Puuttuvuuden mallintaminen FINRISKI -tutkimuksessa

Elli Hirvonen

Tilastotieteen pro gradu -tutkielma

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
24. toukokuuta 2017

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Hirvonen, Elli: Puuttuvuuden mallintaminen FINRISKI -tutkimuksessa

Tilastotieteen pro gradu -tutkielma, 46 sivua, 2 liitettä (11 sivua)
24. toukokuuta 2017

Tiivistelmä:
Puuttuva tieto on ongelma terveystutkimuksissa, koska tutkimukseen osallistujat ja ei-osallistujat usein eroavat toisistaan. Puuttuvuuden mallintaminen on tärkeää, jotta tuloksia pystyttäisiin korjaamaan ja jatkossa ottamaan mallinnettu puuttuvuus paremmin huomioon tutkimusta suunnitellessa. Myös Kansallinen FINRISKI-tutkimus kärsii puuttuvuudesta, sillä tutkimuksen osallistumisprosentti on laskenut jatkuvasti. Tutkimus on Terveyden ja hyvinvoinnin laitoksen väestötutkimussarja, jonka tavoitteena on seurata kansantauteja ja niiden riskitekijöitä Suomessa sekä tarkkailla suomalaisten terveydentilaa. Tässä työssä pyritään mallintamaan Kansallisen FINRISKI-tutkimuksen puuttuvuus tarkastelemalla, kuinka osallistumisaktiivisuus on muuttunut vuosien edetessä miehillä ja naisilla eri tutkimusalueilla.

Kun tarkastellaan osallistumisaktiivisuuden muuttumista eri ihmisillä ajan kuluessa, ollaan kiinnostuneita aikamuuttujista ikä, tutkimusvuosi ja syntymävuosi. Nämä aikamuuttujat ovat matemaattisesti riippuvia toisistaan, joten niiden lisääminen malliin yhtä aikaa on hankalaa. Ikä-periodi-kohortti -analyysi pyrkii tähän, mutta vaatii vahvoja oletuksia, jotka harvoin täyttyvät. Periodi kuvaa analyysissä tutkimusvuotta ja kohortti syntymävuotta. Tässä tutkielmassa ikä-periodi-kohortti -analyysi toteutetaan osissa siten, että tehdään kolme mallia ja jokaisessa mallissa on aina kaksi aikamuuttujaa kerrallaan mukana. Tämä voidaan tehdä, sillä kun kaksi aikamuuttujaa tiedetään, kolmas voidaan laskea kahden muun avulla. Näin päästään eroon kolmen aikamuuttujan vaikutuksien identifioituvuusongelmasta.

Mallien sovittamiseen käytetään additiivista logistista regressiota, joka kuuluu yleistettyihin additiivisiin malleihin. Yleistetyt additiiviset mallit ovat yleistettyjen lineaaristen mallien laajennus, missä lineaariset termit korvataan tasoittavien funktioiden summalla, jolloin epälineaarisia termejä voidaan sovittaa joustavammin. Mallit esitellään graafisesti ja jokainen malli tulkitaan erikseen. Tuloksiksi saatiin mielenkiintoisia eroja miesten ja naisten, eri ikäisten tutkittavien sekä tutkimusvuosien ja -alueiden välillä.

Avainsanat: ikä-periodi-kohortti -analyysi, yleistetyt additiiviset mallit, tasoitusfunktio, kolmannen asteen tasoituspline, takaisinsovitusalgoritmi, lokaali pisteytysalgoritmi, yleistetty logistinen regressio, Kansallinen FINRISKI -tutkimus

Sisältö

1 Johdanto

Puuttuva tieto on kasvava ongelma erilaisissa terveystutkimuksissa. Puuttuvuus koetaan ongelmaksi, koska tutkimukseen osallistujat ja ei-osallistujat usein eroavat toisistaan, jolloin tuloksien yleistettävyys huononee. Puuttuvuuden mallintaminen on tärkeää, jotta tulokset saataisiin koskemaan haluttua kohdepopulaatiota. (Kopra ym., 2015, ss. 1-2). Tuloksia pystyttäisiin tällöin korjaamaan ja tutkimuksia suunnittelemaan jatkossa paremmin, jolloin saataisiin mahdollisesti enemmän vastauksia.

Tämän tutkielman tarkoituksena on mallintaa kansallisen FINRISKI -tutkimuksen puuttuvuutta. Tutkimuksen osallistumisprosentti on laskenut 90 %:sta 64 %:iin vuosien 1972-2012 välillä, joten puuttuvuuden mallintaminen on tarpeellista. Tässä työssä keskitytään siihen, miten osallistumisaktiivisuus on kehittynyt vuosien edetessä miehillä ja naisilla eri tutkimusalueilla.

Kansallinen FINRISKI -tutkimus on Terveyden- ja hyvinvoinninlaitoksen väestötutkimussarja, jossa seurataan sydän- ja verisuonitautien sekä muiden kansantautien ja näiden riskitekijöiden tasoa ja muutosta Suomessa (Terveyden ja hyvinvoinnin laitos, 2015). Tutkimuksen tavoitteena on kerätä tietoa näistä taudeista ja niiden riskitekijöistä. Lisäksi tietoa kerätään kansantautien esiintymisestä väestössä sekä tarkkaillaan suomalaisten terveydentilaa. (Borodulin ym., 2013).

Ensimmäinen tutkimus tehtiin vuonna 1972, jolloin se kantoi nimeä Pohjois-Karjala -projekti. Itä-Suomessa havaittiin 1960-luvulla merkittävästi enemmän sydänkuolleisuutta kuin muualla Suomessa ja siksi Pohjois-Karjala -projekti perustettiin kartoittamaan riskitekijöitä. Myöhemmin Pohjois-Karjala -projekti on jatkunut kansallisena FINRISKI-tutkimuksena ja koko 40 vuoden tutkimusjaksoa kutsutaan tällä nimellä. Tutkimus toteutetaan viiden vuoden välein ja viimeisin tutkimus on tehty vuonna 2012. (Borodulin ym., 2013; Terveyden ja hyvinvoinnin laitos, 2015). Seuraava tutkimus toteutetaan vuonna 2017, jolloin Kansallinen FINRISKI -tutkimus tulee jatkumaan osana Kansallista FinTerveys -tutkimusta (Terveyden ja hyvinvoinnin laitos, 2017).

Tutkimus on Suomen suurin väestötutkimus ja Suomen oloissa ainutlaatuinen. Tutkimuksen avulla vakavien terveysuhkien torjuntaa osataan näin parantaa yhteiskunnan voimavaroilla. Tutkimuksen aineistoa pidetään kansallisesti arvokkaana tietopankkina suomalaisten terveydestä sekä elintavoista tutkimuksen alusta tähän päivään. (Terveyden ja hyvinvoinnin laitos, 2015).

Kansallisen FINRISKI -tutkimuksen tilanteessa halutaan mallintaa puuttuvuutta siten, että katsotaan, kuinka osallistumisaktiivisuus on muuttunut ajan myötä. Tällöin voidaan käyttää ikä-periodi-kohortti -analyysiä, jossa pyritään identifioimaan aikamuuttujien ikä, tutkimusvuosi ja syntymävuosi vaikutukset. Periodilla tarkoitetaan tässä tutkimusvuotta ja kohortilla syntymävuotta. Aikamuuttujien identifiointi on hyvin hankalaa, koska kaikki kolme aikamuutujaa

ovat vahvasti korreloituneita keskenään. Ikä-periodi-kohortti -analyysi on luotu tilanteeseen, jossa kaikki kolme aikamuuttujaa voitaisiin lisätä malliin yhtä aikaa. Näihin ratkaisuihin on kuitenkin esitetty kritiikkiä Bellin ja Jonesin (2014) artikkelissa ja jotta malli toimisi oikein, vahvojen oletusten on oltava voimassa, mikä ei yleensä ole totta.

Tässä työssä pyritään vastaamaan ikä-periodi-kohortti -analyysien ongelmaan toteuttamalla analyysi siten, että sovitetaan kolme mallia, joissa jokaisessa on aina vain kaksi aikamuuttujaa kerrallaan mukana. Tämä perustuu siihen tosiasiaan, että ikä, tutkimusvuosi ja syntymävuosi ovat matemaattisesti riippuvaisia toisistaan, joten kun kaksi näistä muuttujista tiedetään, kolmas muuttuja voidaan laskea kahden muun avulla.

Mallien sovittamiseen käytetään Hastien ja Tibshiranin (1986; 1990; 2006) esittelemää yleistettyä additiivista mallia, joka on yleistetyn lineaarisen mallin laajennus. Koska vastemuuttuja on dikotominen, tutkittava joko osallistui tutkimukseen tai sitten ei, sovitetaan mallit käyttäen additiivista logistista mallia, joka kuuluu yleistettyihin additiivisiin malleihin.

Tämän tutkielman luvussa 2 esitellään työn aineistoa tarkemmin sekä kerrotaan muuttujista. Nähdään hyvin selkeästi, kuinka osallistumisprosentit ovat laskeneet vuosi vuodelta. Lisäksi esitellään dataan lisätyt muuttujat sekä datan rakenteeseen tehdyt muutokset, mitkä olivat tarpeen analyysien kannalta.

Luku 3 käsittelee analyysimenetelmiä. Ensin esitellään ikä-periodi-kohortti -analyysin taustaa. Sen jälkeen tutustutaan yleistettyjen additiivisten mallien teoriaan sekä siihen, kuinka parametrien estimointi suoritetaan.

Luku 4 keskittyy analyysien tuloksiin. Mallien muodostamiseen käytettiin R-ohjelmistoa (R Core Team, 2017). Ennen tuloksien tulkintaa esitellään R-notaatio, jolla malliyhtälöt on esitelty. Tehdyt mallit esitetään graafisesti ja jokainen malli esitellään ja tulkitaan erikseen. Tuloksiksi saatiin mielenkiintoisia ilmiöitä ja kaikki mallit tukivat toistensa tuloksia. Tuloksien kuvaajat on koottu omaksi osiokseen lukuun 5.

Viimeiseksi luvussa 6 esitellään johtopäätökset ja sen jälkeen on maininta hankeympäristöstä. Liitteissä on muokatun datan kuusi ensimmäistä riviä sekä analyysien ja kuvaajien R-koodi.

2 Aineisto

Tässä työssä käytetty aineisto sisältää FINRISKI -tutkimuksen vuosien 1982-2012 rekisteritietoja. Kultakin vuodelta on kerätty seuraavat muuttujat:

- *VUOSI* = Tutkimusvuosi
- *IKA* = Tutkittavan ikä täysinä vuosina
- *SUKUP* = Tutkittavan sukupuoli (1=mies ja 2=nainen)
- *ALUE* = Tutkimusalue (2=Pohjois-Karjala, 3=Pohjois-Savo, 4=Turku/Loimaa, 5=Helsinki/Vantaa ja 6=Oulun lääni)
- *OSAL* = Osallistumisstatus (1=osallistui tutkimukseen ja 0=ei osallistunut tutkimukseen)
- *N* = Kutsuttujen henkilöiden määrä

Otanta-asetelma on vaihdellut hieman tutkimusvuosittain. Tutkittavat poimittiin joka tutkimusvuosi väestörekisteristä ja otanta oli ositettu tutkimusalueittain. Vuonna 1972 tutkittavat poimittiin systemaattisesti syntymäpäivän mukaan ja vuonna 1977 otanta suoritettiin yksinkertaisella satunnaisotannalla. Vuonna 1982 ositettiin otanta 10-vuotisikäryhmän mukaan. Vuodesta 1987 eteenpäin tutkittavat poimittiin väestörekisteristä satunnaisotannalla siten, että kultakin tutkimusalueelta jokaisesta sukupuolen ja 10-vuotisikäryhmän mukaan ositetussa solussa oli 200-250 henkilöä riippuen tutkimusvuodesta (Terveyden ja hyvinvoinnin laitos, 2016; Kopra ym., 2015, ss. 3-4).

Tutkimuksiin valittiin noin 10 000 tutkittavaa, joille lähetettiin kutsut tutkimukseen postitse 2-4 viikkoa ennen tutkimuspäivää. Tutkimukseen osallistujilta mitattiin pituus ja paino, vyötärön ja lantion ympärys sekä verenpaine ja pulssi. Tämän lisäksi osallistujille tehtiin laboratoriomäärityksiä ja heitä pyydettiin täyttämään kyselylomakkeita. Kaikista tutkittavista kerättiin myös terveystietoja kansallisista rekistereistä, kuten sairaaloiden hoitoilmoitus-, lääkekorvaus- ja kuolinsyyrekisteristä. (Terveyden ja hyvinvoinnin laitos, 2015). Osallistuminen tutkimuksessa määriteltiin siten, että tutkittava osallistui tutkimukseen, jos hän oli täyttänyt kyselylomakkeen sekä osallistunut terveystarkastukseen. Jos tutkittava ei ollut osallistunut terveystarkastukseen, vaikka hän olisi täyttänyt kyselylomakkeen, hänet määriteltiin ei-osallistujaksi.

Vuonna 1972 tutkimukseen valittiin 25-59-vuotiaita henkilöitä ja tutkimusalueina olivat Pohjois-Karjala sekä Pohjois-Savo. Samoin vuonna 1977 tutkimusalueina pysyivät kaksi edellä mainittua, mutta tutkittavien ikä oli 30-64 vuotta. Vuosina 1982 ja 1987 mukaan tuli kolmas tutkimusalue, Turku/Loimaa, ja

tutkittavien ikä vaihteli 25 vuodesta 64 vuoteen. Vuonna 1992 ikähaitari pysyi samana, mutta Helsinki/Vantaa lisättiin tutkimusalueisiin. Vuodesta 1997 tutkittavien ikä vakiintui 25-74 vuoteen. Samana vuonna Oulun lääni lisättiin tutkimusalueisiin. Vuosina 2002 ja 2007 vielä Lapin lääni otettiin mukaan, mutta vuonna 2007 siellä järjestettiin vain postikysely (Terveyden ja hyvinvoinnin laitos, 2016, s. 8). Vuonna 2012 Lapin lääni jätettiin pois tutkimusalueista.

Tutkimuksen osallistumisprosentti on laskenut voimakkaasti vuosien kuluessa. Vuosina 1972 ja 1977 osallistumisprosentti oli lähes 90 %. Vuosina 1982 ja 1987 osallistumisprosentti oli vähän yli 80 %, 1992 ja 1997 hieman yli 70 % ja 2002 sekä 2007 vielä 70 %:n tuntumassa, kunnes vuonna 2012 osallistumisprosentti oli enää vain noin 64 %. Lähes jokaisena tutkimusvuotena naisia osallistui tutkimukseen enemmän kuin miehiä. Taulukkoon 1 on koottu tarkempaa tietoa osallistumisprosentista.

Tämän työn analyyseihin otettiin mukaan vain vuodet 1982-2012, sillä vuodet 1972 ja 1977 sisälsivät puuttuvaa tietoa $ALUE$-muuttujan osalta, jos tutkittava ei ollut osallistunut tutkimukseen. Alue 7 eli Lapin lääni jätettiin myös pois, koska sille oli tuloksia vain kahdelta vuodelta. Lisäksi, koska tälle alueelle suoritettiin vain postikysely vuonna 2007, se ei olisi vertailukelpoinen muihin tutkimusalueisiin nähden.

Aineistoon lisättiin uusi muuttuja $S.VUOSI$, joka määräytyi muuttujien $VUOSI$ ja IKA erotuksesta, ja kertoi näin ollen tutkittavan henkilön syntymävuoden. Osallistumisstatuksen ja kutsuttujen määrän N avulla tehtiin uudet muuttujat $K.OSAL$, joka kertoi, kuinka monta tutkittavaa osallistui tutkimukseen sekä $E.OSAL$, joka taas kertoi, kuinka monta tutkittavaa ei osallistunut tutkimukseen.

Taulukko 1: Tietoa osallistumisprosentista. Tutkimusalueet ovat: 2=Pohjois-Karjala,3=Pohjois-Savo, 4=Turku/Loimaa, 5=Helsinki/Vantaa, 6=Oulun lääni ja 7=Lapin lääni.

Vuosi	**Kutsuttujen määrä**	**Ikä**	**Osallistuneita % (määrä)**	**Naisia osallistuneista % (määrä)**	**Miehiä osallistuneista % (määrä)**	**Tutkimus-alueet**
1972	12 440	25-59	87.9 (10 938)	50.8 (5 552)	49.2 (5 386)	2 ja 3
1977	11 359	30-64	89.8 (10 197)	51.8 (5 278)	48.2 (4 919)	2 ja 3
1982	11 395	25-64	82.0 (9 347)	50.6 (4 732)	49.4 (4 615)	2, 3 ja 4
1987	7 931	25-64	81.7 (6 478)	52.0 (3 369)	48.0 (3 109)	2, 3 ja 4
1992	7 927	25-64	76.3 (6 051)	52.9 (3 202)	47.1 (2 849)	2, 3, 4 ja 5
1997	11 500	25-74	73.4 (8 446)	49.6 (4 192)	50.4 (4 253)	2, 3, 4, 5 ja 6
2002	13 498	25-74	71.0 (9 580)	53.2 (5 098)	46.8 (4 482)	2, 3, 4, 5, 6 ja 7
2007	12 000	25-74	66.6 (7 993)	53.2 (4 253)	46.8 (3 740)	2, 3, 4, 5, 6 ja 7
2012	10 000	25-74	64.2 (6 424)	52.7 (3 383)	47.3 (3 041)	2, 3, 4, 5 ja 6

3 Analyysimenetelmät

Analyyseihin valittiin lähtökohdaksi ikä-periodi-kohortti -analyysi (Age-Period-Cohort analysis), jossa ollaan kiinnostuneita kolmesta aikamuuttujasta; ikä, vuosi ja syntymävuosi. Analyyseissä periodi tarkoittaa tutkimusvuotta ja kohortti syntymävuotta. Ikä, tutkimusvuosi ja syntymävuosi ovat tärkeitä syvällisen ymmärryksen kannalta, kun halutaan tietää, miten osallistumisaktiivisuus on muuttunut ajan myötä. Tässä työssä ikä-periodi-kohortti -analyysi toteutetaan siten, että sovitetaan kolme additiivista logistista regressiomallia, joissa jokaisessa jätetään aina yksi aikamuuttuja kerrallaan pois. Näistä malleista tehdään kuvaajia, joilla voidaan tarkastella osallistumiskäyttäytymistä todennäköisyyksien avulla ja sitä kautta mallintaa puuttuvuus. Alalukujen 3.2.1 ja 3.2.2 esiteltävä teoria ja algoritmit ovat Hastien, Tibshiranin ja Friedmanin kirjasta *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2009) luvuista 5.1-5.4 ja 9.1.

3.1 Ikä-periodi-kohortti -analyysi

Muutos tutkimukseen osallistumisessa ilmenee jokaisessa aikamuuttujassa hieman eri tavalla. Yksilöt ensinnäkin vanhenevat, jolloin muutos johtuu ikääntymisestä. Syntymävuosi vaikuttaa taas siten, että kyseisenä vuonna syntynyt ikäryhmä eli kohortti on erilainen kuin aiempana vuonna syntynyt riippumatta iästä. Lisäksi, tutkimusvuoden vaikutus ilmenee aikakautena, joka muokkaa paljon yksilöiden elämää riippumatta siitä, milloin he ovat syntyneet tai minkä ikäisiä he ovat. (Bell & Jones, 2014, s. 334). Tämän vuoksi ikä-periodi-kohortti -analyysi toimisi hyvin Kansallisen FINRISKI -tutkimuksen puuttuvuuden mallintamiseen.

Ryder (1965) oli yksi ensimmäisistä tutkijoista, joka teki eron ikääntymisen, aikakausien muutoksen ja kohorttiryhmien välillä. Näin ollen hän otti huomioon kohortit ennemmin sosiaalisen muutoksen lähteenä kuin peräkkäisinä vuosina esiintyvänä muutoksena. (Bell & Jones, 2014, s. 335).

Aikamuuttujat ikä, tutkimusvuosi ja syntymävuosi riippuvat toisistaan matemaattisesti. Tämä aiheuttaa identifioituvuusongelman. Ikä-periodi-kohortti -analyysi on yritetty kehittää sellaiseen tilanteeseen, jossa sovitetaan yhteen malliin kaikki kolme aikamuuttujaa (ikä, vuosi ja syntymävuosi). Erään ratkaisun ovat esittäneet Yang ja Land teoksessaan *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications* (2013). Bell ja Jones (2014) ovat kritisoineet artikkelissaan Yangin ja Landin hierarkkista ikä-periodi-kohortti -analyysiä ja todenneet, että mikään tekninen toteutus ei pysty rikkomaan loogista ja matemaattista yhteyttä, mikä näiden kolmen aikamuuttujan välillä vallitsee. Jos asiaa lähestyttäisiin Bayes-tilastotieteen keinoin, Bellin ja Jonesin mukaan estimaatteja voitaisiin painaa oikeaan suuntaan, jos tehdään vahva ja

oikea priorioletus. Tämä ei kuitenkaan ole usein tiedossa.

Tässä työssä sovitetaan kolme mallia, joissa jokaisessa jätetään aina yksi aikamuuttujista *IKA, VUOSI* ja *S.VUOSI* kerrallaan pois. Näitä malleja voidaan vertailla ja tulkita, koska aikamuuttujat ovat matemaattisesti riippuvia toisistaan ja, kun aikamuuttujista tiedetään kaksi kolmas voidaan laskea kahden muun avulla. Sosiaalisia ja yksilöllisiä muutoksia pystytään tarkastelemaan merkityksellisellä ja robustilla tavalla ilman, että kaikki kolme aikamuuttujaa ovat samassa mallissa (Bell & Jones, 2014, ss. 336-337).

3.2 Yleistetyt additiiviset mallit

Aikamuuttujien *VUOSI, IKA* ja *S.VUOSI* yhteydestä tutkimuksen osallistumiseen ei ole tarkkaa tietoa. Tämän vuoksi mallit päätettiin sovittaa yleistetyillä additiivisilla malleilla (Generalized Additive Models), joilla voidaan sovittaa epälineaarisia kovariaatteja joustavasti ja mallintaa monimutkaisia riippuvuuksia (Hastie & Tibshirani, 2006, s. 1).

Yleistetty additiivinen malli on laajennus yleistetystä lineaarisesta mallista (Generalized Linear Model), missä lineaarinen kovariaatti korvataan tasoittavien funktioiden summalla. Tämä tarkoittaa sitä, että lineaarinen funktio $\sum_{j=1}^{p} \beta_j X_j$ korvataan additiivisella funktiolla $\sum_{j=1}^{p} s_j(X_j)$, jossa muuttujat s_j ovat datasta estimoitavia tasoitusfunktioita. (Hastie & Tibshirani, 1986, s. 297). Yleistetty additiivinen malli voidaan esittää esimerkiksi muodossa

$$g(\mu) = \beta_0 + s_1(x_1) + \ldots + s_p(x_p),$$

missä g on linkkifunktio, μ on vastemuuttujan y odotusarvo ja muuttujan y jakauma kuuluu eksponentiaaliseen perheeseen (Hastie & Tibshirani, 2006, s. 1).

