hyväksymispäivä arvosana

arvostelija

Maatilapaneeliaineiston analyysi lineaarisella sekamallilla

Alina Sinisalo

Helsinki 27.8.2013

Pro gradu -tutkielma

HELSINGIN YLIOPISTO
Sosiaalitieteiden laitos

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Valtiotieteellinen tiedekunta	Sosiaalitieteiden laitos

Tekijä — Författare — Author
Alina Sinisalo

Työn nimi — Arbetets titel — Title

Maatilapaneeliaineiston analyysi lineaarisella sekamallilla

Oppiaine — Läroämne — Subject
Tilastotiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	27.8.2013	60 sivua + 23 liitesivua

Tiivistelmä — Referat — Abstract

Suomessa maatalousalalla on käynnissä voimakas rakennemuutos, jonka vaikutusta maatalouden tuotantokustannuksiin on tutkittava huomioiden pitemmän aikavälin muutokset. Maatalouden kannattavuuskirjanpitotoiminnalla on Suomessa pitkät perinteet ja tietoja kerätään vuosittain noin tuhannelta toimintaan vapaaehtoisesti liittyneeltä tilalta. Maa- ja elintarviketalouden tutkimuskeskus (MTT) kerää vuosittain maa- ja puutarhatalouden yrityskohtaisen kirjanpitoaineiston. Aineiston tuloksia sopivasti painottamalla pyritään kuvaamaan koko Suomen maatalouden kannattavuutta. Suomen tulokset julkaistaan Taloustohtori–sivuston maa- ja puutarhatalousverkkopalvelussa.

Tutkielman teoriaosassa tarkastellaan mikropaneeliaineiston ja lineaarisen sekamallin ominaisuuksia ja tutustutaan maatalouden kannattavuuskirjanpitoaineistoon. Soveltavassa osassa selvitetään tuotantokustannusten muuttumista suomalaisilla kannattavuuskirjanpitotoimintaan osallistuvilla maatiloilla aikajaksolla 2000–2011 sekä testataan lineaarisen sekamallin käytettävyyttä mallinnettaessa maatalousyrityksen tuotantokustannuksia. Tuotantokustannuksia tarkastellaan kokonaistuotantokustannuksina ja yksikkötuotantokustannuksina.

Kokonaistuotantokustannukset ovat olleet kasvussa koko 2000-luvun ajan kaikissa tuotantosuunnissa. Yksikkötuotantokustannus maitolitraa kohden on pysytellyt tarkastelujakson lähes samalla tasolla tai hieman pienentynyt. Tulosten perusteella lehmien määrän lisääntyminen tiloilla pienentää yksikkökustannusta. Tärkeimmät tuotantokustannuksia selittävät muuttujat liittyvät aikamuutokseen ja tilan suuruutta kuvaaviin tekijöihin, kuten viljelyala ja lehmien määrä. Tutkimuksessa selvitetään myös maatilojen kokoluokan ja maantieteellisen sijainnin merkitystä kustannusten selittäjänä. Tulosten perusteella tilan sijainti ei ole kovin tärkeä selittäjä kustannusten muodostumisessa ja kokoluokista eniten erottuu pienien maatilojen joukko, joka eroaa merkitsevästi keskisuurista ja suurista tiloista siten, että yksikkökustannustaso oli suurempi. Kokonaistuotantokustannukset kasvavat tilakoon kasvaessa.

Mallien toimivuustarkastelujen perusteella lineaarinen sekamalli toimii parhaiten kokonaistasolla ja keskisuurilla ja suurilla kokoluokilla. Pienien tilojen kuvaaminen lineaarisella sekamallilla on epätarkempaa. Tuotantosuunnittain katsottuna malli näyttää antavan aliarvion kokonaiskustannuksista siipikarja-, kasvihuone- ja sikatuotannossa, ja toisaalta yliarvioivan kustannukset viljanviljelyssä, muussa kasvinviljelyssä ja muussa laidunkarjatuotannossa, mutta ero ei kuitenkaan ole merkitsevä. Tilannetta voitaisiin parantaa siten, että kannattavuuskirjanpitotoimintaan pyrittäisiin rekrytoimaan lisää pieniä tiloja ja sellaisien tuotantosuuntien tiloja, joita on nyt aineistossa vähän, esimerkiksi siipikarjatuotannon ja muuta laidunkarjatuotantoa harjoittavia tiloja.

Avainsanat — Nyckelord — Keywords
paneeliaineisto, lineaarinen sekamalli, kannattavuuskirjanpitoaineisto

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

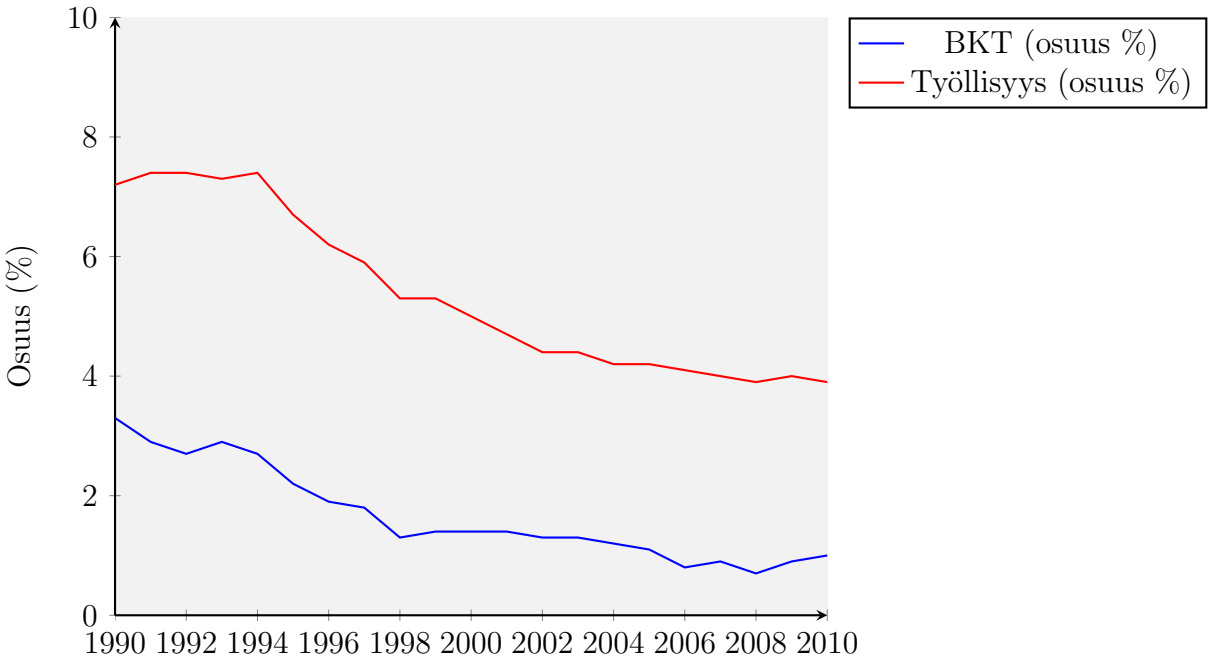
1 Johdanto

Vaikka maatalouden merkitys Suomen kansantaloudelle on ollut pitkään laskussa (kuvio 1), tuotannon bruttoarvo oli 6,1 miljardia euroa vuonna 2010, kun tuotannon tuki 2,1 miljardia euroa otetaan huomioon. Maatalous on pääomavaltainen elinkeino, koska nykyaikaiset maatalouskoneet ja -rakennukset ovat kalliita. Maataloudessa investoidaan selvästi sen suhteellista bruttokansantuoteosuutta enemmän. Vuonna 2010 maatalouden investoinnit olivat 3,3% kansantalouden kokonaisinvestoinneista. Maataloustuotteiden tuottaminen Suomessa on kallista ja tuotannosta aiheutuu merkittäviä kustannuksia, itse asiassa niin merkittäviä, että niitä ei pystytä kattamaan myynnillä. Muun muassa siitä syystä maataloustuotantoa tuetaan. [35]

Maatalousalalla on ollut ja on edelleen käynnissä voimakas rakennemuutos. Maatalousyritysten määrä on vähentynyt 1960-luvun alun 300000 tilasta nykyiseen 60000 tilaan. Suomi liittyi Euroopan Unioniin vuonna 1995, jolloin suomalaiset maataloustuottajat joutuivat kohtaamaan avoimemman kilpailutilanteen. Myös Suomen kansallinen maataloustukipolitiikka piti sopeuttaa EU:n yhteiseen maatalouspolitiikkaan. Rakennemuutoksen vaikutusta maatalouden tuotantokustannuksiin on tutkittava huomioiden pitemmän aikavälin muutokset. Tässä Pro Gradu -tutkielmassa selvitetään tuotantokustannusten muuttumista suomalaisilla kannattavuuskirjanpitotoimintaan osallituvilla maatiloilla aikajaksolla 2000–2011. Muutosta tutkitaan käyttäen kannattavuuskirjanpitoaineistoa, joka on paneeliaineisto.

Paneeliaineistoilla tarkoitetaan toistomittausaineistoja, jossa samoja havaintoyksiköitä mitataan useampaan kertaan. Paneeliaineistoista käytetään myös muita termejä: pitkittäisaineisto (englanniksi longitudinal data), toistomittausaineisto (repeated measurement data). Tärkeimpänä ominaisuutena paneeliaineistoissa pidetään, että niillä voi tehokkaasti tutkia muutosta ajan kuluessa. Paneeliaineistojen käytön etuja ja rajoitteita on käsitelty tutkielman luvussa 2.1. Tässä tutkielmassa keskitytään vain mikropaneeliaineistoihin.

Paneeliaineistojen analysoiminen on tutkimusmetodisesti haastavaa, sillä perinteiset päättelymallit, kuten lineaarinen varianssianalyysi ja kovarianssianalyysi, eivät sellaisenaan toimi. Aineistossa samaan havaintoyksikköön liittyvät mittaukset yleensä korreloivat voimakkaasti keskenään. Toistomittaukset voivat olla tehty eri ajanhetkinä. Lisäksi havaintoyksiköiden mittausten ajankohtien välit voivat vaihdella. Tässä tutkielmassa tarkastellaan paneeliaineiston analysointia lineaarisella sekamallilla, koska se tarjoaa tavan korreloivien havaintojen analysoimisessa. Lineaarinen seka-



Kuvio 1: Maatalouden ja maataloutta palvelevan toiminnan osuus Suomen kansantaloudessa 1990–2010 [47].

malli mahdollistaa sen, että analyysissa voi olla mukana myös sellaisia yrityksiä, jotka ovat voineet lopettaa toimintansa tarkastelujakson aikana, tai tarkasteluaikana on tullut mukaan uusia yrityksiä tai tila on voinut jättää toimittamatta yrityksen kirjanpitotiedot. Lineaarisen sekamallin keskeinen teoria esitetään luvussa 3. Mallia sovelletaan paneeliaineiston analysointiin luvussa 4. Paneeliaineistona tutkielmassa käytetään maatalouden kannattavuuskirjanpitoaineistoa vuosilta 2000–2011, jossa havaintoyksikkönä ovat maatalousyritykset (luku 2.2 ja 4.1).

Tutkielman tutkimuskysymykset voidaan jakaa kahteen osaan:

1. Tilastotieteelliset tutkimuskysymykset:
   - Miten lineaarinen sekamalli soveltuu tutkimusaineitoon?
   - Minkälainen malli sopii tutkimusaineistoon (mallin spesifiointi)?
   - Miten painokertoimien käyttäminen vaikuttaa malliin?
2. Soveltavat tutkimuskysymykset:
   - Miten tuotantokustannukset ovat kehittyneet 2000–2011?

- Onko kirjanpitotilojen tuotantokustannusten kehittymisessä eroja tuotantosuuntien välillä?
- Onko kirjanpitotilojen tuotantokustannusten kehittymisessä alueellisia eroja?

Tuotantokustannuksia tutkitaan maatalousyritystasolla kokonaistuotantokustannuksina (kaikki kustannuslajit yhteensä) ja yksikkötuotantokustannuksina (per tuotettu yksikkö).

Luvussa 5 pohditaan saatuja tuloksia ja tehdään yhteenveto tutkielmasta.

2 Paneeliaineistot ja niiden analysointi

Paneeliaineisto on aineisto, jossa samoista poikkileikkausyksiköistä on mitattu vastemuuttujan $\mathbf{y}$ arvo $y_{it}, i = 1, \ldots, N$, useaan kertaan, $t = 1, \ldots, T_i$. Poikkileikkausyksiköitä voivat olla esimerkiksi ihmiset, kotitaloudet, valtiot, yritykset, osakkeet ja kunnat. Eri havaintoyksiköillä mittausten ajankohta ei välttämättä ole sama, eikä kaikista havaintoyksiköistä ole välttämättä yhtä monta mittausta. Myös saman havaintoyksikön mittausvälit voivat vaihdella.

Vastemuuttujaa koskeva havaintoaineisto voidaan siten koota yhdeksi vektoriksi

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT_i} \end{pmatrix},$$

missä $i = 1, \ldots, N$. Tiiviimmin eri havaintoyksikköjen aineistot voidaan esittää kootusti matriisina pinoamalla kaikki $\mathbf{y}_i$-matriisit vertikaalisesti muodossa

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

Selittävien muuttujien, $X^{(1)}, \ldots, X^{(p)}$ arvot voidaan esittää i:nnelle yksikölle matriisimuodossa siten, että havaintoja on kerätty n_i kappaletta

$$\mathbf{X}_i = \begin{pmatrix} X_{1i}^{(1)} & X_{1i}^{(2)} & \ldots & X_{1i}^{(p)} \\ X_{2i}^{(1)} & X_{2i}^{(2)} & \ldots & X_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i i}^{(1)} & X_{n_i i}^{(2)} & \ldots & X_{n_i i}^{(p)} \end{pmatrix}.$$

Kootusti kaikkien yksiköiden havainnot voidaan esittää pinoamalla kaikki $\mathbf{X}_i$-matriisit vertikaalisesti muodossa

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}.$$

Paneeliaineistot voivat olla keruutavaltaan joko prospektiivisia tai retrospektiivisiä. Prospektiivisten aineistojen osalta voidaan, riippuen yhteydestä, käyttää nimitystä *paneeliaineisto*, *seurantatutkimus*, *kliininen koe* tai *kohorttitutkimus*. Retrospektiivisten aineistojen (*tapaus-verrokkitutkimus*) osalta havaintoyksiköiden tapaus- ja taustahistorian selvittäminen jälkikäteen voi olla vaikeaa ja epäluotettavaa, ellei käytössä ole luotettavia rekistereitä. [39]

2.1 Paneeliaineiston edut ja rajoitteet

Tärkein paneeliaineistojen käytöön liittyvä etu on niiden tehokkuus muutoksen tutkimisessa; miten voidaan analysoida yksilöiden välisiä eroja, kun yksilön ominaisuudet muuttuvat aikajaksolla. [12, 44]

Perinteisesti paneeliaineistojen kohdalla muutoksen analysoimisessa on käytetty lineaarisia malleja, kuten lineaarista varianssianalyysiä ja kovarianssianalyysiä. Näillä menetelmillä saadaan kuitenkin estimoitua malli tarkasti vain silloin, kun havaintojoukko koostuu yhtäsuurista ryhmistä ja toistomittauksista. Koska lineaaristen mallien käyttöä on kritisoitu mallioletusten rikkomisesta, on otettu käyttöön vaihtoehtona lineaariset sekamallit, joilla voidaan tutkia muutosta ajan kuluessa. [44]

Paneeliaineistoja analysoitaessa on otettava huomioon useita seikkoja: paneeliaineiston edut ja rajoitteet, soveltuvat analysointimenetelmät jne. Etuihin kuuluvat muun muassa seuraavat asiat [19, 21] ref. [3]:

(1) Paneeliaineisto antaa *informatiivisempaa tietoa, enemmän vaihtelua, vähemmän kollineaarisuutta muuttujien välillä, enemmän vapausasteita ja lisää tehokkuutta.* Aikasarja-analyysien ongelmana on usein multikollineaarisuus. Tämä on epätodennäköisempää paneeliaineistossa, jossa poikkileikkauskomponentit lisäävät vaihtelua. Aineiston vaihtelu voidaan jakaa esimerkiksi alueiden väliseen vaihteluun (variation between) ja alueiden sisäiseen vaihteluun (variation within), missä ensiksi mainittu vaihtelu on yleensä suurempaa.

(2) Paneeliaineistojen avulla pystytään paremmin analysoimaan *erilaisia sopeutumismekanismeja*, kuten työttömyyden keston muutoksia, yksilöiden tulojen muutos-

ta ja poliittisia muutoksia. Esimerkiksi työttömyyttä mitattaessa, poikkileikkaustiedoilla voidaan arvioida, mikä on työttömien osuus ja osuus muuttuu ajan kuluessa. Paneeliaineistolla voidaan lisäksi arvioida kuinka moni työttömänä yhdellä jaksolla olleista on työttömänä myös toisella jaksolla. Sen avulla voidaan lisäksi tarkkailla onko esimerkiksi köyhyys ohimenevä vai pitkäaikainen ilmiö. Paneeliaineistojen avulla voidaan analysoida ennen-jälkeen-tilanteita tai kahden aikavälin tilanteita (katso [7]).

(3) Paneeliaineistojen avulla voidaan *paremmin tunnistaa ja mitata vaikutuksia*, joita poikkileikkaus- ja aikasarja-aineistoilla ei voida tunnistaa. Esimerkiksi nostaako vai laskeeko ammattiliiton jäsenyys palkkoja? Tähän kysymykseen voidaan vastata tutkimalla työntekijää, joka kuuluu ammattiliittoon ja työntekijää, joka ei kuulu ammattiliitoon työpaikalla. Pitämällä yksilön ominaisuudet vakiona, pystytään selvittämään ammattiliiton jäsenyyden vaikutusta palkkaan ja sen suuruutta. Analyysi perustuu palkkaerojen arviointiin, kun yksilöiden ominaisuudet pidetään vakiona.

(4) Paneeliaineistoihin perustuvien mallien avulla voidaan tehdä ja testata *monimutkaisempia käyttäytymismalleja* kuin aikasarja- tai poikkileikkausaineistoilla pystytään tekemään. Esimerkiksi teknistä tehokkuutta voidaan tutkia käyttäen paneeliaineistoa (katso [4, 8, 24, 5, 22]).

(5) *Mikrotasolla kerätyt* yksilöiden, yritysten ja kotitalouksien *tiedot voivat olla tarkemmin mitattuna* kuin vastaavat muuttujat makrotasolla mitattuna. Vääristymät johtuvat aineistojen aggregoinnista johtuvasta harhasta.

(6) Toisaalta *mikropaneeliaineistoissa* on usein *lyhyempiä aikasarjoja* kuin makrotasolla mitatuissa aineistoissa.

Paneeliaineistoon käyttöön liittyy myös rajoituksia [19, 21] ref. [3]:

(1) *Suunnitteluun ja tiedonkeruuseen liittyvät ongelmat.* Mikropaneeliaineistojen kerääminen voi olla hankalaa ja kallista, koska havaintoja täytyy saada usealta aikajaksolta. Lisäksi aineiston kattavuus, vastaamattomuus, haastattelun väli, vastaajan muistamattomuus, tietämättömyys tai haastattelijasta aiheutuvat virheet voivat olla ongelmallisia.

(2) *Mittausvirheet.* Mittausvirheitä saattaa syntyä, kun kysymys vastaajalle on epäselvä, kysymykseen vastataan vääriin tai vastauksia tahallaan vääristellään, vastaus on sopimaton ja haastattelija vaikuttaa vastaajan antamiin vastauksiin. Lisäksi mittausvirhe ei ole vakio ajassa, vaan se voi vaihdella.

(3) *Valikoitumisharha.* Näitä ovat muun muassa: (a) yksilöitä voi jäädä valikoitumat-

ta aineistoon, jos osa tiedoista puuttuu, koska yhteen tai useampaan kysymykseen voi jäädä vastaus saamatta; (b) yksilön päätös kieltäytyä osallistumasta, yksilön tavoittamattomuus tai muu syy jäädä tutkimuksen ulkopuolelle aiheuttaa aineistossa katoa; (c) poistuma johtuu siitä, että yksilöt voivat kuolla, muuttaa tai kokea vastaamisen raskaaksi; yritykset voivat mennä konkurssiin tai yhdistyä toisen yrityksen kanssa. Poistuman aste vaihtelee paneelitutkimuksittain. Poistumisen vaikutusta voidaan pienentää siten, että kiinteä prosenttiosuus vastaajista korvataan jokaisella tutkimuskerralla.

(4) *Lyhyt tarkastelujakso.* Tyypillisesti mikropaneelit kattavat lyhyen aikavälin tietoja yksilöistä, jolloin yksilöistä ei saada monta havaintoa tutkimukseen. Toisaalta tarkastelujakson pidentäminen ja mittausten lisääminen luo lisäkustannuksia tai saattaa lisätä yksilöiden poistumaa tai valikoitumisharhaa.

Paneeliaineistot eivät ratkaise kaikkia ongelmia, joita aikasarja- tai poikkileikkaustutkimuksissa ei voida käsitellä. Paneeliaineiston keruu on melko kallista ja kysymys liittyy aina siihen kuinka monta mittausta havaintoyksiköistä pitäisi tehdä. Talouselämän muutokset tapahtuvat melko hitaasti ja siten muutoksiin vuodesta seuraavaan liittyy liikaa satunnaisvaihetelua ja aikajänne on liian lyhyt, jotta tiedonkeruusta olisi hyötyä. Paneeliaineiston hyödyt tulevat esiin pitkän aikavälin (5–10 vuotta tai enemmän) aineistoissa [9].

2.2 Maatalouden kannattavuuskirjanpitoaineisto

Suomessa on pitkät perinteet maatalouskirjanpidossa, sillä se käynnistyi jo vuonna 1912 [1]. Maa- ja elintarviketalouden tutkimuskeskus (MTT) kerää vuosittain maa- ja puutarhatalouden yrityskohtaisen kirjanpitoaineiston. Kannattavuuskirjanpitoaineistosta toimitetaan tilakohtaiset aineistot EU:n komission ylläpitämään maatalouden kirjanpidon tietoverkostoon (FADN, Farm Accountancy Data Network), jonka perustana on EU:n lainsäädäntö (79/65/EEC) vuodelta 1965 ja jonka tarkoituksena on kerätä puolueetonta ja tarkoituksenmukaista tilakohtaista tietoa eri maatila- ja puutarhayritysten tuloista ja taloudellisesta toiminnasta. Lainsäädäntöä on esitelty EUR-Lex–verkkosivustolla[1]. Kokoamalla EU-maiden FADN-aineistot yhteen voidaan mitata maatilojen kannattavuutta koko EU:ssa ja arvioida yhteistä maatalouspolitiikkaa.

[1] http://eur-lex.europa.eu/en/legis/latest/chap0330.htm

Tiedonkeruu

Kirjanpitoaineiston tiedot perustuvat tiloilla pidettävään kirjanpitoon. Kirjanpitotoiminta vaatii tiloilta tavanomaista suurempaa panostusta seurantaan, mikä edellyttää viljelijöiltä pitkäjänteisyyttä. Kirjanpito tehdään tarkasti määritellyille tileille yrityksen toiminnot jaoteltuna varsinaiseen maatalouteen, puutarhatalouteen, metsätalouteen, porotalouteen sekä muuhun yritystoimintaan. Yksityistalouteen liittyviä asioita ei kerätä. Yritystoimintaa koskevien tietojen perusteella lasketaan maatilan eri toimialojen tilinpäätökset sekä taloudellista asemaa ja kehitystä kuvaavat tunnusluvut.

Maaseutukeskukset (Proagria) ja Kauppapuutarhaliitto keräävät kannattavuuskirjanpitotilojen tiedot ja tallentavat ne MTT:n tietojärjestelmään. Tallennuksen jälkeen keräävät tahot tarkastavat tiedot MTT:n tarkastusjärjestelmällä. Järjestelmässä on 6000 testiä, jotka tarkastavat monipuolisesti yritysten talouteen ja myös fyysiseen tuotantoprosessiin liittyvät tiedot tilakohtaisesti. Esimerkiksi satotasot, tuotokset ja tilisaldot tarkastetaan olevan tiettyjen vaihtelurajojen sisällä. Lisäksi seurataan testeillä onko tapahtunut poikkeavia muutoksia esimerkiksi kannattavuudessa, tehdyissä työtunneissa ja kustannuksissa. Muutosten seuranta ei tosin ole mahdollista, jos yritys on ensimmäistä kertaa mukana kannattavuuskirjanpitotoiminnassa. Tämän jälkeen tiedonkerääjät toimittavat tiedot MTT:een, jonka yritysanalytiikan tiimi käy ne läpi. Tietojen ja niistä laskettavien tulosten luotettavuuden takaamiseksi tiedot tarkastetaan vielä EU:n Komission Internet-pohjaisella tarkastusjärjestelmällä. Jokaisen yrityksen tiedoille tehdään tuhansia yksittäisiä testejä. [32, 31]

Kirjanpitotoimintaan osallistuvat yritykset saavat vuosittain palauteraportin, joka sisältää tulos- ja taselaskelmat, tarkemman tuotto- ja kustannuserittelyn maa- ja puutarhataloudesta, maksuvalmiuslaskelman sekä talouden tunnuslukuraportin ja yhteenvedot varasto-, työnmenekki-, pellonkäyttö- ja satotiedoista. Lisäksi yritykset saavat vertailuraportin, joka sisältää yrityksestä sekä tarkasteltavan että edellisen tilikauden tilinpäätöksen ja tunnusluvut mukaan lukien vastaavat keskiarvotiedot vertailuryhmästä, jonka tulokset ovat samalla alueella vastaavaa tuotantoa harjoittavien samankokoisten kannattavuuskirjanpitoyritysten painotettuja keskiarvoja. Vertailuraportin avulla yrittäjät voivat tarkastella yrityksen sijoittumista vertailujoukossa. [32]

Ennen kuin tiedoille voidaan tehdä tarkastuksia täytyy yrityskohtaiset tiedot luokitella taloudelliseen koon ja tuotantosuunnan mukaan. Tietojen sijoittamisessa oikeaan luokkaan on oltava erityisen tarkkana silloin, kun tarkasteltava yritys on kasva-

nut merkittävästi vuoden aikana, koska silloin on mahdollista, jos valinnan aikaan ja vuoden tilinpäätöksen hetkellä yritys kuuluisikin eri kokoluokkaan, että tila sijoitettaisiin väärään luokkaan. Seuraavaksi tiedoille tehdään koherenssitestit, joilla pyritään havaitsemaan koodausvirheet, konversiovirheet, puuttuvista tiedoista aiheutuvat virheet ja epätavalliset arvot (outliers), minkä jälkeen tiedoille tehdään homogeenisyystestit. Jatkuvuustesteillä (continuity tests) voidaan havaita poikkeamat ennustettujen arvojen ja havaittujen arvojen välillä. Mikäli poikkeama ylittää asetetun kynnysrajan, yrityksen tiedot otetaan erityistarkasteluun, jos voidaan havaita jokin looginen selitys ilmeisen poikkeaville havainnoille. [14]

Tarkastusten jälkeen tiedot kootaan yhteen tietokantaan. MTT julkaisee tulokset alueittain, kokoluokittain ja tuotantosuunnittain ryhmäkeskiarvoina, joiden laskemisessa on käytetty kalibroituja painokertoimia. Suomen tulokset julkaistaan Taloustohtori–sivuston maa- ja puutarhatalousverkkopalvelussa[2] ja EU:n jäsenmaiden tulokset julkaistaan Euroopan komission maatalouden ja maaseudun kehittämisen pääosaston verkkosivustolla[3].

Otanta

Yritysten valintamenettelytapa perustuu jokaisessa EU:n jäsenvaltiossa kunkin maan omaan otantasuunnitelmaan. Aineistojen yhtenäisyyden vuoksi FADN-hallinto tarkastaa kansalliset otantasuunnitelmat. Valinnan tavoitteena on saada mukaan tiloja, jotka edustavat eri maantieteellisiä alueita, taloudellisia kokoluokkia ja tuotantosuuntia. Yritysten osallistuminen kannattavuuskirjanpitotoimintaan on vapaaehtoista ja siitä syystä rekrytointia ei voi toteuttaa täysin satunnaisesti, koska mahdollisesti kirjanpitotoiminnan kannalta sopiva yritys ei välttämättä ole halukas osallistumaan.

