutukset kaikkina heidän elinvuosinaan Ruotsissa ovat negatiivisia eli pienempiä kuin 0 kruunua per henkilö per vuosi. Korkeamman integraatiopotentiaalin omaavilla puolestaan on joitakin maassaolovuosia, jolloin heidän nettovaikutuksensa ovat positiivisia. Vaikka korkeamman integraatiopotentiaalin mailla on joitakin maassaolovuosia, jolloin heidän nettovaikutuksensa ovat positiivisia, maassaoloajan alku- ja loppupään (vanhuusikä) negatiivisista nettovaikutuksista johtuen heidän (diskonttaamattomat) elinkaarivaikutuksensa jäävät miinuksen puolelle siten, että keskimäärin 58 Ruotsissa vietetyn vuoden aikana nettovaikutusta karttuu keskimäärin -74 000 Ruotsin kruunua per henkilö vuodessa eli noin -7 000 euroa. Matalamman integraatiopotentiaalin lähtömailla – eli Somalialla ja Irakilla – vastaava luku on -94 000 Ruotsin kruunua per henkilö vuodessa eli noin -9 000 euroa. (Ruist 2018, s. 77, 83–84).

Täten yhteenlaskettu 58 maassaolovuodelta karttunut diskonttaamaton elinaariennuste Somaliassa ja Irakissa syntyneille henkilöille on noin -5,45 miljoonaa Ruotsin kruunua eli noin -520 000 euroa per henkilö. Ruist (2018) huomauttaa, että elinkaariennusteet eivät riipu maahanmuuttovuodesta, mikä tulkitaan raportissa niin, että maahanmuuttajien määrä ei vaikuta integroitumiseen ja siten julkisen talou-

den elinkaarivaikutuksiin. Ruist (2018) esittää elinkaariennusteille myös nykyarvoon diskontatut tulokset kolmella vaihtoehtoisella vuotuisella diskonttauskorolla: 1, 2 ja 3 prosenttia. Näistä yhden prosentin diskonttauskorko on lähimpänä tässä tutkimuksessa käytetyn diskonttauskertoimien ennustejakauman odotusarvoa. Yhden prosentin diskonttauskorolla Somaliasta ja Irakista lähtöisin olevien diskontattu elinkaariennuste on -3,883 miljoonaa Ruotsin kruunua eli noin -370 000 euroa per henkilö. (Ruist 2018, s. 80, 84).

Norja

Norjalainen talouslehti *Finansavisen* julkaisi 7.9.2013 elinkaarivaikutusennusteita eri maahanmuuttajaryhmille Norjassa. Nämä luvut perustuivat Norjan Tilastokeskuksen laatimaan tutkimukseen Holmøy and Strøm 2012. Minulla ei ole toistaiseksi käytössä alkuperäistä Finansavisenin artikkelia, mutta muut norjalaislehdet ovat viitanneet mainittuihin Finansavisenin laskemiin elinkaarivaikutusten ennusteisiin. Finansavisenin kanssa samaan konserniin kuuluva *hagnar.no* lainaa Finansavisen artikkelia, jonka mukaan Syyriassa syntyneiden elinkaarivaikutukset ovat arviolta yhtä suuret kuin Irakissa syntyneiden eli -7 miljoonaa Norjan kruunua per henkilö (Hagnar.no). Toisen norjalaisen lehden – joka myös viittaa Finansavisen tuloksiin – mukaan Somaliassa syntyneiden elinkaarivaikutukset ovat -9 miljoonaa Norjan kruunua per henkilö abcnyheter.no.

Mikäli käytetään nykyisiä vaihtokursseja euron ja Norjan kruunun välillä, niin Irakissa syntyneen elinkaarivaikutus (-7 miljoonaa kruunua) on noin -720 000 euroa (vuoden 2012 kruunuissa). Somaliassa syntyneiden mainittu elinkaarivaikutus olisi samalla tavoin laskettuna noin -930 000 euroa (vuoden 2012 kruunuissa). Tilastokeskuksen rahanarvonkertoimen perusteella nämä luvut olisivat vuoden 2015 euroissa noin -738 000 euroa ja -950 000 euroa.

Tässä tutkimuksessa Irakissa syntyneiden, Suomeen 20–24-muuttavien ja koko iäksi Suomeen jäävien elinkaariennusteiden odotusarvo on -690 000 euroa (ks. taulukko 9.28, s. 9.28). Somaliassa syntyneiden, Suomeen 20–24-muuttavien ja koko iäksi Suomeen jäävien elinkaariennusteiden odotusarvo on puolestaan -951 000 euroa (ks. taulukko 9.7, s. 9.7). Nämä luvut on esitetty ilman lasten vaikutuksia. Ainakin 20–24-vuotiaiden muuttavien elinkaariennusteiden odotusarvot vastaavat hyvin mainittuja, muiden norjalaislehtien Finansavisenilta lainaamia lukuja.

Kirjallisuutta

Odd O. Aalen, Ørnulf Borgan, and Håkon K. Gjessing. *Survival and Event History Analysis. A Process Point of View*. Springer, 2008.

abcnyheter.no. Finansavisen med nytt innvandrerregnskap: – hver somalier koster staten 9 millioner kroner. [verkkojulkaisu]. `https://www.abcnyheter.no/nyheter/2013/09/07/181537/hver-somalier-koster-staten-9-millioner-` [julkaistu: 7.9.2013].

Per Kragh Andersen and Niels Keiding. Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11:91–115, 2002.

Torben M. Andersen, Anna Piil Damm, Olof Åslund, Bernt Bratsberg, Oddbjørn Raaum, Knut Røed, Anders Forslund, Linus Liljeberg, Matti Sarvimäki, Marie Louise Schultz-Nielsen, Hans Grönqvist, Susan Niknami, Kristian Thor Jakobsen, Nicolai Kaarsen, Kristine Vasiljeva, and Joakim Ruist. *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*. TemaNord 2017:520. Nordic Council of Ministers, 2017.

R. B. Arellano-Valle, H. Bolfarine, and V. H. Lachos. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 34(6):663–682, 2007.

Reinaldo B. Arellano-Valle and Marc G. Genton. Multivariate extended skew-t distributions and related families. *Metron, International Journal of Statistics*, 68(3):201–234, 2010.

Adelchi Azzalini. sn. the skew-normal and skew-t distributions. versio: 1.3-0, 11 2015. URL `https://cran.r-project.org/package=sn`.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, Yhdysvallat, 2 edition, 1980.

J. Berkson and R. P. Cage. Survival curves for cancer patients following treatment. *Journal of the American Statistical Association*, 47:501–515, 1952.

J.W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B*, 11:15–44, 1949.

Örn B. Bodvarsson and Hendrik Van den Berg. *The Economics of Immigration*. Springer, New York, Yhdysvallat, 2013.

Bernt Bratsberg, Raaum Oddbjørn, and Knut Røed. Immigrant labor market integration across admission classes. In Torben M. Andersen, Anna Piil Damm, and Olof Åslund, editors, *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*, TemaNord 2017:520, chapter 1, pages 17–54. Nordic Council of Ministers, 2017.

Paul-Christian Buerkner. brms. bayesian regression models using stan. versio: 0.6.0, 2015. URL `http://github.com/paul-buerkner/brms`.

Chao Cai, Yubo Zou, Yingwei Peng, and Jiajia Zhang. smcure. fit semiparametric mixture cure model. versio: 2.0, 9 2012.

Chao Cai, Yingwei Peng, and Jiajia Zhang. smcure: An r-package for estimating semiparametric mixture cure models. Published in final edited form as: Comput Methods Programs Biomed. 2012 December ; 108(3): 1255–1260., 12 2013. NIH Public Access. Author Manuscript, available in PMC 2013 December 01.

Ming-Hui Chen, Joseph G. Ibrahim, and Debajyoti Sinha. A new bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94:909–919, 1999.

Ming-Hui Chen, Joseph G. Ibrahim, and Debajyoti Sinha. Bayesian inference for multivariate survival data with a cure fraction. *Journal of Multivariate Analysis*, 80:101–126, 2002.

Siddhartha Chib. Inference in panel data models via gibbs sampling. In László Mátyás and Patrick Sevestre, editors, *The Econometrics of Panel Data: A Handbook of the Theory with Applications*, Advanced Studies in Theoretical and Applied Econometrics, chapter 24, pages 639–651. Kluwer Academic Publishers, Dordrecht, Alankomaat, 2 edition, 1996.

Robert T. Clemen and Terence Reilly. *Making Hard Decisions with Decision Tools*. South-Western, Cengage Learning, Mason, Ohio, Yhdysvallat, 3 edition, 2014.

Richard J. Cook, John D. Kalbfleisch, and Grace Y. Yi. A generalized mover–stayer model for panel data. *Biostatistics*, 3(3):407–420, 2002.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, New Jersey, Yhdysvallat, 2 edition, 2006.

Mário De Castro, Ming-Hui Chen, Joseph G. Ibrahim, and John P. Klein. Bayesian transformation models for multivariate survival data. *Scandinavian Journal of Statistics*, 41:187–199, 2014.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill Book Company, 1970.

Peter J. Diggle, Kung-Yee Liang, and Scott L. Zeger. *Analysis of Longitudinal Data*. Oxforf Statistical Science Series. Oxford University Press, Oxford, Englanti, 1 edition, 1994.

S. Duane, A.D. Kennedy, Pendleton B.J., and D. Roweth. Hybrid monte carlo. *Physics Letters B*, 195: 216–222, 1987.

ETK. Eläketurvakeskuksen tilastotietokanta [verkkojulkaisu]. `http://tilastot.etk.fi/`. Eläketurvakeskus [viitattu: 27.11.2017].

ETK. *Työeläkkeiden rahoitus (vuodet 2008–2011)*. Eläketurvakeskuksen tilastoraportteja. Eläketurvakeskus, 2009–2012.

ETK. *Työeläkkeiden rahoitus vuonna 2011*. Eläketurvakeskuksen tilastoraportteja 05/2012. Eläketurvakeskus, 2012.

ETK. *Kokonaiseläke 2016. Katsaus työeläkkeen, kansaneläkkeen ja verotuksen määräytymiseen*. Eläketurvakeskuksen katsauksia 01/2016. Eläketurvakeskus, 2016. Katsauksen kirjoittajat: Knuuti, Juha ja Ritola, Suvi.

ETK. Tietopyyntö eläketurvakeskukselle (etk) koskien pitkän aikavälin (2016) eläke-ennusteita. Sähköpostitse viestimistä ETK:n virkamiesten kanssa, 5 2018.

Työeläkelakipalvelu (ETK). Rahoituksen ja kustannustenjaon aikasarjat [verkkojulkaisu]. `http://https://www.tyoelakelakipalvelu.fi/`. Eläketurvakeskus [viitattu: 27.11.2017].

James M. Flegal and Galin L. Jones. Implementing mcmc: Estimating with confidence. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, chapter 7, pages 175–197. Chapman & Hall/CRC, Boca Raton, FL, 2011.

Edward Furman. On a multivariate gamma distribution. *Statistics and Probability Letters*, 78, 2007.

Andrew Gelman and Cosma Rohilla Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66:8–38, 2013.

Andrew Gelman and Kenneth Shirley. Inference from simulations and monitoring convergence. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, chapter 6, pages 163–174. Chapman & Hall/CRC, Boca Raton, FL, 2011.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC texts in statistical science, 3 edition, 2014.

J. Geweke. Bayesian treatment of the independent student-t linear model. *Journal of Applied Econometrics*, 8(S1):S19–S40, 1993.

Charles J. Geyer. Importance sampling, simulated tempering, and umbrella sampling. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, chapter 11, pages 295–311. Chapman & Hall/CRC, Boca Raton, FL, 2011.

W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Journal of the American Statistical Association*, 41(2):337–348, 1992.

Mika Gissler, Maili Malin, Petri Matveinen, Matti Sarvimäki, and Aki Kangasharju. *Maahanmuuttajat ja julkiset palvelut: Terveydenhuollon palvelut ja sosiaalihuollon laitospalvelut.* Työpoliittinen tutkimus 296. Stakes ja VATT, 2006.

Ben Goodrich, Jiqiang Guo Guo, Daniel Lee Lee, Krzysztof Sakrejda, Jonah Gabry, Joel de Guzman, Eric Niebler, Thomas Heller Heller, and John Fletcher. rstan. r interface to stan. versiot: 2.7.0 ja 2.9.0, 9 2015. URL http://mc-stan.org.

Hans Grönqvist and Susan Niknami. The school achievements of refugee children: Lessons from sweden. In Torben M. Andersen, Anna Piil Damm, and Olof Åslund, editors, *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*, TemaNord 2017:520, chapter 5, pages 159–184. Nordic Council of Ministers, 2017.

Yin Guosheng. Bayesian transformation cure frailty models with multivariate failure time data. *Statistics in Medicine*, 27:5929–5940, 2008.

Yin Guosheng and Joseph G. Ibrahim. Cure rate models: a unified approach. *The Canadian Journal of Statistics*, 33(4):559–570, 2005.

Hagnar.no. En syria-flyktning koster norge 7 millioner - 8.000 koster flere titalls mrd. [verkkojulkaisu]. https://www.hegnar.no/Nyheter/Naeringsliv/2015/06/En-Syria-flyktning-koster-Norge-7-millione [julkaistu: 18.6.2015].

Alan Hájek. Interpretations of probability. https://plato.stanford.edu/archives/win2012/entries/probabilit 2012. The Stanford Encyclopedia of Philosophy (Winter 2012 Edition), Edward N. Zalta (ed.). Helsinki: Tilastokeskus [viitattu: 3.4.2018].

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

Erling Holmøy and Birger Strøm. *Makroøkonomi og offentlige fi nanser i ulike scenarioer for innvandring*. Statistisk sentralbyrå, 2012.

Cheng Hsiao. *Analysis of Panel Data*. Cambridge University Press, Cambridge, Englanti, 2 edition, 2003.

Wenbo Hu and Alec N. Kercheval. The skewed t distribution for portfolio credit risk. 6 2006.

Timo Hujanen, Mikko Peltola, Unto Häkkinen, and Markku Pekurinen. *Miesten ja naisten terveysmenot ikäryhmittäin 2006*. Stakesin työpapereita. 37/2008. STAKES (Sosiaali- ja terveysalan tutkimus- ja kehittämiskeskus), 2008.

Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer, New York, Yhdysvallat, 2001.

Markus Jäntti. *Tulonjaon ja verojen kohtaannon kehityksestä Suomessa 1985–2006*. Åbo Akademi, 2009.

John D. Kalbfleisch. *The statistical analysis of failure time data*. Wiley, Hoboken (NJ), 2002.

Kansaneläkelaitos. Eläkkeensaajan asumistuen määrä ja maksaminen [verkkojulkaisu]. https://www.kela.fi/elakkeensaajan-asumistuki-maara-ja-maksaminen. [viitattu: 9.3.2019].

Kansaneläkelaitos. Asumistukimenojen kasvu taittui vuonna 2017 [verkkojulkaisu]. http://urn.fi/URN:NBN:fi-fe201802123338, 2018a. [julkaistu: 14.2.2018].