Tasoitusfunktioiden muoto estimoidaan datasta ja tasoitusfunktioilla saadaan paljastettua epälineaarisia regressioefektejä selittäjän x_j vaikutuksesta. Mallin kaikkien kovariaattien ei tarvitse olla epälineaarisia, vaan malleja voidaan tehdä sekoittaen lineaarisia ja epälineaarisia selittäjiä esimerkiksi silloin, kun mukana on diskreettejä selittäjiä. Tällöin diskreetit muuttujat ovat lineaarisia ja jatkuvat muuttujat epälineaarisia. (Hastie & Tibshirani, 2006, s. 1). Tasoitusfunktioiden määrittämiseen on olemassa monia erilaisia tapoja, joita on esitelty Hastien ja Tibshiranin kirjassa (1990), mutta myöhemmin esitellään vain tämän työn analyyseissä käytetty tapa.

Yleistettyjen additiivisten mallien estimoimiseen voidaan käyttää R-ohjelmiston paketin *gam* funktiota *gam*. Paketin dokumentaatio on Hastien (2016) kirjoittama ja yleistettyjen additiivisten mallien sovittaminen tapahtuu lähdekirjallisuudessa esitellyllä esitellyllä tavalla, joka käydään seuraavaksi läpi niiltä osin kuin se on tämän työn kannalta tärkeää.

3.2.1 Additiivisen mallin sovittaminen

Tutustutaan ensin additiivisen mallin sovittamiseen, koska additiivisten mallien teoria antaa pohjan yleistetyille additiivisille malleille ja sitä kautta additiiviseen logistiseen regressioon. Sovittamisen työkaluna käytetään hajontakuvion tasoittajaa (scatterplot smoother). Tässä työssä hajontakuvion tasoittajana käytetään kolmannen asteen tasoitussplinea (cubic smoothing spline).

Kolmannen asteen tasoitussplinen periaatteena on määritellä paloittain määritelty funktio solmukohtien väleille. Solmukohdat ovat paikkoja, joissa edellinen polynomi yhtyy seuraavan kanssa, jolloin muodostuu yhtenäinen käyrä. Polynomeina käytetään kolmannen asteen polynomeja ja näiden polynomien kaarevuutta mitataan niiden toisilla derivaatoilla ja tasoituksen määrä saadaan laskemalla neliöityjä integraaleja toisen asteen derivaatoista.

Kolmannen asteen tasoitussplineltä vaaditaan jatkuvuus, jotta edelliseen solmukohtaan loppunut polynomi yhtyisi saumattomasti seuraavan polynomin kanssa. Lisäksi vaaditaan ensimmäisen ja toisen asteen derivaattojen jatkuvuus, jolloin lopputulos on tasaisen kaareva ja yhteneväinen käyrä läpi havaintopisteiden. Kolmannen asteen tasoitussplinessä päästään solmujen määrittämisen ongelmasta, kun käytetään solmujen maksimaalista määrää.

Tasoitusfunktio estimoidaan minimoimalla penalisoitu jäännösneliösumma (penalized residual sum of squares, PRSS)

$$PRSS(s, \lambda) = \sum_{i=1}^{N} \big(y_i - s(x_i)\big)^2 + \lambda \int_a^b \big(s''(t)\big)^2 dt,$$

missä λ on kiinnitetty tasoitusparametri ja $a \leq x_i \leq \cdots \leq x_n \leq b$. Ensimmäinen termi mittaa läheisyyttä dataan ja toinen osa on sakkotermi, joka rankaisee funktion kaarevuudesta. Tasoitusparametri λ luo kompromissin näiden kahden osan välille, jolloin lopputulos on yhteensopivuuden ja tasaisuuden kompromissi. Jos λ on iso, sakkotermi saa ison painon ja sovite on hyvin tasainen. Eli jos $\lambda = \infty$, saadaan tavallinen regressiosuoran sovite. Pieni λ taas hävittää sakkotermin vaikutusta, jolloin sovitettu käyrä kulkee mahdollisimman lähellä havaintopisteitä.

Nyt additiivinen malli voidaan esittää seuraavanlaisessa muodossa:

$$Y = \beta_0 + \sum_{j=1}^{p} s_j(X_j) + \varepsilon,$$

missä virhetermin ε keskiarvo on nolla ja varianssi σ^2 ja ne ovat riippumattomia muuttujista X_j. Edellä esitetty penalisoitu neliösumma voidaan määritellä tällöin seuraavasti:

$$PRSS(\beta_0, s_1, s_2, ..., s_p) = \sum_{i=1}^{N} \big(y_i - \beta_0 - \sum_{j=1}^{p} s_j(x_{ij})\big)^2 + \sum_{j=1}^{p} \lambda_j \int_a^b \big(s_j''(t)\big)^2 dt,$$

missä $\lambda_j \geq 0$ ovat säätöparametreja.

Otetaan käyttöön merkintä $\{y_i\}_1^N = \{y_1, y_2, ..., y_N\}$. Nyt addditiivinen malli sovitetaan seuraavanlaisella takaisinsovitusalgoritmilla (backfitting algorithm):

Takaisinsovitusalgortimi

Alustus: $\hat{\beta}_0 = \frac{1}{N}\sum_1^N y_i, \quad \hat{s}_j \equiv 0.$

Toisto: $j = (1, 2, ..., p), ..., (1, 2, ..., p), ...,$

$$\hat{s}_j \leftarrow S_j\left[\left\{y_i - \hat{\beta}_0 - \sum_{k\neq j}\hat{s}_k(x_{ik})\right\}_1^N\right],$$

$$\hat{s}_j \leftarrow \hat{s}_j - \frac{1}{N}\sum_{k=1}^N \hat{s}_j(x_{ij}).$$

Lopetusehto: Funktiot $\hat{s}_j$ kaikilla j muuttuvat vähemmän kuin ennalta määritelty raja.

Takaisinsovitusalgoritmissa S_j on kolmannen asteen tasoituisspline, joka esiteltiin aikaisemmin, ja se lisätään kohteille $\left\{y_i - \hat{\beta}_0 - \sum_{k\neq j}\hat{s}_k(x_{ik})\right\}_1^N$ muuttujan x_{ij} funktiona. Näin saadaan uusi estimaatti $\hat{s}_j$. Tämä tehdään jokaiselle selittäjälle vuorollaan käyttämällä muiden sen hetkisten funktioiden estimaatteja $\hat{s}_k$, kun lasketaan termit $\left\{y_i - \hat{\beta}_0 - \sum_{k\neq j}\hat{s}_k(x_{ik})\right\}_1^N$. Tätä jatketaan kunnes $\hat{s}_j$ on vakautettu kaikilla j.

3.2.2 Additiivinen logistinen regressio

Yleistetyissä additiivisissa malleissa käytetään painotettua takaisinsovitusalgoritmia. Seuraavaksi käydään läpi yleistetyn additiivisen mallin sovitus logistisen regression tapauksessa, mitä tässä työssä käytetään osallistumisaktiivisuuden ennustamiseen. Additiivinen logistinen regressio kuuluu yleistettyihin additiivisiin malleihin ja yleinen versio on esitetty Hastien ja Tibshiranin (1990) teoksen luvussa 6.

Additiivisessa logistisessa regressiossa vaste oletetaan binomijakautuneeksi ja malli esitetään seuraavalla tavalla:

$$log\frac{P(Y=1|X)}{P(Y=0|X)} = \beta_0 + s_1(X_1) + \cdots + s_p(X_p).$$

Funktiot $s_1, s_2, ..., s_p$ estimoidaan takaisinsovitusalgoritmilla ja additiivisen mallin sovitukseen tarvitaan painotettu hajontakuvion tasoitin. Yleisetty additiivinen malli sovitetaan käyttäen lokaalia pisteytysalgoritmia (local scoring algorithm), joka additiivisen logistisen regressiomallin tapauksessa on seuraava:

Lokaali pisteytysalgoritmi

Alustus: Laske aloitusarvot $\hat{\beta}_0 = log[\bar{y}/(1-\bar{y})]$, missä $\bar{y}$ on vastemuuttujan keskiarvo, ja asetetaan $\hat{s}_j \equiv 0$.

Määrittele $\hat{\eta}_i = \hat{\beta}_0 + \sum_j \hat{s}_j(x_{ij})$ ja $p_i = 1/[1 + exp(-\hat{\eta}_i)]$
Iteraatio:
(a) Muodosta väliaikainen tavoitemuuttuja

$$z_i = \hat{\eta}_i + \frac{(y_i - \hat{p}_i)}{\hat{p}_i(1-\hat{p}_i)}.$$

(b) Muodosta painot $w_i = \hat{p}_i(1-\hat{p}_i)$.

(c) Sovita additiivinen malli tavoitemuuttujiin z_i painoilla w_i käyttäen painotettua takaisinsovitusalgortimia. Näin saadaan uudet estimaatit $\hat{\beta}_0, \hat{s}_j$, kaikilla j

Lopetusehto: Funktioiden muutokset ovat pienempiä kuin ennalta määritelty raja.

Algoritmi koostuu kahdesta silmukasta. Sisimmäisessä silmukassa toteutetaan takaisinsovitusalgoritmi, jota käytetään lokaalin pisteytysalgoritmin sisällä sen jokaisella askeleella. Tämän painotetun takaisinsovitusalgoritmin estimaatteja käytetään, kun lasketaan uusia painoja ja uusi iteraatio alkaa pisteytysalgoritmissa. Tätä jatketaan, kunnes ennalta määritelty raja saavutetaan.

4 Puuttuvuuden mallintaminen

Puuttuvuuden mallintamista varten luotiin kolme mallia ja jokaisessa mallissa jätettiin aina yksi aikamuuttuja kerrallaan pois. Mallien estimointiin käytettiin luvussa 3 esiteltyä tapaa. Sovitettiin sellaiset mallit, joissa jokaisessa oli mukana kaikki muuttujien väliset interaktiot. Näihin malleihin päädyttiin siksi, koska haluttiin mallintaa juuri tätä nimenomaista tilannetta ja saada selville kaikki mahdolliset selittäjien vaikutukset, mikä oli mahdollista selittäjien pienen määrän vuoksi. Mallit esitetään graafisesti. Malleissa muuttujat *ALUE* sekä *VUOSI* faktoroitiin ja muuttujat *IKA* sekä *S.VUOSI* käsiteltiin additiivisina funktioina. Esitellään ennen tuloksien tulkintaa merkintätapa, jolla malliyhtälöt on esitetty tuloksissa.

4.1 R-notaatio

Esiteltävien mallien yhtälöissä käytetään R-notaatiota, joka on R-ohjelmiston käyttämä merkintätapa ja se on muunnelma McCullaghin ja Nelderin (1989, luku 3.4) esittelemästä notaatiosta. Malliyhtälöissä näkyvä *-operaattori tarkoittaa selittäjien välistä interaktiota ja $\sim$-operaattori kuvaa yhtäsuuruus merkintää. Tasoitusfunktiota merkitään yhtälöissä funktiolla $s()$. Lisäksi jokaiseen muuttujaan kuuluu oma regressiokertoimensa. Malli voidaan esittää esimerkiksi muodossa $Y \sim X * Z * W$, joka tarkoittaa

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 w + \beta_4 xz + \beta_5 xw + \beta_6 zw + \beta_7 xzw + \varepsilon,$$

missä muuttujat β_j kuvaavat regressiokertoimia ja ε virhetermiä.

4.2 Malli 1

Ensimmäisessä mallissa jätettiin ikämuuttuja *IKA* pois ja malliksi saatiin seuraava malli:

$$osallistuminen \sim ALUE * SUKUP * s(S.VUOSI) * VUOSI.$$

Mallista saadaan esitettyä kuvaajat osallistumisen todennäköisyyksistä tutkimusvuosittain syntymävuoden funktiona jokaiselle alue-sukupuoli-ositteelle. Saadaan siis vastattua kysymykseen, miten osallistumisen todennäköisyys on muuttunut tutkimusvuosissa, kun otetaan huomioon syntymävuoden vaikutus. Luvun 5 kuvissa 1-5 on esitelty tulokset jokaiselle alueelle sekä naisille että miehille.

Naisilla on pääsääntöisesti korkeammat osallistumistodennäköisyydet kuin miehillä, mikä huomattiin jo aineiston alkutarkasteluissa. Kuvaajista huomataan

myös heti se, että aikaisempien tutkimusvuosien kohdalla osallistumisen todennäköisyydet ovat korkeammalla kuin myöhempien tutkimusvuosien kohdalla. Trendi on myös laskeva, kun syntymävuosi kasvaa. Vuonna 1997 on suurin lasku osallistumistodennäköisyyksissä, kun syntymävuosi kasvaa, verrattuna muihin tutkimusvuosiin. Tällä vuonna on suurin ero osallistumisaktiivisuudessa ennen vuotta 1950 syntyneiden ja vuoden 1950 jälkeen syntyneiden tutkittavien välillä.

Naisten ja miesten käyrät lähtevät melko samalta tasolta, kun katsotaan ennen vuotta 1950 syntyneitä tutkittavia. Miesten todennäköisyydet laskevat paljon naisten todennäköisyyksiä nopeammin, kun syntymävuosi kasvaa. Täten näyttäisi siltä, että vuoden 1950 jälkeen syntyneet miehet eivät ole kovin innokkaita osallistumaan. Poikkeuksen kuitenkin aiheuttavat Oulun läänin miehet (luku 5, kuva 5); heidän kuvaajansa eroaa selvästi muista alueista. Heillä jokaisen tutkimusvuoden osallistumistodennäköisyydet lähtevät melko samalta tasolta ennen vuotta 1950 syntyneillä ja tutkimusvuodet erottuvat aina vain selkeämmin, kun siirrytään syntymävuosissa eteenpäin. Lisäksi vuonna 2012 vuoden 1950 jälkeen syntyneiden osallistumistodennäköisyydet olivat korkeimmillaan, mutta vuonna 1997 sen sijaan matalimmillaan.

Helsingin/Vantaan (luku 5, kuva 4) alueella on ollut matalimmat osallistumistodennäköisyydet muihin tutkimusalueisiin verrattuna. Tämän alueen naisilla vuosi 2002 on poikkeava, koska tällä tutkimusvuonna osallistumisen todennäköisyydet ovat kasvaneet, kun syntymävuosi kasvaa. Sen sijaan muilla vuosina osallistumistodennäköisyydet laskevat, kun syntymävuosi kasvaa. Miehillä poikkeavaa muihin alueisiin verrattuna on se, että vuonna 2012 on korkeimmat osallistumistodennäköisyydet läpi syntymävuosien. Ennen vuotta 1950 syntyneillä tutkittavilla on aina seuraavana tutkimusvuonna matalammat osallistumistodennäköisyydet kuin edellisenä tutkimusvuonna. Asetelma vaihtuu noin syntymävuoden 1950 kohdalla, jolloin varhaisempina tutkimusvuosina osallistumisen todennäköisyydet ovat matalammalla kuin myöhempinä tutkimusvuosina.

Suurimmat osallistumistodennäköisyydet ovat Pohjois-Karjalassa (luku 5, kuva 1) ja Pohjois-Savossa (luku 5, kuva 2). Vähiten muuttuvat osallistumistodennäköisyydet näyttäisivät olevan Oulun läänin naisilla (luku 5, kuva 5) ja voimakkaimmin laskevat osallistumisen todennäköisyydet Helsingin/Vantaan miehillä (luku 5, kuva 4). Suurimmat erot tutkimusvuosien välillä osallistumistodennäköisyyksissä näyttäisi olevan Turussa/Loimaalla (luku 5, kuva 3); kuvaajissa pudotus osallistumisen todennäköisyyksissä aina seuraavaan tutkimusvuoteen vaikuttaisi olevan isompi kuin muilla alueilla.

4.3 Malli 2

Toisessa mallissa jätettiin pois aikamuuttuja *VUOSI*, eli tutkimusvuosi. Malli on

$$osallistuminen \sim ALUE * SUKUP * s(S.VUOSI) * s(IKA).$$

Tästä mallista tehtiin kolmenlaisia kuvaajia. Ensin tehtiin todennäköisyyskäyräkuvaajat, joissa luvun 5 kuvissa 6-10 osallistumisen todennäköisyys on esitetty ikäryhmittäin syntymävuoden funktiona ja luvun 5 kuvissa 11-15 syntymävuosittain iän funktiona. Luvun 5 kuvissa 16-20 on esitetty vakiotodennäköisyyskäyräkuvaajat, joissa osallistumistodennäköisyys on esitetty iän ja syntymävuoden funktiona, jolloin todennäköisyys on vakio kullakin käyrällä. Kaikki kuvaajat on esitetty jokaiselle alue-sukupuoli-ositteelle.

4.3.1 Todennäköisyyskäyrät

Osallistumisen todennäköisyys ikäryhmittäin syntymävuoden funktiona

Tarkastelussa oli 50 eri ikää, joten käyrät piirrettiin viiden vuoden välein, jotta vältyttiin kuvaajien epäselvyydeltä. Näin saatiin piirrettyä korkeintaan 11 käyrää yhteen kuvaajaan. Nämä käyrät kertovat, miten osallistumisen todennäköisyys on muuttunut ikäryhmissä, kun otetaan huomioon syntymävuoden vaikutus.

Kuvaajista (luku 5, kuvat 6-10) nähdään, että mitä myöhäisempi syntymävuosi ja mitä nuorempi tutkittava sitä pienemmät ovat osallistumistodennäköisyydet. Samansuuntaisia tuloksia saatiin myös mallissa 1. Miehillä osallistumistodennäköisyydet ovat matalampia ja ne laskevat voimakkaammin verrattuna naisten osallistumistodennäköisyyksiin, kun syntymävuosi kasvaa. Kuvaajissa miesten todennäköisyydet ovat melko samalla tasolla naisten todennäköisyyksien kanssa, kun katsotaan syntymävuosien alkupäätä, mutta osallistumistodennäköisyydet laskevat paljon voimakkaammin verrattuna naisten osallistumistodennäköisyyksiin, kun siirrytään syntymävuosissa eteenpäin.

Poikkeuksen aiheuttavat Oulun läänin miehet (luku 5, kuva 10), joilla ikäkäyrät laskevat eri järjestyksessä kuin muilla tutkimusalueilla osallistumistodennäköisyyksiä verrattaessa. Tämä tarkoittaa samaa kuin mallissa 1 eli vuoden 1950 jälkeen syntyneillä viimeinen tutkimusvuosi ei päädykään kaikkein matalimmalle osallistumistodennäköisyyksissä niin kuin muilla tutkimusalueilla.

Pohjois-Karjalan (luku 5, kuva 6) ja Helsingin/Vantaan (luku 5, kuva 9) miesten kuvaajien muoto on samankaltainen. Käyrät ovat suurelta osin päällekkäin, mikä kertoo, etteivät vierekkäiset ikäryhmät eroa merkittävästi toisistaan ja lasku osallistumistodennäköisyyksissä on voimakasta.

Korkeimmat osallistumistodennäköisyydet ovat Pohjois-Karjalassa (luku 5, kuva 6) ja Pohjois-Savossa (luku 5, kuva 7) verrattuna muihin tutkimusalueisiin. Vähiten muuttuvat käyrät ovat Oulun läänin naisilla (luku 5, kuva 10) ja matalimmat sekä voimakkaimmin laskevat Helsingin/Vantaan (luku 5, kuva 9) ja Pohjois-Karjalan (luku 5, kuva 6) miehillä.

Osallistumisen todennäköisyys syntymävuosittain iän funktiona

Tarkastelussa oli yhteensä 70 eri syntymävuotta, joten kuvaajien selkiyttämiseksi käyrät piirrettiin viiden vuoden välein, jolloin saatiin korkeintaan 15 käyrää yhteen kuvaajaan. Näistä todennäköisyyskäyristä saadaan vastaus siihen, miten eri vuosina syntyneiden osallistumistodennäköisyys on muuttunut, kun otetaan huomioon iän vaikutus.

Näissä kuvaajissa (luku 5, kuvat 11-15) kaikkien käyrien yhteinen trendi on kasvava, mutta kuvaajat kuitenkin tukevat edellisiä tulkintoja. Kun ikä kasvaa, niin osallistumistodennäköisyys on sitä korkeampi mitä aikaisemmin on syntynyt ja siten kuvaajien käyrät suuntautuvat alhaalta ylöspäin. Näyttää myös siltä että, kun katsotaan samaa syntymävuosikäyrää, niin osallistumistodennäköisyydet laskevat iän kasvaessa. Tämä tarkoittaa sitä, että lähempänä nykypäivää tutkittavat osallistuvat epätodennäköisemmin tutkimuksiin.

Pohjois-Karjalan (luku 5, kuva 11) ja Helsingin/Vantaan (luku 5, kuva 14) miehillä syntymävuosikäyrät ovat melko muuttumattomia. Nyt iällä ei näyttäisi olevan niin suurta merkitystä osallistumistodennäköisyyksiin, vaan syntymävuoden kasvu vain vaikuttaisi osallistumistodennäköisyyksiin laskevasti. Oulun läänin miesten (luku 5, kuva 15) tapauksessa on mielenkiintoista, että heillä samalla syntymävuosikäyrällä osallistumistodennäköisyydet kasvavat iän kasvaessa. Täten jälleen kerran, heillä lähempänä nykypäivää osallistumisen todennäköisyys on suurempaa vuoden 1950 jälkeen syntyneillä tutkittavilla.

4.3.2 Vakiotodennäköisyyskäyrät

Vakiotodennäköisyyskäyrillä (luku 5, kuvat 16-20) saadaan tarkasteltua iän ja syntymävuoden yhteinen vaikutus osallistumistodennäköisyyteen. Tällaisia kuvaajia kutsutaan korkeuskäyriksi (contour plot), joilla voidaan tarkastella kahta muuttujaa kolmiuloitteisesti. Kuvaajissa punaiset alueet ovat korkeammalla kuin siniset alueet.

Suurimmassa osassa kuvaajista (luku 5, kuvat 16-20) trendinä on, että kun ikä ja syntymävuosi kasvavat, niin osallistumistodennäköisyys pienenee. Syntymävuoden kasvaessa osallistumistodennäköisyys laskee voimakkaammin kuin iän kasvaessa. Naisilla osallistumistodennäköisyydet ovat korkeammat ja laskevat hitaammin kuin miehillä.

Samat poikkeukset ovat nähtävissä vakiotodennäköisyyskäyrissä niin kuin edellisissä malleissakin; Pohjois-Karjalan, Helsingin/Vantaan ja Oulun läänin miesten käyrät. Kun tarkastellaan syntymävuosia ennen vuotta 1950 noin syntymävuoteen 1955, Pohjois-Karjalan miesten (luku 5, kuva 16) osallistumistodennäköisyys laskee nuorimpien tutkittavien kohdalla. Lasku ei kuitenkaan ole kovin voimakasta. Syntymävuodesta 1955 eteenpäin iän kasvaessa osallistumistodennäköisyys laskee. Helsingin/Vantaan miehillä (luku 5, kuva 19) ikä ei vaikuta juurikaan osallistumistodennäköisyyteen, vaan osallistumistodennäköisyys laskee, kun syntymävuosi kasvaa. Oulun läänin miesten (luku 5, kuva 20) kuvaaja poikkeaa muiden alueiden kuvaajista siten, että osallistumistodennäköisyys laskee iän ja syntymävuoden pienentyessä.

Turun/Loimaan kuvaajissa (luku 5, kuva 18) nähdään että, kun ikä ja syntymävuosi kasvavat osallistumistodennäköisyydet laskevat nopeammin kuin muilla alueilla. Siten pudotus osallistumisen todennäköisyyksissä aina seuraavaan tutkimusvuoteen on suurempi kuin muilla tutkimusalueilla. Sama näkyi myös mallin 1 kohdalla.