Otoksen yritykset poimitaan maatalouden rakennetutkimuksessa mukana olleesta tilajoukosta, joka muodostaa tutkimuksen perusjoukon. Rakennetutkimus suoritetaan kymmenen vuoden välein maatalouslaskentana eli kohdejoukon muodostavat kaikki Suomen maatilat ja puutarhayritykset (vastausaste 2010 oli 95%). Maatalouslaskentojen välillä tehdään kaksi tai kolme otostutkimusta. Edellinen rakennetutkimus tehtiin vuonna 2010 ja seuraavat maatalouden rakennetutkimukset tehdään vuosina 2013 ja 2016. Rakennetutkimuksen toteutuksesta vastaa maa- ja metsätalousministeriön tietopalvelukeskus (Tike). [51]

[2] http://www.mtt.fi/taloustohtori/
[3] http://ec.europa.eu/agriculture/rica/

Suomessa maatalousyritysten tilakoko määritetään taloudellisen koon mukaan perustuen standardituotuossummaan. Kannattavuuskirjanpidossa seurataan sellaisia tiloja, joiden standardituotossumma on yli 8000 euroa. Tilojen tuotantosuunta ja kokoluokka perustuu myös standardituotoksiin (Standard Output, SO), millä tarkoitetaan kullekin viljelykasville tai tuotantoeläimelle alueittain laskettua viiden vuoden tietoihin perustuvaa tuotosta, joka vastaa joko yhdestä hehtaarista tai eläimestä saatavaa tuotosta ilman tukia. Kunkin tilan viljelykasvien viljelyala ja tuotantoeläimien lukumäärä kerrotaan tuotekohtaisilla standardituotoksilla ja summataan kokonaisstandardituotokseksi. Jos jokin tuote muodostaa enemmän kuin 2/3-osaa kokonaisstandardituotoksesta, kuuluu tila tuotetta vastaavaan tuotantosuuntaan. Jos tilalla ei ole tällaista tuotetta, katsotaan tilan harjoittavan sekamuotoista tuotantoa. Sekatilojen lisäksi tuotantosuuntia ovat viljanviljely, muu kasvinviljely, kasvihuonetuotanto, avomaatuotanto, lypsykarjatalous, muu nautakarjatalous, muu laidunkarjatalous, sikatalous ja siipikarjatalous. Tuotantosuuntiin sisältyy lisäksi erilaisia tuotantohaaroja. Aluejakona käytetään pääsääntöisesti maataloustukialueisiin perustuvaa jakoa (A, B, C1, C2, C2P, C3 ja C4), mutta mahdollista on käyttää myös LFA-alueita, maaseutukeskusalueita, FADN-alueita, suuralueita ja maaseututyyppialueita. [32]

Kun otosjoukko jaetaan neljään suuralueeseen ja vielä kansallisen tason tuotantosuuntiin, syntyy useita osajoukkoja. Esimerkiksi vuoden 2013 otantasuunnitelman mukaan yritykset jaettiin kymmeneen tuotantosuuntaan neljällä eri suuralueella. Kun Suomessa on käytössä 11 kokoluokkaa, jaetaan havaintojoukko $4 \times 10 \times 11 = 440$ osajoukkoon. Koska yritysten määrä tietyissä osajoukoissa jää liian pieneksi, joudutaan käytännössä kokoluokkia yhdistelemään. Yhden osajoukon minimikokona pidetään viittä yritystä. Kokoluokkien yhdistely vaihtelee vuodesta toiseen ja alueesta toiseen, koska tilojen määrä muuttuu vuosien aikana alan yleisen rakennekehityksen seurauksena ja toisaalta Pohjois-Suomessa on muuta maata kokoluokaltaan pienempiä tiloja.

Otantamenetelmä

Otoksen muodostamiseen käytettävä menetelmä perustuu sekä suhteelliseen että optimaaliseen kiintiöintiin. Koska aineiston keräyskustannukset ovat jokaisessa ositteessa yhtä suuret, optimaalisessa kiintiöinnissä on kyseessä Neyman-kiintiöinti [33]. Ositteiden lopullisena otoskokona käytetään näiden molempien menetelmien otoskokojen keskiarvoa. Yhden osittеen otoskoon minimikoko on viisi tilaa.

Suhteellinen kiintiöinti tarkoittaa sitä, että kustakin ositteesta poimitaan yhtä suuri osuus (%) tiloista. Suhteellisessa kiintiöinnissä otetaan huomioon ositteen koko. Suuremmasta ositteesta poimitaan otokseen enemmän tiloja kuin pienemmästä.

Optimaalisessa kiintiöinnissä otoskoko painotetaan tilojen lukumäärän lisäksi kiintiömuuttujan keskihajonnalla. Optimaalisessa kiintiöinnissä ositteen otoskooksi tulee sitä suurempi, mitä suurempia ovat ositteen koko ja hajonta.

Suhteellisen kiintiöinti lasketaan yhtälöllä

$$n_h = n\frac{N_h}{N}$$

ja optimaalinen (Neyman-) kiintiöinti yhtälöllä

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{k} N_h S_h},$$

joissa h on ositteen numero, k on ositteiden lukumäärä, N_h on yritysten lukumäärä ositteessa h, N on tilojen lukumäärä yhteensä, n_h on otoskoko ositteessa h, n on otoskoko yhteensä ja S_h on hajonta (AWU) ositteessa h.

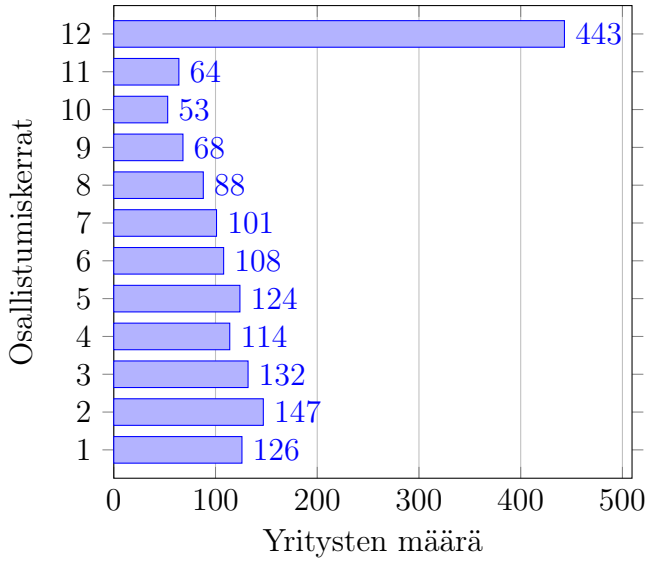
Lopullinen otos määritetään suhteellisen ja optimaalisen kiintiöinnin keskiarvona, jolloin pieniin tilakokoluokkiin tulee tämän hetkisen tilarakenteen kannalta riittävästi tiloja sekä myös keskikokoiset että suuremmat tilakokoluokat saavat riittävän edustavuuden.

Puutarhatilojen osalta otanta kiintiöidään erikseen, koska puutarhatilat ovat keskenään erittäin heterogeenisiä, ja jotta puutarhatilaotoksen suhteellinen osuus kaikista puutarhatiloista ei kasvaisi liian suureksi verrattuna muihin tilaryhmiin.

Tästä syystä ositteisiin poimitaan satunnaisesti kahdeksankertainen määrä sopivia tiloja, joiden joukosta sopivuuden mukaan järjestyksessä kysytään yrityksiltä, jos ne haluavat osallistua kirjanpitoon. Kirjanpitotoimintaan osallistumisesta on yrityksille hyötyä, koska se ohjaa suunnitelmalliseen toimintaan ja lisäksi tilat saavat vuosittain kootun raportin yrityksen taloudellisista tunnusluvuista ja tietoa kannattavuudesta suhteessa muihin vastaavan kokoluokan tiloihin ja alaan yleensä.

Vuosina 2000–2011 on ollut mukana yhteensä 1568 yritystä, joista 443 yritystä on ollut mukana koko ajanjakson (kuvio 2). Keskimäärin mukana joka vuosi on ollut 926 maa- ja puutarhataloutta harjoittavaa yritystä.

Uudet kirjanpitotilat valitaan perustuen Tiken laatiman otantasuunnitelman mu-

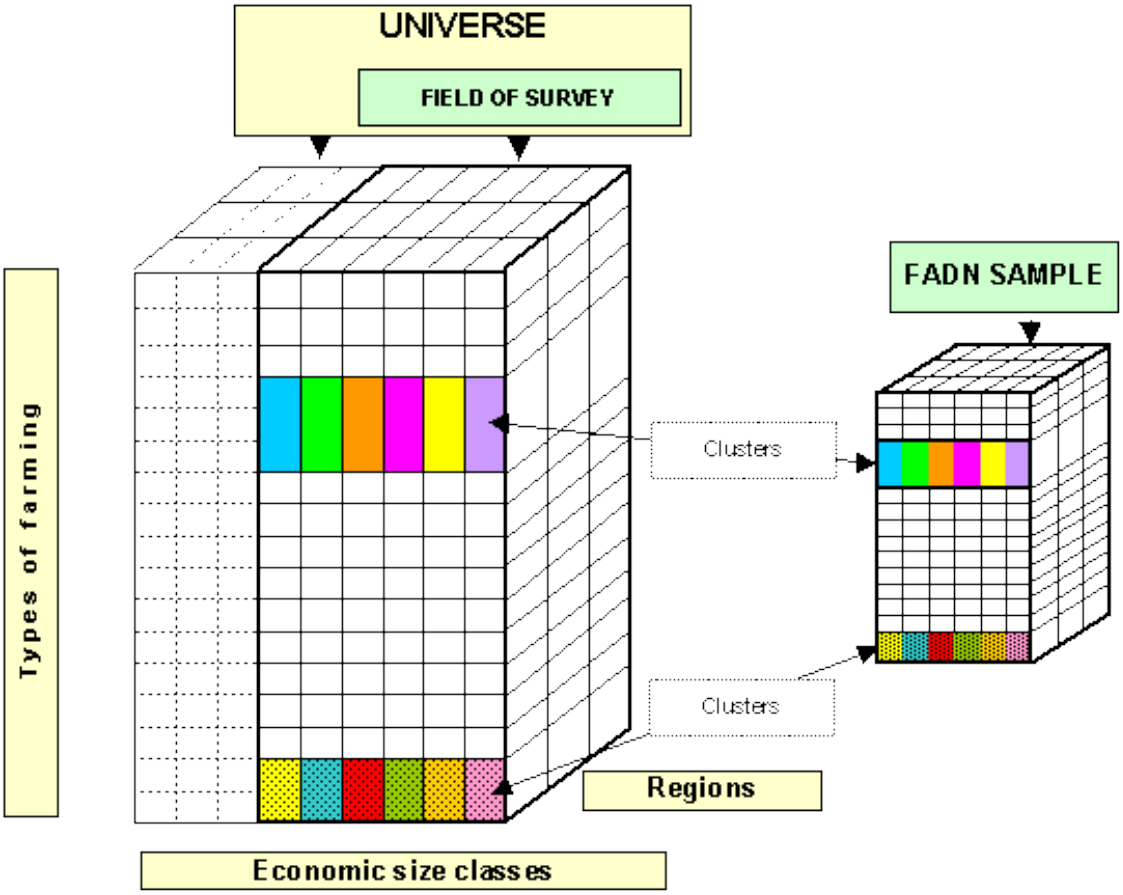


Kuvio 2: Yritysten osallistumiskerrat kannattavuuskirjanpitotoimintaan 2000–2011.

kaiseen otantaan. Käytännössä tämä tarkoittaa sitä, että päätoimisista maa- tai puutarhatalousyrityksistä valitaan satunnaisesti tilajoukko, joka edustaa sellaisia alueita, tuotantosuuntia ja tilakokoja, joista kirjanpidossa on mukana tavoitetta vähemmän tiloja. Tästä joukosta rekrytoidaan halukkaat tilat mukaan kirjanpitotoimintaan. Kiinnostuneet tilat voivat myös itse ilmaista halukkuutensa liittyä mukaan kannattavuuskirjanpitoon. Proagria-keskukset, Kauppapuutarhaliitto hoitavat tilojen rekrytoinnin ja neuvovat kirjanpidon aloittamisessa. [32]

Kirjanpitotoimintaan osallistuvat maatalousyritykset ovat keskimääräistä suurempia ja mahdollisesti kannattavampia. Sopivasti painottamalla voidaan kirjanpitotilojen tulosten perusteella laskea suomalaisten maatilojen taloustulokset. Painotus perustuu maa- ja puutarhatalousyritysten lukumäärään eri tukialueilla, tuotantosuunnissa ja tilakokoluokissa. Tiedot tilojen lukumääristä saadaan maatila- ja puutarharekisteristä. Painot lasketaan jokaiselle vuodelle erikseen. [32]

Osajoukkojen aggregointi voidaan havainnollistaa kolmeulotteisena matriisina (kuvio 3), jonka dimensio määräytyy tuotantosuunnan, taloudellisen kokoluokan ja alueen mukaan. Painokerroin lasketaan yksittäiselle maatilalle seuraavasti: ajatellaan, että taloudelliselta kooltaan suuri viljanviljelytila sijaitsee Etelä-Suomessa ja ryhmään kuuluu FADN-otoksessa 10 maatilaa ja maatilarekisterissä ryhmän koko populaatio on 100, saa jokainen yksittäinen kyseisen ryhmän maatila FADN-otoksessa painokertoimen $100/10 = 10$.



Kuvio 3: Painokertoimien laskennan perusteet FADN-järjestelmässä [15].

3 Lineaariset sekamallit

Sekamalleista (mixed models) käytetään erilaisia nimityksiä. Joskus käytetään nimitystä toistomittausmalli (repeated-measures model), joskus nimitystä hierarkinen malli (hierarchical model) tai monitasomalli (multi-level model). Sekamallit voidaan jakaa kolmeen eri ryhmään: lineaarisiin, yleistettyihin lineaarisiin ja epälineaarisiin sekamalleihin [11]. Lineaariset sekamallit ovat tilastollisia malleja jatkuville selitettäville muuttujille, missä residuaalit ovat normaalisti jakautuneita, mutta eivät välttämättä riippumattomia tai niillä ei ole vakiosuuruinen varianssi [53]. Tässä tutkielmassa käsitellään lineaarisia sekamalleja (LSM).

Tutkimuksissa, joissa analysoidaan ryväsaineistoja, pitkittäisaineistoja tai paneeliaineistoja, voidaan hyödyntää LSM:ja. Tällaisia aineistoja hyödynnetään lääketieteen, biologian, fysiikan, sosiaalitieteen ja taloustieteen tutkimuksessa [53].

3.1 Lineaaristen sekamallien yleinen muoto

Eräs yleisin ja eniten käytetyistä tilastollisista malleista on lineaarinen malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

jossa vektorissa $\mathbf{y}$ esitetään havaintotiedot, $\boldsymbol{\beta}$ on tuntematon kiinteiden vaikutusten parametrit sisältävä vektori, $\mathbf{X}$ on selittävien muuttujien matriisi ja $\boldsymbol{\epsilon}$ on jäännöstermi. Lineaarisessa mallissa tavoitteena on kuvata keskimäärin $\mathbf{y}$ käyttäen kiinteiden vaikutusten parametreja $\boldsymbol{\beta}$.

Malli voi kuitenkin sisältää myös satunnaisvaikutuksia. Lineaarinen satunnaisvaikutusmalli voidaan ilmaista yhtälönä

$$\mathbf{y} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{2}$$

jossa $\mathbf{b}$ on tuntematon satunnaisvaikutusten parametrit sisältävä vektori, $\mathbf{Z}$ on selittävien muuttujien matriisi.

Kun kiinteät vaikutukset ja satunnaisvaikutukset yhdistetään samaan malliin, puhutaan sekamallista. Kiinteiden vaikutusten ja satunnaisvaikutusten mallia voidaan pitää sekamallin erikoistapauksina. Lineaaristen sekamallien alkulähteenä pidetään julkaisua [17].

LSM voidaan esittää havaintoyksikölle i muodossa [25]:

$$\begin{cases} \mathbf{y}_i = \underbrace{\mathbf{X}_i\boldsymbol{\beta}}_{\text{kiinteä osa}} + \underbrace{\mathbf{Z}_i\mathbf{b}_i}_{\text{satunnaisosa}} + \underbrace{\boldsymbol{\epsilon}_i}_{\text{satunnaisosa}}, \\ \mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}), \\ \boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i), \\ \mathbf{b}_1, \ldots, \mathbf{b}_n, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n \text{ riippumattomia}, \end{cases} \tag{3}$$

jossa $\mathbf{y}_i$ on i:nnen yksikön $n_i \times 1$-vastevektori, $\mathbf{X}_i$ on kiinteän osan selittävien muuttujien $n_i \times p$-koematriisi, jonka ensimmäinen sarake koostuu ykkösistä (vakiotermit), $\boldsymbol{\beta}$ on kiinteiden vaikutusten regressiokertoimien $p \times 1$-vektori, $\mathbf{Z}_i$ on satunnaisosan selittävien muuttujien $n_i \times q$-koematriisi, jonka ensimmäinen sarake koostuu ykkösistä (vakiotermit), $\mathbf{b}_i$ on satunnaisvaikutusten regressiokertoimien $q \times 1$-vektori, n_i on aineiston yksiköiden havaintojen määrä, joka voi vaihdella yksiköiden välillä, p on mallin kiinteiden parametrien lukumäärä, q on mallin satunnaisten parametrien lukumäärä ja $\boldsymbol{\epsilon}_i$ on $n_i \times 1$-jäännösvektori.

Yhtälön (3) oletusten ollessa voimassa odotusarvo ja kovarianssi ovat:

$$\begin{aligned} \mathrm{E}(\mathbf{y}_i) &= \mathbf{X}_i\boldsymbol{\beta} \\ \\ \mathbf{V}_i &= \mathrm{Cov}(\mathbf{y}_i) \\ &= \mathrm{Cov}(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}) \\ &= \mathrm{Cov}(\mathbf{Z}_i\mathbf{b}_i) + \mathrm{Cov}(\boldsymbol{\epsilon}) \\ &= \mathbf{Z}_i\mathrm{Cov}(\mathbf{b}_i)\mathbf{Z}_i' + \mathbf{R}_i \\ &= \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i. \end{aligned} \tag{4}$$

Edellisestä seuraa, että

$$\mathbf{y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i). \tag{5}$$

Jos N kpl yksittäisiä vektoreita, jotka ovat muotoa (3), kootaan matriisimuotoon siten, että

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_N \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_N \end{pmatrix}$$

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \vdots \\ \mathbf{b}_N \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_N \end{pmatrix},$$

voidaan yhtälö (3) kirjoittaa tiivistettynä muodossa [11, 27]:

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \\ \mathbf{b} \sim N(\mathbf{0}, \mathbf{G}), \\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R}), \\ \mathrm{Cov}[\mathbf{b}, \boldsymbol{\epsilon}] = \mathbf{0}. \end{cases} \tag{6}$$

Yhden yhtälön mallissa (6) täytyy kuitenkin pitää mielessä, että havainnot yksiköiden i välillä ovat riippumattomia [11]. Mallin (6) parametrit ovat kiinteiden vaikutusten vektori $\boldsymbol{\beta}$ ja kaikki tuntemattomat kovarianssimatriiseissa $\mathbf{G}$ ja $\mathbf{R}$. Satunnaisvaikutukset $\mathbf{b}$ eivät ole parametreja, koska ne eivät ole vakioita. Matriisit $\mathbf{X}$ ja $\mathbf{Z}$ voivat sisältää joko jatkuvia tai luokka-asteikollisia muuttujia, mutta vektorissa $\mathbf{b}$ on vain satunnaismuuttujia. Matriisien $\mathbf{G}$ ja $\mathbf{R}$ kaikki tuntemattomat arvot voidaan esittää kootusti vektorissa $\boldsymbol{\Theta}$. [27]

Vastaavasti kuin yhtälössä (5), voidaan esittää koottu malli (6) muodossa:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{D}\mathbf{Z}' + \mathbf{R}). \tag{7}$$

Kiinteät vaikutukset, eli regressiokertoimet, kuvaavat selitettävän ja selittävien muuttujien välisiä suhteita. Kiinteät vaikutukset ovat tuntemattomia lukuarvoja, jotka estimoidaan kerätyn aineiston perusteella. Ne koskevat LSM:ssa joko koko populaatiota (esim. koko väestö) tai sen osajoukkoa (naiset) ja vaikuttavat vain vastemuuttujan keskiarvoon. Käytännössä vaikutus on kiinteä, jos aineistossa on joko vähän tai paljon populaation eri tasoja ja vaikutus käyttäytyy systemaattisesti. [10].

Vaikutus on satunnainen, jos aineistossa on suuri määrä eri populaation tasoja ja vaikutus käyttäytyy satunnaisesti. Satunnaisvaikutukset ovat satunnaisia lukuarvoja, jotka edustavat satunnaisia eroavaisuuksia kiinteillä vaikutuksilla kuvatuista selitettävän ja selitettävien muuttujien välisistä suhteista. Satunnaisvaikutukset esiinty-

vät hierarkisissa tutkimuksissa (esimerkiksi koulu/luokka/oppilas) tai pitkittäistutkimuksissa. Satunnaisvaikutukset koskevat eroavaisuuksia yksikkötasolla suhteessa koko populaatioon. Satunnaisvaikutukset vaikuttavat vain vastemuuttujan varianssiin. [10, 53]

3.2 Kovarianssirakenteet

Satunnaisvaikutukset ovat satunnaismuuttujia. Oletetaan, että q satunnaisvaikutukset vektorissa $\mathbf{b}_i$ ovat multinormaalijakautuneita keskiarvon ollessa $\mathbf{0}$ ja varianssi-kovarianssimatriisin ollessa $\mathbf{D}$:

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}),$$

$$\mathbf{D} = \mathrm{Var}(\mathbf{b}_i) = \begin{pmatrix} \mathrm{Var}(b_{1i}) & \mathrm{Cov}(b_{1i}, b_{2i}) & \dots & \mathrm{Cov}(b_{1i}, b_{qi}) \\ \mathrm{Cov}(b_{1i}, b_{2i}) & \mathrm{Var}(b_{2i}) & \dots & \mathrm{Cov}(b_{2i}, b_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(b_{1i}, b_{qi}) & \mathrm{Cov}(b_{2i}, b_{qi}) & \dots & \mathrm{Var}(b_{qi}) \end{pmatrix}.$$

Saman yksikön residuaalit, jotka liittyvät useisiin havaintoihin, voivat korreloida keskenään. Yksikön i residuaaleista tehdään vastaavat oletukset kuin satunnaisvaikutuksista:

$$\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i).$$

Eri yksiköiden residuaalit sen sijaan ovat riippumattomia toisistaan. Lisäksi yksiköiden residuaalit ja satunnaisvaikutukset ovat keskenään riippumattomia:

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\epsilon}_i) = \begin{pmatrix} \mathrm{Var}(\epsilon_{1i}) & \mathrm{Cov}(\epsilon_{1i}, \epsilon_{2i}) & \dots & \mathrm{Cov}(\epsilon_{1i}, \epsilon_{qi}) \\ \mathrm{Cov}(\epsilon_{1i}, \epsilon_{2i}) & \mathrm{Var}(\epsilon_{2i}) & \dots & \mathrm{Cov}(\epsilon_{2i}, \epsilon_{qi}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\epsilon_{1i}, \epsilon_{qi}) & \mathrm{Cov}(\epsilon_{2i}, \epsilon_{qi}) & \dots & \mathrm{Var}(\epsilon_{qi}) \end{pmatrix}.$$

Sekä $\mathbf{D}$ ja $\mathbf{R}_i$ matriiseille on olemassa erilaisia kovarianssirakenteita, joiden parametrit voidaan esittää vektoreissa $\boldsymbol{\Theta}_D$ ja $\boldsymbol{\Theta}_R$.

Rakenteettomaksi kovarianssirakenteeksi (UN, unstructured) kutsutaan rakennetta, joka ei aiheuta muita rajoitteita kuin, että matriisin tulee olla positiividefiniitti

ja symmetrinen. Sellaista käytetään erityisesti satunnaisvaikutusmalleissa. Se sopii myös hyvin pitkittäisaineistoihin, koska sen käyttöön ei liity mitään oletuksia virheiden rakenteesta [45]. Koska tämä $q \times q$-matriisi on symmetrinen, $\mathbf{\Theta}_D$-vektorilla on yhteensä $(q \times (q+1))/2$ parametria. Esimerkiksi LSM, jossa on kaksi satunnaisvaikutusta liittyen yksikköön i:

$$\mathbf{D} = \mathrm{Var}(\mathbf{b}_i) = \begin{pmatrix} \sigma^2_{b_1} & \sigma_{b_1,b_2} \\ \sigma_{b_1,b_2} & \sigma^2_{b_2} \end{pmatrix}.$$

Näin ollen $\mathbf{\Theta}_D$ sisältää kolme parametria:

$$\mathbf{\Theta}_D = \begin{pmatrix} \sigma^2_{b_1} \\ \sigma_{b_1,b_2} \\ \sigma^2_{b_2} \end{pmatrix}.$$

Toinen yleinen kovarianssirakenne matriisille $\mathbf{D}$ on *varianssikomponenttirakenne* (VC, variance components), jossa jokaisella satunnaisvaikutuksella on oma varianssi, mutta kaikki kovarianssit ovat nollia. Siten $\mathbf{\Theta}_D$-vektorissa on yhteensä q parametria eli matriisin $\mathbf{D}$ diagonaalin varianssit. Esimerkiksi LSM, jossa on kaksi satunnaisvaikutusta liittyen yksikköön i:

$$\mathbf{D} = \mathrm{Var}(\mathbf{b}_i) = \begin{pmatrix} \sigma^2_{b_1} & 0 \\ 0 & \sigma^2_{b_2} \end{pmatrix}.$$

Tässä tapauksessa $\mathbf{\Theta}_D$ sisältää kaksi parametria:

$$\mathbf{\Theta}_D = \begin{pmatrix} \sigma^2_{b_1} \\ \sigma^2_{b_2} \end{pmatrix}.$$

Rakenteeton kovarianssirakenne ja *varianssikomponenttirakenne* ovat yleisimmin käytössä olevat rakenteet matriisille $\mathbf{D}$, mutta muitakin rakenteita on olemassa. Niitä on esitelty muun muassa teoksessa [2].

Diagonaalirakenne (DIAG, diagonal) on yksinkertaisin $\mathbf{R}_i$-matriisin rakenne, jossa saman yksikön residuaalien oletetaan olevan korreloimattomia ja niillä oletetaan olevan yhtä suuri varianssi. Diagonaalinen $\mathbf{R}_i$ jokaiselle yksikölle i voidaan esittää muodossa:

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\epsilon}_i) = \sigma^2\mathbf{I} = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}.$$

Diagonaalisessa rakenteessa $\boldsymbol{\Theta}_R$ sisältää vain yhden parametrin:

$$\boldsymbol{\Theta}_R = \begin{pmatrix} \sigma^2 \end{pmatrix}.$$

Eräs usein käytetty $\mathbf{R}_i$-matriisin rakenne on *tasakorrelaatiorakenne* (CS, compound symmetry). Rakennetta voidaan käyttää esimerkiksi silloin, kun havaintokohteesta otetaan useita näytteitä vastaavissa olosuhteissa, jolloin sekä kovarianssi että varianssi ovat vakioita. Kyseistä kovarianssirakennetta voidaan käyttää, kun tutkitaan, ovatko varianssi ja korrelaatio havaintoparien välillä vakioita kaikilla mittauskerroilla. Yleinen muoto $\mathbf{R}_i$ jokaiselle yksikölle i voidaan esittää muodossa:

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\epsilon}_i) = \begin{pmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \dots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \dots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \dots & \sigma^2 + \sigma_1 \end{pmatrix}.$$

Tällöin $\boldsymbol{\Theta}_R$ sisältää kaksi parametria:

$$\boldsymbol{\Theta}_R = \begin{pmatrix} \sigma^2 \\ \sigma_1 \end{pmatrix}.$$

Eräs toinen usein käytetty $\mathbf{R}_i$-matriisin rakenne on *autoregressiivinen* (AR(1), first-order autoregressive). Sen yleinen muoto voidaan esittää jokaiselle yksikölle i muodossa:

$$\mathbf{R}_i = \mathrm{Var}(\boldsymbol{\epsilon}_i) = \begin{pmatrix} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \dots & \sigma^2 \end{pmatrix}.$$

Tällöin $\boldsymbol{\Theta}_R$ sisältää kaksi parametria:

$$\mathbf{\Theta}_R = \begin{pmatrix} \sigma^2 \\ \rho \end{pmatrix}.$$

AR(1) rakennetta käytetään yleensä silloin, kun havaintoja on tehty tasaisin määrävälein havaintoyksiköistä. Rakenteessa havainnot, jotka on otettu ajallisesti lähimpinä toisiaan korreloivat keskenään enemmän kuin havainnot, jotka on otettu ajallisesti kaukana toisistaaan.

