JYVÄSKYLÄN YLIOPISTO
MATEMATIIKAN JA
TILASTOTIETEEN LAITOS

Mauno Keto

ALUEKIINTIÖINTI OSITETUSSA OTANNASSA

Tilastotieteen lisensiaattitutkielma

2014

TIIVISTELMÄ

Aluetilastoja tuotetaan paljon kyselytutkimuksilla, joissa tietojen keruu nojautuu erilaisiin otanta-asetelmiin. Näissä asetelmissa alue määrittyy yleensä ositteeksi ja aluetunnusluvut ositetason estimaateiksi. Keskeinen kysymys tällaisessa tilanteessa on, miten otos kiintiöityy alueiden kesken. Taustalla on siis kiintiöintiongelma, jonka seurauksena joillekin alueille tulee vähän tai ei lainkaan otoshavaintoja. Tästä syystä alue-estimoinnissa on yleistä käyttää suorien otanta-asetelmaperusteisten estimaattien sijasta malliavusteisia tai –perusteisia estimaatteja, jotta kaikille alueille saataisiin riittävän tarkat estimaatit halutuista tunnusluvuista.

Aluekiintiöinti sinällään on monitahoinen ongelma, josta osoituksena ovat kirjallisuudesta poimitut useat ehdotukset. Niissä optimointikriteerit ovat erilaisia. Eräissä sellaisia ei ole lainkaan, toisissa ne on asetettu aluetasolle, toisissa perusjoukkotasolle ja eräissä jopa samanaikaisesti sekä perusjoukko- että aluetasolle. Yhteistä niille kaikille on kuitenkin, että niissä ei ole estimointia tehostavaa mallia lainkaan mukana. Edellisistä poiketen tässä työssä on kehitetty kolme aluekiintiöintiä, joista yksi perustuu perusjoukosta poimittavaan pieneen esiotokseen ja regressiomalliin ja kaksi on ehdollistettu estimointivaiheessa käyttöön otettavaan yksikkötason lineaariseen sekamalliin ja paljon käytössä olevaan paras lineaarinen ennuste (EBLUP) –estimointiin.

Kiintiöintiratkaisujen toimivuus on testattu simulointikokein. Kokeilussa mukana ovat kolme tässä työssä kehitettyä malliavusteista kiintiöintiä. Näiden vertailukiintiöinteinä ovat kirjallisuudesta poimitut 1) tasakiintiöinti, 2) suhteellinen kiintiöinti, 3) perusjoukon suhteen optimaalinen eli Neyman-kiintiöinti, 4) alueoptimaalinen potenssikiintiöinti ja 5) epälineaarinen optimikiintiöinti (NLP). Koealustana on reaalinen perusjoukko, joka on osa maamme asuntomyyntirekisteriä vuodelta 2011. Asunnoista koostuva perusjoukko käsittää 34 aluetta, joiden välinen vaihtelu on kohtalaisen voimakasta. Aluemallin käytölle on tällöin edellytyksiä. Lisävertailuja varten alueita on yhdistelty siten, että on saatu rakenteeltaan erilainen, 14 alueen perusjoukko. Näistä perusjoukoista on simuloitu 1500 ositettua yksinkertaista satunnaisotosta kiintiöintimenetelmittäin ja laskettu kaikista otoksista samat malliperusteiset estimaatit ja tunnusluvut sekä estimoinnin tehokkuutta ja luotettavuutta kuvaavat laatumittarit. Laatumittareista keskeinen on prosenttilukuna ilmaistava aluekohtainen ja koko perusjoukon suhteellinen keskivirhe (*$RRMSE_d$ %* ja *RRMSE%*). Estimointitulosten ja laatumittarien pohjalta on arvioitu eri kiintiömenetelmien toimivuutta. Johtopäätös on, että optimaalinen kiintiöinti riippuu asetettavista optimointikriteereistä ja perusjoukon aluerakenteesta.

Avainsanat: Aluekiintiöinti, aluemalli, alueoptimaalinen kiintiöinti, esiotos, sijaismuuttuja ja alue-estimoinnin laatumittarit.

KIITOKSET

Kiitän seuraavia henkilöitä jatko-opintojeni tukemisesta sekä lisensiaattitutkielmani ohjauksesta: emeritusprofessori Erkki Pahkinen, joka on perehdyttänyt minut otannan mielenkiintoiseen maailmaan ja joka on väsymättä ja aikaansa säästämättä ohjannut minua eteenpäin tässä tutkimuksessa, professori Antti Penttinen, joka on tukenut merkittävästi jatko-opintojeni käynnistymistä ja joka on ollut toinen tutkimukseni ohjaaja sekä YTT Kari Nissinen, joka luovutti ystävällisesti käyttööni simulointikokeissa käyttämiäni SAS-ohjelmia ja joka on neuvonut monissa pienalue-estimoinnin teoriaan liittyvissä ongelmissa.

Mikkelissä heinäkuun 30. päivänä 2014

Mauno Keto

SISÄLTÖ

1 JOHDANTO

Aluetilastojen käyttö yhteiskunnallisessa päätöksenteossa on lisääntynyt merkittävästi Suomessa ja muissakin maissa viimeisten 15 - 20 vuoden aikana. Alue voi tarkoittaa maantieteellisesti rajattua aluetta tai jotain perusjoukosta määriteltyä osajoukkoa, joka muodostetaan esim. tietyn demografisen, sosiaalisen tai talouselämään liittyvän ominaisuuden (sukupuoli, ikä, etninen ryhmä, tuloluokka, yritysmuoto jne.) tai useamman eri ominaisuuden ristiinluokittelun perusteella. Aluetilastoiksi määritellään yleensä alueita kuvaavia tunnuslukuja kuten keskiarvoja, kokonaismääriä ja erilaisia osuuslukuja. Edellä mainittujen tunnuslukujen laskennan kannalta paras tietolähde olisi ilman muuta perusjoukon kaikki alueet kattava tietorekisteri. Näin ei kuitenkaan aina ole. Monesti on turvauduttava otanta-asetelmalla kerättyyn kyselyaineistoon, jolloin esille nousee kysymys siitä, miten hyvin näin poimittu aineisto edustaa aluetasolla tutkittavia ilmiöitä. Taustalla on silloin yhtenä tekijänä se, miten otos on jakautunut eli kiintiöitynyt eri alueille, jotta aluetason tunnusluvut saadaan mahdollisimman luotettavasti lasketuiksi tai tarkkaan ottaen estimoiduiksi, kun kysymyksessä on otanta-aineisto.

Otantateorian näkökulmasta tutkimuskohteeksi valikoituu otoksen kiintiöinti tilanteessa, jossa optimointikriteerit asetetaan ositetasolle, koska osite on määriteltävissä myös alueeksi tai päinvastoin. Näin ollen tämän tutkimuksen keskeiseksi ongelmaksi nousee, miten optimaalinen aluekiintiöinti ratkaistaan. Aiempia ratkaisuja on haettu kirjallisuuskatsauksessa. Niistä on saatu tähän tutkimukseen hyödynnetyiksi menetelmiä, jotka pääosin liittyvät otanta-asetelmassa kerätyn alueaineiston estimointivaiheeseen. Niistä on kuvauksia luvuissa 3–4. Niiden mukaan alueestimointi voidaan toteuttaa kolmella eri tavalla kuten suorana, malliavusteisena ja malliperusteisena estimointina. Perusratkaisu on otanta-asetelman mukainen suora estimointi, jossa keskeisenä ovat otosalkioiden sisältymistodennäköisyyksistä johdetut otantapainot. Tällöin estimointi on toteutettavissa aluekohtaisten otoshavaintojen perusteella. Edellytys on tietenkin se, että havaintoja on riittävästi kaikilta alueilta. Tämä on ehto, joka ei aina toteudu kiintiöintimenetelmästä johtuen ja estää näin tällaisen suoran estimointimenetelmän käytön. Tämä ongelma on nostanut alue-estimointiin monesti liitetyn käsitteen ”pienalue”.

Käsitteellä ”pienalue” on kaksi tulkintaa. Aluetta voidaan nimittää pienalueeksi perusjoukon tasolla, kun siihen kuuluu vähän tai ei lainkaan tilastoyksikköjä. Otanta-asetelman tasolla pienalue syntyy tilanteessa, jossa alueelle kiintiöityy vähän tai lainkaan otosyksikköjä. Tällöin alueen otoskoko on niin pieni, ettei ole mahdollista saada riittävän luotettavia alue-estimaatteja, jotka perustuisivat suoraan asetelmaperusteiseen estimointiin. Sen sijaan on käytettävä epäsuoraa mal-

lipohjaista estimointia, joka käyttää kaikilta alueilta saatavaa otos- ja aputietoa. Olennaista pienalueelle on kuitenkin sille tuleva pieni otoskoko, ei niinkään alueen koko. Pienalueiden määrä voi vaihdella suurestikin riippuen siitä, millainen jako on tutkimuksen kannalta relevantti. Sama perusjoukko voidaan jakaa pienalueisiin monella eri tavalla. Joskus taas pienempiä alueita yhdistellään tiettyjen yhteisien ominaisuuksien perusteella sellaisessa tilanteessa, jossa kaikille pienille alueille ei ole mahdollista kohdistaa otantaa kustannus- tai aikarajoituksien vuoksi.

Aluetasolla hyviin estimointituloksiin johtavaa otoskiintiöintiä on tutkittu ja siitä on keskusteltu jo pitkään. Tutkimus on painottunut selvästi enemmän asetelmaperusteisen suoran estimoinnin puolelle. Tutkimuksissa, joissa on käytetty malliperusteista (epäsuoraa) estimointia, on keskitytty lähinnä vain keskineliövirheen (MSE) minimointiin, mutta sellaiset asiat kuin ennusteiden mahdollisimman pieni virhe ja harha ovat jääneet vähemmälle huomiolle. Näiden ääripäiden väliin jää malliavusteinen suora estimointi, joka on tavanomainen tekniikka otanta-aineistojen analyysissä.

Muutamia ratkaisuja otoksen kiintiöinnistä aluetasolle on kuvattu luvussa 4. Perusratkaisu lähtee määrittelystä, jossa alue tulkitaan otanta-asetelman ositteeksi. Kiintiöinnin toteutus edellyttää tietoja perusjoukosta ja tietyn optimointikriteerin. Vähiten tietoja tarvitaan tasa- ja suhteellisessa kiintiöinnissä. Edellisessä tarvitaan vain yksi lukumäärätieto, joka on alueiden lukumäärä *D*. Jälkimmäisessä on tiedettävä jokaisen ositteen tilastoyksiköiden lukumäärä N_d . Kummassakaan kiintiöinnissä ei ole mitään optimointikriteeriä. Perusratkaisuina ne ovat jatkossa viitteellisiä aluekiintiöintejä. Sitä vastoin Neyman-, potenssi- ja NLP-kiintiöinneissä tarvitaan tiedot perusjoukon ositetason hajonnan ($S_d(y)$) tai vaihtelukertoimen ($CV_d(y)$) parametreista. Nämä tiedot eivät yleensä ole tutkittavasta vastemuuttujasta tiedossa, jolloin ne korvataan jollakin sijaismuuttujatiedolla. Näissä parametriperusteisissa kiintiöinneissä on optimointikriteerit. Neyman-kiintiöinnissä optimointikriteeri on asetettu niin, että kiintiöinti minimoi perusjoukkotason estimaatin keskivirheen. Potenssikiintiöinnin optimointi tähtää pelkästään siihen, että aluetason (osite) keskivirheet olisivat samaa suuruusluokkaa. Epälineaariseen ohjelmointialgoritmiin perustuva NLP-kiintiöinti sisältää optimiehdon sekä perusjoukko- että aluetasoille. Edellisistä poiketen sillä on ratkaistavissa myös nämä ehdot täyttävä otoskoko, kun sen sijaan muissa edellisissä otoskoko oletetaan annetuksi.

