JYVÄSKYLÄN YLIOPISTO
MATEMATIIKAN JA TILASTO-
TIETEEN LAITOS

Pro gradu -tutkielma

nhmgrid: R-paketti epähomogeenisten Markovin mallien todennäköisyyksien visualisointiin

Miro Jäntti

11. maaliskuuta 2024

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Jäntti, Miro: nhmgrid: R-paketti epähomogeenisten Markovin mallien todennäköisyyksien visualisointiin

Tilastotieteen pro gradu -tutkielma, 33 sivua, 4 liitettä (6 sivua)
11. maaliskuuta 2024

Tiivistelmä

Homogeenisten Markovin mallien siirtymätodennäköisyydet voidaan esittää joko tila-avaruuden graafina tai siirtymämatriisina. Tämä on helppoa, koska jokaista tilasiirtymää vastaa yksittäinen siirtymätodennäköisyys. Epähomogeenisten Markovin mallien tapauksessa siirtymätodennäköisyyksien esittäminen on hankalampaa, koska ne voivat muuttua ajassa.

Tässä tutkielmassa esitellään tapa visualisoida ajasta riippuvia siirtymätodennäköisyyksiä siirtymäkuviomatriisina. Tällainen kuviomatriisi muistuttaa pitkälti homogeenista siirtymämatriisia, mutta sen alkiot ovat lukujen sijaan kuvioita. Siirtymätodennäköisyyksien piirtäminen kuvioina mahdollistaa suuren määrän informaatiota, kuten ryhmien välisten siirtymätodennäköisyyksien tai luottamusvälien esittämisen kompaktissa ja helposti tulkittavassa muodossa.

Tutkielmassa toteutetaan R-paketti, jota voidaan käyttää työkaluna epähomogeenisten Markovin mallien analyyseissa. Kyseiseen pakettiin on implementoitu sekä siirtymätodennäköisyyksien estimointi usealle eri mallityypille että tilasiirtymien suhteellisten frekvenssien laskeminen aineistosta ilman mallia. Paketissa on myös helppokäyttöinen piirtofunktio, jolla siirtymätodennäköisyydet voi visualisoida siirtymäkuviomatriisina.

Sen lisäksi, että pakettia käsitellään tutkielman lukuisissa kuvitteellisissa esimerkeissä, sovelletaan sitä tässä tutkielmassa myös todelliseen aineistoon. Sovellusesimerkissä on käytössä yhdysvaltalaisten nuorten tupakointikäyttäytymistä käsittelevän Minnesota Adolescent Community Cohort -tutkimuksen osa-aineisto. Tämän sovellusesimerkin tavoitteena on osoittaa paketin soveltuvuus ja hyödyllisyys tilasiirtymien analysoinnissa.

Avainsanat: epähomogeeninen Markovin malli, Markovin ketju, R-paketti, siirtymäkuviomatriisi, siirtymätodennäköisyys, tilasiirtymä, visualisointi

Sisällys

1 Johdanto

Pitkittäistutkimuksissa on tavanomaista, että havaintoja tehdään yksittäisistä havaintoyksiköistä enemmän kuin vain kerran. Usein mittauksia toistetaan tietyn ajan kuluessa, jolloin havaintoyksiköistä on tehty havaintoja monessa aikapisteessä. Tällaisia aineistoja kutsutaan pitkittäisaineistoiksi (Hsiao, 2022). Usein tutkimuksen kohteena on ihminen, jonka ominaisuuksia halutaan seurata henkilön ikääntyessä.

Tutkittaessa ihmisen ominaisuuksia, voidaan joitakin havaintoja luokitella kategorisiksi tiloiksi. Esimerkiksi tupakoinnin määrä voidaan luokitella tiloihin ei tupakoi, tupakoi vähän tai tupakoi paljon. Nämä tilat muodostavat tila-avaruuden ja henkilön on mahdollista siirtyä yhdestä tilasta toiseen jollakin siirtymätodennäköisyydellä. Siirtymää tilasta toiseen kutsutaan tilasiirtymäksi.

Tavallisesti siirtymätodennäköisyyksiä tutkitaan ajasta riippumattomina, jolloin ne on helppo esittää matriisimuodossa. Sen sijaan ajasta riippuvien siirtymätodennäköisyyksien esittäminen matriisimuodossa ei ole yhtä helppoa. Hankaluus johtuu siitä, että jokaisella aikapisteellä on oma siirtymämatriisi, joten siirtymätodennäköisyyksiä ei voi lukea vain yksittäisestä siirtymämatriisista. Ratkaisuna tähän on siirtymäkuviomatriisi, joka tiivistää eri aikapisteiden siirtymämatriisien informaation yhteen kuvioista koostuvaan matriisiin. Siirtymätodennäköisyyksien esittäminen kuviomatriisina avaa mahdollisuuksia esittää suuren määrän informaatiota kerralla.

Tämän tutkielman tavoitteena on luoda ja esitellä R-ohjelmointiympäristöön (R Core Team, 2023) paketti, jolla käyttäjä voi piirtää ja tarkastella epähomogeenisen Markovin mallin tilojen välisiä siirtymätodennäköisyyksiä visuaalisesti. Paketin tarkoituksena on toimia työkaluna analyysissä riippumatta siitä, millaisen Markovin mallin käyttäjä on analyysiä varten valinnut. Siirtymätodennäköisyyksien visuaalisen tarkastelun tarkoituksena on antaa käsitystä todennäköisyyksien muutoksista ajan kuluessa tilojen välillä ja miten todennäköisyydet eroavat mahdollisten ryhmien välillä. Tutkielman tavoitteen mukaisia paketteja ei ole tarjolla R-ympäristössä, mutta luvussa 4.2.3 nostetaan esiin muutamia olemassa olevia vastaavankaltaisia Markovin mallien siirtymätodennäköisyyksiin liittyviä paketteja.

Luku 2 keskittyy tila-avaruuksiin ja Markovin malleihin, joita voidaan käyttää mallintamaan siirtymiä tilojen välillä. Markovin malleja on erilaisia ja

kyseisessä luvussa tutustutaan yksinkertaisiin Markovin malleihin. Markovin malleissa oletetaan, että yksilö voi olla tietyllä ajanhetkellä vain yhdessä tilassa ja todennäköisyyteen siirtyä seuraavaan tilaan vaikuttaa ainoastaan tämänhetkinen tila. Tavanomaisesti Markovin mallit ovat aikahomogeenisia eli siirtymätodennäköisyydet ovat ajan suhteen vakiot. Tässä tutkielmassa keskitytään kuitenkin ajasta riippuvia siirtymätodennäköisyyksiä mallintaviin epähomogeenisiin Markovin malleihin.

Luvussa 3 tutustutaan erilaisiin lähestymistapoihin Markovin mallien sovittamiseen pitkittäisaineistoihin R-ympäristössä. Näitä lähestymistapoja ovat esimerkiksi multinomilogistiset regressiomallit ja dynaamiset paneelimallit. Luvussa keskitytään erityisesti R-pakettiin **dynamite** (Tikka ja Helske, 2023), joka käyttää Bayes-lähestymistapaa pitkittäisaineistojen analysointiin.

Luvussa 4 käsitellään tutkielman aikana tehtyä R-pakettia **nhmgrid**. Ensiksi tutustutaan paketteihin, jotka toimivat tutkielman paketin taustalla. Tämän jälkeen esitellään esimerkkien avulla tutkielman paketin ominaisuudet ja ohjeita sen käyttämiseen. Esimerkeissä käytetään paketin sisältämää kuvitteellisista henkilöistä simuloitua terveydentila-aineistoa.

Luku 5 sisältää sovellusesimerkin, jossa käytetään **dynamite**-pakettia Markovin mallin sovittamiseen sekä **nhmgrid**-pakettia siirtymätodennäköisyyksien estimointiin ja visualisointiin. Sovelluksen aineisto on peräisin vuosien 2000–2013 aikana toteutetusta Minnesota Adolescent Community Cohort -tutkimuksesta (Forster, 2016), jossa tutkittiin nuorten suhtautumista tupakointiin ja tupakointikäyttäytymistä. Tutkimuksen tavoitteena oli selvittää oliko Minnesota Youth Tobacco Prevention -aloite onnistunut vähentämään nuorten tupakoimista. Tässä sovellusesimerkissä kuitenkin keskitytään enimmäkseen **nhmgrid**-paketin soveltamiseen todelliseen aineistoon ja erityisesti tupakointitilojen välisten siirtymätodennäköisyyksien visuaaliseen tarkasteluun.

2 Tilasiirtymämallit

Seuraavaksi kerrotaan tiloista, tila-avaruuksista ja siirtymistä tilojen välillä. Tämän jälkeen tutustutaan erilaisiin Markovin malleihin ja niiden välisiin eroihin tilasiirtymien mallinnuksessa.

2.1 Tila-avaruus ja tilasiirtymä

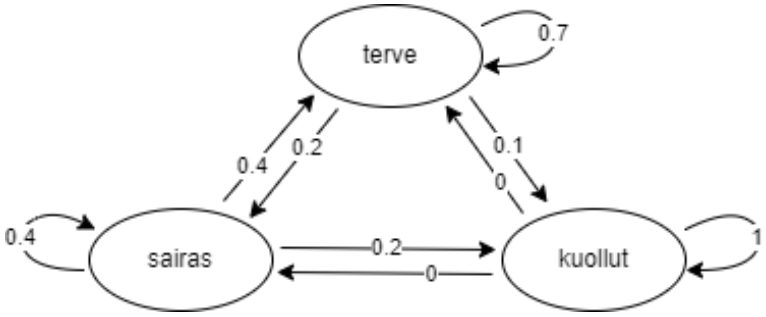
Tilasiirtymämalleissa tilalla viitataan jonkinlaiseen suureeseen, jonka arvo voi muuttua ajan kuluessa (Patterson ym., 2008). Tutkittaessa esimerkiksi ihmisen terveydentilaa, voidaan henkilön terveydentila luokitella tiloihin terve, sairas ja kuollut.

Tila-avaruus koostuu erilaisista tiloista, joiden välillä yksilö voi siirtyä. Tietyllä ajanhetkellä yksilö voi kuulua tila-avaruuden tiloista vain yhteen ja tilasta toiseen siirtyminen tapahtuu tietyllä todennäköisyydellä $0 \leq p \leq 1$. Yksilön on myös mahdollista pysyä nykyisessä tilassaan, mikä voidaan ajatella myös tilasiirtymänä.

Tilasiirtymiä tutkiessa mittausten välisellä ajalla on vaikutusta siirtymätodennäköisyyksiin. Esimerkiksi henkilön terveydentilassa tuskin havaitaan muutoksia tuntitasolla mitattuna, joten todennäköisyys siirtyä toiseen terveydentilaan saattaa olla pieni. Sen sijaan vuosittain mitattuna terveydentilojen väliset siirtymätodennäköisyydet ovat mahdollisesti suurempia.

Tila-avaruudesta ja sen siirtymistä voidaan piirtää graafi, jossa tilojen välisiä siirtymiä kuvataan nuolilla, joiden viereen merkitään siirtymän todennäköisyys. Kuviossa 1 on piirretty esimerkki tila-avaruudesta ja sen tilojen välisistä tilasiirtymistä. Kyseisen graafin siirtymätodennäköisyydet on mielivaltaisesti valittu esimerkkiä varten.

Erikoistapauksissa siirtymän todennäköisyys voi olla joko 0 tai 1. Todennäköisyyden ollessa 1, on tilasta mahdollista siirtyä vain tiettyyn tilaan ja siirtymä tapahtuu varmasti. Esimerkiksi siirtyminen kuollut-tilasta on mahdollista vain samaan tilaan, eli tila ei vaihdu. Vastaavasti kuollut-tilasta todennäköisyys siirtyä tiloihin terve tai sairas on 0, eli on mahdotonta siirtyä näihin tiloihin.



Kuvio 1: Tiloista terve, sairas ja kuollut koostuvan tila-avaruuden graafi. Tilasta lähtevä nuoli ja siihen merkitty luku esittää todennäköisyyttä siirtyä nuolen osoittamaan tilaan.

Siirtymien todennäköisyyksien on summauduttava ykköseksi (Rolski ym., 2009). Tässä tapauksessa esimerkiksi tilan sairas siirtymätodennäköisyydet, eli tilasta sairas lähteviä nuolia vastaavat todennäköisyydet summautuvat ykköseksi ($0.4 + 0.4 + 0.2 = 1$) samoin kuin tilasta kuollut ($0 + 0 + 1 = 1$). Sen sijaan tilaan saapuvia nuolia vastaavat todennäköisyydet voivat summautua miksi tahansa ei-negatiiviseksi luvuksi. Esimerkiksi tilaan terve saapuvien nuolien todennäköisyydet summautuvat yli ykköseksi ($0.4 + 0 + 0.7 = 1.1$).

2.2 Markovin mallit

Markovin malleista yksinkertaisin on Markovin ketju (Gagniuc, 2017). Tilasiirtymäprosessia kutsutaan ensimmäisen asteen Markovin ketjuksi, jos se toteuttaa niin kutsutun Markov-ominaisuuden. Tämä edellyttää, että seuraavan tilan todennäköisyys riippuu ainoastaan tämänhetkisestä tilasta eikä yhdestäkään aiemmasta tilasta (Rolski ym., 2009). Markov-ominaisuus voidaan esittää matemaattisesti ehdollisten todennäköisyyksien yhtälönä

$$\mathrm{P}(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \mathrm{P}(X_{n+1} = x | X_n = x_n),$$

missä X_i on yksilön tilaa kuvaava satunnaismuuttuja ajanhetkellä i, x tarkoittaa mielivaltaista tilaa ja x_i on tila, jossa yksilö oli ajanhetkellä i. Ajanhetki n tarkoittaa nykyhetkä ja $n + 1$ ajanhetkeä, jolloin seuraava mittaus tehdään.