4.4 Malli 3

Viimeisenä käsitellään malli, jossa on jätetty pois aikamuuttuja *S.VUOSI* eli syntymävuosi. Malli näyttää seuraavalta:

$$osallistuminen \sim ALUE * SUKUP * VUOSI * s(IKA).$$

Mallin kuvaajissa saadaan esitettyä osallistumisen todennäköisyys tutkimusvuosittain jokaiselle alue-sukupuoli-ositteelle iän funktiona. Kuvaajista nähdään, miten osallistumistodennäköisyys on muuttunut tutkimusvuosissa, kun otetaan huomioon tutkittavien iän vaikutus. Luvun 5 kuvissa 21-25 on esitetty tulokset jokaisen alueen naisille ja miehille.

Kuvaajien (luku 5, kuvat 21-25) trendi on nouseva eli, mitä vanhempi tutkittava on sitä korkeampi osallistumistodennäköisyys on jokaisena tutkimusvuotena. Pääsääntöisesti aikaisempina tutkimusvuosina on korkeammat osallistumistodennäköisyydet kuin myöhempinä tutkimusvuosina. Kuten mallissa 1 nähtiin, että tutkimusvuonna 1997 oli suurin ero ennen vuotta 1950 syntyneiden ja vuoden 1950 jälkeen syntyneiden välillä osallistumistodennäköisyyksissä verrattuna muihin tutkimusvuosiin, sama asia nähdään myös tässä mallissa. Vanhimmilla tutkittavilla on huomattavasti korkeammat osallistumistodennäköisyydet kuin nuorimmilla tutkittavilla vuonna 1997 verrattuna muihin tutkimusvuosiin.

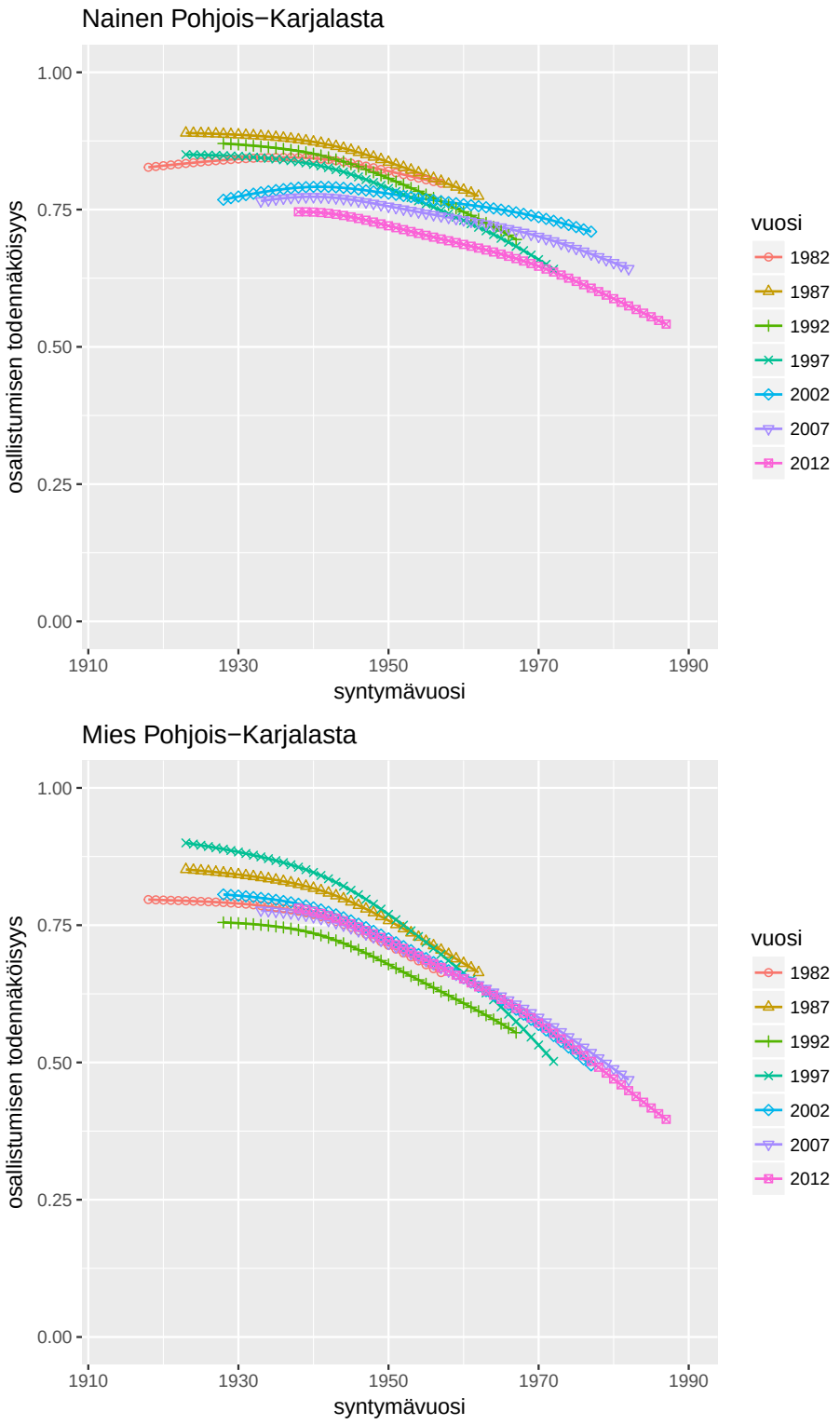
Naisilla on vähemmän muuttuvat käyrät ja korkeammat osallistumistodennäköisyydet kuin miehillä. Nuorilla miehillä osallistumisen todennäköisyydet ovat selvästi matalampia kuin nuorilla naisilla, kun taas vanhimmilla tutkittavilla naisilla ja miehillä osallistumistodennäköisyydet ovat lähes samalla tasolla.

Nuorimmilla tutkittavilla on isommat erot osallistumistodennäköisyyksissä tutkimusvuosien välillä verrattuna vanhimpiin tutkittaviin. Tutkimusvuosien erot tasoittuvat, kun ikä kasvaa. Erityisen hyvin tämä nähdään Pohjois-Savon miesten kuvaajasta (luku 5, kuva 22).

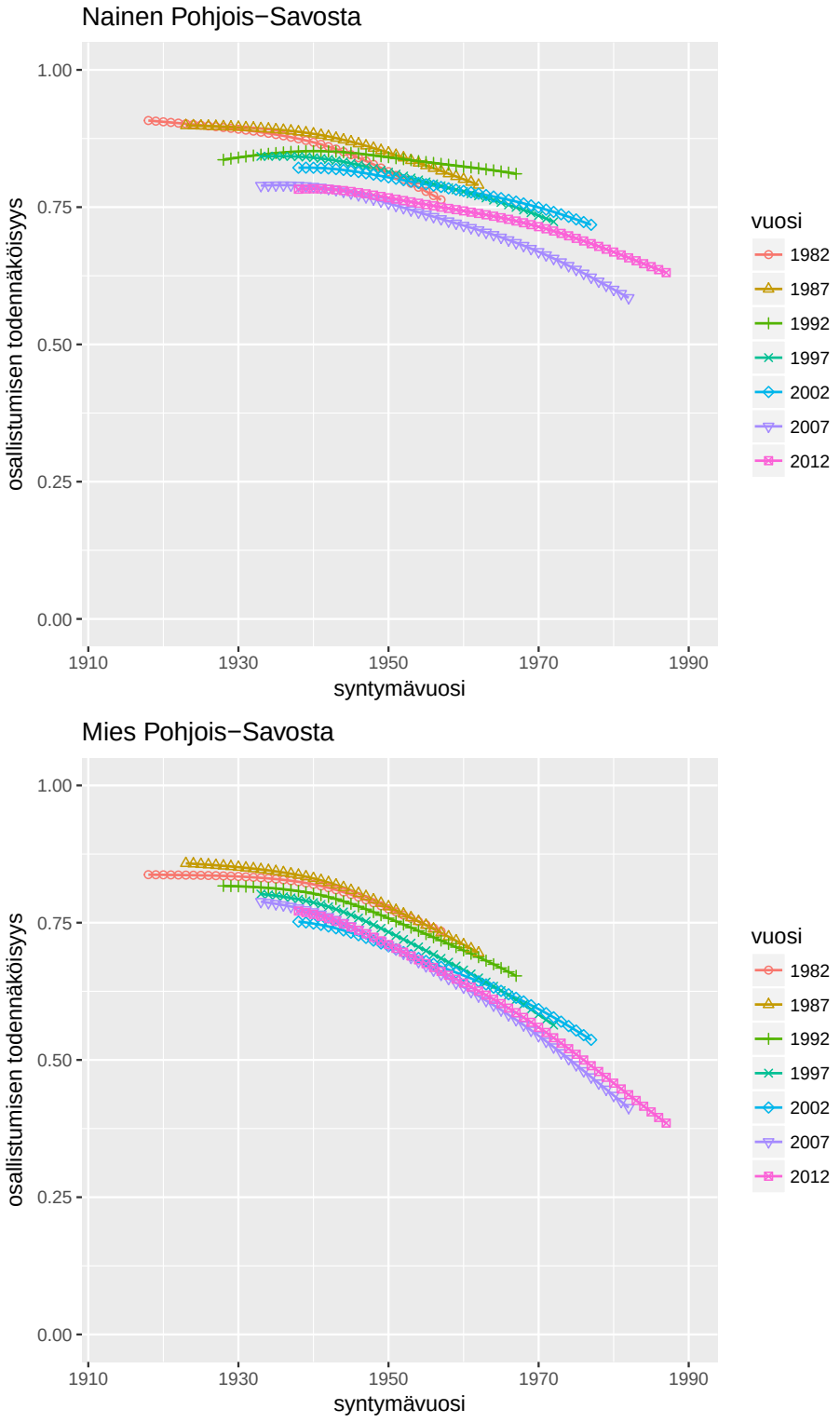
Helsingissä/Vantaalla vanhimmilla tutkittavilla miehillä (luku 5, kuva 24) ei tutkimusvuosi 2012 ole selkeästi alimpana osallistumistodennäköisyyksissä, niin kuin muilla tutkimusalueilla, vaan vuodet 2002 ja 2007. Näillä vuosilla vanhimmilla ja nuorimmilla tutkittavilla ei ole niin suurta eroa osallistumistodennäköisyyksissä, kun verrataan vuoteen 2012, jolla ero on todella suuri vanhimpien ja nuorimpien tutkittavien välillä. Helsingin/Vantaan Naisilla (luku 5, kuva 24) tutkimusvuotena 2002 nuorimmat tutkittavat ovat osallistuneet tutkimukseen paremmin kuin vanhimmat tutkittavat.

Oulun läänissä (luku 5, kuva 25) kahdella viimeisellä tutkimusvuotena nuorimmilla osallistujilla ei ole tapahtunut juurikaan muutosta tutkimukseen osallistumisessa. Vanhimmilla tutkittavilla taas on selkeä pudotus osallistumistodennäköisyyksissä kahden viimeisen tutkimusvuoden välillä.

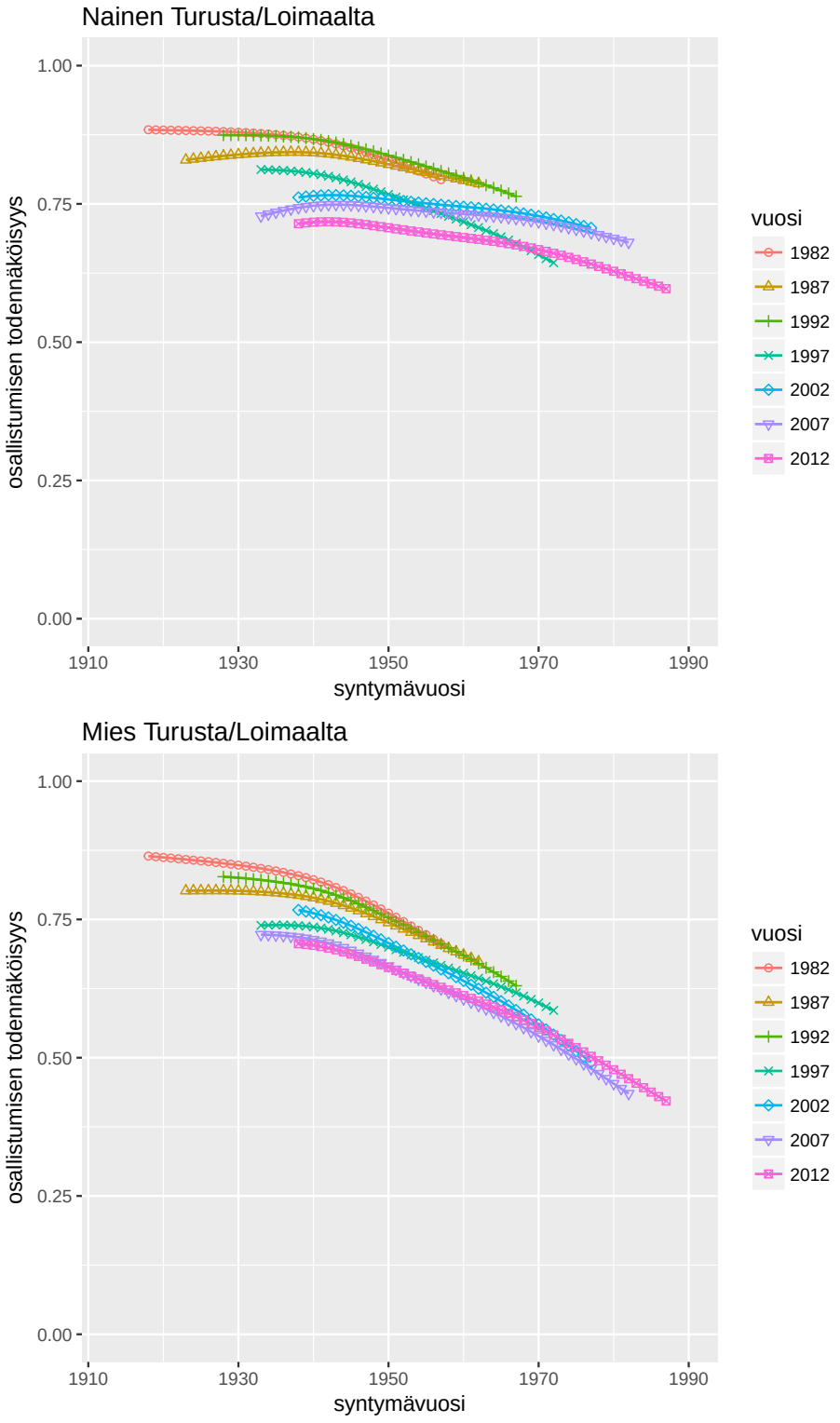
5 Kuvaajat



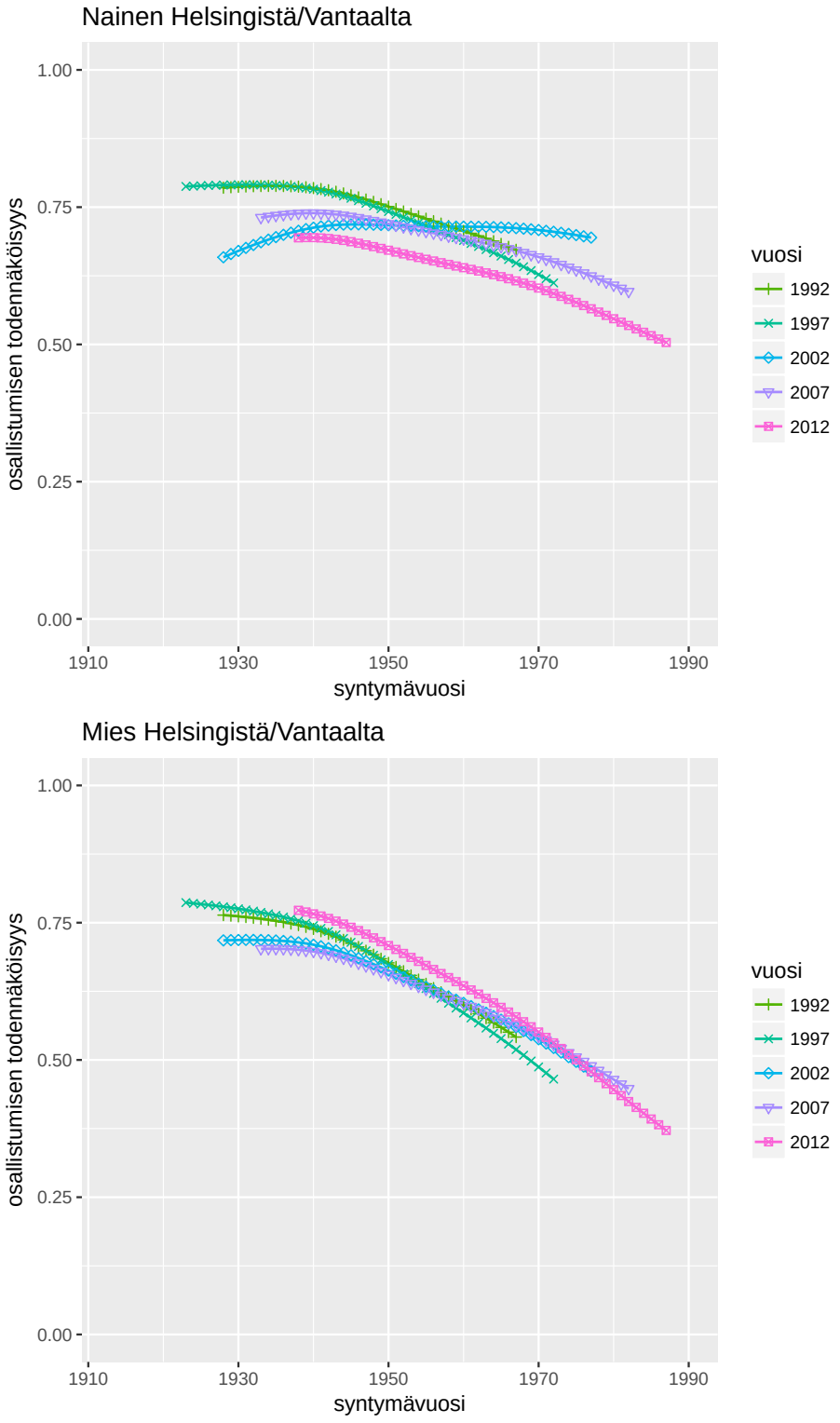
Kuva 1: Ensimmäisen mallin ja alueen kaksi osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



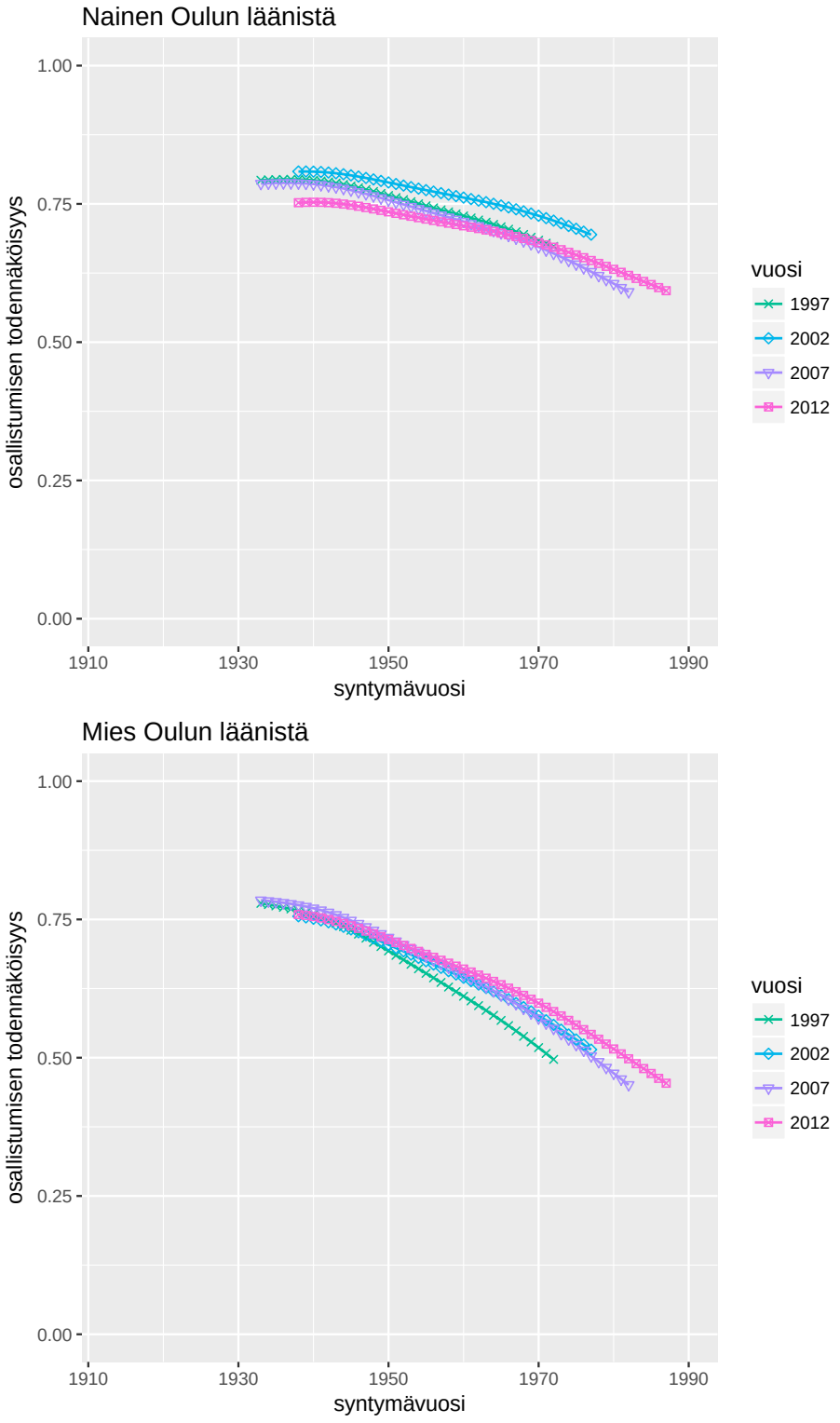
Kuva 2: Ensimmäisen mallin ja alueen kolme osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