Silloin, kun havaintoyksiköiden vastemuuttujien mittaukset suoritetaan epäsäännöllisesti siten, että mittausten aikavälit ovat enemmän tai vähemmän yksilölliset kullekin havaintoyksikölle, tarvitaan mallivirheille sellainen kovarianssirakenne, jossa aika kuvataan jatkuvana muuttujana. Tällainen rakenne on esimerkiksi *spatiaalinen potenssi* eli SP(POW) (spatial power):

$$\mathrm{Cov}(\epsilon_{ij}, \epsilon_{i'j}) = \sigma^2_\epsilon \rho^{d_{ii'}},$$

jossa $d_{ii'}$ mittausväli ajan hetkille i ja i' ja ρ on korrelaatiokerroin, joka mittaa mallivirheiden lineaarisuutta, kun mittaukset on tehty yhden yksikön aikavälillä. Silloin, kun aikavälit ovat kokonaislukuja, malli redusoituu muotoon AR(1).

3.3 Estimointimenetelmät

Tässä luvussa esitellään LSM:n parametrien estimointimenetelmistä suurimman uskottavuuden menetelmä (ML, maximum likelihood) ja rajoitettu suurimman uskottavuuden menetelmä (REML, restricted maximum likelihood). Esitys mukailee teoksen [53] lukua 2.4. On olemassa muita estimointimenetelmiä, kuten momenttimenetelmät (MM, method of moments) ja MIVQUE0 (minimum variance quadratic unbiased estimation), mutta niitä ei käsitellä tässä tutkielmassa. REML-menetelmä ei ole yhtä laskennallisesti vaativa kuin ML-menetelmä ja lisäksi REML-estimaatteja pidetään vähemmän harhaisina kuin ML-estimaatteja. Kuten ML-menetelmässä, REML-menetelmän jakauman ominaisuuksia ei tunneta kuin asymptoottisesti. Simulaatiotulosten perusteella REML- ja ML-menetelmiä pidetään parhaina vaihtoehtoina estimoinnissa [50].

ML-estimointi

Rakennetaan parametrien $\boldsymbol{\beta}$ (kiinteiden vaikutusten parametrit) ja $\boldsymbol{\Theta}$ (kaikki kovarianssiparametrit) uskottavuusfunktio käyttäen pohjana vasteen $\mathbf{y}_i$ rajajakaumaa. Mallin (5) mukaisen multinormaalijakauman havaintoaineiston $\mathbf{y}_i$ määräämä todennäköisyystiheysfunktio on muotoa

$$L_i(\mathbf{y}_i) = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sqrt{\det(\mathbf{V}_i)}} \exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}, \tag{8}$$

jossa ”det” tarkoittaa determinanttia ja ”exp” eksponenttifunktiota. Kootaan uskottavuusfunktiot yhteen kertomalla havaintoyksiköiden (n kpl) vaikutus yhtälön (8) mukaisesti:

$$L(\boldsymbol{\beta}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} L_i(\boldsymbol{\beta}, \boldsymbol{\Theta}). \tag{9}$$

Vastaava log-uskottavuusfunktio on

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\Theta}) &= \ln L(\boldsymbol{\beta}, \boldsymbol{\Theta}) \\ &= -\frac{1}{2} n \ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} \ln\{\det(\mathbf{V}_i)\} \\ &\quad - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}), \end{aligned} \tag{10}$$

jossa $n = \sum n_i$ on havaintojen lukumäärä ja ”ln” on luonnollinen logaritmi. Vaikka on mahdollista etsiä yhtälöstä (10) estimaatit parametreille $\boldsymbol{\beta}$ ja $\boldsymbol{\Theta}$ samanaikaisesti, voidaan tehtävää yksinkertaistaa profiloimalla parametri $\boldsymbol{\beta}$ ulos siten, että oletetaan parametrin $\boldsymbol{\Theta}$ ja sitä seuraten $\mathbf{V}_i$ olevan tunnettuja. Näin ei käytännössä kuitenkaan ole, vaan $\mathbf{V}_i$ täytyy estimoida ensin. Siis log-uskottavuusfunktio on ainoastaan muuttujan $\boldsymbol{\beta}$ funktio ja sen optimointi tapahtuu etsimällä minimi funktiolle $q(\boldsymbol{\beta})$, joka on yhtä kuin viimeinen termi yhtälössä (10):

$$q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{11}$$

Yhtälön (11) optimointi muuttujan $\boldsymbol{\beta}$ suhteen voidaan etsiä yleistetyn pienimmän neliösumman menetelmällä ja muuttujan $\boldsymbol{\beta}$ optimi voidaan ratkaista analyyttisesti:

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i. \tag{12}$$

Estimaatti $\hat{\boldsymbol{\beta}}$ on muuttujan $\boldsymbol{\beta}$ paras lineaarinen harhaton estimaattori (BLUE).

Seuraavaksi tarkastellaan tilannetta, jossa estimoidaan parametrit $\boldsymbol{\beta}$ ja $\boldsymbol{\Theta}$ olettaen, että $\boldsymbol{\Theta}$ on tuntematon. Vastaavasti, kuten edellä, profiloidaan nyt vuorostaan $\boldsymbol{\Theta}$ ulos log-uskottavuusfunktiosta (10).

Ensiksi rakennetaan profiloitu log-uskottavuusfunktio, joka johdetaan funktiosta $l(\boldsymbol{\beta}, \boldsymbol{\Theta})$ korvaamalla $\boldsymbol{\beta}$ sen estimaattorilla $\hat{\boldsymbol{\beta}}$ yhtälön (12) mukaisesti. Tuloksena on funktio

$$l_{ML}(\boldsymbol{\Theta}) = -\frac{1}{2} n \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \ln \{\det(\mathbf{V}_i)\} - \frac{1}{2} \sum_{i=1}^{n} \mathbf{r}_i' \mathbf{V}_i^{-1} \mathbf{r}_i, \tag{13}$$

jossa

$$\mathbf{r}_i = \mathbf{y}_i - \mathbf{X}_i \left\{ (\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{y}_i \right\}. \tag{14}$$

Yhtälöä (13) ei voi ratkaista suljetussa muodossa, joten estimaatti $\hat{\boldsymbol{\Theta}}$ ratkaistaan iteroimalla. Iteroinnissa käytetään esimerkiksi Newton–Raphson-, scoring- tai DFP-algoritmeja [39]. Kun iterointialgoritmeilla on saatu ratkaistua estomaattorit variansseille ja kovariansseille matriisissa $\mathbf{D}$ ja $\mathbf{R}_i$, voidaan laskea estimaattori $\hat{\boldsymbol{\beta}}$. Ensiksi korvataan $\mathbf{D}$ ja $\mathbf{R}_i$ yhtälössä (4) niiden ML-estimaateilla $\hat{\mathbf{D}}$ ja $\hat{\mathbf{R}}_i$ ja lasketaan $\hat{\mathbf{V}}_i$ ($\mathbf{V}_i$ estimaatti):

$$\hat{\mathbf{V}}_i = \mathbf{Z}_i \hat{\mathbf{D}} \mathbf{Z}_i' + \hat{\mathbf{R}}_i. \tag{15}$$

Sitten käytetään yhtälössä (12) esitettyä yleistetyn pienimmän neliösumman ratkaisua, jossa $\mathbf{V}_i$ korvataan sen estimaatilla (15)

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{n} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{y}_i, \tag{16}$$

joka on muuttujan $\boldsymbol{\beta}$ paras empiirinen lineaarinen harhaton estimaattori (EBLUE). Estimaattorin $\hat{\boldsymbol{\beta}}$ varianssi on $p \times p$ varianssi-kovarianssi-matriisi:

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{n} \mathbf{X}_i' \hat{\mathbf{V}}_i^{\ -1} \mathbf{X}_i)^{-1}. \tag{17}$$

REML-estimointi

REML-estimaatit parametrille $\boldsymbol{\Theta}$ pohjautuvat REML log-uskottavuusfunktion optimointiin:

$$\begin{aligned} l_{REML}(\boldsymbol{\Theta}) = &-\frac{1}{2}(n-p)\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n} \ln\left\{\det(\mathbf{V}_i)\right\} \\ &- \frac{1}{2}\sum_{i=1}^{n} \mathbf{r}_i' \mathbf{V}_i^{-1} \mathbf{r}_i - \frac{1}{2}\sum_{i=1}^{n} \ln\left\{\det(\mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)\right\}, \end{aligned} \tag{18}$$

jossa $\mathbf{r}_i$ on kuten yhtälössä (14). Sen jälkeen, kun $\hat{\mathbf{V}}_i$ on laskettu jollakin soveltuvalla iterointialgoritmilla, voidaan laskea REML-estimaatit kiinteiden vaikutusten parametreille $\hat{\boldsymbol{\beta}}$ ja $\mathrm{Var}(\hat{\boldsymbol{\beta}})$. REML-menetelmässä ei ole erityistä yhtälöä kiinteiden vaikutusten parametrien estimoinnille, joten käytetään ML-menetelmän yhtälöitä (15) ja (16). Vaikka estimoinnissa käytetään molemmissa menetelmissä samoja yhtälöitä, $\hat{\boldsymbol{\beta}}$ ja $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ eroavat toisistaan ML- ja REML-menetelmissä, koska $\hat{\mathbf{V}}_i$ on eri molemmissa menetelmissä.

3.4 Mallin valinta ja diagnostiikka

Päämääränä on valita sellainen malli, joka selittää tutkittavaa ilmiötä hyvin, mutta on samaan aikaan mahdollisimman yksinkertainen. Riippuen aineistosta erilaisia mahdollisuuksia sisällyttää kiinteitä vaikutuksia ja satunnaisvaikutuksia voi olla useita. Myös $\mathbf{D}$ ja $\mathbf{R}_i$ matriisille on olemassa useita erilaisia vaihtoehtoja. Mallin valintaa tehtäessä pitää kokeilla useita erilaisia vaihtoehtoja kiinteille vaikutuksille ja kovarianssirakenteille. Lopullisen mallin valinta riippuu sekä tilastotieteellisten että tutkimusalan tietojen käyttämisestä. Kaikkiin tilanteisiin yksiselitteisesti sopivia ratkaisuja ei ole olemassa ja siitä syystä lopullisen mallin valinnan tekee tutkija. Mallin valintaan on olemassa kaksi perusstrategiaa: ”ylhäältä alaspäin” ja ”alhaalta ylöspäin” askeltavat strategiat. [53]

Ylhäältä alaspäin -strategia

Mallin rakentaminen aloitetaan lisäämällä niin paljon kiinteitä vaikutuksia (ja niiden yhdysvaikutuksia) kuin aineisto antaa mahdollisuuksia ja on tutkimusasetelman kannalta mielekästä. Seuraavaksi malliin ryhdytään lisäämään satunnaisvaikutuksia. Niiden tarvetta voidaan arvioida tilastollisilla testeillä. Kun sekä kiinteät vaikutukset että satunnaisvaikutukset on lisätty malliin, jäljelle jäänyt mallin vaihtelu johtuu residuaaleista, joille valitaan soveltuva kovarianssirakenne. Seuraavaksi mallia ryhdytään yksinkertaistamaan, mikä tarkoittaa sitä, että tilastollisten testien avulla tutkitaan kiinteiden vaikutusten tarvetta olla mallissa selittäjinä. Kiinteiden vaikutusten poistaminen aloitetaan monimutkaisimmista selittäjistä, joita ovat yhdysvaikutukset. Yleensä on vielä niin, että jos jollekin tietylle kiinteälle vaikutukselle ei ole merkitsevää tarvetta olla mallissa, voidaan kyseisen muuttujan satunnaisvaikutuskin poistaa jatkotarkasteluista. [53, 52]

Alhaalta ylöspäin -strategia

Alhaalta ylöspäin -strategia on peräisin hierarkisien lineaaristen mallien kirjallisuudesta ja sitä on esitelty muun muassa teoksissa [40, 46] ref. [53]. Mallin rakentaminen aloitetaan sijoittamalla tason 1 mallin kiinteisiin vaikutuksiin ainoastaan vakiotermi. Satunnaisvaikutukset, jotka liittyvät tason 2 ja 3 muuttujiin, sisällytetään malliin mukaan. Näin voidaan arvioida tasojen välistä vaihtelua pitkittäisaineistossa ilman, että tarkastellaan selittäjien vaikutusta.

Seuraavassa vaiheessa ryhdytään lisäämään selittäviä muuttujia tason 1 malliin. Tason 2 malliin tutkitaan mahdollisuuksia lisätä tason 1 kiinteitä vaikutteita satunnaisvaikutuksina. Kolmannessa vaiheessa lisätään tason 2 malliin selittäviä muuttujia ja tutkitaan mahdollisuuksia lisätä niitä vastaavia satunnaisvaikutuksia tason 3 malliin. Kun on löydetty sopivat yhtälöt tason 1 mallin selittäjille tason 2 mallissa on saatu määriteltyä, voidaan arvioida tason 2 mallin satunnaisvaikutuksien toimivuutta.

Tilastolliset testit mallin valinnassa

Sisäkkäisten mallien tapa on vaihtoehto vertailla kahta kilpailevaa mallia (A ja B) keskenään. Malli A on sisäkkäinen mallille B, jos A on yleisemmän mallin B osajoukko. Yleinen menetelmä sisäkkäisten suurimman uskottavuuden mallien vertailemi-

seen on uskottavuussuhdetesti (LR-testi). Testisuure, joka noudattaa χ^2 -jakaumaa, määritellään yhtälöllä

$$-2\ln\frac{L_{pienempi}}{L_{suurempi}} = -2\ln(L_{pienempi}) + 2\ln(L_{suurempi}) \sim \chi^2_{df}, \tag{19}$$

jossa $L_{pienempi}$ ja $L_{suurempi}$ tarkoittavat joko ML- tai REML-menetelmällä laskettuja parametriestimaatteja. *Suuremmalla* mallilla tarkoitetaan vertailuparin mallia, missä on enemmän selittäviä muuttujia kuin *pienemmässä* mallissa. Vapausaste *df* saadaan vähentämällä suuremman mallin parametrien määrästä pienemmän mallin parametrien määrä. Jos testisuureen arvo on riittävän suuri, nollahypoteesia (H_0: ”pienempi malli on parempi”) vastaan on riittävästi todistetta ja pitäydytään suuremmassa mallissa (H_A: ”suurempi malli on parempi”). Jos taas testisuure on pieni, todistusta on pienemmän mallin hyväksi. Kun testi suoritetaan REML-estimaateille pitää molemmissa vertailtavassa mallissa olla samanlainen $\mathbf{X}$-matriisi. Siten kiinteiden parametrien testaamisessa käytetään ML-estimaatteja [52].

Toinen tapa mallin valinnassa on käyttää erilaisia informaatiokriteereitä, jolloin vertailtavien mallien ei tarvitse olla sisäkkäisiä. Informaatiokriteerit tulkitaan siten, että mitä pienempi on kriteerin arvo, sitä paremmin malli sopii aineistoon. Akaiken informaatiokriteeri (AIC, Akaike information criteria) voidaan laskea joko käyttäen ML- tai REML-estimaatteja yhtälöllä

$$AIC = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}) + 2p \tag{20}$$

ja Bayesin informaatiokriteeri (BIC, Bayesian information criteria) voidaan laskea yhtälöllä

$$BIC = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}}) + p\ln(n), \tag{21}$$

joissa n on aineiston havaintojen määrä ja p on mallin parametrien määrä.

Mallin valinnan jälkeen tehdään mallille asetettujen oletuksien tarkastaminen, jota kutsutaan mallin diagnostiikaksi. Erityisesti huomioidaan jakaumaoletukset ja onko valittu malli herkkä epätavallisille havainnoille.

Residuaalit

Residuaali on havaitun arvon ja sen estimoidun tai ennustetun arvon välinen erotus. Kun kyseessä ovat lineaariset sekamallit, voidaan residuaalit jakaa reunaresiduaaleiksi (marginal residuals) ja ehdollisiksi residuaaleiksi (conditional residuals). LSM:n reunakeskiarvo on $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ ja ehdollinen keskiarvo $\mathrm{E}[\mathbf{Y}|\mathbf{b}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$. Reunaresiduaali $\mathbf{r}_m$ on havaitun arvon ja estimoidun keskiarvon välinen erotus:

$$\mathbf{r}_m = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

ja ehdollinen residuaali $\mathbf{r}_c$ on havaitun arvon ja ennustetun arvon välinen erotus [42]:

$$\mathbf{r}_c = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}} = \mathbf{r}_m - \mathbf{Z}\hat{\mathbf{b}}.$$

Residuaaleja käytetään mallin oletusten tarkistamisessa ja vieraiden ja mahdollisesti vaikutusvaltaisten havaintojen havaitsemisessa. Residuaalit $\mathbf{r}_m$ ja $\mathbf{r}_c$ eivät yleensä sovellu sellaisenaan kovin hyvin näihin tarkoituksiin. Vaikka todellisen mallin virheet eivät korreloisi keskenään ja niiden varianssit olisivat yhtä suuret, residuaalit voivat korreloida keskenään ja niiden varianssi voi vaihdella.

Satunnaismuuttuja voidaan standardoida jakamalla muuttujan arvon ja keskiarvon erotus keskihajonnalla. Edellä mainittujen residuaalien keskiarvo on nolla, mutta varianssi on tuntematon, koska se riippuu vektorin $\boldsymbol{\Theta}$ todellisista arvoista. Studentoitu residuaali lasketaan jakamalla residuaalin arvo sen keskihajonnan estimaatilla. Pearson-residuaalissa havainnon ja ennusteen erotuksen jakajana vastemuuttujan keskihajonta. Näin voidaa toimia, jos $\hat{\boldsymbol{\beta}}$ vaihtelua ei oteta huomioon. Voidaan myös tarkastella residuaalien vektoria $\boldsymbol{\Theta}$ ja sen varianssiestimaattia $\mathbf{V}(\hat{\boldsymbol{\Theta}})$.

Havaintojen vaikutusvaltaisuus

ML- ja REML-menetelmät ovat herkkiä epätavallisille havainnoille tutkimusaineistossa. Vaikutusvaltaisuutta (influence diagnostics) tutkitaan tekniikoilla, joilla voidaan löytää havaintoja, jotka vaikuttavat voimakkaasti vektorien $\boldsymbol{\beta}$ ja $\boldsymbol{\Theta}$ parametrien arvoihin. Ideana on mitata yksittäisen havainnon tai havaintojoukon poistamisen vaikutusta koko tutkimusaineiston analysoinnin tuloksiin.

Vaikutusvaltaisuuden diagnostiikka voidaan jakaa pääpiirteittäin neljään ryhmään [42]:

- Kokonaisvaikutusvaltaisuus,
  - Likelihood-etäisyys /-siirtymä
  - Rajoitettu likelihood-etäisyys /-siirtymä
- Parametrien estimaattien muutos,
  - Cookin etäisyys
  - Multivariate DFFITS statistic
- Parametrien estimaattien tarkkuuden muutos ja
  - Hajontamatriisin jälki
  - Kovarianssisuhde
- Vaikutus sovitettuihin ja ennustettaviin arvoihin.

Puuttuvat havainnot

Puuttuvat havainnot ovat melko tyypillisiä pitkittäis- tai paneeliaineistoissa, koska havaintojen kohteet voivat poistua kokeesta. Tutkimuksen kohteet voivat myös unohtaa vastata tai olla haluttomia vastaamaan. Myös tietojen tallentamisessa ja kirjaamisessa on voinut tapahtua virheitä. Käytännössä kaikkiin pitkittäistutkimuksiin liittyy puuttuneisuutta ja aineiston epätasapainoa. Jotkut pitkittäistutkimukset on kuitenkin suunniteltu sellaisiksi, että eri poikkileikkausyksiköiltä mitataan havaintoja eri määrä tai mittausmäärä voi olla satunnainen. Lisäksi mittausten aikaväli voi vaihdella tai olla satunnainen. Tällaisia tutkimuksia kutsutaan epätasapainoisiksi. Päinvastoin tasapainoisissa tutkimuksissa mittausten määrä on eri poikkileikkausyksiköillä sovittu kiinteäksi ja mittausten ajankohta on mahdollisimman yhtenäinen. Puuttuneisuuden arvioimisessa on siten tärkeää tietää tutkimuksen lähtötavoitteet. [12, 43]

Puuttuneisuuden arvioimiseksi määritellään seuraavat termit [41, 28]:

- Täydelliset havainnot: tulosvektori $\mathbf{Y}_i$, joka tallennettaisiin tilanteessa, jossa ei olisi lainkaan puuttuvia havaintoja.
- Puuttumisen indikaattorit: binaarimuotoiset indikaattorit tallennetaan vektoriin $\mathbf{R}_i$, joka ilmaisee josko Y_{ij} havaittiin ($R_{ij} = 1$) tai ei havaittu ($R_{ij} = 1$).

- Havaitut havainnot: vektori $\mathbf{Y}_i^O$, joka sisältää ne mittaukset Y_{ij}, joille $R_{ij} = 1$.
- Puuttuvat havainnot: vektori $\mathbf{Y}_i^M$, joka sisältää ne mittaukset Y_{ij}, joille $R_{ij} = 0$.
- Poistumisaika: jos puuttuneisuus on monotonista (jos kerran $R_{ij} = 0$, niin $R_{ik} = 0 \forall k > j$), voidaan määrittää aika, jolloin yksikkö poistuu kokeesta: $D_i = min_k(R_{ik} = 0)$).

Puuttuneisuus on täysin satunnaista (MCAR, missing completely at random), jos $P(\mathbf{R}_i|\mathbf{Y}_i, \mathbf{X}_i) = P(\mathbf{R}_i|\mathbf{X}_i)$, mikä tarkoittaa, että $E(Y_{ij}|R_{ij} = 1, \mathbf{X}_i) = E(Y_{ij}|\mathbf{X}_i)$.

Puuttuneisuus on satunnaista (MAR, missing at random), jos $P(\mathbf{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M, \mathbf{X}_i) = P(\mathbf{R}_i|\mathbf{Y}_i^O, \mathbf{X}_i)$. Tässä tapauksessa puuttuvan tiedon todennäköisyys riippuu vain havaituista arvoista ei puuttuvista arvoista. Ongelmallista sen sijaan on, että mahdollisesti $E(Y_{ij}|R_{ij} = 1, \mathbf{X}_i) \neq E(Y_{ij}|\mathbf{X}_i)$.

Puuttuneisuus voi olla luonteeltaan myös muuta kuin satunnaista (MNAR, missing not at random), jos todennäköisyys $P(\mathbf{R}_i|\mathbf{Y}_i^O, \mathbf{Y}_i^M, \mathbf{X}_i)$ riippuu puuttuvista arvoista $\mathbf{Y}_i^M$.

Usein puuttuneisuuden jakautuminen voidaan jättää huomioimatta suoritettaessa analyyseja aineistolle, jos puuttuneisuus on MCAR tai MAR. Sen sijaan, jos puuttuneisuus ei ole satunnaista (MNAR), sitä ei voida jättää huomioimatta jatkoanalyyseissa. Englanninkielisessä kirjallisuudessa puuttuneisuus jaetaankin usein sen huomioimisen tarpeen mukaan (ignorable tai non-ignorable).

Puuttuneisuus voidaan jakaa myös yksikkökatoon (unit non-response) ja eräkatoon (item non-response). Yksikkökato tarkoittaa, kyselyn kohteelta ei saada mitään tietoja, lähinnä siksi, että kohde ei vastaa kyselyyn. Eräkato tarkoittaa sitä, että kyselyn kohteesta on saatu tiedot kerättyä, mutta niissä on puutteita. Tässä tutkielmassa käsitellään puuttuneisuutta eräkadon kannalta. Yksikkökadon huomioimiseen liittyy kyselyn tulosten painottaminen.

Puuttuneisuuden ongelmien korjaamiseksi on olemassa useita erilaisia menetelmiä [18].

Puuttuvien havaintojen poistaminen. Yksinkertainen lähestymistapa on poistaa analysoitavien havaintojen joukosta ne havaintoyksiköt, joista on puuttuvia tietoja analyysiin sisältyvissä muuttujissa. Tällainen ratkaisu voi kuitenkin pienentää ratkaisevasti saadun otoksen kokoa. Ennen poistamista on syytä tarkistaa keskittyykö puuttuneisuus joihinkin tiettyihin havaintoryhmiin, koska siinä tapauksessa puuttuvien

havaintojen poistaminen vääristää analyysin avulla tehtäviä päätelmiä.

Muuttujien poistaminen. Analyysiin ei ehkä ole järkevää ottaa mukaan sellaista muuttujaa, jossa on merkittävä määrä havainnoista puutteellisia. Tällainen vaihtoehto ei tietenkään silloin tule kyseeseen, jos kyseinen muuttuja on olennainen tutkimusongelman ratkaisemisen kannalta. Jos kuitenkin tällainen muuttuja on mahdollista poistaa, ja esimerkiksi jokin toinen muuttuja mittaa lähes samaa asiaa, on hyvänä puolena se, että havaintoyksikköjen määrä ei vähene.

Puuttuvien havaintojen parittainen poistaminen. Useat monimuuttujamenetelmät perustuvat muuttujien kovarianssi- tai korrelaatiomatriisin analysointiin (esimerkiksi faktori- ja regressioanalyysi). Tällaisessa tapauksessa puuttuvia havaintoja voidaan poistaa analyysista parittaisesti (pairwise deletion). Silloin korrelaatiomatriisia laskettaessa otetaan huomioon kaikki ne havaintoyksiköt, joista on tiedot niillä kahdella muuttujalla, joista korrelaatio lasketaan. Korrelaatiomatriisissa jokainen korrelaatioarvo voi siten perustua erilaiseen havaintoyksikköjen määrään. Seurauksena aineisto pienenee, mutta ei läheskään yhtä paljon verrattuna tilanteeseen, jossa kaikki puuttuvia tietoja sisältävät havaintoyksiköt poistettaisiin analyysista.

Keskiarvon käyttö. Jos puuttuvia havaintoja ei voida poistaa, voidaan koodata puuttuvien muuttujan arvojen tilalle jokin ennalta päätetty arvo ja sisällyttää siten kaikki havaintoyksiköt analyysiin. Yleensä puuttuvien havaintojen tilalle koodataan muuttujan keskiarvo. Keskiarvon käyttöä perustellaan sillä, että jos tutkijalla ei ole etukäteen mitään tietoa puuttuvan havainnon arvosta, paras arvio täksi arvoksi on juuri koko aineiston keskiarvo. Ilmeinen etu tämän menetelmän käytössä on, että se ei pienennä aineiston kokoa. Huono puoli on, että keskiarvojen käyttö johtaa muuttujien hajonnan pienenemiseen. Jos puuttuvia havaintoja on paljon, voi tällä olla suuri merkitys jatkoanalyysin kannalta. Käytännössä muuttujien hajonnan pienenemisestä seuraa, että niiden välinen korrelaatio pienenee eli havaitut yhteydet muuttujien välillä eivät ole niin vahvoja, kuin jos puuttuvia havaintoja olisi aineistossa vähemmän.

Ryhmäkeskiarvojen käyttö. Puuttuvat muuttujan arvot voidaan korvata myös ryhmäkeskiarvoilla. Tämän menetelmän ongelma on, että se korostaa ryhmien sisäistä samankaltaisuutta ja ryhmien välisiä eroja. Seuraukset ovat päinvastaiset kuin koko muuttujan keskiarvojen käytössä puuttuvien havaintojen tilalla. Ryhmäkeskiarvojen käyttö voi vääristää tuloksia kasvattamalla muuttujien välisiä korrelaatioita.

Muita tapoja. Eräs keino on jakaa aineisto ryhmiin ja koodata puuttuvan arvon kohdalle havaintomatriisissa edellisen havainnon arvo. Tämä tarkoittaa, että puuttuvien

arvojen tilalle koodataan useita eri arvoja, ei ainoastaan keskiarvoja. Menetelmän etu on, että se ei vähennä muuttujien hajontaa niin kuin pelkkien keskiarvojen käyttö. Sen sijaan on melko karkea oletus, että muuttujan arvo pysyisi muuttumattomana edellisestä mittauskerrasta seuraavaan. Myös regressioanalyysia voidaan käyttää puuttuvien havaintojen sopivien arvojen löytämiseksi.

LSM:ssa voidaan hyödyntää myös sellaisia havaintoja yksiköistä, jotka ovat epätäydellisiä, tai siis, yksiköillä voi olla eri määrä havaintoja eri ajanhetkiltä. Havaintojen keräämisen ajanhetkien ei myöskään tarvitse olla samoja yksiköiden välillä. Tällainen joustavuus LSM:ssa on kuitenkin voimassa vain silloin, kun havaintojen puuttuminen on satunnaista (MCAR tai MAR).

3.5 Painotus

Painotuksessa on kyse siitä, että voitaisiin mahdollisimman hyvin estimoida halutun tavoiteperusjoukon tunnuslukuja. Painotus on siten tarpeellista silloin, kun aineisto ei kata koko tavoiteperusjoukkoa, mutta tuloksia halutaan esittää tavoiteperusjoukon tasolla. Painottamalla voidaan kompensoida sekä otannasta että puuttuneisuudesta aiheutuvia vaikutuksia.