Äärellisen tila-avaruuden tapauksessa tilasiirtymien lukumäärä on äärellinen, jolloin siirtymätodennäköisyydet voidaan esittää matriisimuodossa. Tällaista

matriisia kutsutaan siirtymämatriisiksi $\mathbf{T}$. Sen alkiot sisältävät todennäköisyyden siirtyä riviä vastaavasta tilasta saraketta vastaavaan tilaan.

Kuvion 1 siirtymätodennäköisyydet matriisimuodossa ovat

$$\mathbf{T} = \begin{bmatrix} 0.7 & 0.4 & 0.1 \\ 0.2 & 0.4 & 0.2 \\ 0 & 0 & 1 \end{bmatrix},$$

jos tilat terve, sairas ja kuollut indeksoidaan vastaavasti 1, 2 ja 3. Tilan indeksi kertoo miltä riviltä lähtötila ja miltä sarakkeelta saapumistila luetaan. Tässä tapauksessa esimerkiksi siirtymä tilasta terve tilaan kuollut tapahtuu todennäköisyydellä 0.1, joka on matriisin ensimmäisen rivin ja kolmannen sarakkeen alkion arvo.

Joissakin lähteissä Markovin ketjulla viitataan prosessiin, jossa aikamuuttuja on diskreetti ja sen sijaan prosessia, jossa aikamuuttuja on jatkuva, kutsutaan Markovin prosessiksi (Dodge, 2003). Joskus nähdään myös eksplisiittisesti mainittavan joko diskreettiaikainen Markovin ketju (engl. *discrete-time Markov chain*, DTMC) tai jatkuva-aikainen Markovin ketju (engl. *continuous-time Markov chain*, CTMC) (Barbot ym., 2011).

Eräs tunnettu esimerkki Markovin ketjusta on satunnaiskävely. Siinä tilaa merkitsee kokonaisluku, joka kasvaa tai pienenee yksikön verran jokaisessa siirtymässä seuraavaan aikapisteeseen. Siispä uuden tilan arvo riippuu ainoastaan edellisestä tilasta, eli prosessi toteuttaa Markov-ominaisuuden ja se on yksinkertainen Markovin ketju (Rolski ym., 2009).

Markovin ketju on oleellinen osa Markovin Monte Carlo (MCMC) -menetelmää (Roberts ja Rosenthal, 2004), jota käytetään Bayes-mallinnuksessa. Siinä posteriorijakaumasta poimitaan Markovin ketjun avulla näytteitä, joita käytetään parametrien estimoinnissa. MCMC-menetelmä perustuu siihen, että ideaalisessa tilanteessa tarpeeksi pitkään simuloitu Markovin ketju ei ole riippuvainen käyttäjän määrittämästä aloitustilasta ja siitä poimitut näytteet ovat näytteitä halutusta posteriorijakaumasta.

2.3 Epähomogeeniset Markovin mallit

Edellisessä luvussa mainituissa malleissa oletuksena on, että siirtymätodennäköisyydet ovat aikahomogeenisia eli todennäköisyys siirtyä tilasta toiseen ei muutu ajan kuluessa. Tämän tutkielman tavoitteena on visualisoida tilasiirtymien todennäköisyyksien muutoksia ajassa, joten homogeenisuus ei ole linjassa tavoitteen kanssa.

Epähomogeeniset Markovin mallit toteuttavat Markov-ominaisuuden, mutta niiden siirtymätodennäköisyydet ovat ajasta riippuvia (Brémaud ja Brémaud, 2020). Näin ollen epähomogeeniset Markovin mallit ovat mielenkiintoisia tämän tutkielman tavoitteen kannalta. Epähomogeenisessa tilanteessa malli ei estimoi yksittäistä siirtymämatriisia. Sen sijaan jokaiselle aikapisteelle estimoidaan siirtymämatriisi $\mathbf{T}(n)$, joka sisältää siirtymätodennäköisyydet tilojen välillä ajanhetkellä n (Brémaud ja Brémaud, 2020).

Visualisoinnin kannalta siirtymämatriiseja voi ajatella myös toisella tavalla. Sen sijaan, että jokaisella ajanhetkellä on oma siirtymämatriisi, voi siirtymämatriisin $\mathbf{T}$ ajatella koostuvan vektoreista todennäköisyyksiä, jotka sisältävät siirtymätodennäköisyydet kaikissa aikapisteissä. Tällöin esimerkiksi kolmen tilan tapauksessa siirtymämatriisin voi esittää muodossa

$$\mathbf{T} = \begin{bmatrix} \boldsymbol{p_{11}} & \boldsymbol{p_{12}} & \boldsymbol{p_{13}} \\ \boldsymbol{p_{21}} & \boldsymbol{p_{22}} & \boldsymbol{p_{23}} \\ \boldsymbol{p_{31}} & \boldsymbol{p_{32}} & \boldsymbol{p_{33}} \end{bmatrix},$$

missä $\boldsymbol{p_{ij}} = (\pi_{ij,1}, \pi_{ij,2}, ..., \pi_{ij,N})$ lähtötilalle i ja saapumistilalle j. Todennäköisyysvektorin alkio $\pi_{ij,n}$ on todennäköisyys tilasiirtymälle $i \rightarrow j$ ajanhetkellä n ja N on aikapisteiden lukumäärä.

Vektoreiden alkiot ovat ajallisesti järjestettyjä siirtymätodennäköisyyksiä, joten niiden visualisointi on luonnollista tehdä piirtämällä todennäköisyydet ajan funktiona. Piirtämällä vektoreista koostuvan siirtymämatriisin jokaisen alkion tilalle kuvion saadaan aikaan siirtymäkuviomatriisi. Tällöin jokaiseen yksittäiseen kuviomatriisin alkioon, eli osakuvioon voidaan sisällyttää paljon informaatiota, kuten siirtymätodennäköisyyden luottamusvälin tai jopa ryhmitellä todennäköisyydet jonkin muuttujan mukaan ja piirtää näiden ryhmien todennäköisyydet erikseen samaan kuvioon.

3 Tilasiirtymämallien sovittaminen R-ympäristössä

Tässä luvussa tutustutaan kategoriseen regressioanalyysiin, jota voidaan käyttää tilasiirtymiä tutkiessa. Aluksi esitellään Markovin mallin toteuttaminen kategorisena regressiomallina, minkä jälkeen tutustutaan eräisiin paketteihin, joilla kategorisia regressiomalleja voidaan sovittaa R-ympäristössä. Näistä paketeista tutustutaan syvemmin pakettiin **dynamite**, jota käytetään luvun 5 sovellusesimerkissä.

3.1 Regressioanalyysi tilasiirtymille

Kategorisessa regressioanalyysissä mallinnetaan vastemuuttujaa, jonka arvot ovat kategorisia (Tutz, 2011). Tila-avaruus koostuu kategorisista tiloista, joten tällainen regressiomalli sopii erinomaisesti mallintamaan tilasiirtymiä. Kategorisen regressiomallin määritteleminen Markovin mallina vaatii Markov-ominaisuuden noudattamista. Homogeenisen Markovin mallin rakenne koostuu yksinkertaisimmillaan nykyistä tilaa kuvaavasta vastemuuttujasta ja edellistä tilaa kuvaavasta selittäjästä.

Regressiomallin selittäjiksi voidaan määritellä muitakin muuttujia. Homogeenisuudesta voidaan luopua määrittelemällä aikamuuttuja mallin selittäjäksi, jolloin kyseessä on epähomogeeninen Markovin malli. Esimerkiksi ihmisen terveydentilaa tutkittaessa voi olla kiinnostavaa tutkia sukupuolen määrittämiä ryhmiä erikseen, jolloin yhdeksi mallin selittäjäksi määriteltäisiin henkilön sukupuoli. Selittäjä voi olla myös jatkuva, kuten henkilön paino.

Siirtymätodennäköisyyksien esittäminen visuaalisessa muodossa on kuitenkin hankalampaa, jos mallissa on ylimääräisiä selittäjiä mukana. Henkilön terveydentilojen väliset siirtymätodennäköisyydet voidaan estimoida ja piirtää helposti erikseen miehille ja naisille, mutta henkilön painon huomioiminen samassa kuvassa on vaikeaa. Jos tavoitteena ei kuitenkaan ole erityisemmin tutkia henkilön painon vaikutusta tilasiirtymiin, voidaan siirtymätodennäköisyyksiä estimoitaessa ylimääräiset muuttujat marginalisoida havaintojakauman yli (Hernan ja Robins, 2023). Tällöin tutkitaan henkilön edellistä tilaa kuvaavan muuttujan reunavaikutusta, joka voidaan myös laskea mahdollisille ryhmille erikseen.

Kun regressiomalli on sovitettu alkuperäiselle aineistolle, voidaan siirtymätodennäköisyydet laskea reunavaikutuksina estimoimalla mallilla todennäköisyydet kontrafaktuaaliselle aineistolle. Kontrafaktuaalisessa aineistossa tehdään interventio edellistä tilaa kuvaavaan muuttujaan. Interventiolla tutkitaan, mitä olisi tapahtunut, jos henkilön edellinen tila olisikin ollut jokin toinen tila (Hagmayer ym., 2007). Tietystä lähtötilasta lähtevien tilasiirtymien marginalisoidut siirtymätodennäköisyydet saadaan keskiarvoistamalla sellaisia estimaatteja, jotka on laskettu kontrafaktuaaliselle aineistolle, jossa jokaiseen havaintoriviin on asetettu edellistä tilaa kuvaavan muuttujan arvoksi kyseinen lähtötila. Tämä menettely toistetaan jokaiselle tila-avaruuden tilalle, jolloin saadaan lasketuksi marginalisoidut siirtymätodennäköisyyksien estimaatit jokaiselle tila-avaruuden tilasiirtymälle.

Eräs kategorinen regressiomalli on multinomilogistinen regressiomalli, joka on yleistys logistisesta regressiosta useammalle kuin kahdelle luokalle (Kwak ja Clayton-Matthews, 2002). R-ympäristössä multinomilogistisia regressiomalleja voidaan sovittaa pitkittäisaineistoihin esimerkiksi paketilla **mlogit** (Croissant, 2020) tai **nnet** (Venables ja Ripley, 2002). Osassa tämän tutkielman esimerkeissä käytetään paketilla **nnet** sovitettua mallia.

Pitkittäisaineistojen analysointiin on esitetty myös erilaisia paneelimalleja, kuten dynaamiset paneelimallit (Arellano ja Bond, 1991), ristiviivästetyt paneelimallit (Allison ym., 2017) sekä näiden erilaiset muunnelmat (Mulder ja Hamaker, 2021; Zyphur ym., 2020). Dynaamisten monimuuttujapaneelimallien (Helske ja Tikka, 2022) tavoitteena on ratkaista ongelmia ja rajoituksia, jotka johtuvat aiempien menetelmien vaatimista oletuksista.

Dynaamiset monimuuttujapaneelimallit sallivat aikamuuttumattomia ja -muuttuvia sekä yksilökohtaisia vaikutuksia. Tämän lisäksi ne mahdollistavat usean vastemuuttujan samanaikaisen mallintamisen sekä sallivat mielivaltaisia viivästettyjä riippuvuussuhteita. Näitä malleja voidaan sovittaa Bayeslähestymistavalla R-paketilla **dynamite** (Tikka ja Helske, 2023), joka tukee myös kategorisen vastemuuttujan mallintamista.

3.2 Tilasiirtymien mallintaminen R-paketilla dynamite

Seuraavaksi tutustutaan dynaamisten monimuuttujapaneelimallien sovittamiseen paketilla **dynamite**. Malli sovitetaan `dynamite`-funktiolla, jonka rakenne on

```
dynamite(dformula, data, time, group = NULL, ...),
```

missä ensimmäinen parametri `dformula` on tyyppiä `dynamiteformula`, josta kerrotaan lisää luvussa 3.2.1. Mallille syötettävän aineiston (`data`) tulee sisältää mallin lausekkeessa käytetyt muuttujat. Aikamuuttujan nimi asetetaan `time`-parametrin arvoksi ja pitkittäisaineiston tapauksessa yksilön tai ryhmän tunnisteeseen viittaavan muuttujan nimi tulee merkitä parametrin `group` arvoksi. Jos arvon jättää oletusarvoon `NULL`, olettaa `dynamite` kaikkien havaintojen kuuluvan samaan ryhmään, mikä ei ole ominaista pitkittäistutkimuksissa.

Kolme pistettä parametrilistan lopussa vihjaa, että funktiolle voi syöttää ylimääräisiä parametreja, jotka ohjataan `dynamite`-funktion taustalla käyttämiin funktioihin. Esimerkiksi MCMC-menetelmän parametrit voidaan määrittää tässä, koska `dynamite` käyttää oletuksena taustalla **rstan**-pakettia (Stan Development Team, 2024) posteriorijakauman simuloinnissa. Näitä MCMC-menetelmän parametreja ovat esimerkiksi käytettävien Markovin ketjujen lukumäärä (`chains`, oletuksena 4), iteraatioiden lukumäärä (`iter`, oletuksena 2 000) ja lämmittelyjakson pituus (`warmup`, oletuksena $\lfloor \texttt{iter}/2 \rfloor$).

3.2.1 Mallin rakenteen kirjoittaminen

Mallin lausekkeen rakentamiseen voidaan käyttää funktiota `obs`, joka on funktion `dynamiteformula` alias. Tämä funktio luo R-ympäristön perinteisestä kaavasyötteestä `dynamiteformula`-tyyppisen olion, joka sopii funktiolle `dynamite` parametriksi. Funktion `obs` rakenne on

```
obs(formula, family),
```

missä `formula` on kaavasyöte muodossa $y \sim x$ ja `family` on kaavassa käytetyn vastemuuttujaan `y` liittyvän virhetermin jakaumaperheen nimi. Käytettävissä olevia jakaumaperheitä ovat muun muassa gaussian (normaalijakauma), binomial (binomijakauma), categorical (luokkajakauma) ja multinomial

(multinomijakauma). Kaikki mahdolliset jakaumaperheet on listattu paketin **dynamite** dokumentaatiossa. Tilasiirtymiä mallinnettaessa vastemuuttujan jakaumaksi määritellään luokkajakauma.