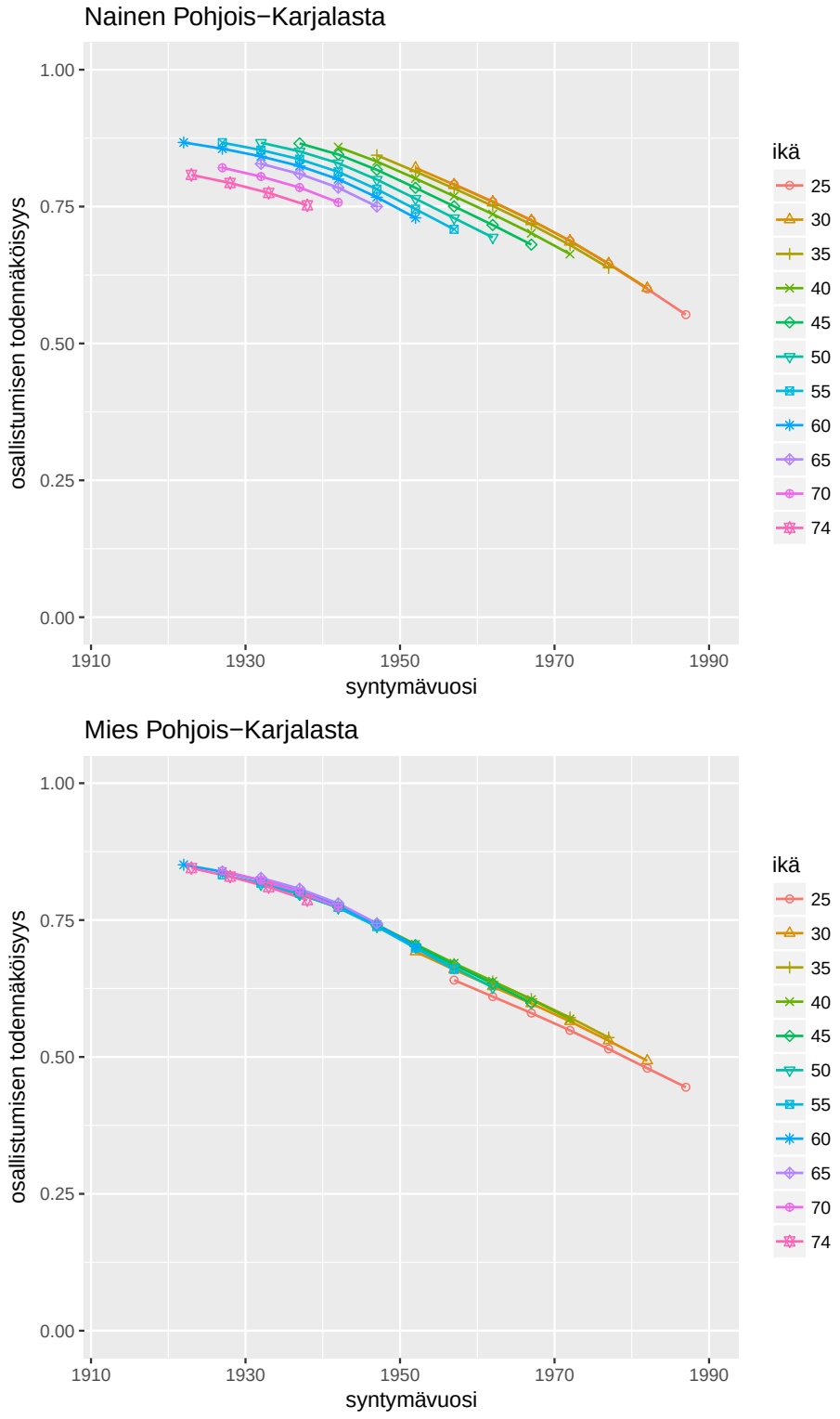
Kuva 3: Ensimmäisen mallin ja alueen neljä osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



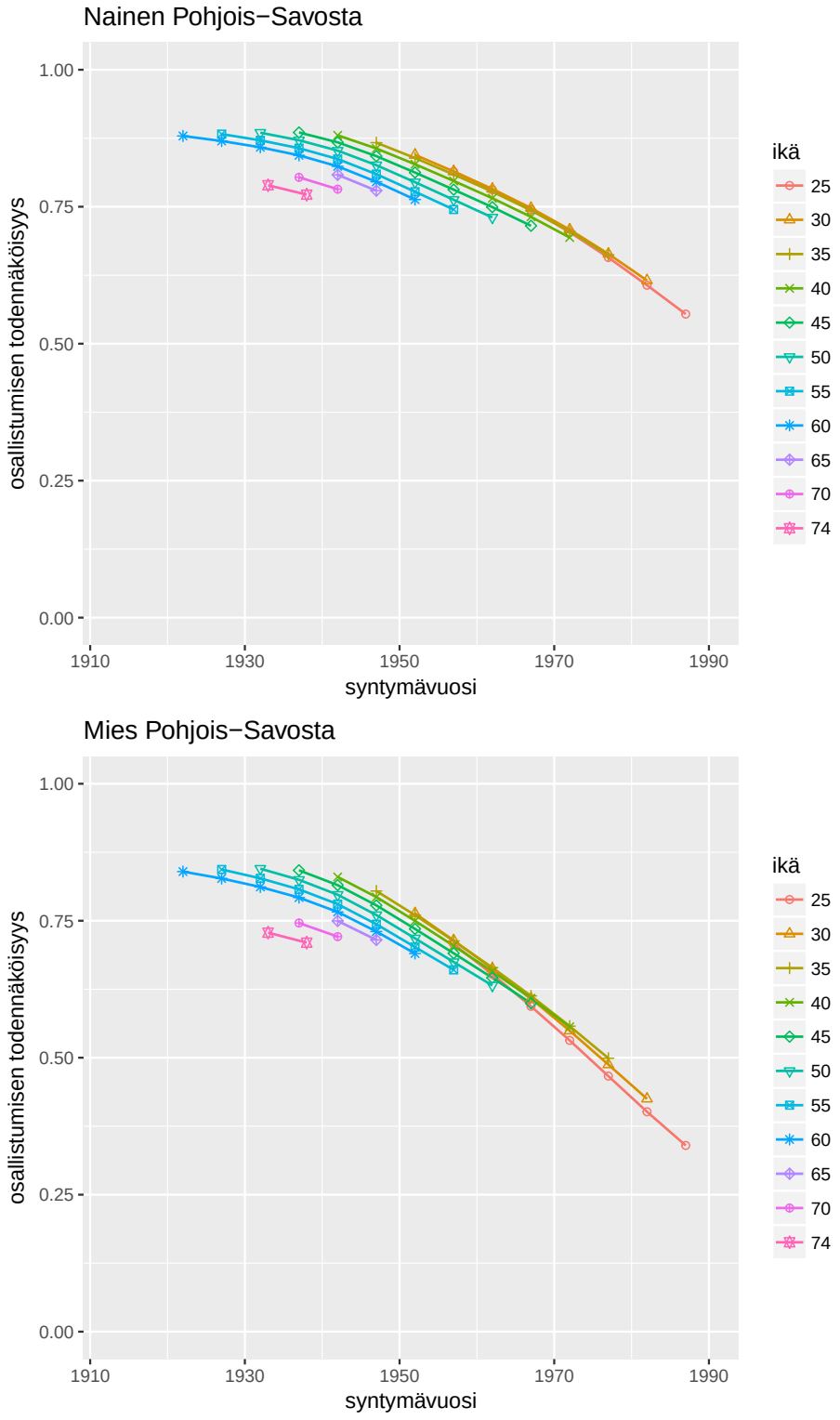
Kuva 4: Ensimmäisen mallin ja alueen viisi osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



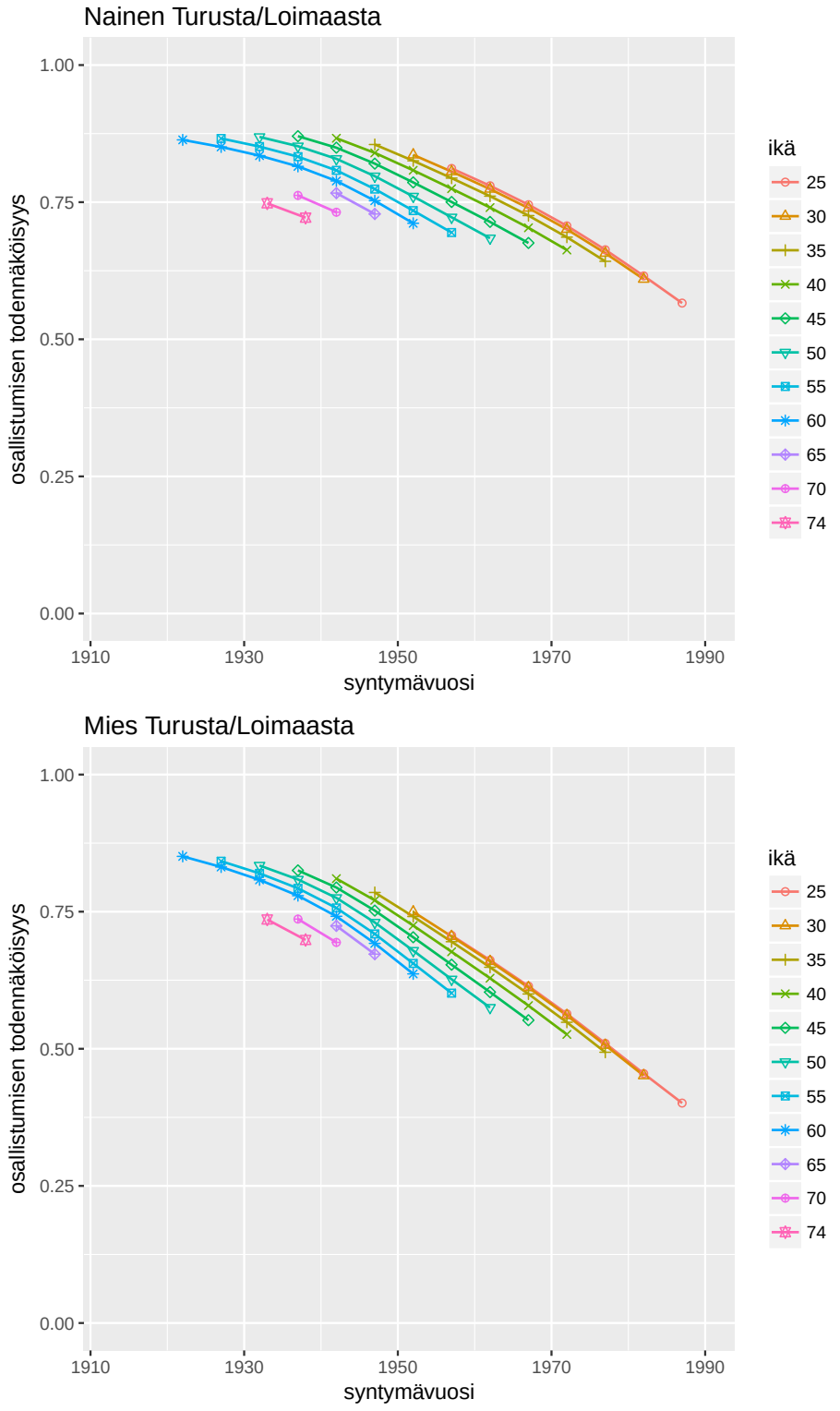
Kuva 5: Ensimmäisen mallin ja alueen kuusi osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



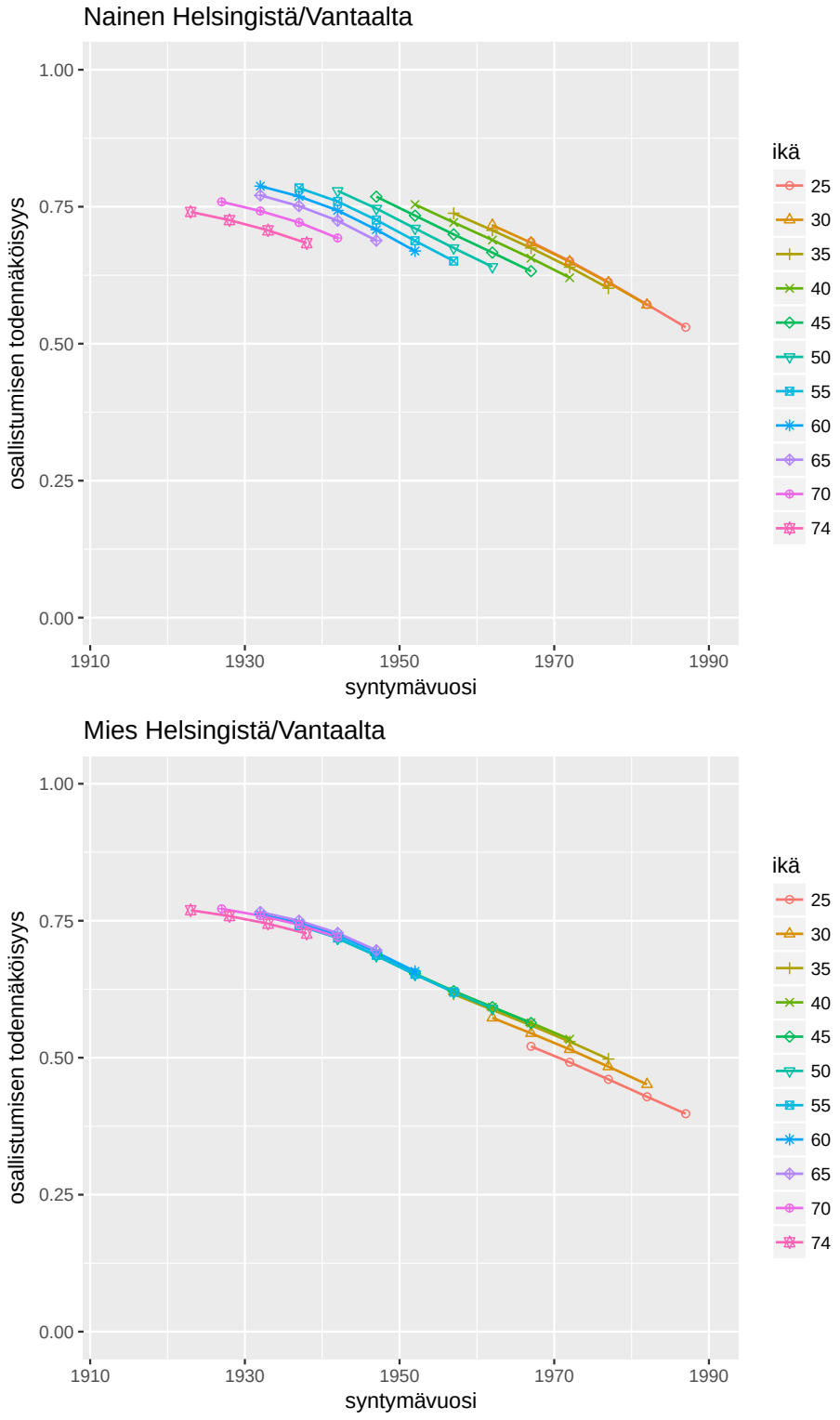
Kuva 6: Toisen mallin ja alueen kaksi todennäköisyyskäyrät ikäryhmittäin syntymävuoden funktiona naisille ja miehille



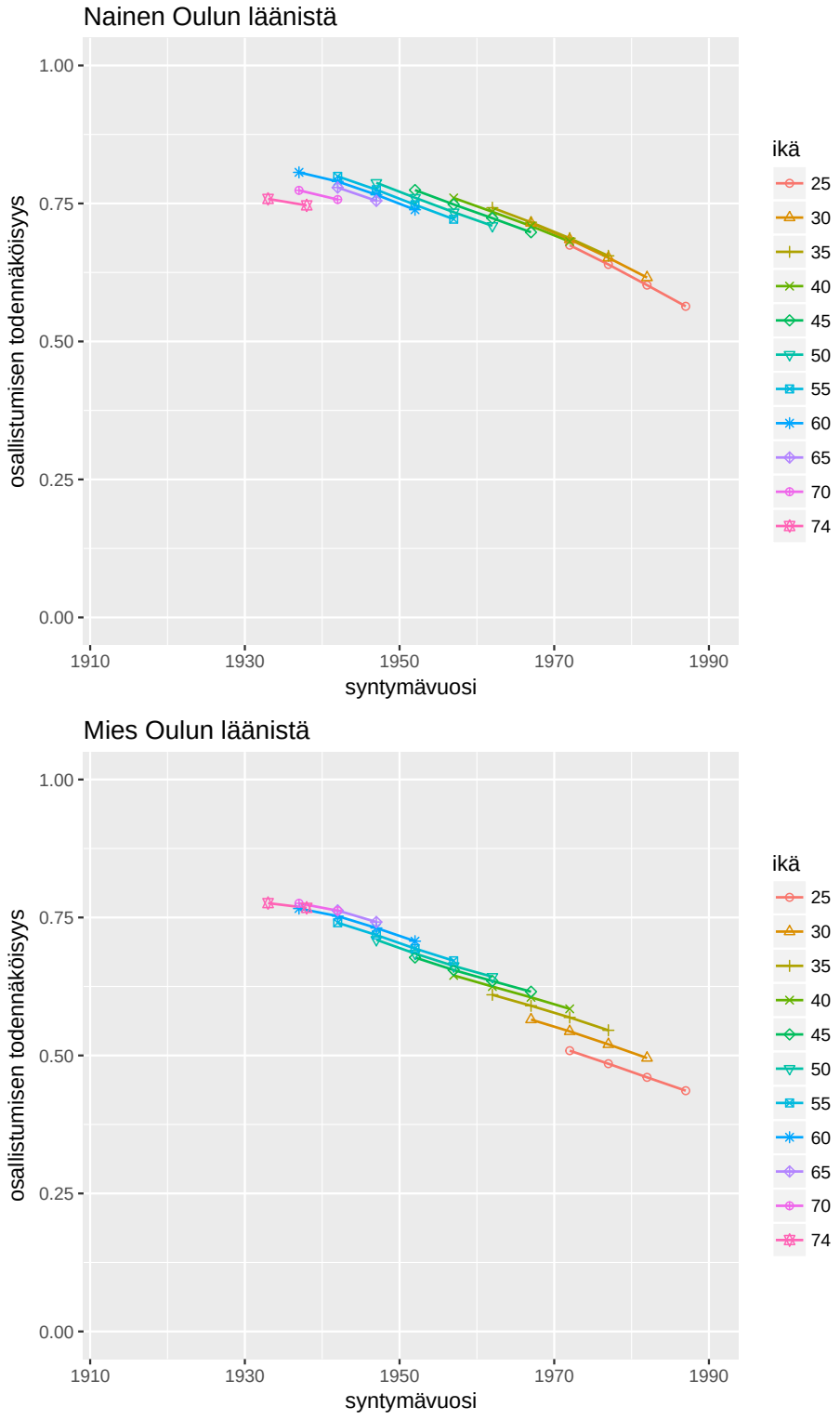
Kuva 7: Toisen mallin ja alueen kolme todennäköisyyskäyrät ikäryhmittäin syntymävuoden funktiona naisille ja miehille



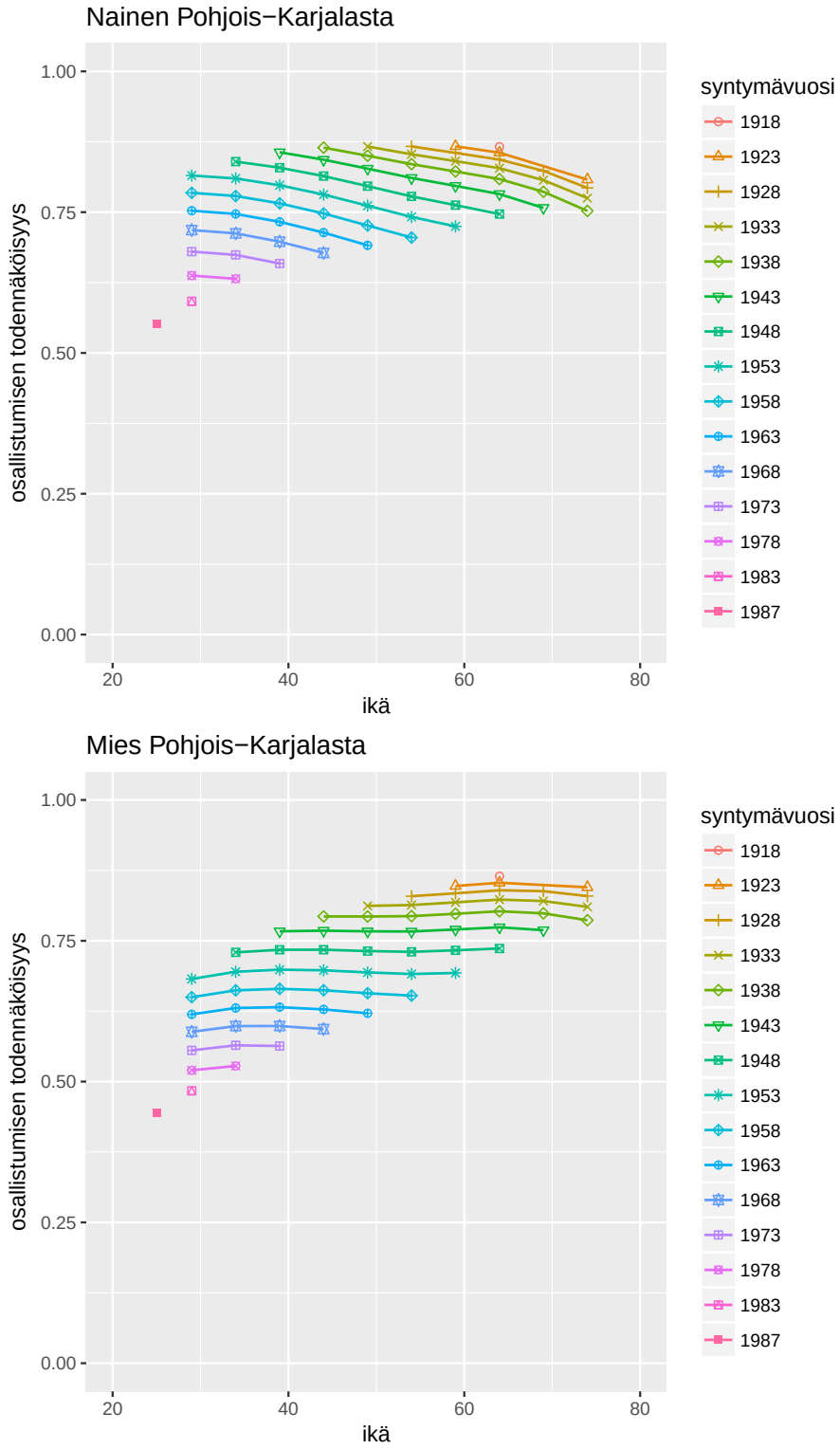
Kuva 8: Toisen mallin ja alueen neljä todennäköisyyskäyrät ikäryhmittäin syntymävuoden funktiona naisille ja miehille