Otantatutkimuksen painotus ei ole yksinkertaista. Ei ole selvää, miten käyttää painoja arvioitaessa monimutkaisempia kokonaisuuksia kuin yksinkertaisia keskiarvoja tai suhteita ja keskivirheiden laskeminen on hankalaa jopa painotetulle keskiarvolle. Otantatutkimuksen painot eivät yleensä vastaa valikoitumisen käänteistodennäköisyyksiä vaan pikemminkin painot perustuvat todennäköisyyksien ja kadon yhdistelmään. Painojen rakentaminen on prosessi, jota ei ole tarkoin määritelty.[16].

Yleensä ei ole selvää, miten painoja pitäisi käyttää vähänkin monimutkaisemmille estimaateille, kuten regressiokertoimille (asiaa käsitelty esimerkiksi artikkeleissa [13, 20, 37]). Regressiokertoimien määrittämisessä tulisi analyysiin sisällyttää "X-muuttujina" kaikki sellaiset muuttujat, jotka vaikuttavat otoksen valintaa tai katoon. Käytännössä rajoitutaan sellaisiin muuttujiin, joilla voidaan ajatella merkittävästi vaikuttavan näytteenottoon tai katoon ja jotka ovat myös kiinnostavia selittäviä muuttujia. Hyvä lähtökohta on sisällyttää analyysiin painotuksen perustana käytetyt muuttujat. Eräs suositeltu lähestymistapa, kun estimoidaan y:n regressiota z:lla, on käyttää painotettua pienimmän neliösumman menetelmää tai tehdä painottamaton regressio huomioiden X-muuttujat, joita on käytetty painotuksessa. Lisäksi on syytä huomioida, että painoja käytettäessä mallin muuttujien keskivirheitä on vai-

keampi tulkita luotettavasti. Tilastollisten mallien rakentamiseen liittyy usein subjektiiviset valinnat, mutta kirjallisuudessa olevat ohjeet painojen muodostamiseen ovat epämääräisempiä kuin menetelmäoppaat muihin tilastollisiin menetelmiin [29].

Painotettujen estimaattien keskivirheiden laskeminen ei ole yksinkertaista, koska painot itsessään ovat yleensä satunnaismuuttujia, jotka riippuvat aineistosta [54]. Klassisia keskivirheitä ei edes voida saada käyttäen aineistoa ja painoja, vaan lisäksi tarvitaan tietoa painojen laskemistavasta. Lisäksi, painotetut regressiot eivät yleensä anna oikeita keskivirheitä.

Simulaatiotutkimuksessa [30] on havaittu painojen huomiomatta jättäminen tai pitäminen painoja vakiona on havaittu aliarvioivan epävarmuutta (eli keskivirheet ovat liian pieniä), kun taas epävarmuus tulee yliarvioitua, jos painoja käsitellään käänteistodennäköisyyksinä. Tarkempia keskivirheitä voidaan saada käyttäen linkkuveitsimenetelmiä, jotka huomioivat painojen muodostamisen.

Yleensä y voi riippua sekä muuttujista X ja z ja siten muuttujien X populaation yhteisjakauman keskiarvosta. Selitettävän muuttujan y regressiivinen estimointi edellyttää muuttujien z ja X yhteisvaikutuksen estimointia [23]. Tästä syystä voi olla tarpeen sisällyttää regressiomalliin näiden muuttujien yhteisvaikutus selittäjäksi, vaikka tavoitteena olisikin selvittää vain muuttujien y ja z välistä riippuvuutta. Jos malliin sisällytetään yhteisvaikutuksia selittäjiksi, jälkiositus on välttämätöntä, jotta populaation regressiokertoimet voidaan estimoida.

Regressiomallin rakentaminen on verrattain helppoa, mutta otanta-aineiston kanssa tehtävä yleistys koko populaatioon johtaa usein monimutkaisiin malleihin. Liian monimutkaisen mallin tuloksia on vaikea hyödyntää ja tulkita.

Vaikka LSM soveltuu osite- ja ryväsaineistojen käsittelyyn, on epäselvää pitäisikö kyselyaineiston painoja käyttää ylipäätään ja miten painot pitäisi LSM:iin sisällyttää. Painojen lisäämiseen voidaan käyttää Pfeffermann–Skinner–Holmes–Goldstein–Rasbash-menetelmää (PSHGR) [36], Rabe-Hesketh–Skrondal-menetelmää (RHS) [38] ja Korn–Graubard-menetelmää (KG) [23]. Näitä kolmea lähestymistapaa on analysoitu ja vertailtu väitöskirjassa [6]. Kaikissa menetelmissä painot sijoitetaan LSM:n estimaattiin käyttäen PML-menetelmää (pseudo-maximum likelihood) estimoinnin eri vaiheissa johtaen siten erilaisiin mallin estimaattoreihin. KG-menetelmän heikkoudeksi väitetään, että se ei ole käytännönläheinen vaihtoehto, koska silloin tarvitaan korkeamman asteen painokertoimia, joita ei yleensä ole saatavilla. PSHGR- ja RHS-menetelmässä käytetään samoja painokertoimia keskenään, mutta eri kertoimia kuin KG-menetelmässä. LSM:lle, jossa on satunnaisvaikutuksena vain va-

kiotermi ja jossa populaation elementtien lukumäärä kaikille ryväksille yhtä suuri, toimivat PSHGR- ja RHS-menetelmät samalla tavalla. Silloin, kun populaation elementtien lukumäärä ryväksille vaihtelee, syntyy menetelmien välille selvä ero. Väitöskirjassa todetaan, että tarvitaan teoreettista tutkimusta siitä, miten painotusta pitäisi käyttää pitkittäisaineistojen analysoinnissa.

4 Lineaarisen sekamallin sovellus maatilapaneeliaineistoon

4.1 Tutkimusaineisto

Tutkielmassa on käytetty maa- ja puutarhatalouden kannattavuuskirjanpitoaineistoa vuosilta 2000–2011 eli paneeliaineistoa. Havaintoyksikkönä ovat maatalouden yritykset. Toistomittaukset yrityksistä on tehty vuoden välein. Tarkasteluvälillä 2000–2011 osa yrityksistä on poistunut vapaaehtoisesti kannattavuuskirjanpitotoiminnasta, osa on lopettanut toimintansa ja osa tullut mukaan kesken tarkastelujakson. Paneeliaineisto ei ole siis täydellinen. Tarkastelussa mukana ovat lypsykarjataloutta, viljanviljelyä, muuta kasvinviljelyä, muuta nautakarjataloutta, sikataloutta, kasvihuonetuotantoa, sekamuotoista tuotantoa, avomaatuotantoa, muuta laidunkarjataloutta ja siipikarjataloutta harjoittavat yritykset (taulukko 1).

Tutkittavana vastemuuttujana tutkielmassa on tuotantokustannus, joka on luonteeltaan jatkuva muuttuja. Tuotantokustannusta ja sen kehitystä tutkitaan yritystasolla kokonaistuotantokustannuksina ja yksikkökustannuksina per tuotettu yksikkö. Kustannukset on mitattu aineiston yritysten kirjanpitotuloksista. Hintojen analysoinnissa ja tulkinnassa käytetään kuluttajahintaindeksillä [48] vuoteen 2011 deflatoituja kustannus- ja hintatietoja.

Tuotantokustannus on summa seuraavista kustannuslajeista: tarvike-, kotieläin-, kone-, rakennus-, työ- ja korkokustannus sekä muu kustannus.

Tarvikekustannuksiin sisältyvät lannoitteet, kalkitus, siemenet, kasvinsuojelu, polttoaineet, sähkö ja ostorehut. Kotieläinkustannuksiin sisältyvät eläintenostokulut ja muita kotieläimiin liittyviä kustannuksia, kuten eläinlääkkeet, kotieläinten tarvikkeet, eläinlääkärikulut ja siemennyskulut. Konekustannuksiin sisältyvät konepoistot ja muita koneisiin liittyviä kustannuksia, kuten kunnossapito, vuokraus, kalustohankinta. Rakennuskustannus muodostuu rakennuspoistoista ja muista rakennuskustannuksista. Muu kustannus pitää sisällään vakuutukset, kiinteät vuokrat, peltovuokran, muut poistot ja muita kustannuslajeja. Työkustannus koostuu maksetuista palkoista ja yrittäjän palkkavaatimuksesta. Palkkavaatimus on yrityksen työkirjanpitoon perustuva yrittäjäperheen työtuntimäärä kerrottuna ennalta asetetulla tuntipalkkavaatimuksella, jota on muutettu vuosittain maataloustyöntekijän tuntipalkan muutoksen mukaisesti [32]. Korkokustannus koostuu korkokuluista ja oman pääoman korkovaatimuksesta.

Taulukko 1: Tilojen määrä tuotantosuunnittain 2000–2011.

Tuotantosuunta	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Lypsykarjatalous	359	332	331	324	344	371	365	366	367	358	353	335
Viljanviljely	132	134	142	155	161	167	164	158	171	171	154	147
Muu kasvinviljely	72	81	74	85	83	102	95	105	86	93	101	112
Muu nautakarjatalous	73	67	62	65	71	75	89	96	95	98	102	102
Sikatalous	97	92	88	81	75	77	70	70	66	56	52	45
Kasvihuonetuotanto	70	71	70	68	68	58	61	66	65	64	65	60
Sekamuotoinen tuotanto	63	60	52	53	57	53	65	65	67	71	67	55
Avomaatuotanto	36	32	25	19	14	10	13	15	18	13	13	14
Muu laidunkarja	12	11	10	10	14	16	17	23	21	21	21	19
Siipikarjatalous	15	14	13	13	16	18	13	13	12	14	12	15
Yhteensä	929	894	867	873	903	947	952	977	968	959	940	904

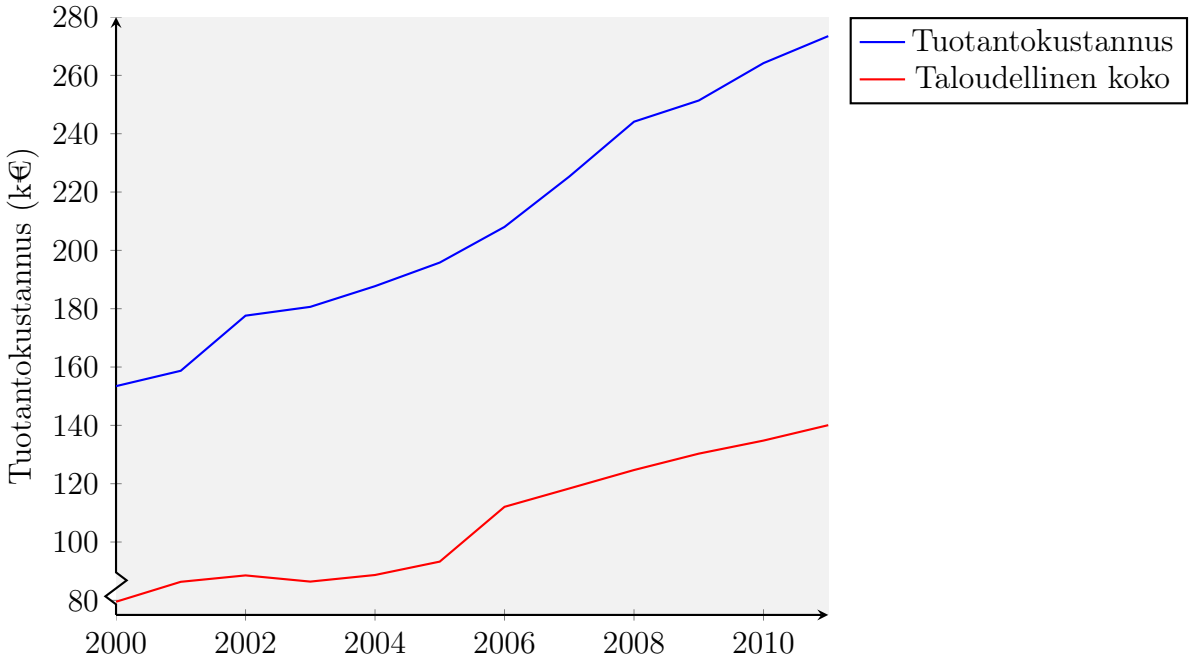
4.2 Kokonaistuotantokustannusmalli

Kun tarkastellaan kirjanpitotilojen keskimääräisiä kokonaistuotantokustannuksia 2000–2011, havaitaan, että kustannukset ovat kasvaneet tasaisesti koko ajanjakson (kuvio 4). Kasvu on ollut voimakasta (78%), mikä tarkoittaa keskimäärin yli 10 000 euron kasvua joka vuosi.

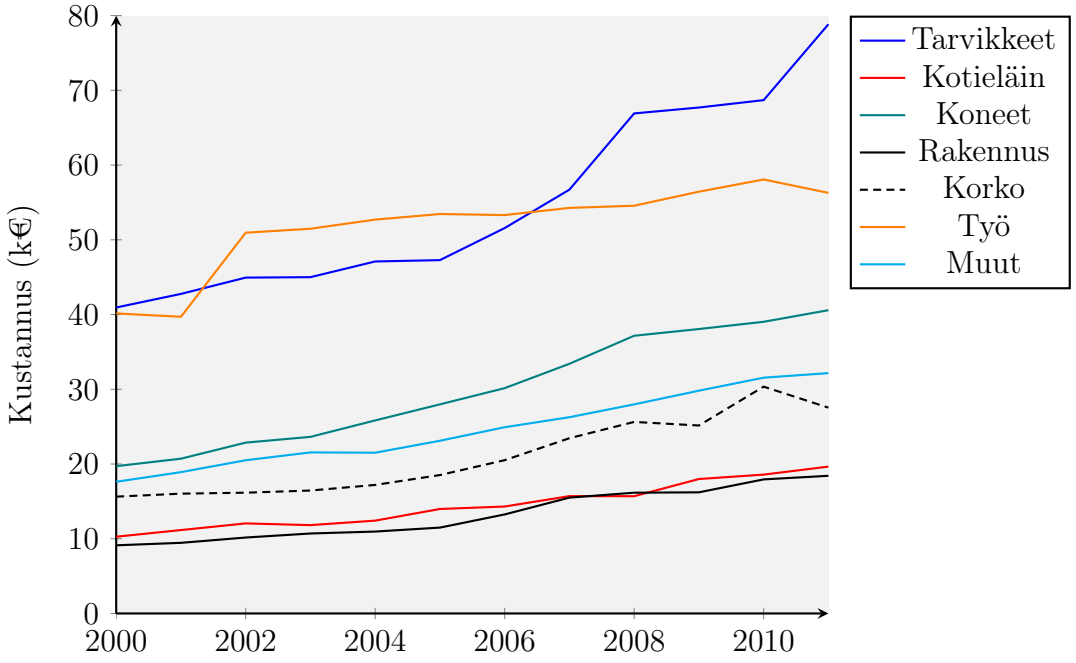
Joinakin vuosina kustannusten kasvu on ollut nopeampaa, esimerkiksi vuodesta 2001 vuoteen 2002 (keskimääräinen kasvu on ollut yli 18 000 euroa), vuodesta 2006 vuoteen 2007 (yli 17 000 euroa) ja vuodesta 2007 vuoteen 2008 (yli 18 000 euroa). Kustannuskasvua 2007–2008 selittää öljyn raaka-ainehinnan noususta johtuva energian ja poltto- ja voiteluainekustannuksen kasvu (+24% maatalouden tuotantovälineiden ostohintaindeksillä [49] mitattuna), rehujen (+14%) sekä erityisesti lannoitteiden ja maanparannusaineiden (+79%) voimakas kallistuminen. Lisäksi kallistunut korkotaso ja yritysten kasvanut velkataakka lisäsi pääomakustannuksia. Myös investointeihin oleellisesti liittyvä rakentamisen kustannustaso nousi yleistä inflaatioa nopeammin. Kasvun voidaan ajatella johtuvan myös osittain tilakoon kasvusta, koska keskimäärin tilojen koko on kasvanut. Tuotantokustannukset ovat kuitenkin kasvaneet tilakoon kasvua nopeammin (kuvio 4). [26, 34]

Kaikki osakustannukset ovat kasvaneet tarkastelujaksolla (kuvio 5); 12 vuoden aikana kustannukset ovat noin kaksinkertaistuneet lukuun ottamatta työkustannusta, joka on 1,4-kertaistunut. Suurimmat tuotantokustannukset maataloudessa ovat työkustannus ja tarvikekustannus, mitkä myös vaihtelevat eniten. Kuviosta 6 nähdään, että pääosin kirjanpitotiloilla edelleen työt tehdään viljelijäperheen voimin ja palkatun työvoiman osuus on pieni. Tämä johtuu siitä, että Suomessa maatilat ovat kooltaan pieniä, eikä monilla tiloilla ole tarvetta tai mahdollisuuksia palkata työvoimaa. Toisaalta kasvihuone- ja avomaatuotantoa harjoittavilla puutarhatiloilla palkatun työvoiman osuus koko työkustannuksista on merkittävästi muuta maataloutta suurempi, 60% ja 47%, vastaavasti. Vaikka työkustannus on ollut 2005–2011 suhteellisen vakiintunut, on yrittäjän palkkavaatimuksen osuus jatkuvasti pienentynyt. Aikavälillä 2001–2002 näkyvä tasomuutos johtuu palkkavaatimuksen laskentatavan muuttumisesta. Vuodesta 2002 lähtien on huomioitu palkkauksen sivukulut, jos työt teetätettäisiin ulkopuolisella viljelijän sijaan.

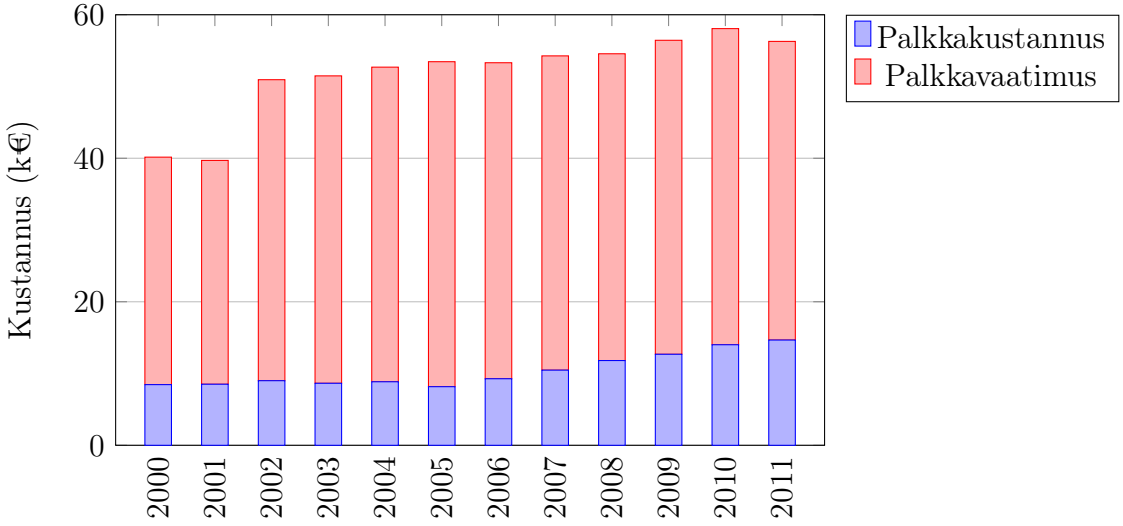
Kannattavuuskirjanpitotilojen tuotantokustannusten kehityssuunta 2000–2011 näyttää olevan sama eri tuotantosuunnissa (kuvio 7). Kustannukset ovat nousseet kaikissa tuotantosuunnissa. Kasvu on ollut voimakkainta kotieläintuotantoon erikoistuneilla tiloilla. Muuta nautakarjataloutta harjoittavissa yrityksissä kustannuskas-



Kuvio 4: Keskimääräisten tuotantokustannusten ja kirjanpitotilojen taloudelliseen koon kehitys 2000–2011.



Kuvio 5: Keskimääräisten tuotantokustannusten rakenteen kehitys 2000–2011.

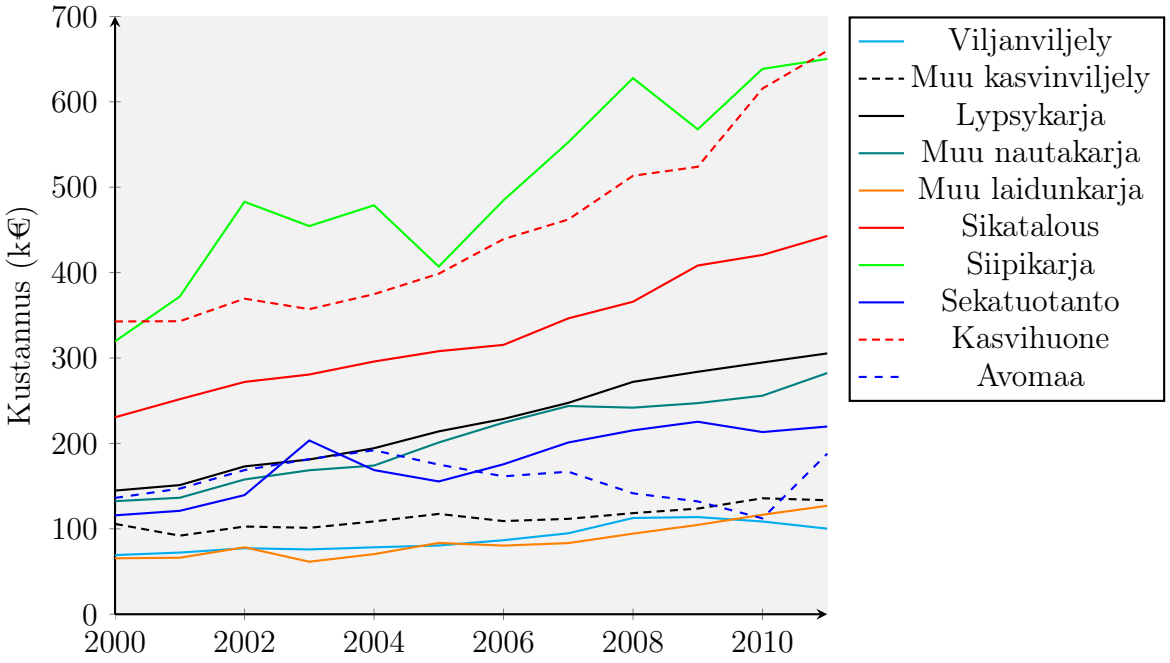


Kuvio 6: Työkustannuksen rakenne 2000–2011.

vu on ollut 113%, lypsykarjatiloilla 111%, siipikarjatiloilla 103%, sikatiloilla 92% ja muuta laidunkarjanpitoa harjoittavilla tiloilla 94%. Kasvihuonetiloilla kasvu on ollut 92% ja avomaatiloilla 38%. Sekatiloilla kokonaistuotantokustannukset ovat kasvaneet 90%, viljanviljelytiloilla 45% ja muuta kasvinviljelyä harjoittavilla tiloilla 26%. Siipikarjatiloilla on selkeimmin havaittavissa suurimmat vaihtelut vuosien välillä ja muutenkin tuotantokustannukset ovat suurimmat. Tutkimusaineistossa on kaikkein vähiten juuri siipikarjatiloja, joten muutokset näkyvät kuvaajassa helpoiten.

Tuotannon erilaisuudesta johtuen eri tuotantosuunnissa kustannusrakenne on erilainen (taulukko 2). Muuta laidunkarjaa pitävillä tiloilla ja lypsykarjatiloilla suurin kustannuserä oli työkustannus (36% ja 29% kaikista tuotantokustannuksista keskimäärin 2000–2011), siipikarjatiloilla työkustannus oli pienin (9%). Siipikarja- ja sikatiloilla tarvikekustannus oli merkittävin kustannus (43% ja 29%), josta rehut muodostavat suurimman osan kustannuksesta. Myös kasvihuone- ja avomaatiloilla suurimpia kustannuseriä olivat työkustannus (32% ja 33%) ja tarvikekustannus (43% ja 23%).

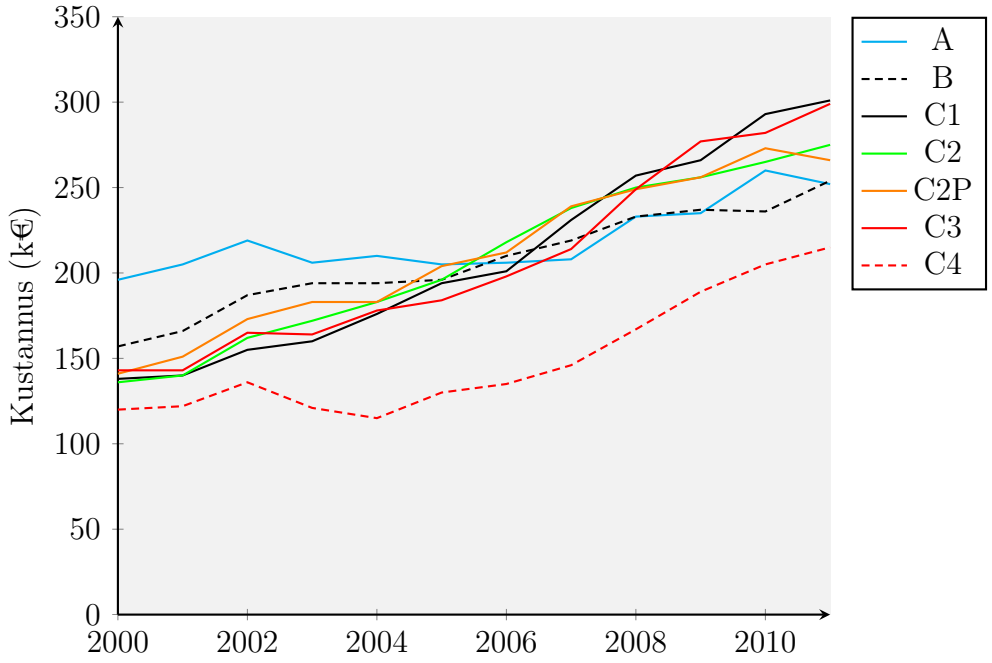
Suomi jakautuu seitsemään eri maataloustukialueeseen (liite 1). Tässä tutkielmassa aluetarkastelut tehdään tukialuejakoon perustuen. Alueittain tarkasteltuna tuotantokustannusten kehityssuunta vuosina 2000–2011 näyttää olevan sama eri alueilla (kuvio 8). Kustannukset ovat nousseet, mutta eri alueilla eri tavalla. Eniten kustannukset ovat kasvaneet C1- (119%), C3- (110%), C2- (102%), C2P- (89%) ja C4-tukialueella (80%). Vähiten kustannukset kasvaneet A- (29%) ja B-tukialueella



Kuvio 7: Keskimääräisten tuotantokustannusten kehitys tuotantosuunnittain 2000–2011.

Taulukko 2: Tuotantokustannusten rakenne keskimäärin tuotantosuunnittain 2000–2011.

Tuotantosuunta	Työ	Tarvike	Kotieläin	Rakennus	Kone	Korko	Muu	Yht.
Lypsykarja	29%	21%	5%	7%	17%	10%	11%	100%
Viljanviljely	17%	21%	0%	5%	21%	18%	18%	100%
Muu kasvinviljely	20%	22%	1%	5%	20%	14%	19%	100%
Muu nautakarja	21%	20%	13%	7%	16%	11%	11%	100%
Sikatalous	15%	29%	19%	7%	11%	10%	8%	100%
Kasvihuone	32%	43%	0%	6%	5%	4%	11%	100%
Sekatuotanto	20%	23%	10%	5%	16%	12%	14%	100%
Avomaa	33%	23%	0%	5%	15%	9%	15%	100%
Muu laidunkarja	36%	13%	4%	7%	16%	11%	14%	100%
Siipikarja	9%	43%	17%	7%	9%	7%	8%	100%



Kuvio 8: Keskimääräisten tuotantokustannusten kehitys tukialueittain 2000–2011.

(62%). Alueellisen eron kustannusten kehityksessä selittää osittain tilan tuotantosuunta. Esimerkiksi C1-, C2-, C2P- ja C3-tukialueilla on suhteellisen paljon lypsykarjatiloja ja A- ja B-tukialueilla on suhteellisen paljon viljanviljely- ja kasvinviljelytiloja. Jatkoanalyysia varten pohjoiset tukialueet C2P, C3 ja C4 on yhdistetty (C2P–C4), koska tilojen määrä jää vähäiseksi. Tilojen määrä tuotantosuunnittain eri tukialueilla on esitetty liitteessä 2.