Kaavan formalisointi noudattaa R-ympäristön tuttuja käytänteitä. Vastemuuttujan nimi kirjoitetaan ~ -merkin vasemmalle puolelle ja selittävät muuttujat merkin oikealle puolelle. Mikäli selittäviä muuttujia on useita, erotellaan ne + -merkein. Vakiotermin voi halutessaan jättää mallista pois asettamalla luvun -1 yhdeksi selittäjäksi.

Oletusarvoisesti selittäjät ovat aikamuuttumattomia, mutta ne voidaan merkitä `obs`-funktiossa aikamuuttuviksi funktiolla `varying`. Esimerkiksi mallissa, joka on määritelty

```
obs(y ~ x + varying(~ -1 + z), "categorical") +
   splines(df = 10),
```

mallinnetaan ehdollisesti luokkajakautunutta vastemuuttuujaa `y` aikamuuttumattomalla selittäjällä `x` ja aikamuuttuvalla selittäjällä `z`. Mallissa voi olla mukana vain joko aikamuuttumaton tai -muuttuva vakiotermi. Yllä olevassa mallissa on aikamuuttumaton vakiotermi.

Aikamuuttuvien selittäjien tapauksessa `dynamite` käyttää kuutiosplinejä regressiokertoimien estimoinnissa. Tämän takia mallin rakenteeseen tulee määritellä funktiolla `splines` kuutiosplinin vapausaste (oletuksena 4). Yllä olevaan malliin vapausasteeksi on määritelty 10. Mikäli aikamuuttuvia selittäjiä olisi useita, käyttäisi `dynamite` jokaisen tällaisen selittäjän vaikutuksen estimoinnissa 10-asteista spliniä, mutta jokaiselle selittäjälle estimoitaisiin omat spliniparametrit.

Malliin voidaan lisätä viivästettyjä selittäjiä funktiolla `lag`. Kyseiselle funktiolle syötetään parametreiksi haluttu muuttuja, jota viivästetään, ja vapaaehtoisesti myös haluttu viive. Oletusarvoisesti `lag` käyttää yhden askeleen viivettä, joka on sopiva ensimmäisen asteen Markovin mallin sovittamiseen.

3.2.2 Usean vastemuuttujan rakenne

Mikäli mallinnettavia vastemuuttujia on useita, voidaan useita tyyppiä `dynamiteformula` olevia olioita yhdistää + -merkeillä, jolloin `dynamite`-funktio pyrkii mallintamaan kaikkia vastemuuttujia samanaikaisesti. Tämä on mah-

dollista, mikäli muuttujien väliset riippuvuussuhteet eivät muodosta silmukkaa. Esimerkiksi rakenne

```
obs(y ~ x, "categorical") +
   obs(x ~ z, "categorical"),
```

missä muuttujaa `y` mallinnetaan `x`-muuttujalla ja samanaikaisesti `x`-muuttujaa mallinnetaan muuttujalla `z`, on validi. Sen sijaan rakenne

```
obs(y ~ x, "categorical") +
   obs(x ~ z, "categorical") +
   obs(z ~ y, "categorical"),
```

ei ole, koska päättelyketju muodostaa silmukan. Päättelyketju saa muodostaa silmukan siinä tapauksessa, jos selittävä muuttuja on viivästetty. Esimerkiksi rakenne

```
obs(y ~ x, "categorical") +
   obs(x ~ lag(y), "categorical")
```

on validi, sillä muuttujaa `x` selitetään viivästetyllä `y`-muuttujan arvolla, joka voidaan olettaa tunnetuksi tässä vaiheessa.

4 nhmgrid-paketti

Tässä luvussa käsitellään tutkielman aikana tehtyä R-pakettia **nhmgrid**. Ensiksi kerrotaan kolmannen osapuolen paketeista, joita tutkielman paketti käyttää taustalla toimiakseen. Tämän jälkeen tutustutaan **nhmgrid**-paketin ominaisuuksiin ja toimintaperiaatteisiin.

4.1 Riippuvuus toisista paketeista

Kuvien piirtämiseen R-ympäristössä löytyy useita paketteja, kuten **ggplot2** (Wickham, 2016), **lattice** (Sarkar, 2008) ja **base** (R Core Team, 2023). Näistä viimeisin eroaa muista siinä, että **base** on sisäänrakennettu osa R-ympäristöä ja **ggplot2** sekä **lattice** ovat kolmannen osapuolen tahojen kehittämiä.

Siirtymäkuviomatriisien piirtämiseen käytetään **ggplot2**-pakettia, jolla piirretyt kuviot ovat helposti muokattavissa käyttäjän tarpeiden mukaan. Tämän lisäksi siirtymätodennäköisyyksien estimoimiseen käytetään apuna **marginaleffects**-pakettia (Arel-Bundock, 2024). Seuraavaksi esitellään **ggplot2**-paketin oleellisimmat ominaisuudet. Tämän jälkeen tutustutaan pakettiin **marginaleffects** ja kerrotaan sen tarjoamista hyödyistä siirtymätodennäköisyyksien estimoinnissa.

4.1.1 ggplot2

Grafiikkapaketti **ggplot2** (Wickham, 2016) tarjoaa käyttäjälle monipuolisen rajapinnan sekä yksinkertaisten että monimutkaisten kuvioiden piirtämiseen R-ympäristössä. Se on yksi suosituimmista visualisointityökaluista R-ympäristöön sisäänrakennetun **base**-paketin piirtotyökalujen ohella (Teutonico, 2015). Pohjimmiltaan **ggplot2** perustuu grafiikan kielioppiin, jonka Wilkinson (2005) esittelee kirjassaan. Wickham (2010) esittelee artikkelissaan vaihtoehtoisen parametrisaation grafiikan kieliopille, jossa lopullinen grafiikka on kokoelma erilaisia päällekäisiä tasoja tai komponentteja, joiden perusteet esitellään seuraavaksi.

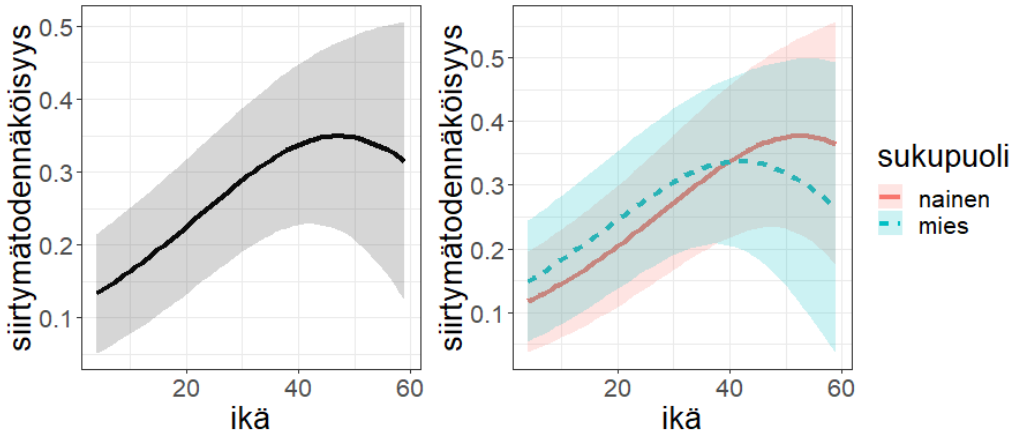
Kuvio alustetaan `ggplot`-funktiolla, jolla määritellään oletuksena käytettävä aineisto ja joukko kuvauksia aineiston muuttujista kuvassa näkyvien elementtien ominaisuuksiin (engl. *aesthethics*). Tämän jälkeen kuvioon lisätään + -merkein eroteltuna geometrisia tasoja, jotka määrittävät, millaisessa muodossa aineisto piirretään. Paketissa **ggplot2** on useita geometrisia komponentteja, joita käyttäjä voi sisällyttää kuvaan `geom_`-alkuisilla funktioilla. Eräs tällainen geometrinen komponentti on `geom_point`, joka piirtää aineistossa olevat havaintopisteet niiden x- ja y-koordinaattikuvausten mukaan. Kuhunkin geometriseen tasoon käyttäjä voi halutessaan määrittää matemaattisen muunnosfunktion sekä oletuksesta poikkeavan aineiston ja ominaisuuskuvaukset.

Ominaisuuskuvausten avulla kuvioon voi sisällyttää paljon informaatiota. Kuvaus voidaan tehdä esimerkiksi aineiston diskreetistä muuttujasta kuvioon piirrettävien havaintopisteiden muotoon tai väriin, jolloin havainnot ovat tunnistettavia kuviosta. Vastaavasti jatkuva muuttuja voidaan kuvata esimerkiksi havaintopisteiden kokoon, mikä helpottaa havaintojen vertailua.

Visualisointi **ggplot2**-paketilla noudattaa siis hyvin systemaattista kaavaa. Johdonmukaisuutensa ansiosta jopa monimutkaisten kuvioiden piirtäminen on tehty kyseisellä paketilla helpoksi ja kuvioiden muotoilu halutun näköiseksi on selkeää. Tästä syystä **ggplot2** valittiin toimimaan **nhmgrid**-paketin piirtofunktion taustalla.

Epähomogeenisten Markovin mallien siirtymätodennäköisyydet **nhmgrid** visualisoi jatkuvana käyränä. Käyrän piirtoon käytetään funktiota `geom_line`, jolle määritellään seuraavanlaiset ominaisuuskuvaukset. Vaaka-akselin arvoiksi kuvataan aineiston aikamuuttujan arvot, pystyakselin arvoiksi estimoidut siirtymätodennäköisyydet ja mikäli käyttäjä on estimoinut siirtymätodennäköisyydet eri ryhmille erikseen, niin käyrän väriksi ja viivan tyyliksi kuvataan diskreetin ryhmämuuttujan arvo. Jos käyttäjä on estimoinut todennäköisyyksille luottamus- tai posteriorivälit, niin **nhmgrid** piirtää ne `geom_ribbon`-funktiolla.

Kuviossa 2 on esimerkkejä siirtymätodennäköisyyskuvioista, jotka on piirretty **ggplot2**-paketilla. Vasemmassa kuviossa on piirretty todennäköisyyskäyrä sekä sen luottamusväli. Oikeassa kuviossa käyrä ja sen luottamusväli on piirretty sukupuolen määrittämille ryhmille erikseen. Oletuksena **ggplot2** lisää selitteet ominaisuuskuvauksille kuvion oikealle puolelle. Poikkeuksena on vaaka- ja pystyakselien arvojen kuvaukset, koska näiden arvot näkyvät kuvion akseleissa.



Kuvio 2: Esimerkkejä **ggplot2**-paketilla piirretyistä todennäköisyyskäyristä ja niiden luottamusväleistä. Vasemmassa kuviossa käyrä on piirretty ilman ryhmittelyä ja oikeassa kuviossa ryhmittely on tehty sukupuolen mukaan.

Tutkielman paketin tavoitteena on koota monta kuviota ruudukkomaisesti matriisiin, joka sisältää siirtymätodennäköisyyskuviot jokaiselle siirtymälle tila-avaruuden tilojen välillä. Paketissa **ggplot2** on funktio `facet_grid`, joka soveltuu tähän erinomaisesti. Kyseistä funktiota käytetään lisäämällä se tasoksi **ggplot2**-olioon ja funktion parametreilla määritellään, mitä muuttujia käytetään kuvioiden erottelussa. Lopullinen kuva on ruudukkomainen ja koostuu yhtä monesta kuviosta kuin erottelevien muuttujien kombinaatioita on. Tutkielman paketin tapauksessa erotteleviksi muuttujiksi asetetaan lähtö- sekä saapumistila, jolloin lopullisessa kuviomatriisissa on oma osakuvio jokaiselle tilasiirtymälle.

4.1.2 marginaleffects

Mallien sovittamiseen on tarjolla R-ympäristössä paljon erilaisia paketteja. Yhden mallityypin oliorakenne ei välttämättä vastaa jonkin toisen mallityypin rakennetta, joten tavallisesti käyttäjän täytyy kirjoittaa R-koodia mallityyppikohtaisesti analyysissään. Ratkaisuna tähän on paketti **marginaleffects** (Arel-Bundock, 2024), jonka avulla usean eri mallityypin käyttäminen onnistuu yhden rajapinnan kautta. Tutkielman paketissa hyödynnetään tämän rajapinnan marginalisoitujen ennusteiden laskemiseen tarkoitettua `avg_predictions`-funktiota.

Eräs erityisen hyödyllinen ominaisuus tässä paketissa on tuki usealle eri mallityypille. Uusille mallityypeille **marginaleffects**-paketin tuen lisääminen on tehty helpoksi. Uusien mallityyppien ilmaantuessa myös **nhmgrid** tukee niitä, mikäli niihin lisätään **marginaleffects**-tuki. Tämän ansiosta tutkielman pakettia ei tarvitse erikseen päivittää, jotta se toimisi tällaisten uusien mallityyppien kanssa.

Dynaamiset monimuuttujapaneelimallit ovat esimerkki eräästä mallityypistä, jolla ei tämän tutkielman kirjoitushetkellä ole **marginaleffects**-tukea. Tästä syystä **nhmgrid**-pakettiin on implementoitu **dynamite**-tuki erikseen. Parempi ratkaisu kuitenkin olisi, jos tuki **dynamite**-malleille implementoitaisiin paketin **marginaleffects** sisäisenä ominaisuutena. Vaikka tuki on toteutettu **dynamite**-malleille erikseen, on niiden käyttäminen toteutettu yhdenmukaisesti muiden mallityyppien kanssa.