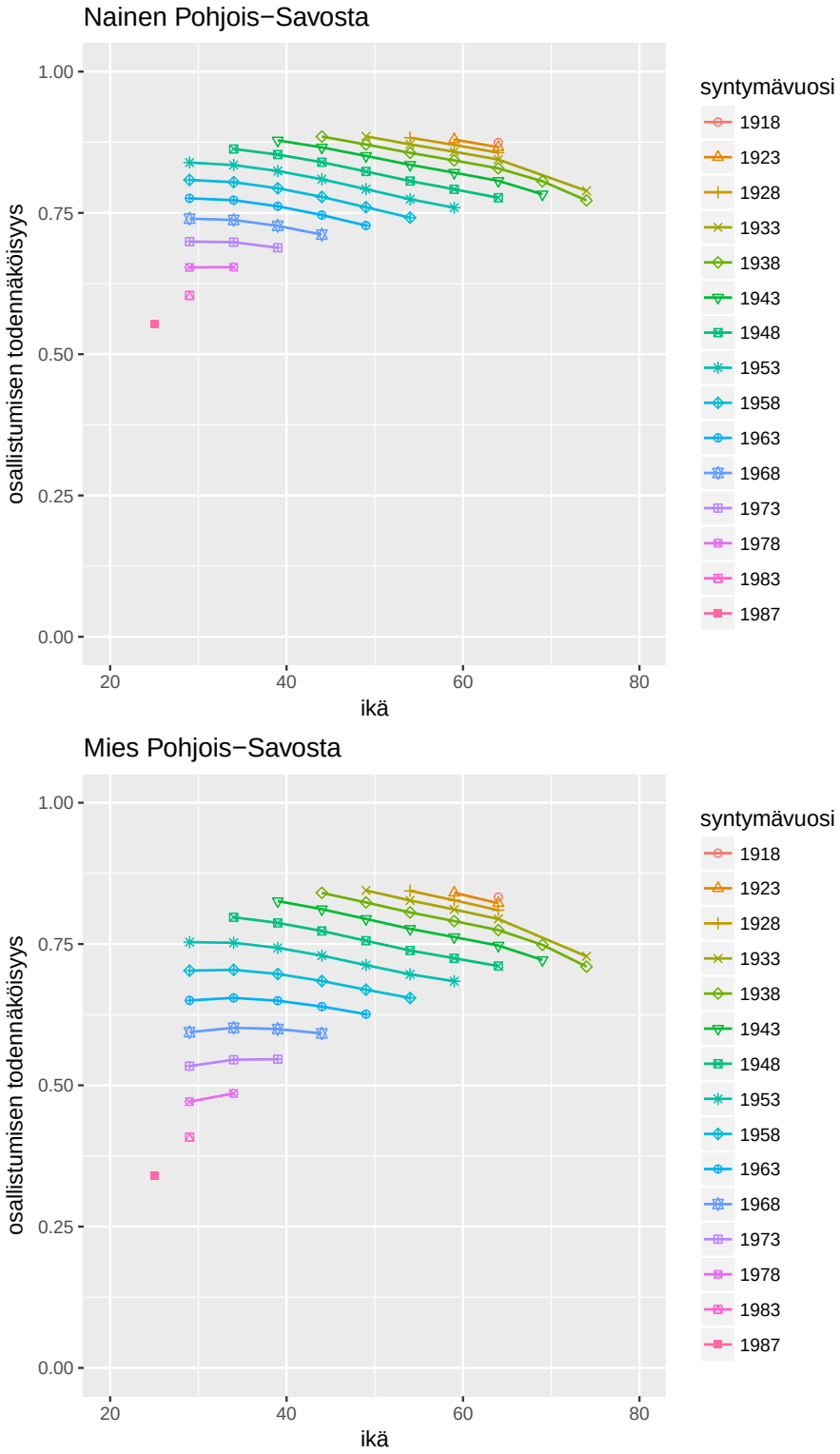
Kuva 9: Toisen mallin ja alueen viisi todennäköisyyskäyrät ikäryhmittäin syntymävuoden funktiona naisille ja miehille



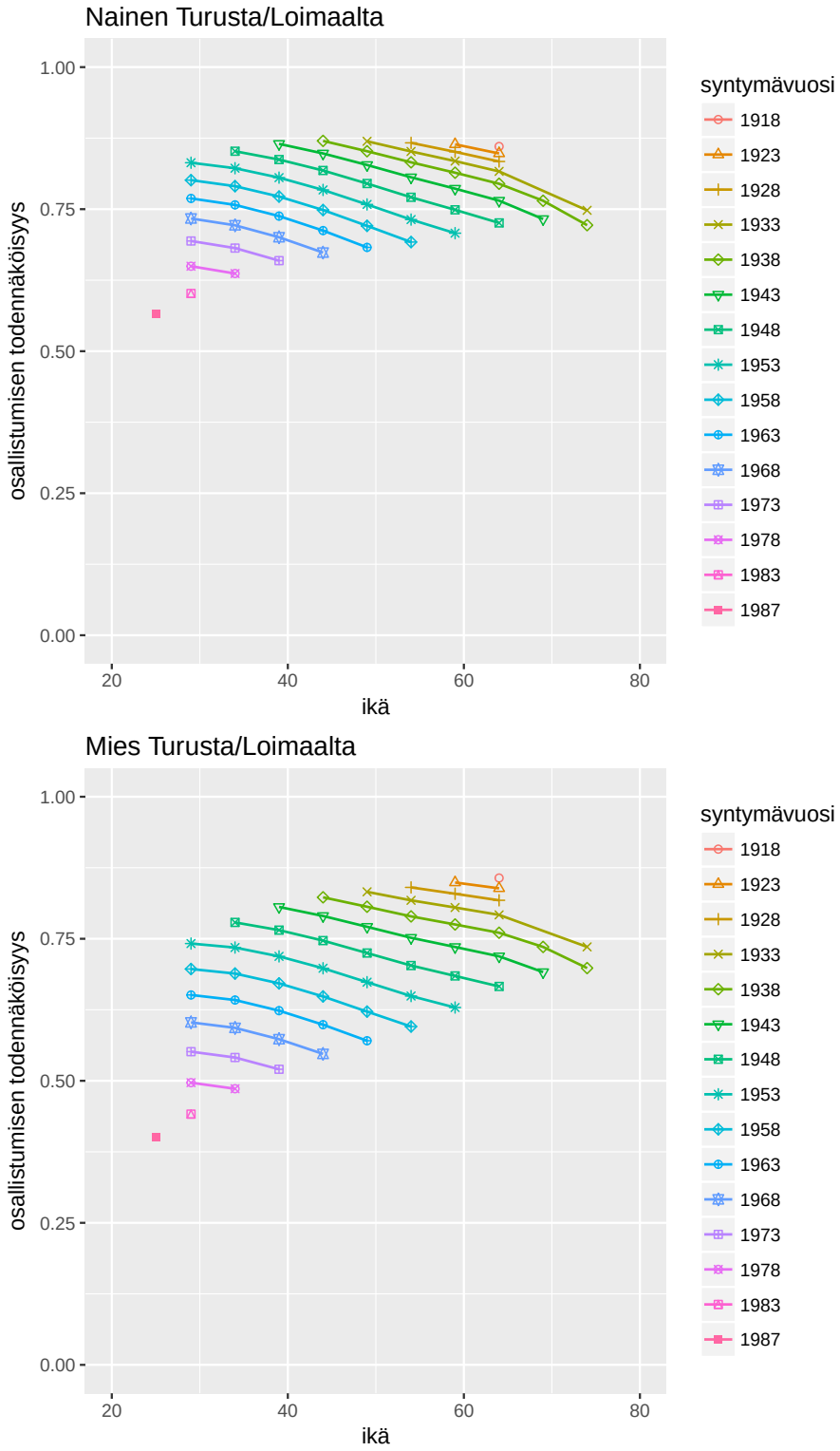
Kuva 10: Toisen mallin ja alueen kuusi todennäköisyyskäyrät ikäryhmittäin syntymävuoden funktiona naisille ja miehille



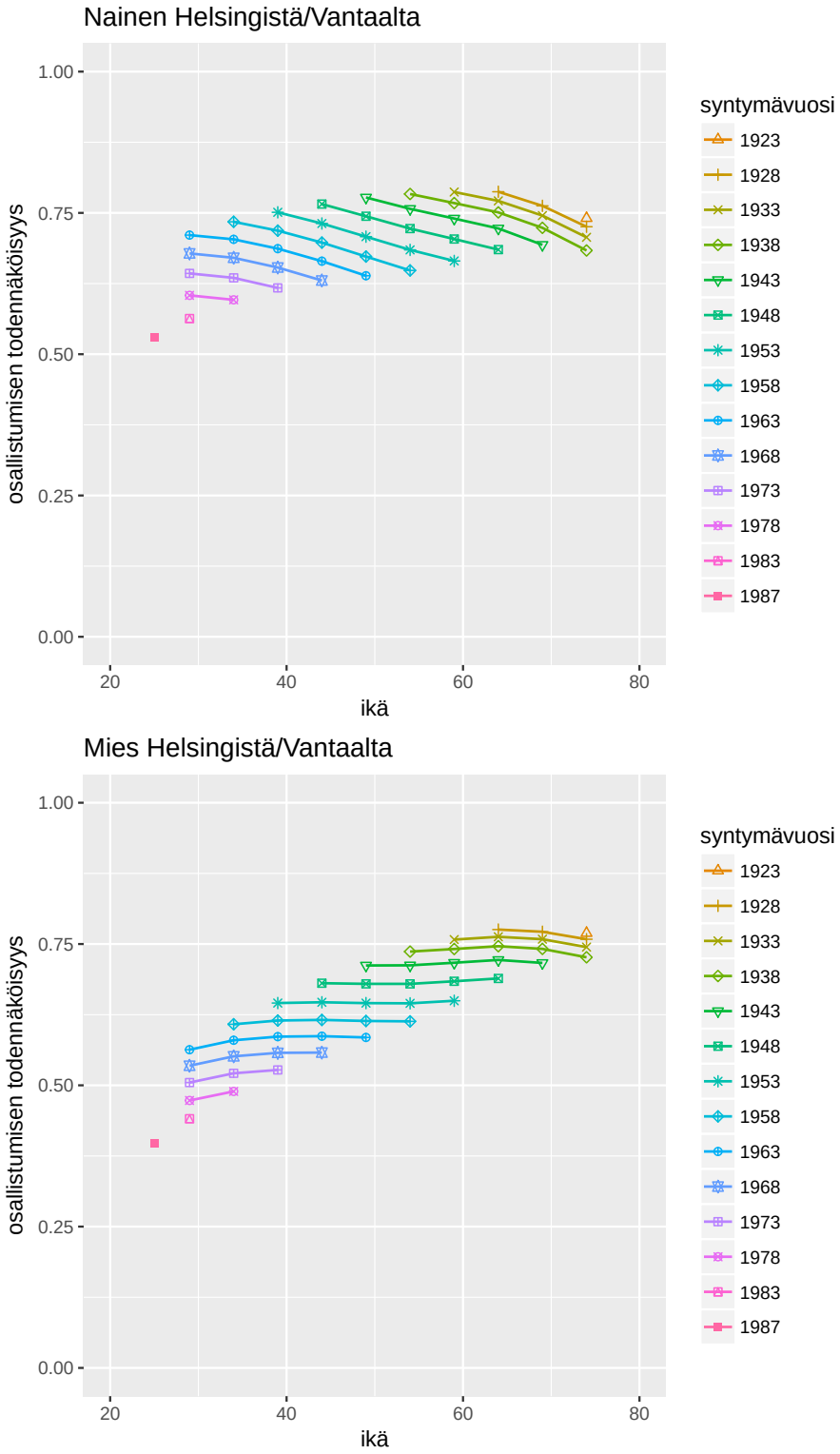
Kuva 11: Toisen mallin ja alueen kaksi todennäköisyyskäyrät syntymävuosittain iän funktiona naisille ja miehille



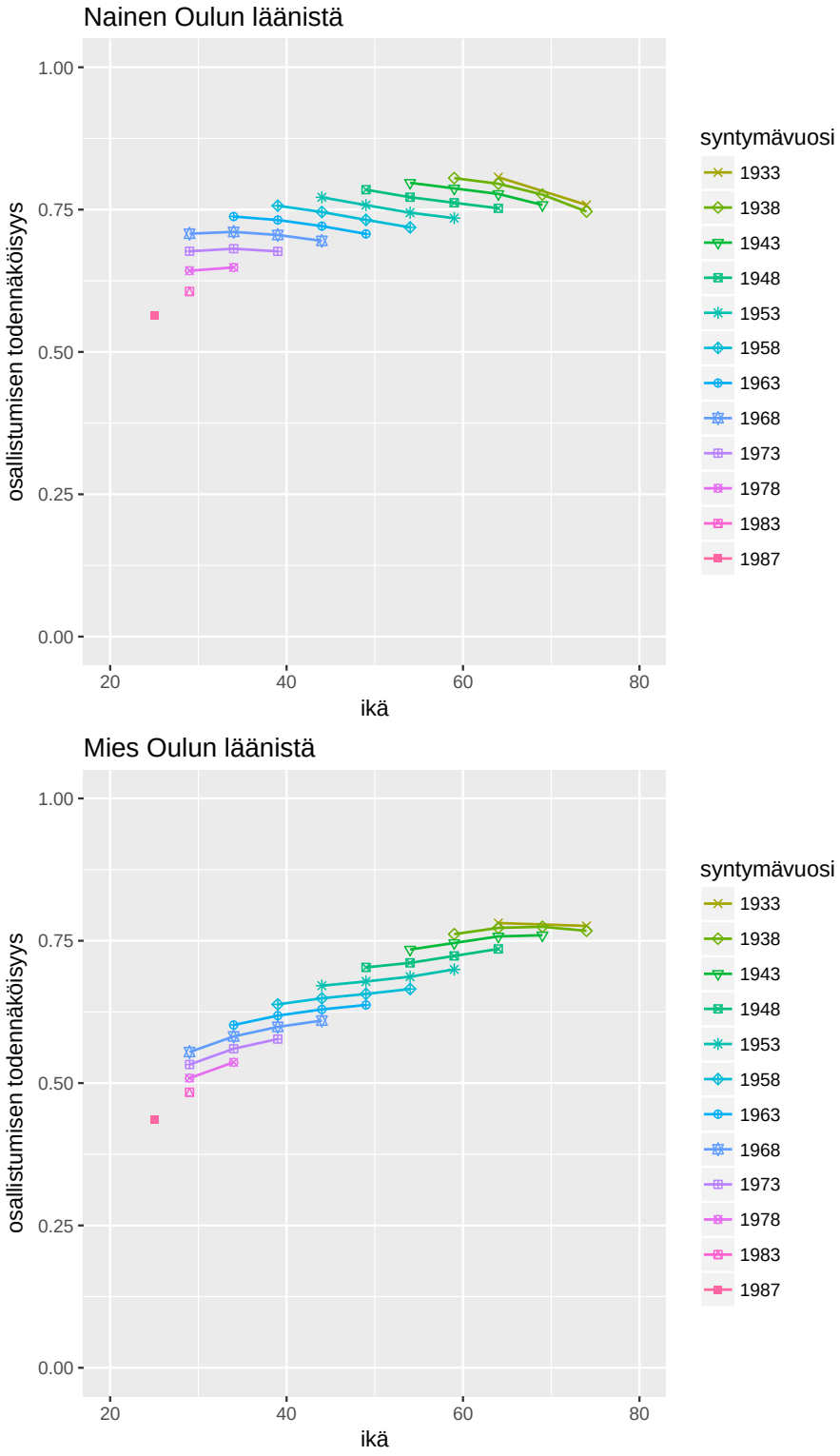
Kuva 12: Toisen mallin ja alueen kolme todennäköisyyskäyrät syntymävuosittain iän funktiona naisille ja miehille



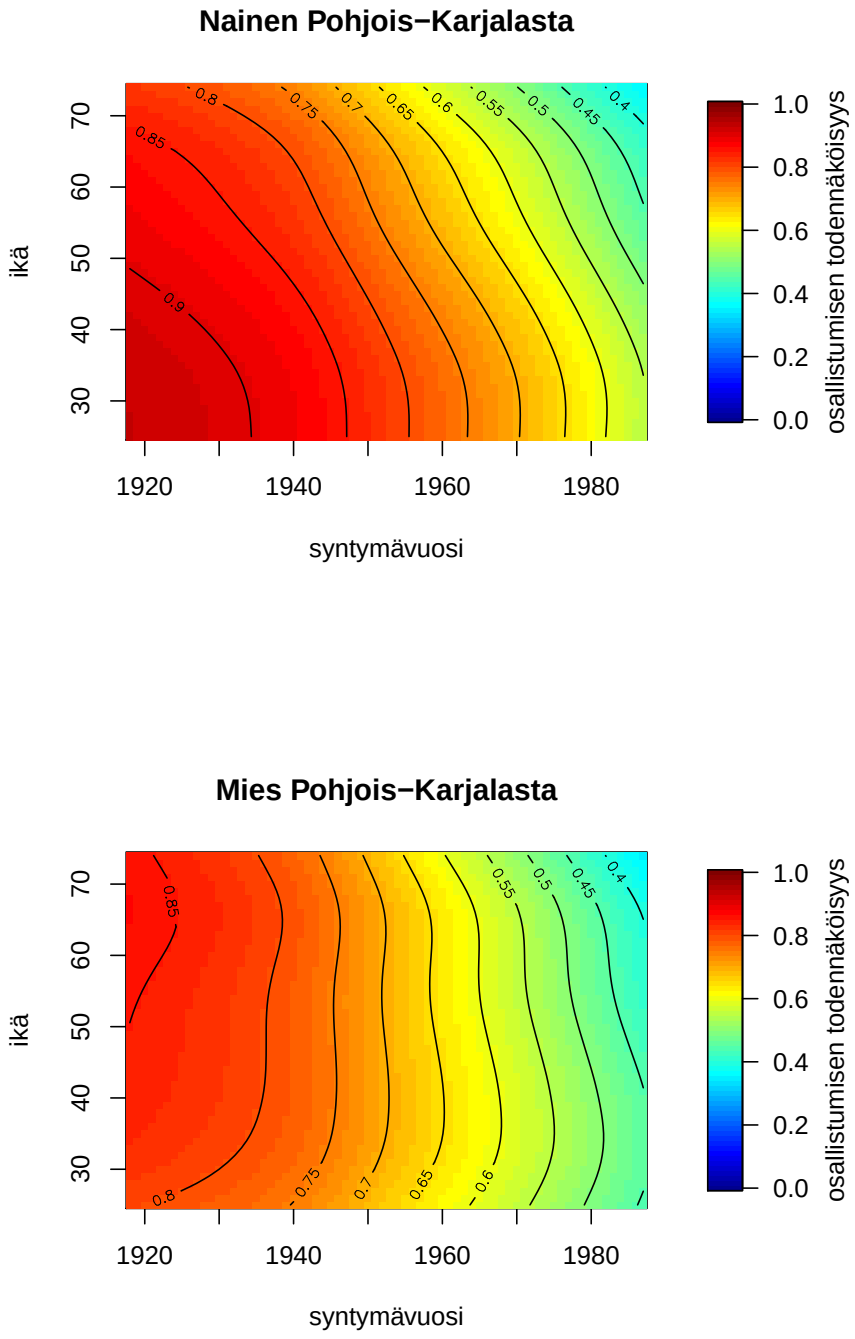
Kuva 13: Toisen mallin ja alueen neljä todennäköisyyskäyrät syntymävuosittain iän funktiona naisille ja miehille



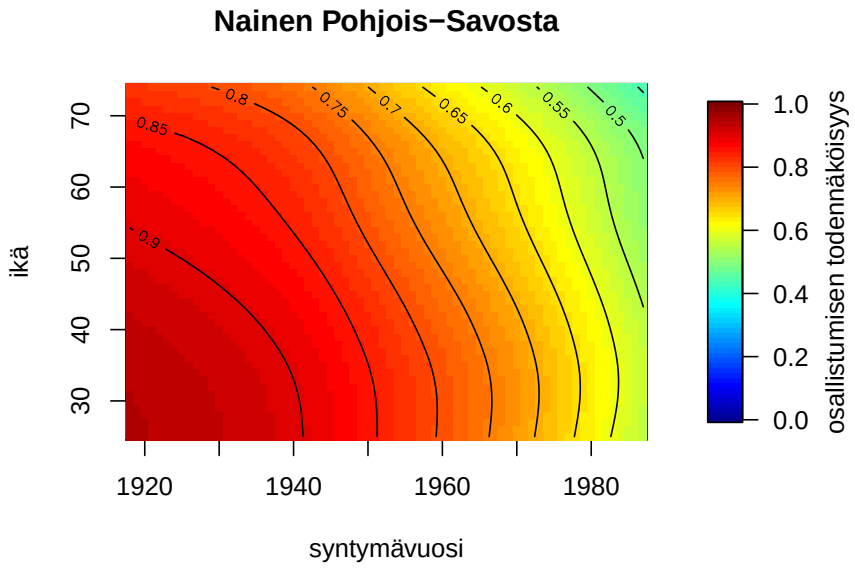
Kuva 14: Toisen mallin ja alueen viisi todennäköisyyskäyrät syntymävuosittain iän funktiona naisille ja miehille

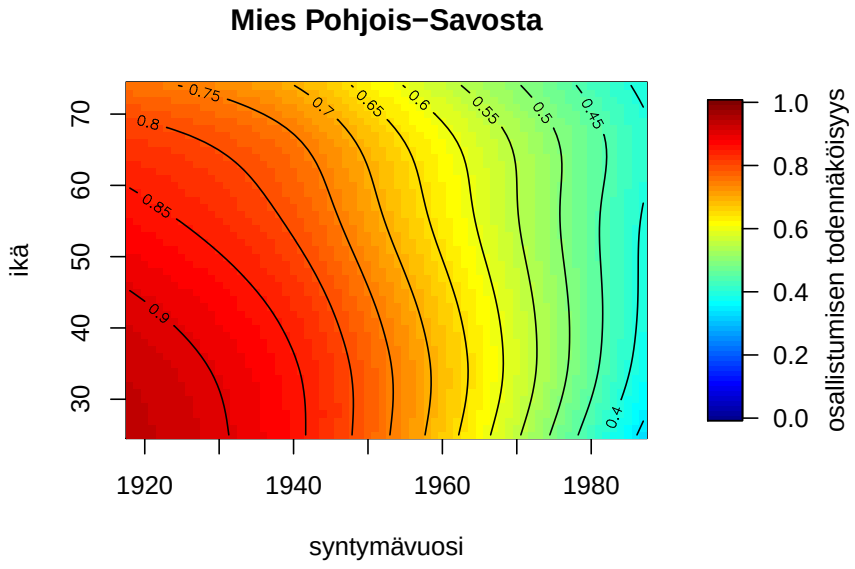


Kuva 15: Toisen mallin ja alueen kuusi todennäköisyyskäyrät syntymävuosittain iän funktiona naisille ja miehille