Yksikään edellä luetelluista kiintiöintimenetelmistä ei ratkaisuissaan hyödynnä alue-estimoinnin tukena yleisesti käytettyjä malleja ja estimointimenetelmiä. Tämä tutkimus lähti aikanaan liikkeelle tarpeesta täydentää aluekiintiöintiä siten, että siinä olisi mukana alue-estimoinnissa käy-

tössä oleva mallinnus. Varsinaiseksi aluemalliksi on valittu tässä aluevaikutukset huomioon ottava lineaarinen, yhden apumuuttujan sisältävä sekamalli, josta käytetään nimeä hierarkkinen yksikkötason regressiomalli. Mallia voidaan siis pitää yhtenä annettuna ennakkotietona ja on aivan samassa asemassa kuin miten aputieto eli yleensä apumuuttuja on mukana malliavusteisessa otanta-aineiston analyysissä. Tämä ratkaisu on johtanut kolmen uuden aluekiintiöintiratkaisun johtamiseen, joista yksi perustuu perusjoukosta poimittavaan pieneen esiotokseen ja regressiomalliin ja kaksi perustuvat estimointivaiheessa käyttöön otettavaan ja edellä mainittuun sekamalliin ja paljon käytössä olevaan paras lineaarinen ennuste (EBLUP) –estimointiin. Ratkaisut on esitetty luvussa 5. Näistä ensimmäinen edellyttää, että perusjoukosta poimitaan pieni esiotos, jonka avulla voidaan estimoida vastemuuttujaa *y* korvaava sijaismuuttuja ja sitä kautta päästä suhteellisten aluevarianssien painotetun keskiarvon minimointiin. Toinen tässä kehitetty aluekiintiöintiratkaisu perustuu mallin avulla lasketun keskineliövirheen (MSE) tärkeimmän komponentin *g1* yli alueiden lasketun keskiarvon minimointiin (Keto 2012). Myös kolmas kiintiöintiratkaisu käyttää hyväkseen esiotoksen avulla johdettua sijaismuuttujaa, mutta sen lisäksi myös otossimulointia ja valittuun aluemalliin liittyvää EBLUP-estimointia.

Luvussa 6 on suoritettu tehokkuusvertailu kolmen uuden aluekiintiöintiratkaisun ja viiden aikaisemmasta kirjallisuudesta poimitun kesken. Alustana on käytetty kotimaista kiinteistövälittäjien asuntomyyntitilastoa vuodelta 2011. Vastemuuttujana *y* on asunnon hinta (1000 €), apumuuttujana *x* asunnon koko (m^2) ja alueena asunnon sijaintikunta (34 aluetta). Kiinteäksi otoskooksi on valittu $n = 170$ ja aluekiintiöinti on suoritettu mainituilla kahdeksalla eri menetelmällä. Tämän jälkeen on simuloitu kiintiöintikohtaisesti 1500 satunnaisotosta, joille on suoritettu valittuun aluemalliin soveltuva EBLUP-estimointi tunnuslukuna olleelle vastemuuttujan kokonaismäärälle. Eri kiintiöintien otoksista lasketaan estimaattien lisäksi sellaisia tunnus- ja tehokkuuslukuja, joilla voidaan mitata kiintiöintimenetelmien laatua ja toimivuutta. Näistä tärkein laatuvertailuluku tässä tutkimuksessa on prosenttimuodossa ilmaistava kokonaismäärän neliöperusteinen suhteellinen virhe ($RRMSE_d\%$). Tarkemmat laskentatulokset löytyvät luvusta 6. Liitteissä B, G, J ja K on esitetty eri kiintiöintien johtamiseen liittyviä laskelmia.

Johtopäätökset on esitetty luvussa 7. Niistä ilmenee, että käytetty malli, siihen perustuva estimointi ja simulointimenetelmä asettavat omat rajoituksensa saatujen tulosten käyttökelpoisuudelle. Tarkoituksena ei ole ollut löytää jotain yleispätevää ”laskukaavaa”, jonka avulla voitaisiin laskea optimaalinen aluetason otoskiintiöinti, vaan tutkia sitä, onko mahdollista saavuttaa hyviä estimointituloksia tavallisuudesta poikkeavien kiintiöintien avulla, kun valittua mallia käytetään. Toinen asia, jota ei ole otettu huomioon, on optimaalisuuskriteerien mahdollinen painotus, eli

onko jokin kriteeri tärkeämpi saavuttaa kuin muut kriteerit. Tässä suunnassa eräs vaihtoehto on tutkia, onko mahdollista sisällyttää NLP-kiintiöintiin lisäinformaatioksi mallitus. Kaiken kaikkiaan tutkimusongelma on varsin laaja, ja yhdellä kiintiöinnillä tuskin voidaan optimoida kaikkia kriteerejä samanaikaisesti. Kaikkia mahdollisia kiintiöintivaihtoehtoja ei voida kokeilla ja käytetyllä havaintoaineistolla on oma vaikutuksensa, mutta vertailujen tuloksista saadaan kuitenkin käyttökelpoista tietoa otanta-asetelman suunnitteluun ja otoskiintiöintiin.

2 ALUE-ESTIMOINTI

2.1 Alue ja pienalue

Alue määritellään tilastollisessa tutkimuksessa siten, että kohteena oleva perusjoukko jaetaan maantieteellisesti tai jonkin ominaisuuden perusteella toisensa poissulkeviin osajoukkoihin. Alueiden määrittelyt perustuvat yleensä luokittelutason muuttujiin, joiden arvot identifioivat käytössä olevat alueet. Jos otantamenetelmä on ositettu otanta, muodostetut alueet määrittyvät suoraan ositteiksi. Alueet voivat olla hyvinkin poikkeavia toisiinsa nähden tilastoyksiköiden määrän ja muiden erilaisten ominaisuuksien suhteen, mikä on tietysti luonnollista, koska muutenhan estimoinnissa voitaisiin käyttää yksinkertaisempia menetelmiä. Estimointi voi kohdistua pienalueiden lisäksi myös koko perusjoukkoon, jolloin luotettavat tulokset on saatava sekä ylemmällä että alemmalla tasolla, mutta tämä on usein erittäin vaikea tehtävä.

Alue voidaan määritellä pieneksi, jos siihen sisältyy perusjoukon tasolla vähän tilastoyksikköjä. Ero alueen ja pienalueen välillä on Raon (2003) mukaan myös se, että alueelta tuleva otos on riittävän suuri. Tällöin asetelmaperusteinen suora estimointi tuottaa tarpeeksi luotettavat alueestimaatit, kun taas pienalueelta saatava otos on liian pieni, jotta tarvittavaan tarkkuuteen päästäisiin. Tässä tutkimuksessa käsitellään estimointia lähinnä pienaluetasolla.

2.2 Estimoitavat aluetason tunnusluvut

Perusasetelmaan kuuluu vastemuuttuja (y), johon liittyviä estimoitavia parametreja ovat useimmiten lukumääriin tai mitattaviin suureisiin liittyvät aluetason tunnusluvut kuten kokonaismäärät, keskiarvot, mediaanit, kvartiilit ja osuusluvut. Alue-estimoinnissa tärkeitä ovat myös alueiden sisäistä ja välistä varianssia mittaavat tunnusluvut kuten sisäkorrelaatio ja erilaiset homogeenisuusmitat. Eräs standardimitta on vaihtelukerroin, joka määritellään vastemuuttujan keskihajonnan ja keskiarvon osamääränä. Jos estimointi perustuu regressiomalliin, alueiden otoshavainnoista voidaan estimoida aluekohtaiset regressiokertoimet.

Aluetason estimoinnin problematiikka ehdollistuu siihen, miten paljon otoshavaintoja tulee alueelle. Tästä otannan satunnaisuudesta johtuen esiintyy alueita, joille ei tule lainkaan otoshavaintoja. Tässä on syy, miksi alue-estimointia on pakko tehdä malliperusteisesti siten, että käytettävissä olevaa otantatietoa ”lainataan” muilta alueilta. Tällaista ratkaisua sovelletaan yleisesti alueestimoinnissa. Malliperusteisuus jakaantuu kahteen alalajiin, joista toisessa otanta-

asetelmaperusteista alue-estimaattoria korjataan malliavusteisesti ja toisessa lasketaan alue-estimaatit pelkästään malliperusteisesti.

Käytännön tutkimuksessa voidaan estimoida esimerkiksi seuraavia numeerisia tunnuslukuja:

- kotitalouksien keskimääräinen tai kokonaisvelka maakunnittain tai kunnittain
- keskimääräisen kk-palkan estimointi eri ammattiryhmissä
- puolueiden kannatusprosentit maakunnittain koko maan lisäksi
- NATO-mielipiteen kannatus aikuisväestön eri ikäryhmissä sukupuolittain.

2.3 Alue-estimoinnin laatumitat

Tilastollisessa laskennassa estimoinnin tuloksia arvioidaan tietynlaisin laatumitoin. Alue-estimoinnissa on niin paljon erikoistilanteita, että siitä on syntynyt suuri määrä alan kirjallisuutta (Lehtonen ym. 2006). Alue-estimoinnin eräs perusongelma on estimaattorien harhaisuus, joka johtuu useimmiten siitä, että suoria asetelmaperusteisia estimaattoreita ei voida käyttää, koska otoshavaintoja on liian vähän tai ei ollenkaan. Käytössä olevat alue-estimoinnin laatumitat on rakennettava tämän tosiasian varaan. Tästä syystä tavanomaiset estimaattien keskivirheet on vaihdettava keskineliövirheiksi tai niiden estimaateiksi.

Oletetaan, että θ_d on estimoitava tunnusluku (esimerkiksi keskiarvo tai kokonaismäärä) alueella *d* ja $\hat{\theta}_d$ on tunnusluvun estimaattori. Tällöin estimaattorin harha on erotus

$$B(\hat{\theta}_d) = E(\hat{\theta}_d) - \theta_d .$$

Estimaattorin $\hat{\theta}_d$ tarkkuutta mitataan sen keskineliövirheellä

$$MSE(\hat{\theta}_d) = E(\hat{\theta}_d - \theta_d)^2 = V(\hat{\theta}_d) + (E(\hat{\theta}_d) - \theta_d)^2 ,$$

eli estimaattorin varianssin ja harhan neliön summa. Jos estimaattori on harhaton, sen keskineliövirhe on sama kuin varianssi. Keskineliövirheen sijasta käytetään joskus sen estimaattia tai approksimaatiota.

Estimaattorin $\hat{\theta}_d$ vaihtelukerroin (CV*)* on osamäärä

$$CV(\hat{\theta}_d) = \sqrt{MSE(\hat{\theta}_d)} / \hat{\theta}_d ,$$

joka voidaan ilmaista myös prosenttimuodossa. Tämä on vastine yleiselle satunnaismuuttujan vaihtelukertoimelle (keskihajonta jaettuna keskiarvolla).

Alueiden välisen vaihtelun mittaamiseen voidaan käyttää myös homogeenisuusmittaa, joka on ryväsotantaan liittyvän sisäkorrelaation vastine erisuurten rypäiden tapauksessa:

$$R_a^2 = 1 - R^2 = 1 - MSW / S^2 \ ,$$

missä *MSW* tarkoittaa rypäiden (tässä tapauksessa ositteiden) sisäistä keskineliösummaa ja S^2 muuttujan varianssia. Homogeenisuusmitta on nollan ja ykkösen välillä ja kuvaa alueiden välisen vaihtelun osuutta kokonaisvaihtelusta.

Estimaattorin tehokkuus on suhteellinen käsite, koska siihen liittyy vertailu. Yleisessä tapauksessa estimaattori $\hat{\theta}_A$ on tehokkaampi kuin estimaattori $\hat{\theta}_B$, jos niiden keskineliövirheiden välillä on voimassa epäyhtälö

$$MSE(\hat{\theta}_A) < MSE(\hat{\theta}_B).$$

Jos estimaattorit ovat harhattomia, epäyhtälössä esiintyvät estimaattorien varianssit.

Malliperusteisen estimaattorin tehokkuutta voidaan mitata pienalue-estimoinnissa suhdeluvulla *EFF* (Rao 2003), jossa jaettavana on jälkiositetun asetelmaperusteisen estimaattorin keskivirhe ja jakajana malliperusteisen alue-estimaattorin keskineliövirheen neliöjuuri.

Otossimulointien yhteydessä käytetään erilaisia laatumittareita, jotka esitellään alaluvussa 6.2.

2.4 Lähestymistapoja estimointimenetelmiin

Pienalueiden estimointimenetelmät ja estimaattorit voidaan luokitella eri tavoin sen mukaan, käytetäänkö tutkittavan alueen estimoinnissa vain tältä alueelta ja vain tutkimusajanjaksolla saatavaa informaatiota vai lainataanko sitä alueen ulkopuolisilta alueilta ja käytetäänkö myös tutkimusajanjakson ulkopuolella kerättyä aputietoa. Tämän jaon perusteella on estimointi joko suoraa tai epäsuoraa (Federal Committee on Statistical Methodology, 1993). Alueet on suunniteltu ennen otantaa.