Kansaneläkelaitos. Eläkkeensaajan asumistuen saajat ja asumismenot [verkkojulkaisu]. Kelasto-raportit | Kela/ Tilasto- ja tietovarastoryhmä/ tilastot@kela.fi / NIT116A, 2018b. [julkaistu: 15.5.2018].

Kauppalehti. Kaksi kolmesta ei maksanut lainkaan yhteisöveroa [verkkojulkaisu]. https://www.kauppalehti.fi/uutiset/kaksi-kolmesta-ei-maksanut-lainkaan-yhteisoveroa/pDqX3bpH 2014. Julkaistu 4.11.2014 [viitattu: 10.5.2017].

Petri Koistinen. *Computational Statistics*. Department of Mathematics and Statistics. University of Helsinki., 2012.

Andrej Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berliini, 1933.

Gary Koop. *Bayesian Econometrics*. John Wiley & Sons Ltd, West Sussex, England, 2003.

Sixten Korkman, Jukka Lassila, Niku Määttänen, and Tarmo Valkonen. *Hyvinvointivaltion rahoitus – Riittävätkö rahat, kuka maksaa?* ETLA, Elinkeinoelämän Tutkimuslaitos. Kustantaja: Taloustieto Oy, 2007.

Hannu Koskinen and Rami Vainio. *Klassinen mekaniikka*. Limes ry, Helsinki, 2010.

Suprateek Kundu and David B. Dunson. Latent factor models for density estimation [verkkojulkaisu]. arXiv:1108.2720v2, 2011. [julkaistu: 19.9.2011].

Tsung I. Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100(2):257–265, 2009.

Helmut Lütkepohl and Markus Krätzig. *Applied Time Series Econometrics*. Cambridge University Press, Cambridge, Englanti, 2004.

Yanyuan Ma and Guosheng Yin. Cure rate model with mismeasured covariates under transformation. *Journal of the American Statistical Association*, 103(482):743–756, 2008.

Maahanmuuttovirasto. Kiintiöpakolaiset [verkkojulkaisu]. http://www.migri.fi/turvapaikka_suomesta/kiintiopa 2017. [viitattu: 23.10.2017].

Ying C. Macnab, Zhenguo Qiu, Paul Gustafson, B. Dean, Charmaine, Arne Ohlsson, and Shoo K. Lee. Hierarchical bayes analysis of multilevel health services data: A canadian neonatal mortality study. *Health Services & Outcomes Research Methodology*, 5:5–26, 2004.

Timo Matala. Kysymyksiä ja vastauksia eläkemaksuista. Sähköpostitse viestimistä Tilastokeskuksen Timo Matalan kanssa, 2 2016.

Antti Mielonen. *Kokonaiseläke 2011. Työeläke, kansaneläke ja verotus*. Eläketurvakeskuksen katsauksia 2011:1. Eläketurvakeskus (ETK), 2011.

Radford M. Neal. Probabibilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 9 1993.

Radford M. Neal. Mcmc using hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, Handbooks of Modern Statistical Methods, chapter 5, pages 113–162. Chapman & Hall/CRC, Boca Raton, FL, 2011.

Council of Economic Advisers. *Long-term Interest Rates: A Survey*. Executive Office of the President of the United States, 2015.

Opetus- ja kulttuuriministeriö. Koulutuksellisen tasa-arvon edistämiseen 22,5 miljoonaa euroa [verkkojulkaisu]. `http://minedu.fi/artikkeli/-/asset_publisher/koulutuksellisen-tasa-arvon-edistamiseen-2` 2013. Opetus- ja kulttuuriministeriö [viitattu: 4.11.2017].

Martyn Plummer, Nicky Best, Kate Cowles, Karen Vines, Deepayan Sarkar, Douglas Bates, and Arni Almond, Russell amd Magnusson. coda. output analysis and diagnostics for mcmc. versio: 0.17-1, 3 2015. URL `https://cran.r-project.org/package=coda`.

Riikka Purra. *Kansainvälinen pakolaisinstituutio ja valtioiden mahdollisuudet. Imagon suojelusta rajojen suojeluun*. Suomen Perusta, 2016.

Marja Riihelä, Risto Vaittinen, and Reijo Vanne. *Changing Patterns of Intergenerational Resource Allocation in Finland*. Reports 2011:1. Finnish Centre for Pensions (Eläketurvakeskus, ETK), 2011.

David Rios Insua, Fabrizio Ruggeri, and Michael P. Wiper. *Bayesian Analysis of Stochastic Process Models*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd, West Sussex, England, 2012.

Ismo Risku. *Yksityisalojen palkansaajien työeläkkeet syntymävuoden ja sukupuolen mukaan*. Eläketurvakeskuksen raportteja 09/2015. Eläketurvakeskus (ETK), 2015.

Ismo Risku, Jukka Appelqvist, Mikko Sankala, Hannu Sihvonen, Heikki Tikanmäki, and Risto Vaittinen. *Lakisääteiset eläkkeet – pitkän aikavälin laskelmat 2013*. Eläketurvakeskuksen raportteja 04/2013. Eläketurvakeskus (ETK), 2013.

Christian P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, New York, Yhdysvallat, 2 edition, 2007.

Paulino Perez Rodriguez. ars. adaptive rejection sampling. versio: 0.4, 3 2009.

Thierry Roncalli and Thibault Lagache. A gauss implementation of skew normal/studentdistributions (sn, st, msn and mst). the skew library. Gauss-ohjelman Skew-kirjaston ohje., 11 2004.

Joakim Ruist. The fiscal cost of refugee immigration: The example of sweden. *Population and Development Review*, 41:567–581, 2015.

Joakim Ruist. The fiscal impact of refugee immigration. In Torben M. Andersen, Anna Piil Damm, and Olof Åslund, editors, *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*, TemaNord 2017:520, chapter 7, pages 211–232. Nordic Council of Ministers, 2017.

Joakim Ruist. *Tid för integration – en ESO-rapport om flyktingars bakgrund och arbetsmarknadsetablering*. Expertgruppen för studier i offentlig ekonomi, 2018:3. Finansdepartment, Regeringskansliet, 2018.

Sujit K. Sahu, Dipak K. Dey, and Márcia D. Branco. A new class of multivariate skew distributions to bayesian regression models. *The Canadian Journal of Statistics*, 31(2):129–150, 2003.

Samuli Salminen. *Maahanmuutot ja Suomen julkinen talous. Osa 1: Toteutuneet julkisen talouden tulot ja menot*. Suomen Perusta, 2015a.

Samuli Salminen. Turvapaikanhakijoiden työllistymisestä [verkkojulkaisu]. `http://www.suomenperusta.fi/ajatus/turvapaikanhakijoiden-tyollistymisesta`, 2015b. [julkaistu: 7.10.2015].

Matti Sarvimäki. Assimilation to a welfare state: Labor market performance and use of social benefits by immigrants to finland. *Scandinavian Journal of Economics*, 113(3):665–688, 2011.

Matti Sarvimäki. Labor merket integration of refugees in finland. In Torben M. Andersen, Anna Piil Damm, and Olof Åslund, editors, *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*, TemaNord 2017:520, chapter 3, pages 91–114. Nordic Council of Ministers, 2017a.

Matti Sarvimäki. *Labor Market Integration of Refugees to Finland.* VATT Research Reports 185. VATT. Valtion taloudellinen tutkimuskeskus., 2017b.

Marie Louise Schultz-Nielsen. Labour market integration of refugees in denmark. In Torben M. Andersen, Anna Piil Damm, and Olof Åslund, editors, *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*, TemaNord 2017:520, chapter 2, pages 55–90. Nordic Council of Ministers, 2017.

Olof Åslund, Anders Forslund, and Linus Liljeberg. Labour market entry of non-labour migrants – swedish evidence. In Torben M. Andersen, Anna Piil Damm, and Olof Åslund, editors, *Nordic Economic Policy Review – Labour Market Integration in the Nordic Countries*, TemaNord 2017:520, chapter 4, pages 115–158. Nordic Council of Ministers, 2017.

Sosiaali- ja terveysministeriö (STM). Sosiaalivakuutusmaksut, maksut 2011–2017 [verkkojulkaisu]. `http://stm.fi/sosiaalivakuutusmaksut`. Sosiaali- ja terveysministeriö [viitattu: 27.11.2017].

Stan Development Team. *Stan Modeling Language. User's Guide and Reference Manual. Stan Version 2.9.0.*, 12 2015. mc-stan.org.

Joseph E. Stiglitz. *Economics of the Public Sector*. W.W. Norton & Company, Inc., New York, Yhdysvallat, 3 edition, 2000.

K. Sumathi and Aruna Rao K. Cure rate models – a partial review with an application to recurrent event or count data. 2008.

Suomen Pankki. Peruskoron muutokset vuodesta 1867 [verkkojulkaisu]. `https://www.suomenpankki.fi/globalassets/fi/tilastot/korot/documents/peruskoron_muutokset_fi` 2017. [viitattu: 21.2.2017].

Brajendra C. Sutradhar. On familial longitudinal poisson mixed models with gamma random effects. *Journal of Multivariate Analysis*, 87(2):398–412, 2003.

Terry M. Therneau and Thomas Lumley. survival. a package for survival analysis in s. versio: 2.38, 2015. URL https://cran.r-project.org/package=survival.

THL. Nuorten syrjäytyminen käy myös yhteiskunnalle kalliiksi [verkkojulkaisu, julkaistu: 11.1.2018]. https://thl.fi/fi/-/nuorten-syrjaytyminen-kay-myos-yhteiskunnalle-kalliiksi, 2018. Terveyden ja hyvinvoinnin laitos (THL) [viitattu: 25.2.2019].

Heikki Tikanmäki, Jukka Appelqvist, Kaarlo Reipas, Mikko Sankala, and Hannu Sihvonen. *Lakisääteiset eläkkeet – pitkän aikavälin laskelmat 2016*. Eläketurvakeskuksen raportteja 08/2016. Eläketurvakeskus (ETK), 2016.

THL Tilasto-ja indikaattoripankki Sotkanet.fi. Ikääntyneiden tavallisen ja tehostetun palveluasumisen sekä vanhainkotien asiakkaat 31.12.2016, kotihoidon asiakkaat 30.11.2016 [verkkojulkaisu]. sotkanet.fi. [viitattu: 1/2019].

Tilastokeskus. Aineistokuvaus. yhdistetty työntekijä-työnantaja-aineisto (fleed), otos, henkilöt. https://tilastokeskus.fi/static/media/uploads/tup/mikroaineistot/, 2 2016. Viitattu 9/2016.

Tilastokeskus. Käsitteet [verkkojulkaisu]. http://www.stat.fi/meta/kas, 2017a. Tilastokeskus [viitattu: 19.2.2017].

Tilastokeskus. Tilastojen kuvaukset: Kuluttajahintaindeksi [verkkojulkaisu]. http://tilastokeskus.fi/meta/til/khi.html, 2017b. Tilastokeskus [viitattu: 19.2.2017].

Pekka Tuominen. *Todennäköisyyslaskenta I*. Limes ry, Kruununhaka, Helsinki, 7 edition, 2004.

Juhani Turkkila. *Tuloverotuksen reaalinen muuttuminen Suomessa: Kuusikymmentä vuotta verojen vuoristorataa*. Sarja A 46. ETLA, Elinkeinoelämän Tutkimuslaitos. Kustantaja: Taloustieto Oy, 2011.

Gerhard Tutz. *Regression for Categorical Data*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, Yhdysvallati, 2012.

Valtioneuvoston kanslia, valtiovarainministeriö. *Hallituksen vuosikertomuksen 2015 liite 2. Tilinpäätöslaskelmat*. Hallituksen julkaisusarja. 5/2016. Valtioneuvoston kanslia, 2016.

Valtioneuvoston kanslia, valtiovarainministeriö. *Hallituksen vuosikertomuksen 2016 liite 2. Tilinpäätöslaskelmat*. Hallituksen julkaisusarja. 3a/2017. Valtioneuvoston kanslia, 2017.

Valtiovarainministeriö. Valtion talousarvioesitykset, 2018 [verkkojulkaisu]. http://budjetti.vm.fi/indox/index.jsp, 2017. Tilastokeskus [julkaistu 19.9.2017, viitattu: 2.11.2017].

VATT. *Maahanmuuttajien integroituminen Suomeen*. VATT-analyysi 1-2014. Valtion taloudellinen tutkimuskeskus, 2014a.

VATT. *Maahanmuuttajien integroituminen Suomeen. Liite VATT-analyysin lukuun 5.* VATT-analyysi 1-2014. Valtion taloudellinen tutkimuskeskus, 2014b.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Efficient implementation of leave-one-out cross-validation and waic for evaluating
tted bayesian models. 9 2015.

Aki Vehtari, Andrew Gelman, Jonah Gabry, Juho Piironen, and Ben Goodrich. loo. efficient leave-one-out cross-validation and waic for bayesian models. versio: 0.1.5, 2 2016. URL https://github.com/jgabry/loo.

Verohallinto. Kuntien ja kuntayhtymien arvonlisäverotusohje [verkkojulkaisu]. https://www.vero.fi/syventavat-vero-ohjeet/ohje-hakusivu/48201/kuntien_ja_kuntayhtymien_arvo a. Verohallinto [viitattu: 2.11.2017].

Verohallinto. Arvonlisäveron vähennys ostoista [verkkojulkaisu]. https://www.vero.fi/yritykset-ja-yhteisot/tietoa-yritysverotuksesta/arvonlisaverotus/ostoihi b. Verohallinto [viitattu: 2.11.2017].

Verohallinto. Arvonlisäverotunnisteen, arvonlisäveronumeron ja nimen merkitseminen valtion virastojen ja laitosten antamiin laskuihin [verkkojulkaisu]. https://www.vero.fi/syventavat-vero-ohjeet/ohje-hakusivu/48121/arvonlisaverotunnisteen_arvon c. Verohallinto [viitattu: 2.11.2017].

Verohallinto. Energiaverotus [verkkojulkaisu]. https://www.vero.fi/syventavat-vero-ohjeet/ohje-hakusivu/56 d. Verohallinto [viitattu: 2.11.2017].

Verohallinto. *Taskutilasto 2013*. 2013.

Veronmaksajain keskusliitto. Kiinteistöveroa kertyy runsaat 1,5 miljardia euroa [verkkojulkaisu]. https://www.veronmaksajat.fi/ajankohtaista/Ajankohtaista/Uutiset/Kiinteistoveroa-kertyy-runs 2014. Tilastokeskus [viitattu: 1.11.2017].

Suomen virallinen tilasto (SVT). Asuntojen vuokrat [verkkojulkaisu]. http://www.stat.fi/til/asvu. ISSN=1798-100X. Helsinki: Tilastokeskus [viitattu: 10.3.2019].

Suomen virallinen tilasto (SVT). Kuluttajahintaindeksi [verkkojulkaisu]. http://www.stat.fi/til/khi, 2015. ISSN=1796-3524=1798-5137. Helsinki: Tilastokeskus [viitattu: 27.3.2015].

Suomen virallinen tilasto (SVT). Muuttoliike [verkkojulkaisu]. http://www.stat.fi/til/muutl, 2016a. ISSN=1797-6766. Helsinki: Tilastokeskus [viitattu: 21.9.2016].