4.2 Paketin nhmgrid ominaisuudet

Paketin ominaisuuksiin kuuluu siirtymätodennäköisyyksien estimointi mallipohjaisesti sekä näiden todennäköisyyksien visualisointi siirtymäkuviomatriisina. Paketilla voidaan myös laskea tilojen välisiä siirtymäosuuksia suoraan aineistosta. Siirtymäosuuksien visualisointi toimii samalla tavalla kuin mallilla estimoitujen siirtymätodennäköisyyksien visualisointi, mutta tuottaa oletuksena eri geometrisista komponenteista koostuvia siirtymäkuvioita.

Paketti **nhmgrid** sisältää keinotekoisen pitkittäisaineiston `health`, joka sisältää simuloituja havaintoja ihmisten terveydentiloista. Kuvitteellisia ihmisiä aineistossa on 100 ja jokaisesta henkilöstä tiedossa on sekä heidän sukupuoli että ikä. Mittauksia on jokaisesta henkilöstä 10 ja ne on tehty vuoden välein. Kyseinen aineisto on käytössä sekä tämän tutkielman luvun 4 että paketin dokumentaation esimerkeissä. Aineiston tarkoitus on toimia yksinkertaisena esimerkkinä tilasiirtymäaineistosta sekä olla apuna paketin käytön aloittamisessa. Sen sijaan aineiston tarkoituksena ei ole vastata todellisten ihmisten terveydentilojen muutoksia.

Tutkielman paketti on asennettavissa GitHub-versionhallintapalvelusta käyttämällä apuna esimerkiksi **devtools**-pakettia (Wickham ym., 2022). Tällöin paketin voi asentaa komennolla

```
devtools::install_github("mirojantti/nhmgrid").
```

Seuraavaksi siirrytään kuvailemaan toimenpiteitä paketin **nhmgrid** käytön aloittamiseksi siinä tapauksessa, että tilasiirtymämalli on jo sovitettu. Tämän jälkeen esitellään paketin piirtofunktio, jonka jälkeen käsitellään siirtymäosuuksien laskemista ja visualisointia sekä vertaillaan muiden pakettien vastaavankaltaisia ominaisuuksia.

4.2.1 Siirtymätodennäköisyyksien estimointi mallilla

Tämän luvun esimerkeissä `health`-aineistoon on sovitettu yksinkertainen multinomilogistinen regressiomalli paketin **nnet** (Venables ja Ripley, 2002) funktiolla `multinom`. Vastemuuttuujaa terveydentila selitetään muuttujilla ikä, sukupuoli ja henkilön edellinen terveydentila.

Oleellisessa roolissa **nhmgrid**-paketissa on `stprob`-olio, joka noudattaa R-kielen S3-olio-ohjelmointiperiaatteita. S3-järjestelmä mahdollistaa luokkien lisäämisen olioille ja geneeristen funktioiden implementoinnin näille luokille (Wickham, 2019). Olio `stprob` sisältää estimaatit kaikkien tilojen välisistä siirtymätodennäköisyyksistä jokaisessa aikapisteessä. Voidakseen visualisoida siirtymätodennäköisyyksiä **nhmgrid**-paketilla, on käyttäjän ensin estimoitava ne.

Mikäli käyttäjän tilasiirtymämalli on sovitettu **dynamite**-paketilla tai se on jokin **marginaleffects**-paketin tukema malli, voi käyttäjä estimoida siirtymätodennäköisyydet funktiolla `stprobs` (*state transition probabilities*), joka palauttaa käyttäjälle `stprob`-olion. Tämän funktion rakenne on

```
stprobs(model, x = NULL, lag_state = NULL,
        group = NULL, interval = 0.95),
```

missä ainoat vaaditut parametrit ovat `model` ja `x`. Parametri `model` on joko **dynamite**-malli tai jokin **marginaleffects**-tuen omaava malli. Parametrin `x` arvoksi tulee asettaa aikamuuttujan nimi. Paketilla **dynamite** sovitettujen mallien tapauksessa parametrin `x` arvon voi kuitenkin jättää määrittämättä, koska `stprobs`-funktio osaa tunnistaa aikamuuttujan **dynamite**-mallista automaattisesti.

Muut parametrit ovat vapaavalintaisia. Funktio `stprobs` pyrkii tunnistamaan edellistä tilaa kuvaavan muuttujan mallista automaattisesti. Mikäli automaattinen tunnistus ei jostain syystä onnistu, kehottaa funktion tuottama virheviesti käyttäjää asettamaan manuaalisesti parametrin `lag_state`

arvoksi edellistä tilaa kuvaavan muuttujan nimen. Parametrin `group` arvoksi voi käyttäjä halutessaan asettaa ryhmittelevän muuttujan nimen, jolloin `stprobs` estimoi siirtymätodennäköisyydet ryhmille erikseen. Funktio `stprobs` laskee oletuksena estimaattien 95 %:n luottamus- tai posteriorivälin, mutta välin leveys on määriteltävissä parametrilla `interval`.

Funktio `stprobs` estimoi siirtymätodennäköisyydet käyttäen paketin **marginaleffects** funktiota `avg_predictions`, joka laskee marginalisoidut ennusteet ja niiden epävarmuudet mallista tilasiirtymille. Luottamusvälien laskemiseen **marginaleffects** käyttää delta-menetelmää (Arel-Bundock, 2024). Jos parametrin `model` arvoksi on syötetty **dynamite**-malli, laskee `stprobs` siirtymätodennäköisyydet ja niiden posteriorivälit posteriorinäytteiden keskiarvoina ja kvantiileina. Posteriorinäytteet saadaan paketin **dynamite** geneerisen `fitted`-funktion implementaatiolla.

Malleille, jotka eivät ole **dynamite**-malleja tai eivät ole tuettuja paketissa **marginaleffects**, täytyy käyttäjän estimoida siirtymätodennäköisyydet manuaalisesti. Tällöin `stprob`-olio luodaan funktiolla `manual_stprob`, jonka rakenne on

```
manual_stprob(prob),
```

missä parametri `prob` on siirtymätodennäköisyysestimaatit sisältävä tyyppiä `data.frame` oleva taulukko. Taulukko 1 esittää esimerkin funktiolle syötettävän taulukon rakenteesta. Taulukon tulee sisältää estimaatit kaikille mahdollisille tilasiirtymäkombinaatioille kaikissa aikapisteissä. Aikapisteet merkitään sarakkeeseen `x`. Siirtymän lähtötila merkitään sarakkeeseen `from` ja saapumistila sarakkeeseen `to`. Siirtymätodennäköisyyden estimaatti tietylle tilasiirtymälle `from` $\rightarrow$ `to` aikapisteessä `x` merkitään sarakkeeseen `mean`. Tähdellä merkityt sarakkeet ovat valinnaisia, jotka tulee täyttää, jos estimaatit on laskettu ryhmille erikseen (`group`) tai siirtymätodennäköisyyksien estimaateille on laskettu luottamus- tai posteriorivälit (`lower` ja `upper`). Kyseinen funktio tarkastaa sille syötetyn taulukon sisällön ja antaa virheilmoituksen käyttäjälle mikäli taulukko ei vastaa vaadittua muotoa.

Taulukko 1: Esimerkki funktiolle `manual_stprob` syötettävästä taulukkorakenteesta. Tähdellä merkityt sarakkeet ovat valinnaisia.

x	from	to	group*	mean	lower*	upper*
4	terve	terve	mies	0.86	0.79	0.92
4	terve	terve	nainen	0.89	0.84	0.95
4	terve	sairas	mies	0.14	0.07	0.20
4	terve	sairas	nainen	0.11	0.05	0.16
4	terve	kuollut	mies	0.00	0.00	0.01
4	terve	kuollut	nainen	0.00	0.00	0.00
4	sairas	terve	mies	0.84	0.74	0.93
...	...	...	...	...	...	...
5	terve	terve	mies	0.85	0.78	0.92
...	...	...	...	...	...	...

4.2.2 Siirtymätodennäköisyyksien visualisointi

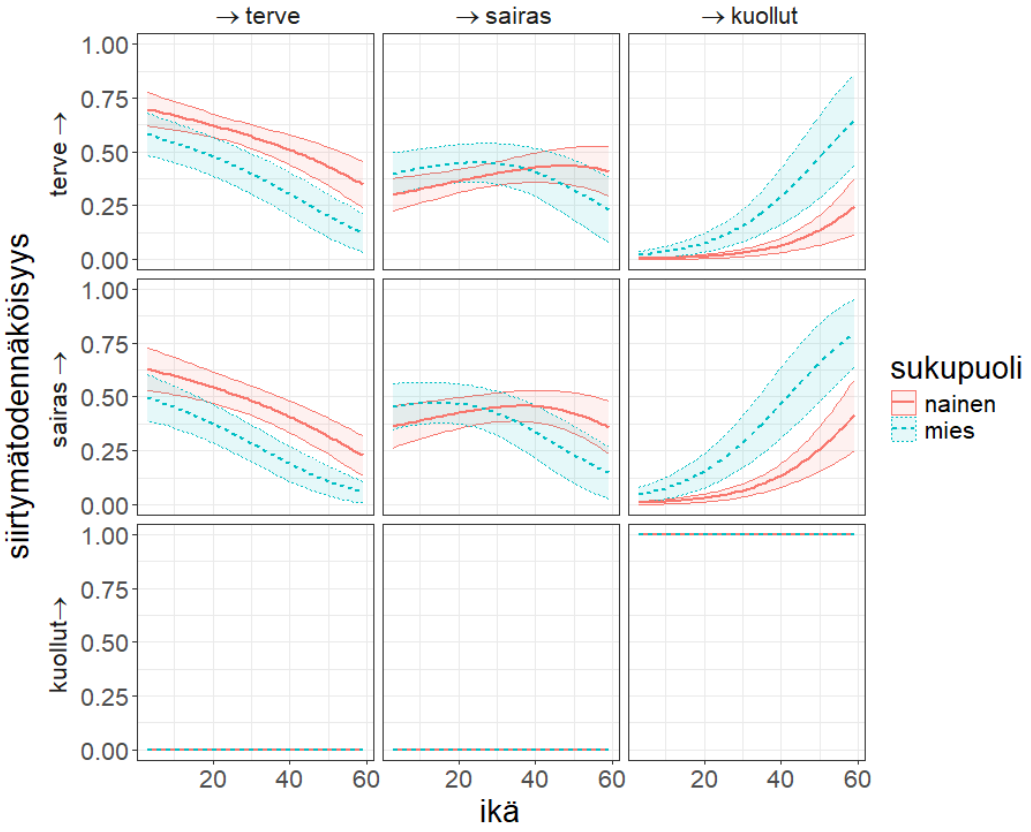
Siirtymätodennäköisyydet piirretään geneerisen `plot`-funktion implementaatiolla, jonka rakenne on

```
plot.stprob(x, default_geoms = TRUE),
```

missä ainut vaadittu parametri `x` on luvussa 4.2.1 kuvatulla tavalla luotu `stprob`-olio. Oletuksena piirtofunktio lisää siirtymäkuviomatriisin osakuvioihin geometrisia komponentteja, mutta halutessaan käyttäjä voi määrittää parametrin `default_geoms` arvoksi `FALSE`, jolloin piirtofunktio ainoastaan luo siirtymäkuviomatriisin rakenteen, jossa jokaista tilasiirtymää vastaava osakuvio on tyhjä.

Funktion palauttama olio on **ggplot2**-kuvio, joten siihen voi kohdistaa muotoiluja tai lisätä uusia geometrisia komponentteja helposti. Jos esimerkiksi parametrin `default_geoms` avulla oletuskomponentit on jätetty piirtämättä, voi käyttäjä määrittää itse, mitä geometrisia komponentteja osakuvioiden piirtämisessä käytetään lisäämällä piirtofunktion palauttamaan olioon tasoja kuten luvussa 4.1.1 on kuvailtu.

Kuvio 3 esittää siirtymäkuviomatriisin, joka on estimoitu `health`-aineistolle käyttäen multinomilogistista regressiomallia. Kuviomatriisi sisältää sukupuolen määrittämille ryhmille erikseen piirretyt tilasiirtymien todennäköisyyksien estimaatit 95 %:n luottamusväleineen.



Kuvio 3: Siirtymäkuviomatriisi `health`-aineiston terveydentilojen terve, sairas ja kuollut välisille siirtymille. Ryhmittely on tehty sukupuolen mukaan.

Siirtymäkuviomatriisia tulkitaan seuraavasti. Lähtötila luetaan rivin alusta. Tässä tapauksessa ylimmän rivin (terve→) osakuviot kuvaavat siirtymätodennäköisyyksiä, joissa lähtötila on terve. Sarakkeen alkuun merkitty tila ilmaisee saapumistilaa. Tässä tapauksessa keskimmäisen sarakkeen (→sairas) osakuviot kertovat todennäköisyyksistä, kun saapumistila on sairas. Tiettyjen tilojen välinen siirtymätodennäköisyys katsotaan kuviomatriisin tietystä alkiosta tällä logiikalla. Esimerkiksi tilasiirtymän terve→sairas siirtymätodennäköisyys on piirretty kuviomatriisin ylimmän rivin keskimmäiseen alkioon.

4.2.3 Siirtymäosuuksien laskeminen aineistosta

Mallipohjaisen siirtymätodennäköisyyksien tarkastelun lisäksi **nhmgrid** kykenee laskemaan ja visualisoimaan siirtymäosuuksia suoraan aineistosta ilman Markovin mallia. Siirtymäosuudella tarkoitetaan tilasiirtymien suhteellisia frekvenssejä. Seuraavaksi tutustutaan siirtymäosuuksien laskemiseen ja piirtämiseen **nhmgrid**-paketilla, jonka jälkeen vertaillaan vastaavanlaisia toteutuksia toisista paketeista.