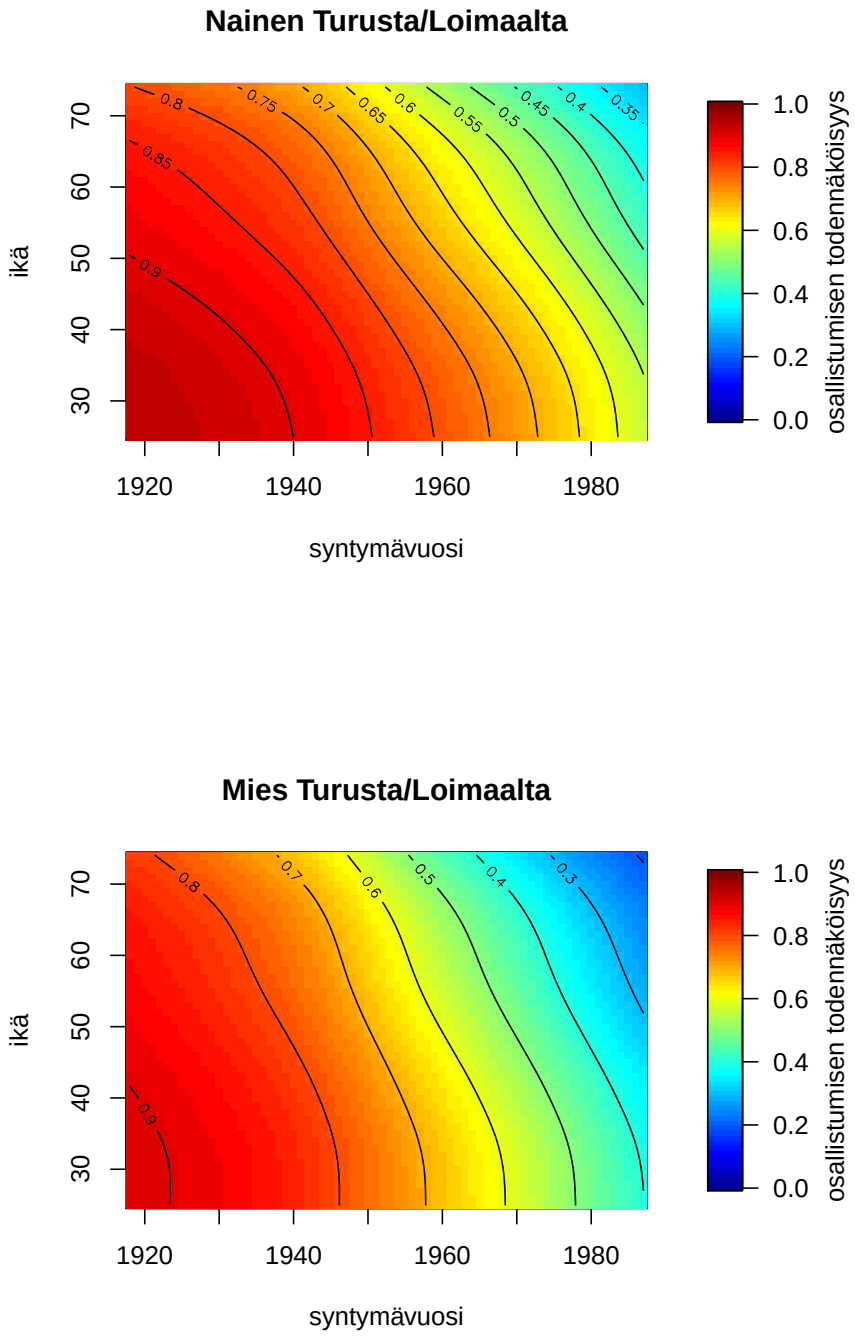


Kuva 16: Toisen mallin ja alueen kaksi vakiotodennäköisyyskäyrät syntymävuoden ja iän suhteen naisille ja miehille

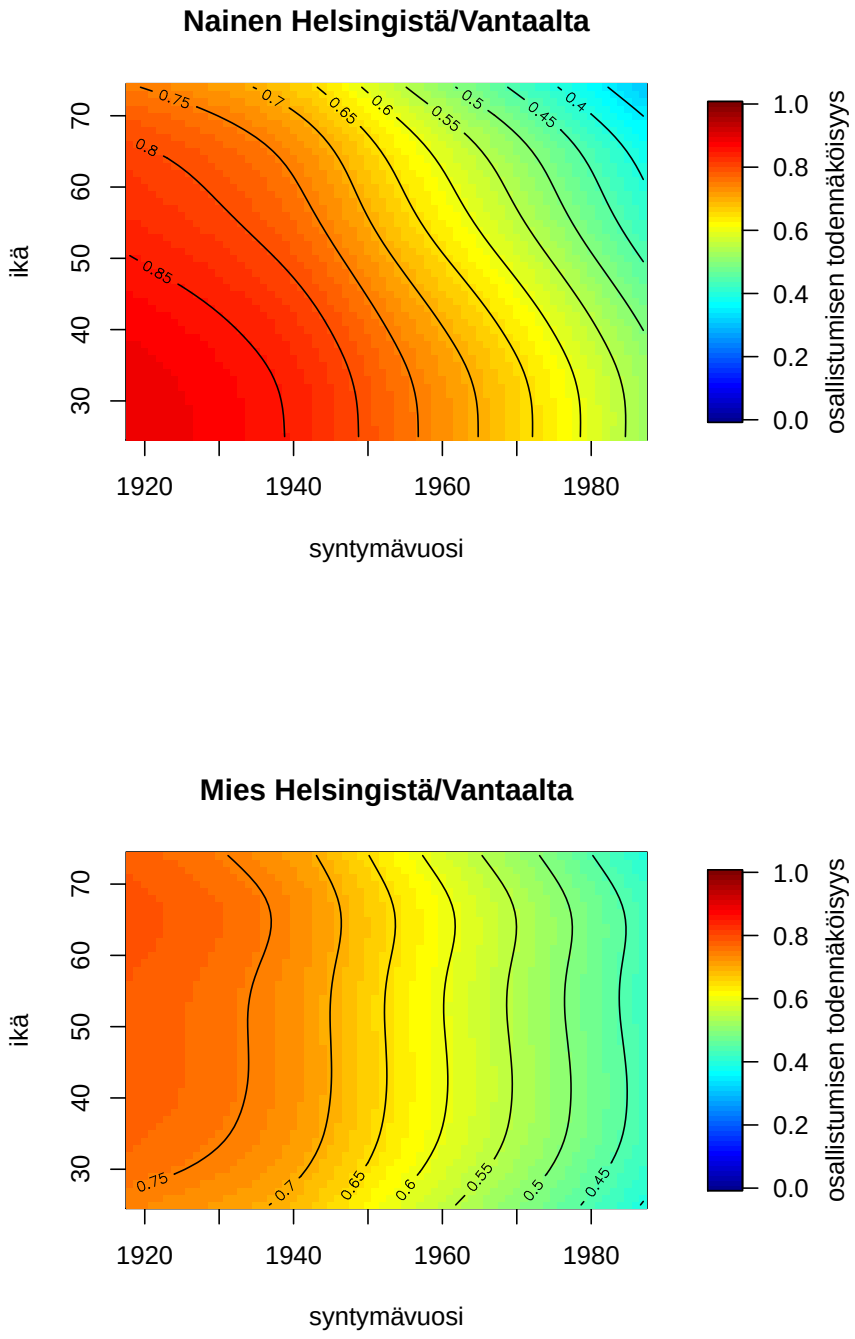




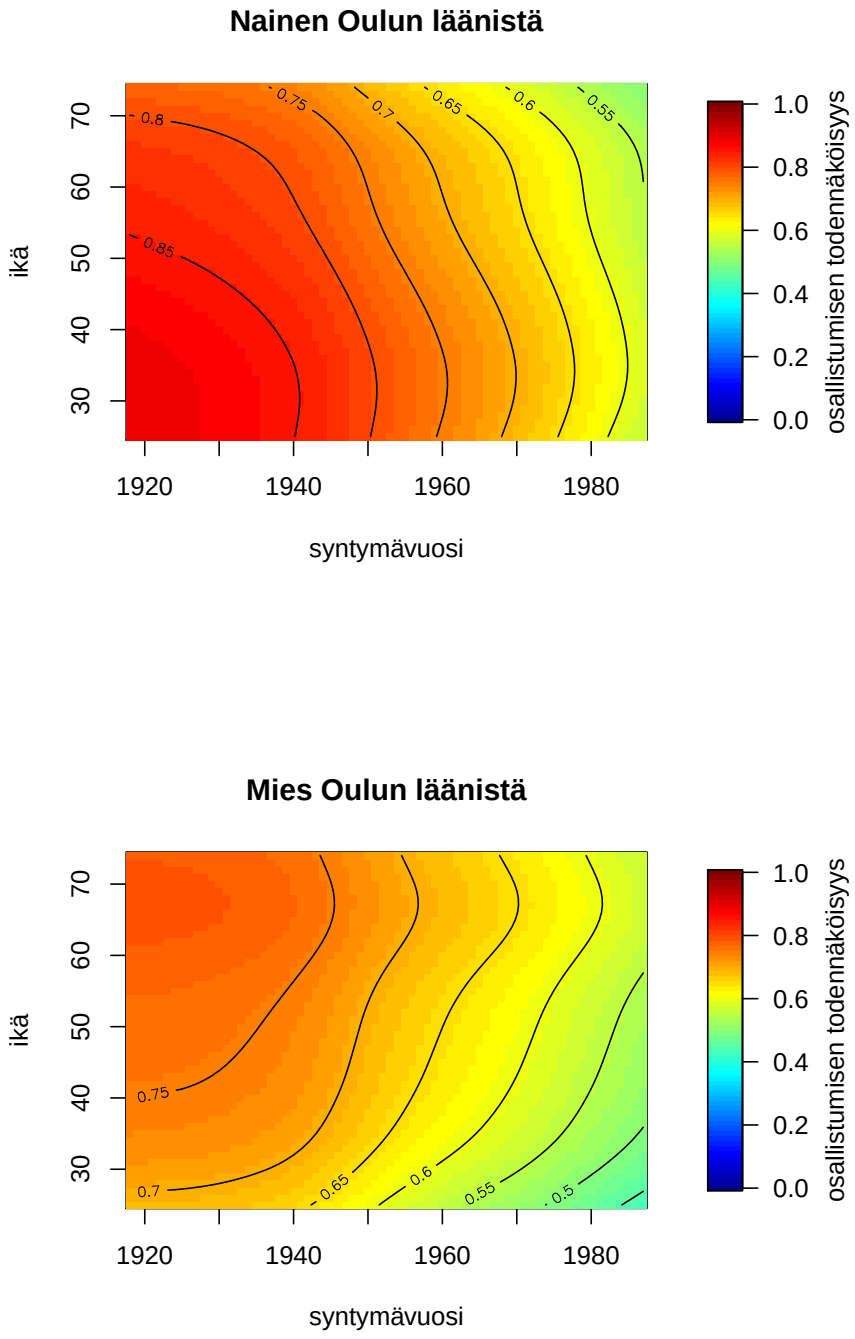
Kuva 17: Toisen mallin ja alueen kolme vakiotodennäköisyyskäyrät syntymävuoden ja iän suhteen naisille ja miehille



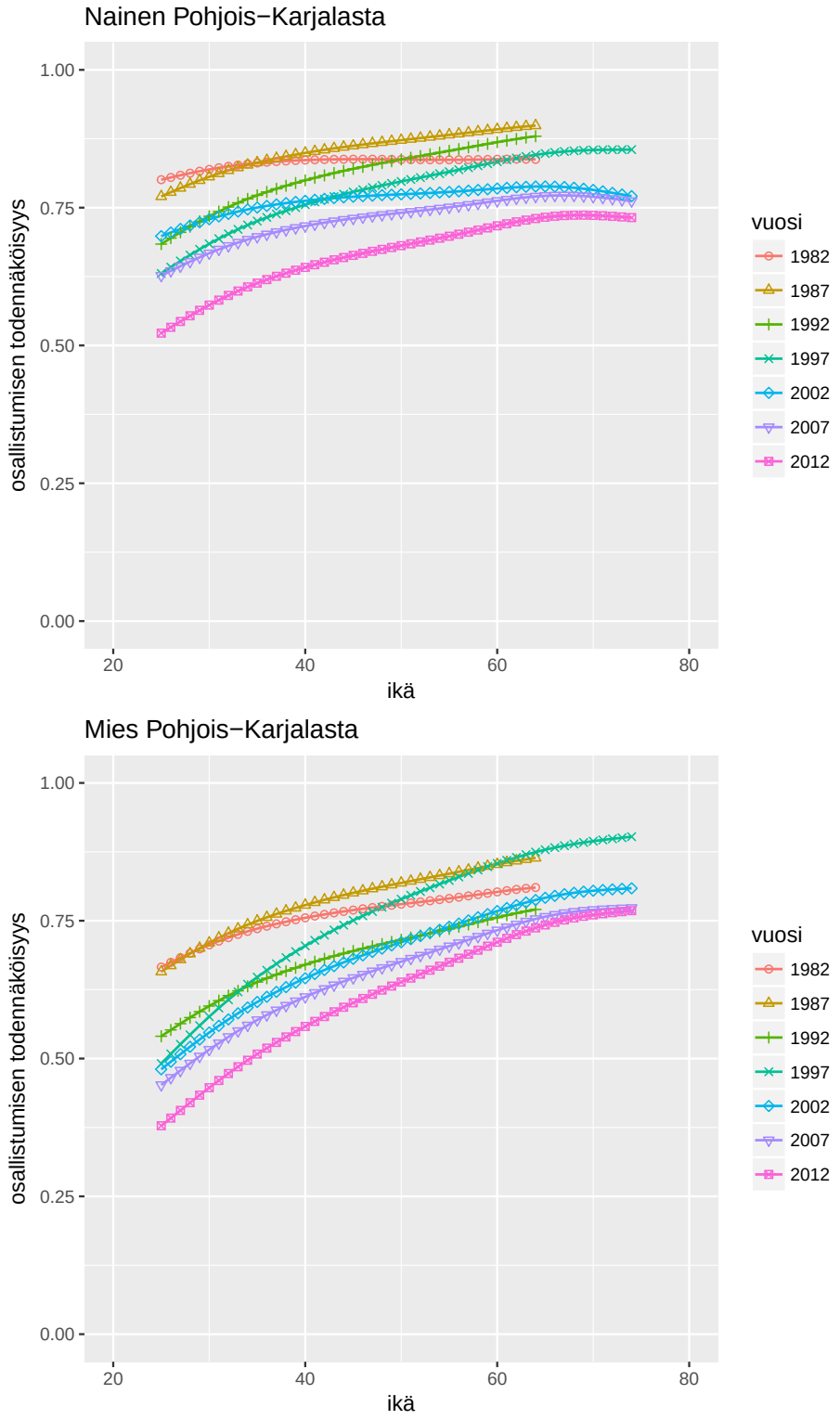
Kuva 18: Toisen mallin ja alueen neljä vakiotodennäköisyyskäyrät syntymävuoden ja iän suhteen naisille ja miehille



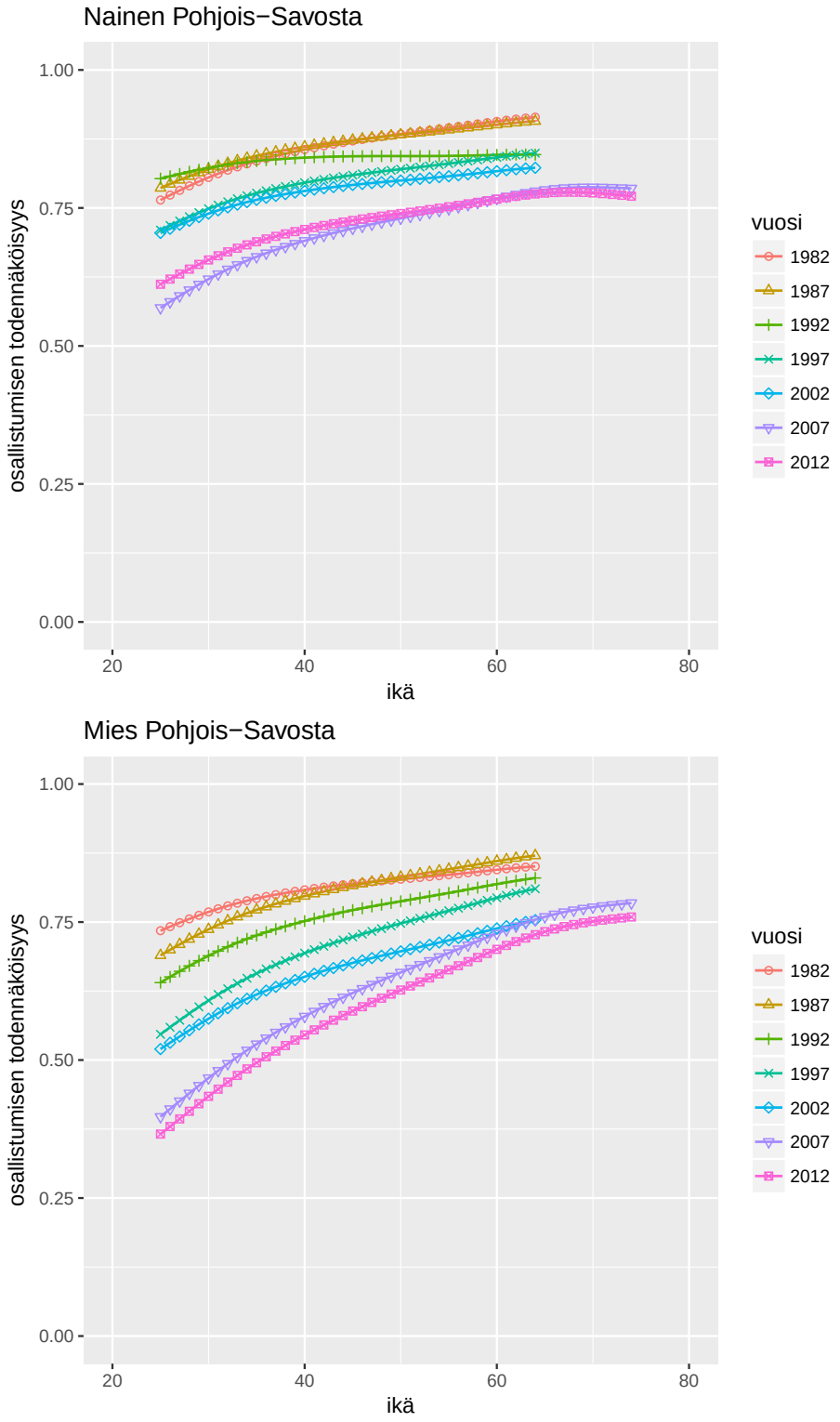
Kuva 19: Toisen mallin ja alueen viisi vakiotodennäköisyyskäyrät syntymävuoden ja iän suhteen naisille ja miehille



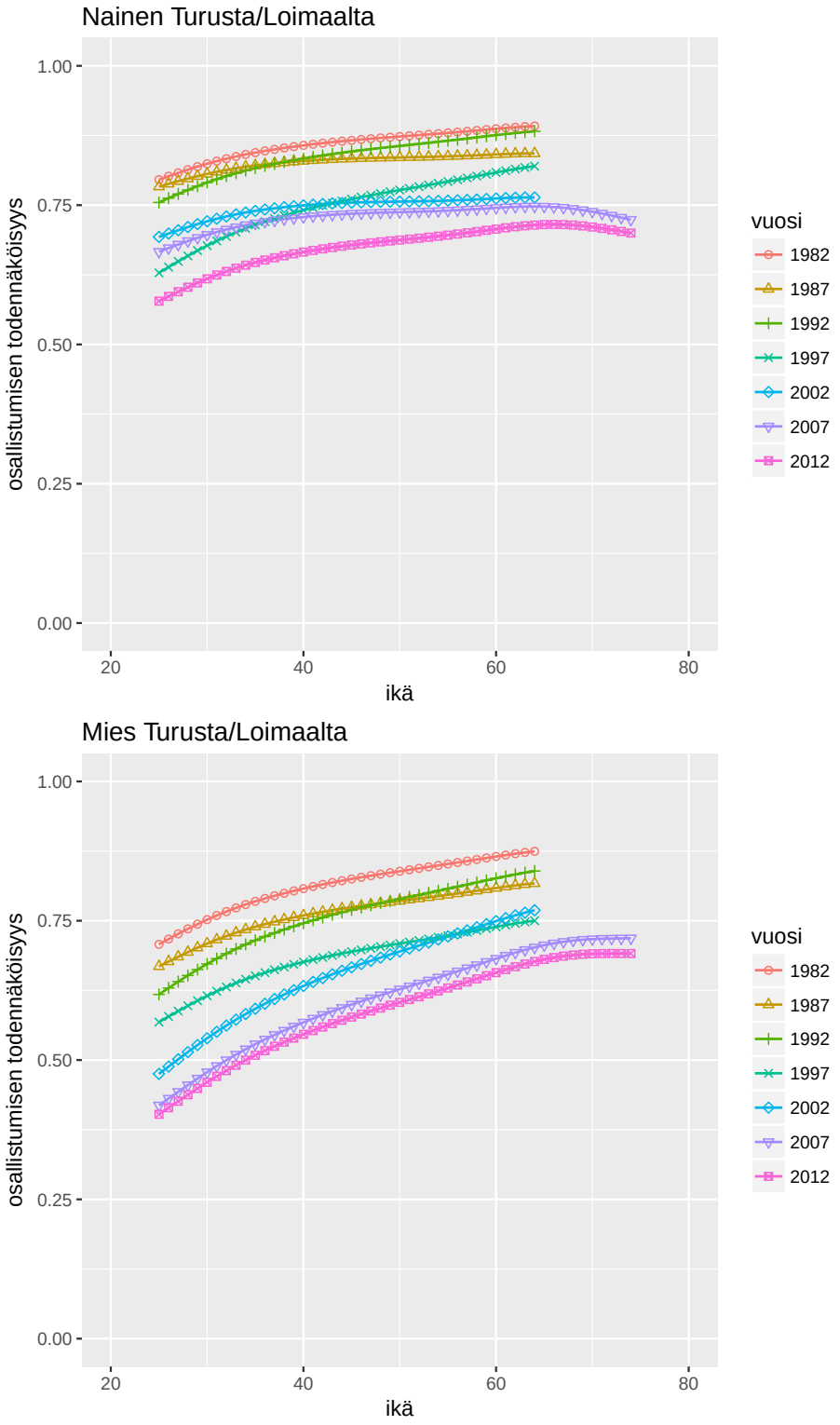
Kuva 20: Toisen mallin ja alueen kuusi vakiotodennäköisyyskäyrät syntymävuoden ja iän suhteen naisille ja miehille



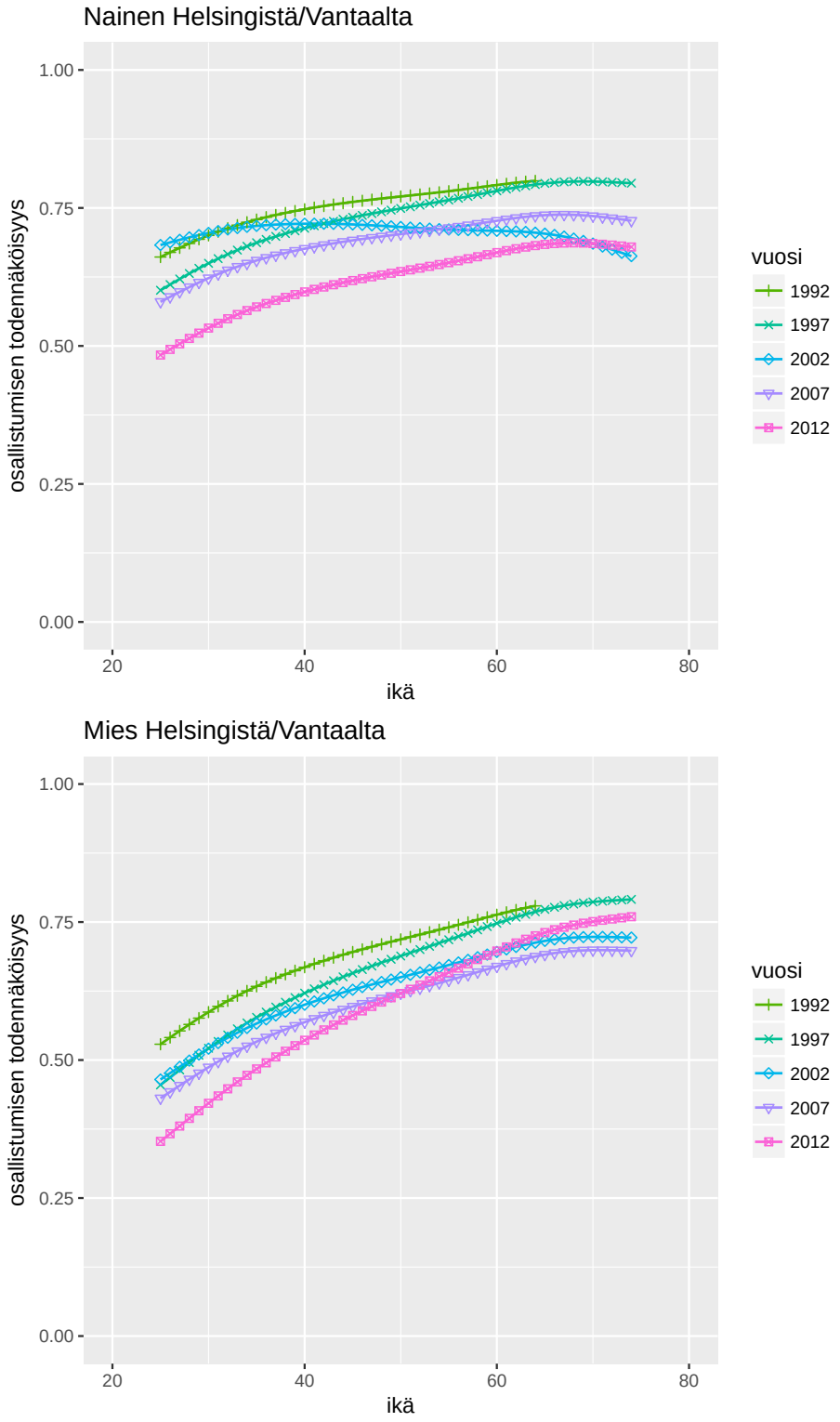
Kuva 21: Kolmannen mallin ja alueen kaksi osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