Suora asetelmaperusteinen pienalue-estimointi pohjautuu perinteiseen otantateoriaan ja siihen, että äärellisestä, N tilastoyksikköä sisältävästä kiinteästä perusjoukosta poimitaan n kappaleen suuruinen satunnaisotos *s*, jonka poimintatodennäköisyys *p(s)* määritellään otanta-asetelman

(SRSWOR-, systemaattinen, ositettu otanta jne.) mukaan. Estimaattorien laskennassa tarvitaan lisäksi alueen *d* yksittäisen tilastoyksikön *k* sisältymistodennäköisyyttä π_{dk} , joka riippuu otanta-asetelmasta. Jakaumaoletuksia ei tehdä, vaan ainut satunnainen tekijä on otoksen *s* koostumus. Koska asetelmaperusteiseen estimointiin ei kuulu malli, ei voida varsinaisesti puhua mallin parametrien, vaan vastemuuttuja *y*:n havaintoarvojen funktioiden $h(y_1, y_2, ..., y_N)$, kuten summan tai keskiarvon, estimoinnista. Estimaattorin ominaisuudet johdetaan otanta-asetelman määräämästä otantajakaumasta. Vastemuuttujan *y* jonkin estimaattorin (esimerkiksi kokonaismäärän tai keskiarvon) varianssi ja harha lasketaan kaikkien mahdollisten kiinteän havaintomäärän kokoisten otosten yli, jolloin voidaan puhua asetelmavarianssista ja –harhasta.

Suora malliavusteinen estimointi perustuu aluekohtaisiin malleihin (esimerkiksi regressiomalli), jolloin vastemuuttujan *y* lisäksi on käytettävissä erilaista apumuuttujatietoa jokaiselta alueelta. Yksittäisen alueen estimoinnissa voidaan kuitenkin joskus käyttää myös alueen ulkopuolista ylemmän tason apumuuttujainformaatiota, jos se on ainoa käytettävissä oleva tällainen informaatio (Federal Committee on Statistical Methodology, 1993). Epäsuora estimaattori käyttää myös tutkimuskohteena olevan alueen ulkopuolista vaste- ja apumuuttujatietoa, joka on voitu kerätä myös useana eri ajankohtana.

Malliperusteinen pienalue-estimointi perustuu sellaisten tilastollisten mallien käyttöön, joiden avulla vastemuuttujan *y* aluekohtaiset estimaatit voidaan laskea malliennusteina. Mallin soveltamista varten tarvitaan rekisteritasoista apumuuttujatietoa kaikista tilastoyksiköistä. Apumuuttujat korreloivat vastemuuttujan *y* kanssa. Kaikki malliperusteinen pienalue-estimointi on epäsuoraa, mutta kaikki epäsuora estimointi ei ole välttämättä malliperusteista (Nissinen 2009). Käytettävä malli tuottaa alueelle *d* ennusteet otoksen ulkopuolisille *y*-arvoille tai niiden summalle, kun estimoidaan *y*:n summaa tai keskiarvoa koko alueella. Muita epäsuoria estimaattoreita ovat Raon (2003) mukaan mm. seuraavat: synteettiset estimaattorit, jotka eivät noudata eksplisiittisesti mitään varsinaista mallia, mutta jotka perustuvat oletukseen, että pienalueiden ominaisuudet ovat likimain samat kuin koko perusjoukon vastaavat sekä suoran ja synteettisen estimaattorin yhdistelmät. Malliperusteinen pienalue-estimointi yhdistää alueet ja käyttää sekä otos- että apumuuttuja-aineistoa hyväkseen.

Pienalue-estimointiin liittyviä ongelmia esiintyy kahdella eri tasolla, joista ensimmäinen liittyy otanta-asetelmaan ja otoksen kiintiöitymiseen aluetasolla. Onnistunut kiintiöinti edellyttäisi, että kaikille alueille saataisiin samanlaatuiset estimaatit. Tämä ei ole ensisijaisesti yhteydessä otos-

kokoon, vaan vastemuuttujan sisäiseen vaihteluun eri alueilla (Choudry ym. 2012). Toinen ongelma on tutkittavan ilmiön alueiden sisäinen ja välinen vaihtelu, joita ei tunneta.

3 ALUE-ESTIMOINNIN MENETELMÄT

3.1 Perusoletukset

Estimointimenetelmien kuvaamisessa käytetään seuraavia oletuksia ja merkintöjä:

1) Perusjoukkoon *U* kuuluu *N* kpl tilastoyksiköitä.
2) Perusjoukko on jaettu toisensa poissulkeviin alueisiin, joita on *D* kpl.
3) U_d tarkoittaa alueen *d* perusjoukkoa (*d = 1,2,…, D).*
4) Perusjoukko on pienalueiden yhdiste: $U = U_1 \cup U_2 \cup ... \cup U_D$.
5) Tilastoyksiköiden määrä alueen *d* perusjoukossa = N_d, joten $N = \sum_{d=1}^{D} N_d$.
6) Otokseen poimitaan tilastoyksiköitä kiinteä määrä *n*.
7) Merkintä s_d tarkoittaa alueelle *d* tulevaa otosta.
8) Alueen *d* otoskoko = n_d, joten $n = \sum_{d=1}^{D} n_d$.
9) Apumuuttujat $\boldsymbol{x} = x_1, x_2, ..., x_p$ (*p* kpl).
10) Vastemuuttujan *y* estimoitavat tunnusluvut: Alueen *d* kokonaismäärä $Y_d = \sum_{k \in U_d} y_{dk}$ tai keskiarvo $\bar{Y}_d = Y_d / N_d$.
11) Vastemuuttujan *y* vastaavat otostunnusluvut: Alueen *d* kokonaismäärä $y_d = \sum_{k \in (s_d)} y_{dk}$ ja keskiarvo $\bar{y}_d = \sum_{k \in (s_d)} y_{dk} / n_d$.
12) Apumuuttujista $\boldsymbol{x}$ tunnetaan mahdollisesti kaikki arvot perusjoukossa, mutta ainakin alueiden kokonaismäärät $X_d = \sum_{k \in U_d} x_{dk}$ tai keskiarvot $\bar{X}_d = X_d / N_d$.

3.2 Suora estimointi

Yksinkertainen esimerkki on vastemuuttujan *y* alueen *d* kokonaismäärän Y_d suora asetelmaperusteinen estimaattori

$$\hat{Y}_d = \sum_{k \in s_d} w_{dk} y_{dk} \,, \tag{3.1}$$

missä on w_{dk} tarkoittaa otantapainoa ja s_d alueelta *d* tulevaa otosta. Jos $w_{dk} = 1/\pi_{dk}$, missä π_{dk} on alueen *d* tilastoyksikön *k* sisältymistodenäköisyys, saadaan yksinkertaisen satunnaisotannan

(SRSWOR) tapauksessa, jossa sisältymistodennäköisyydet ovat samat kaikille tilastoyksiköille eli n/N (Thompson 2002), estimaattorille lauseke

$$\begin{aligned} \hat{Y}_d &= \frac{N}{n}\sum_{k\in s_d} y_{dk}, & n_d \geq 1 \\ &= 0, & n_d = 0. \end{aligned} \tag{3.2}$$

Rao ja Choudry (1999) määrittelevät tämän laajennusestimaattoriksi. Se on aluetasolle sovellettu Horvitz-Thompson –estimaattori. Tätä nimeä käytetään jatkossa, kun estimaattoriin viitataan.

Estimaattorista (3.2) saadaan tehokkaampi (pienempi varianssi), kun *N* ja *n* korvataan eksaktilla tiedolla alueelta *d:*

$$\begin{aligned} \hat{Y}_d &= \frac{N_d}{n_d}\sum_{k\in s_d} y_{dk} = N_d \bar{y}_d, & n_d \geq 1 \\ &= 0, & n_d = 0. \end{aligned} \tag{3.3}$$

Suoran asetelmaperusteisen estimaattorin varsin suuri suosio johtuu siitä, että se on ainakin likimain asetelmaperusteisesti harhaton, ts. sen odotusarvoksi tulee estimoitava suure, mutta sillä on kaksi merkittävää puutetta: tehottomuus, mikä ilmenee estimaattorin suurena varianssina ja se, että alueille, joilta ei otoksessa ole lainkaan havaintoja, ei saada estimaattia. Asetelmaperusteista lähestymistapaa on Raon (2003) mukaan kritisoitu myös sen johdosta, että estimoinnin perusteella tehtävät päätelmät, vaikkakin ne ovat vapaat jakaumaoletuksista, viittaavat toistettaviin otospoimintoihin eikä poimittuun yhteen otokseen *s*. Vaihtoehdoksi on esitetty rajoitettua mahdollisten otosten joukkoa, mikä johtaisi ehdollisesti kelvollisiin päätelmiin. Jälkiositukseen perustuva estimointi on esimerkki tästä periaatteesta (Rao 2003).

3.3 Malliavusteinen suora estimointi

Estimoinnin tehokkuutta voidaan lisätä käyttämällä asetelmaperusteista mallitehosteista estimaattoria, johon liittyy apumuuttujien $x_1, x_2, ..., x_p$ antama lisäinformaatio kaikilta pienalueilta. Mikäli apumuuttujia on vain yksi, voidaan käyttää suhdetehosteista estimaattoria. Vastemuuttujan *y* kokonaismäärä alueella *d* estimoidaan lausekkeella

$$\hat{Y}_{d,rat} = \bar{y}_d / \bar{x}_d \times X_d, \tag{3.4}$$

joten alue-estimaatin laskemiseksi on tunnettava vaste- ja apumuuttujan otoskeskiarvot ja apumuuttujan kokonaismäärä alueella *d*.

Regressiotehosteinen estimointi on mahdollinen yhden tai useamman apumuuttujan tapauksessa, ja estimaattori saadaan korjaamalla puhdasta asetelmaperusteista estimaattoria regressiotermillä, jonka kerroin on joko aluekohtainen ja laskettu alueen otoshavainnoista tai yhteinen, koko otosaineistosta laskettu.

Mallitehosteisen estimoinnin toinen hyvä puoli tehokkuuden lisääntymisen ohella on se, että estimaattori on asetelmaperusteisesti harhaton, olipa malli oikea tai ei. Yksi puute kuitenkin säilyy: jos jonkin alueen havaintomäärä on nolla, ei estimaattia voida edelleenkään laskea.

3.4 Epäsuora ja malliperusteinen alue-estimointi

Epäsuoran estimoinnin perusidea on käyttää tietyn alueen estimoinnissa apuna muilta alueilta saatavaa informaatiota, joka on peräisin vastemuuttujasta ja siihen yhteydessä olevista täydentävää tietoa sisältävistä apumuuttujista, jolloin voidaan puhua efektiivisen otoskoon lisäyksestä (Rao 2003). Käytettävissä oleva otosaineisto ja sitä täydentävä aineisto tuodaan pienalueestimointiin valitun implisiittisen tai eksplisiittisen mallin avulla. Estimoinnin tuloksia kutsutaan yleensä alue-ennusteiksi eikä piste-estimaateiksi.

Implisiittinen epäsuora estimaattori ei varsinaisesti sisällä erityistä aluemallia, vaan alueita yhdistävän tekijän, jota voidaan käyttää estimoinnissa hyväksi. Esimerkkinä voidaan mainita synteettinen suhde-estimaattori yhden apumuuttujan tapauksessa:

$$\hat{Y}_{d,syn} = \bar{y} / \bar{x} \times X_d \,, \tag{3.5}$$

joka rakentuu sen olettamuksen varaan, että tulos- ja apumuuttujan suhde on kaikilla alueilla vakio. Kaavassa (3.5.) $\bar{y}$ ja $\bar{x}$ tarkoittavat koko otoksesta laskettujen tulos- ja apumuuttujan keskiarvoja sekä X_d apumuuttujan kokonaismäärää alueella *d*. Nyt voidaan laskea ennuste niillekin alueille, joista ei ole havaintoja. Sen sijaan näiden varianssien laskennassa täytyy turvautua yhden tai useamman muun alueen variansseihin.

Malliperusteinen pienalue-estimointi perustuu satunnaisotokseen, joka poimitaan jostakin stokastista mallia noudattavasta äärettömästä perusjoukosta (superpopulaatiosta), jolloin vastemuuttujan *y* oletetaan noudattavan jotain todennäköisyysjakaumaa. Käytännössä otos poimitaan kuitenkin äärellisestä, kiinteän kokoisesta perusjoukosta. Lisäksi taustalla ovat tietyt olettamukset niistä tekijöistä, joiden mukaan *y* vaihtelee alueiden välillä ja sisällä. Toinen malliin olennaisesti kuuluva oletus on, että vastemuuttuja *y* on riippuvuussuhteessa yhteen tai useampaan apumuuttu-

jaan $\mathbf{x} = x_1, x_2, ..., x_p$, joista on käytettävissä monipuolista informaatiota eri alueilta. Malli liittää vastemuuttujan ja apumuuttujat yhteen. Malliin liittyvät oletukset voivat olla sellaisia, että 1) alueilla on jokin yhteinen ominaisuus, 2) alueiden vaikutukset tutkittaviin suureisiin voidaan estimoida tai 3) vastemuuttujan ja apumuuttujien välinen riippuvuus voidaan ilmaista mallissa. Estimoitavat suureet ovat samoja kuin asetelmaperusteisessa estimoinnissa eli vastemuuttujan *y* kokonaismäärä tai keskiarvo alueella *d*.