Suomen virallinen tilasto (SVT). Tulonjakotilasto [verkkojulkaisu]. http://www.stat.fi/til/tjt, 2016b. ISSN=1795-8121. Helsinki: Tilastokeskus [viitattu: 21.9.2016].

Suomen virallinen tilasto (SVT). Työvoimatutkimus [verkkojulkaisu]. http://www.stat.fi/til/tyti/, 2016c. ISSN=1798-7830. Helsinki: Tilastokeskus [viitattu: 17.9.2016].

Suomen virallinen tilasto (SVT). Väestöennuste [verkkojulkaisu]. http://www.stat.fi/til/vaenn/, 2016d. ISSN=1798-5137. Helsinki: Tilastokeskus [viitattu: 14.9.2016].

Suomen virallinen tilasto (SVT). Väestörakenne [verkkojulkaisu]. http://www.stat.fi/til/vaerak, 2016e. ISSN=1797-5379. Helsinki: Tilastokeskus [viitattu: 21.9.2016].

Suomen virallinen tilasto (SVT). Ansiotasoindeksi [verkkojulkaisu]. http://www.stat.fi/til/ati, 2017a. ISSN=1796-3737. Helsinki: Tilastokeskus [viitattu: 25.9.2017].

Suomen virallinen tilasto (SVT). Koulutuksen talous [verkkojulkaisu]. http://www.stat.fi/til/kotal, 2017b. ISSN=1799-0947. Helsinki: Tilastokeskus [viitattu: 3.12.2017].

Suomen virallinen tilasto (SVT). Syntyneet [verkkojulkaisu]. http://www.stat.fi/til/synt/, 2017c. ISSN=1798-2391. Helsinki: Tilastokeskus [viitattu: 19.2.2017].

Suomen virallinen tilasto (SVT). Väestörakenne [verkkojulkaisu]. http://www.stat.fi/til/vaerak/2015/01/vaerak_2015_01_2016-09-23_laa_001_fi.html, 2017d. ISSN=1797-5379. Vuosikatsaus 2015, Laatuseloste: Väestörakenne 2015. Helsinki: Tilastokeskus [viitattu: 13.2.2017].

S. Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, and RStudio. ggplot2: Create elegant data visualisations using the grammar of graphics. URL https://cran.r-project.org/package=ggplot2.

Andreas Wienke. *Frailty Models in Survival Analysis*. Chapman & Hall/CRC biostatistics series. CRC Press, Boca Raton, FL, 2011.

A.Y. Yakovlev, B. Asselain, V.J. Bardou, A. Fourquet, T. Hoang, A. Rochefediere, and Tsodikov A.D. A simple stochastic model of tumor recurrence and its applications to data on premenopausal breast cancer. In B. Asselain, M. Boniface, C. Duby, C. Lopez, JP. Masson, and J. Tranchefort, editors, *Biometrie et Analyse de Dormees Spatio-Temporelles 12*, pages 66–82. Société Francaise de Biométrie, ENSA Rennes, France, 1993.

Kelvin K.W. Yau, Andy H. Lee, and Angus S.K. Ng. Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis*, 41: 359–366, 2003.

Yleisradio. Maahanmuuttajanuorilla yli viisinkertainen syrjäytymisriski. http://yle.fi/uutiset/3-6021551, 2012. Uutinen julkaistu: 25.4.2012. [viitattu: 16.2.2017].

Yleisradio. Koulutusvertailun karu fakta: Joka viides nuori suomalaismies ei opiskele ja on ilman työtä. http://yle.fi/uutiset/3-9169450, 2016. Uutinen julkaistu: 15.9.2016. [viitattu: 16.2.2017].