Tilojen kokonaistuotantokustannuksia pyritään selittämään lineaarisella sekamallilla huomioimalla tilojen väliset erot, kun havaintoyksiköiden ominaisuudet muuttuvat ajanjaksolla 2000–2011. LSM:n rakentamiseen käytetään askeltavaa ”alhaalta ylöspäin” -strategiaa. Mallissa selitettävänä muuttujana (`ind_tuotantokust`) on maatalousyrityksen kokonaistuotantokustannus, joka kuvaa vuoteen 2011 deflatoituja kokonaistuotantokustannuksia tiloilla. Analysoinnissa käytetään SPSS-ohjelmiston `MIXED`-komentoa. Malliin sisällytetään kiinteät (`/FIXED`) ja satunnaisvaikutukset (`/RANDOM`). Mallia (1) lähdetään rakentamaan sisällyttämällä kiinteä vakiotermi (`intercept`) ja satunnaisvaikutukset vakio ja aika-muuttuja (`intercept` ja `year`). Muuttuja `year` on jatkuva muuttuja ja se on laskettu kaavalla $\texttt{year} = \texttt{vuosi} - 2000$ eli vuotta 2000 vastaa arvo 0, vuotta 2001 arvo 1 ja niin edelleen. Satunnaisvaikutukset määritellään havaintoyksikkönä olevan maatilanumeron (`mtnro`) suhteen.

Apukomennolla /RANDOM lasketaan yksilöiden välisiä eroja, kun yksilön ominaisuudet muuttuvat aikajaksolla. Satunnaisvaikutusten kovarianssirakenne määritetään COVTYPE-argumentilla. Tässä tapauksessa rakenteeksi on valittu rakenteeton malli (UN), joka soveltuu pitkittäisaineistoihin. Residuaalille on valittu ensimmäisen asteen autoregressiivinen rakenne (AR1), koska se soveltuu hyvin aineistoon, jossa samasta havaintoyksiköstä on otettu peräkkäisiä havaintoja ja voidaan olettaa, että lähekkäin olevat havainnot korreloivat keskenään voimakkaammin kuin kaukana toisistaan otetut havainnot. Mallissa on mukana 11113 havaintoa ja 1568 eri maatilaa. Mallin tulosten perusteella vakio ($\beta_0 = 157314$, $S.E. = 3679$) on erittäin merkitsevä (p<0,001) ja perusteltu osa mallia. Yleensä mikä tahansa tuotanto edellyttää kiinteää tuotantopanoksista riippumatonta kustannusosaa.

Seuraavaksi malliin ryhdytään lisäämään kiinteiden vaikutusten muuttujia, jotka kuvaavat selitettävän ja selittävien muuttujien välisiä suhteita. Lisätään malliin kiinteäksi vaikutukseksi tuotantosuunta (ts), joka on luonteeltaan luokka-asteikollinen muuttuja. Kuvion 7 ja taulukon 2 perusteella tuotantosuunta näyttää vaikuttavan merkittävästi tuotantokustannusten rakenteeseen. SPSS:n MIXED-komento asettaa kiinteiden vaikutusten liittyvä parametrin korkeimman arvon nollaksi eli muita arvoja verrataan nollaksi asetettuun arvoon. Tuotantosuunnista viljanviljely valitaan testauksen perustaksi, joten mallissa kaikkia muita tuotantosuuntia verrataan viljanviljelytiloihin. Mallin (2) tulosten perusteella tuotantosuuntien välillä on eroja (liite 3). Tuloksista voidaan havaita, että lähes kaikki tuotantosuunnat, lukuunottamatta muuta kasvinviljelyä (p=0,527) ja muuta laidunkarjataloutta harjoittavia tiloja (p=0,338), erosivat merkitsevästi viljanviljelyä harjoittavista tiloista tarkasteltaessa niiden tuotantokustannuksia. Viljanviljelytiloja suuremmat kokonaistuotantokustannukset ovat kasvihuone- ja avomaatuotannossa, lypsykarjatiloilla, sika- ja siipikarjataloudessa ja muuta nautakarjataloutta ja sekamuotoista tuotantoa harjoittavilla tiloilla. Tuotantosuuntien välisiä eroja testataan myös parittain (apukomennolla /EMMEANS, liite 4). Kasvihuonetuotanto eroaa merkitsevästi kaikista muista tuotantosuunnista. Kasvihuonetiloilla kokonaistuotantokustannukset ovat suuremmat kuin muilla tiloilla. Tilastollisesti merkitsevää eroa kokonaistuotantokustannuksissa ei havaita tuotantosuuntapareissa siipikarjatalous ja sikatalous, viljanviljely ja muu kasvinviljely, viljanviljely ja muu laidunkarja, muu kasvinviljely ja muu laidunkarja, avomaa ja muu laidunkarja, avomaa ja sekatuotanto, avomaa ja muu kasvinviljely, avomaa ja muu nautakarjatalous, muu nautakarjatalous ja sekatuotanto, muu nautakarjatalous ja muu laidunkarja, lypsykarjatalous ja muu nautakarjatalous, muu laidunkarja ja sekatuotanto.

Kun malliin (2a) lisätään vielä aika-muuttuja `year` kiinteisiin vaikutuksiin, huomataan, että malli paranee ja tuotantosuuntamuuttujien kertoimet hieman pienenevät eli aika selittää tilojen kokonaistuotantokustannuksia. Aika-muuttuja on myös merkitsevä ($p<0{,}001$). Kustannukset kasvavat ajan myötä. Aikamuuttuja on mallissa sekä kiinteä vaikutus että satunnaisvaikutus.

Lisätään malliin (3) seuraavaksi muuttuja `sokokoluokka_E2` eli taloudellinen kokoluokka, joka on tärkeä tilan ominaisuutta kuvaava tekijä. Lisäksi taloudellista kokoluokkaa käytetään kannattavuuskirjanpitoaineiston painotuksen yhtenä perusteena. Pienin kokoluokka valitaan testauksen perustaksi. Kuten oli odotettua, tilan taloudellinen koko selittää merkitsevästi ($p<0{,}001$) tilan kokonaistuotantokustannuksia. Periaatteena, mitä suurempi on maatilan kokoluokka, sitä enemmän tilalla on myös kustannuksia.

Jatketaan mallin (4) kehittämistä muuttujalla `tukialue5`, joka määrittää tilan sijainnin tukialueittain jaoteltuna ja joka on myös yksi painotuksen peruste. Muuttujassa tukialueet C2P, C3 ja C4 on yhdistetty (C2P–C4) vähäisen havaintomäärän takia. Näin on toimittu myös kannattavuuskirjanpitoaineiston painotuksissa. Tukialueista pohjoisin eli C2P–C4-yhdistelmä valitaan testauksen perustaksi. Havaitaan, että alueella on merkitystä tuotantokustannuksiin. Tukialue A eroaa merkitsevästi kaikista muista alueista (pari A–B $p=0{,}007$, muut parit $p<0{,}001$). Tukialue B eroaa alueesta C1 ($p=0{,}041$) ja C2P–C4 ($p=0{,}006$). Pareittain tarkasteltuna merkitsevää eroa ei havaita pareissa B ja C2 ($p=0{,}100$), C1 ja C2 ($p=0{,}632$), C1 ja C2P–C4 ($p=0{,}235$), C2 ja C2P–C4 ($p=0{,}114$) väliltä.

Lisätään malliin (5) vielä muuttuja `viljelyala`, joka on jatkuva selittävä muuttuja. Viljelyala on tärkeä ominaisuus, joka kuvaa kuinka suuri maatila on kyseessä. Viljelyala on tilastollisesti erittäin merkitsevä ($p<0{,}001$) selittäjä mallissa.

Taulukkoon 4 on kerätty mallien 1–5 ML-estimoinnilla lasketut -2LL–arvot. Reunimmaisella sarakkeella on esitetty χ^2-jakaumaan perustuvan testin p-arvo luvun 3.4 mukaisesti. Pidetään 5% riskitasoa rajana hylätä nollahypoteesi H_0: ”pienempi malli on parempi” virheellisesti. Kun huomioidaan periaate ”mitä pienempi arvo, sen parempi malli”, havaitaan, että selittävien muuttujien lisääminen parantaa mallia.

Malli 5 valitaan lopulliseksi malliksi. Mallin parametrien estimaatit on esitetty taulukossa 3. Kokonaistuotantokustannukset kasvavat mallin perusteella ajan myötä ($p<0{,}001$) ja viljelyalan kasvaessa ($p<0{,}001$). Mitä suuremmasta taloudelliselta kooltaan olevasta tilasta on kyse, sitä suuremmat ovat kokonaistuotantokustannukset ($p<0{,}001$; taulukko 3 ja liite 5).

Taulukko 3: Mallin 5 parametrien estimaatit ja niiden keskivirheet.

Vaikutus	Parametri	Estimaatti	S.E.	p-arvo
Vakio	β_0	60650	9890	<0,001
Tuotantosuunta:				
Muu kasvinviljely	β_1	740	2344	0,752
Kasvihuonetuotanto	β_2	215501	10262	<0,001
Avomaatuotanto	β_3	11899	6070	0,050
Lypsykarjatalous	β_4	33235	3936	<0,001
Muu nautakarjatalous	β_5	27797	4039	<0,001
Muu laidunkarja	β_6	9699	8877	0,275
Sikatalous	β_7	40015	4597	<0,001
Siipikarjatalous	β_8	53169	7402	<0,001
Sekamuotoinen tuotanto	β_9	20273	2931	<0,001
Viljanviljely	β_{10}	0	0	
Tukialue:				
A	β_{11}	36850	11710	0,002
B	β_{12}	19767	10643	0,064
C1	β_{13}	7601	10982	0,489
C2	β_{14}	12086	10638	0,256
C2P–C4	β_{15}	0	0	
Kokoluokka:				
50000–100000	β_{16}	10631	1873	<0,001
100000–	β_{17}	27570	2394	<0,001
0–50000	β_{18}	0	0	
year	β_{19}	6428	473	<0,001
viljelyala	β_{20}	814	47	<0,001
Kovarianssiparametrit:				
UN (1,1)	σ_0^2	$4{,}420\times10^9$	$0{,}990\times10^9$	<0,001
UN (2,1)	$\sigma_0\sigma_{19}$	$0{,}862\times10^9$	59301914	<0,001
UN (2,2)	σ_{19}^2	$0{,}169\times10^9$	12487751	<0,001
Mallivirhe:				
AR1 diagonaalinen	σ^2	$6{,}590\times10^9$	$0{,}789\times10^9$	<0,001
AR1 rho	ρ	0,897	0,012	<0,001
Havainnot		11113		
-2 REML log-likelihood		269615		
AIC		269625		
BIC		269662		

Taulukko 4: Mallien 1–5 ML-estimoidut Likelihood-arvot.

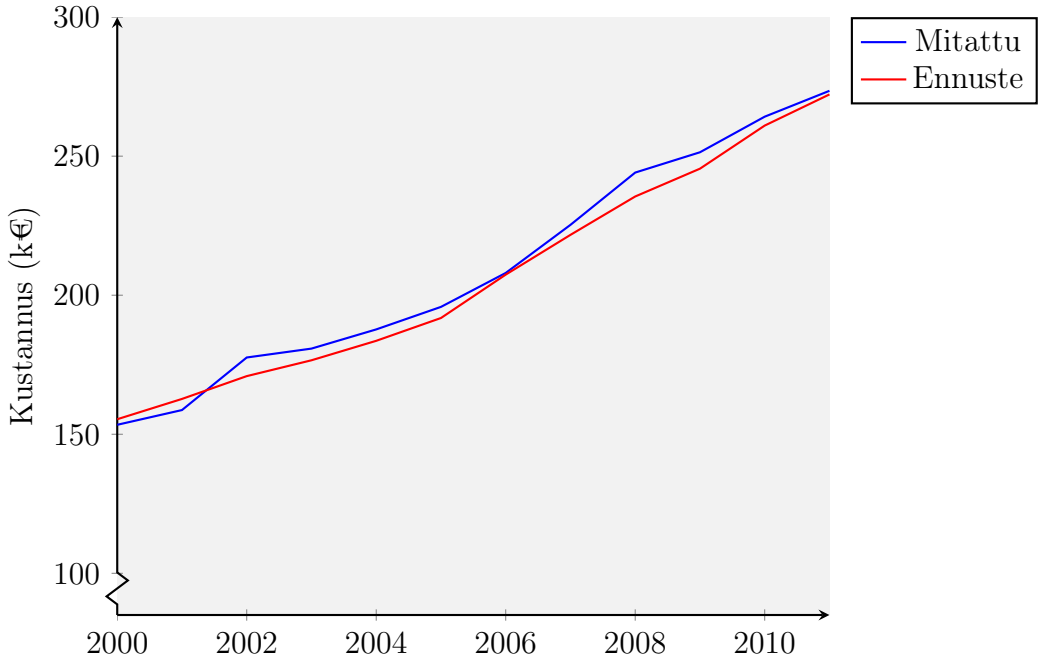
Malli	-2LL	p-arvo
1	270948	–
2	270663	<0,001
2a	270429	<0,001
3	270243	<0,001
4	270219	<0,001
5	269937	<0,001

Selkeimmin tilat erottaa toisistaan niiden tuotantosuunta. Pääsääntöisesti kokonaistuotantokustannukset eroavat merkitsevästi eri tuotantosuuntien välillä ($p<0{,}001$). Viljanviljelytiloilla ja muuta kasvinviljelyä harjoittavilla tiloilla näyttää olevan pienimmät kokonaistuotantokustannukset kuin muissa tuotantosuunnissa. Kasvihuonetuotannossa, siipikarja-, sika- ja lypsykarjataloudessa on suurimmat kokonaistuotantokustannukset. Parittaistestin perusteella (liite 5) merkitsevää eroa ei havaita tuotantosuuntapareissa viljanviljely ja muu kasvinviljely, viljanviljely ja muu laidunkarja, muu kasvinviljely ja avomaatuotanto, muu kasvinviljely ja muu laidunkarja, avomaatuotanto ja muu laidunkarja, avomaatuotanto ja sekatuotanto, lypsykarjatalous ja muu nautakarjatalous, lypsykarjatalous ja sikatalous, muu laidunkarja ja sekatuotanto, sikatalous ja siipikarjatalous.

Tukialueella näyttää olevan vähän merkitystä kokonaistuotantokustannusten muodostumisessa, sillä vain A-tukialue eroaa muista tukialueista ($p=0{,}035$ tai vähemmän). Muut tukialueet eivät parittain vertailun perusteella eroa toisistaan ($p>0{,}064$) (liite 5).

Mallivirheestä voidaan havaita, että tilojen kokonaistuotantokustannusten korrelaatio vuosien välillä on merkittävän suuri (0,897). Vuosi on lyhyt aika tilan aikajänteessä, maatalouden investoinnit ovat suuria, eikä tiloilla yleensä tapahdu nopeita muutoksia tuotannossa. Kovarianssiparametrien kertoimista voidaan tulkita, että kustannukset muuttuvat tiloilla ajan kuluessa eri nopeudella.

Kuviossa 9 on esitetty vertailuna aineistosta mitatut keskimääräiset kokonaistuotantokustannukset ja mallilla 5 ennustetut arvoja. Ennuste näyttää olevan hyvin lähellä mitattuja arvoja. Mallin toimivuuden tarkastelu tuotantosuunnittain ja tukialueittain on esitetty liitteessä 6. Taulukossa 5 ja 6 on esitetty mallin (5) kiinteiden vaikutusten testit ja satunnaisvaikutusten G-matriisi.



Kuvio 9: Keskimääräisten tuotantokustannusten kehitys 2000–2011, mitatut arvot ja mallilla 5 ennustetut arvot.

Taulukko 5: Mallin 5 kiinteiden vaikutusten parametrien tyypin 3 testit.

Kerroin	df_1	df_2	F	p-arvo
vakio	1	2404,0	1065,1	<0,001
ts	9	8174,8	56,9	<0,001
sokokoluokkaE2	2	9521,8	72,8	<0,001
tukialue5	4	1471,4	3,4	0,009
year	1	1290,5	184,6	<0,001
viljelyala	1	10230,3	306,3	<0,001

Taulukko 6: Mallin 5 satunnaisvaikutusten G-matriisi.

	vakio|mtnro	year|mtnro
vakio|mtnro	$4{,}420\times10^9$	$0{,}862\times10^9$
year|mtnro	$0{,}862\times10^9$	$0{,}169\times10^9$

UN-rakenne

Kokonaiskustannusmallissa (kuvio 9) mallilla sovitetut arvot seurasivat hyvin mitattuja arvoja keskimääräisellä tasolla (ero keskimäärin 2,0%, min 0,3%, max 3,8%). Tukialueittain (liite 6, kuvio 12) malli näyttää toimivan melko hyvin (ka 2,9–3,5%, min 0,0–1,1%, max 5,0–9,0%). A- ja B-tukialueilla malli näyttää aliarvioivan kustannukset selvästi vuosina 2002–2004. Malli näyttää toimivan paremmin ajan kuluessa. Tuotantosuunnittain tarkasteltuna mallin toimivuudessa oli selviä eroja (liite 6, kuvio 13 ja 14). Parhaiten malli näytti toimivan tuotantosuunnissa muu nautakarjatalous (ka 2,9%, min 0,3%, max 9,3%), lypsykarjatalous (ka 3,1%, min 1,5%, max 6,1%), sekamuotoinen tuotanto (ka 4,5%, min 0,0%, max 11,3%) ja kasvihuonetuotanto (ka 5,7%, min 2,9%, max 9,3%). Suurimmat poikkeamat olivat tuotantosuunnissa muu laidunkarja (ka 16,5%, min 5,4%, max 32,1%), siipikarjatalous (ka 12,1%, min 0,4%, max 27,1%) ja viljanviljely (ka 21,2%, min 6,5%, max 37,3%). Kun mitatut kustannukset olivat yli 200000 euroa, malli aliarvioi kustannukset ja kun kustannukset olivat alle 130000 euroa, malli yliarvioi kustannukset. Kun kustannukset olivat suuruusluokaltaan 140000–200000 euroa, malli välillä yliarvioi ja välillä aliarvioi kustannuksia. Ero mitattujen ja malleilla sovitettujen arvojen välillä osui keskivirheiden sisään.

Kokonaistuotantokustannusmallin diagnostiikka

Diagnostiikassa käytetään REML-estimoitua mallia 5. Tallennetaan SPSS:n `MIXED`-komennon apukomennolla `/SAVE` ehdolliset ennustetut arvot ja ehdolliset residuaalit. Muunnetaan ehdolliset residuaalit standardoiduiksi arvoiksi. Tarkastellaan kvantiilikuviolla ja histogrammilla (liite 7, kuvio 15) standardoitujen residuaalien normaalisuutta. Residuaalien arvojen pitäisi sijoittua suurin piirtein samalle suoralle. Histogrammista voidaan havaita, että jakauma on huipukkaampi kuin normaalijakauma (huipukkuus 25,5 ja vinous 2,8).

Residuaalien homoskedastisuutta voidaan tutkia sirontakuviolla (liite 7, kuvio 16), jossa ovat standardoidut residuaalit ja sovitetut arvot. Kuvio näyttää melko hyvältä, vaikkakin sovitettujen arvojen kasvaessa residuaalit näyttävät hajaantuvan. Epätavallisien havaintojen olemassaoloa voidaan havainnollistaa sirontakuviolla (liite 7, kuvio 16), jossa on standardoitu residuaali ja havaintonumero. Havainnot, joiden standardoitu residuaali on yli 3, ovat epätavallisia. Havaitaan, että aineistossa on tällaisia havaintoja, mutta niiden lukumäärä (230) suhteessa kokonaisuuteen (11113) on vähäinen (noin 2%). Toisin sanoen 98% standardoiduista residuaaleista osuu välille [–3,3]. Epätavallisten havaintojen vaikutusvaltaisuutta voidaan kokeilla poista-

malla ne mallista. Havaitaan, että epätavallisten havaintojen poistaminen vaikuttaa mallin kertoimiin hyvin vähän.

Tutkitaan vaikutusten lineaarisuutta sirontakuviolla (liite 7, kuvio 17), jossa ovat reunaresiduaali ja selittäjistä muuttujista `year` ja `viljelyala`. Kuviosta voidaan havaita, että aineistossa on muutamia epätavallisia havaintoja, mutta pääsääntöisesti reunaresiduaalit sijoittuvat melko tasaisesti nollan ympärille. Vuotuista hajontaa näyttää olevan hieman enemmän pienimmässä kokoluokassa. Havaintokeskittymä pienissä hehtaarimäärissä voi johtua siitä, että tilan viljelyalasta suurin osa on vuokrattua. Myös kasvihuonetuotannossa on vähän viljelyalaa.

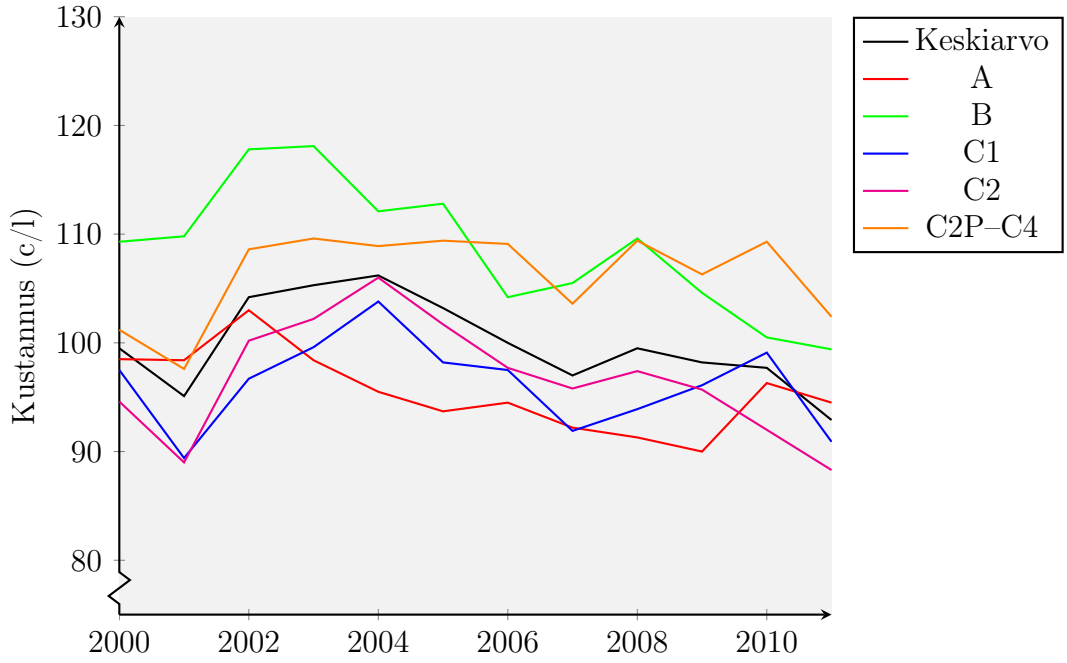
4.3 Yksikkötuotantokustannusmalli

Tässä tutkielmassa tutkitaan kokonaistuotantokustannusmallin lisäksi yksikkötuotantokustannusmallia, jossa selitettävänä muuttujana on kannattavuuskirjanpitotoimintaan osallistuvien lypsykarjatilojen kokonaistuotantokustannus per tuotettu maitolitra. Malli rakennetaan samoilla periaatteilla kuin kokonaistuotantokustannusmalli. Hintatiedot deflatoidaan vuoden 2011 arvoon.

Tuotantokustannus maitolitraa kohti laskettuna (tukia ei ole huomioitu) on ollut tarkasteluvälillä 2000–2011 keskimäärin 99,9 senttiä (keskihajonta 32,3). Kustannus sisältää kaikki tuotannosta aiheutuvat kulut, joten kyseessä ei ole maitolitran yksikkötuotantokustannus. Vuodesta 2000 vuoteen 2011 mennessä yksikkökustannus on pienentynyt noin 7%. Lypsykarjatilojen määrä on pienentynyt suhteessa saman verran, mutta lehmien määrä tiloilla on lisääntynyt.

Kuviosta 10 voidaan havaita, että tuotantokustannukset ovat vaihdelleet eri tukialueilla. A-, C1- ja C2-tukialueilla keskimääräinen tuotantokustannus on ollut lähestulkoon sama (95,5 c/l, 96,2 c/l ja 96,7 c/l, vastaavasti). Tukialueilla B ja C2P–C4 tuotantokustannus on ollut 108,6 c/l ja 106,3 c/l. Kun tarkastellaan tuotantokustannusta maitolitraa kohden tilojen kokoluokkien mukaan jaoteltuna, havaitaan selvä tasoero kustannuksissa. Pienillä tiloilla on selvästi suurempi yksikkökustannustaso ja myös vaihtelu on suurempaa (kuvio 11).

Mallin rakentamiseen käytetään askeltavaa ”alhaalta ylöspäin” -strategiaa. Mallissa selitettävänä muuttujana (`kust_per_maitoL`) on lypsykarjatilojen maidon tuotantokustannus. Mallia (6) lähdetään rakentamaan sisällyttämällä kiinteä vakiotermi (`intercept`), kiinteät jatkuvat vaikutukset aika (`year`) ja lehmien lukumäärä (`lehma`) ja satunnaisvaikutukset vakio (`intercept`) ja aika (`year`). Satunnaisvai-

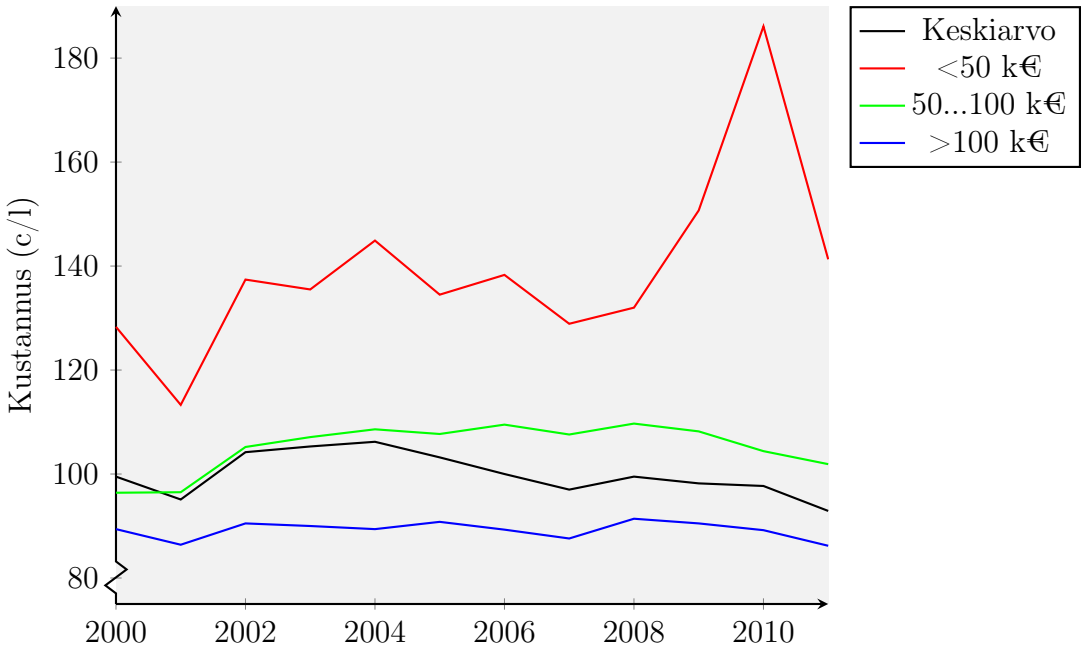


Kuvio 10: Keskimääräisten tuotantokustannusten kehitys maitolitraa kohden 2000–2011 tukialueittain.

kutusten kovarianssirakenteeksi valitaan UN ja residuaalin kovarianssirakenteeksi AR(1). Mallin tulosten perusteella vakio on erittäin merkitsevä ($p<0{,}001$). Kiinteä aikamuuttuja on erittäin merkitsevä ($p<0{,}001$) ja sen kerroin positiivinen (1,525), joten kustannukset kasvavat ajan kuluessa. Lehmien lukumäärää kuvaava kiinteä vaikutus on erittäin merkitsevä ($p<0{,}001$) ja sen kerroin on negatiivinen (-0,84) eli lehmien lukumäärän lisääminen pienentää yksikkötuotantokustannuksta. Tämä on luontevaa, sillä yleensä suuremmasta tuotannosta saadaan mittakaavaetuja ja tehokkuutta.

Lisätään malliin (7) seuraavaksi luokitteleva kokoluokkamuuttuja (`sokokoluokka_E2`) kiinteisiin vaikutuksiin. Parittaisvertailussa havaitaan, että pienet tilat (taloudellinen koko alle 50000 euroa) eroavat merkitsevästi ($p<0{,}001$) muista kokoluokista, mutta keskisuuret ja suuret tilat eivät eroa toisistaan ($p=0{,}136$). Aika- ja lehmämuuttujat ovat edelleen merkitseviä ($p<0{,}001$), mutta kertoimet ovat hieman muuttuneet (1,46 ja -0,71, vastaavasti).

Seuraavaksi lisätään malliin (8) tukialuemuuttuja `tukialue5`. Havaitaan, että tukialue selittää heikosti vaihtelua mallissa. Parittain tarkasteltuna merkitseviä eroja on kuitenkin pareissa B ja C1 ($p=0{,}015$), B ja C2 ($p=0{,}002$), B ja C2P–C4 ($p=0{,}050$)



Kuvio 11: Keskimääräisten tuotantokustannusten kehitys maitolitraa kohden 2000–2011 tilakokoluokittain.