Pienalue-estimoinnin mallit voidaan luokitella mm. seuraavasti (Rao 2003): 1) aluetason mallit, jossa on käytettävissä vain aluetason tietoa ja 2) yksikkötason mallit, joista on ensimmäisenä esimerkkinä 2-tasoinen varianssikomponenttimalli

$$y_{dk} = \mathbf{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk},\ k = 1,\ldots, N_d,\ d = 1,\ldots, D, \tag{3.6}$$

joka sisältää vektoriarvoisen apumuuttujan $\mathbf{x} = x_1, x_2, ..., x_p$, regressiokertoimet $\boldsymbol{\beta}$, toisistaan riippumattomat satunnaismuuttujat v_d, jotka edustavat aluevaikutuksia sekä satunnaistekijän e_{dk}. Vastemuuttujan *y* kokonaismäärän tai keskiarvon ennuste alueelle *d* on otosarvojen ja otoksen ulkopuolisten *y*-arvojen ennusteiden yhdistelmä. Malli esitellään tarkemmin alaluvussa 3.5.3. Toinen esimerkki on satunnaiskertoimia sisältävä regressiomalli (Torabi-Rao 2008)

$$y_{dk} = \mathbf{x}'_{dk}\boldsymbol{\beta}_d + e_{dk};\ \boldsymbol{\beta}_d = \mathbf{Z}_d\boldsymbol{\beta} + v_d,\ k = 1,\ldots, N_d;\ d = 1,\ldots, D. \tag{3.7}$$

Tässä mallissa $p \times q$ –matriisi $\mathbf{Z}_d$ sisältää aluevaikutuskovariaatit ja $\boldsymbol{\beta}$ on regressiokerroinvektori. Vastemuuttujan *y* ennusteiden laskennassa käytetään jokaiselle alueelle omaa regressiokerrointa, toisin kuin mallissa (3.6).

Raon (2003) mukaan eräs käsitys on, että käytettäessä epäsuoraa estimointia sen tulisi perustua eksplisiittisten mallien käyttöön. Mallit määrittelevät selkeästi tavan, jonka mukaan käytettävissä oleva havaintoaineisto liitetään itse estimointiprosessiin. Malliperusteinen lähestymistapa tarjoaa seuraavia etuja: 1) optimaaliset estimaattorit voidaan johtaa mallin mukaan, 2) aluekohtaiset vaihtelua kuvaavat mittaluvut voidaan liittää jokaiseen estimaattoriin, toisin kuin perinteisessä epäsuorassa estimoinnissa käytetyt koko perusjoukkoon liittyvät vaihtelumittarit, 3) mallien validiteetti (sopivuus) voidaan testata otosaineiston avulla ja 4) erilaisia malleja on käytettävissä monipuolinen valikoima riippuen vastemuuttujan luonteesta ja havaintoaineiston rakenteen monimutkaisuudesta (mm. spatiaalinen riippuvuus ja aikasidonnaisuus).

Epäsuorien tai aidosti malliperusteisten estimaattorien etuna on vielä asetelmaperusteisiin verrattuna pienempi varianssi (tai keskineliövirhe MSE), mutta ongelmana on harhaisuus, joka voi olla hyvinkin voimakas erityisesti jälkimmäisten estimaattorien tapauksessa, erityisesti silloin, kun valittu malli on väärä. Estimaattorin valinta pienalue-estimoinnissa on eräänlaista tasapainoilua varianssin ja harhan välillä. Tärkeää on myös se, miten voimakkaasti varianssi tai keskineliövirhe pienenee, kun otoskoko kasvaa. Joka tapauksessa on näyttöä sen käsityksen puolesta, että malliperusteinen estimaattori on asetelmaperusteista parempi, kun mittareina käytetään keskineliövirhettä (MSE) tai tarkkuutta mittavaa keskimääräistä suhteellista virhettä (*ARE* tai *RRMSE*). Yleisesti käytössä olevia alue-estimoinnin laatumittareita analysoidaan myöhemmin tässä tutkimuksessa, jolloin vertaillaan erilaisten otoskiintiöintien vaikutusta niihin.

3.5 Aluemalliin pohjautuva EBLUP-estimointi

3.5.1. Estimointimenetelmän yleisyys ja soveltuvuus

EBLUP *(Empirical Best Linear Unbiased Predictor*) on yleiskäsite kokonaiselle estimaattoriperheelle, jonka taustalla voi olla toisistaan poikkeavia, tilannesovitteisia aluemallituksia. Toinen vastaavanlainen perhe on EB- eli Empirical Bayes –estimaattorit. EBLUP-estimoinnin pohjana on jo kymmeniä vuosia vanha yleinen lineaarinen malli (alaluku 3.5.2), johon sisältyy kiinteä efekti ja satunnaisvaikutus. Tästä menetelmästä on ilmestynyt artikkeleita jo 1990-luvun alkupuolelta lähtien, ja sen keskeisistä lähteistä voidaan mainita aiemmin mainitut Raon kirja (2003) sekä seuraavat artikkelit: Prasad ja Rao (1990), joiden artikkelissa esiteltiin estimointimenetelmä ensimmäisen kerran seikkaperäisesti, Ghosh ja Rao (1994), Torabi ja Rao (2008), Jiang ja Lahiri (2006) sekä Meza ja Lahiri (2005). Nissisen väitöskirjassa (2009) EBLUP-estimointimenetelmä on keskeinen työkalu.

EBLUP-estimointia käytetään tänä päivänä monipuolisesti erilaisissa tutkimuksissa esimerkiksi maataloudessa (mm. satomäärien arviointi), yhteiskuntatieteissä (esim. köyhyys- ja työvoimatutkimukset) ja biologiassa (esim. eri eläinkantojen arviointi). Näissä tutkimuksissa esiintyy sekä suuria että pieniä alueita, joille kaikille on saatava luotettavat ennusteet. Yksittäisen alueen otoskoko voi olla erittäin pieni tai jopa nolla.

EBLUP-estimointi soveltuu mallipohjaisena menetelmänä tilanteisiin, joissa alueiden välillä on havaittavissa merkittävää vaihtelua alueiden sisäisen vaihtelun lisäksi. Malliin kuuluu piilevä (ei

näy havaintoaineistossa) aluevaikutuskomponentti, joka voidaan kuitenkin estimoida havaintoaineistosta (Rao 2003).

3.5.2 Lineaarinen sekamalli

Tässä tutkimuksessa käytettävä lineaarinen hierarkkinen sekamalli, joka esitellään alaluvussa 3.5.3, on erikoistapaus hyvin tunnetusta lineaarisesta sekamallista, joka määritellään matriisimuodossa seuraavasti:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zv} + \boldsymbol{e}\,. \tag{3.8}$$

Tässä mallissa n×1 -vektori $\boldsymbol{y}$ sisältää vastemuuttujan havaintoarvot, $\boldsymbol{\beta}$ on kiinteiden vaikutusten p×1 -vektori, $\boldsymbol{X}$ on selittävien muuttujien arvot sisältävä täyden asteen n×p –matriisi sekä q×1 -vektori $\boldsymbol{v}$ ja n×1-vektori $\boldsymbol{e}$ edustavat satunnaisvaikutuksia ja satunnaisvirheitä. Malliin kuuluu vielä matriisi $\boldsymbol{Z}$, joka liitetään satunnaisvektoriin $\boldsymbol{v}$ ja jonka rakenne ja koko ovat malliin sopivia. Matriisi $\boldsymbol{Z}$ sisältää usein vain nollia ja ykkösiä, jolloin se määritellään insidenssimatriisiksi, mutta se voi joskus sisältää myös selittäviä muuttujia, jotka normaalisti sisältyvät matriisiin $\boldsymbol{X}$. Jos $\boldsymbol{Z}$ sisältää vain nollia ja ykkösiä sekä vektorin v sisältämät satunnaisvaikutukset ovat keskenään korreloimattomia, kutsutaan tätä mallia varianssikomponenttimalliksi. Vektoreista $\boldsymbol{v}$ ja $\boldsymbol{e}$ oletetaan, että E($\boldsymbol{v}$) = E($\boldsymbol{e}$) = $\mathbf{0}$ ja että niiden kovarianssimatriisit ovat Cov($\boldsymbol{v}$) = $\boldsymbol{G}$ ja Cov($\boldsymbol{e}$) = $\boldsymbol{R}$, jotka puolestaan riippuvat eräistä varianssiparametreista $\boldsymbol{\delta} = (\delta_1,...,\delta_q)'$. Edelleen Cov($\boldsymbol{v},\boldsymbol{e}$) = $\boldsymbol{0}$ eli $\boldsymbol{v}$ ja $\boldsymbol{e}$ ovat toisistaan riippumattomia. Varianssikomponenttimallissa matriisit $\boldsymbol{G}$ ja $\boldsymbol{R}$ ovat diagonaalisia.

Edellisten oletusten voimassa ollessa $\boldsymbol{y}$:n odotusarvo on

$$E(\boldsymbol{y}) = \boldsymbol{X\beta}\,,$$

ja $\boldsymbol{y}$:n odotusarvo ehdolla $\boldsymbol{v}$ on

$$\mathrm{E}(\boldsymbol{y} \mid \boldsymbol{v}) = \boldsymbol{X\beta} + \boldsymbol{Zv}\,.$$

Matriisi $\boldsymbol{V}$ = Cov($\boldsymbol{y}$) = $\boldsymbol{R} + \boldsymbol{ZGZ'}$ on $\mathbf{y}$:n varianssi-kovarianssimatriisi, ja jos $\boldsymbol{v}$ on annettu, on voimassa Cov($\boldsymbol{y} \mid \boldsymbol{v}$) = $\boldsymbol{R}$.

Mallin (3.8) kiinteä osa $\boldsymbol{X\beta}$ määrittelee $\boldsymbol{y}$:n keskiarvorakenteen ja satunnaisosa $\boldsymbol{Zv} + \boldsymbol{e}$ kovarianssirakenteen. Oikeilla satunnaisosan sekä kovarianssimatriisien $\boldsymbol{G}$ ja $\boldsymbol{R}$ täsmennyksillä voidaan määritellä lukuisa joukko erilaisia kovarianssirakenteita. Lineaariset sekamallit antavat täten vankan taustan tilastollisen datan mallintamiseen.

Satunnaistermit $\boldsymbol{v}$ ja $\boldsymbol{e}$ oletetaan tavallisesti normaalisti jakautuneiksi:

$$\boldsymbol{v} \sim N_q(\boldsymbol{0}, \boldsymbol{G}) \text{ ja } \boldsymbol{e} \sim N_n(\boldsymbol{0}, \boldsymbol{R}) .$$

Nämä satunnaistermit ovat itse asiassa piilomuuttujia eivätkä siis näy havaintoaineistossa. Normaalisuusoletus on välttämätön, jos käytetään uskottavuuteen perustuvia estimointimenetelmiä (ML, REML). Kovarianssimatriisit $\boldsymbol{G}$ ja $\boldsymbol{R}$ ovat tässä skalaarimuodossa olevien varianssiparametrien $\boldsymbol{\delta}$ funktioita. Normaalioletuksista seuraa, että myös $\boldsymbol{y}$ noudattaa normaalijakaumaa:

$$\boldsymbol{y} \sim \mathrm{N}_n(\boldsymbol{X\beta}, \boldsymbol{ZGZ'} + \boldsymbol{R}) . \tag{3.9}$$

Mallista estimoidaan ensimmäisenä varianssiparametrit $\boldsymbol{\delta}$ ja kiinteiden vaikutusten vektori $\boldsymbol{\beta}$. Varianssi-kovarianssimatriisi $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ'}$ voidaan kirjoittaa muotoon $\mathbf{V}(\boldsymbol{\delta})$, koska se riippuu varianssiparametreista, jotka tarkoittavat tässä yksinkertaisesti variansseja ja kovariansseja.