Liite A

Olotilamallinnus: De Castron et al. (2014) kehittämän posteriorijakauman yksiulotteisen tapauksen simulointi Gibbsin otannalla.

```
##########################################################
# R-ohjelman koodi:
# De Castron et al. (2014) kehittämän posteriorijakauman
# yksiulotteisen tapauksen simulointi Gibbsin otannalla.
##########################################################

# Sisältö:

# 1. Tarvittavat R-paketit (asennettava aluksi).
# 2. Selviytymisaika-aineisto paranemisosuudella.

# Valmistelu.
# 3. Määritetään intervallit.
# 4. Selittävät muuttujat ja regressiokertoimet.
# 5. Estimoitavien parametrien priorijakaumat.

# Gibbsin otanta.
# 6. Parametrien alkuarvot Gibbsin otantaan.
# 7. Poiminnat ehdollisista jakaumista.
# 8. Kiinnostavia parametreja koskevat posterioripäättelyt.

# Posterioriennusteet.
# 9. Posterioriennusteiden valmistelu.
# 10. Posterioriennusteiden simulointi.
# 11. Posterioriennustetarkistukset.

#####


# 1. Tarvittavat R-paketit (asennettava aluksi).

# survival-paketti selviytymisfunktioiden Kaplan-Meier-estimaatteja varten.
# ars-paketti Adaptive Rejection Sampling -otantaa varten Gibbsin otannassa.
# coda-paketti posterioripäättelyistä varten.
# smcure-paketista mallinnusaineisto.

library(survival)       # (Therneau and Lumley 2015).
library(ars)            # (Rodriguez 2009).
library(coda)           # (Plummer et al. 2015).

library(smcure)         # (Cai et al. 2012).

#####
```

```r
# 2. Selviytymisaika-aineisto paranemisosuudella.

# 2.1 Aineisto.
data(e1684)

# Poistetaan puuttuvia arvoja sisältävä havainto.
# Huom! Poistaminen toimintatapana ei yleisesti hyvä.
aineisto <- na.omit(e1684)
attach(aineisto)

# 2.2 Aineistontarkastelua.

# Havaintomäärä
N <- nrow(aineisto)
N

# Yhteenveto aineistosta:
summary(aineisto)
# FAILTIME kertoo tapahtuman ajan.
# FAILCENS on indikaattori, joka kertoo sensoroitiinko havainto.
# TRT, AGE ja SEX ovat selittäviä muuttujia.

# Kaplan-Meier-estimaatit (K-M-estimaatit).
# Poikkiviivat ovat sensorointiaikoja.

# Koko aineisto:
km_koko = survfit(Surv(FAILTIME,FAILCENS)~1,data=aineisto)
plot(km_koko)

# Selittävän muuttujan arvojen mukaan:

# TRT
km_trt <- survfit(Surv(FAILTIME,FAILCENS)~TRT,data=aineisto)
plot(km_trt, col=1:2)
legend("topright", c("TRT = 0","TRT = 1"),lty=c(1,1),col=c(1,2))

# AGE
km_age <-
   survfit(Surv(FAILTIME,FAILCENS)~cut(aineisto$AGE,c(-35,-15,0,15,35)),
   data=aineisto)
plot(km_age, col=1:4)
legend("topright",
```

```
   levels(cut(aineisto$AGE,c(-35,-15,0,15,35))),lty=rep(1,4),col=1:4)

# SEX
km_sex <- survfit(Surv(FAILTIME,FAILCENS)~SEX,data=aineisto)
plot(km_sex, col=1:2)
legend("topright", c("SEX = 0","SEX = 1"),lty=c(1,1),col=c(1,2))

##########################################################################

#################
# Valmistelu.
#################

# 3. Määritetään intervallit
# (ks. luvut 3.2.2.7 ja 3.2.3.2).

# Intervalleja yhteensä J kappaletta
J <- 6

# Intervallien rajat:
FAILRAJA <- rep(0,J+1)
FAILRAJA[1] <- 0
FAILRAJA[2] <- 0.3
FAILRAJA[3] <- 0.6
FAILRAJA[4] <- 2.2
FAILRAJA[5] <- 4.5
# Sensoroimattoman tapahtuman maksimiaika
FAILRAJA[6] <- max(1*(aineisto$FAILCENS==1)*FAILTIME)
FAILRAJA[7] <- 100

# Intervallien indikaattorit havainnoille, J kpl:
Fail_1 <- 1*(FAILTIME<0.3)
Fail_2 <- 1*(FAILTIME>=0.3 & FAILTIME<0.6)
Fail_3 <- 1*(FAILTIME>=0.6 & FAILTIME<2.2)
Fail_4 <- 1*(FAILTIME>=2.2 & FAILTIME<4.5)
Fail_5 <- 1*(FAILTIME>=4.5 & FAILTIME<=8.27)
Fail_6 <- 1*(FAILTIME>8.27)

#####

# 4. Selittävät muuttujat ja regressiokertoimet.
# Selittäjämatriisi, keskistetään indikaattorimuuttujat:
X <- cbind(AGE,SEX-mean(SEX),TRT-mean(TRT))
Betat_lkm <- ncol(X)
```

```
#####

# 5. Estimoitavien parametrien priorijakaumat.
# (ks. luku 3.2.4 ja De Castro et al. 2014, 191--192).

# Betoille epäsopivat priorit ~ 1.

# Theta, IG-priorijakauma:
 gamma_1 <- 1
 gamma_2 <- 1

# r, IG-priorijakaumat:
 gamma_3 <- 1
 gamma_4 <- 1

# lambda_j, gamma-priorijakaumat:
 gamma_5 <- c(rep(0,J-1),1)
 gamma_6 <- c(rep(0,J-1),1)

##########################################################################

#####################
# Gibbsin otanta.
#####################

# Simulaatioiden määrä
S <- 30000
# Kun simulaatiomäärä S = 30000, simulaatioissa kuluu n. 10 min.

# 6. Parametrien alkuarvot Gibbsin otantaan.

# 6.1 Lambda_j
lambda_1.initial <- 0.1
lambda_2.initial <- 0.1
lambda_3.initial <- 0.01
lambda_4.initial <- 0.01
lambda_5.initial <- 0.01
lambda_6.initial <- 0.01

# 6.2 Z_i
Z.initial <- rep(0.05,N)

# 6.3 Betat
```

```
Beta.initial <- rep(0.5,Betat_lkm)

# 6.4 u_i
U.initial <- rep(1,N)

# 6.5 eeta = theta^-1
eeta.initial <- 3

# 6.6 ksii = log(r*), r* = r/theta
ksii.initial <- 1


# Parametreille simulaatiovektorit

# Tapahtumaintervallit
FAIL_ALUE <- numeric(N)
for (i in 1:N) {
FAIL_ALUE[i] <- sum(1*(FAILTIME[i]>FAILRAJA))
}

# 1. Lambda_j
 # lambda_j
 lambda <- matrix(rep(0,S+1),nrow=J,ncol=S+1)
 lambda[1,1] <- lambda_1.initial
 lambda[2,1] <- lambda_2.initial
 lambda[3,1] <- lambda_3.initial
 lambda[4,1] <- lambda_4.initial
 lambda[5,1] <- lambda_5.initial
 lambda[6,1] <- lambda_6.initial
 lambda.nyk <- lambda[,1]

# 2. Zeta_i
 zeta <- matrix(rep(0,S+1),nrow=N,ncol=S+1)
 zeta[,1] <- Z.initial
 zeta.nyk <- zeta[,1]

# 3. Betat
 beta <- matrix(rep(0,S+1),nrow=Betat_lkm,ncol=S+1)
 beta[,1] <- Beta.initial
 beta.nyk <- beta[,1]

# 4. U_i
 U_i <- matrix(rep(0,S+1),nrow=N,ncol=S+1)
 U_i[,1] <- U.initial
```

```
 U_i.nyk <- U_i[,1]

# 5. eeta
 eeta <- matrix(rep(0,S+1),nrow=1,ncol=S+1)
 eeta[,1] <- eeta.initial
 eeta.nyk <- eeta[,1]

# 6. ksii
 ksii <- matrix(rep(0,S+1),nrow=1,ncol=S+1)
 ksii[,1] <- ksii.initial
 ksii.nyk <- ksii[,1]

# 7. Paranemisosuus rhoo
 rhoo <- matrix(rep(0,S+1),nrow=N,ncol=S+1)


##########################################################################

#########################################
# 7. Poiminnat ehdollisista jakaumista.

#########################################

# Ehdolliset jakaumat on esitetty De Castron et al. (2014)
# artikkelin liitteessä B.

#################################################################################
# Poimintakoodin ALKU on tässä.
# VALITSE JA AJA TÄSTÄ kerralla siihen asti, kunnes tulee POIMINTAKOODIN LOPPU.
# S on edellä määritetty simulaatioiden lukumäärä.
#################################################################################

#################################################################################
# Huom! Varsinkin simulaation alkuvaiheessa seuraavia virheilmoituksia esintyy:
#
# Error in sobroutine initial_...
# ifault= 4
# Error in sobroutine initial_...
# ifault= 3
#
# Tämä johtuu siitä, että parametrien nykyiset arvot ovat vielä
# kaukana ars-otannan määritellyistä aloituspisteistä.
# Ks. lisätietoja: ?ars
#################################################################################
```

```
# (Ajoaika talteen)
kello_alku <- proc.time()
for (s in 1:S) {

# 0. Hazardin kertymäfunktio, Ht ( ks. De Castro et al. 2014, Liite B).

apusumma <- numeric(N)
for (i in 1:N) {
apusumma[i] <- sum(c(0,lambda.nyk[1:J-1]*(FAILRAJA[2:J] -
   FAILRAJA[1:J-1]))[1:FAIL_ALUE[i]])
}

Ht <- lambda.nyk[FAIL_ALUE] *
   ((FAILTIME-FAILRAJA[FAIL_ALUE])^(1-get(paste('Fail_',J,sep=")))) *
((1-exp(-FAILTIME+FAILRAJA[J]))^get(paste('Fail_',J,sep=''))) +
apusumma

# 1. Lambdat

for (j in 1:J) {
vali_indik <- get(paste('Fail_',j,sep=''))
loppu_indik <- 1*(FAILTIME>FAILRAJA[j+1])
A <- zeta.nyk*(FAILTIME - FAILRAJA[j])^(1-get(paste('Fail_',J,sep='')))
B <- ((1-exp(-FAILTIME+FAILRAJA[J]))^get(paste('Fail_',J,sep=''))) *
   exp(X%*%beta.nyk)
C <- zeta.nyk*(FAILRAJA[j+1]-FAILRAJA[j]) * exp(X%*%beta.nyk)

shape <- gamma_5[j] + sum(get(paste('Fail_',j,sep=''))*FAILCENS)
rate <- gamma_6[j]/((eeta.nyk^-1)*exp(ksii.nyk)) + sum(vali_indik*A*B) +
   sum(loppu_indik*C)

lambda.nyk[j] <- rgamma(n=1,shape=shape,rate=rate)
}

# 2. Zetat

# zeta_i:t
for (i in 1:N) {
shape_z_i <- FAILCENS[i] + U_i.nyk[i]/exp(ksii.nyk)
rate_z_i <- as.numeric(1 + (exp(X[i,]%*%beta.nyk) * Ht[i]))

zeta.nyk[i] <- rgamma(n = 1, shape = shape_z_i, rate = rate_z_i)
}
```

```
# 3. Betat

# betat_p
for (p in 1:Betat_lkm)  {

f <- function(beta) {sum((X[,p]*rep(beta,N) +
   X[,-p]%*%as.matrix(beta.nyk[-p]))*FAILCENS) +
sum((-FAILCENS-U_i.nyk/exp(ksii.nyk)) * log(1 + Ht * exp(X[,p]*rep(beta,N) +
   X[,-p]%*%as.matrix(beta.nyk[-p]))))}

# ars vaatii ehdollisen jakauman derivaatan määrittämistä.
fpilkku <- function(beta) { sum(X[,p]*FAILCENS) +
   sum(((-FAILCENS*Ht*X[,p]*exp(X[,p]*rep(beta,N) +
   X[,-p]%*%as.matrix(beta.nyk[-p])))-
(U_i.nyk/exp(ksii.nyk)*Ht*X[,p]*exp(X[,p]*rep(beta,N) +
   X[,-p]%*%as.matrix(beta.nyk[-p]))))  / (1 + Ht*exp(X[,p]*rep(beta,N) +
   X[,-p]%*%as.matrix(beta.nyk[-p]))))}

f_vec <- Vectorize(f)
fpilkku_vec <- Vectorize(fpilkku)

beta.nyk[p] <- ars(1,f_vec,fpilkku_vec,x=c(-7,2.5,7))
}

# 4. U:t

# U_i:t
for (i in 1:N) {

shape_u_i <- eeta.nyk + FAILCENS[i]
rate_u_i <- 1 + log(1 + Ht[i]*exp(X[i,]%*%beta.nyk))/exp(ksii.nyk)
U_i.nyk[i] <- rgamma(n = 1, shape = shape_u_i, rate = rate_u_i)
}

# 5. Eeta (theta^-1)

f <- function(eeta) { (gamma_1 + sum(gamma_3) + sum(gamma_5) - 1) * log(eeta) -
eeta*(gamma_2 + sum(gamma_4)/exp(ksii.nyk) +
   sum(gamma_6)*sum(lambda.nyk)/exp(ksii.nyk)) -
eeta*sum(log(1 + log(1+Ht*exp(X%*%beta.nyk))/exp(ksii.nyk))) +
sum(log(rep(eeta,N)^FAILCENS)) }

# ars vaatii ehdollisen jakauman derivaatan määrittämistä.
```

```
fpilkku <- function(eeta) { ((gamma_1 + sum(gamma_3) +
sum(gamma_5) - 1) / eeta) - (gamma_2 + sum(gamma_4)/exp(ksii.nyk) +
sum(gamma_6)*sum(lambda.nyk)/exp(ksii.nyk)) -
sum(log(1 + log(1+Ht*exp(X%*%beta.nyk))/exp(ksii.nyk))) +
sum((1/eeta)*FAILCENS) }

f_vec <- Vectorize(f)
fpilkku_vec <- Vectorize(fpilkku)

eeta.nyk <- ars(1,f_vec,fpilkku_vec,x=c(0.01,2.5,10),lb=TRUE,xlb=0)

# 6. Ksii (log(rt))

f <- function(ksii) {-ksii*(sum(gamma_3) + sum(gamma_5) + sum(FAILCENS)) -
exp(-ksii)*eeta.nyk*(gamma_4+sum(gamma_6%*%lambda.nyk)) +
sum((-eeta.nyk-FAILCENS)*log(1 +
   exp(rep(-ksii,N))*log(1+Ht*exp(X%*%beta.nyk))))}

# ars vaatii ehdollisen jakauman derivaatan määrittämistä.
fpilkku <- function(ksii) {-(sum(gamma_3) + sum(gamma_5) + sum(FAILCENS)) +