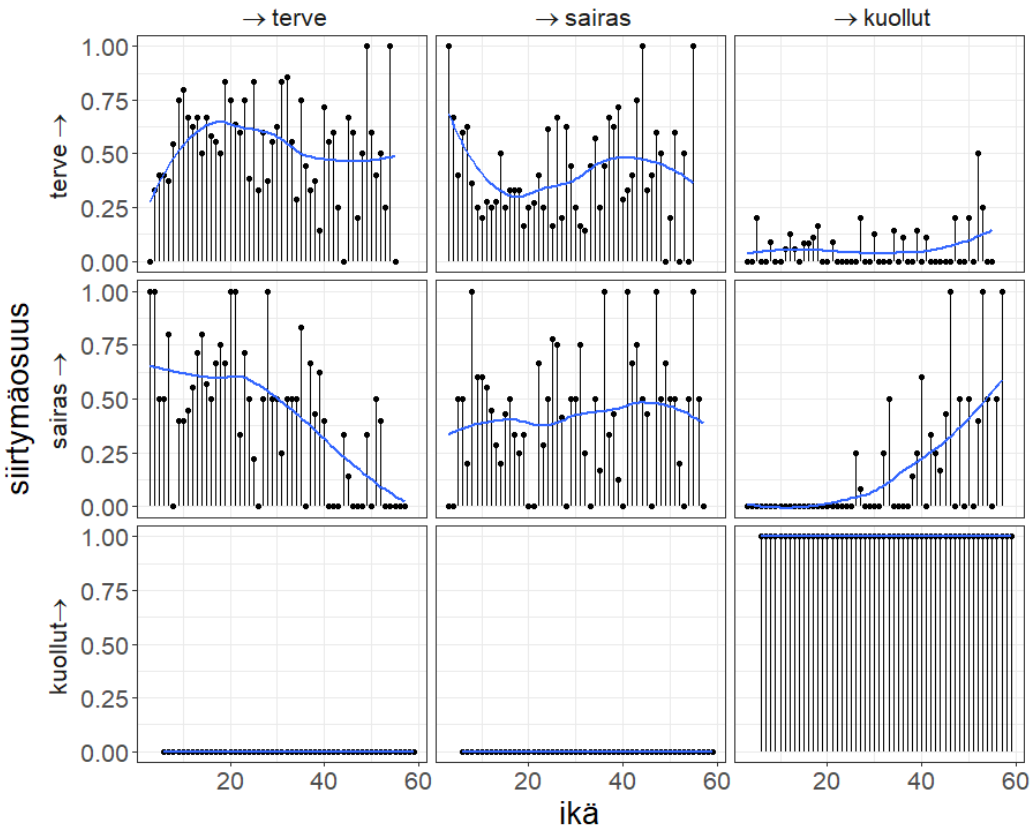
Osuudet lasketaan funktiolla `stprops` (*state transition proportions*), jonka rakenne on

```
stprops(data, id, state, x, group = NULL),
```

missä `data` on pitkittäismuotoinen aineisto, `id` on yksilöä kuvaavan muuttujan nimi, `state` on mitattua tilaa vastaavan muuttujan nimi ja `x` on aikaa kuvaavan muuttujan nimi. Näiden lisäksi tarkastelu voidaan halutessaan ryhmitellä jonkin luokittelevan muuttujan mukaan, jonka sarakkeen nimi asetetaan parametrin `group` arvoksi. Funktio laskee siirtymien osuudet aineistosta ja palauttaa `stprob`-olion, jonka voi piirtää kuten luvussa 4.2.2.

Mikäli aineistossa on aikapisteitä, joissa ei ole lainkaan havaittu siirtymiä jostakin lähtötilasta, asettaa `stprops`-funktio kyseisen lähtötilan jokaiselle siirtymälle näissä aikapisteissä osuudeksi `NA` eli puuttuvan tiedon. Puuttuvat siirtymäosuudet **nhmgrid** jättää piirtämättä kuvioon. Osuuskuvioiden piirtäminen eroaa todennäköisyyskuvioiden piirtämisestä siinä, että jatkuvan käyrän sijaan osuudet piirretään tikkarikuvioina (engl. *lollipop plot*). Tikkarikuvio koostuu paketin **ggplot2** geometrisista komponenteista `geom_point` ja `geom_segment`.

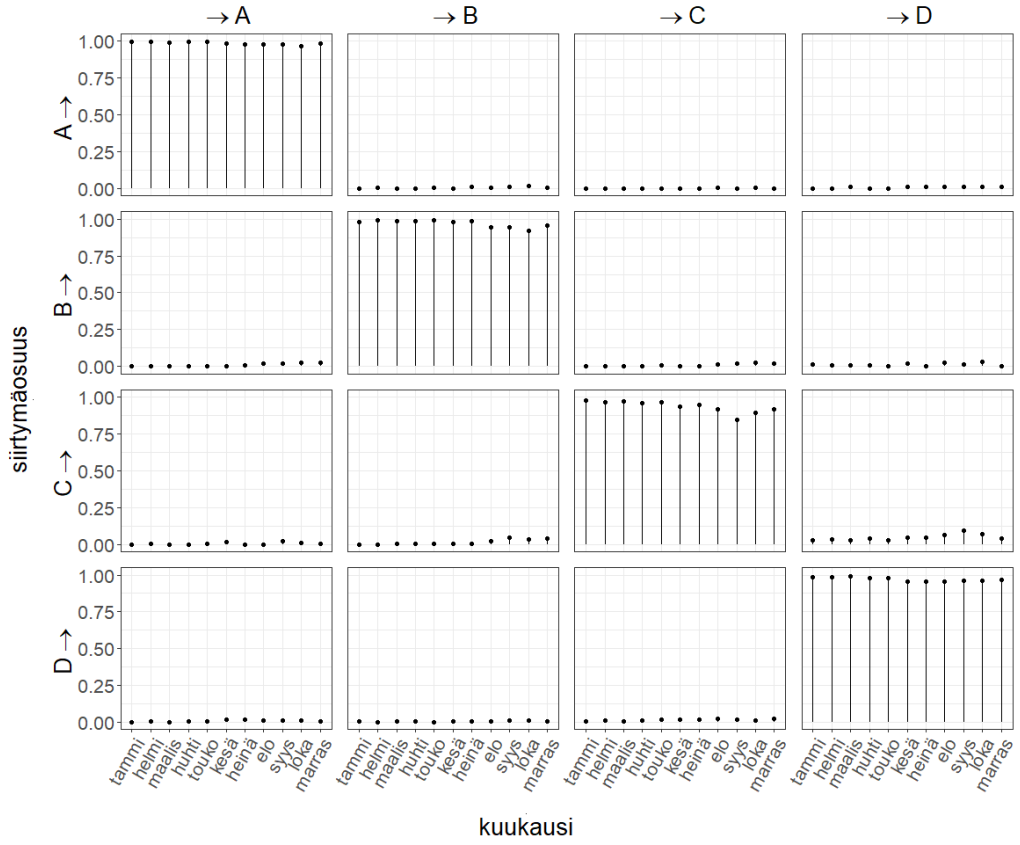
Kuvio 4 esittää **nhmgrid**-paketin sisältämästä `health`-aineistosta lasketut siirtymäosuudet. Havaintojen lukumäärä on melko pieni suhteessa aikamuuttujan arvojen lukumäärään, joten osuudet ovat melko epätasaiset. Tästä syystä kuvioihin on lisätty tasoituskäyrät käyttäen paikallisesti estimoitua sirontakuviosilotusta (engl. *locally estimated scatterplot smoothing*, LOESS) (Cleveland, 1979) paketin **ggplot2** funktiolla `geom_smooth`. Tasoituskäyrät muistuttavat hieman aiemman esimerkin (kuvio 3) siirtymätodennäköisyyskäyriä, mikä on erityisesti havaittavissa liitteen A siirtymäkuviomatriisista A.1, johon on piirretty `health`-aineistosta estimoidut ryhmittelemättömät siirtymätodennäköisyydet ja -osuudet yhdessä.



Kuvio 4: Siirtymäkuviomatriisi `health`-aineistosta lasketuille terveydentilojen välisille siirtymäosuuksille. Kuvioon on lisätty LOESS-menetelmällä tasoituskäyrät.

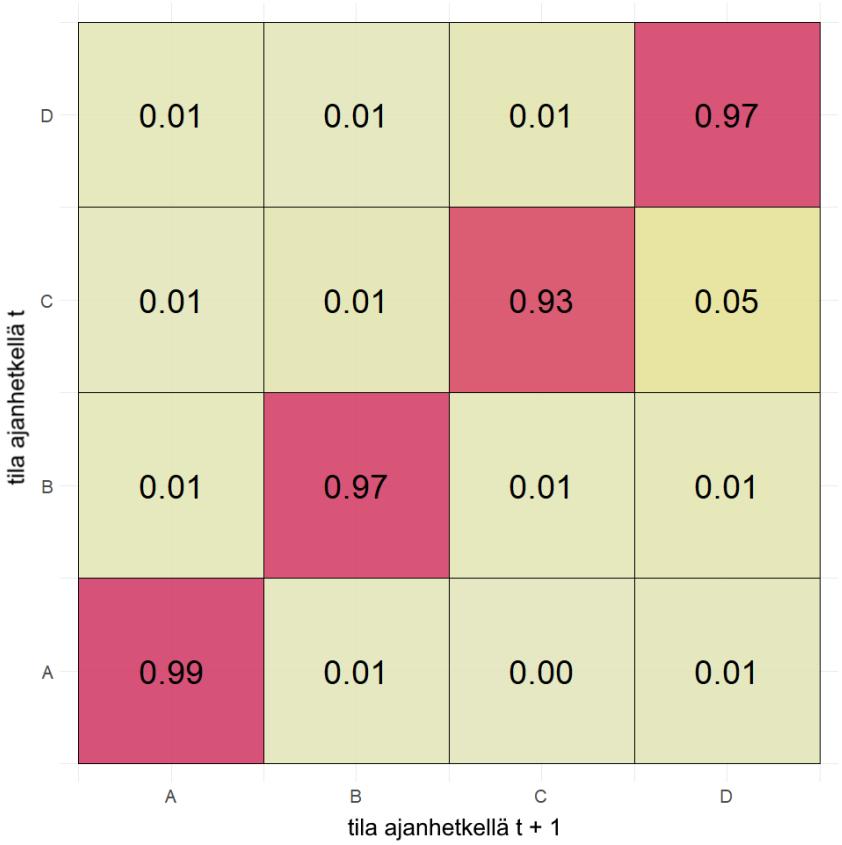
Vastaavanlainen ominaisuus osuuksien laskemiseen löytyy myös paketista **TraMineR** (Gabadinho ym., 2011), jossa siirtymäosuudet lasketaan funktiolla `seqtrate`. Sen sijaan kyseinen paketti ei tarjoa työkaluja näiden osuuksien visualisoimiseen tämän tutkielman tavoitteen mukaisesti. Siirtymäosuudet, jotka on laskettu funktiolla `seqtrate`, voidaan kuitenkin visualisoida muuntamalla ne ensin `stprob`-olioksi paketin **nhmgrid** funktiolla `as.stprob`. Kyseinen muuntofunktio on implementoitu ainoastaan olioille, jotka on luotu `seqtrate`-funktiolla käyttäen parametrin `time.varying` arvona `TRUE`. Kun kyseisen parametrin arvo on asetettu näin, kohtelee `seqtrate` tilasiirtymiä epähomogeenisena Markovin ketjuna.

Kuviossa 5 on visualisoitu funktion `seqtrate` dokumentaation mukaisesti lasketut tilojen väliset siirtymäosuudet `actcal`-aineistolle, joka on sisäänrakennettu osa **TraMineR**-pakettia. Tämän esimerkin tarkoituksena on osoittaa **TraMineR**-yhteensopivuus, eikä suinkaan tehdä johtopäätöksiä saaduista tuloksista. Todettakoon kuitenkin, että siirtymäosuudet näyttävät melko homogeenisilta, mikä vihjaa siitä, että todennäköisyys siirtyä tilasta toiseen ei riipu ajasta.



Kuvio 5: Funktiolla `seqtrate` lasketut tilojen väliset siirtymäosuudet paketin **TraMineR** `actcal`-aineistolle. Siirtymäkuviomatriisi on piirretty **nhmgrid**-paketilla.

TraMineR-paketissa on kuitenkin piirtofunktioita toisenlaisiin tarkoituksiin, mutta ne on toteutettu **base**-paketin piirtotyökaluilla. Paketissa **ggseqplot** (Raab, 2022) on **ggplot2**-versiot paketin **TraMineR** piirtofunktioista. Yhtenä eroavaisuutena alkuperäisiin piirtofunktioihin on funktio `ggseqtrplot`, jolla voi piirtää funktion `seqtrate` laskemat siirtymäosuudet matriisimuodossa. Se ei vaadi alkuperäisen `seqtrate`-funktion palauttaman olion muuntamista toiseen muotoon, kuten **nhmgrid** vaatii. Kuten kuviossa 6 on havaittavissa, se ei kuitenkaan piirrä siirtymäosuuksia ajan funktiona, eli sillä ei voi visualisoida epähomogeenisten Markovin mallien siirtymätodennäköisyyksiä tämän tutkielman tavoitteiden mukaisesti.



Kuvio 6: Funktiolla `seqtrate` lasketut tilojen väliset siirtymäosuudet paketin **TraMineR** `actcal`-aineistolle. Siirtymämatriisi on piirretty **ggseqplot**-paketin funktiolla `ggseqtrplot`.

5 Sovellusesimerkki

Tässä luvussa esitellään Minnesota Adolescent Community Cohort (MACC) -paneelitutkimusta (Forster, 2016) ja siinä kerättyä aineistoa. Kyseinen tutkimuksen aineisto on saatavilla osoitteessa `https://www.icpsr.umich.edu/web/ICPSR/studies/36282`. Tässä sovellusesimerkissä MACC-tutkimuksen osa-aineistoon sovitetaan Markovin malli dynaamisilla monimuuttujapaneelimalleilla käyttäen **dynamite**-pakettia ja tavoitteena on tutkia osa-aineistosta estimoituja siirtymätodennäköisyyksiä visuaalisesti paketilla **nhmgrid**.