Taulukko 7: Mallien 6–8 ML estimoidut Likelihood-arvot.

Malli	-2LL	p-arvo
6	37584	–
7	37459	<0,001
8	37449	<0,001

(liite 8). Tutkimuksen kannalta tukialue on kuitenkin mielenkiintoinen luokitteleva muuttuja, joten se sisällytetään malliin.

Taulukkoon 7 on kerätty mallien 6–8 ML-estimoinnilla lasketut -2LL Likelihood-arvot. Reunimmaisella sarakkeella on esitetty χ^2-jakaumaan perustuvan testin p-arvo luvun 3.4 mukaisesti. Pidetään 5% riskitasoa rajana hylätä nollahypoteesi H_0: ”pienempi malli on parempi” virheellisesti. Kun otetaan huomioon periaate ”mitä pienempi arvo, sen parempi malli”, voidaan havaita, että selittävien muuttujien lisääminen parantaa mallia.

Valitaan malli 8 lopulliseksi malliksi (taulukko 8, liite 8). Malli 8 sisältää yhteensä 4205 havaintoa ja 633 eri maatilaa vuosilta 2000–2011. Jatkuvat kiinteät vaikutukset ovat erittäin merkitseviä (p<0,001). Aikamuuttujan kerroin (1,48) on kaksi kertaa

lehmämuuttujan arvo (-0,71), mikä voidaan tulkita niin, että tilaa pitäisi kasvattaa vuosittain kahdella lehmällä, jotta voitaisiin kompensoida vuotuinen ajasta johtuva kustannusten kasvu. Tukialueella on vähän merkitystä tuotantokustannusten muodostumisessa. Pienin kokoluokka eroaa muista kokoluokista merkitsevästi. Mallivirheestä voidaan havaita, että tilojen tuotantokustannusten korrelaatio vuosien välillä on suuri (0,492). Kovarianssiparametrien kertoimista voidaan tulkita, että kustannukset muuttuvat tiloilla ajan kuluessa eri nopeudella. Mallin toimivuuden tarkastelu tilakokoluokittain ja tukialueittain on esitetty liitteessä 9.

Taulukossa 9 on esitetty mallin (8) kiinteiden vaikutusten testit ja taulukossa 10 satunnaisvaikutusten G-matriisi.

Yksikkömallissa ennustetut arvot seurasivat melko hyvin mitattuja arvoja (liite 9, kuvio 18 ja 19), kun kyseessä oli vertailu keskisuureen (ero keskimäärin 2,9%, min 0,4%, max 6,9%) ja suureen (ka 1,5%, min 0,4%, max 3,1%) kokoluokkaan. Myös kokonaiskeskiarvoon nähden malli toimi hyvin (ka 2,0%, min 0,1%, max 5,6%). Sen sijaan pienen kokoluokan ennustettavuus oli heikko joinakin vuosina (ka 6,2%, min 0,2%, max 18,1%). Tukialueittain tarkasteltuna sovitetut arvot erosivat mitatuista seuraavasti: A-tukialue (ka 2,7%, min 0,7%, max 7,4%), B-tukialue (ka 2,3%, min 0,3%, max 6,7%), C1-tukialue (ka 2,5%, min 0,4%, max 6,9%), C2-tukialue (ka 2,7%, min 0,2%, max 8,8%) ja C2P–C4-tukialue (ka 2,1%, min 0,2%, max 4,8%). Ero mitattujen ja malleilla sovitettujen arvojen välillä osui keskivirheiden sisään.

Yksikkötuotantokustannusmallin diagnostiikka

Diagnostiikassa käytetään REML-estimoitua mallia 8. Tallennetaan SPSS:n `MIXED`-komennon apukomennolla `/SAVE` ehdolliset ennustetut arvot ja ehdolliset residuaalit. Muunnetaan ehdolliset residuaalit standardoiduiksi arvoiksi. Tarkastellaan kvantiilikuviolla ja histogrammilla (liite 10, kuvio 20) standardoitujen residuaalien normaalisuutta. Residuaalien arvojen pitäisi sijoittua suurin piirtein samalle suoralle. Histogrammista voidaan havaita, että jakauma on huipukkaampi kuin normaalijakauma.

Residuaalien homoskedastisuutta voidaan tutkia sirontakuviolla (liite 10, kuvio 21), jossa ovat standardoidut residuaalit ja sovitetut arvot. Kuvio näyttää melko hyvältä, vaikkakin sovitettujen arvojen kasvaessa residuaalit näyttävät hajaantuvan. Epätavallisien havaintojen olemassaoloa voidaan havainnollistaa sirontakuviolla (liite 10, kuvio 21), jossa on standardoitu residuaali ja havaintonumero. Havainnot,

Taulukko 8: Mallin 8 parametrien estimaatit ja niiden keskivirheet.

Vaikutus	Parametri	Estimaatti	S.E.	p-arvo
Vakio	β_0	134,3	2,92	<0,001
Tukialue:				
A	β_1	-1,86	4,86	0,702
B	β_2	7,08	3,61	0,050
C1	β_3	-1,20	3,18	0,707
C2	β_4	-2,74	2,89	0,343
C2P–C4	β_5	0	0	
Kokoluokka:				
50000–100000	β_6	-19,5	1,77	<0,001
100000–	β_7	-21,4	2,20	<0,001
0–50000	β_8	0	0	
year	β_9	1,48	0,18	<0,001
lehma	β_{10}	-0,71	0,04	<0,001
Kovarianssiparametrit:				
UN (1,1)	σ_0^2	344,6	57,8	<0,001
UN (2,1)	$\sigma_0\sigma_9$	5,09	6,87	0,459
UN (2,2)	σ_9^2	2,79	1,20	<0,020
Mallivirhe:				
AR1 diagonaalinen	σ^2	434,3	27,3	<0,001
AR1 rho	ρ	0,492	0,03	<0,001
Havainnot		4205		
-2 REML log-likelihood		37432		
AIC		37442		
BIC		37474		

Taulukko 9: Mallin 8 kiinteiden vaikutusten parametrien tyypin 3 testit.

Kerroin	df_1	df_2	F	p-arvo
vakio	1	944,6	5174,0	<0,001
year	1	327,6	71,5	<0,001
lehma	1	964,2	253,0	<0,001
sokokoluokkaE2	2	3797,9	62,7	<0,001
tukialue5	4	565,1	2,5	0,039

Taulukko 10: Mallin 8 satunnaisvaikutusten G-matriisi.

	vakio|mtnro	year|mtnro
vakio|mtnro	344,6	5,091
year|mtnro	5,091	2,793

UN-rakenne

joiden standardoitu residuaali on yli 3, ovat epätavallisia. Havaitaan, että aineistossa on tällaisia havaintoja, mutta niiden lukumäärä (46) suhteessa kokonaisuuteen (4205) on vähäinen (noin 1%). Toisin sanoen 99% standardoiduista residuaaleista osuu välille [–3,3]. Kun epätavalliset havainnot poistetaan ja ajetaan malli uudestaan, havaitaan, että mallin kertoimet pysyvät lähestulkoon ennallaan.

Tutkitaan vaikutusten lineaarisuutta sirontakuviolla (liite 9, kuvio 22), jossa ovat reunaresiduaali ja selittäjistä muuttujista `year` ja `lehma`. Kuviosta voidaan havaita, että aineistossa on muutamia epätavallisia havaintoja, mutta pääsääntöisesti reunaresiduaalit sijoittuvat melko tasaisesti nollan ympärille. Vuotuista hajontaa näyttää olevan hieman enemmän pienimmässä kokoluokassa.

4.4 Painot mallissa

Tutkitaan, miten painojen käyttäminen vaikuttaa lopullisten mallien (malli 5 ja 8) tulkintaan ja mallien kertoimiin. Malleihin kokeillaan kolmea eri painotustapaa.

1. Painokertoimet (jatkuva muuttuja `ppmi`) on laskettu jokaiselle tilalle huomioiden yrityksen tuotantosuunta, taloudellinen kokoluokka ja tukialue. Painot on laskettu jokaiselle tilalle ositeindikaattoreiden mukaisesti jokaiselle vuodelle erikseen.

2. Kalibroidaan kohdan 1 painokertoimia siten, että viljelyalan koko otetaan huomioon painokertoimissa (jatkuva muuttuja `paino`). Kalibroinnissa käytetään tietoa koko Suomen viljelyalasta, jonka on kerännyt Tike.

3. Luokitellaan kalibroidut painomuuttujat yhtäsuuriksi luokiksi pieniin (paino alle 22), keskikokoisiin (23–56) ja suuriin (yli 57) (muuttuja `painoluokat`; 1, 2 ja 3).

Lisättäessä kokonaistuotantokustannusmalliin 5 kiinteäksi vaikutukseksi painokertoimia `ppmi`, `paino` tai `painoluokat` havaitaan, että yksikään painokertoimista ei ole merkitsevä selittäjä. Mallin parametrien kertoimet muuttuvat hieman painon myötä, mutta muutos on hyvin pieni, noin prosentin verran tai vähemmän suuntaan tai toiseen.

Yksikkökustannusmallissa 8 painokerroin `ppmi` on hyvin lähellä merkitsevyyden rajaa (p=0,053) ja painoluokka 1 eroaa merkitsevästi (p=0,003) painoluokasta 3. Painokerroin `ppmi` vaikuttaa mallin 8 parametrien kertoimiin. Kokoluokkien kertoimet pienenivät yli 15% ja tukialueiden kertoimet kasvoivat yli 9%. Jatkuvien muuttujien muutos oli vähäistä, aikamuuttuja pieneni 3% ja lehmämuuttujassa ei tapahtunut muutosta. Painoluokkien vaikutus mallin 8 kertoimiin oli hyvin vähäinen, pääasiassa alle prosentin.

5 Johtopäätökset ja yhteenveto

Tutkielman tavoitteena oli tutkia, miten LSM soveltuu tutkimusaineistona käytettyyn mikropaneelimuotoiselle maatalouden kannattavuuskirjanpitoaineistoon, miten LSM pitäisi spesifioida aineistoon ja miten painokertoimien käyttäminen vaikuttaa malliin. Soveltavan osan tavoitteena oli tutkia, miten tuotantokustannukset ovat kehittyneet 2000–2011, onko kirjanpitotilojen tuotantokustannusten kehittymisessä eroja tuotantosuuntien välillä ja onko kirjanpitotilojen tuotantokustannusten kehittymisessä alueellisia eroja.

Tarkasteltava tutkimusaineisto oli suuri, sillä tarkastelussa olivat kaikkien vuosien 2000–2011 maatalouden kannattavuuskirjanpitoon osallistuneiden maatilojen kirjanpitotiedot. Kirjanpitotilojen kokonaistuotantokustannuksia ja lypsykarjatilojen tuotantokustannuksia per tuotettu maitolitra selitettiin lineaarisella sekamallilla. Yhteensä aineistossa oli 11113 havaintoa 1568 eri tilalta ja lypsykarjatilojen osaaineistossa 4205 havaintoa 633 eri tilalta. Tutkimusaineisto oli muodoltaan paneeliaineisto, jossa samasta havaintoyksiköstä oli useita tietoja. Suurin osa yrityksistä oli osallistunut kirjanpitotoimintaan kaikkina tarkasteluvuosina.

Luvussa 2 tarkasteltiin paneeliaineistojen käyttöä, niiden etuja ja rajoituksia. Paneeliaineistojen tärkein etu on niiden tehokkuus muutoksen tutkimisessa. Kannattavuuskirjanpitoaineiston kohdalla haittapuolena voivat olla paneelin valikoitumisharha ja poikkileikkausriippuvuus. Valikoitumisharha on siinä mielessä ilmeinen, koska osallistuminen kannattavuuskirjanpitotoimintaan on tiloille vapaaehtoista, mutta vaatii toisaalta panostuksia. Tutkimuksen kannalta kaikentyyppisiä tiloja ei välttämättä saada rekrytoitua toimintaan mukaan.

Maatalousyritykset eivät ole täysin riippumattomia toisistaan. Esimerkiksi viljelymaan lisääminen ei välttämättä ole mahdollista, koska laajentumishaluinen tila ei välttämättä onnistu hankkimaan lisäalueita tilan ympäristöstä, jos halukkaita myyjiä tai vuokraajia ei ole. Maatalousyritykset voidaan todeta riippuvaisiksi toisistaan myös siinä suhteessa, että ne voivat yhdistyä tai toinen yritys voi ostaa toisen yrityksen toimintoja. Poikkileikkausriippuvuudet voivat johtaa harhaisiin päätelmiin. Tässä tutkielmassa kuitenkin oletettiin maatilojen olevan toisistaan riippumattomia.

Luvussa 4 kuvattiin tutkimusaineiston sisältöä ja esitettiin miten kirjanpitotilojen tuotantokustannukset ovat kehittyneet tarkastelujaksolla. Eri tuotantosuuntien tuotantokustannusrakennetta selvitettiin tutkimusaineiston perusteella ja havaittiin, et-

tä kaikissa tuotantosuunnissa kustannukset ovat kasvaneet, osin tilakoon kasvustakin johtuen, mutta kasvussa on ollut erilaista vaihtelua. Kuvailevien tarkastelujen perusteella pääteltiin, että tuotantosuunta on tärkeä luokitteleva muuttuja sisällytettäväksi lineaariseen sekamalliin. Tukialueen vaikutusta tuotantokustannusten selittäjänä on siinä mielessä mielenkiintoista tutkia, koska tilan maantieteellinen sijainti on yksi peruste maataloustukien suuruudelle. Tällaisella tukiratkaisulla pyritään muun muassa tasoittamaan maantieteellisistä syistä johtuvia tilojen välisiä eroja. Kuvailevan tarkastelun perusteella kustannukset ovat kasvaneet eri tahtia eri tukialueilla. Siispä tukialue katsottiin tärkeäksi luokittelijaksi lineaariseen sekamalliin. Myös standardituotokseen perustuva taloudellinen kokoluokka päätettiin sisällyttää malliin mukaan, koska esimerkiksi EU-tasolla maatiloja on tapana luokitella sen perusteella. Kustannusten ajallisen muutoksen kuvaamiseksi malliin lisättiin jatkuvaksi muuttujaksi aikamuuttuja. Tilan suuruutta parhaiten kuvaavat viljelyalan määrä ja eläinten määrä, joten ne lisättiin malliin jatkuviksi selittäjiksi.

Erilaisia malleja kokeiltiin useita vaihtelemalla kiinteitä selittäjiä kuitenkin pitämällä satunnaismuuttujat ja kovarianssirakenne samana. Satunnaisvaikutuksena pidettiin mallissa vakiotermi ja aikamuuttuja. Satunnaisvaikutusten kovarianssirakenteeksi valittiin rakenteeton (UN) ja jäännösvaihtelun rakenteeksi autoregressiivinen (AR1). Molemmat rakenteet toimivat mallissa hyvin. Lopulta valittiin kaksi lopullista mallia, yksi kokonaistuotantokustannusmalli ja yksi yksikkötuotantokustannusmalli. Selittäjiksi valittiin malleihin luokka-asteikollisia muuttujia: tuotantosuunta, taloudellinen kokoluokka ja tilan maataloustukialuejakoon perustuva sijainti. Mallien jatkuvia selittäjiä olivat aikamuuttuja ja maatilan suuruutta kuvaavat viljelyala ja lehmien lukumäärä. Mallien laatimisen tavoite oli se, että ne olisivat mahdollisimman yksinkertaisia, mutta kuitenkin selittäisivät tuotantokustannuksia riittävän hyvin ottaen huomioon ajallisen muutoksen ja maatilan ominaisuuksia.

Mallien toimivuutta tutkittiin diagnostisin menetelmin käyttäen apuna erilaisia mallin oletuksiin perustuvia graafisia tarkasteluja. Pääpaino oli residuaalien tarkastelussa. Mallien toimivuutta tutkittiin myös tulostamalla samaan kuvioon aineiston mitatut tuotantokustannukset ja mallien avulla sovitettujen reunamallien tulokset. Tarkastelut tehtiin keskiarvotasolla ja luokittelevien selittäjien mukaan. Kokonaiskustannusmallissa (ero keskimäärin 2,0%, min 0,3%, max 3,8%) ja yksikkökustannusmallissa (ka 2,0%, min 0,1%, max 5,6%) mallilla sovitetut arvot seurasivat hyvin mitattuja arvoja keskimääräisellä tasolla. Ero mitattujen ja malleilla sovitettujen arvojen välillä osui keskivirheiden sisään.

Mallien tulosten perusteella jatkuvat muuttujat, aika ja viljelyala kokonaismallissa ja aika ja lehmien lukumäärä yksikkömallissa selittävät hyvin tuotantokustannuksia. Maataloustukialue selittää heikosti tuotantokustannusten muodostumista. Kokonaismallissa vain A-tukialue erottui muista tukialueista ja yksikkömallissa vain B-tukialue erottui pohjoisimmista tukialueista (C1, C2 ja C2P–C4). Maatilan taloudellista kokoa voidaan mitata tilan standardituotoksen avulla. Tutkimusaineiston tilat luokiteltiin näin kolmeen kokoluokkaan. Kokonaismallin kohdalla kokoluokan vaikutus selittäjänä on selkeä, mitä suurempi tila on kyseessä, sitä suuremmat ovat myös kokonaistuotantokustannukset. Sen sijaan yksikkömallissa keskisuuri ja suuri kokoluokka eivät eronneet toisistaan merkitsevästi.

Painojen käytöstä malleissa ei saatu kattavasti tuloksia. Niiden käyttäminen lineaarisissa sekamalleissa on hankalaa muun muassa siksi, että mallin oletusten ja keskivirheiden tulkinta muuttuu vaikeaksi. Painojen merkitystä päätettiin tutkia siten, että painot sisällytettiin malleihin jatkuvana muuttujana ja seurata, miten ne vaikuttavat parametrien kertoimiin ja ovatko ne tilastollisesti merkitseviä muuttujia. Muutokset parametrien kertoimissa olivat pieniä ja pääasiassa painot eivät olleet tilastollisesti merkitseviä.

Tutkielmassa saatiin vastauksia asetettuihin tutkimuskysymyksiin. Jatkotutkimusaiheena olisi mielenkiintoista tutkia painojen sisällyttämisen vaihtoehtoja lineaariseen sekamalliin. Toinen jatkotutkimuksen kohde voisi olla lineaarisen sekamallin hyödyntäminen tulevien vuosien tuotantokustannusten ennustamisessa. Myös muiden tuotantosuuntien kuin lypsykarjatalouden yksikkökustannuksia voisi olla hyödyllistä mallintaa. Lisäksi malleissa voisi kokeilla selittävinä muuttujina sellaisia ominaisuuksia, jotka ovat ulkoisia tekijöitä tiloille. Tällaisia tekijöitä voisivat olla esimerkiksi tuotteiden ja panosten hinnat, pellon hinta ja erilaiset yhdysvaikutukset.

Lähteet

1 *Maatalouden kannattavuustutkimus 75 vuotta.* Numero 53 sarjassa *MTTL:n julkaisuja.* MTTL, Helsinki, 1987.

2 *SAS/STAT 9.3 User's Guide: Mixed Modeling.* SAS Institute, 2011, ISBN 978-1-60764-921-2.

3 Baltagi, Badi H.: *Econometric analysis of panel data.* Wiley, Chichester, 3 painos, 2005, ISBN 978-0-470-01456-1.

4 Baltagi, Badi H. ja James M. Griffin: *A General Index of Technical Change.* The Journal of Political Economy, 96(1):20–41, February 1988.

5 Baltagi, Badi H., James M. Griffin ja Daniel P. Rich: *Airline Deregulation: The Cost Pieces of the Puzzle.* International Economic Review, 36(1):245–258, February 1995.

6 Bertolet, Marianne: *To weight or not to weight? Incorporating sampling designs into model-based analyses.* väitöskirja, Carnegie Mellon University, Pittsburgh, July 2008.

7 Bertrand, Marianne, Esther Duflo ja Sendhil Mullainathan: *How Much Should We Trust Differences-in-Differences Estimates?* The Quarterly Journal of Economics, 119(1):249–275, February 2004.

8 Cornwell, Christopher, Peter Schmidt ja Robin C. Sickles: *Production Frontiers with Cross-sectional and Time-series Variation in Efficiency Levels.* Journal of Econometrics, 46(1–2):185–200, October–November 1990, ISSN 0304-4076.

9 Deaton, Angus: *Data and Econometric Tools for Development Analysis.* Nide 3A sarjassa *Handbook of Development Economics*, luku 33, sivut 1785–1882. Elsevier, 1995, ISBN 978-0-444-82301-4.

10 Demétrio, Clarice G. B.: *An Introduction to Mixed Models.* Luentomoniste, May 2009.

11 Demidenko, Eugene: *Mixed Models: Theory and Applications.* Wiley Series in Probability and Statistics. Wiley, Hoboken, 1 painos, 2004, ISBN 978-0-471-60161-6.

12 Diggle, Peter J., Kung Yee Liang ja Scott L. Zeger: *Analysis of longitudinal data*, nide 13 sarjassa *Oxford statistical science series*. Clarendon Press, Oxford, 1 painos, 1994, ISBN 0198522843.

13 Dumouchel, William H. ja Greg J. Duncan: *Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples*. Journal of the American Statistical Association, 78(383):535–543, 1983. http://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478006.

14 EC: *Data Collection from Agricultural Holdings*, 2010. http://ec.europa.eu/agriculture/rica/collect_en.cfm.

15 EC: *Weighting*, 2010. http://ec.europa.eu/agriculture/rica/methodology3_en.cfm.

16 Gelman, Andrew: *Struggles with survey weighting and regression modeling*. Statistical Science, 22(2):153–164, 2007.

17 Henderson, Charles R: *Estimation of variance and covariance components*. Biometrics, 9(2):226–252, 1953.

18 Hertel, Bradley R.: *Minimizing Error Variance Introduced By Missing Data Routines in Survey Analysis*. Sociological Methods Research, 4(4):459–474, 1976. http://smr.sagepub.com/content/4/4/459.abstract.

19 Hsiao, Cheng: *Analysis of Panel Data*. Numero 34 sarjassa *Econometric Society monographs*. Cambridge University Press, Cambridge, 2 painos, 2003, ISBN 978-0-521-52271-4.

20 Kish, Leslie: *Weighting for unequal P i*. Journal of Official Statistics, 8(2):183–200, 1992.

21 Klevmarken, N. Anders: *Introduction*. European Economic Review, 33(2–3):523–529, March 1989, ISSN 0014-2921.

22 Koop, Gary ja Mark F. J. Steel: *Bayesian Analysis of Stochastic Frontier Models*. Teoksessa Baltagi, Badi H. (toimittaja): *A Companion to Theoretical Econometrics*, Blackwell Companions to Contemporary Economics, luku 24, sivut 520–573. Blackwell, Oxford, 1 painos, 2001, ISBN 0-631-21254-X.

23 Korn, Edward L. ja Barry I. Graubard: *Estimating variance components by using survey data*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65(1):175–190, 2003, ISSN 1467-9868. http://dx.doi.org/10.1111/1467-9868.00379.

24 Kumbhakar, Subal C. ja C. A. Knox Lovell: *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge, 1 painos, 2000, ISBN 978-0-521-48184-7.

25 Laird, Nan M. ja James H. Ware: *Random-Effects Models for Longitudinal Data*. Biometrics, 38(4):963–974, December 1982.

26 Lehtinen, Ilkka: *Liian kalliita jauhoja kassissa*, 2009. http://www.stat.fi/artikkelit/2009/art_2009-09-08_001.html.

27 Littell, Ramon C., George A. Milliken, Walter W. Stroup, D. Wolfinger Russell ja Oliver Schabenberger: *SAS for Mixed Models*. SAS Institute, Cary, 2 painos, 2006, ISBN 978-1-59047-500-3.

28 Little, Roderick J. A. ja Donald B. Rubin: *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, Hoboken, 2 painos, 2002.

29 Lohr, Sharon L.: *Sampling: Design and Analysis*. Brooks/Cole, Boston, 2 painos, 2010.

30 Lu, Hao ja Andrew Gelman: *A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking*. Journal of Official Statistics, 19(2):133–151, 2003.

31 MTT: *Tilakohtaisten tietojen testit*. 2011.

32 MTT: *Taustatiedot*, 2013. https://portal.mtt.fi/portal/page/portal/taloustohtori/kannattavuuskirjanpito/taustatiedot.

33 Neyman, Jerzy: *On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection*. Journal of the Royal Statistical Society, 97(4):558–625, 1934.

34 Niemi, Jyrki ja Jaana Ahlstedt (toimittajat): *Suomen maatalous ja maaseutuelinkeinot 2009*. Numero 109 sarjassa *MTT julkaisuja*. MTT Taloustutkimus, Helsinki, 2009, ISBN 978-951-687-149-6.

35 Niemi, Jyrki ja Jaana Ahlstedt (toimittajat): *Suomen maatalous ja maaseutuelinkeinot 2011*. Numero 111 sarjassa *MTT julkaisuja*. MTT Taloustutkimus, 2011, ISBN 978-951-687-158-8.

36 Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein ja J. Rasbash: *Weighting for unequal selection probabilities in multilevel models*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(1):23–40, 1998, ISSN 1467-9868. http://dx.doi.org/10.1111/1467-9868.00106.

37 Pfeffermann, Danny: *The Role of Sampling Weights When Modeling Survey Data*. International Statistical Review / Revue Internationale de Statistique, 61(2):317–337, 1993, ISSN 03067734.

38 Rabe-Hesketh, Sophia ja Anders Skrondal: *Multilevel modelling of complex survey data*. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(4):805–827, 2006.

39 Rahiala, Markku: *Lineaaristen sekamallien käyttö paneeliaineistojen analysoinnissa*. Luentomoniste, Syksy 2009. http://cc.oulu.fi/~mrahiala/mixed.pdf.

40 Raudenbush, Stephen W. ja Anthony S. Bryk: *Hierarchical Linear Models: Applications and Data Analysis Methods*, nide 1 sarjassa *Advanced Quantitative Techniques in the Social Sciences*. SAGE Publications, 2 painos, 2002.

41 Rubin, Donald B.: *Inference and missing data*. Biometrika, 63(3):581–592, 1976. http://biomet.oxfordjournals.org/content/63/3/581.abstract.

42 Schabenberger, Oliver: *Mixed Model Influence Diagnostics*. Teoksessa *Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference*, numero 189-29, sivu 17, Cary, 2004. SAS Institute.

43 Schafer, Joseph L. ja John W. Graham: *Missing data: Our View of the State of the Art*. Psychological Methods, 7(2):147–177, 2002.

44 Shek, Daniel T. L. ja Cecilia Ma: *Longitudinal Data Analyses Using Linear Mixed Models in SPSS: Concepts, Procedures and Illustrations*. The Scientific World Journal, 11:42–76, 2011.

45 Singer, Judith D.: *Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models*. Journal of Educational and Behavioral Statistics, 23(4):323–355, December 1998.

46 Snijders, Tom A. B. ja Roel J. Bosker: *Multilevel Modeling: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications, 1 painos, 1999.

47 SVT: *Maatilatalouden tuotannon arvo ja rahavirrat*. Teoksessa *Maatilatilastollinen vuosikirja 2011*, luku 9, sivu 197. Tike, Helsinki, 2011.

48 SVT: *Kuluttajahintaindeksi 1995=100*, 2013. http://www.stat.fi/til/khi/.

49 SVT: *Maatalouden tuotantovälineiden ostohintaindeksi. 4. Vuosineljännes 2012.*, 2013. http://www.stat.fi/til/ttohi/2012/04/ttohi_2012_04_2013-02-15_tie_001_fi.html.

50 Swallow, William H. ja John F. Monahan: *Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components*. Technometrics, 26(1):47–57, 1984, ISSN 0040-1706.