exp(-ksii)*eeta.nyk*(gamma_4+sum(gamma_6%*%lambda.nyk)) +
sum((-eeta.nyk-FAILCENS)*
(-exp(rep(-ksii,N))*log(1+Ht*exp(X%*%beta.nyk)))/
(1 + exp(rep(-ksii,N))*log(1+Ht*exp(X%*%beta.nyk))))}

f_vec <- Vectorize(f)
fpilkku_vec <- Vectorize(fpilkku)

ksii.nyk <- ars(1,f_vec,fpilkku_vec,x=c(-15,-8,3))

# 7. Paranemisosuus rhoo (De Castro et al. 2014, 191).

r <- exp(ksii.nyk)/eeta.nyk
beta_0_inf <- log(sum((lambda.nyk/r)[1:J-1]*(FAILRAJA[2:J] - FAILRAJA[1:J-1]))
   + (lambda.nyk/r)[J])
rhoo.nyk <- 1/((1 + (eeta.nyk^-1) * (log(1 + (exp(ksii.nyk) / eeta.nyk) *
   exp(beta_0_inf + X%*%beta.nyk)))/(exp(ksii.nyk) / eeta.nyk))^eeta.nyk)

#########################################
# Nykyiset parametrien arvot talteen.
#########################################

# 1. Lambdat
lambda[,s+1] <- lambda.nyk
```

```
# 2. Zetat
zeta[,s+1] <- zeta.nyk

# 3. Betat
beta[,s+1] <- beta.nyk

# 4. U_i:t
U_i[,s+1] <- U_i.nyk

# 5. eeta
eeta[,s+1] <- eeta.nyk

# 6. ksii
ksii[,s+1] <- ksii.nyk

# 7. rhoo
rhoo[,s+1] <- rhoo.nyk

}
# (ajoaika)
ajoaika <- proc.time() - kello_alku
ajoaika[3]/60   # Ajoaika minuutteina.
ajoaika[3]/S    # Ajoaika sekunteina per simulaatio.

##########################################################
# POIMINTAKOODIN LOPPU ON TÄSSÄ.
# VALITSE JA AJA POMINTAKOODIN ALUSTA TÄHÄN ASTI.
##########################################################

#######################################################################

# 8. Kiinnostavia parametreja koskevat posterioripäättelyt.

# Ei käytetä lämmittelyvaiheen (warmup) simulaatioita päättelyssä.
warmup <- 2000

# "Time-series SE" on Monte Carlo -keskivirhe
# parametrien odotusarvoestimaateille.

# 8.1 Lambdat

lambda.mcmc <- mcmc(t(lambda[,-(1:warmup)]))
summary(lambda.mcmc)
```

```
plot(lambda.mcmc)

# 8.2 Betat

beta.mcmc <- mcmc(t(beta[,-(1:warmup)]))
summary(beta.mcmc)
plot(beta.mcmc)

# 8.3 eeta
eeta.mcmc <- mcmc(eeta[-(1:warmup)])
summary(eeta.mcmc)
plot(eeta.mcmc)

# 8.4 ksii
ksii.mcmc <- mcmc(ksii[-(1:warmup)])
summary(ksii.mcmc)
plot(ksii.mcmc)

# 8.5 parenemisosuuden rhoo keskiarvo
rhoo_ka.mcmc <- mcmc(apply(rhoo[,-(1:warmup)],2,mean))
summary(rhoo_ka.mcmc)
plot(rhoo_ka.mcmc)

##########################################################################
##########################################################################

#######################
# Posteriorienusteet.
#######################
# (Ks. luku 3.2.6.3.)

# 9. Posteriorienusteiden valmistelu.

# 9.1 Ennusteintervallit

Ennusteajat <- seq(0.02,10.2,0.02)
ennustelkm <- length(Ennusteajat)

# Ennusteiden intervallit
Ennuste_ALUE <- numeric(ennustelkm)
for (i in 1:(ennustelkm)) {
Ennuste_ALUE[i] <- sum(1*(Ennusteajat[i]>FAILRAJA))
}
```

```
#  Mallinnusintervallit ennusteille, J kpl:
Ennuste_1 <- 1*(Ennusteajat<0.3)
Ennuste_2 <- 1*(Ennusteajat>=0.3 & Ennusteajat<0.6)
Ennuste_3 <- 1*(Ennusteajat>=0.6 & Ennusteajat<2.2)
Ennuste_4 <- 1*(Ennusteajat>=2.2 & Ennusteajat<4.5)
Ennuste_5 <- 1*(Ennusteajat>=4.5 & Ennusteajat<=8.27)
Ennuste_6 <- 1*(Ennusteajat>8.27)

# 9.2 Parametrien posteriorisimulaatiot

# Ennusteiden nopeuttamiseksi otetaan mukaan joka 50. simulaatio
thinning_interval <- 50
thinning_S <- (S/thinning_interval)+1

# Ei käytetä posterioriennusteissa ensimmäistä 30 %:ia simulaatioista.
thinning_warmup <- round(0.3*thinning_S)

# Ennusteissa käytettävät simulaatiot
kaytett_sim <- thinning_S - thinning_warmup + 1

thin_lambda <- lambda[,seq(1, ncol(lambda),thinning_interval)]
thin_beta <- beta[,seq(1, ncol(beta),thinning_interval)]
thin_eeta <- eeta[,seq(1, ncol(eeta),thinning_interval)]
thin_ksii <- ksii[,seq(1, ncol(ksii),thinning_interval)]
thin_rhoo <- rhoo[,seq(1, ncol(rhoo),thinning_interval)]

eeta.post <- c(thin_eeta[thinning_warmup:thinning_S])
ksii.post <- c(thin_ksii[thinning_warmup:thinning_S])
lambda.post <- thin_lambda[,thinning_warmup:thinning_S]
beta.post <- cbind(thin_beta[,thinning_warmup:thinning_S])

# 9.3 Ennusteajoille ja -sensoroinneille matriisit.
# Sensorointi tapahtuu, jos "selviää"
# koko seuranta-ajan (10,2 aikayksikköä) yli
# ilman tapahtumaa.

FAILCENS_ennuste <- matrix(rep(0,kaytett_sim),nrow=N,ncol=kaytett_sim)
FAILTIME_ennuste <- matrix(rep(0,kaytett_sim),nrow=N,ncol=kaytett_sim)

#####

# 10. Posterioriennusteiden simulointi.

# ALKU. AJA SIIHEN ASTI KUNNES "LOPPU".
```

```
for (e in 1:(kaytett_sim)) {

r.post <- exp(ksii.post[e])/eeta.post[e]

# Hazardin kertymäfunktio ennustekohdille: Ht_enn

ennuste_apusumma <- numeric(ennustelkm)
for (i in 1:ennustelkm) {
ennuste_apusumma[i] <- sum(c(0,(lambda.post[1:J-1,e]/r.post)*(FAILRAJA[2:J] -
   FAILRAJA[1:J-1]))[1:Ennuste_ALUE[i]])
}

Ht_enn <- (lambda.post[Ennuste_ALUE,e] / r.post) * ((Ennusteajat -
   FAILRAJA[Ennuste_ALUE])^(1-get(paste('Ennuste_',J,sep=""))))*
((1 - exp(-Ennusteajat + FAILRAJA[J]))^get(paste('Ennuste_',J,sep=""))) +
ennuste_apusumma

# Hazardin kertymäfunktio suurimmalle ajalle (= seurannan loppu): Ht_loppu
# maksimiaika eli seurannan loppu:
tau <- 10.2
Ht_loppu <-
   (lambda.post[J,e]/r.post)*(1-exp(-tau+FAILRAJA[J]))+max(ennuste_apusumma)
beta_0 <- log(Ht_loppu)

# Jokaiselle henkilölle ennuste tämän muuttointervallista:

   for (i in 1:N)       {

   # Survival alusta (t = 0, S = 1) ennusteajankohtaan,
   # (De Castro et al. 2014, lausekkeeseen 2 pohjautuen).

   # W_i:n (frailty) poiminta gamma(eeta,eeta)-jakaumasta kullekin hlölle
   w_i.post <- rgamma(1, shape=eeta.post[e],rate=eeta.post[e])

   # Alusta t:hen Survival
   Surv_ennuste_0_t <- (1 + r.post*(Ht_enn/Ht_loppu) * c(exp(beta_0 +
    X[i,]%*%beta.post[,e])))^(-w_i.post/r.post)

   # Muuttointervalliarvonta
   arpa <- runif(1)
   muuttointervalli <- sum(1*(arpa<Surv_ennuste_0_t))+1

        if (muuttointervalli==ennustelkm+1) {
                FAILCENS_ennuste[i,e] <- 0
```

```
                FAILTIME_ennuste[i,e] <- tau
        } else {
                FAILCENS_ennuste[i,e] <- 1
                FAILTIME_ennuste[i,e] <- Ennusteajat[muuttointervalli]
        }
   }
}

FAILCENS_ennusteet <- as.vector(FAILCENS_ennuste)
FAILTIME_ennusteet <- as.vector(FAILTIME_ennuste)

ennusteet <- as.data.frame(cbind(FAILTIME_ennusteet,FAILCENS_ennusteet))

##################################
# Posteriorienustesimulaatiot.
# LOPPU.
##################################

#####################################

# 11. Posteriorienustetarkistukset.

# 11.1 Koko aineisto/kaikki ennusteet erittelemättä

km_aineisto_koko <- survfit(Surv(FAILTIME,FAILCENS)~1,data=aineisto)
km_ennuste_koko <-
   survfit(Surv(FAILTIME_ennusteet,FAILCENS_ennusteet)~1,data=ennusteet)
plot(km_aineisto_koko)
lines(km_ennuste_koko$time,km_ennuste_koko$surv, lty=2,col="red")
legend("topright",
   c("Havaintoaineisto","Posteriorienusteet"),lty=c(1,2),col=c(1,2))

# 11.2 Selittävien muuttujien mukaan
post.ennusteet <-
   as.data.frame(cbind(aineisto,FAILCENS_ennuste,FAILTIME_ennuste))

# TRT
ennusteet.trt.0.failcens <-
   as.vector(t(post.ennusteet[which(post.ennusteet$TRT==0) ,
   6:(kaytett_sim+5)]))
ennusteet.trt.0.failtime <-
   as.vector(t(post.ennusteet[which(post.ennusteet$TRT==0) ,
   (kaytett_sim+6):(2*kaytett_sim+5)]))
ennusteet.trt.0 <- as.data.frame(cbind(ennusteet.trt.0.failtime,
```

```
    ennusteet.trt.0.failcens))

ennusteet.trt.1.failcens <-
    as.vector(t(post.ennusteet[which(post.ennusteet$TRT==1) ,
    6:(kaytett_sim+5)]))
ennusteet.trt.1.failtime <-
    as.vector(t(post.ennusteet[which(post.ennusteet$TRT==1) ,
    (kaytett_sim+6):(2*kaytett_sim+5)]))
ennusteet.trt.1 <- as.data.frame(cbind(ennusteet.trt.1.failtime,
    ennusteet.trt.1.failcens))

aineisto.km_trt <- survfit(Surv(FAILTIME,FAILCENS)~TRT,data=aineisto)
ennusteet.km_trt.0 <- survfit(Surv(ennusteet.trt.0.failtime,
    ennusteet.trt.0.failcens) ~ 1 , data = ennusteet.trt.0)
ennusteet.km_trt.1 <- survfit(Surv(ennusteet.trt.1.failtime,
    ennusteet.trt.1.failcens) ~ 1 , data = ennusteet.trt.1)

plot(km_trt, col=1:2)
lines(ennusteet.km_trt.0$time,ennusteet.km_trt.0$surv, lty=2,col=1)
lines(ennusteet.km_trt.1$time,ennusteet.km_trt.1$surv, lty=2,col=2)
legend("topright", c("Havaintoaineisto, TRT = 0","Havaintoaineisto, TRT = 1",
"Posterioriennusteet, TRT = 0","Posterioriennusteet, TRT =
    1"),lty=c(1,1,2,2),col=c(1,2,1,2))

# Havaintojen ja posterioriennusteiden mukaan TRT = 1 -ryhmällä paranemisosuus
# on suurempi kuin verrokkiryhmällä TRT = 0.
# Posterioriennusteet eivät aivan täysin tavoita TRT = 0 -ryhmän
# selviytymisfunktion muotoa K-M-kuvioin perusteella.

# SEX
ennusteet.sex.0.failcens <-
    as.vector(t(post.ennusteet[which(post.ennusteet$SEX==0) ,
    6:(kaytett_sim+5)]))
ennusteet.sex.0.failtime <-
    as.vector(t(post.ennusteet[which(post.ennusteet$SEX==0) ,
    (kaytett_sim+6):(2*kaytett_sim+5)]))
ennusteet.sex.0 <- as.data.frame(cbind(ennusteet.sex.0.failtime,
    ennusteet.sex.0.failcens))

ennusteet.sex.1.failcens <-
    as.vector(t(post.ennusteet[which(post.ennusteet$SEX==1) ,
    6:(kaytett_sim+5)]))
ennusteet.sex.1.failtime <-
    as.vector(t(post.ennusteet[which(post.ennusteet$SEX==1) ,
```

```
    (kaytett_sim+6):(2*kaytett_sim+5)]))
ennusteet.sex.1 <-
    as.data.frame(cbind(ennusteet.sex.1.failtime,ennusteet.sex.1.failcens))

aineisto.km_sex <- survfit(Surv(FAILTIME,FAILCENS)~SEX,data=aineisto)
ennusteet.km_sex.0 <-
    survfit(Surv(ennusteet.sex.0.failtime,ennusteet.sex.0.failcens) ~ 1 ,
    data=ennusteet.sex.0)
ennusteet.km_sex.1 <-
    survfit(Surv(ennusteet.sex.1.failtime,ennusteet.sex.1.failcens) ~ 1 ,
    data=ennusteet.sex.1)

plot(km_sex, col=1:2)
lines(ennusteet.km_sex.0$time,ennusteet.km_sex.0$surv, lty=2,col=1)
lines(ennusteet.km_sex.1$time,ennusteet.km_sex.1$surv, lty=2,col=2)
legend("topright", c("Havaintoaineisto, SEX = 0","Havaintoaineisto, SEX = 1",
"Posterioriennusteet, SEX = 0","Posterioriennusteet, SEX =
    1"),lty=c(1,1,2,2),col=c(1,2,1,2))

#########################################################
# R-ohjelman koodi:
# De Castron et al. (2014) kehittämän posteriorijakauman
# yksiulotteisen tapauksen simulointi Gibbsin otannalla.
# LOPPU.
#########################################################
```