5.1 MACC-paneelitutkimus

MACC-tutkimus toteutettiin vuosien 2000–2013 aikana Yhdysvalloissa. Tutkimuksen tavoitteena oli tutkia murrosikäisten nuorten tupakointitottumuksia ja onnistuvatko ennaltaehkäisevät toimenpiteet vaikuttamaan nuorten tupakoinnin aloittamiseen. Tutkimuksen aikaan Minnesotassa oli meneillään Minnesota Youth Tobacco Prevention Initiative -aloite, jonka tavoitteena oli vähentää nuorten tupakointia 30 prosentilla vuoteen 2005 mennessä.

Tutkimukseen osallistui 4 824 nuorta, joista noin 3 600 oli Minnesotan osavaltiosta. Osallistujiksi valittiin henkilöitä, jotka olivat tutkimuksen aloitusvuonna 12–16-vuotiaita. Aineisto kerättiin puhelinhaastatteluina, ja haastattelut toteutettiin puolivuosittain, joten kyseessä on pitkittäisaineisto. Aineisto on melko laaja, sillä osallistujia haastateltiin 13 vuoden aikana 26 kertaa ja haastattelukysymyksiä oli paljon.

Puhelinhaastattelujen kysymykset eivät olleet jokaisella kerralla samat. Kysymykset koskivat pääosin henkilön ja hänen perheensä tupakointitottumuksia, mutta sen lisäksi osa kysymyksistä koski muun muassa alkoholinkäyttöä ja sosioekonomista asemaa.

5.2 Osa-aineisto

Tässä sovellusesimerkissä käytössä on MACC-tutkimuksen osa-aineisto, joka sisältää vähemmän haastattelukertoja ja muuttujia kuin alkuperäinen aineisto. Osallistujia tässä osa-aineistossa on 4 395. Joissain tapauksissa osallistujaa oli haastateltu vuoden aikana kolme kertaa, mikä rikkoo aikamuuttujan

tasavälisyyden. Nämä 426 havaintoriviä päätettiin poistaa osa-aineistosta, vaikka havaintojen poistaminen ei ole paras lähestymistapa tässä tilanteessa. Kyseinen päätös tehtiin, koska tämän sovellusesimerkin ensisijaisena tavoitteena on esitellä paketin **nhmgrid** käyttöä.

Aineistosta on karsittu suuri osa muuttujista pois. Sukupuolen ja iän lisäksi tiedossa on, kuinka usein osallistuja näkee tupakointia elokuvissa ja kuinka usein hän itse polttaa tupakkaa. Jälkimmäiset määrät on luokiteltu alkuperäisen aineiston luokkia yksinkertaisemmin. Osa-aineistossa on huomioitu vain vuosien 2004–2008 aikana toteutetut kahdeksan haastattelukertaa, sillä ainoastaan tällä aikavälillä haastattelut sisälsivät kysymyksiä sekä osallistujan tupakoinnin määrästä että elokuvissa nähdyn tupakoinnin määrästä.

Haastateltavalta kysyttiin, kuinka usein hän näkee näyttelijöiden tupakoivan elokuvissa. Vastaukset on luokiteltu luokkiin ei näe koskaan, näkee harvoin, näkee joskus ja näkee usein. Osa-aineistoa varten luokka ei näe koskaan yhdistettiin luokkaan näkee harvoin luokkien yksinkertaistamiseksi. Haastateltavalta kysyttiin, kuinka usein hän käyttää tupakkatuotteita. Vastaukset luokiteltiin alkuperäisen aineiston kuuden luokan sijaan neljään luokkaan. Luokiksi määriteltiin ei tupakoi, kuukausittain, viikottain ja päivittäin.

Tutkimukseen osallistujan tupakoinnin määrä on luokiteltu yllä mainittuihin neljään luokkaan. Kukin luokka voidaan nähdä tila-avaruuden tilana. Osallistujan tupakointi saattaa muuttua osa-aineistossa olevien neljän vuoden haastattelujen välillä puolivuosittain, jolloin tapahtuu tilasiirtymä yhdestä tupakointitilasta toiseen.

Tämän sovellusesimerkin tavoitteena on osa-aineiston perusteella selvittää, miten nuorten tupakointikäyttäytyminen muuttuu iän mukaan, ja tarkastella, onko sukupuolella tai tupakoinnin näkemisellä elokuvissa vaikutusta tupakoinnin vähentymiseen. Erityisesti kiinnostavaa on tutkia siirtymätodennäköisyyksien muutoksia eri tupakointitilojen välillä ajan kuluessa käyttäen **nhmgrid**-paketin visualisointityökaluja.

5.3 Markovin malli ja sen visualisointi

Seuraavaksi sovitetaan MACC-tutkimuksen osa-aineistoon Markovin malli tavoitteena mallintaa siirtymiä henkilön tupakoinnin määrää kuvaavien tilojen välillä. Käytetään mallin sovitukseen dynaamista monimuuttujapaneelimallia, joka sovitetaan R-ympäristössä käyttäen **dynamite**-pakettia.

Mallin vastemuuttujana on henkilön tupakoinnin määrä, joka on neliluokkainen (ei tupakoi, tupakoi kuukausittain, tupakoi viikottain tai tupakoi päivittäin). Selittäviksi muuttujiksi valitaan henkilön sukupuoli, tupakoinnin määrä edellisellä mittauskerralla sekä kolmiluokkainen muuttuja, joka kuvastaa kuinka usein kyseinen henkilö altistuu tupakoinnin näkemiselle elokuvissa (harvoin, joskus tai usein). Näistä selittäjistä sukupuoli ja altistuminen määritellään aikamuuttumattomiksi selittäjiksi malliin. Aikamuuttuvana selittäjänä mallissa on henkilön edellinen tupakoimistila, ja tälle aikamuuttuvalle selittäjälle on määritelty splini vapausasteella 4 (oletus **dynamite**-paketin funktiossa `splines`). Henkilön ikä on asetettu mallin aikaa kuvaavaksi muuttujaksi.

R-koodi mallin sovittamiseen ja MCMC-menetelmän diagnostiikkatarkasteluihin on liitteessä B. Mallissa käytetään neljää Markovin ketjua ja MCMC-menetelmässä käytetään **cmdstanr**-pakettia (Gabry ym., 2023). Kussakin ketjussa iteraatioiden lukumäärä on 2 000, joista puolet käytetään algoritmin lämmittelyjaksoon, eli yhteensä käytettäviä näytteitä posteriorijakaumasta on 4 000.

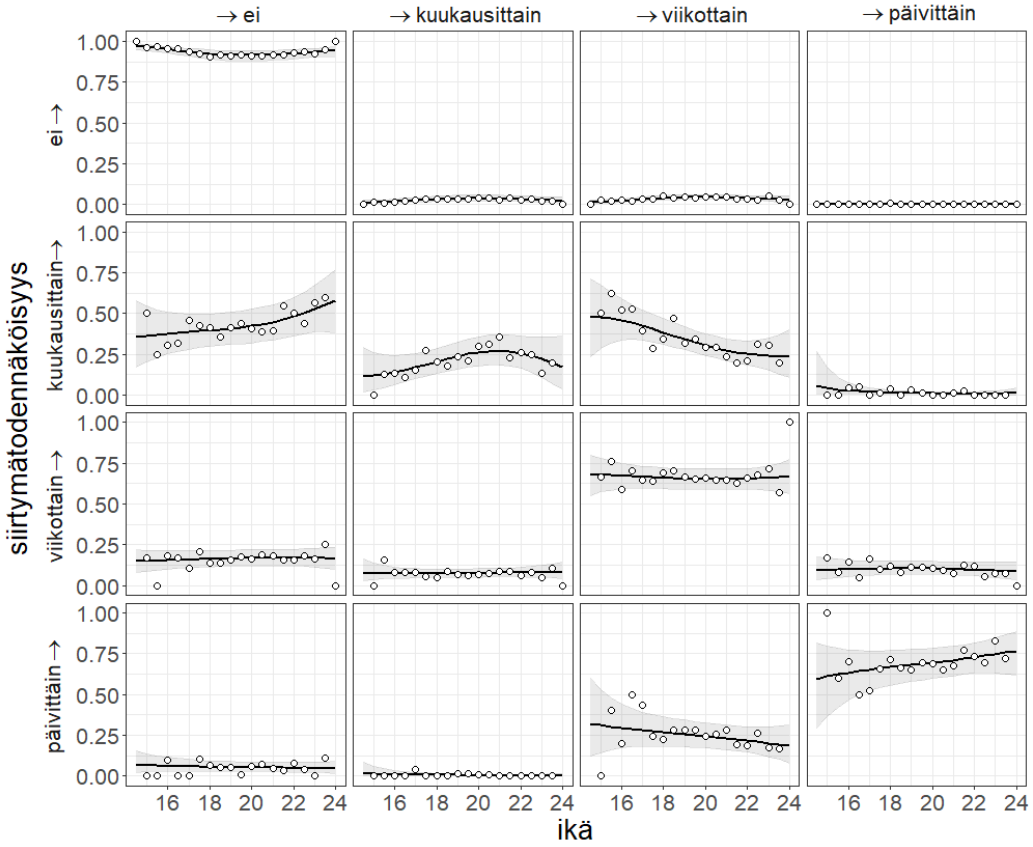
MCMC-menetelmän diagnostiikasta käy ilmi, että vain kaksi simulaation iteraatioista päättyi divergenssiin. Parametrien $\hat{R}$-arvot ovat korkeimmillaan 1, mikä viittaa siihen, että malli on konvergoinut. Posteriorinäytteiden tehollliset otoskoot ovat pienimmilläänkin riittävän suuria. Diagnostiikan mukaan mallin tuloksiin voidaan siis luottaa. Mallin regressiokertoimien posteriorikeskiarvot 95 %:n posteriorivälеineen on piirretty kuvioihin, jotka ovat tutkielman liitteessä C. Aikamuuttumattomat regressiokertoimet ovat kuviossa C.1 ja aikamuuttuvat kuviossa C.2.

Visualisoidaan mallin mukaiset siirtymätodennäköisyydet käyttäen pakettia **nhmgrid**. Siirtymätodennäköisyydet estimoidaan `stprobs`-funktiolla. Ensiksi kokeiltiin piirtää siirtymätodennäköisyydet ryhmiteltynä sekä sukupuolen (kuvio D.1) että tupakoinnin näkemiselle altistumisen (kuvio D.2) mukaan. Nämä kaksi siirtymäkuviomatriisia on liitteessä D. Ryhmittelyllä ei kuitenkaan havaittu riittävän mielenkiintoisia visuaalisia eroja ryhmien välillä, joten päätettiin tutkia aineistoa yhtenä kokonaisuutena ja marginalisoida ryhmien yli.

Kuvio 7 esittää mallilla estimoidut tupakointitilojen väliset siirtymätodennäköisyydet 95 %:n posteriorivälеineen. Vaaka-akselilla on henkilön ikä ja pystyakselilla todennäköisyys siirtyä rivin tilasta sarakkeen tilaan. Mallinnettu aineisto sisältää mittauksia puolen vuoden välein, joten siirtymäkuviomatrii-

sin siirtymätodennäköisyydet kertovat todennäköisyydestä siirtyä tupakointitilasta toiseen puolen vuoden aikana.

Yhtenä mittarina mallin hyvyydelle voidaan pitää mallin sopivuutta aineistoon. Tästä syystä siirtymätodennäköisyysestimaattien lisäksi osakuvioihin on piirretty valkoisilla pisteillä siirtymäosuudet, jotka on laskettu aineistosta käyttäen funktiota `stprops`. Siirtymäosuuksien lisääminen kuviomatriisiin on toteutettu samaan tapaan kuin liitteessä A. Todennäköisyyskäyrät ja niiden posteriorivälit näyttävät sopivan hyvin yhteen osuuspisteiden kanssa, eli malli vaikuttaa sopivan aineistoon hyvin.



Kuvio 7: Henkilön tupakoinnin määrää kuvailevien tilojen väliset siirtymätodennäköisyydet ja niiden 95 %:n posteriorivälit visualisoitu **nhmgrid**-paketilla. Valkoiset pisteet kuvaavat aineistosta laskettuja siirtymäosuuksia.

Lyhennetään tupakointitilat seuraavasti: ei tupakoi lainkaan (EI), tupakoi kuukausittain (KK), tupakoi viikottain (VK) ja tupakoi päivittäin (PV). Käytetään näitä lyhenteitä seuraavissa tulkinnoissa. Suurin osa siirtymätodennäköisyyksistä näyttää melko vakiolta. Tämä vihjaa siitä, että näiden tilojen väliset siirtymätodennäköisyydet ovat homogeenisia, eli ne eivät riipu ajasta. Tästä syystä nämä tilasiirtymät eivät ole varsinaisen kiinnostavia visuaalisen tarkastelun kannalta.

Osassa kuvioista on sen sijaan havaittavissa epähomogeenisuutta. Erityisesti lähtötilasta KK siirtymätodennäköisyydet näyttävät riippuvan henkilön iästä. Myöskin siirtymät tilasta PV tiloihin VK tai PV näyttävät sisältävän monotonisen trendin. Erityisesti huomattavaa on, että todennäköisyys äärimmäiselle muutokselle henkilön tupakoinnissa on hyvin pieni, eli esimerkiksi ei ole todennäköistä paljon tupakoivalle henkilölle lopettaa tupakointia kokonaan kerralla. Kuitenkin siirtymän VK→EI todennäköisyys on melko suuri verrattuna muihin tilasiirtymiin, joissa niin sanotusti hypätään vierekkäisen tilan yli.