Yleisessä lineaarisessa mallissa regressiokertoimet $\boldsymbol{\beta}$ voidaan estimoida pienimmän neliösumman menetelmällä:

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X'V^{-1}X})^{-1}\boldsymbol{X'V^{-1}y}, \tag{3.10}$$

joka on $\boldsymbol{\beta}$:n harhaton estimaattori ja jota kutsutaan BLUE-estimaattoriksi (*Best Linear Unbiased Estimator*). Jos kovarianssimatriisi $\boldsymbol{V}(\boldsymbol{\delta}) = \boldsymbol{R}(\boldsymbol{\delta}) + \boldsymbol{ZG}(\boldsymbol{\delta})\boldsymbol{Z'}$ on tuntematon, se korvataan otosaineistosta tuotetulla estimaatilla $\boldsymbol{V}(\hat{\boldsymbol{\delta}}) = \boldsymbol{R}(\hat{\boldsymbol{\delta}}) + \boldsymbol{ZG}\ (\hat{\boldsymbol{\delta}})\boldsymbol{Z'}$, jossa $\hat{\boldsymbol{\delta}}$ saadaan jollakin sopivalla menetelmällä (alaluku 3.5.4).

Normaalisuusoletuksen (3.9) perusteella voidaan estimoinnissa käyttää $\boldsymbol{y}$:n tiheysfunktiosta johdettua uskottavuusfunktiota ja sen logaritmia, joka maksimoidaan parametrivektorin $\boldsymbol{\beta}$ suhteen derivoinnin avulla. Tällöin saadaan $\boldsymbol{\beta}$:lle suurimman uskottavuuden (ML) –estimaatori. Jos $\boldsymbol{V}$ on tunnettu, saadaan $\boldsymbol{\beta}$:n ML-estimaattori $\hat{\boldsymbol{\beta}}_{ML}$, ja jos $\boldsymbol{V}$ on tuntematon, on estimaattori

$$\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X'}\hat{\boldsymbol{V}}_{ML}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'}\hat{\boldsymbol{V}}_{ML}^{-1}\boldsymbol{y}.$$

Varianssi-kovarianssimatriisi $\mathbf{V}(\boldsymbol{\delta})$ estimoidaan maksimoimalla sen suurimman uskottavuuden lauseke jollakin numeerisella optimointimenetelmällä (mm. Nissinen 2009).

Toinen uskottavuusfunktion optimointiin perustuva estimointimenetelmä on REML, joka on kehitetty 1970-luvun alussa (esim. Nissinen 2009). Se perustuu vastemuuttujan $\boldsymbol{y}$ sellaiseen lineaariseen muunnokseen, että tuloksena oleva jakauma ei riipu $\boldsymbol{\beta}$:sta, joka jää pois uskottavuusfunktiosta. Tämän johdosta menetettävät vapausasteet otetaan huomioon estimoitaessa matriisia

$\mathbf{V}(\boldsymbol{\delta})$. REML-menetelmä tuottaa harhattoman tai lähes harhattoman varianssiestimaattorin. Sillä on samat hyvät ominaisuudet kuin ML-estimaattorilla, jolla puolestaan on eräitä merkittäviä puutteita, ja käytettävät laskentaoperaatiot eivät ole olennaisesti monimutkaisempia. Tästä syystä REML on suosittu estimointimenetelmä. Yleisesti käytössä olevat tilasto-ohjelmistot (SAS ja SPSS) sisältävät sekä ML- että REML-menetelmän.

Varianssikomponenttien $\boldsymbol{\delta}$ (satunnais- ja aluevaihtelu) estimoinnissa voidaan käyttää myös 1950-luvulla kehitettyä Henderson 3 –menetelmää, joka perustuu lineaaristen regressiomallien soveltamiseen vaste- ja apumuuttujan otosarvoihin sekä otosarvojen poikkeamiin aluekeskiarvoista. Menetelmä on klassisen ANOVA-menetelmän sovellus sekamallitapaukseen.

Satunnaisvaikutukset $\boldsymbol{v}$ eivät itse asiassa ole oikeita parametreja, vaikka ne käyttäytyvät kuten parametrit, ja koska niitä ei voi havaita konkreettisesti, ne on estimoitava havainnoista. Pienalueestimoinnissa on kysymys näiden kohdalla aluevaikutuksista, kun ennustetaan sopivan mallin avulla alueiden kokonaismääriä ja keskiarvoja. Normaalisuusoletuksen ollessa voimassa mallissa (3.8) on $\boldsymbol{v}$:n odotusarvo, kun $\boldsymbol{y}$ on tunnettu,

$$\mathrm{E}(\boldsymbol{v} \mid \boldsymbol{y}) = \boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \tag{3.11}$$

ja se on samalla $\boldsymbol{v}$:n paras lineaarinen ennustin. Käytännössä lausekkeessa (3.10) oleva tuntematon $\tilde{\boldsymbol{\beta}}$ korvataan sen estimaattorilla $\hat{\boldsymbol{\beta}}$, jolloin saadaan $\boldsymbol{v}$:n paras empiirinen lineaarinen harhaton ennustin (*EBLUP).* Sen matriisimuotoinen lauseke on

$$\hat{\boldsymbol{v}} = \boldsymbol{G}\,\boldsymbol{Z}'\hat{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{3.12}$$

3.5.3 Hierarkkinen lineaarinen sekamalli

Tässä tutkimuksessa käytettävä malli on yksikkötason varianssikomponenttimalli

$$y_{dk} = \boldsymbol{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk};\ k = 1,\ldots,\ N_d;\ d = 1,\ldots, D. \tag{3.13}$$

Mallin perusedellytyksenä on, että kaikkien apumuuttujien $\boldsymbol{x} = x_1, x_2, ..., x_p$ arvot ovat käytettävissä alueen *d* perusjoukon jokaiselle tilastoyksikölle *k.*

Mallissa (3.13) y_{dk} tarkoittaa vastemuuttujan *y* havainnon *k* arvoa alueella *d,* $\mathbf{x}_{dk}$ on apumuuttujien arvojen muodostama vektori alueella *d,* v_d on alueen *d* (*d* = 1,…, *D*) satunnaisvaikutus

mallissa ja estimoidaan havainnoista, ja e_{dk} edustaa mallin satunnaisvirhettä. Aluevaikutukset v_d ovat riippumattomia satunnaismuuttujia, joiden keskiarvo on nolla ja yhteinen varianssi σ_v^2. Vastaavasti satunnaisvirheet e_{dk} ovat riippumattomia satunnaismuuttujia, joiden keskiarvo on nolla ja yhteinen varianssi σ_e^2. Lisäksi v_d ja e_{dk} oletetaan toisistaan riippumattomiksi ja niiden oletetaan tavallisesti noudattavan normaalijakaumaa. Varianssit σ_v^2 ja σ_e^2, regressiokertoimet $\boldsymbol{\beta}$ ja aluevaikutukset v_d estimoidaan otoshavainnoista yleisen lineaarisen mallin (3.8) mukaan. Otoshavaintoja ei tietenkään ole käytettävissä kiintiöintivaiheessa.

Edellä tehtyjen oletusten mukaisesti saadaan vastemuuttujan *y* yksittäisen havaintoarvon odotusarvolle ja varianssille seuraavat lausekkeet:

$$E(y_{dk}) = \boldsymbol{x}'_{\boldsymbol{dk}}\boldsymbol{\beta} \text{ ja } V(y_{dk}) = \sigma_v^2 + \sigma_e^2 . \tag{3.14}$$

Estimoinnissa tarvitaan seuraavia matriiseja:

1) Vastemuuttujan *y* otosarvot sisältävä *n*×1 -vektori **y**,

2) Apumuuttujien $\boldsymbol{x} = x_1, x_2, ..., x_p$ otosarvojen muodostama n×(p+1) -matriisi **X**, jonka ensimmäinen sarake koostuu ykkösistä

3) Vastemuuttujan *y* varianssi-kovarianssimatriisi ***V***, jonka koko on $n \times n$ ja joka sisältää havaintoaineistosta estimoitavat alue- ja satunnaisvarianssin,

4) Insidenssimatriisi ***Z*** (malli 3.8),

5) Aluevaikutusvektori ***v***, joka estimoidaan kaavan (3.12) mukaisesti.

Varianssi-kovarianssimatriisi ***V***, jolla on lävistäjämatriisirakenne, määritellään seuraavasti:

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V_1} & 0 & \dots & \dots & 0 \\ 0 & \boldsymbol{V_2} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \boldsymbol{V_q} \end{bmatrix}, \tag{3.15}$$

joka muodostuu neliöalimatriiseista $\boldsymbol{V_d}$ (d = 1,2,…,h) ja nollista. Alimatriisin d koko on $n_d \times n_d$, ja $\sum_d n_d = n$. Alimatriisin $\boldsymbol{V_d}$ sisältö on seuraavanlainen:

$$\mathbf{V_d} = \begin{bmatrix} \sigma_v^2 + \sigma_e^2 & \sigma_v^2 & \dots & \sigma_v^2 \\ \sigma_v^2 & \sigma_v^2 + \sigma_e^2 & \dots & \sigma_v^2 \\ \dots & \dots & \dots & \dots \\ \sigma_v^2 & \dots & \sigma_v^2 & \sigma_v^2 + \sigma_e^2 \end{bmatrix}. \qquad (3.16)$$

Päälävistäjä muodostuu kokonaisvarianssista ja muut elementit vain aluevaikutusvarianssista.

Estimoinnissa tarvitaan mallin (3.8) mukaisesti myös $n \times h$ -matriisia ***Z***, jonka rivien määrä n = kokonaisotoskoko ja sarakkeiden määrä h = niiden alueiden lukumäärä, joilta on havaintoja. Ykkösten määrä sarakkeella d $(1 \le d \le h)$ = alueen d havaintojen määrä.

3.5.4 Estimoitavat suureet ja ennusteet

Estimoinnin ensimmäinen vaihe on alue- ja satunnaisvarianssien σ_v^2 ja σ_e^2 estimointi, jossa voidaan käyttää useita eri menetelmiä, kuten on edellä todettu (ML, REML, Henderson 3). Näiden antamien tulosten välillä voi olla huomattaviakin eroja (lähes 10 %), mikä ilmeni, kun menetelmiä sovellettiin kokeilumielessä samaan aineistoon. Aluevarianssin σ_v^2 estimaatti voi saada negatiivisen arvon, jolloin se asetetaan nollaksi. Lisäksi on estimoitava em. varianssien asymptoottiset varianssit ja kovarianssi (Rao 2003), mikä sekin on varsin työläs operaatio. Tuloksena saadaan varianssien estimaatit $\hat{\sigma}_v^2$ ja $\hat{\sigma}_e^2$ sekä niiden asymptoottiset varianssit ja kovarianssi (Rao 2003):

$$\begin{aligned} V(\hat{\sigma}_v^2) &= 2\eta_1^{-2}\left[(n-h-1)^{-1} \times (h-1)(n-2)\hat{\sigma}_e^4 + 2\eta_1\hat{\sigma}_e^2\hat{\sigma}_v^2 + \eta_2\hat{\sigma}_v^4\right] \\ V(\hat{\sigma}_e^2) &= 2(n-h-2)^{-1}\hat{\sigma}_e^4 \\ Cov(\hat{\sigma}_e^2, \hat{\sigma}_v^2) &= -(h-1)\eta_1^{-1} Var(\hat{\sigma}_e^2), \end{aligned} \qquad (3.17)$$

missä n = otoskoko, h = niiden alueiden (ositteiden) lukumäärä, joista on havaintoja, ja termit η_1 ja η_2 määritellään matriisimuodossa seuraavasti:

$$\eta_1 = n - tr\left[(\boldsymbol{X}'\boldsymbol{X})^{-1} \sum_{d=1}^{D} n_d^2 \bar{\boldsymbol{x}}_{\boldsymbol{d}} \bar{\boldsymbol{x}}_d'\right]$$

$\eta_2 = tr(\boldsymbol{MZZ}')^2$, missä matriisi ***Z*** on määritelty edellä ja $n \times n$ –matriisi

$\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X(X'X)}^{-1}\boldsymbol{X}'$ (***I*** on identiteettimatriisi).