Liite B

Vinon Studentin t-jakauman simulointi Stan-ohjelmalla. Nettovaikutusmallinnuksessa käytetyt Stan- ja BRMS-koodit.

```
#######################################################
# R-ohjelman koodi:
# Vinon Studentin t-jakauman simulointi Stan-ohjelmalla.
#######################################################


# Ks. vinosta t-jakaumasta luku 4.3.3.5.

# Sisältö:

# 1. Tarvittavat R-paketit (asennettava aluksi).
# 2. Poimitaan otos havaintoja vinosta t-jakaumasta.
# 3. Aineisto Stan-ohjelmaa varten.
# 4. Stan-ohjelman koodi.
# 5. Apumuuttujan W alkuarvot simulaatioihin.
# 6. Simulaatioketjuja koskevat valinnat.
# 7. Simulointi Stan-ohjelmalla.
# 8. Posterioriennustejakauma estimoiduilla parametreilla.
# 9. Posterioriennustetarkistus:
# posterioriennusteet vs. alkuperäinen havaintoaineisto (vino_t).

# 10. Vertailu suhteessa normaalijakaumaoletukseen
# ja ei-vinoon t-jakaumaoletukseen.


#####

# 1. Tarvittavat R-paketit (asennettava aluksi).

# sn:llä saadaan poimittua havaintoja vinosta t-jakaumasta.
# rstan on Stanin R-ohjelman versio.
# loo ja brms mallivertailua varten (kohta 10.).

library(sn)             # (Azzalini 2015).
library(rstan)          # (Goodrich et al. 2015).

library(brms)           # (Buerkner 2015).
library(loo)            # (Vehtari et al. 2016).

#####

# 2. Poimitaan N kappaletta havaintoja vinosta t-jakaumasta.
# Huom! "Iso" N kasvattaa simulointiin tarvittavaa aikaa.
```

```
# xi määrää jakauman sijaintia, omega jakauman hajontaa,
# alpha vinoutta ja
# nu t-jakauman vapausasteita.

N <- 2000
vino_t <- rst(n=N,xi=0,omega=1.5,alpha=-1,nu=2.5)

hist(vino_t, breaks=50, xlim=c(-15,15))

#####

# 3. Aineisto Stan-ohjelmaa varten:

Y <- vino_t[1:N]
N <- N
X <- rep(1,N)

#####

# 4. Stan-ohjelman koodi:

stan_koodi_vino_t <- '
functions {
}
data {
  int<lower=1> N;
  vector[N] Y;
  vector[N] X;
}
transformed data {
}
parameters {

  real b_Intercept;  // Vakiotermi.

  real<lower=0> sigma;  // Havaintojen hajontaparametri.

  real g ; // t-jakauman vinousparametri.

    vector<lower=0>[N] W; // Apumuuttuja, ks. s. 154.

  real<lower=0> v; // t-jakauman vapausasteparametri.

}
```

```
transformed parameters {

  vector[N] myy ;
  vector<lower=0>[N] vino_sigma;

  // W:llä ehdollisen normaalijakauman odotusarvo,
  // vrt. lauseke 4.13, s. 154:
  myy = (b_Intercept * X) + (g * W);

  // W:llä ehdollisen normaalijakauman keskihajonta,
  // vrt. lauseke 4.13, s. 154:
  for (n in 1:N) {
  vino_sigma[n] = sigma * sqrt(W[n]);
  }

}

model {

  // Priorijakaumat (heikosti informatiivisia koskien parametreja)
  b_Intercept ~ normal(0, 10);

  sigma ~ cauchy(0, 10);
  g ~ normal(0, 10);

  // Otosjakauman vapausasteet:
  v ~ cauchy(0, 10);

  // Apumuuttujalla W on käänteisgamma-jakauma:
  W ~ inv_gamma(v/2, v/2);

  // W:llä ehdollinen normaalijakautunut otosjakauma,
  // vrt. lauseke 4.13, s. 154:
  Y ~ normal(myy, vino_sigma);

}

// Log-uskottavuusfunktion arvot talteen LOO:n ja WAIC:n laskemista varten.
generated quantities {
  vector[N] log_lik;
    for (n in 1:N) {
   log_lik[n] = normal_lpdf(Y[n]| myy[n], vino_sigma[n]);
   }
}
```

```
,

#####

# 5. Apumuuttujan W alkuarvot simulaatioihin.
# Koska W positiivinen, tarvitsee määrittää erikseen.

init_W <- function() {list(W=runif(N,0.01,2))}
init_W()

#####

# 6. Simulaatioketjuja koskevat valinnat.

simulaatiot_yht <- 2000
simulaatiot_warmup <- 500
simulaatioketjut <- 2
kaytettavat_simulaatiot <-
   (simulaatiot_yht-simulaatiot_warmup)*simulaatioketjut
kaytettavat_simulaatiot

#####

# 7. Simulointi Stan-ohjelmalla.
# Simulaation käynnistyminen saattaa viedä muutaman minuutin,
# kun Stan kirjoittaa C++-koodia.

vino_t_simulointi <- stan(model_name= "vino_t_simulaatio", model_code =
   stan_koodi_vino_t,
data=c("Y","N","X"), init=init_W, iter=simulaatiot_yht,
   warmup=simulaatiot_warmup, chains=simulaatioketjut)

# Simulaatiotulokset, missä "se_mean"
# on MC-keskivirhe parametrien odotusarvoestimaatille (mean)
# ja "Rhat" on R-hattu eli parametrien simulaatioketjujen konvergoitumissuure:

summary(vino_t_simulointi, pars=c('b_Intercept', 'sigma', 'g', 'v'))$summary


#####

# 8. Posterioriennustejakauma estimoiduilla parametreilla.

# Parametrien simuloidut arvot:
```

```
simuloidut_parametriarvot <- extract(vino_t_simulointi, pars=c('b_Intercept',
    'sigma', 'g', 'v','W'))
str(simuloidut_parametriarvot)

posterioriennusteet <- matrix(rep(0,N*kaytettavat_simulaatiot),nrow=N)

for (j in 1:kaytettavat_simulaatiot) {
Z <- rnorm(N,0,simuloidut_parametriarvot$sigma[j])
w <- sample(simuloidut_parametriarvot$W[j,])
posterioriennusteet[,j] <- rep(simuloidut_parametriarvot$b_Intercept[j],N) +
    simuloidut_parametriarvot$g[j]*w + sqrt(w)*Z
}

#####

# 9. Posterioriennustetarkistus:
# posterioriennusteet vs. alkuperäinen havaintoaineisto (vino_t).

# Keskiarvot:
mean_sim <- mean(posterioriennusteet)
mean_sim
mean_vino_t <- mean(vino_t)
mean_vino_t

# Keskihajonnat:
sd_sim <- sd(posterioriennusteet)
sd_sim
sd_vino_t <- sd(vino_t)
sd_vino_t

# Jakaumien prosenttipisteet:
p_pisteet_posterioriennusteet <- quantile(posterioriennusteet,
    c(0.01,seq(0.05,0.95,0.05),0.99))
p_pisteet_vino_t <- quantile(vino_t, c(0.01,seq(0.05,0.95,0.05),0.99))

# Vertaillaan prosenttipisteiden sopivuutta.
# Kun jakaumat suoralla, niin jakaumat vastaavat toisiaan.
plot(p_pisteet_posterioriennusteet,p_pisteet_vino_t)
abline(0,1)

# Vielä histogrammeilla jakaumien vertailu:
par(mfrow=c(2,1))
hist(posterioriennusteet,breaks=1000, xlim=c(-10,5))
hist(vino_t, breaks=350, xlim=c(-10,5))
```

```
#####

# 10. Vertailu suhteessa normaalijakaumaoletukseen
# ja ei-vinoon t-jakaumaoletukseen.

# 10.1 Normaalijakaumaoletus:

# Epäinformatiiviset priorit normaalijakaumamalliin:
normal_priorit <- c(set_prior("normal(0,10)", class = "Intercept"),
   set_prior("cauchy(0,10)", class = "sigma"))

# Simulaatio BRMS:llä, joka käyttää Stan-ohjelmaa:
normaali_oletus_data <- as.data.frame(Y)
normaali_oletus <- brm(Y ~ 1, family=gaussian, prior = normal_priorit,
chains = simulaatioketjut, iter = simulaatiot_yht, warmup = simulaatiot_warmup,
data=normaali_oletus_data)

summary(normaali_oletus)

posterioriennusteet_normaali_oletus <- as.vector(predict(normaali_oletus,
   summary=F))
p_pisteet_vino_normaali_oletus <-
   quantile(posterioriennusteet_normaali_oletus,
   c(0.01,seq(0.05,0.95,0.05),0.99))

# Vertaillaan prosenttipisteiden sopivuutta.
# Kun jakaumat suoralla, niin jakaumat vastaavat toisiaan.
par(mfrow=c(1,1))
plot(p_pisteet_vino_normaali_oletus,p_pisteet_vino_t)
abline(0,1)


# 10.2 Ei-vino Studentin t-jakauma -oletus:

# Epäinformatiiviset priorit t-jakauma-malliin:
student_priorit <- c(set_prior("normal(0,10)", class = "Intercept"),
   set_prior("cauchy(0,10)", class = "sigma"),
set_prior("cauchy(0,10)", class = "nu"))

# Simulaatio BRMS:llä, joka käyttää Stan-ohjelmaa:
student_oletus_data <- as.data.frame(Y)
ei_vino_student_t_oletus <- brm(Y ~ 1, family=student, prior = student_priorit,
chains = simulaatioketjut, iter = simulaatiot_yht, warmup = simulaatiot_warmup,
```

```
data=student_oletus_data)

summary(ei_vino_student_t_oletus)

posterioriennusteet_ei_vino_student_t_oletus <-
    as.vector(predict(ei_vino_student_t_oletus, summary=F))
p_pisteet_ei_vino_student_t_oletus <-
    quantile(posterioriennusteet_ei_vino_student_t_oletus,
    c(0.01,seq(0.05,0.95,0.05),0.99))

# Vertaillaan prosenttipisteiden sopivuutta.
# Kun jakaumat suoralla, niin jakaumat vastaavat toisiaan.
plot(p_pisteet_ei_vino_student_t_oletus,p_pisteet_vino_t)
abline(0,1)


# 10.3 Vertaillaan mallien ennustetarkkuutta,
# pienempi WAIC ja LOO viittaavat parempaan tarkkuuteen:

# Vino-t-malli vs. normaalijakaumamalli.
compare(loo(extract_log_lik(vino_t_simulointi)),
loo(logLik(normaali_oletus)))
compare(waic(extract_log_lik(vino_t_simulointi)),
waic(logLik(normaali_oletus)))

# Vino-t-malli vs. ei-vino t-malli.
compare(loo(extract_log_lik(vino_t_simulointi)),
loo(logLik(ei_vino_student_t_oletus)))
compare(waic(extract_log_lik(vino_t_simulointi)),
waic(logLik(ei_vino_student_t_oletus)))

# Ei-vino t-malli vs. normaalijakaumamalli.
compare(loo(logLik(ei_vino_student_t_oletus)),
loo(logLik(normaali_oletus)))
compare(waic(logLik(ei_vino_student_t_oletus)),
waic(logLik(normaali_oletus)))

#######################################################
# R-ohjelman koodi:
# Vinon Studentin t-jakauman simulointi Stan-ohjelmalla.
# LOPPU.
#######################################################
```

```
#######################################################
# Nettovaikutusmallinnuksessa käytetyt
# Stan-ohjelman (rstan) (Goodrich et al. 2015)
# ja R-paketin BRMS (Buerkner 2015) koodit.
#######################################################


# Esitän koodit ilman pitkiä selityksiä.
# Kehotan lukijaa tutustumaan tarvittaessa
# näiden R-pakettien ohjekirjoihin.

# Esitän koodit malleittain/mallinnusaskeleittain.

# Tulokset ja mallien numerot on esitetty esimerkiksi
# Irakin osalta taulukossa 4.4, s. 177.

# Mallinnusaskeleet on esitetty numeroineen luvussa:
# Aineistomallin mallinnusaskeleet (4.3.3).

# Parametrien selitteet löytyvät taulukosta 4.1, s. 170
# ja parametrien priorijakaumista
# ks. lukuEstimoitavien parametrien priorijakaumat (4.3.4).


######

# Malli1/mallinnusaskel1.
# Tyhjä hierarkkinen normaalimalli (luku 4.3.3.1):

brm(Y ~ 1 + (1|hlot), data=nettovaikutusaineisto_irak,
   family="gaussian",prior=priorit,
chains=2,iter=1500,warmup=700)

# Missä priorijakaumat:
  b ~ uniform(-3,3);    # (β)
  sigma ~ cauchy(0,5);  # (σ)
  sd_1 ~ cauchy(0,5);   # (τ_i)
  pre_1 ~ normal(0,1);

# ja "hlot" on tieto, joka erittelee aineiston henkilöt.

######
```

```
# Malli2/mallinnusaskel2.
# Tyhjä hierarkkinen Studentin t-malli (luku 4.3.3.2):

brm(Y ~ 1 + (1|hlot), data=nettovaikutusaineisto_irak,
    family="student",prior=priorit,
chains=2,iter=1500,warmup=700)

# Missä priorijakaumat:
  b ~ uniform(-3,3);    # (β)
  nu ~ cauchy(0,10);    # (ν)
  sigma ~ cauchy(0,5);  # (σ)
  sd_1 ~ cauchy(0,5);   # (τ_i)
  pre_1 ~ normal(0,1);

# ja "hlot" on tieto, joka erittelee aineiston henkilöt.

######

# Malli3/mallinnusaskel3.
# Selittävät muuttujat hierarkkisessa Studentin t-mallissa (luku 4.3.3.3):

brm(Y ~ b_1 + ... + b_p +
(1|hlot), data=nettovaikutusaineisto_irak, family="student", autocor =
    cor_arma(formula = ~1|hlot,r=2), prior=priorit,
chains=2,iter=1500,warmup=700)

# Missä priorijakaumat:
  b[1] ~ uniform(-3,3); # (β_0), vakiotermi
  b ~ normal(0,1);      # (β)
  arr ~ uniform(-1,1);  # (β_u)
  nu ~ cauchy(0,10);    # (ν)
  sigma ~ cauchy(0,5);  # (σ)
  sd_1 ~ cauchy(0,5);   # (τ_i)
  pre_1 ~ normal(0,1);

# ja "hlot" on tieto, joka erittelee aineiston henkilöt.

######

# Malli4, osa I/mallinnusaskel4.
# Satunnaisvaikutusten lisääminen hierarkkiseen Studentin t-malliin
# (luku 4.3.3.4):
```

```
brm(Y ~  b_1 + ... + b_p +
(1 + maassaoloaika + ian_nelio|hlot) + (1|maahanmuuttovuosi) +
   (1|maassaoloaika), data=nettovaikutusaineisto_irak, family="student",
   autocor = cor_arma(formula = ~1|hlot,r=2), prior=priorit,
chains=2,iter=1500,warmup=700)

# Missä priorijakaumat:
  b[1] ~ uniform(-3,3); # (β0), vakiotermi
  b ~ normal(0,1);      # (β)
  nu ~ cauchy(0,10);    # (ν)
  arr ~ uniform(-1,1);  # (βu)
  sigma ~ cauchy(0,5);  # (σ)
  sd_1 ~ cauchy(0,5);   # (τi^2)
  L_1 ~ lkj_corr_cholesky(1);
  pre_1 ~ normal(0,1);
  sd_2 ~ cauchy(0,5);   # (τm^2)
  sd_3 ~ cauchy(0,5);   # (τt^2)

# ja "hlot" on tieto, joka erittelee aineiston henkilöt.

######

# Malli4, osa II/mallinnusaskel4.
# Vino otosjakauma ja puolison nettovaikutukset (luku 4.3.3.5):

# Tässä esitettävä Stan-ohjelman koodi
# on tehty BRMS-ohjelman tuottamien Stan-koodien pohjalta.
# Vrt. myös edellä tässä liitteessä esitetty
# koodi tyhjälle vinon t-jakauman mallille.
# Priorijakaumista puuttuu määritelmä autoregressiivisille termeille:
# siten niiden oletuspriorijakaumaksi tuli: tasajakauma(-1,1).
# Tässä esitetty vanha versio mallille 4, lopuksi 4. mallinnus-
# askeleessa käytettiin koodia, joka alla olevan 5:nnen kaltainen.

functions {
}
data {
  int<lower=1> N;  # number of observations
  vector[N] Y;  # response variable
  int<lower=1> K;  # number of fixed effects
  matrix[N, K] X;  # FE design matrix

  # data for random effects of hlot
  int<lower=1> J_1[N];  # RE levels
```

```
  int<lower=1> N_1;  # number of levels
  int<lower=1> K_1;  # number of REs
  row_vector[K_1] Z_1[N];  # RE design matrix
  int NC_1;  # number of correlations

  # data for random effects of maahanmuuttovuosi_kesk
  int<lower=1> J_2[N];  # RE levels
  int<lower=1> N_2;  # number of levels
  int<lower=1> K_2;  # number of REs
  real Z_2[N];  # RE design matrix

  # data for random effects of maassaolo_sum_kesk
  int<lower=1> J_3[N];  # RE levels
  int<lower=1> N_3;  # number of levels
  int<lower=1> K_3;  # number of REs
  real Z_3[N];  # RE design matrix

  # data needed for ARR effects
  int<lower=1> Karr;
  matrix[N, Karr] Yarr;  # ARR design matrix
}
transformed data {
}
parameters {
  real b_Intercept;  # fixed effects Intercept
  vector[K] b;  # fixed effects

  matrix[N_1, K_1] pre_1;  # unscaled REs
  vector<lower=0>[K_1] sd_1;  # RE standard deviation
  cholesky_factor_corr[K_1] L_1;  # cholesky factor of correlations matrix

  vector[N_2] pre_2;  # unscaled REs
  real<lower=0> sd_2;  # RE standard deviation

  vector[N_3] pre_3;  # unscaled REs
  real<lower=0> sd_3;  # RE standard deviation

  vector[Karr] arr;  # autoregressive effects of the response

  real<lower=0> sigma;  # residual SD without skewness
  real<lower=1> v;  # degrees of freedom
  real g ; # Y skewness parameter

  vector<lower=0>[N] W; # Y skew-t-apumuuttuja.
```

```
}
transformed parameters {
  vector[N] eta;  # linear predictor

  vector[K_1] r_1[N_1];  # REs
  vector[N_2] r_2;  # REs
  vector[N_3] r_3;  # REs

  vector<lower=0>[N] skew_sigma; # skew-Y SD

  # compute linear predictor
  eta <- (X * b + b_Intercept) + (Yarr * arr) + (g * W); # (g * W) due to
    skewness

  # RE1
  for (i in 1:N_1) {
    r_1[i] <- sd_1 .* (L_1 * to_vector(pre_1[i]));  # scale REs
  }

  # RE2
  r_2 <- sd_2 * (pre_2);  # scale REs

  # RE3
  r_3 <- sd_3 * (pre_3);  # scale REs

  # if available add REs to linear predictor
  for (n in 1:N) {
    eta[n] <- eta[n] + Z_1[n] * r_1[J_1[n]] + Z_2[n] * r_2[J_2[n]] + Z_3[n] *
   r_3[J_3[n]];
  }

  # Skew Y SD
   for (n in 1:N) {
   skew_sigma[n] <- sigma * sqrt(W[n]);
   }

}
model {
  # 1. prior specifications
  b_Intercept ~ uniform(-3,3);
  b ~ normal(0,1);

  sd_1 ~ cauchy(0,5);
```

```
  L_1 ~ lkj_corr_cholesky(1);
  to_vector(pre_1) ~ normal(0, 1);

  sd_2 ~ cauchy(0,5);
  pre_2 ~ normal(0, 1);

  sd_3 ~ cauchy(0,5);
  pre_3 ~ normal(0, 1);

  sigma ~ cauchy(0,5);
  v ~ cauchy(0,10); # degrees of freedom
  g ~ uniform(-10,10); # skewness

  # 2. likelihood contribution

    # skew-Y, apumuuttuja W:
  W ~ inv_gamma(v/2, v/2);

  # skew-t jakautunut Y
  Y ~ normal(eta, skew_sigma);

}
generated quantities {
  vector[N] log_lik; # Pointwise Log-likelihood
  for (n in 1:N) log_lik[n] <- normal_log(Y[n], eta[n], skew_sigma[n]);

}

######

# Malli5, mallinnusaskel5.
# Otosjakauman vinouden mallinnus (luku 4.3.3.6):

# Priorijakaumista puuttuu määritelmä autoregressiivisille termeille:
# siten niiden oletuspriorijakaumaksi tuli: tasajakauma(-1,1).

functions {
}
data {
  int<lower=1> N;  // total number of observations
  vector[N] Y;  // response variable
  int<lower=1> K;  // number of population-level effects
  matrix[N, K] X;  // population-level design matrix
  // data for group-specific effects of ID 1
```

```
  int<lower=1> J_1[N];
  int<lower=1> N_1;
  int<lower=1> M_1;
  vector[N] Z_1_1;
  vector[N] Z_1_2;
  vector[N] Z_1_3;
  int<lower=1> NC_1;
  // data for group-specific effects of ID 2
  int<lower=1> J_2[N];
  int<lower=1> N_2;
  int<lower=1> M_2;
  vector[N] Z_2_1;
  vector[N] Z_2_2;
  int<lower=1> NC_2;
  // data for group-specific effects of ID 3
  int<lower=1> J_3[N];
  int<lower=1> N_3;
  int<lower=1> M_3;
  vector[N] Z_3_1;
  vector[N] Z_3_2;
  int<lower=1> NC_3;
  // data needed for ARR effects
  int<lower=1> Karr;
  matrix[N, Karr] Yarr;  // ARR design matrix

    # skewness model
  int<lower=1> K_skewness;  # number of skewness fixed effects
  matrix[N, K_skewness] X_skewness;  # FE design matrix
}
transformed data {
  int Kc;
  matrix[N, K - 1] Xc;  // centered version of X
  vector[K - 1] means_X;  // column means of X before centering
  Kc = K - 1;  // the intercept is removed from the design matrix
  for (i in 2:K) {
    means_X[i - 1] = mean(X[, i]);
    Xc[, i - 1] = X[, i] - means_X[i - 1];
  }
}
parameters {
  vector[Kc] b;  // population-level effects
  real temp_Intercept;  // temporary Intercept
  real<lower=0> sigma;  // residual SD
  real<lower=0> v;  // degrees of freedom
```

```
  vector<lower=0>[M_1] sd_1;  // group-specific standard deviations
  matrix[M_1, N_1] z_1;  // unscaled group-specific effects
  // cholesky factor of correlation matrix
  cholesky_factor_corr[M_1] L_1;

  vector<lower=0>[M_2] sd_2;  // group-specific standard deviations
  matrix[M_2, N_2] z_2;  // unscaled group-specific effects
  // cholesky factor of correlation matrix
  cholesky_factor_corr[M_2] L_2;

  vector<lower=0>[M_3] sd_3;  // group-specific standard deviations
  matrix[M_3, N_3] z_3;  // unscaled group-specific effects
  // cholesky factor of correlation matrix
  cholesky_factor_corr[M_3] L_3;

  vector[Karr] arr;  // autoregressive effects of the response

  # Vinous:
  vector[K_skewness] b_skewness; # skewness model parameter(s)
  vector<lower=0>[N] W; # Y skew-t-apumuuttuja, ks. skew t distribution.pdf,
    s. 4 (5)
}
transformed parameters {
  // group-specific effects
  matrix[N_1, M_1] r_1;
  vector[N_1] r_1_1;
  vector[N_1] r_1_2;
  vector[N_1] r_1_3;
  // group-specific effects
  matrix[N_2, M_2] r_2;
  vector[N_2] r_2_1;
  vector[N_2] r_2_2;
  // group-specific effects
  matrix[N_3, M_3] r_3;
  vector[N_3] r_3_1;
  vector[N_3] r_3_2;
  r_1 = (diag_pre_multiply(sd_1, L_1) * z_1)';
  r_1_1 = r_1[, 1];
  r_1_2 = r_1[, 2];
  r_1_3 = r_1[, 3];
  r_2 = (diag_pre_multiply(sd_2, L_2) * z_2)';
  r_2_1 = r_2[, 1];
  r_2_2 = r_2[, 2];
```

```
  r_3 = (diag_pre_multiply(sd_3, L_3) * z_3)';
  r_3_1 = r_3[, 1];
  r_3_2 = r_3[, 2];

}
model {
  vector[N] eta;
  vector[N] skew_sigma; # skew-Y SD

  eta = (Xc * b) + temp_Intercept + (Yarr * arr) + ((X_skewness * b_skewness)
    .* W);
  for (n in 1:N) {
    eta[n] = eta[n] + r_1_1[J_1[n]] * Z_1_1[n] + r_1_2[J_1[n]] * Z_1_2[n] +
   r_1_3[J_1[n]] * Z_1_3[n] + r_2_1[J_2[n]] * Z_2_1[n] + r_2_2[J_2[n]] *
   Z_2_2[n] + r_3_1[J_3[n]] * Z_3_1[n] + r_3_2[J_3[n]] * Z_3_2[n];
    skew_sigma[n] = sigma * sqrt(W[n]);
  }

    # 1. prior specifications
  temp_Intercept ~ uniform(-3,3);
  b ~ normal(0,1);

  sd_1 ~ cauchy(0,5);
  L_1 ~ lkj_corr_cholesky(1);
  to_vector(z_1) ~ normal(0, 1);

  sd_2 ~ cauchy(0,5);
  L_2 ~ lkj_corr_cholesky(1);
  to_vector(z_2) ~ normal(0, 1);

  sd_3 ~ cauchy(0,5);
  L_3 ~ lkj_corr_cholesky(1);
  to_vector(z_3) ~ normal(0, 1);

  sigma ~ cauchy(0,5);
  v ~ cauchy(0,10); # degrees of freedom

  b_skewness ~ normal(0,1);

  # 2. likelihood contribution

    # skew-Y, apumuuttuja W:
  W ~ inv_gamma(v/2, v/2);
```

```
  # skew-t jakautunut Y
  Y ~ normal(eta, skew_sigma);

}
generated quantities {
  real b_Intercept;  // population-level intercept

  vector[N] eta_enn;
  vector[N] skew_sigma_enn; # skew-Y SD
  vector[N] log_lik; # Pointwise Log-likelihood

  b_Intercept = temp_Intercept - dot_product(means_X, b);

  eta_enn = (Xc * b) + temp_Intercept + (Yarr * arr) + ((X_skewness *
   b_skewness) .* W);
  for (n in 1:N) {
    eta_enn[n] = eta_enn[n] + r_1_1[J_1[n]] * Z_1_1[n] + r_1_2[J_1[n]] *
   Z_1_2[n] + r_1_3[J_1[n]] * Z_1_3[n] + r_2_1[J_2[n]] * Z_2_1[n] +
   r_2_2[J_2[n]] * Z_2_2[n] + r_3_1[J_3[n]] * Z_3_1[n] + r_3_2[J_3[n]] *
   Z_3_2[n];
    skew_sigma_enn[n] = sigma* sqrt(W[n]);
        log_lik[n] <- normal_log(Y[n], eta_enn[n], skew_sigma_enn[n]);
  }
}


##############################################################
# Nettovaikutusmallinnuksessa käytetyt Stan-ohjelman (rstan)
# ja R-paketin BRMS koodit.
# LOPPU.
##############################################################
```