Siirtymätodennäköisyyksillä on tapana olla suurimpia siirtymämatriisin diagonaalin alkioilla (Raab, 2022), mikä tarkoittaa, että on todennäköisempää pysyä nykyisessä tilassaan kuin siirtyä toiseen tilaan. Tämä väite näyttää pitävän paikkansa myös tämän sovelluksen tapauksessa poislukien tilasiirtymää KK→KK, jonka todennäköisyys on pienempi kuin sen vieressä olevien tilasiirtymien KK→EI ja KK→VK todennäköisyydet kaikissa aikapisteissä. Tämä tarkoittaa, että kuukausittain tupakoivat henkilöt todennäköisemmin joko lopettavat tupakoinnin kokonaan tai tupakoivat jatkossa viikottain kuin jatkaisivat kuukausittain tupakoimista. Tilasiirtymän KK→EI todennäköisyys kasvaa iän myötä, eli kuukausittain tupakoivista henkilöistä vanhemmat lopettavat tupakoinnin nuorempia todennäköisemmin. Tämä taas johtaa siihen, että vastaavasti kuukausittain tupakoivista henkilöistä nuoremmat tupakoivat jatkossa viikottain vanhempia henkilöitä todennäköisemmin, mikä näkyy tilasiirtymän KK→VK laskevasta trendistä.

Tilasiirtymien PV→VK ja PV→PV kuvaajien perusteella näyttää siltä, että mitä vanhempi henkilö on, sitä todennäköisempää on jatkaa päivittäin tupakoimista. On kuitenkin huomattava, että siirtymien PV→VK ja PV→PV todennäköisyyksien posterioriväleille voisi piirtää vakiosuoran, jonka mukaan todennäköisyydet olisivat homogeenisia. Siispä on hankala tehdä varmaa tulkintaa iän vaikutuksesta näiden kahden tilasiirtymän siirtymätodennäköisyyksiin.

6 Johtopäätökset

Tässä tutkielmassa esitettiin tapaan, jolla epähomogeenisen Markovin mallin siirtymämatriiseja voi esittää visuaalisessa muodossa. Ajatuksena on yhdistää jokaisen aikapisteen siirtymämatriisi yhdeksi siirtymäkuviomatriisiksi, ja jokainen kuviomatriisin alkio sisältää tietyn tilasiirtymän todennäköisyyskuvion. Tällaisella lähestymistavalla voidaan visualisoida suuri määrä informaatiota kompaktisti. Siirtymäkuviomatriisi voi sisältää esimerkiksi estimaattien lisäksi niiden luottamusvälin. Myös samaan kuviomatriisiin voi sisällyttää estimaatit ryhmiteltynä jonkin aineiston muuttujan, kuten sukupuolen mukaan. Näin suuri määrä informaatiota olisi hankala esittää tavallisessa matriisimuodossa, jossa on kuvioiden sijasta lukuja.

Tämä ajatusmalli implementoitiin R-pakettiin **nhmgrid**, joka on avoimesti saatavilla GitHub-versionhallintapalvelussa osoitteessa `https://github.com/mirojantti/nhmgrid`. Paketilla on mahdollista estimoida siirtymätodennäköisyydet usealla eri mallityypillä tai laskea tilasiirtymien suhteelliset frekvenssit aineistosta. Näiden visualisointi on toteutettu **ggplot2**-paketilla, ja siirtymäkuviomatriisien muokkaaminen käyttäjän tarpeiden mukaan on tarvittaessa helppoa noudattaen **ggplot2**-käytänteitä. Paketin **nhmgrid** käytön aloittaminen on helppoa seuraamalla sen dokumentaation esimerkkejä, joissa käytetään paketin sisältämää simuloitua terveydentila-aineistoa.

Paketin sisäänrakennetun aineiston lisäksi pakettia sovellettiin vuonna 2016 julkaistuun Minnesota Adolescent Community Cohort -paneelitutkimukseen. Sovelluksessa oli käytössä tutkimuksen osa-aineisto ja tarkoituksena oli tarkastella visuaalisesti, miten nuorten tupakointikäyttäytyminen muuttuu iän myötä. Osa-aineistoon sovitettiin dynaaminen monimuuttujapaneelimalli R-ympäristössä käyttäen pakettia **dynamite**. Malliin otettiin ryhmitteleviä selittäjiä mukaan ryhmien mahdollisten erojen havaitsemiseksi. Ryhmämuuttujat olivat henkilön sukupuoli ja kuinka usein henkilö altistuu tupakoinnin näkemiselle elokuvissa.

Sovellusesimerkin ensisijaisena tarkoituksena oli kokeilla tutkielman pakettia **nhmgrid** oikeaan aineistoon, ja varsinaisten tulosten tulkinta oli toissijaista. Tässä tutkielmassa ei havaittu suuria eroja ryhmien välillä, joten siirtymätodennäköisyydet estimoitiin marginalisoiden ryhmien yli. Osa-aineistosta estimoitujen siirtymätodennäköisyyksien kuvioista havaittiin, että ainakin joidenkin tilasiirtymien osalta aineistossa on havaittavissa epähomogeenisuutta iän suhteen. Siirtymäkuviomatriisin estimointi ja visualisointi oli helppoa

paketilla **nhmgrid**. Tutkielman paketista vaikuttaa siis olevan käytännön hyötyä tilasiirtymiä sisältävien pitkittäisaineistojen tutkimisessa.

Paketin **nhmgrid** tärkein vahvuus on joustavuus käyttäjän tarpeiden mukaan. Siirtymätodennäköisyyksien estimointi ei vaadi käytettävän jotain tiettyä mallityyppiä, vaan käyttäjä voi estimoida siirtymätodennäköisyydet joko **dynamite**-mallilla tai jollakin paketin **marginaleffects** tukemalla mallilla. Halutessaan käyttäjä voi estimoida siirtymätodennäköisyydet myös manuaalisesti ja silti käyttää **nhmgrid**-pakettia siirtymäkuviomatriisin piirtoon.

Tutkielman pakettia voisi kuitenkin jatkokehittää joidenkin ominaisuuksien osalta. Kuten tämän tutkielman siirtymäkuviomatriiseista on havaittavissa, kaikkien tilasiirtymien visuaalinen tarkastelu ei ole mielekästä, jos siirtymätodennäköisyys on ajasta riippumaton. Siispä paketin piirtofunktiolla voisi olla hyvä mahdollistaa vain tiettyjen tilasiirtymien kuvioiden piirtäminen. Tällöin käyttäjä voisi keskittyä vain mielenkiintoisten kuvioiden tarkasteluun ja kuvioiden piirtämiselle isompana jäisi enemmän tilaa.

Kuviomatriisissa olevien tekstien sisällön muuttaminen on lähtökohtaisesti helppoa käyttämällä paketin **ggplot2** funktiota `labs`. Tällä funktiolla voidaan asettaa esimerkiksi kuviomatriisille otsikko tai nimetä akselien tunnisteet uudelleen. Tilojen uudelleennimeäminen on kuitenkin hankalaa, mikä on havaittavissa liitteen A R-koodista. Pakettiin voisi implementoida ratkaisun, joka tekisi sisällön muuttamisesta helppoa sellaisille teksteille, joita ei voi muuttaa funktiolla `labs`.

Viitteet

P. D. Allison, R. Williams & E. Moral-Benito. Maximum likelihood for cross-lagged panel models with fixed effects. *Socius*, 3:2378023117710578, 2017. doi: 10.1177/2378023117710578. URL `https://doi.org/10.1177/2378023117710578`.

V. Arel-Bundock. *marginaleffects: Predictions, comparisons, slopes, marginal means, and hypothesis tests*, 2024. URL `https://marginaleffects.com/`. R-paketin versio 0.17.0.9002.

M. Arellano & S. Bond. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *The Review of Economic Studies*, 58(2):277–297, 1991. ISSN 00346527, 1467937X. URL `http://www.jstor.org/stable/2297968`.

B. Barbot, T. Chen, T. Han, J.-P. Katoen & A. Mereacre. Efficient CTMC model checking of linear real-time objectives. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, sivut 128–142. Springer, 2011.

P. Brémaud & P. Brémaud. Non-homogeneous Markov chains. *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*, sivut 399–422, 2020.

W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Y. Croissant. Estimation of random utility models in R: the mlogit package. *Journal of Statistical Software*, 95(11):1–41, 2020. doi: 10.18637/jss.v095.i11. URL `https://www.jstatsoft.org/index.php/jss/article/view/v095i11`.

Y. Dodge. *The Oxford dictionary of statistical terms*. OUP Oxford, 2003.

J. Forster. Minnesota adolescent community cohort (MACC) study 2000–2013, 2016.

A. Gabadinho, G. Ritschard, N. S. Müller & M. Studer. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37, 2011. doi: 10.18637/jss.v040.i04.

J. Gabry, R. Češnovar & A. Johnson. *cmdstanr: R Interface to 'CmdStan'*, 2023. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

P. A. Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017.

Y. Hagmayer, S. A. Sloman, D. A. Lagnado & M. R. Waldmann. Causal reasoning through intervention. *Causal learning: Psychology, philosophy, and computation*, sivut 86–100, 2007.

J. Helske & S. Tikka. Estimating causal effects from panel data with dynamic multivariate panel models, 2022. URL `osf.io/preprints/socarxiv/mdwu5`.

M. Hernan & J. Robins. *Causal inference: what if*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023. ISBN 9781420076165. URL `https://books.google.fi/books?id=_KnHIAAACAAJ`.

C. Hsiao. *Analysis of panel data*. Numero 64. Cambridge University Press, 2022.

C. Kwak & A. Clayton-Matthews. Multinomial logistic regression. *Nursing research*, 51(6):404–410, 2002.

J. D. Mulder & E. L. Hamaker. Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(4):638–648, 2021. doi: 10.1080/10705511.2020.1784738. URL `https://doi.org/10.1080/10705511.2020.1784738`.

T. A. Patterson, L. Thomas, C. Wilcox, O. Ovaskainen & J. Matthiopoulos. State–space models of individual animal movement. *Trends in Ecology and Evolution*, 23(2):87–94, 2008. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2007.10.009. URL `https://www.sciencedirect.com/science/article/pii/S0169534707003588`.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Wien, Itävalta, 2023. URL `https://www.R-project.org/`.

M. Raab. *ggseqplot: Render sequence plots using 'ggplot2'*, 2022. URL `https://maraab23.github.io/ggseqplot/`.

G. O. Roberts & J. S. Rosenthal. General state space Markov chains and MCMC algorithms. 2004.

T. Rolski, H. Schmidli, V. Schmidt & J. L. Teugels. *Stochastic processes for insurance and finance*. John Wiley & Sons, 2009.

D. Sarkar. *lattice: Multivariate data visualization with R*. Use R! Springer New York, 2008. ISBN 9780387759692. URL https://books.google.fi/books?id=gXxKFWkE9h0C.

Stan Development Team. RStan: the R interface to Stan, 2024. URL https://mc-stan.org/. R-paketin versio 2.32.5.

D. Teutonico. *ggplot2 essentials*. Packt Publishing Ltd, 2015.

S. Tikka & J. Helske. dynamite: An R package for dynamic multivariate panel models, 2023. URL https://arxiv.org/abs/2302.01607.

G. Tutz. *Regression for categorical data*, volyymi 34. Cambridge University Press, 2011.

W. N. Venables & B. D. Ripley. *Modern applied statistics with S*. Springer, New York, neljäs painos, 2002. URL https://www.stats.ox.ac.uk/pub/MASS4/. ISBN 0-387-95457-0.

H. Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28, 2010. doi: 10.1198/jcgs.2009.07098.

H. Wickham. *ggplot2: Elegant graphics for data analysis*. Use R! Springer International Publishing, 2016. ISBN 9783319242750. URL https://books.google.fi/books?id=RTMFswEACAAJ.

H. Wickham. *Advanced R*. CRC press, 2019.

H. Wickham, J. Hester, W. Chang & J. Bryan. *devtools: Tools to make developing R packages easier*, 2022. https://devtools.r-lib.org/, https://github.com/r-lib/devtools.

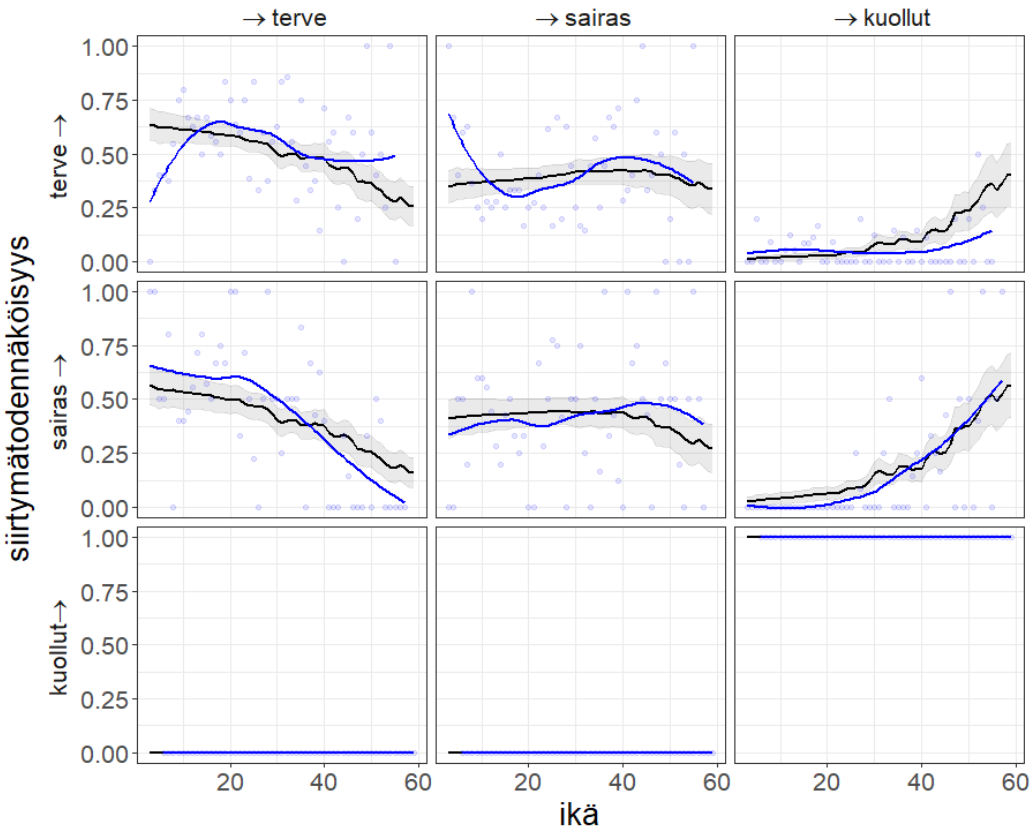
L. Wilkinson. Guides. *The Grammar of Graphics*, sivut 347–356, 2005.