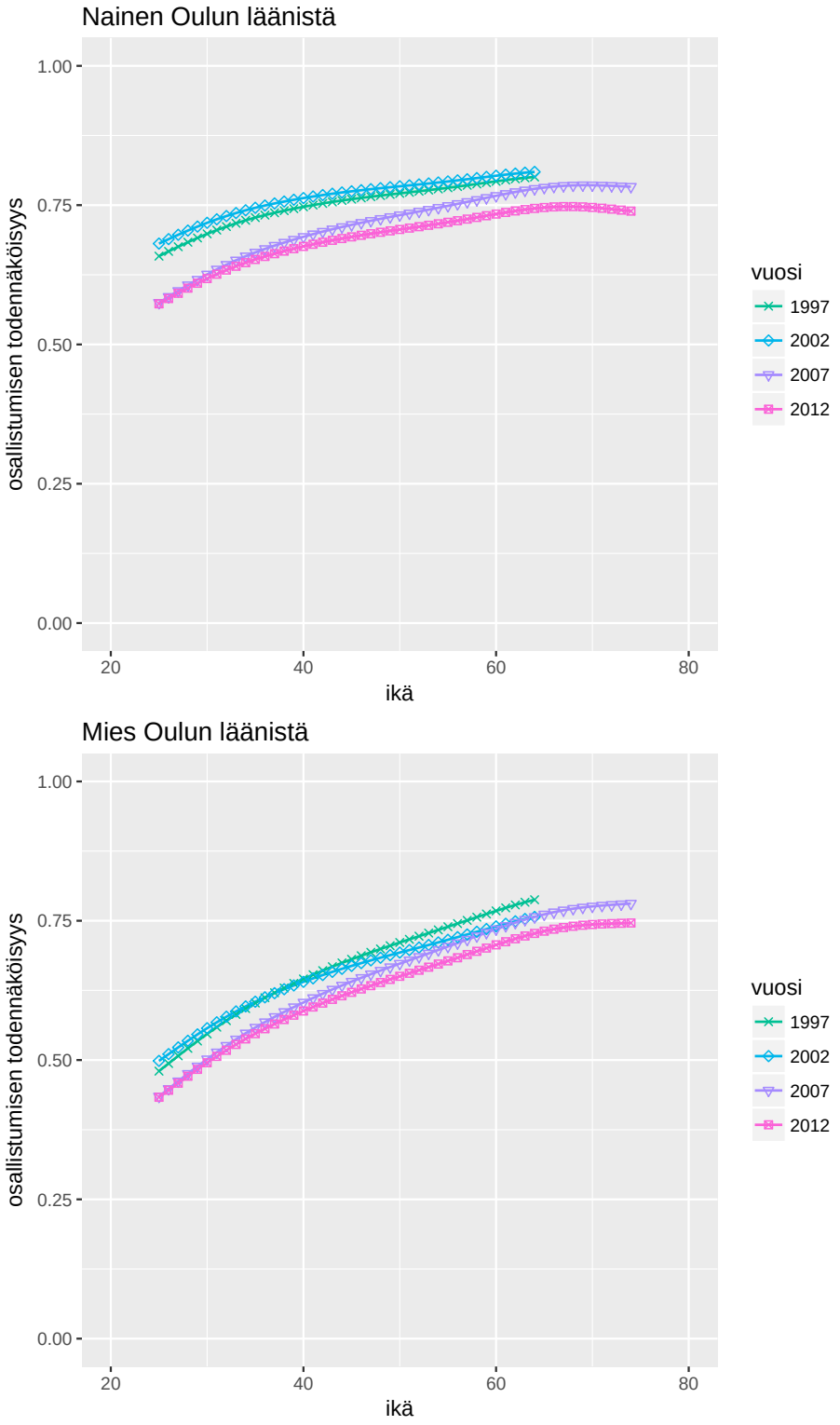
Kuva 22: Kolmannen mallin ja alueen kolme osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



Kuva 23: Kolmannen mallin ja alueen neljä osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



Kuva 24: Kolmannen mallin ja alueen viisi osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille



Kuva 25: Kolmannen mallin ja alueen kuusi osallistumistodennäköisyydet tutkimusvuosittain naisille ja miehille

6 Johtopäätökset

Tässä työssä tutkittiin Kansallisen FINRISKI -tutkimuksen osallistumisaktiivisuutta ja pyrittiin mallintamaan tutkimuksen puuttuvuutta. Tutkimuksen osallistumisprosentti on laskenut vuosi vuodelta, joten puuttuvuuden mallintaminen oli tärkeää, jotta tuloksia voidaan korjata ja tutkimusta suunnitella paremmin jatkossa siten, että saataisiin mahdollisimman paljon osallistujia.

Työn aihetta lähestyttiin ikä-periodi-kohortti -analyysin kautta, jossa pyritään identifioimaan aikamuuttujien ikä, tutkimusvuosi ja syntymävuosi vaikutukset. Tässä periodi kuvaa tutkimusvuotta ja kohortti syntymävuotta. Ikä-periodi-kohortti -analyysi on luotu tilanteeseen, jossa kaikki kolme aikamuuttujaa voitaisiin lisätä malliin yhtä aikaa. Tämä on kuitenkin hankalaa aikamuuttujien matemaattisen riippuvuuden takia ja ratkaisuihin on kohdistettu kritiikkiä. Täten päädyttiin toteuttamaan ikä-periodi-kohortti -analyysi siten, että sovitetaan kolme mallia, joissa jokaisessa jätetään aina yksi aikamuuttuja kerrallaan pois. Tämä perustuu siihen, että kun kaksi aikamuuttujaa tiedetään, kolmas voidaan laskea näiden kahden tiedossa olevan aikamuuttujan avulla.

Mallit sovitettiin käyttäen Hastien ja Tibshiranin (1986; 1990; 2006) kehittämää yleistettyä additiivista mallia ja tarkemmin additiivista logistista regressiota. Yleistettyjen additiivisten mallien vahvuutena on, että ne eivät tee vahvoja lineaarisia oletuksia. Vastaavaa aihetta kuin tässä tutkielmassa ei ole tutkittu aikaisemmin, joten valittiin lähestymistapa, jossa epälineaaristen selittäjien sovittaminen on joustavaa. Koska kiinnostus kohdistui kaikkiin mahdollisiin selittäjien vaikutuksiin, sovitettiin sellaiset mallit, joissa oli mukana kaikki selittäjien väliset interaktiot. Mallit esitettiin graafisesti jokaiselle tutkimusalue-sukupuoliositteelle. Kuvaajat tulkittiin ja niitä vertailtiin keskenään.

Vaikka yleistetyt additiiviset mallit vaikuttavat hyvin ylivoimaisilta yleistettyihin lineaarisiin malleihin nähden, on niitäkin käytettävä harkitusti. Yleistetty additiivinen malli on kuitenkin vain yleistettyjen lineaaristen mallien laajennus. Koska yleistetyt additiiviset mallit ovat niin joustavia ja ne sovittuvat epälineaarisiin vaikutuksiin hyvin, on oltava varovainen, kun selittäjien määrä mallissa kasvaa. Tällöin on olemassa ylisovittumisen vaara. Tässä työssä selittäjiä oli vain neljä jokaista mallia kohden, joten ylisovittumisen vaaraa ei pitäisi olla. (Hastie & Tibshirani, 1990, s. 6 ja luku 10).

Tulokset olivat hyvin mielenkiintoisia. Aina seuraava tutkimusvuosi oli edellistään heikompi osallistumistodennäköisyyksien suhteen ja nuoremmat sekä vuoden 1950 jälkeen syntyneet tutkittavat osallistuivat huonommin kuin vanhemmat ja ennen vuotta 1950 syntyneet tutkittavat. Tutkimusvuosi 1997 nousi erityisesti esiin tuloksissa. Nuorimpien ja vuoden 1950 jälkeen syntyneiden tutkittavien osallistumistodennäköisyydet olivat vuonna 1997 huomattavasti matalammat kuin vanhimpien ja ennen vuotta 1950 syntyneiden tutkittavien todennäköisyydet verrattuna muihin tutkimusvuosiin.

Miesten osallistumistodennäköisyydet olivat paljon matalampia kuin naisten todennäköisyydet. Miesten osallistumisen todennäköisyydet myös laskivat paljon voimakkaammin kuin naisilla, kun ikä ja syntymävuosi kasvoi. Vanhimpien sekä ennen vuotta 1950 syntyneiden miesten ja naisten osallistumistodennäköisyydet olivat lähes samalla tasolla, mutta nuorten ja vuoden 1950 jälkeen syntyneiden miesten osallistumistodennäköisyydet olivat huomattavasti matalammat kuin nuorten ja 1950 jälkeen syntyneiden naisten todennäköisyydet.

Joitakin poikkeuksia löytyi eri tutkimusalueilla. Oulun läänin miehillä tulokset olivat päinvastaiset kuin muiden alueiden miesten tulokset. Heillä tutkimukseen osallistuminen oli todennäköisempää nuorimmilla ja vuoden 1950 jälkeen syntyneillä kuin vanhimmilla ja ennen 1950 syntyneillä tutkittavilla miehillä. Pohjois-Karjalassa ja -Savossa olivat korkeimmat osallistumistodennäköisyydet ja matalimmat todennäköisyydet taas olivat Helsingin/Vantaan alueella. Helsingin/Vantaan miehillä iällä ei ollut niin suurta vaikutusta osallistumistodennäköisyyksien laskuun kuin muilla alueilla. Myös Pohjois-Karjalan miehillä tämä oli havaittavissa, vaikkakaan ei niin voimakkaana kuin Helsingin/Vantaan miehillä. Turussa/Loimaalla oli havaittavissa isoimmat erot osallistumistodennäköisyyksissä tutkimusvuosittain verrattuna muihin tutkimusalueisiin. Heillä seuraavan tutkimisvuoden osallistumistodennäköisyydet putosivat edelliseen tutkimusvuoteen verrattuna enemmän kuin muilla tutkimusalueilla.

Kuvaajat eivät ole ristiriidassa toisiinsa nähden ja ne tukevat toisiaan hyvin. Tuloksista saatiin kokonaisvaltainen kuva puuttuvuudesta Kansallisessa FINRISKI -tutkimuksessa. Voidaan tietenkin miettiä, olisiko kuitenkin parempi saada kaikki kolme aikamuuttujaa samaan malliin, jolloin tarvitsisi tulkita vain yksi malli ja sen kuvaajat. Jos halutaan kaikki mahdolliset muuttujien väliset interaktiot mukaan, kuten tämän työn tilanteessa haluttiin, ja mukana olisivat kaikki kolme aikamuuttujaa, voisi mallista tulla turhan monimutkainen. Tällöin mallissa olisi todella paljon regressiokertoimia.

Hankeympäristö

Työ liittyy Suomen Akatemian rahoittamaan Jyväskylän yliopiston ja Terveyden ja hyvinvoinnin laitoksen yhteishankkeeseen ”Non-participation in health examination surveys”/ ”Kato terveystutkimuksissa”[rahoituspäätös 266251]. Aineiston käyttöön on saatu lupa Terveyden ja hyvinvoinnin laitoksen FINRISKI-tutkimukselta.

Viitteet

Bell, J. & Jones, K. Another 'futile quest'? a simulation study of Yang and Land's hierarchical age-period-cohort model. *Demographic Research*, 30: 333–360, 2014.

Borodulin, K., Saarikoski, L., Lund, L., Juolevi, A., Grönholm, M., Helldán, A., Peltonen, M., Laatikainen, T., & Vartiainen, E. *Kansallinen FINRISKI 2012 -terveystutkimus - Osa I: Tutkimuksen toteutus ja menetelmät*. Juvenes Print – Suomen Yliopistopaino Oy, 2013.

Hastie, T. Package 'gam'. `https://cran.r-project.org/web/packages/gam/gam.pdf`, 2016. Accessed: 27.10.2016.

Hastie, T. & Tibshirani, R. Generalized additive models. *Statistical Science*, 1(3): 297–310, 1986.

Hastie, T. & Tibshirani, R. Generalized additive models. In *Encyclopedia of Statistical Sciences*. Wiley Online Library, 2006.

Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2. edition, 2009.

Hastie, T.J. & Tibshirani, R.J. *Generalized Additive Models*. Chapman & Hall, 1990.

Kopra, J., Härkänen, T., Tolonen, H., & Karvanen, J. Correcting for non-ignorable missingness in smoking trends. *Stat*, 4(1): 1–14, 2015.

McCullagh, P. & Nelder, J. *Generalized Linear Models*. Chapman and Hall, 2 edition, 1989.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

Ryder, N.B. The cohort as a concept in the study of social-change. *American Sociological Review*, 30(6): 843–861, 1965.

Terveyden ja hyvinvoinnin laitos. Kansallinen finriski-tutkimus. `https://www.thl.fi/fi/aiheet/tietopaketit/thl-biopankki/thl-biopankin-naytekokoelmat/`, 2015. Viitattu: 27.10.2016.

Terveyden ja hyvinvoinnin laitos. Tieteellisen tutkimuksen rekisteriseloste. `https://www.thl.fi/documents/10531/862648/FINRISKIn+rekisteriselosteet+1992-2012.pdf/f5567a5b-6134-4013-9691-0c2280296c2d`, 2016. Viitattu: 19.01.2017.

Terveyden ja hyvinvoinnin laitos. Finriski-tutkimus. https://www.thl.fi/fi/tutkimus-ja-asiantuntijatyo/hankkeet-ja-ohjelmat/hankkeet/25761, 2017. Viitattu: 26.04.2017.

Yang, Y. & Land, K.C. *Age-Period-Cohort Analysis: New Models, Methods, and Empirical Applications*. Chapman and Hall, 2013.

Liite A: Datan alku

Taulukko 2: Muokatun datan kuusi ensimmäistä riviä

	VUOSI	ALUE	SUKUP	IKA	S.VUOSI	E.OSAL	K.OSAL
1	1982	2	1	25	1957	12	18
2	1982	2	1	26	1956	12	24
3	1982	2	1	27	1955	21	33
4	1982	2	1	28	1954	10	34
5	1982	2	1	29	1953	12	32
6	1982	2	1	30	1952	19	31