Liite C

Hurdle-gamma-sekoitusjakauman simulointi Stan-ohjelmalla. Palkkasummamallinnuksessa käytetyt Stan- ja BRMS-koodit.

```
##############################################################
# R-ohjelman koodi:
# Hurdle-gamma-sekoitusjakauman simulointi Stan-ohjelmalla.
##############################################################


# Katso hurdle-gamma-sekoitusjakaumasta luvut:
# Tyhjä hurdle-gamma (5.3.1.1) ja
# Tyhjä hurdle-gamma-sekoitus (5.3.1.2).

# Sisältö:

# 1. Tarvittavat R-paketit (asennettava aluksi).
# 2. Poimitaan otos havaintoja hurdle-gamma-sekoitusjakaumasta
# (2 gamma-jakauman sekoitus).
# 3. Aineisto Stan-ohjelmaa varten.
# 4. Stan-ohjelman koodi.
# 5. Simulaatioketjuja koskevat valinnat.
# 6. Simulointi Stan-ohjelmalla.
# 7. Posteriorisimulaatio estimoiduilla parametreilla.
# 8. Posterioriennustetarkistus: posterioriennusteet vs.
# alkuperäinen havaintoaineisto (hurdle_gamma_sekoitus_aineisto).

# 9. Vertailu suhteessa hurdle-gamma-jakaumaan (ei gamma-jakaumien sekoitusta).

#####


# 1. Tarvittavat R-paketit (asennettava aluksi).

# rstan on Stanin R-ohjelman versio.
# ggplot2 kuvion piirtämistä varten (kohta 8.).
# loo ja brms mallivertailua varten (kohta 9.).

library(rstan)          # (Goodrich et al. 2015).

library(ggplot2)        # (Wickham et al.).

library(brms)           # (Buerkner 2015).
library(loo)            # (Vehtari et al. 2016).


#####
```

```r
# 2. Poimitaan Otos_lkm kappaletta havaintoja hurdle-gamma-sekoitusjakaumasta
# (2 gamma-jakauman sekoitus):

Otos_lkm <- 2000


# 2.1 Määritetään parametriarvot.

# p_hurdle (ks. lauseke 5.5, s.250)
# ja p_theta (ks. lauseke 5.9, s.251):
p_hurdle <- 0.3
p_theta <- 0.4

# 1. gamma-jakauman parametit:
muoto_gamma1 <- 0.8
rate_gamma1 <- 1        # skaala_gamma1 = 1/rate_gamma1

# Esimerkkiotos 1. gamma-jakaumasta:
hist(rgamma(10000,muoto_gamma1,rate_gamma1),breaks=100)

# 2. gamma-jakauman parametit:
muoto_gamma2 <- 10
rate_gamma2 <- 2        # skaala_gamma2 = 1/rate_gamma2

# Esimerkkiotos 2. gamma-jakaumasta:
hist(rgamma(10000,muoto_gamma2,rate_gamma2),breaks=100)


# 2.2 Poiminta hurdle-gamma-sekoitusjakaumasta (2 gamma-jakauman sekoitus.
hurdle_gamma_sekoitus_aineisto <- numeric(Otos_lkm)

for (s in 1:Otos_lkm) {

        hurdle_toteuma <- rbinom(1,1,p_hurdle) # Poiminta Bernoulli-jakaumasta.

        if (hurdle_toteuma==1) {
          hurdle_gamma_sekoitus_aineisto[s] <- 0
        } else {
            theta_toteuma <- rbinom(1,1,p_theta) # Poiminta
                # Bernoulli-jakaumasta.
                if (theta_toteuma==1) {
                hurdle_gamma_sekoitus_aineisto[s] <-
   rgamma(1,muoto_gamma1,rate_gamma1)
```

```
                } else {
                hurdle_gamma_sekoitus_aineisto[s] <-
    rgamma(1,muoto_gamma2,rate_gamma2)
                }
        }

}

# Hurdlen keskiarvo poimitussa aineistossa:
mean(hurdle_gamma_sekoitus_aineisto==0)

# Koko poimittu jakauma:
plot(density(hurdle_gamma_sekoitus_aineisto))

# Yli nollan olevien poimittujen havaintojen jakauma:
hist(hurdle_gamma_sekoitus_aineisto[hurdle_gamma_sekoitus_aineisto>0],
    breaks=30)

#####


# 3. Aineisto Stan-ohjelmaa varten:

N <- 3*Otos_lkm
N_trait <- Otos_lkm
Y <- hurdle_gamma_sekoitus_aineisto
X <- cbind(c(rep(1,N_trait),rep(0,2*N_trait)),
c(rep(0,N_trait),rep(1,N_trait),rep(0,N_trait)),
c(rep(0,N_trait),rep(0,N_trait),rep(1,N_trait)))
K <- ncol(X)

#####

# 4. Stan-ohjelman koodi:

# Perustuu BRMS-paketin Hurdle-gamma-jakaumaa
# (ei sekoitusjakaumaa) koskevaan koodiin (ks. kohta 9.).
# Parametrien priorijakaumat ovat heikosti informatiivisia.
# Koodissa 2 gamma-jakaumaa: "gamma_a" ja "gamma_b".
# Vasta simulaatiotuloksissa nähdään,
# kumpi näistä vastaa 1. gamma-jakaumaa ja kumpi 2. gamma-jakaumaa,
# koska malli itsessään ei tee tätä identifioimista millään tavalla.
# Tästä syystä ajetaan vain yksi simulaatioketju:
# ks. sekoitusjakaumien komponenttijakaumien
```

```
# identifioituvuudesta luku 5.3.2 sekä
# Stan Development Team (2015), s. 202--204.

stan_koodi_hurdle_gamma_sekoitus <- '
functions {
  /* Funktio: Hurdle-gamma-sekoitusjakauman log-tiheysfunktio
   * yhden havainnon osalta.
   *
   * Funktion argumentit:
   * Ks. funktion argumenteista luvut 5.3.1.1 - 5.3.1.3.
   *   y: yksi havainto yllä poimitusta hurdle-gamma-sekoitusjakaumasta.
   *   theta: Indikaattori havainnolle kuulua gamma-jakaumaan a
   *   (eikä gamma-jakaumaan b).
   *   muoto_gamma_b: Gamma-jakauman a muotoparametri.
   *   muoto_gamma_b: Gamma-jakauman b muotoparametri.
   *   myy_gamma_a: Gamma-jakauman a sovite, jakauman log-odotusarvo.
   *   myy_gamma_b: Gamma-jakauman b sovite, jakauman log-odotusarvo.
   *   myy_hurdle: Hurdlen odotusarvo, eli todennäköisyys sille,
    *  että havainnon arvo on nolla.
   *
   * Funktio palauttaa arvona:
   *   Havaintokohtainen log-tiheysfunktion arvo,
   *   joka lisätään koko jakaumaa käsittelevään summaan.
   *   --> Stan-ohjelman posteriorisimulaatio perustuu
   *   log-posterioritiheysfunktion määrittämiseen (ks. luku 4.3.6.1).
   */

   real hurdle_gamma_lpdf(real y, real theta, real muoto_gamma_a, real
    muoto_gamma_b, real myy_gamma_a,
                          real myy_gamma_b, real myy_hurdle) {
     if (y == 0) {
       return bernoulli_logit_lpmf(1| myy_hurdle);
     } else {
       return bernoulli_logit_lpmf(0| myy_hurdle) +
             log_sum_exp(log(theta) + gamma_lpdf(y| muoto_gamma_a,
    muoto_gamma_a / exp(myy_gamma_a)),
             log1m(theta) + gamma_lpdf(y| muoto_gamma_b, muoto_gamma_b /
    exp(myy_gamma_b)));
     }
   }
}
data {
  int<lower=1> N;   // 3*havaintojen lukumäärä
        // (hurdlea, gamma_a-jakaumaa ja gamma_b-jakaumaa varten).
```

```
  int<lower=1> N_trait;  // Havaintojen lukumäärä.
  real Y[N_trait];  // Havaintoaineisto, selitettävä muuttujat.
  int<lower=1> K;  // Vakiotermien lukumäärä (= 3)
        // hurdlen, gammma_a-jakauman ja gamma_b-jakauman malleissa.
  matrix[N, K] X;  // Ykkösten ja nollien matriisi vakioita varten.
}
transformed data {
}
parameters {
  vector[K] b;  // Hurdlen, gamma_a-jakauman ja gamma_b-jakauman
        // tyhjien regressiomallien vakiotermit.
  real<lower=0, upper=1> p_theta_a; // Todennäköisyys havainnolle kuulua
        // gamma_a-jakaumaan (eikä gamma_b-jakaumaan).
  real<lower=0> muoto_gamma_a;  // Gamma_a-jakauman muotoparametri.
  real<lower=0> muoto_gamma_b;  // Gamma_b-jakauman muotoparametri.
}
transformed parameters {
  vector[N] myy;  // Hurdlen, gamma_a-jakauman, ja gamma_b-jakauman
        // lineaariset sovitteet.
  myy = X * b;
}
model {
  // Priorijakaumat:
  muoto_gamma_a ~ uniform(0.1,100); // Heikosti informatiivinen priori-jakauma.
  muoto_gamma_b ~ uniform(0.1,100); // Heikosti informatiivinen priori-jakauma.

  b ~ uniform(-10, 10); // Heikosti informatiivinen priori-jakauma..
  p_theta_a ~ uniform(0,1); // Heikosti informatiivinen priori-jakauma.

  // Summataan havaintokohtaiset log-posteriorijakauman arvot.
  // Ks. funktio "hurdle_gamma" yllä.
  for (n in 1:N_trait) {
    Y[n] ~ hurdle_gamma(p_theta_a, muoto_gamma_a, muoto_gamma_b, myy[n], myy[n
   + N_trait], myy[n + (2*N_trait)]);
  }
}
generated quantities {
// Log-uskottavuusfunktion arvot talteen LOO:n ja WAIC:n laskemista varten.
vector[N_trait] log_lik;
for (n in 1:N_trait) log_lik[n] = hurdle_gamma_lpdf(Y[n]| p_theta_a,
   muoto_gamma_a, muoto_gamma_b, myy[n], myy[n + N_trait], myy[n +
   (2*N_trait)]);
}
'
```

```
#####


# 5. Simulaatioketjua koskevat valinnat.

simulaatiot_yht <- 1000
simulaatiot_warmup <- 300
kaytettavat_simulaatiot <- simulaatiot_yht-simulaatiot_warmup

#####


# 6. Simulointi Stan-ohjelmalla.
# Vain yksi simulaatioketju.

hurdle_gamma_sekoitus <- stan(model_name = "hurdle_gamma_sekoitus", model_code
   = stan_koodi_hurdle_gamma_sekoitus,
data=c("N","N_trait","Y","K","X"),
iter=simulaatiot_yht, warmup=simulaatiot_warmup, chains=1)

# Simulaatiotulokset, missä "se_mean"
# on MC-keskivirhe parametrien odotusarvoestimaatille (mean)
# ja "Rhat" on R-hattu eli parametrien konvergoitumistestisuure:

summary(hurdle_gamma_sekoitus,
   pars=c('b','p_theta_a','muoto_gamma_a','muoto_gamma_b'))$summary
# Vakiotermien (b) selitteet:
# exp(b[1]) on gamma_a-jakauman odotusarvo.
# exp(b[2]) on gamma_b-jakauman odotusarvo.
# logistic(b[3]) eli 1/1+exp(-b[3]) on hurdlen odotusarvo.

#####


# 7. Posterioriennustejakauma estimoiduilla parametreilla.

# Parametrien posteriorijakaumien simuloidut arvot:
post.parametriarvot <- extract(hurdle_gamma_sekoitus,
   pars=c('myy','p_theta_a','muoto_gamma_a','muoto_gamma_b'))
str(post.parametriarvot)

# 7.1 Määritetään aluksi, onko simuloitu gamma_a-jakauma
# sekoitusjakauman 1. gamma-jakauma vai 2. gamma-jakauma
# (ja gamma_b-jakauma on 1. gamma-jakauma,
```

```
# jos gamma_a-jakauma on 2. gamma-jakauma, jne).

# p_theta on alussa määritelty todennäköisyys
# havainnolle kuulua 1. gamma-jakaumaan.

# Tämän perusteella simuloinnin gamma_a-jakauma on 1. gamma-jakauma
# (ja gamma_b-jakauma 2. gamma-jakauma), jos :

if((p_theta<0.5 & mean(post.parametriarvot$p_theta)<0.5) | (p_theta>=0.5 &
   mean(post.parametriarvot$p_theta)>=0.5)) {

post.muoto_gamma1 <- post.parametriarvot$muoto_gamma_a
post.myy_gamma1 <- post.parametriarvot$myy[,1]
post.p_theta <- post.parametriarvot$p_theta_a
post.muoto_gamma2 <- post.parametriarvot$muoto_gamma_b
post.myy_gamma2 <- post.parametriarvot$myy[,(Otos_lkm+1)]
} else {

# muutoin gamma_jakauma on 2. gamma-jakauma
# (ja gamma_b-jakauma 1. gamma-jakauma):

post.muoto_gamma1 <- post.parametriarvot$muoto_gamma_b
post.myy_gamma1 <- post.parametriarvot$myy[,(Otos_lkm+1)]
post.p_theta <- 1-post.parametriarvot$p_theta_a
post.muoto_gamma2 <- post.parametriarvot$muoto_gamma_a
post.myy_gamma2 <- post.parametriarvot$myy[,1]
}

# 7.2 Ensimmäinen gamma-jakauma parametrien posteriorijakaumien
# simuloiduilla arvoilla.
post.p_gamma1 <- rgamma(Otos_lkm*kaytettavat_simulaatiot, post.muoto_gamma1,
   post.muoto_gamma1/exp(post.myy_gamma1))
hist(post.p_gamma1,breaks=100, xlim=c(0,10))

# 7.3 Ensimmäinen gamma-jakauma parametrien posteriorijakaumien
# simuloiduilla arvoilla.
post.p_gamma2 <- rgamma(Otos_lkm*kaytettavat_simulaatiot, post.muoto_gamma2,
   post.muoto_gamma2/exp(post.myy_gamma2))
hist(post.p_gamma2,breaks=100, xlim=c(0,15))

# 7.4 Hurdle parametrin posteriorijakauman simuloiduilla arvoilla.
post.p_hurdle <-
   rep(1/(1+exp(-post.parametriarvot$myy[,(2*Otos_lkm+1)])),Otos_lkm)
mean(post.p_hurdle)
```

```
hist(post.p_hurdle)

# 7.5 Thetan (havainnon kuuluminen 1. gamma-jakaumaan) parametri p_theta.
hist(post.p_theta)
mean(post.p_theta)
post.p_theta2 <- rep(post.p_theta,Otos_lkm)

# 7.6 Posterioriennusteet hurdle-gammasekoitusjakaumalle.
posterioriennusteet <- numeric(kaytettavat_simulaatiot*Otos_lkm)

for (n in 1:(kaytettavat_simulaatiot*Otos_lkm)) {

  hurdle_testi <- runif(1,0,1)

        if (hurdle_testi<post.p_hurdle[n]) {
        posterioriennusteet[n] <- 0
        } else {
            theta_testi <- runif(1,0,1)
                if (theta_testi<post.p_theta2[n]) {
                posterioriennusteet[n] <- post.p_gamma1[n]
                } else {
                posterioriennusteet[n] <- post.p_gamma2[n]
                }
        }

}