Komponenttien estimoinnissa tarvitaan lisäksi lauseketta

$$\gamma_d = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 n_d^{-1}) = n_d \sigma_v^2 / (n_d \sigma_v^2 + \sigma_e^2) . \quad (3.18)$$

Termille γ_d saadaan otoksesta estimaatti varianssikomponenttien estimoinnin jälkeen. Tässä tutkimuksessa alue ja osite ovat identtiset. Tähän liittyen Meza ja Lahiri (2005) ovat määritelleet erityisen alueiden (ositteiden) välisen sisäkorrelaation, jonka kaava on seuraava:

$$\varphi = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2) = 1/(1 + \sigma_e^2 / \sigma_v^2) . \quad (3.19)$$

Sisäkorrelaatio mittaa otoksesta sitä, kuinka voimakas alueiden välinen vaihtelu on suhteessa kokonaisvaihteluun. Kun varianssikomponentit korvataan otoksesta lasketuilla estimaateilla, saadaan sisäkorrelaatiolle φ estimaatti:

$$\hat{\varphi} = \hat{\sigma}_v^2 / (\hat{\sigma}_v^2 + \hat{\sigma}_e^2) = 1/(1 + \hat{\sigma}_e^2 / \hat{\sigma}_v^2) .$$

Lausekkeesta (3.19) nähdään helposti, että sisäkorrelaatio vaihtelee välillä 0 - 1. Korrelaatio jää nollaksi, kun alueiden välistä vaihtelua ei ole, eli kun aluevaihtelukomponentin σ_v^2 arvo on nolla. Kun varianssikomponentit ovat yhtä suuria, on sisäkorrelaation arvo 0.5. Mitä suurempi on aluevaihtelun osuus kokonaisvaihtelusta, sitä lähempänä ykköstä sisäkorrelaatio on, mutta tämä tilanne lienee harvinaista, koska satunnaisvaihtelua on aina olemassa. Käytännössä sisäkorrelaatiota ei tiedetä, koska se riippuu otoksen vaste- ja apumuuttujan arvoista. Koska sisäkorrelaatiota kuitenkin tarvitaan alaluvussa 5.5 esiteltävän alueoptimaalisen *g1*-kiintiöinnin johtamisessa, se korvataan alaluvussa 6.1.5 kuvatulla apumuuttujan homogeenisuusmitalla (lauseke 6.1 ja taulukko 6.3).

Regressiokertoimet $\boldsymbol{\beta}$ ja aluevaikutukset $\boldsymbol{v}$ estimoidaan havaintoaineistosta lausekkeiden (3.10) ja (3.12) mukaisesti, jolloin saadaan niiden estimaatit $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{v}}$. $\hat{\boldsymbol{\beta}}$:n analyyttinen lauseke on monimutkainen siinäkin tapauksessa, että käytössä on vain yksi apumuuttuja. Sen sijaan yksittäisen tilastoyksikön EBLUP-ennusteeseen sisältyvä estimoitu aluevaikutus $\hat{v}_d$ on yhden apumuuttujan tapauksessa varsin yksinkertainen:

$$\hat{v}_d = \gamma_d (\bar{y}_d - \hat{\boldsymbol{\beta}} \bar{x}_d) . \quad (3.20)$$

Vastemuuttujan *y* kokonaismäärän Y_d ennustaminen alueelle *d* on itse asiassa otoksen ulkopuolisten *y*-arvojen summan ennustamista. Summaennuste lasketaan apumuuttujien, regressiokertoimien $\boldsymbol{\beta}$ ja aluevaikutusten v_d avulla. Mikäli joltakin alueelta ei ole havaintoja, saadaan sille joka tapauksessa ennuste. Yhdestä otoksesta laskettu vastemuuttujan *y* kokonaismäärän BLUP-

ennuste alueelle *d* on yksinkertaisesti *y*:n otosarvojen summan ja otoksen ulkopuolisten *y*-arvojen summan ennusteen summa:

$$\hat{Y}_{d,BLUP} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \tilde{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \boldsymbol{x}'_{dk}\tilde{\boldsymbol{\beta}} + (N_d - n_d^*)\tilde{v}_d \,. \qquad (3.21)$$

Kaavassa (3.21) on termi N_d kiinteä alueen *d* tilastoyksiköiden lukumäärä ja menetelmäspesifi termi n_d^* tapauksesta riippuva alueelle *d* tuleva otoskoko (ei kiinteä). Kun BLUE-estimaattorit $\tilde{\boldsymbol{\beta}}$ ja $\tilde{v}_d$ korvataan otosestimaateillaan $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{v}}_d$, saadaan summalle EBLUP-ennuste, jota merkitään tässä lyhyesti symbolilla $\hat{Y}_{d,EBLUP}$. Merkintä ”s_d” lausekkeessa (3.21) tarkoittaa alueelta *d* saatavaa otosta ja ”$\bar{s}_d$” otoksen ulkopuolisia tilastoyksiköitä. Edelleen lausekkeessa esiintyvät apumuuttujien $\boldsymbol{x} = (x_1, x_2, ..., x_p)'$ summat, mutta yksittäisiä *x*-arvoja ei tarvita. Aluevaikutus liitetään jokaiseen otokseen s_d kuulumattomaan tilastoyksikköön.

Koska alueen *d* ennuste (3.21) on harhainen, käytetään ennusteen keskivirheen tilalla sen keskineliövirhettä MSE, joka on MSE:n määritelmän mukaan estimaattorin varianssin ja harhan neliön summa:

$$MSE(\hat{Y}_{d,EBLUP}) = E(\hat{Y}_{d,EBLUP} - Y_d)^2 = V(\hat{Y}_{d,EBLUP}) + (E(\hat{Y}_{d,EBLUP}) - Y_d)^2 \,. \qquad (3.22)$$

Vastemuuttujan *y* kokonaismäärän MSE:n estimoinnissa käytetään äärelliselle perusjoukoille kehitettyä Prasad-Rao -approksimaatiota (Prasad-Rao 1990, Rao 2003)

$$mse(\hat{Y}_{d,EBLUP}) = g_{1d}(\sigma_v^2, \sigma_e^2) + g_{2d}(\sigma_v^2, \sigma_e^2) + 2g_{3d}(\sigma_v^2, \sigma_e^2) + g_{4d}(\sigma_v^2, \sigma_e^2) \,, \quad (3.23)$$

jonka neljä komponenttia määritellään seuraavasti:

$$\begin{aligned}
&g_{1d}(\sigma_v^2, \sigma_e^2) = (N_d - n_d^*)^2 (1 - \gamma_d)\sigma_v^2 \,, \\
&g_{2d}(\sigma_v^2, \sigma_e^2) = (N_d - n_d^*)^2 (\bar{\boldsymbol{x}}_{\boldsymbol{d}}^* - \gamma_d \bar{\boldsymbol{x}}_{\boldsymbol{d}})' (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1} (\bar{\boldsymbol{x}}_{\boldsymbol{d}}^* - \gamma_d \bar{\boldsymbol{x}}_{\boldsymbol{d}}), \\
&g_{3d}(\sigma_v^2, \sigma_e^2) = \quad (N_d - n_d^*)^2 n_d^{-2} (\sigma_v^2 + \sigma_e^2 / n_d^*)^{-3} [\sigma_e^4 V(\hat{\sigma}_v^2) + \sigma_v^4 V(\hat{\sigma}_e^2) \\
&- 2\sigma_e^2 \sigma_v^2 Cov(\hat{\sigma}_e^2, \hat{\sigma}_v^2)], \\
&g_{4d}(\sigma_v^2, \sigma_e^2) = (N_d - n_d^*)\sigma_e^2 \,.
\end{aligned} \qquad (3.24)$$

Komponentin g_{3d} kaavassa esiintyvä termi $\bar{\boldsymbol{x}}_{\boldsymbol{d}}^*$ tarkoittaa apumuuttujien ***x*** otoksen ulkopuolisten havaintoarvojen keskiarvomatriisia.

MSE:n approksimaation (3.23) estimaatti saadaan sijoittamalla lausekkeisiin (3.24) otoksesta laskettavien varianssikomponenttien σ_v^2 ja σ_e^2 estimaatit $\hat{\sigma}_v^2$ ja $\hat{\sigma}_e^2$. Kaikki kaavoihin kuuluvat osat on määritelty aikaisemmin. Komponentit kuvaavat erilaisia epävarmuuksia seuraavasti (mm. Rao 2003): g_{1d} liittyy aluevaikutuksiin v_d, g_{2d} regressiokertoimiin $\boldsymbol{\beta}$, g_{3d} varianssiparametreihin $\boldsymbol{\delta}$ ja g_{4d} alueen *d* otoksen ulkopuolisiin *y*-arvoihin liittyvään epävarmuuteen.

Alue-ennusteen $\hat{Y}_{d,EBLUP}$ vaihtelukerroin (CV) suhteuttaa ennusteeseen liittyvän epävarmuuden itse ennusteeseen ja on prosenttilukuna ilmaistava ennusteen MSE-approksimaation neliöjuuren ja ennusteen suhde:

$$\mathrm{CV}(\hat{Y}_{d,EBLUP})\% = 100 \times \sqrt{mse\,(\hat{Y}_{d,EBLUP})} / \hat{Y}_{d,EBLUP} \ . \tag{3.25}$$

Jos lähtökohtana on EBLUP-estimaattorien asymptoottinen normaalisuus, voidaan alue-ennusteen ja sen estimoidun keskineliövirheen avulla laskea 95 %:n luottamusväli (CI_{95}):

$$\mathrm{CI}_{95}(\hat{Y}_{d,EBLUP}) = \hat{Y}_{d,EBLUP} \pm 1.96 \times \sqrt{mse\,(\hat{Y}_{d,EBLUP})}. \tag{3.26}$$

Komponentin g_{1d} %-osuus MSE:n arvosta on ollut monissa alue-estimointiin liittyvissä tutkimuksissa raporttien mukaan yleisesti yli 90 %, kuten Nissinen (2009) toteaa. Muiden komponenttien osuudet ovat olleet luonnollisesti pieniä. Tämän tutkimuksen simuloinnit erilaisten otoskiintiöintien mukaan vahvistavat Nissisen toteamusta. Korkea %-osuus edellyttää voimakasta alueiden välistä vaihtelua, koska g_{1d}-komponentti mittaa juuri tämän vaihtelun epävarmuutta. Komponentin merkitys tulee esille myöhemmin, kun aluekiintiöintiin etsitään analyyttista ratkaisua otoskokojen n_d^* (*d = 1,2,..., D)* funktiona ilmaistavan komponentin aluekohtaisten arvojen keskiarvon minimoinnin avulla.

3.5.5 Apumuuttujan alueominaisuuksien ja otoskokojen vaikutus estimointituloksiin

Alue-ennusteiden luotettavuuden arvioinnissa voidaan valitun mallin (3.13) ja EBLUP-estimointimenetelmän perusteella pitää tärkeimpinä kriteereinä alue-ennusteiden MSE-arvoja, alue-ennusteista ja MSE-arvoista laskettuja CV-arvoja sekä alue-ennusteiden suhteellista virhettä ja harhaa. Otossimulointeja käyttävissä tutkimuksissa voidaan ottaa huomioon myös se, kuinka hyvin vastemuuttujan todelliset aluekeskiarvot tai -summat sijoittuvat ennusteista laskettavien luottamusvälien sisälle. EBLUP-ennusteen lausekkeen (3.21) sekä MSE:n lausekkeen approk-

simaation (3.23) ja sen komponenttien lausekkeiden (3.24) perusteella voidaan päätellä asioita, joita otoskiintiöinnissä tulee ottaa huomioon.

Jos alueen *d* otoskoko $n_d^* = 0$, saa komponentti g_{3d} arvon nolla, ja myös alueen *d* aluevaikutuskomponentti v_d jää nollaksi. Tällainen alue on useimmiten pieni (vähän tilastoyksiköitä). Alueen estimointi onnistuu, mutta ennuste on täysin muilta alueilta tulevan informaation varassa ja perustuu vain synteettiseen regressio-osaan (kaavassa 3.21). Jos apumuuttujan alue-ominaisuudet (keskiarvo ja CV) ovat lähellä perusjoukon vastaavia ja apumuuttujan vaihteluväli ei ole kovin suppea, on mahdollista saada hyvinkin tarkka alue-ennuste vastemuuttujalle, mutta otoskoko nolla johtaa alueella hyvin todennäköisesti suuriin MSE- ja CV-arvoihin, ja sitä varmemmin, mitä enemmän apumuuttuja poikkeaa ominaisuuksiltaan muista alueista.

Alueiden kokonaismäärien estimoinnin erikoistapaus on se, että aluevarianssikomponentin σ_v^2 estimaatti jää nollaksi, mistä Raon (2003) mukaan aiheutuu suuria ongelmia. Tällöin sekä g_{1d} että g_{3d} tulevat nolliksi. MSE koostuu nyt vain kahdesta muusta komponentista, joten regressiokertoimissa ja alueen ulkopuolisissa havainnoissa on entistä enemmän epävarmuutta. Aluevaikutuksia ei voida estimoida, joten vastemuuttujan alue-ennusteiden laskennassa käytetään vain mallin regressio-osaa, jonka kertoimet supistuvat tässä tapauksessa tavallisiksi pns-kertoimiksi. MSE:lle saadaan kyllä yleensä pieni arvo, mutta alue-ennusteissa paljon virhettä, koska regressiokertoimesta puuttuu varianssikomponenttien vaikutus. Jos on syytä olettaa vaihtelua alueiden välille, vaihtelun tulisi näkyä myös otoksessa.