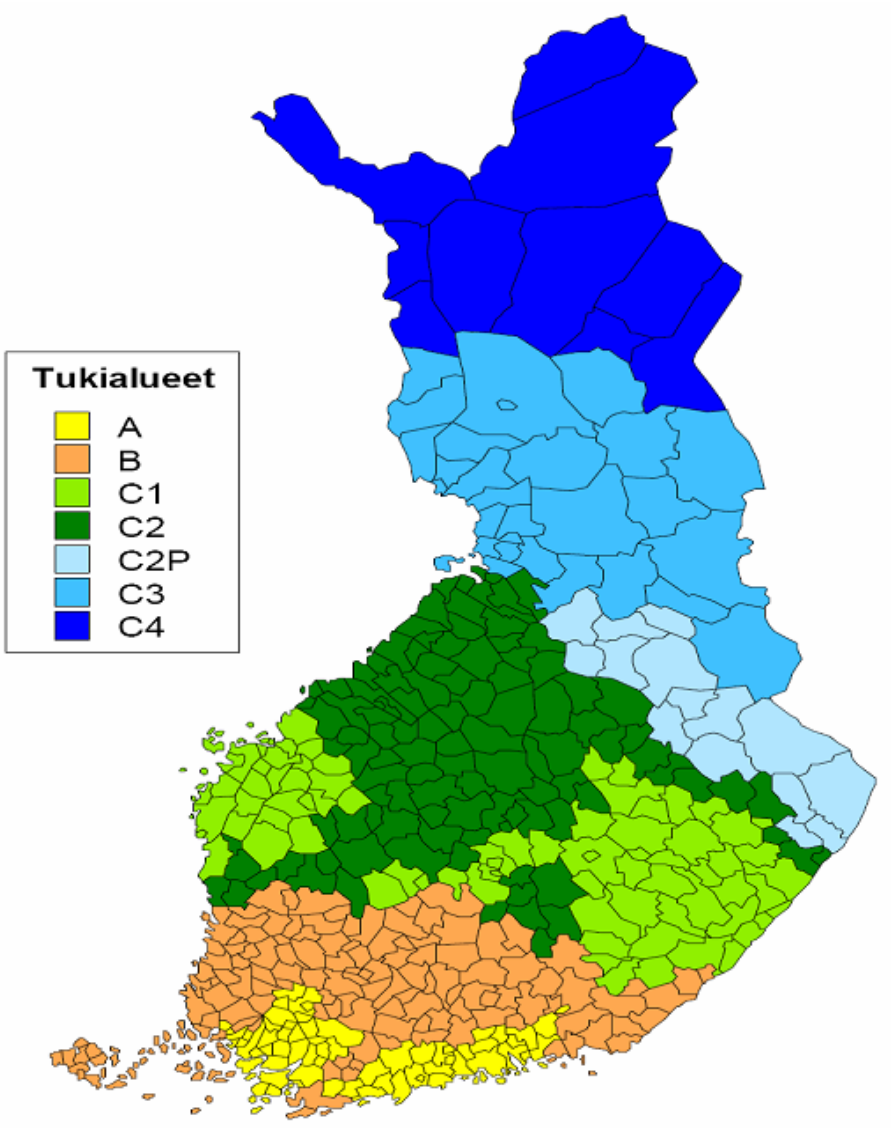
51 Tike: *Rakennetutkimus*, 2013. http://www.maataloustilastot.fi/tiedonkeruu-rakennetutkimus_fi.

52 Verbeke, Geert ja Geert Molenberghs: *Linear Mixed Models for Longitudinal Data*. Springer series in statistics. Springer, New York, 1 painos, 2000, ISBN 978-1-44190-299-3.

53 West, Brady, Kathleen Welch ja Andrzej Galecki: *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman Hall/CRC, Boca Raton, 1 painos, 2006, ISBN 978-1-58488-480-4.

54 Yung, W. ja J. N. K. Rao: *Jackknife Variance Estimation under Imputation for Estimators Using Poststratification Information*. Journal of the American Statistical Association, 95(451):903–915, 2000. http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2000.10474281.

Liite 1. Kartta Suomen tukialueista 2008



Suomi jaetaan seitsemään eri maataloustukialueeseen. Tukialueilla pyritään tasoittamaan olosuhteista johtuvia eroja. Tuen määrä kasvaa etelästä pohjoiseen päin eli tukialueesta A kohti tukialuetta C4. Pääsääntöisesti tukialueiden rajat noudattelevat kuntarajoja, mutta näin ei välttämättä ole aina.

Liite 2. Havaintomäärät eri tukialueilla

Erityyppinen maatalous sijoittuu eri tavoin ympäri Suomea. Kasvinviljely on yleisempää maan eteläosassa, kun taas eläinten pitoon erikoistuneet tilat sijaitsevat pohjoisessa. Kirjanpitotilojen sijoittuminen eri tukialueille tuotantosuunnittain on esitetty taulukossa 11.

Taulukko 11: Kirjanpitotilojen määrä eri tukialueilla tuotantosuunnittain.

Tuotantosuunta / Tukialue	A	B	C1	C2	C2P	C3	C4	Yht.
Muu kasvinviljely	193	454	194	207	25	14	2	1089
Kasvihuonetuotanto	219	229	207	90	24	17	0	786
Avomaatuotanto	26	85	65	42	4	0	0	222
Lypsykarjatalous	247	567	917	1719	213	432	110	4205
Muu nautakarjatalous	72	157	264	371	53	78	0	995
Muu laidunkarja	17	38	39	43	2	40	16	195
Sikatalous	133	374	158	191	3	10	0	869
Siipikarjatalous	3	137	22	2	4	0	0	168
Sekamuotoinen tuotanto	105	318	132	133	26	14	0	728
Viljanviljely	572	744	300	227	8	5	0	1856
Yhteensä	1587	3103	2298	3025	362	610	128	11113

Liite 3. Malli 2

Malli 2 ajettiin SPSS-ohjelmiston MIXED-komennolla. Tässä liitteessä on esitetty mallin syntaksi ja mallin tulokset.

```
* Malli 2.

MIXED ind_tuotantokust BY ts WITH year
/CRITERIA = CIN(95) MXITER(100) MXSTEP(5) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0,
ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED = ts | SSTYPE(3)
/METHOD = REML
/PRINT = R G SOLUTION TESTCOV
/RANDOM intercept year | SUBJECT(mtnro) COVTYPE(UN)
/REPEATED = year | SUBJECT(mtnro) COVTYPE(AR1)
/EMMEANS=TABLES(ts) COMPARE ADJ(LSD).
```

Taulukko 12: Mallin 2 parametrien estimaatit ja niiden keskivirheet.

Vaikutus	Parametri	Estimaatti	S.E.	p-arvo
Vakio	β_0	123345	4264	<0,001
Tuotantosuunta:				
Muu kasvinviljely	β_1	1486	2350	0,527
Kasvihuonetuotanto	β_2	175977	11024	<0,001
Avomaatuotanto	β_3	12193	6137	0,047
Lypsykarjatalous	β_4	28286	3897	<0,001
Muu nautakarjatalous	β_5	24583	4025	<0,001
Muu laidunkarja	β_6	8641	9021	0,338
Sikatalous	β_7	40963	4592	<0,001
Siipikarjatalous	β_8	48343	7452	<0,001
Sekamuotoinen tuotanto	β_9	20166	2921	<0,001
Viljanviljely	β_{10}	0	0	
Kovarianssiparametrit:				
UN (1,1)	σ_0^2	$5{,}985\times10^9$	$1{,}144\times10^9$	<0,001
UN (2,1)	$\sigma_0\sigma_9$	$1{,}216\times10^9$	$0{,}082\times10^9$	<0,001
UN (2,2)	σ_9^2	$0{,}248\times10^9$	$0{,}016\times10^9$	<0,001
Mallivirhe:				
AR1 diagonaalinen	σ^2	$8{,}71\times10^9$	$0{,}893\times10^9$	<0,001
AR1 rho	ρ	0,922	0,008	<0,001
Havainnot		11113		
-2 REML log-likelihood		270477		
AIC		270487		
BIC		270524		

Liite 4. Mallin 2 parittaiset testit

Taulukko 13: Mallin 2 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Muu kasvinviljely	Kasvihuonetuotanto	-174491	11053	<0,001
	Avomaatuotanto	-10707	6162	0,082
	Lypsykarjatalous	-26801	3934	<0,001
	Muu nautakarjatalous	-23097	4063	<0,001
	Muu laidunkarja	-7155	8998	0,427
	Sikatalous	-39477	4677	<0,001
	Siipikarjatalous	-46857	7500	<0,001
	Sekamuotoinen tuotanto	-18680	3012	<0,001
	Viljanviljely	1486	2350	0,527
Kasvihuonetuotanto	Muu kasvinviljely	174491	11053	<0,001
	Avomaatuotanto	163783	11974	<0,001
	Lypsykarjatalous	147690	10931	<0,001
	Muu nautakarjatalous	151394	11046	<0,001
	Muu laidunkarja	167336	13768	<0,001
	Sikatalous	135014	11367	<0,001
	Siipikarjatalous	127634	12531	<0,001
	Sekamuotoinen tuotanto	155811	10881	<0,001
	Viljanviljely	175977	11024	<0,001
Avomaatuotanto	Muu kasvinviljely	10707	6162	0,082
	Kasvihuonetuotanto	-163784	11974	<0,001
	Lypsykarjatalous	-16093	6573	0,014
	Muu nautakarjatalous	-12390	6637	0,062
	Muu laidunkarja	3552	10436	0,733
	Sikatalous	-28770	6951	<0,001
	Siipikarjatalous	-36150	9059	<0,001
	Sekamuotoinen tuotanto	-7973	5900	0,177
	Viljanviljely	12193	6137	0,047

Taulukko 13: Mallin 2 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Lypsykarjatalous	Muu kasvinviljely	26801	3934	<0,001
	Kasvihuonetuotanto	-147690	10931	<0,001
	Avomaatuotanto	16093	6573	0,014
	Muu nautakarjatalous	3703	2865	0,196
	Muu laidunkarja	19645	9112	0,031
	Sikatalous	-12677	4919	0,010
	Siipikarjatalous	-20057	7641	0,009
	Sekamuotoinen tuotanto	8120	3412	0,017
	Viljanviljely	28287	3897	<0,001
Muu nautakarjatalous	Muu kasvinviljely	23097	4063	<0,001
	Kasvihuonetuotanto	-151394	11046	<0,001
	Avomaatuotanto	12390	6637	0,062
	Lypsykarjatalous	-3703	2865	0,196
	Muu laidunkarja	15942	9077	0,079
	Sikatalous	-16380	4991	0,001
	Siipikarjatalous	-23760	7678	0,002
	Sekamuotoinen tuotanto	4417	3450	0,201
	Viljanviljely	24583	4025	<0,001
Muu laidunkarja	Muu kasvinviljely	7155	8998	0,427
	Kasvihuonetuotanto	-167336	13768	<0,001
	Avomaatuotanto	-3552	10436	0,733
	Lypsykarjatalous	-19645	9112	0,031
	Muu nautakarjatalous	-15942	9077	0,079
	Sikatalous	-32322	9472	0,001
	Siipikarjatalous	-39702	11115	<0,001
	Sekamuotoinen tuotanto	-11525	8721	0,186
	Viljanviljely	8641	9021	0,338

Taulukko 13: Mallin 2 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Sikatalous	Muu kasvinviljely	39477	4677	<0,001
	Kasvihuonetuotanto	-135014	11367	<0,001
	Avomaatuotanto	28770	6951	<0,001
	Lypsykarjatalous	12677	4919	0,010
	Muu nautakarjatalous	16380	4991	0,001
	Muu laidunkarja	32322	9472	0,001
	Siipikarjatalous	-7380	7900	0,350
	Sekamuotoinen tuotanto	20797	3846	<0,001
	Viljanviljely	40963	4592	<0,001
Siipikarjatalous	Muu kasvinviljely	46857	7500	<0,001
	Kasvihuonetuotanto	-127634	12531	<0,001
	Avomaatuotanto	36150	9059	<0,001
	Lypsykarjatalous	20057	7641	0,009
	Muu nautakarjatalous	23760	7678	0,002
	Muu laidunkarja	39702	11115	<0,001
	Sikatalous	7380	7900	0,350
	Sekamuotoinen tuotanto	28177	6941	<0,001
	Viljanviljely	48343	7452	<0,001
Sekamuotoinen tuotanto	Muu kasvinviljely	18680	3012	<0,001
	Kasvihuonetuotanto	-155811	10881	<0,001
	Avomaatuotanto	7973	5900	0,177
	Lypsykarjatalous	-8120	3412	0,017
	Muu nautakarjatalous	-4417	3450	0,201
	Muu laidunkarja	11525	8721	0,186
	Sikatalous	-20797	3846	<0,001
	Siipikarjatalous	-28177	6941	<0,001
	Viljanviljely	20166	2921	<0,001

Taulukko 13: Mallin 2 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Viljanviljely	Muu kasvinviljely	-1486	2350	0,527
	Kasvihuonetuotanto	-175977	11024	<0,001
	Avomaatuotanto	-12193	6137	0,047
	Lypsykarjatalous	-28287	3897	<0,001
	Muu nautakarjatalous	-24583	4025	<0,001
	Muu laidunkarja	-8641	9021	0,338
	Sikatalous	-40963	4592	<0,001
	Siipikarjatalous	-48343	7452	<0,001
	Sekamuotoinen tuotanto	-20166	2921	<0,001

Liite 5. Mallin 5 parittaiset testit

Malli 5 ajettiin SPSS-ohjelmiston MIXED-komennolla. Tässä liitteessä on esitetty mallin syntaksi ja mallin parittaiset testit tuotantosuunnittain, kokoluokittain ja tukialueittain.

```
* Malli 5.

MIXED ind_tuotantokust
BY ts sokokoluokka_E2 tukialue5 WITH year viljelyala
/CRITERIA = CIN(95) MXITER(100) MXSTEP(5) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0, ABSOLUTE) LCONVERGE(0,
ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
/FIXED = ts sokokoluokka_E2 tukialue5 year viljelyala | SSTYPE(3)
/METHOD = REML
/PRINT = R G SOLUTION TESTCOV
/RANDOM intercept year | SUBJECT(mtnro) COVTYPE(UN)
/REPEATED = year | SUBJECT(mtnro) COVTYPE(AR1)
/EMMEANS=TABLES(ts) COMPARE ADJ(LSD)
/EMMEANS=TABLES(sokokoluokka_E2) COMPARE ADJ(LSD)
/EMMEANS=TABLES(tukialue5) COMPARE ADJ(LSD).
```

Taulukko 14: Mallin 5 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Muu kasvinviljely	Kasvihuonetuotanto	-214761	10294	<0,001
	Avomaatuotanto	-11159	6090	0,067
	Lypsykarjatalous	-32495	3957	<0,001
	Muu nautakarjatalous	-27057	4064	<0,001
	Muu laidunkarja	-8960	8854	0,312
	Sikatalous	-39276	4667	<0,001
	Siipikarjatalous	-52429	7444	<0,001
	Sekatuotanto	-19533	3011	<0,001
	Viljanviljely	740	2344	0,752

Taulukko 14: Mallin 5 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Kasvihuonetuotanto	Muu kasvinviljely	214761	10294	<0,001
	Avomaatuotanto	203601	11243	<0,001
	Lypsykarjatalous	182265	10228	<0,001
	Muu nautakarjatalous	187703	10338	<0,001
	Muu laidunkarja	205801	13092	<0,001
	Sikatalous	175485	10625	<0,001
	Siipikarjatalous	162332	11872	<0,001
	Sekatuotanto	195228	10139	<0,001
	Viljanviljely	215501	10265	<0,001
Avomaatuotanto	Muu kasvinviljely	11159	6090	0,067
	Kasvihuonetuotanto	-203601	11243	<0,001
	Lypsykarjatalous	-21336	6488	0,001
	Muu nautakarjatalous	-15897	6550	0,015
	Muu laidunkarja	2200	10275	0,830
	Sikatalous	-28116	6846	<0,001
	Siipikarjatalous	-41270	8949	<0,001
	Sekatuotanto	-8374	5820	0,150
	Viljanviljely	11899	6070	0,050
Lypsykarjatalous	Muu kasvinviljely	32495	3957	<0,001
	Kasvihuonetuotanto	-182265	10228	<0,001
	Avomaatuotanto	21336	6488	0,001
	Muu nautakarjatalous	5439	2847	0,056
	Muu laidunkarja	23536	8964	0,009
	Sikatalous	-6780	4864	0,163
	Siipikarjatalous	-19934	7579	0,009
	Sekatuotanto	12963	3407	<0,001
	Viljanviljely	33235	3936	<0,001

Taulukko 14: Mallin 5 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Muu nautakarjatalous	Muu kasvinviljely	27057	4064	<0,001
	Kasvihuonetuotanto	-187704	10338	<0,001
	Avomaatuotanto	15897	6550	0,015
	Lypsykarjatalous	-5439	2847	0,056
	Muu laidunkarja	18097	8931	0,043
	Sikatalous	-12219	4943	0,013
	Siipikarjatalous	-25372	7615	0,001
	Sekatuotanto	7524	3438	0,029
	Viljanviljely	27797	4039	<0,001
Muu laidunkarja	Muu kasvinviljely	8960	8854	0,312
	Kasvihuonetuotanto	-205801	13092	<0,001
	Avomaatuotanto	-2200	10275	0,830
	Lypsykarjatalous	-23536	8964	0,009
	Muu nautakarjatalous	-18097	8931	0,043
	Sikatalous	-30316	9348	0,001
	Siipikarjatalous	-43470	10976	<0,001
	Sekatuotanto	-10573	8586	0,218
	Viljanviljely	9700	8877	0,275
Sikatalous	Muu kasvinviljely	39276	4667	<0,001
	Kasvihuonetuotanto	-175485	10625	<0,001
	Avomaatuotanto	28116	6846	<0,001
	Lypsykarjatalous	6780	4864	0,163
	Muu nautakarjatalous	12219	4943	0,013
	Muu laidunkarja	30316	9348	0,001
	Siipikarjatalous	-13154	7811	0,092
	Sekatuotanto	19743	3822	<0,001
	Viljanviljely	40015	4597	<0,001

Taulukko 14: Mallin 5 tuotantosuuntamuuttujan parittaiset testit.

(I) Tuotantosuunta	(J) Tuotantosuunta	Erotus (I-J)	S.E.	p-arvo
Siipikarjatalous	Muu kasvinviljely	52429	7444	<0,001
	Kasvihuonetuotanto	-162332	11872	<0,001
	Avomaatuotanto	41270	8949	<0,001
	Lypsykarjatalous	19935	7579	0,009
	Muu nautakarjatalous	25372	7615	0,001
	Muu laidunkarja	43470	10976	<0,001
	Sikatalous	13154	7811	0,092
	Sekatuotanto	32896	6884	<0,001
	Viljanviljely	53169	7402	<0,001
Sekatuotanto	Muu kasvinviljely	19533	3011	<0,001
	Kasvihuonetuotanto	-195228	10139	<0,001
	Avomaatuotanto	8374	5820	0,151
	Lypsykarjatalous	-12963	3407	<0,001
	Muu nautakarjatalous	-7524	3438	0,029
	Muu laidunkarja	10573	8586	0,218
	Sikatalous	-19743	3822	<0,001
	Siipikarjatalous	-32896	6884	<0,001
	Viljanviljely	20273	2931	<0,001
Viljanviljely	Muu kasvinviljely	-740	2344	0,752
	Kasvihuonetuotanto	-215501	10262	<0,001
	Avomaatuotanto	-11899	6070	0,050
	Lypsykarjatalous	-33235	3936	<0,001
	Muu nautakarjatalous	-27797	4039	<0,001
	Muu laidunkarja	-9699	8877	0,275
	Sikatalous	-40015	4597	<0,001
	Siipikarjatalous	-53169	7402	<0,001
	Sekatuotanto	-20273	2931	<0,001

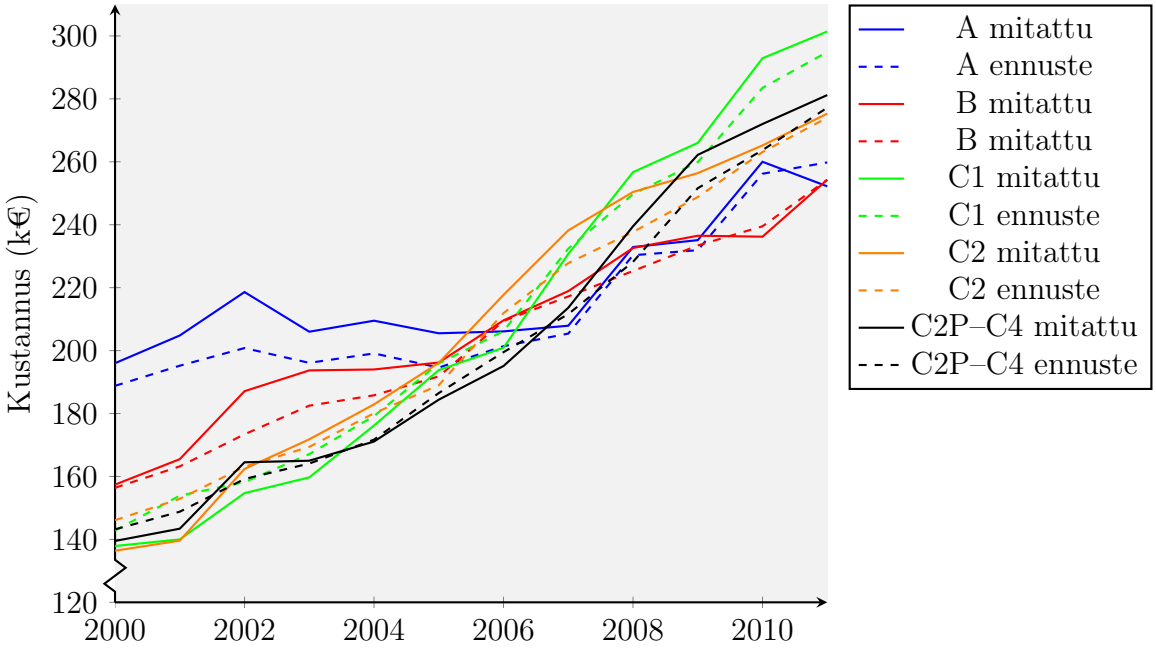
Taulukko 15: Mallin 5 kokoluokkamuuttujan parittaiset testit.

(I) Kokoluokka	(J) Kokoluokka	Erotus (I-J)	S.E.	p-arvo
0–50000	50000–100000	-10631	1873	<0,001
	100000–	-27570	2394	<0,001
50000–100000	100000–	-16939	1688	<0,001
	0–50000	10631	1873	<0,001
100000–	50000–100000	16939	1688	<0,001
	0–50000	27570	2394	<0,001

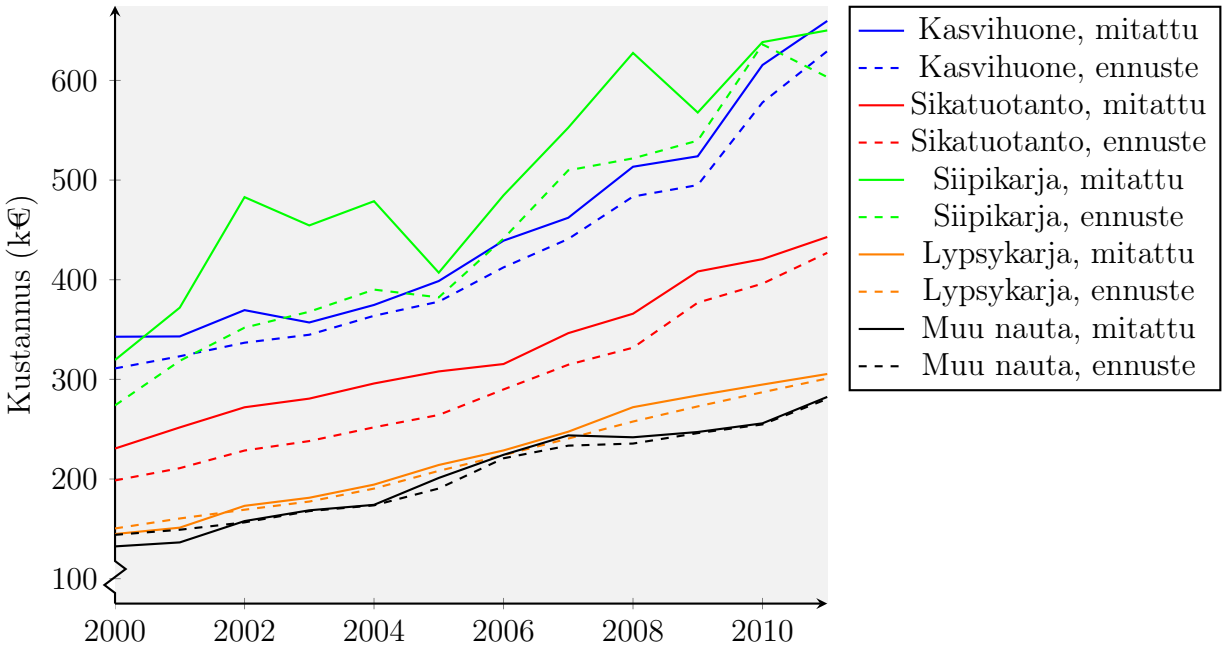
Taulukko 16: Mallin 5 tukialuemuuttujan parittaiset testit.

(I) Tukialue	(J) Tukialue	Erotus (I-J)	S.E.	p-arvo
A	B	17082	8084	0,035
	C1	29248	9347	0,002
	C2	24763	9117	0,007
	C2P–C4	36850	11710	0,002
B	A	-17082	8084	0,035
	C1	12166	8032	0,130
	C2	7681	7710	0,319
	C2P–C4	19767	10643	0,064
C1	A	-29248	9347	0,002
	B	-12166	8032	0,130
	C2	-4484	7931	0,572
	C2P–C4	7601	10982	0,489
C2	A	-24764	9117	0,007
	B	-7681	7710	0,319
	C1	4484	7931	0,572
	C2P–C4	12086	10638	0,256
C2P–C4	A	-36850	11710	0,002
	B	-19767	10643	0,064
	C1	-7601	10982	0,489
	C2	-12086	10638	0,256

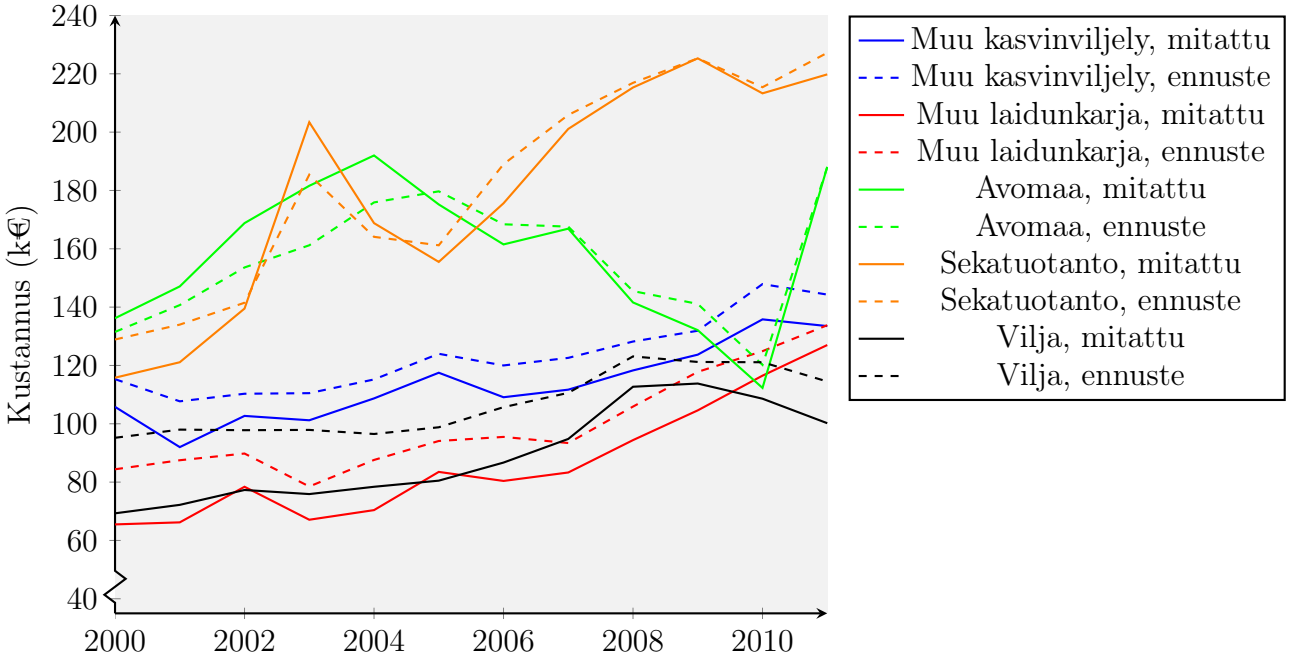
Liite 6. Mallin 5 toimivuus



Kuvio 12: Tuotantokustannusten kehitys 2000–2011 tukialueittain, mitatut arvot ja mallilla 5 ennustetut arvot.