M. J. Zyphur, P. D. Allison, L. Tay, M. C. Voelkle, K. J. Preacher, Z. Zhang, E. L. Hamaker, A. Shamsollahi, D. C. Pierides, P. Koval & E. Diener. From data to causes I: building a general cross-lagged panel model (GCLM). *Organizational Research Methods*, 23(4):651–687, 2020. doi: 10.1177/1094428119847278. URL https://doi.org/10.1177/1094428119847278.

Liitteet

Liite A: Todennäköisyyksien ja osuuksien piirtäminen samaan kuvaan

```
R> library(nhmgrid)
R> library(ggplot2)
R> library(nnet)

R> # Sovitetaan health-aineistoon yksinkertainen multinomi-
R> # logistinen regressiomalli funktiolla nnet::multinom.
R> fit <- multinom(state ~ lagstate + age + sex, data = health)

R> # Lasketaan siirtymäosuudet aineistosta ja
R> # estimoidaan siirtymätodennäköisyydet mallilla.
R> props <- stprops(health, id = "id", state = "state", x = "age")
R> probs <- stprobs(fit, x = "age")

R> # Käännetään tilat suomeksi.
R> list(terve = "healthy", sairas = "sick", kuollut = "deceased") ->
+     levels(props$from) -> levels(props$to) ->
+     levels(probs$from) -> levels(probs$to)

R> # Piirretään siirtymätodennäköisyydet ja -osuudet.
R> plot(probs) +
+     geom_point(data = props, color = "blue", alpha = 0.1) +
+     geom_smooth(data = props, color = "blue", se = FALSE) +
+     labs(x = "ikä", y = "siirtymätodennäköisyys")
```

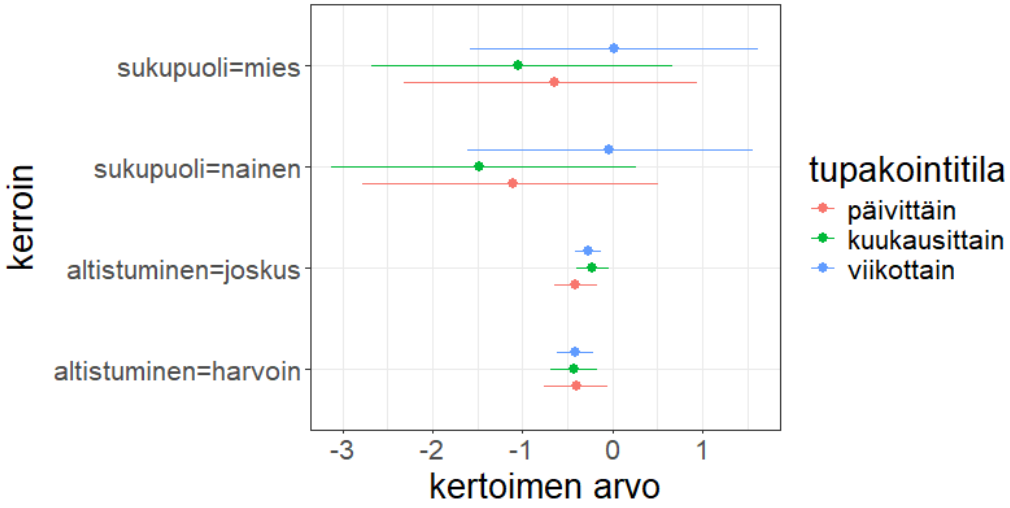


Kuvio A.1: Siirtymäkuviomatriisi `health`-aineistolle. Siirtymätodennäköisyydet (musta) piirrettynä siirtymäosuuksien (sininen) kanssa päällekkäin. Sininen käyrä on siirtymäosuuksille LOESS-menetelmällä laskettu tasoituskäyrä.

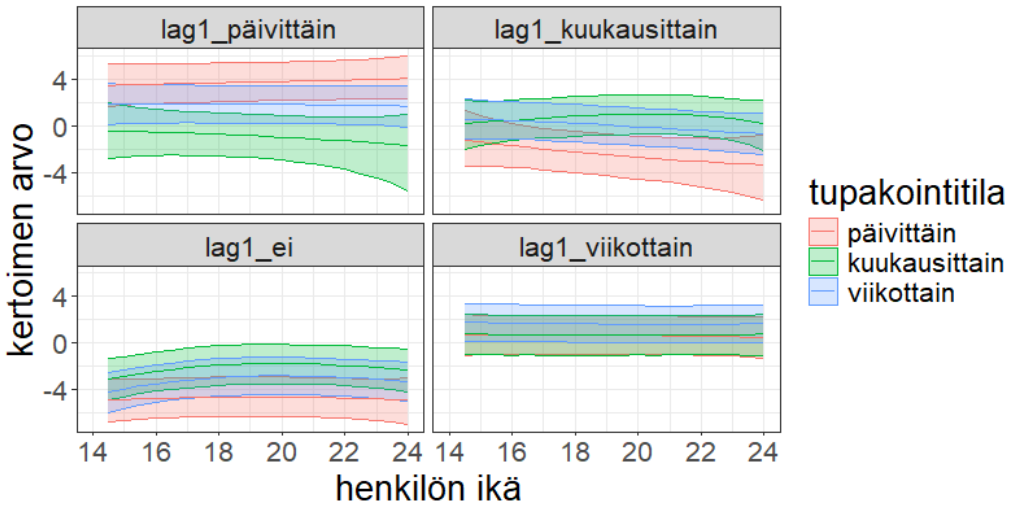
Liite B: Tupakointimallin sovittamisen ja diagnostiikan R-koodi

```
R> library(dynamite)

R> # Sovitetaan dynamite-malli.
R> f <- obs(smoking ~ -1 + sex + exposure + varying(~ -1 + lag(smoking)),
+           family = "categorical")
R> fit <- dynamite(f + splines(noncentered = TRUE),
+    data = smoking, time = "age", group = "id", backend = "cmdstanr",
+    chains = 4, parallel_chains = 4, iter_warmup = 1000,
+    iter_sampling = 1000, refresh = 100, seed = 202403)

R> # Tarkastellaan MCMC-tulosteita.
R> mcmc_diagnostics(fit)

NUTS sampler diagnostics:
2 out of 4000 iterations ended with a divergence. See Stan documentation
for details.

Smallest bulk-ESS values:
delta_smoking_smoking_lag1no[20.5]_weekly  1245
delta_smoking_smoking_lag1no[21]_weekly    1245
delta_smoking_smoking_lag1no[20]_weekly    1247

Smallest tail-ESS values:
tau_smoking_smoking_lag1monthly_daily       1511
tau_smoking_smoking_lag1monthly_monthly     1531
delta_smoking_smoking_lag1no[15.5]_monthly  1785

Largest Rhat values:
beta_smoking_exposurerarely_monthly             1
delta_smoking_smoking_lag1monthly[23]_weekly    1
delta_smoking_smoking_lag1monthly[22.5]_weekly  1
```

Liite C: Tupakointimallin regressiokertoimien kuviot

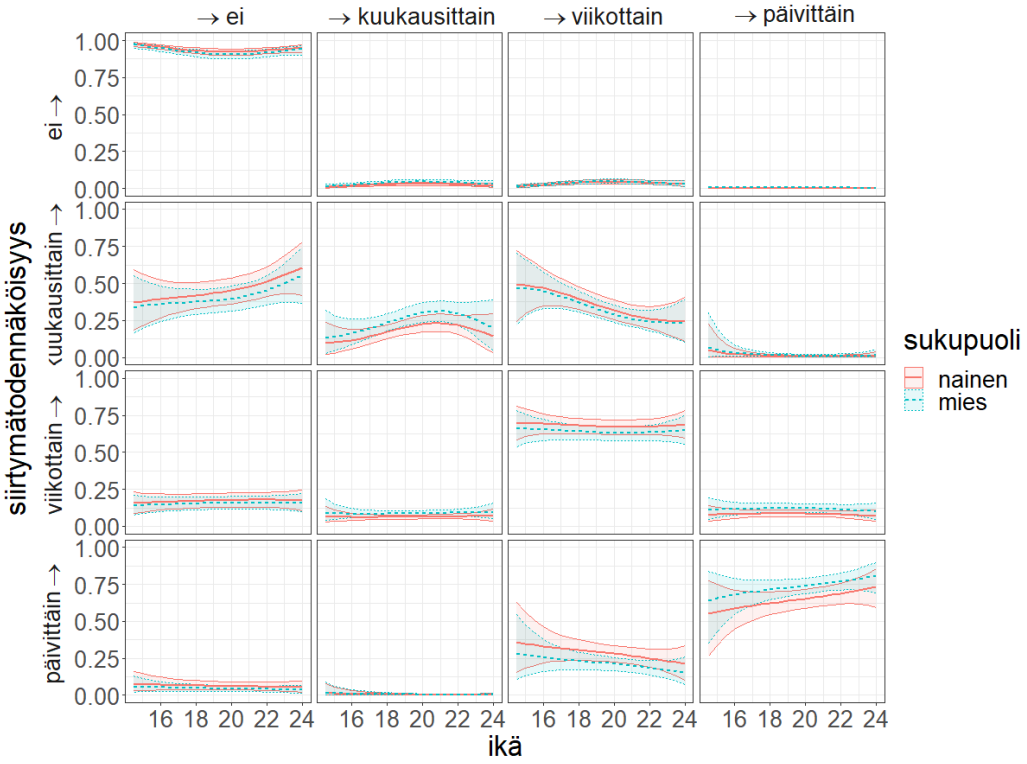


Kuvio C.1: Aikamuuttumattomien regressiokertoimien posteriorikeskiarvot (pisteet) 95 %:n posteriorivälein (viivat). Väri ilmaisee tupakointitilaa.

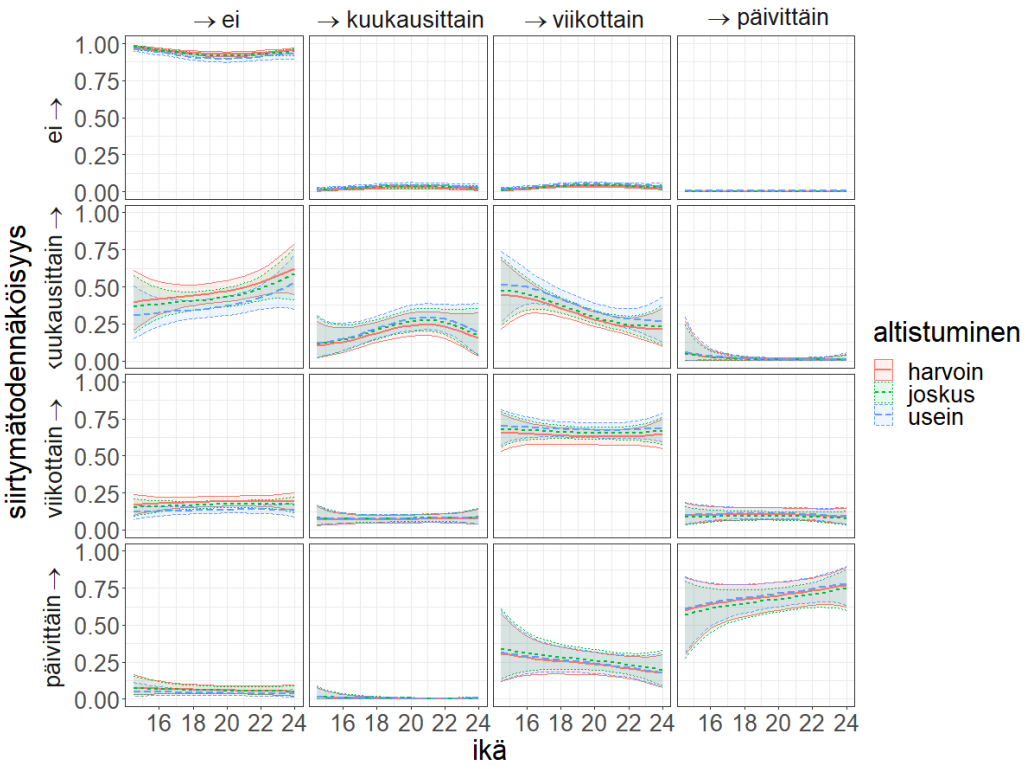


Kuvio C.2: Aikamuuttuvien regressiokertoimien posteriorikeskiarvot 95 %:n posteriorivälein. Käyrän väri ilmaisee nykyistä tupakointitilaa ja kuvioruudun otsikko kertoo edellisen tupakointitilan.

Liite D: Ryhmitellyt tupakointitilojen siirtymäkuviomatriisit



Kuvio D.1: Henkilön tupakoinnin määrää kuvaavien tilojen siirtymäkuviomatriisi. Todennäköisyydet on ryhmitelty sukupuolen mukaan.



Kuvio D.2: Henkilön tupakoinnin määrää kuvaavien tilojen siirtymäkuviomatriisi. Todennäköisyydet on ryhmitelty elokuvissa tupakoinnin näkemiselle altistumisen mukaan. Ryhmät on näkee usein (punainen), näkee joskus (vihreä) ja näkee harvoin (sininen).