Koska otoskoko *n* on usein pieni suhteessa perusjoukon kokoon *N*, on apumuuttujalla *x* keskeinen osuus alue-ennusteessa. Jokaisen yksittäisen alueen apumuuttujan ominaisuuksia kannattaa tarkastella ja verrata perusjoukon vastaaviin sekä muiden alueiden ominaisuuksiin. Jos ominaisuudet ovat lähellä perusjoukon vastaavia, saattaa alueen suhteellista osuutta pienempi otoskoko riittää. Jos sen sijaan voimakkaita poikkeamia apumuuttujan ominaisuuksissa suhteessa perusjoukkoon ja muihin alueisiin esiintyy, on otoskokoa pohdittava erityisen huolellisesti.

Alueelle, jonka apumuuttujan keskiarvo on selvästi muita alueita alhaisempi ja vaihtelu vähäistä, kannattaa harkita otoskokoa, joka on alueen suhteellista osuutta suurempi. Jos tällaisen alueen otoskoko on pieni, saattaa muiden alueiden vaikutus alue-ennusteeseen olla liian voimakas, mistä on hyvin todennäköisesti seurauksena suuri ennustevirhe, MSE- ja CV-arvo. Muut alueet ikään kuin ”vetävät” niistä hyvin poikkeavaa aluetta väärään suuntaan.

Jos alueen apumuuttujan keskiarvo ja vaihtelu ovat selvästi muita alueita suurempia, on tälle alueelle harkittava otoskokoa, joka on pienempi kuin sen suhteellista kokoa vastaava. Jos otoskoko on suuri, voi tämä alue vaikuttaa vääristävästi muiden alueiden ennusteisiin siten, että niistä tulee liian suuria, mistä puolestaan voi olla seurauksena mm. suuret MSE- ja CV-arvot.

Yksi erikoistilanne on se, että pieniä estimoitavia alueita on useita. Nämä voivat olla ominaisuuksiltaan lähellä muita alueita ja perusjoukkoa tai poiketa niistä hyvinkin suuresti (matala keskiarvo ja pieni vaihtelu, suuri keskiarvo ja voimakas vaihtelu tai näiden välimuoto). Niiden otoskoot ovat tuskin koskaan kovinkaan suuria, ellei sitten otantaresurssia haluta käyttää niihin jonkin erityisen syyn takia.

Lopullinen otoskiintiöinti on aina enemmän tai vähemmän kompromissi useiden eri tavoitteiden välillä, ja kaikkien alueiden kohdalla tuskin koskaan päästään samanaikaisesti optimaaliseen lopputulokseen. mutta kaikille alueille on tavoitteena saada edes kohtuullisen hyvät tulokset. Kriteerinä voi olla esimerkiksi se, että alueiden CV-arvoille, suhteellisille virheille ja harhoille on ennalta määrätty rajat, joita ne eivät saa ylittää. Joskus tavoitellaan myös näiden keskiarvojen minimointia, mutta EBLUP-estimoinnissa ei analyyttinen ratkaisu ole mahdollinen lausekkeiden monimutkaisuuden vuoksi, vaan niiden tilalle on kehitettävä muita vaihtoehtoja, kuten tässä tutkimuksessa on tehty. Se on tietysti oma kysymyksensä, miten hyvät tulokset määritellään. Otoskiintiöinnin suunnitteluun kannattaa joka tapauksessa käyttää aikaa eikä kannata välttämättä pyrkiä kiintiöintiongelmasta heti eroon päätymällä suoraan perinteisiin ratkaisuihin, kuten suhteelliseen tai tasakiintiöintiin.

3.6 Erilaisten estimaattorien tehokkuuden vertailu

Monissa tutkimuksissa on vertailtu erityyppisten estimaattorien tehokkuutta. Alue-estimoinnin yksi peruskysymys on, onko malliperusteinen estimointi parempi kuin suora estimointi. Mielipide näyttää olevan vahvasti sidoksissa siihen, kumpaa sen esittäjä pääasiassa käyttää.

Malec ym. (1999) ovat kuvanneet tutkimusta, jossa sovellettiin malliperusteista analyysiä estimoitaessa aikuisten ylipainon yleisyyttä USA:n osavaltioissa. Tutkimuksessa oli otettu huomioon myös se tieto, miten otos on poimittu ja muutettu poimintatodennäköisyydet Bayes-uskottavuuksiksi. Vastemuuttuja oli painoindeksi (paino/pituus^2). Aineistona oli aiemmin kerätty terveys- ja ravintotutkimuksen (NHANES III) aineisto. Käytössä oli kaksivaiheinen hierarkkinen logit-malli, jonka avulla pyrittiin saamaan selville alueellinen osavaltio-piirikunta -vaihtelu.

Otantamenetelmänä oli ryväsotanta. Myös asetelmaperusteista ja synteettistä estimointimenetelmää sovellettiin, ja estimointituloksia vertailtiin kansallisella tasolla. Malliperusteista estimointia sovellettiin ilman otanta-asetelmakorjausta ja sen kanssa. Estimointi oli Bayes-tyyppinen, ja laskelmat tehtiin SAS –ohjelmistolla. Iteraatiokierrokset toteutettiin Gibbsin algoritmin avulla. Estimointiin käytettiin 16 523 painoindeksiarvoa, ja taustamuuttujat olivat demografisia ominaisuuksia (väestöryhmä, ikäryhmä jne.).

Tuloksista voidaan mainita mm. seuraavaa:

1) Poimintatodennäköisyyksien korjauksella saatiin tarkempia estimaatteja vastemuuttujalle.
2) Alue-estimaatit, jotka perustuivat vain oman alueen informaatioon, olivat epätarkempia kuin estimaatit, jotka perustuivat myös muilta alueilta lainattuun informaatioon.
3) Jotta alue-estimointi ei tasoita liikaa estimaatteja, tarkistettiin mallin sisältämä vaihtelu otosaineiston vaihtelun kanssa.
4) Malliperusteiset estimaatit olivat synteettisiin verrattuina lähempänä asetelmaperusteisia estimaatteja osavaltioissa, joista oli poimittu paljon havaintoja otokseen.
5) Malliperusteinen posteriorivarianssi pienenee otoskoon kasvaessa.
6) Malliperusteisella estimoinnilla on taipumus käyttää vain yhden alueen dataa.
7) Synteettinen estimointi oli artikkelin tutkimuksessa riittävä, mutta hierarkkisen mallin käyttö oli tarpeen vahvistamaan johtopäätös.
8) Suuren otoskoon saaneille alueille tuli lähellä toisiaan olevat estimaatit riippumatta siitä, millaisella menetelmällä ne oli saatu.
9) Vastemuuttuja vaihteli alueittain vähän: ylipainoisten estimoitu osuus oli 32–40 %, ja erot näkyivät ulottuvuudella pohjoiset–eteläiset osavaltiot.

Pfefferman ja Sverchkov (2004) vertailivat tutkimuksessaan erilaisten estimaattorien tehokkuuksia. Tässä tutkimuksessa ei perusjoukkoa ollut jaettu alueisiin, mutta tutkimuksen merkitys on siinä, että estimoinnissa käytettiin otoksen ulkopuolisten havaintojen ennusteita apuna sekä erilaisia estimaattoreita, jotka joko eivät käyttäneet tai käyttivät apumuuttujatietoja.

Tutkimus koski vastemuuttujan (y) kokonaismäärän estimointia, sen estimoitua MSE:tä ja mahdollista harhaa. MSE:n estimaattina toimi keskimääräinen neliöity ennustevirhe ja tulosten tarkastelussa käytettiin myös sen neliöjuurta (RMSE). Estimointi perustui vastemuuttujan otoshavaintoihin, otantayksiköiden otantapainoihin ja apumuuttujiin ($\boldsymbol{x}$). Tutkimuksessa johdettiin ensin otantajakauman parametrit ja sitten otantajakauma otoksen ulkopuolisille havainnoille,

joiden arvot ennustettiin ehdollistamalla otoksen ja otantapainojen kanssa. Kolmentyyppisiä estimaattoreita käytettiin: 1) suoria eli ***x***:n arvoja käyttämättömiä kuten Horvitz-Thompson- ja Hajek-estimaattoria, 2) korkeamman asteen regressioestimaattoreita ja GREG-estimaattoria ja 3) lineaarista regressiomallia ennustettaessa syntymäpainon keskiarvoa. Selittävänä muuttujana oli raskausviikkojen määrä. Simuloinnin alustana oli erään aikaisemman tutkimuksen aineisto, jossa oli lähes 10 000 havaintoa. Simuloinnissa otanta toistettiin 1000 kertaa tästä aineistosta kolmella eri otosmäärällä: 232, 1 145 ja 2 429. Tulosten analysoinnin perusteella havaittiin mm. seuraavat seikat: 1) suorat estimaattorit olivat tehottomia pienellä otoskoolla, 2) koko populaation ennusteista saatiin parempia (alhaisemmat RMSE:t), kun estimoinnissa käytettiin otoshavaintojen lisäksi otoksen ulkopuolisten havaintojen ennusteita ja 3) apumuuttujatietoa käyttävät estimaattorit olivat huomattavasti tehokkaampia suoriin verrattuina, varsinkin pienten otoskokojen kohdalla.

Pfeffermanin ja Sverchkovin toisessa tutkimuksessa (2007) oli sikäli uudenlainen lähestymistapa, että alueiden poimintatodennäköisyydet olivat erisuuria ja ne olivat (mahdollisesti) sidoksissa vastemuuttujan todellisiin aluekeskiarvoihin, ja erisuuria olivat myös otokseen poimittujen alueiden tilastoyksiköiden poimintatodennäköisyydet, jotka taas olivat (mahdollisesti) sidoksissa vastemuuttujan arvoihin, siinäkin tapauksessa, että malliin otettiin mukaan kovariaatit (apumuuttujat). Ongelma on siinä, että malli, jota perusjoukon tilastoyksiköt noudattavat, ei olekaan enää voimassa otoshavainnoissa. Termi informatiivinen otanta on kehittynyt tältä pohjalta. Artikkelissa kuvataan sitä, että mikäli informatiivisen otanta-asetelman vaikutuksia ei kyetä selvittämään, on tuloksena harhaisia estimaattoreita, joiden keskineliövirhe MSE kasvaa.

Artikkelissa johdettiin kolmeen eri aluemalliin perustuvat harhattomat estimaattorit, joista yksi oli EBLUP-estimaattori, vastemuuttujan keskiarvoille erikseen pienalueille, joista tuli havaintoja otokseen ja erikseen otokseen kuulumattomille pienalueille. Mallit eivät kuitenkaan sisältäneet aluevaikutuskomponenttia. Lisäksi on kehitetty tilastollinen testi, jonka avulla voidaan testata otanta-asetelman informatiivisuuden hyöty, ts. johtiko otanta-asetelman huomiotta jättäminen suurempiharhaisiin estimaattoreihin. Estimointimenetelmiä testattiin keinotekoisella aineistolla.

Artikkelissa kuvattiin ensin, miten johdetaan otoksen ja perusjoukon jakauma sekä otokseen kuulumattomien havaintojen (lukumäärä = $N - n$) jakauma. Sitten kuvattiin optimaalisten keskiarvoestimaattorien johtaminen otantaan kuuluville ja kuulumattomille alueille, samoin kuin harhan johtaminen silloin, kun informatiivinen otanta-asetelma jätetään huomioon ottamatta.

Aluekeskiarvojen ja niiden MSE-estimaattoreiden johtaminen esitettiin lopuksi täydennettynä otoksen informatiivisuuden testauksen periaatteilla.

Simulointikokeissa generoitiin ensin synteettisesti perusjoukon kolmen (3) ositteen yhteensä 50 alueeseen vastemuuttujan y ja usean apumuuttujan $\boldsymbol{x}$ havainnot ja näille alue-efektit. Otokseen poimittiin PPS-otannalla jokaisesta ositteesta 10 aluetta ja näiden ositteiden sisällä jokaiselta alueelta 5 (osite 1), 25 (osite 2) tai 50 havaintoa (osite 3), yhteensä 800 havaintoa. Jokaiselle alueelle laskettiin tavallinen EBLUP-, GREG-tyyppinen sekä uusi informatiivisuuden huomioon ottava EBLUP-keskiarvoestimaatti sekä näille harhat ja RMSE:t (keskineliövirheen neliöjuuri).