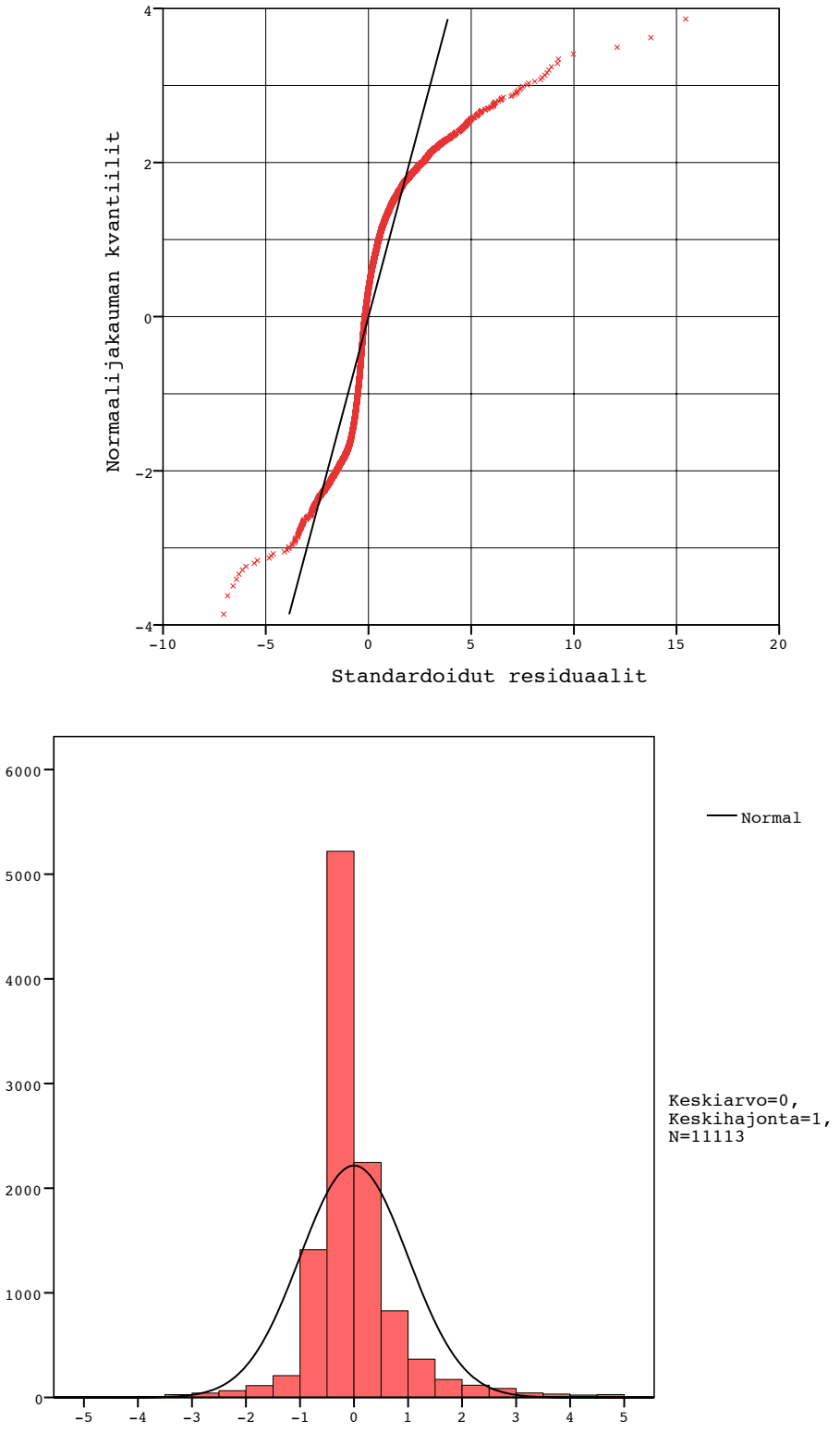


Kuvio 13: Tuotantokustannusten kehitys 2000–2011 tuotantosuunnittain, mitatut arvot ja mallilla 5 ennustetut arvot.

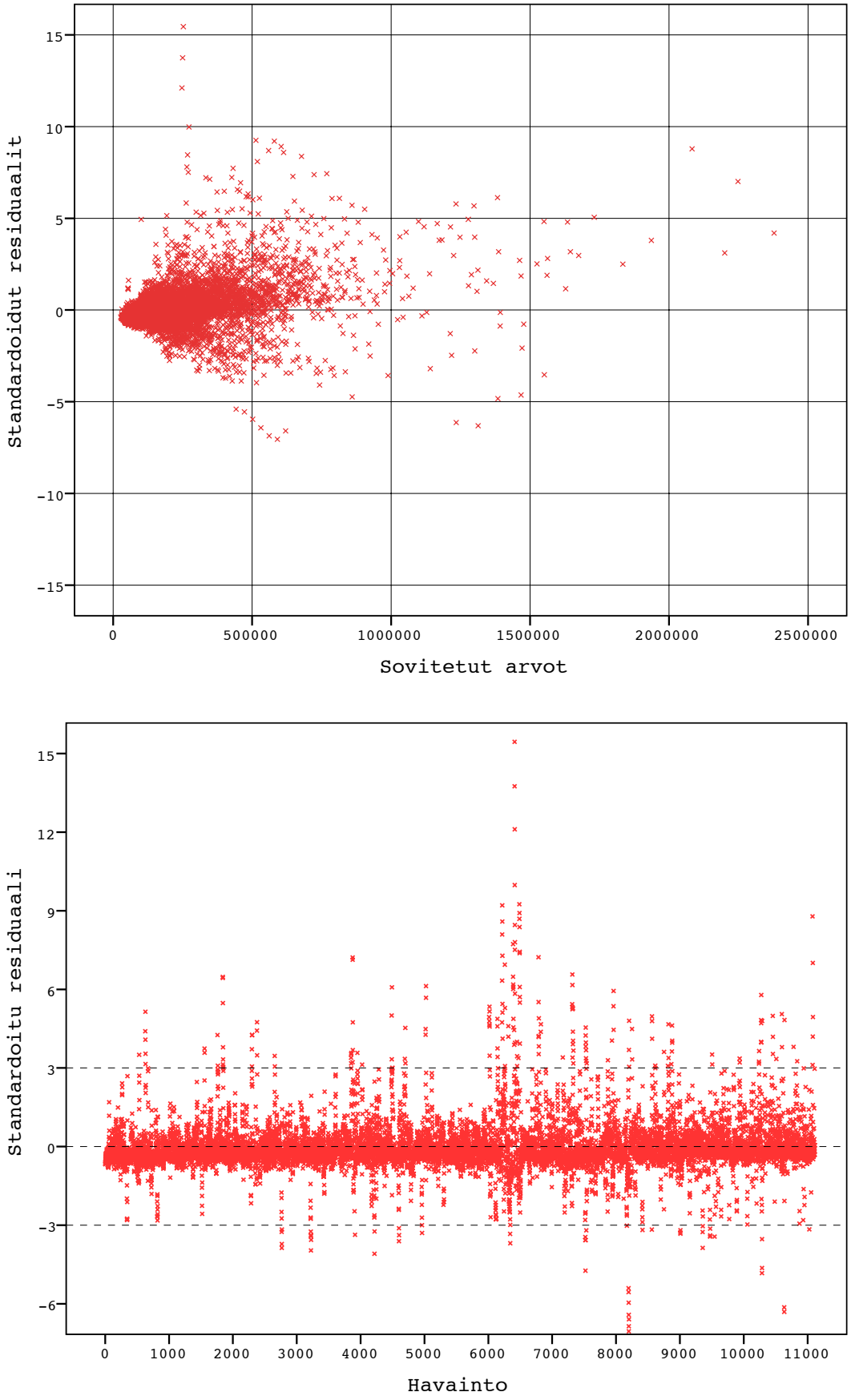


Kuvio 14: Tuotantokustannusten kehitys 2000–2011 tuotantosuunnittain, mitatut arvot ja mallilla 5 ennustetut arvot.

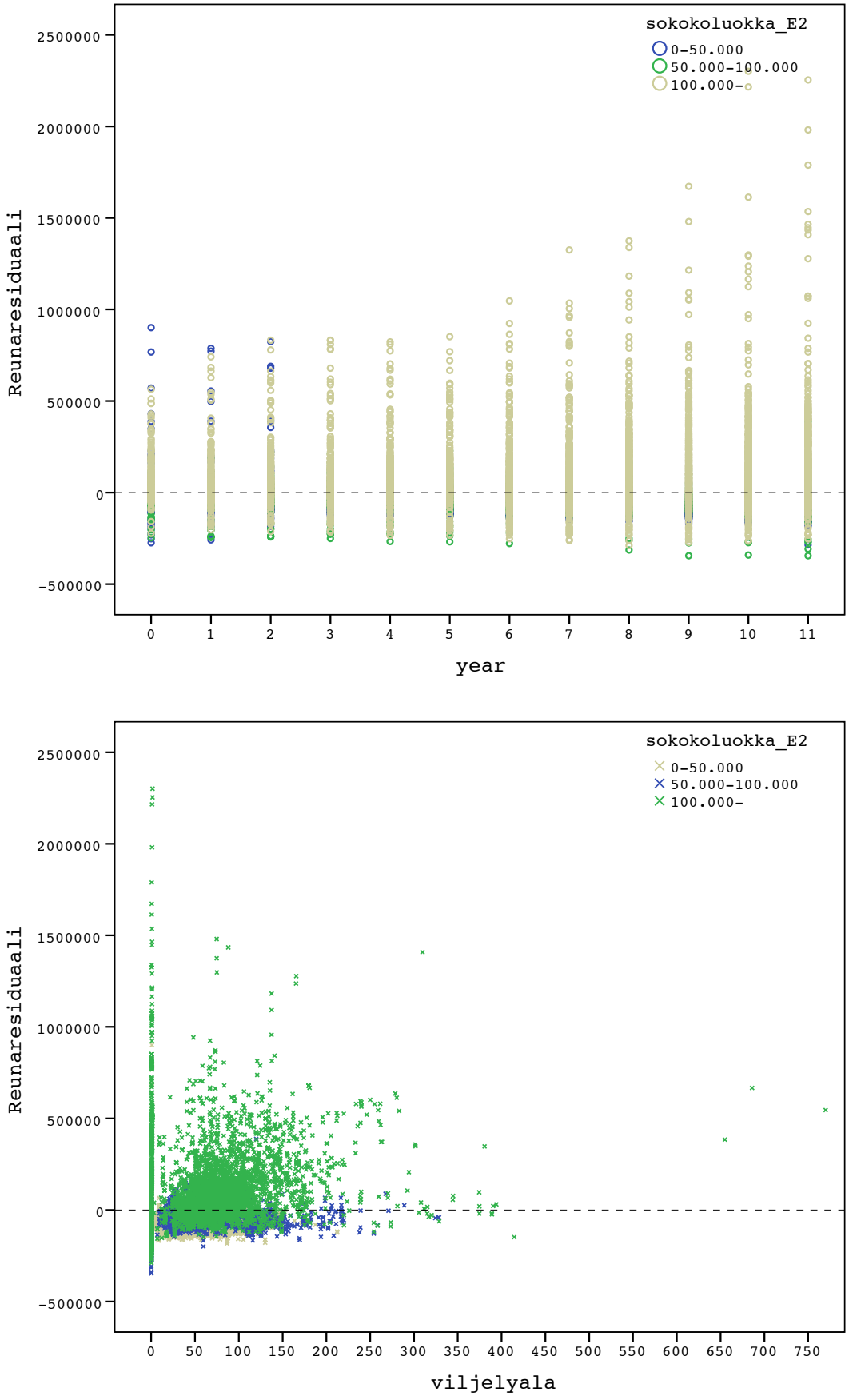
Liite 7. Mallin 5 diagnostiikka



Kuvio 15: Residuaalien kvantiilikuvio ja histogrammi.



Kuvio 16: Residuaalien sirontakuvio.



Kuvio 17: Reunaresiduaalin ja selittävän muuttujan sirontakuvio.

Liite 8. Mallin 8 parittaiset testit

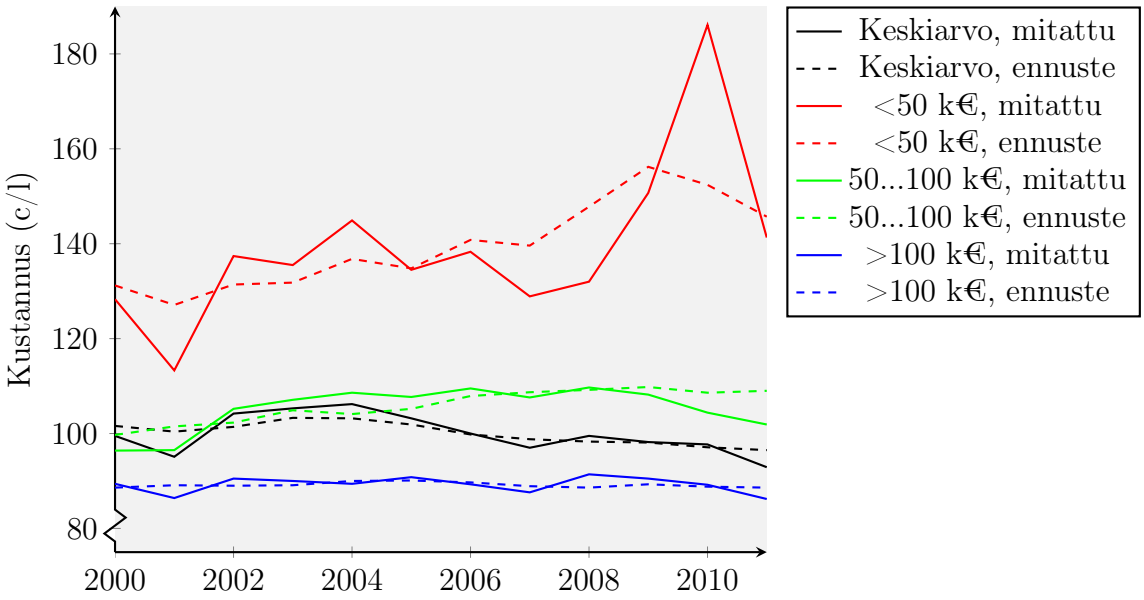
Taulukko 17: Mallin 8 kokoluokkamuuttujan parittaiset testit.

(I) Kokoluokka	(J) Kokoluokka	Erotus (I-J)	S.E.	p-arvo
0–50000	50000–100000	19,5	1,77	<0,001
	100000–	21,4	2,20	<0,001
50000–100000	100000–	1,95	1,39	0,161
	0–50000	-21,4	1,77	<0,001
100000–	50000–100000	-1,95	1,39	0,161
	0–50000	-21,4	2,20	<0,001

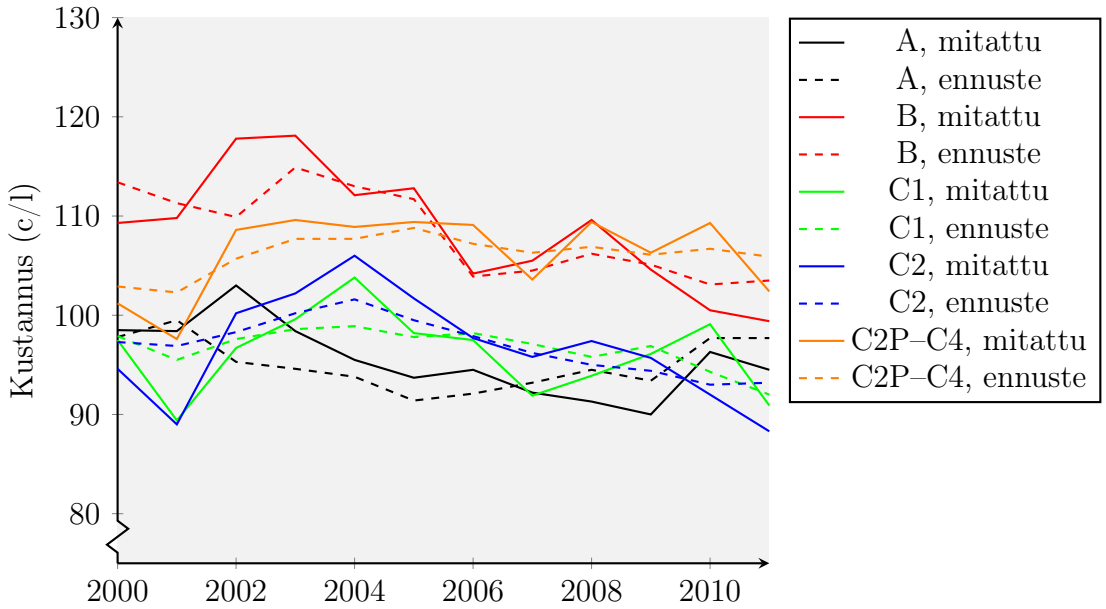
Taulukko 18: Mallin 8 tukialuemuuttujan parittaiset testit.

(I) Tukialue	(J) Tukialue	Erotus (I-J)	S.E.	p-arvo
A	B	-8,94	5,01	0,075
	C1	-0,67	4,70	0,887
	C2	0,88	4,51	0,845
	C2P–C4	-1,87	4,87	0,702
B	A	8,94	5,01	0,075
	C1	8,28	3,40	0,015
	C2	9,82	3,13	0,002
	C2P–C4	7,08	3,61	0,050
C1	A	0,67	4,70	0,887
	B	-8,28	3,40	0,015
	C2	1,55	2,59	0,551
	C2P–C4	-1,20	3,18	0,707
C2	A	-0,89	4,51	0,845
	B	-9,82	3,13	0,002
	C1	-1,55	2,59	0,551
	C2P–C4	-2,75	2,89	0,343
C2P–C4	A	1,87	4,87	0,702
	B	-7,08	3,61	0,050
	C1	1,20	3,18	0,707
	C2	2,75	2,89	0,343

Liite 9. Mallin 8 toimivuus

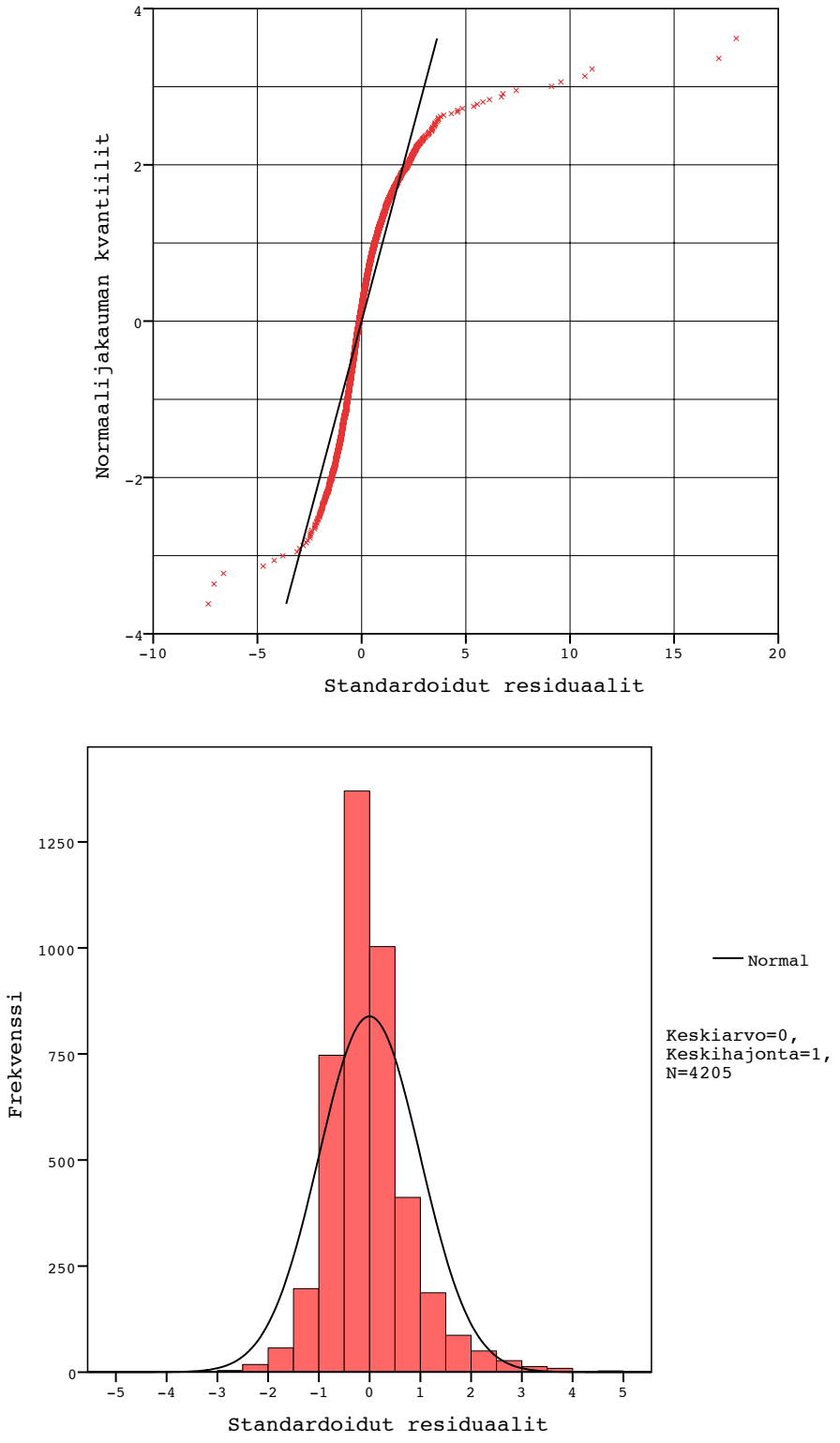


Kuvio 18: Tuotantokustannusten kehitys maitolitraa kohden 2000–2011 tilakokoluokittain, mitatut arvot ja mallilla 8 ennustetut arvot.

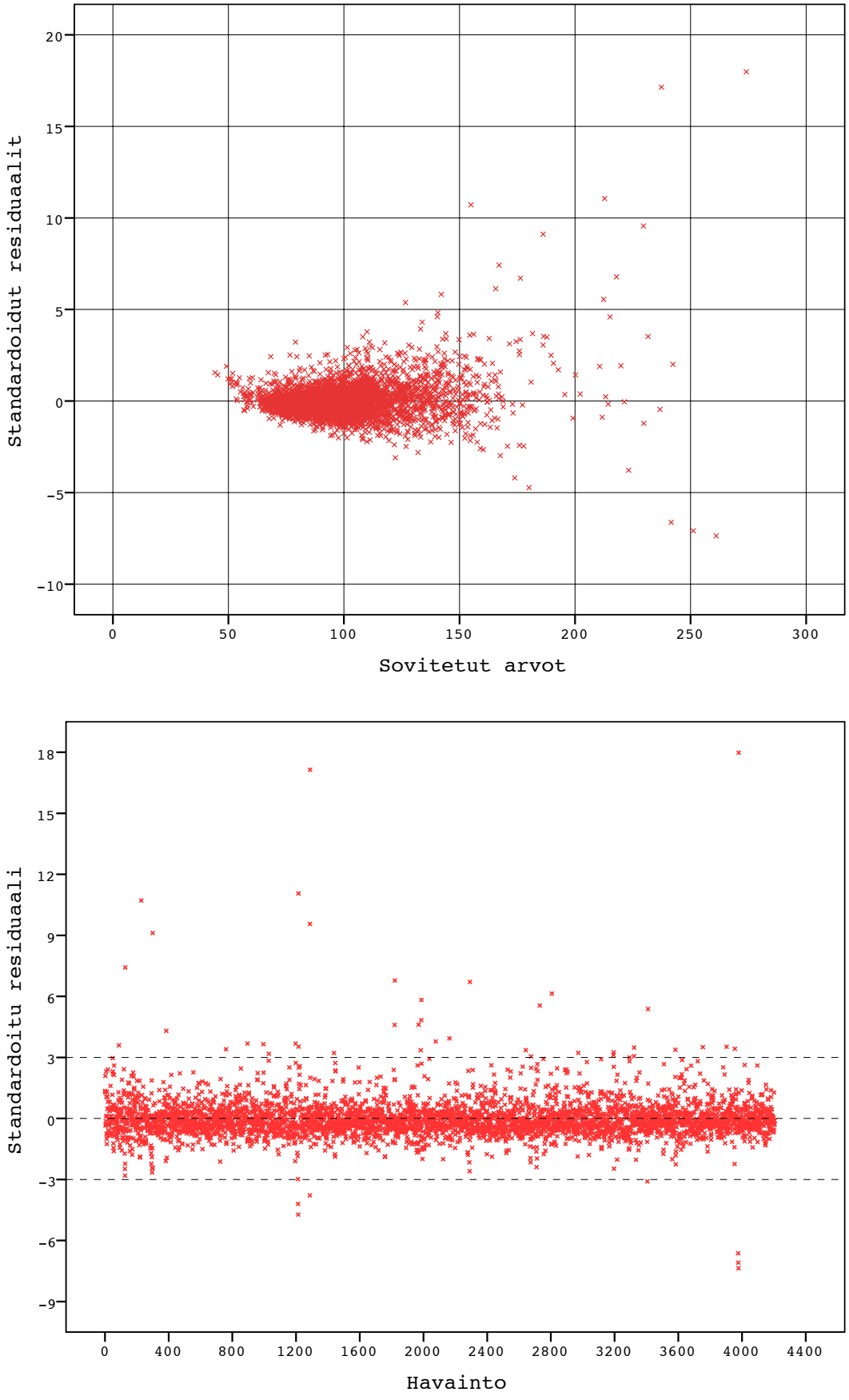


Kuvio 19: Tuotantokustannusten kehitys maitolitraa kohden 2000–2011 tukialueittain, mitatut arvot ja mallilla 8 ennustetut arvot.

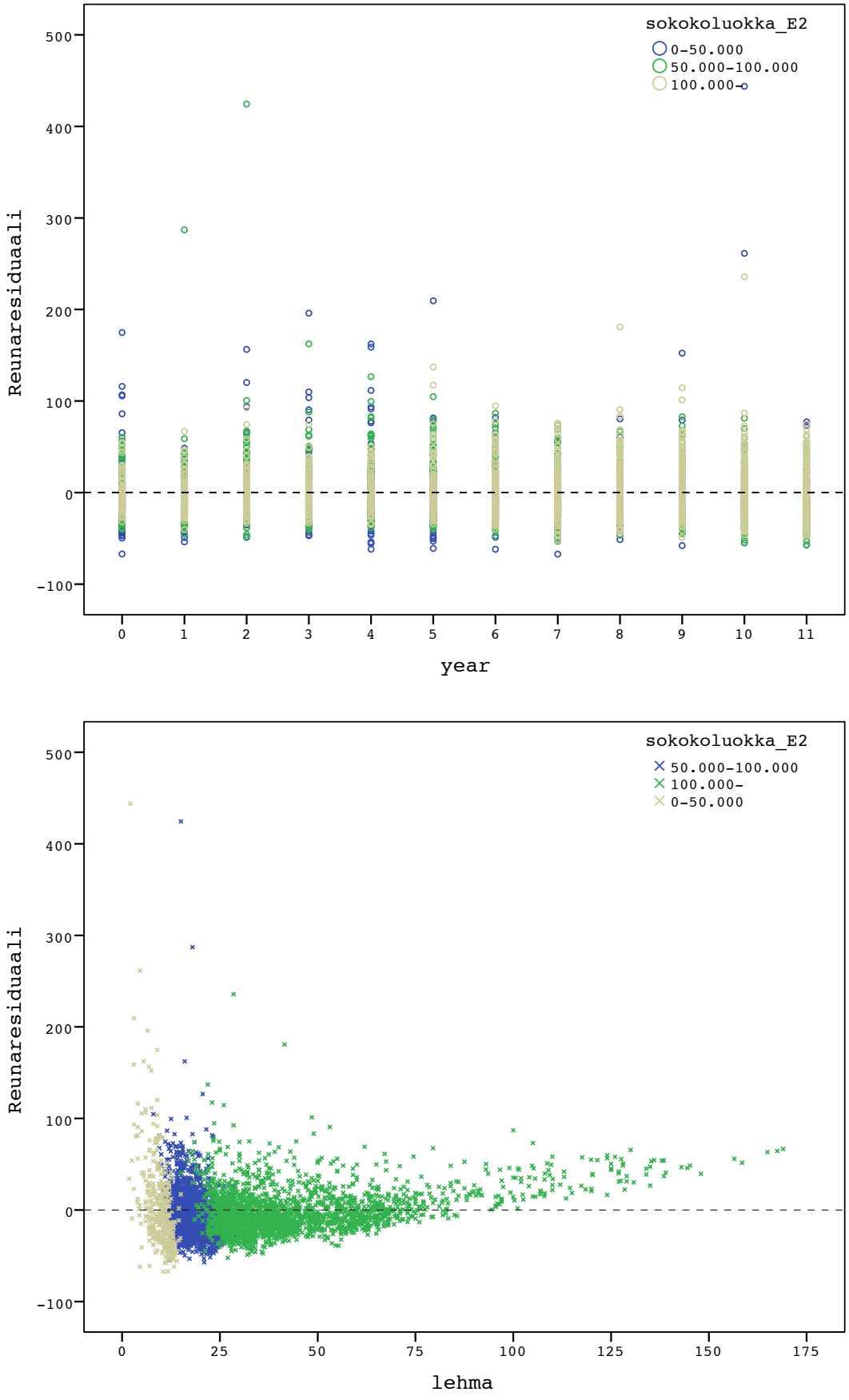
Liite 10. Mallin 8 diagnostiikka



Kuvio 20: Residuaalien kvantiilikuvio ja histogrammi.



Kuvio 21: Residuaalien sirontakuvio.



Kuvio 22: Reunaresiduaalin ja selittävän muuttujan sirontakuvio.