#####


# 8. Posterioriennustetarkistus: posterioriennusteet vs.
# alkuperäinen havaintoaineisto (hurdle_gamma_sekoitus_aineisto).

# Keskiarvot:
mean_sim <- mean(posterioriennusteet)
mean_sim
mean_hurdle_gamma_sekoitus_aineisto <- mean(hurdle_gamma_sekoitus_aineisto)
mean_hurdle_gamma_sekoitus_aineisto

# Keskihajonnat:
sd_sim <- sd(posterioriennusteet)
sd_sim
sd_hurdle_gamma_sekoitus_aineisto <- sd(hurdle_gamma_sekoitus_aineisto)
sd_hurdle_gamma_sekoitus_aineisto
```

```
# Nollaosuudet:
mean(posterioriennusteet==0)
mean(hurdle_gamma_sekoitus_aineisto==0)

# Jakaumien prosenttipisteet:
p_pisteet_posterioriennusteet <- quantile(posterioriennusteet,
   c(0.01,seq(0.05,0.95,0.05),0.99))
p_pisteet_hurdle_gamma_sekoitus_aineisto <-
   quantile(hurdle_gamma_sekoitus_aineisto, c(0.01,seq(0.05,0.95,0.05),0.99))

# Vertaillaan prosenttipisteiden sopivuutta. Kun pisteet suoralla,
# niin jakaumat vastaavat toisiaan.
plot(p_pisteet_posterioriennusteet,p_pisteet_hurdle_gamma_sekoitus_aineisto)
abline(0,1)

# Vielä histogrammeilla jakaumien vertailu:

# Koko jakaumat:
par(mfrow=c(2,1))
hist(posterioriennusteet,breaks=30, xlim=c(0,10))
hist(hurdle_gamma_sekoitus_aineisto, breaks=30, xlim=c(0,10))

# Koko jakaumat:
par(mfrow=c(2,1))
hist(posterioriennusteet[posterioriennusteet>0],breaks=100, xlim=c(0,10))
hist(hurdle_gamma_sekoitus_aineisto[hurdle_gamma_sekoitus_aineisto>0],
   breaks=50, xlim=c(0,10))

# Vertailu: Aineiston histogrammi vs. parametrien posterioriennusteiden
# keskiarvojen tuottama jakauma:

# Histogrammiaineisto, ylinollat:
hist_ylinollat <-
   hist(hurdle_gamma_sekoitus_aineisto[hurdle_gamma_sekoitus_aineisto>0],
   breaks=30)

havai_hist <- rep(hist_ylinollat$density,each=10)

Xt <- seq(0.01,max(hist_ylinollat$breaks),length.out=
   length(hist_ylinollat$density)*10)

theta.m <- 1/(1+(exp(mean(post.p_theta))))
```

```
gamma_1.m <- theta.m*dgamma(Xt, shape= mean(post.muoto_gamma1), rate=
   mean(post.muoto_gamma1) / exp(mean(post.myy_gamma1)) )
gamma_2.m <- (1-theta.m)*dgamma(Xt, shape= mean(post.muoto_gamma2), rate=
   mean(post.muoto_gamma2) / exp(mean(post.myy_gamma2)) )

gamma_1_2.m <- gamma_1.m + gamma_2.m
plot(Xt, gamma_1_2.m)

gammat.df.m <- as.data.frame(cbind(gamma_1_2.m,gamma_1.m,gamma_2.m,havai_hist))


# Piirretään jakaumien vertailu ggplot2:lla:
havainnot_vs_mean_posteriori_enn <- ggplot(gammat.df.m, aes(x=Xt, y =
   havai_hist))+
geom_line()+
geom_area(aes(x=Xt, y = havai_hist, fill="Aineiston jakauma"), alpha=0.3)+
geom_line(aes(y= gamma_1.m,
color="1. gamma-jakauma",
linetype="1. gamma-jakauma",
size="1. gamma-jakauma"))+
geom_line(aes(y= gamma_2.m,
color="2. gamma-jakauma",
linetype="2. gamma-jakauma",
size="2. gamma-jakauma"))+
geom_line(aes(y= gamma_1_2.m,
color="Sekoitusjakauma",
linetype="Sekoitusjakauma",
size="Sekoitusjakauma"))+
labs(x = "", y = element_blank())+
scale_y_continuous(limits=c(0,0.5))+
scale_fill_manual(name="Alkuperäinen havaintoaineisto:",
values =c("Aineiston jakauma"="green"))+
scale_size_manual(name="Odotusarvoestimaatit:",
values=c("2. gamma-jakauma"=1.2,
"1. gamma-jakauma"=1.2, "Sekoitusjakauma"=1.5))+
scale_linetype_manual(name="Odotusarvoestimaatit:",
values=c("2. gamma-jakauma"=2,
"1. gamma-jakauma"=2, "Sekoitusjakauma"=1))+
scale_colour_manual(name="Odotusarvoestimaatit:",
values=c("2. gamma-jakauma"="red",
"1. gamma-jakauma"="blue", "Sekoitusjakauma"="black"))+
theme_minimal(base_size=12, base_family = "") %+replace%
theme(legend.text=element_text(size=16),
        legend.position = c(0.8, 0.8),
```

```
        legend.key.width=unit(1,"cm"),
        legend.background = element_rect(fill="white",
        linetype="solid",
        colour="black"))+
guides(fill= guide_legend(order = 1))

havainnot_vs_mean_posteriori_enn


#####


# 9. Vertailu suhteessa hurdle-gamma-jakaumaan (ei gamma-jakaumien sekoitusta).

# Hurdle-gamma-jakauman simulointi BRMS-ohjelman avulla.
# Määritetään samankaltaiset heikosti informatiiviset priorijakaumat
# kuin sekoitusjakaumamallin yhteydessä.
priorit <- c(set_prior("normal(0,10)", class = "Intercept"),
set_prior("gamma(1,1)", class = "shape"),
set_prior("beta(1,1)", class = "hu"))

hurdle_gamma_oletus <- brm(hurdle_gamma_sekoitus_aineisto ~ 1,
data=as.data.frame(hurdle_gamma_sekoitus_aineisto),prior=priorit,
family=hurdle_gamma,iter=1500, warmup=500, chains=2)
summary(hurdle_gamma_oletus)

# Stan-ohjelman koodi hurdle-gamma-mallille, jonka pohjalta yllä oleva
# hurdle-gammasekoitusmallin koodi on luotu:
stancode(hurdle_gamma_oletus)

posterioriennusteet_hurdle_gamma_oletus <-
   as.vector(predict(hurdle_gamma_oletus, summary=F))
par(mfrow=c(1,1))
hist(posterioriennusteet_hurdle_gamma_oletus, breaks=100, xlim=c(0,10))

p_pisteet_hurdle_gamma_oletus <-
   quantile(posterioriennusteet_hurdle_gamma_oletus,
   c(0.01,seq(0.05,0.95,0.05),0.99))

# Vertaillaan prosenttipisteiden sopivuutta.
# Kun jakaumat suoralla, niin jakaumat vastaavat toisiaan.
plot(p_pisteet_hurdle_gamma_oletus,p_pisteet_hurdle_gamma_sekoitus_aineisto)
abline(0,1)
```

```
# 9.2 Vertaillaan mallien ennustetarkkuutta,
# pienempi WAIC ja LOO viittaavat parempaan tarkkuuteen:
compare(waic(extract_log_lik(hurdle_gamma_sekoitus)),
waic(logLik(hurdle_gamma_oletus)))
compare(loo(extract_log_lik(hurdle_gamma_sekoitus)),
loo(logLik(hurdle_gamma_oletus)))


##############################################################
# R-ohjelman koodi:
# Hurdle-gamma-sekoitusjakauman simulointi Stan-ohjelmalla.
# LOPPU.
##############################################################
```

```
##############################################################
# Palkkasummamallinnuksessa käytetyt
# Stan-ohjelman (rstan) (Goodrich et al. 2015)
# ja R-paketin BRMS (Buerkner 2015) koodit.
##############################################################


# Esitän koodin tässä vain viimeiselle eli neljännelle mallinnusaskeleelle,
# joka on hurdle-gamma-sekoitusjakauma, joka sisältää satunnaisvaikutuksia:
# ks. luku 5.3.1.4.

# Ensimmäinen mallinnusaskel (luku 5.3.1.1) vastaa käytännössä
# yllä esitettyä hurdle-gamma-jakauman
# simulointia BRMS-paketilla.

# Toinen mallinnusaskel (luku 5.3.1.2) vastaa käytännössä
# yllä esitettyä hurdle-gammasekoitusjakauman
# simulointia Stan-ohjelmalla.

#  Kolmannessa mallinnusaskeleessa (luku 5.3.1.3) lisätään kaikille
   havainnoille
# yhteiset selittävät muuttujat, joten sen Stan-koodi on käytännössä
   samanlainen
# toisen mallinnusaskeleen kanssa.

# Esitän lopullisen mallin koodin ilman pitkiä selityksiä.
# Kehotan lukijaa tutustumaan tarvittaessa
# näiden R-pakettien ohjekirjoihin sekä
# hurdle-gammasekoitusjakauman Stan-koodiin edellä.

# Parametrien selitteet löytyvät taulukosta 5.1, s. 262.
# Ks. myös edellä hurdle-gammasekoitusjakauman simulointi.
# Parametrien priorijakaumista (sekoitusmallissa)
# ks. luku Estimoitavien parametrien priorijakaumat (5.3.2).
# Sekoitusjakauman komponentit identifioidaan tässä
# melko informatiivisten priorijakaumien avulla.

######

# Tässä esitettävä Stan-ohjelman koodi
# on tehty BRMS-ohjelman hurdle-gamma-jakaumaperhettä koskevan
# Stan-koodin pohjalta.
```

```
functions {
  /* hurdle gamma log-PDF of a single response
   * Args:
   *   y: the response value
   *   shape: shape parameter of gamma distribution
   *   eta: linear predictor for gamma part
   *   eta_hu: linear predictor for hurdle part
   * Returns:
   *   a scalar to be added to the log posterior
   */
   real hurdle_gamma_log(real y, real theta, real shape1, real shape2, real
    eta1,
                         real eta2, real eta_hu) {
     if (y == 0) {
       return bernoulli_logit_log(1, eta_hu);
     } else {
       return bernoulli_logit_log(0, eta_hu) +
             log_sum_exp(log(theta) + gamma_log(y, shape1, shape1 / exp(eta1)),
             log1m(theta) + gamma_log(y, shape2, shape2 / exp(eta2)));
     }
   }
}
data {
  int<lower=1> N;  # number of observations
  int<lower=1> N_trait;  # number of obs per response
  real Y[N_trait];  # response variable
  int<lower=1> K;  # number of fixed effects
  matrix[N, K] X;  # FE design matrix

  # Data for prediction of latent class membership
  int<lower=1> K_theta;
  matrix[N_trait, K_theta] X_theta;  # Theta design matrix

  # data for random effects of hlot
  int<lower=1> J_1[N];  # RE levels
  int<lower=1> N_1;  # number of levels
  int<lower=1> K_1;  # number of REs
  row_vector[K_1] Z_1[N];  # RE design matrix
  int NC_1;  # number of correlations

}
transformed data {
}
parameters {
```

```
  vector[K] b;  # fixed effects

  vector[K_theta] b_theta;  # latent group membership

  real<lower=0> shape1;  # 1.-gammajakauman shape parameter
  real<lower=0> shape2;  # 2.-gammajakauman shape parameter

  # RE1
    # REgamma1
  vector<lower=0>[N_1] pre_gamma_1;  # unscaled REs
  real<lower=0> r_lambda1;

  # REgamma2
  vector<lower=0>[N_1] pre_gamma_2;  # unscaled REs
  real<lower=0> r_lambda2;

 # hurdle
  vector<lower=0>[N_1] pre_hurdle;  # unscaled REs
  real<lower=0> r_hurdle;

  # RE1 gamma class
  real<lower=0> sd_theta;
  vector[N_trait] pre_theta;  # unscaled REs

}
transformed parameters {
  vector[N] eta_kiint;  # linear predictor
  vector[N] eta;  # linear predictor + REs
  vector[N_trait] eta_theta; # logit predictor for gamma mixture membership

  # RE1-theta
  vector[N_1] r_theta;

  # compute linear predictor
  eta_kiint <- X * b ;

 # Adding random effects
  for (n in 1:N) {
    eta[n] <- eta_kiint[n] + Z_gamma_1[n] * pre_gamma_1[J_1[n]] + Z_gamma_2[n]
   * pre_gamma_2[J_1[n]] + Z_hurdle[n] * pre_hurdle[J_1[n]];
  }

  r_theta <- pre_theta*sd_theta;
```

```
  # compute linear predictor for gamma class and add RE-theta
  for (n_trait in 1:N_trait) {
        eta_theta[n_trait] <- inv_logit((X_theta[n_trait] * b_theta) +
    r_theta[J_1[n_trait]]);
  }

}
model {
  # prior specifications

  shape1 ~ gamma(256, 256);
  shape2 ~ gamma(200, 2);

 # Huom! Satunnaisvaikutuksia sisältävän mallin 2 vakiota
 # jouduttiin kiinnittämään, jotta kummassakin ajetussa
 # simulaatioketjussa samat gamma1- ja gamma2-jakaumat
 # eli etteivät vaihdu päittäin simulaatioketjuittain.
  b[1] ~ normal(0.8101, 0.0001);
  b[2] ~ normal(1.3036, 0.0001);
  b[3] ~ normal(-4.12,1);
  b ~ normal(0,10);

  b_theta[1] ~ normal(0, 2);
  b_theta ~ normal(0, 10);

  # Satunnaisvaikutukset
  r_lambda1 ~ cauchy(0, 10);
  r_lambda2 ~ cauchy(0, 10);
  r_hurdle ~ cauchy(0, 10);

  to_vector(pre_gamma_1) ~ gamma(1/r_lambda1, 1/r_lambda1);
  to_vector(pre_gamma_2) ~ gamma(1/r_lambda2, 1/r_lambda2);
  to_vector(pre_hurdle) ~ gamma(1/r_hurdle, 1/r_hurdle);

  # RE-theta
  sd_theta ~ cauchy(0, 10);
  to_vector(pre_theta) ~ normal(0, 1);

  # likelihood contribution
  for (n in 1:N_trait) {
    Y[n] ~ hurdle_gamma(eta_theta[n], shape1, shape2, eta[n], eta[n +
    N_trait], eta[n + (2*N_trait)]);
  }
}
```

```
generated quantities {
vector[N_trait] log_lik; # Pointwise Log-likelihood
for (n in 1:N_trait) log_lik[n] <- hurdle_gamma_log(Y[n], eta_theta[n],
   shape1, shape2, eta[n], eta[n + N_trait], eta[n + (2*N_trait)]);
}

##############################################################
# Palkkasummamallinnuksessa käytetyt
# Stan-ohjelman (rstan) (Goodrich et al. 2015)
# ja R-paketin BRMS (Buerkner 2015) koodit.
# LOPPU.
##############################################################
```