Liite B: R-koodi

```
## Kuvaajien muodostaminen uudelleen muokatusta datasta ##

data <- read.table("G:/GRADU/Finrisk_participation_aggregated.csv", header=TRUE, sep=",")
attach(data)

## Lisätään syntymävuosi dataan ###

data$S.VUOSI <- VUOSI-IKA

### Muodostetaan data, josta on poistettu vuodet 1972 ja 1977 seka alue 7 ####
datFull <- subset(data, VUOSI != 1972 & VUOSI != 1977 & ALUE != 7)
attach(datFull)

### Muokataan E.OSAL ja K.OSAL dataan ###
library(reshape2)
datRe <- dcast(datFull, VUOSI+ALUE+SUKUP+IKA+S.VUOSI ~ osal, value.var="N")
names(datRe) <- c("VUOSI","ALUE","SUKUP","IKA","S.VUOSI","E.OSAL","K.OSAL")
datRe[is.na(datRe)] <- 0
attach(datRe)

library(gam)
library(ggplot2)

####################################
# Malli 1: IKA-muuttuja jätetty pois #
####################################

m7 <- gam(cbind(K.OSAL,E.OSAL)~factor(ALUE)*SUKUP*s(S.VUOSI)*factor(VUOSI),family=binomial, data=datRe)

# lisätään mallin m7 ennusteet yhatm7 dataan:
datRe$yhatm7<-predict(m7,type="response")

# Muodostetaan kuvaajat:

# Alue 2 (Pohjois-Karjala) nainen
nainen2 <- subset(datRe, SUKUP==2 & ALUE==2)
ggplot(nainen2, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI), colour=factor(VUOSI), shape=factor(VUOSI)))+
  scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1), xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point() +
  labs(title="Nainen Pohjois-Karjalasta", x="syntymävuosi",y="osallistumisen todennäköisyys",colour="vuosi",shape="vuosi")

# Alue 2 mies
mies2 <- subset(datRe, SUKUP==1 & ALUE==2)
ggplot(mies2, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI),colour=factor(VUOSI), shape=factor(VUOSI)))+
  scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1), xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Mies Pohjois-Karjalasta", x="syntymävuosi",y="osallistumisen todennäköisyys",colour="vuosi",shape="vuosi")

# Alue 3 (Pohjois-Savo) nainen
nainen3 <- subset(datRe, SUKUP==2 & ALUE==3)
ggplot(nainen3, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI), colour=factor(VUOSI), shape=factor(VUOSI)))+
  scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1), xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point() +
  labs(title="Nainen Pohjois-Savosta", x="syntymävuosi",y="osallistumisen todennäköisyys",colour="vuosi",shape="vuosi")
```

```
55
56  # Alue 3 mies
57  mies3 <- subset(datRe, SUKUP==1 & ALUE==3)
58  ggplot(mies3, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI),colour=factor(VUOSI),
        shape=factor(VUOSI)))+
59    scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(1913,1990))+
60    geom_line(size=0.6) + geom_point()+
61    labs(title="Mies Pohjois-Savosta", x="syntymävuosi",y="osallistumisen todennäkö
        isyys",colour="vuosi",shape="vuosi")
62
63  # Alue 4 (Turku/Loimaa) nainen
64  nainen4 <- subset(datRe, SUKUP==2 & ALUE==4)
65  ggplot(nainen4, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
66    scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(1913,1990))+
67    geom_line(size=0.6) + geom_point() +
68    labs(title="Nainen Turusta/Loimaalta", x="syntymävuosi",y="osallistumisen todennäk
        öisyys",colour="vuosi",shape="vuosi")
69
70  # Alue 4 mies
71  mies4 <- subset(datRe, SUKUP==1 & ALUE==4)
72  ggplot(mies4, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI),colour=factor(VUOSI),
        shape=factor(VUOSI)))+
73    scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(1913,1990))+
74    geom_line(size=0.6) + geom_point()+
75    labs(title="Mies Turusta/Loimaalta", x="syntymävuosi",y="osallistumisen todennäkö
        isyys",colour="vuosi",shape="vuosi")
76
77  # Alue 5 (Helsinki/Vantaa) nainen
78  nainen5 <- subset(datRe, SUKUP==2 & ALUE==5)
79  ggplot(nainen5, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
80    scale_shape_manual(values=3:nlevels(factor(VUOSI)))+
81    scale_colour_manual(values=c("#53B400","#00C094","#00B6EB","#A58AFF","#FB61D7"))+
82    coord_cartesian(ylim=c(0,1),xlim=c(1913,1990))+geom_line(size=0.6) + geom_point()
        +
83    labs(title="Nainen Helsingistä/Vantaalta", x="syntymävuosi",y="osallistumisen
        todennäköisyys",colour="vuosi",shape="vuosi")
84
85  # Alue 5 mies
86  mies5 <- subset(datRe, SUKUP==1 & ALUE==5)
87  ggplot(mies5, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI),colour=factor(VUOSI),
        shape=factor(VUOSI)))+
88    scale_shape_manual(values=3:nlevels(factor(VUOSI)))+
89    scale_colour_manual(values=c("#53B400","#00C094","#00B6EB","#A58AFF","#FB61D7"))+
        coord_cartesian(ylim=c(0,1),xlim=c(1913,1990))+
90    geom_line(size=0.6) + geom_point()+
91    labs(title="Mies Helsingistä/Vantaalta", x="syntymävuosi",y="osallistumisen todenn
        äköisyys",colour="vuosi",shape="vuosi")
92
93  # Alue 6 (Oulun lääni) nainen
94  nainen6 <- subset(datRe, SUKUP==2 & ALUE==6)
95  ggplot(nainen6, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
96    scale_shape_manual(values=4:nlevels(factor(VUOSI)))+
97    scale_colour_manual(values=c("#00C094","#00B6EB","#A58AFF","#FB61D7"))+coord_
        cartesian(ylim=c(0,1),xlim=c(1913,1990))+
98    geom_line(size=0.6) + geom_point() +
99    labs(title="Nainen Oulun läänistä", x="syntymävuosi",y="osallistumisen todennäkö
        isyys",colour="vuosi",shape="vuosi")
100
101 # Alue 6 mies
102 mies6 <- subset(datRe, SUKUP==1 & ALUE==6)
103 ggplot(mies6, aes(x=S.VUOSI, y=yhatm7, group=factor(VUOSI),colour=factor(VUOSI),
        shape=factor(VUOSI)))+
```

```
104   scale_shape_manual(values=4:nlevels(factor(VUOSI)))+
105   scale_colour_manual(values=c("#00C094","#00B6EB","#A58AFF","#FB61D7"))+coord_
          cartesian(ylim=c(0,1),xlim=c(1913,1990))+
106   geom_line(size=0.6) + geom_point()+
107   labs(title="Mies Oulun läänistä", x="syntymävuosi",y="osallistumisen todennäkö
          isyys",colour="vuosi",shape="vuosi")
108
109 ######################################
110 # Malli 2: VUOSI-muuttuja jätetty pois #
111 ######################################
112
113 mv7 <- gam(cbind(K.OSAL,E.OSAL)~factor(ALUE)*SUKUP*s(S.VUOSI)*s(IKA),family=binomial
        , data=datRe)
114
115 # lisätään mallin mv7 ennusteet yhatmv7 dataan
116 datRe$yhatmv7<-predict(mv7,type="response")
117
118
119 ### Osallistumisen todennäköisyys ikäryhmittäin syntymävuoden funktiona ###
120
121 # Ehto ikäjoukon muodostamiseen, jossa iät valittu viiden vuoden välein
122 ikaehto<-IKA==25| IKA==30| IKA==35| IKA==40| IKA==45| IKA==50| IKA==55| IKA==60| IKA
        ==65| IKA==70| IKA==74
123
124 # Muodostetaan kuvaajat:
125
126 # Alue 2 (Pohjois-Karjala) nainen
127 nainen2ika <- subset(datRe, SUKUP==2 & ALUE==2 & (ikaehto))
128 ggplot(nainen2ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
        shape=factor(IKA)))+
129   scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
          xlim=c(1913,1990))+
130   geom_line(size=0.6) + geom_point()+
131   labs(title="Nainen Pohjois-Karjalasta", x="syntymävuosi",y="osallistumisen todennä
          köisyys",colour="ikä",shape="ikä")
132
133 # Alue 2 mies
134 mies2ika <- subset(datRe, SUKUP==1 & ALUE==2 & (ikaehto))
135 ggplot(mies2ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
        shape=factor(IKA)))+
136   scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
          xlim=c(1913,1990))+
137   geom_line(size=0.6) + geom_point()+
138   labs(title="Mies Pohjois-Karjalasta", x="syntymävuosi",y="osallistumisen todennäkö
          isyys",colour="ikä",shape="ikä")
139
140 # Alue 3 (Pohjois-Savo) nainen
141 nainen3ika <- subset(datRe, SUKUP==2 & ALUE==3 & (ikaehto))
142 ggplot(nainen3ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
        shape=factor(IKA)))+
143   scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
          xlim=c(1913,1990))+
144   geom_line(size=0.6) + geom_point()+
145   labs(title="Nainen Pohjois-Savosta", x="syntymävuosi",y="osallistumisen todennäkö
          isyys",colour="ikä",shape="ikä")
146
147 # Alue 3 mies
148 mies3ika <- subset(datRe, SUKUP==1 & ALUE==3 & (ikaehto))
149 ggplot(mies3ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
        shape=factor(IKA)))+
150   scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
          xlim=c(1913,1990))+
151   geom_line(size=0.6) + geom_point()+
152   labs(title="Mies Pohjois-Savosta", x="syntymävuosi",y="osallistumisen todennäkö
          isyys",colour="ikä",shape="ikä")
153
154 # Alue 4 (Turku/Loimaa) nainen
155 nainen4ika <- subset(datRe, SUKUP==2 & ALUE==4 & (ikaehto))
```

```
ggplot(nainen4ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
    shape=factor(IKA)))+
  scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
      xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Nainen Turusta/Loimaasta", x="syntymävuosi",y="osallistumisen todennäk
      öisyys",colour="ikä",shape="ikä")

# Alue 4 mies
mies4ika <- subset(datRe, SUKUP==1 & ALUE==4 & (ikaehto))
ggplot(mies4ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
    shape=factor(IKA)))+
  scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
      xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Mies Turusta/Loimaasta", x="syntymävuosi",y="osallistumisen todennäkö
      isyys",colour="ikä",shape="ikä")

# Alue 5 (Helsinki/Vantaa) nainen
nainen5ika <- subset(datRe, SUKUP==2 & ALUE==5 & (ikaehto))
ggplot(nainen5ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
    shape=factor(IKA)))+
  scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
      xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Nainen Helsingistä/Vantaalta", x="syntymävuosi",y="osallistumisen
      todennäköisyys",colour="ikä",shape="ikä")

# Alue 5 mies
mies5ika <- subset(datRe, SUKUP==1 & ALUE==5 & (ikaehto))
ggplot(mies5ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
    shape=factor(IKA)))+
  scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
      xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Mies Helsingistä/Vantaalta", x="syntymävuosi",y="osallistumisen todenn
      äköisyys",colour="ikä",shape="ikä")

# Alue 6 (Oulun lääni) nainen
nainen6ika <- subset(datRe, SUKUP==2 & ALUE==6 & (ikaehto))
ggplot(nainen6ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
    shape=factor(IKA)))+
  scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
      xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Nainen Oulun läänistä", x="syntymävuosi",y="osallistumisen todennäkö
      isyys",colour="ikä",shape="ikä")

# Alue 6 mies
mies6ika <- subset(datRe, SUKUP==1 & ALUE==6 & (ikaehto))
ggplot(mies6ika, aes(x=S.VUOSI, y=yhatmv7, group=factor(IKA), colour=factor(IKA),
    shape=factor(IKA)))+
  scale_shape_manual(values=1:nlevels(factor(IKA))) +coord_cartesian(ylim=c(0,1),
      xlim=c(1913,1990))+
  geom_line(size=0.6) + geom_point()+
  labs(title="Mies Oulun läänistä", x="syntymävuosi",y="osallistumisen todennäkö
      isyys",colour="ikä",shape="ikä")

##################################

### Osallistumisen todennäköisyys syntymävuosittain iän funktiona ###

# Ehto syntymävuosijoukon muodostamiseen, jossa syntymävuodet valittu viiden vuoden
    välein:
synt.ehto<-S.VUOSI==1918| S.VUOSI==1923| S.VUOSI==1928| S.VUOSI==1933| S.VUOSI
    ==1938| S.VUOSI==1943| S.VUOSI==1948|
  S.VUOSI==1953| S.VUOSI==1958| S.VUOSI==1963| S.VUOSI==1968| S.VUOSI==1973| S.VUOSI
      ==1978| S.VUOSI==1983| S.VUOSI==1987
```

```
203
204 # Muodostetaan kuvaajat:
205
206 # Alue 2 (Pohjois-Karjala) nainen
207 nainen2synt <- subset(datRe, SUKUP==2 & ALUE==2 & (synt.ehto))
208 ggplot(nainen2synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(
        S.VUOSI), shape=factor(S.VUOSI)))+
209   scale_shape_manual(values=1:nlevels(factor(S.VUOSI))) +coord_cartesian(ylim=c(0,1)
        ,xlim=c(20,80))+
210   geom_line(size=0.6) + geom_point()+
211   labs(title="Nainen Pohjois-Karjalasta", x="ikä",y="osallistumisen todennäköisyys",
        colour="syntymävuosi",shape="syntymävuosi")
212
213 # Alue 2 mies
214 mies2synt <- subset(datRe, SUKUP==1 & ALUE==2 & (synt.ehto))
215 ggplot(mies2synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(S.VUOSI
        ), shape=factor(S.VUOSI)))+
216   scale_shape_manual(values=1:nlevels(factor(S.VUOSI))) +coord_cartesian(ylim=c(0,1)
        ,xlim=c(20,80))+
217   geom_line(size=0.6) + geom_point()+
218   labs(title="Mies Pohjois-Karjalasta", x="ikä",y="osallistumisen todennäköisyys",
        colour="syntymävuosi",shape="syntymävuosi")
219
220 # Alue 3 (Pohjois-Savo) nainen
221 nainen3synt <- subset(datRe, SUKUP==2 & ALUE==3 & (synt.ehto))
222 ggplot(nainen3synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(
        S.VUOSI), shape=factor(S.VUOSI)))+
223   scale_shape_manual(values=1:nlevels(factor(S.VUOSI))) +coord_cartesian(ylim=c(0,1)
        ,xlim=c(20,80))+
224   geom_line(size=0.6) + geom_point()+
225   labs(title="Nainen Pohjois-Savosta", x="ikä",y="osallistumisen todennäköisyys",
        colour="syntymävuosi",shape="syntymävuosi")
226
227 # Alue 3 mies
228 mies3synt <- subset(datRe, SUKUP==1 & ALUE==3 & (synt.ehto))
229 ggplot(mies3synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(S.VUOSI
        ), shape=factor(S.VUOSI)))+
230   scale_shape_manual(values=1:nlevels(factor(S.VUOSI))) +coord_cartesian(ylim=c(0,1)
        ,xlim=c(20,80))+
231   geom_line(size=0.6) + geom_point()+
232   labs(title="Mies Pohjois-Savosta", x="ikä",y="osallistumisen todennäköisyys",
        colour="syntymävuosi",shape="syntymävuosi")
233
234 # Alue 4 (Turku/Loimaa) nainen
235 nainen4synt <- subset(datRe, SUKUP==2 & ALUE==4 & (synt.ehto))
236 ggplot(nainen4synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(
        S.VUOSI), shape=factor(S.VUOSI)))+
237   scale_shape_manual(values=1:nlevels(factor(S.VUOSI))) +coord_cartesian(ylim=c(0,1)
        ,xlim=c(20,80))+
238   geom_line(size=0.6) + geom_point()+
239   labs(title="Nainen Turusta/Loimaalta", x="ikä",y="osallistumisen todennäköisyys",
        colour="syntymävuosi",shape="syntymävuosi")
240
241 # Alue 4 mies
242 mies4synt <- subset(datRe, SUKUP==1 & ALUE==4 & (synt.ehto))
243 ggplot(mies4synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(S.VUOSI
        ), shape=factor(S.VUOSI)))+
244   scale_shape_manual(values=1:nlevels(factor(S.VUOSI))) +coord_cartesian(ylim=c(0,1)
        ,xlim=c(20,80))+
245   geom_line(size=0.6) + geom_point()+
246   labs(title="Mies Turusta/Loimaalta", x="ikä",y="osallistumisen todennäköisyys",
        colour="syntymävuosi",shape="syntymävuosi")
247
248 # Alue 5 (Helsinki/Vantaa) nainen
249 nainen5synt <- subset(datRe, SUKUP==2 & ALUE==5 & (synt.ehto))
250 ggplot(nainen5synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(
        S.VUOSI), shape=factor(S.VUOSI)))+
251   scale_shape_manual(values=2:nlevels(factor(S.VUOSI)))+
```

```
252   scale_colour_manual(values=c("#E58700","#C99800","#A3A500","#6BB100","#00BA38","
          #00BF7D","#00C0AF","#00BCD8","#00B0F6","#619CFF","#B983FF","#E76BF3","#FD61D1"
          ,"#FF67A4"))+
253   coord_cartesian(ylim=c(0,1),xlim=c(20,80))+ geom_line(size=0.6) + geom_point()+
254   labs(title="Nainen Helsingistä/Vantaalta", x="ikä",y="osallistumisen todennäkö
          isyys",colour="syntymävuosi",shape="syntymävuosi")
255
256 # Alue 5 mies
257 mies5synt <- subset(datRe, SUKUP==1 & ALUE==5 & (synt.ehto))
258 ggplot(mies5synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(S.VUOSI
        ), shape=factor(S.VUOSI)))+
259   scale_shape_manual(values=2:nlevels(factor(S.VUOSI)))+
260   scale_colour_manual(values=c("#E58700","#C99800","#A3A500","#6BB100","#00BA38","
          #00BF7D","#00C0AF","#00BCD8","#00B0F6","#619CFF","#B983FF","#E76BF3","#FD61D1"
          ,"#FF67A4"))+
261   coord_cartesian(ylim=c(0,1),xlim=c(20,80))+geom_line(size=0.6) + geom_point()+
262   labs(title="Mies Helsingistä/Vantaalta", x="ikä",y="osallistumisen todennäköisyys"
          ,colour="syntymävuosi",shape="syntymävuosi")
263
264 # Alue 6 (Oulun lääni) nainen
265 nainen6synt <- subset(datRe, SUKUP==2 & ALUE==6 & (synt.ehto))
266 ggplot(nainen6synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(
        S.VUOSI), shape=factor(S.VUOSI)))+
267   scale_shape_manual(values=4:nlevels(factor(S.VUOSI)))+
268   scale_colour_manual(values=c("#A3A500","#6BB100","#00BA38","#00BF7D","#00C0AF","
          #00BCD8","#00B0F6","#619CFF","#B983FF","#E76BF3","#FD61D1","#FF67A4"))+
269   coord_cartesian(ylim=c(0,1),xlim=c(20,80))+geom_line(size=0.6) + geom_point()+
270   labs(title="Nainen Oulun läänistä", x="ikä",y="osallistumisen todennäköisyys",
          colour="syntymävuosi",shape="syntymävuosi")
271
272 # Alue 6 mies
273 mies6synt <- subset(datRe, SUKUP==1 & ALUE==6 & (synt.ehto))
274 ggplot(mies6synt, aes(x=IKA, y=yhatmv7, group=factor(S.VUOSI), colour=factor(S.VUOSI
        ), shape=factor(S.VUOSI)))+
275   scale_shape_manual(values=4:nlevels(factor(S.VUOSI)))+
276   scale_colour_manual(values=c("#A3A500","#6BB100","#00BA38","#00BF7D","#00C0AF","
          #00BCD8","#00B0F6","#619CFF","#B983FF","#E76BF3","#FD61D1","#FF67A4"))+
277   coord_cartesian(ylim=c(0,1),xlim=c(20,80))+geom_line(size=0.6) + geom_point()+
278   labs(title="Mies Oulun läänistä", x="ikä",y="osallistumisen todennäköisyys",colour
          ="syntymävuosi",shape="syntymävuosi")
279
280 ###################################
281
282 ### Vakiotodennäköisyyskäyrät ###
283
284 library(fields)
285 # Funktio, joka tekee matriisin kahdesta muuttujasta ja sijoittaa halutun arvon
        matriisin ruutuihin
286 c.kuva <- function(sarakkeet, arvo){
287   N <- length(unique(sarakkeet[,1]))
288   M <- length(unique(sarakkeet[,2]))
289   c.matriisi <- matrix(0, nrow = N, ncol = M)
290   sarakkeet2 <- sarakkeet
291   sarakkeet2[,1] <- as.numeric(factor(sarakkeet2[,1]))
292   sarakkeet2[,2] <- as.numeric(factor(sarakkeet2[,2]))
293   for (i in 1:nrow(sarakkeet2))
294     c.matriisi[sarakkeet2[i,1], sarakkeet2[i,2]] <- arvo[i]
295   return(c.matriisi)
296 }
297
298 # Muodostetaan kuvaajat:
299
300 # Alue 2 (Pohjois-Karjala) nainen
301 x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
302 y <-  seq(min(datRe$IKA), max(datRe$IKA))
303 fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
304 fit$ALUE<-2
305 fit$SUKUP<-2
```

```
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn) #matriisi, jossa
    osallistumistodennäköisyydet joka ikä-syntymävuosi-leikkauksessa
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Nainen Pohjois-Karjalasta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="
    Nainen Pohjois-Karjalasta", add=T)

# Alue 2 (Pohjois-Karjala) mies
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-2
fit$SUKUP<-1
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Mies Pohjois-Karjalasta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="Mies
    Pohjois-Karjalasta", add=T)

# Alue 3 (Pohjois-Savo) nainen
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-3
fit$SUKUP<-2
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Nainen Pohjois-Savosta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="
    Nainen Pohjois-Savosta", add=T)

# Alue 3 (Pohjois-Savo) mies
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-3
fit$SUKUP<-1
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Mies Pohjois-Savosta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="Mies
    Pohjois-Savosta", add=T)

# Alue 4 (Turku/Loimaa) nainen
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-4
fit$SUKUP<-2
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Nainen Turusta/Loimaalta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="
```

```
    Nainen Turusta/Loimaalta", add=T)

# Alue 4 (Turku/Loimaa) mies
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-4
fit$SUKUP<-1
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Mies Turusta/Loimaalta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="Mies
    Turusta/Loimaalta", add=T)

# Alue 5 (Helsinki/Vantaa) nainen
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-5
fit$SUKUP<-2
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Nainen Helsingistä/Vantaalta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="
    Nainen Helsingistä/Vantaalta", add=T)

# Alue 5 (Helsinki/Vantaa) mies
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-5
fit$SUKUP<-1
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Mies Helsingistä/Vantaalta",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="Mies
    Helsingistä/Vantaalta", add=T)

# Alue 6 (Oulun lääni) nainen
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-6
fit$SUKUP<-2
tn <-  predict(mv7, newdata = fit, type="response")
c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
    ikä", main="Nainen Oulun läänistä",
           legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
               line=2.5, cex=1))
contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="
    Nainen Oulun läänistä", add=T)

# Alue 6 (Oulun lääni) mies
x <-  seq(min(datRe$S.VUOSI), max(datRe$S.VUOSI))
y <-  seq(min(datRe$IKA), max(datRe$IKA))
fit <-  data.frame(expand.grid(S.VUOSI = x, IKA = y))
fit$ALUE<-6
fit$SUKUP<-1
```

```
414 tn <-  predict(mv7, newdata = fit, type="response")
415 c.kuvaaja <- c.kuva(fit[,c("S.VUOSI", "IKA")], tn)
416 image.plot(x = x, y = y, z = c.kuvaaja,zlim=c(0,1), xlab = "syntymävuosi", ylab = "
        ikä", main="Mies Oulun läänistä",
417            legend.args=list(text='osallistumisen todennäköisyys', side=4, font=1,
                   line=2.5, cex=1))
418 contour(x = x, y = y, z = c.kuvaaja, xlab = "syntymävuosi", ylab = "ikä",main="Mies
        Oulun läänistä", add=T)
419
420 ########################################
421 # Malli 3: S.VUOSI-muuttuja jätetty pois #
422 ########################################
423
424 ms7 <- gam(cbind(K.OSAL,E.OSAL)~factor(ALUE)*SUKUP*s(IKA)*factor(VUOSI),family=
        binomial, data=datRe)
425
426 # lisätään mallin ms7 ennusteet yhatms7 dataan:
427 datRe$yhatms7<-predict(ms7,type="response")
428
429 # Muodostetaan kuvaajat:
430
431 # Alue 2 (Pohjois-Karjala) nainen
432 nainen2 <- subset(datRe, SUKUP==2 & ALUE==2)
433 ggplot(nainen2, aes(x=IKA, y=yhatms7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
434   scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(20,80))+
435   geom_line(size=0.6) + geom_point() +
436   labs(title="Nainen Pohjois-Karjalasta", x="ikä",y="osallistumisen todennäköisyys",
        colour="vuosi",shape="vuosi")
437
438 # Alue 2 mies
439 mies2 <- subset(datRe, SUKUP==1 & ALUE==2)
440 ggplot(mies2, aes(x=IKA, y=yhatms7, group=factor(VUOSI),colour=factor(VUOSI),shape=
        factor(VUOSI)))+
441   scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(20,80))+
442   geom_line(size=0.6) + geom_point()+
443   labs(title="Mies Pohjois-Karjalasta", x="ikä",y="osallistumisen todennäköisyys",
        colour="vuosi",shape="vuosi")
444
445 # Alue 3 (Pohjois-Savo) nainen
446 nainen3 <- subset(datRe, SUKUP==2 & ALUE==3)
447 ggplot(nainen3, aes(x=IKA, y=yhatms7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
448   scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(20,80))+
449   geom_line(size=0.6) + geom_point() +
450   labs(title="Nainen Pohjois-Savosta", x="ikä",y="osallistumisen todennäköisyys",
        colour="vuosi",shape="vuosi")
451
452 # Alue 3 mies
453 mies3 <- subset(datRe, SUKUP==1 & ALUE==3)
454 ggplot(mies3, aes(x=IKA, y=yhatms7, group=factor(VUOSI),colour=factor(VUOSI),shape=
        factor(VUOSI)))+
455   scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(20,80))+
456   geom_line(size=0.6) + geom_point()+
457   labs(title="Mies Pohjois-Savosta", x="ikä",y="osallistumisen todennäköisyys",
        colour="vuosi",shape="vuosi")
458
459 # Alue 4 (Turku/Loimaa) nainen
460 nainen4 <- subset(datRe, SUKUP==2 & ALUE==4)
461 ggplot(nainen4, aes(x=IKA, y=yhatms7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
462   scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
        xlim=c(20,80))+
463   geom_line(size=0.6) + geom_point() +
```

```
464   labs(title="Nainen Turusta/Loimaalta", x="ikä",y="osallistumisen todennäköisyys",
          colour="vuosi",shape="vuosi")
465
466 # Alue 4 mies
467 mies4 <- subset(datRe, SUKUP==1 & ALUE==4)
468 ggplot(mies4, aes(x=IKA, y=yhatms7, group=factor(VUOSI),colour=factor(VUOSI),shape=
        factor(VUOSI)))+
469   scale_shape_manual(values=1:nlevels(factor(VUOSI)))+coord_cartesian(ylim=c(0,1),
          xlim=c(20,80))+
470   geom_line(size=0.6) + geom_point()+
471   labs(title="Mies Turusta/Loimaalta", x="ikä",y="osallistumisen todennäköisyys",
          colour="vuosi",shape="vuosi")
472
473 # Alue 5 (Helsinki/Vantaa) nainen
474 nainen5 <- subset(datRe, SUKUP==2 & ALUE==5)
475 ggplot(nainen5, aes(x=IKA, y=yhatms7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
476   scale_shape_manual(values=3:nlevels(factor(VUOSI)))+
477   scale_colour_manual(values=c("#53B400","#00C094","#00B6EB","#A58AFF","#FB61D7"))+
          coord_cartesian(ylim=c(0,1),xlim=c(20,80))+
478   geom_line(size=0.6) + geom_point() +
479   labs(title="Nainen Helsingistä/Vantaalta", x="ikä",y="osallistumisen todennäkö
          isyys",colour="vuosi",shape="vuosi")
480
481 # Alue 5 mies
482 mies5 <- subset(datRe, SUKUP==1 & ALUE==5)
483 ggplot(mies5, aes(x=IKA, y=yhatms7, group=factor(VUOSI),colour=factor(VUOSI),shape=
        factor(VUOSI)))+
484   scale_shape_manual(values=3:nlevels(factor(VUOSI)))+
485   scale_colour_manual(values=c("#53B400","#00C094","#00B6EB","#A58AFF","#FB61D7"))+
          coord_cartesian(ylim=c(0,1),xlim=c(20,80))+
486   geom_line(size=0.6) + geom_point()+
487   labs(title="Mies Helsingistä/Vantaalta", x="ikä",y="osallistumisen todennäköisyys"
          ,colour="vuosi",shape="vuosi")
488
489 # Alue 6 (Oulun lääni) nainen
490 nainen6 <- subset(datRe, SUKUP==2 & ALUE==6)
491 ggplot(nainen6, aes(x=IKA, y=yhatms7, group=factor(VUOSI), colour=factor(VUOSI),
        shape=factor(VUOSI)))+
492   scale_shape_manual(values=4:nlevels(factor(VUOSI)))+
493   scale_colour_manual(values=c("#00C094","#00B6EB","#A58AFF","#FB61D7"))+coord_
          cartesian(ylim=c(0,1),xlim=c(20,80))+
494   geom_line(size=0.6) + geom_point() +
495   labs(title="Nainen Oulun läänistä", x="ikä",y="osallistumisen todennäköisyys",
          colour="vuosi",shape="vuosi")
496
497 # Alue 6 mies
498 mies6 <- subset(datRe, SUKUP==1 & ALUE==6)
499 ggplot(mies6, aes(x=IKA, y=yhatms7, group=factor(VUOSI),colour=factor(VUOSI),shape=
        factor(VUOSI)))+
500   scale_shape_manual(values=4:nlevels(factor(VUOSI)))+
501   scale_colour_manual(values=c("#00C094","#00B6EB","#A58AFF","#FB61D7"))+coord_
          cartesian(ylim=c(0,1),xlim=c(20,80))+
502   geom_line(size=0.6) + geom_point()+
503   labs(title="Mies Oulun läänistä", x="ikä",y="osallistumisen todennäköisyys",colour
          ="vuosi",shape="vuosi")
```