Suoritettujen simulointien perusteella tehtiin mm. seuraavia johtopäätöksiä:

1) Jos informatiivinen otanta-asetelma jätetään huomioon ottamatta, lisääntyy ennusteharha sekä otokseen tulleilla ja sen ulkopuolelle jääneillä alueilla, samoin kuin RMSE:t.
2) Asetelmaperusteiset estimaattorit ovat lähes harhattomia alueilla, joiden otoskoko oli riittävän iso (esimerkkitapauksessa 25). Tämä ei liene mikään uusi tieto.
3) Uudenlaiset keskiarvoestimaattorit olivat kirjaimellisesti harhattomia molemman tyyppisillä alueilla (kohta 1).
4) Uuden keskiarvoestimaattorin RMSE:t olivat otantaan kuulumattomilla alueilla korkeammat verrattuina otantaan sisältyneiden alueiden RMSE-arvoihin, mutta pienempiä kuin muun tyyppisten estimaattorien vastaavat arvot. Informatiivisen otanta-asetelman mukaan lasketut estimoitujen aluekeskiarvojen luottamusvälit (95 % ja 99 %) sisälsivät todellisen aluekeskiarvon 94 prosentissa otoksista, otoskoosta riippumatta.

Testi osoitti, että otanta-asetelma oli informatiivinen sekä alueiden että alueiden sisäisten tilastoyksiköiden poiminnassa. Se piti siis ottaa huomioon. Lopuksi voidaan kirjoittajien näkemyksenä pitää sitä, että mallien käyttö on suoraa estimointia parempi alue-estimoinnissa.

Longford (2007) käsittelee tutkimuksessaan malliperusteisten pienalue-estimaattorien keskivirheitä. Artikkelissa kuvataan simulaatiokokeita, joita tehtiin keinotekoisella aineistolla EURAREA-projektiin liittyen. Populaatio käsitti 100 aluetta. Kokonaisotoskoko oli 3 698 ja pienin alueellinen otoskoko 15. Tarkoituksena oli johtaa aluekeskiarvon empiirisen Bayes-estimaattorin ja asetelmaperusteisen estimaattorin lineaarisen painotetun yhdistelmäestimaattorin keskineliövirheen (MSE) estimaattori standardissa pienalueympäristössä. Estimaattorista saatiin erilaisia variaatioita painoja muuttamalla. Uuden estimaattorin tehokkuutta mitattiin otossimu-

lointien avulla, koska sen ominaisuuksia ei voitu johtaa analyyttisesti. Keinotekoisesta havaintoaineistosta poimittiin 1 000 kertaa 3 698:n suuruinen otos alueittain yksinkertaisella satunnaisotannalla. Aluekeskiarvoille laskettiin ensin suorat estimaatit $\hat{\mu}_d$ ja kehitettiin EB-estimaattorit $\tilde{\mu}_d$ kullekin alueelle *d* sekä jälkimmäisten MSE-estimaattorit eri tavoin johdettuina kuten keskiarvoistettuna, naiivina ja kahtena eri yhdistelmäestimaattorina. Estimoinnissa käytettiin sekä ML- että REML-menetelmää, joiden antamat tulokset erosivat toisistaan hyvin vähän.

Simulointien perusteella havaittiin, että tutkituilla estimaattorityypeillä on vahvat ja heikot ominaisuutensa, mutta yhdistelmäestimaattori oli kokonaisvaltaisesti arvioituna tehokkain. Erityisesti pienten alueiden estimointitulokset paranivat. Kirjoittajalla on harhasta se näkemys, että sen tulisi olla mieluummin positiivinen kuin negatiivinen – siis mieluummin yliestimointia kuin aliestimointia.

Artikkeli ei perustunut tässä tutkimuksessa käytettyyn hierarkkiseen sekamalliin, mutta käsitteli joka tapauksessa malliperusteista alue-estimointia. Malliperusteisessa estimoinnissa usein hyvin tärkeitä alueiden ominaisuuksia ei kuvattu lainkaan. Mitään kaavoja ei johdettu optimaaliselle otoskoolle.

Longford toteaa, että alue-estimointi on usein vain eräänlainen sivutulos, kun pääasiana on koko populaation estimointi, joten kaikkien alueiden optimaalinen estimointi tulee harvoin kysymykseen. Yhden alueen tehokas estimointi on aina ongelmallista, kun sieltä poimittava otos on pieni. Hän viittaa tässäkin artikkelissa alueiden tärkeyskertoimiin, joiden avulla voidaan säädellä alueiden otoskokojen määräytymistä kiintiöinnissä.

Torabi ja Rao (2008) vertasivat artikkelissaan uutta mallitehosteista kaksitasoista GREG-estimaattoria ja EBLUP-estimaattoria pienalue-estimoinnissa. Vaikka artikkelissa ei ole käsitelty otoskiintiöintiä, on se siitä huolimatta tämän tutkimuksen kannalta merkittävä. Tekijät päätyivät empiirisen havaintoaineiston analysointiin liittyvien simulointikokeiden perusteella siihen, että malliperusteinen EBLUP-estimaattori on mallitehosteista regressioestimaattoria (GREG) tehokkaampi, olipa kyse tavanomaisesta tai R. Lehtosen ja A. Veijasen (1999) kehittämästä uudesta GREG-estimaattorista. Vertailussa on käytetty keskineliövirheen (MSE) lisäksi monia laatumittareita, joiden avulla saadaan monipuolisempi kuva estimoinnin tehokkuudesta ja joita on käytetty myös tässä tutkimuksessa. Lehtonen ym. (2003 ja 2005) ovat käsitelleet samaa asiaa.

Rao ja You (2002) käyttivät alue-ennusteiden tuottamiseen pseudo-EBLUP –estimointia, jossa käytettiin erilaisia otantapainoja. Estimointituloksia verrattiin tavanomaiseen EBLUP-estimointiin (ei käytä painoja) ja toiseen, Prasadin ja Raon aikaisemmin kehittämään Pseudo-EBLUP-menetelmään. MSE-vertailussa tavanomainen EBLUP-estimointi tuotti pienimmät MSE-arvot, mutta erot eivät olleet suuria. Artikkelissa korostettiin myös otannan voimakasta keskittämistä tietyille alueille. Myös käytetyt otoskoot olivat kiinnostavia, sillä ne olivat varsin pieniä eli välillä 1–5, kun alueiden koot olivat sentään välillä 402–965 (tilastoyksikköä). Keskimääräinen otoskoko oli siis noin kolme eli huomattavan pieni.

Gelman (2007) tarkastelee artikkelissaan mm. jälkiositukseen liitetyn hierarkkisen sekamallin käytön perustelua korjattaessa otoksen ja perusjoukon välisiä eroja. Estimoitavana voi olla koko populaation keskiarvo tai regressiokertoimet. Hierarkkisessa regressioanalyysissa, jota sovellettiin kerättyihin aineistoihin, johdettiin estimaatti perusjoukon keskiarvolle ilman jälkiositusta ja sitten jälkiosituksen kanssa, joka on välttämätöntä silloin, kun malliin kuuluu vuorovaikutuksia. Estimointiin sisältyi myös painojen estimointi.

Jälkiositusta käytettäessä on kuitenkin vaarana liian monen ositteen syntyminen, jolloin estimaattien tuottaminen niihin kaikkiin voi tulla ylivoimaiseksi. Ositteiden koon vaikutusta ei suoraan arvioida, mutta jos niitä tulee paljon, otoskoot pienenevät tai häviävät kokonaan. Artikkelin hyöty tulee kirjoittajan näkemyksestä, jonka mukaan liian monimutkaiset mallit vaikeuttavat estimointia.

Lehtonen, Myrskylä, Särndal ja Veijanen (2006) ovat verranneet erilaisten GREG- (model-assisted) ja EBLUP-tyyppisten (model-dependent) mallien antamia estimointituloksia tutkimuksessa, jossa käytettiin keinotekoisesti generoitua perusjoukkoa. Vastemuuttujan lisäksi oli käytössä kaksi apumuuttujaa, joita käytettiin joko yksitellen tai yhdessä. GREG-malleissa käytettiin joko pelkästään perusjoukon tason vaikutuksia tai lisäksi myös aluevaikutuksia. EBLUP-malleissa oli aluevaikutus mukana koko ajan. Jälkimmäisten mallien osalta oli tämän tutkimuksen kannalta merkittävä tulos se, että joidenkin mallien valinnan seurauksena oli voimakkaasti harhainen estimointitulos, jota ei parantanut edes otoskoon lisääminen alueille.

Nissinen (2009) on käyttänyt lineaarisia sekamalleja pienalue-estimoinnissa pitkittäisaineistojen käsittelyyn (samojen tilastoyksiköiden toistuvat mittaukset ja rotatointi). Hän tulee otossimulointien perusteella johtopäätökseen, että estimoitaessa pienalueiden kokonaismääriä rotatoinnin ansiosta saadaan uudenlaista havaintoaineistoa, mikä parantaa piste-estimaattien tarkkuutta sekä

pienentää niiden harhaa ja keskineliövirhettä (MSE). Rotatoinnin antama lisähyöty on vielä se, että se eliminoi väärän mallivalinnan aiheuttaman mahdollisen harhan.

Väitöskirjassa on esitelty hyvin perusteellisesti lineaarisiin sekamalleihin perustuvan pienalue-estimoinnin teoreettinen perusta ja erityisesti EBLUP-estimointi, mutta alueiden otoskiintiöinti ei ole esillä. Joka tapauksessa tässäkin on osoitus epäsuoran pienalue-estimoinnin käyttökelpoisuudesta.

Tässä alaluvussa esitellyissä tutkimuksissa ei otettu suoranaisesti kantaa siihen, millainen aluekiintiöinti olisi estimointulosten kannalta optimaalinen. Käytettyjä otoskokoja oli kyllä mainittu sekä analysoitu otoskoon lisäämisen vaikutuksia. Monista tutkimuksista nousee kuitenkin esille aluekiintiöinnin kannalta tärkeä kysymys, miten alue- ja perusjoukon tasoa pitäisi painottaa, jos estimoidaan kummankin tason tunnuslukuja.

4 KIINTIÖINTI ALUE-ESTIMOINTIA VARTEN

4.1. Alueet ositteina

Aluekiintiöinti voi olla suunnittelematonta tai suunniteltua. Suunniteltu kiintiöinti tarkoittaa sitä, että alueet ovat ositetun otannan ositteita, joista otos poimitaan. Alueiden otoskoot voidaan määrätä enemmän teknisesti tai ne voidaan johtaa jonkin tehokkuutta kuvaavan tunnusluvun optimointituloksista. Kiintiöinti noudattaa joka tapauksessa jotain ennalta valittua selkeää periaatetta. Otanta voi tilanteesta riippuen keskittyä vahvasti muutamille alueille, kun taas joiltakin alueilta voi tulla vain yksi havainto tai ei ollenkaan.

Onnistunut kiintiöinti on sidoksissa ainakin vastemuuttujan *y* ja estimointimenetelmästä riippuen joskus myös apumuuttujien $\mathbf{x} = (x_1, x_2, ..., x_p)'$ eri ominaisuuksien omaan ja yhteisvaihteluun sekä alueiden tilastoyksiköiden määriin. Lisäksi on tärkeää pystyä jakamaan kokonaisvaihtelu alueiden väliseen ja sisäiseen vaihteluun. Jos näistä tekijöistä on riittävästi tietoja käytettävissä, on mahdollista löytää sellainen otoskiintiöinti, jonka ansiosta päästään tehokkaaseen estimointiin ja jonka kriteerinä ovat aluetasolla lasketut laatumittarit tai niiden keskiarvot. Muita kriteerejä voivat olla esimerkiksi joidenkin alueiden suurempi tärkeys estimoinnin onnistumisessa muiden alueiden kustannuksella tai mahdollisimman tarkkojen ennusteiden (pieni ennustevirhe) tuottaminen kaikille alueille. Mutta valitun mallin mukaan tuominen jo kiintiöintivaiheeseen antaa mahdollisuuden käyttää laajemmin alueisiin liittyvää informaatiota ja myös alueiden välisiä yhteyksiä.

4.2. Sovellettuja kiintiöintiratkaisuja

Seuraavissa alaluvuissa esitellään joukko aiemmin kehiteltyä aluekiintiöintiratkaisuja. Olennaista niille on, miten ne käyttävät perusjoukosta olevaa tietoa ja miten niiden optimointikriteerit on asetettu. Ensimmäiseksi on otettu tilastoyksikköjen lukumääriin perustuvat tasakiintiöinti ja suhteellinen kiintiöinti, jotka on esitetty alaluvussa 4.2.1. Näistä kehittyneempään kiintiöintiryhmään kuuluvat sellaiset ratkaisut, joissa tarvitaan parametritietoja perusjoukosta. Alaluvussa 4.2.2. on kaksi, joista toisessa eli Neyman – kiintiöinnissä optimointikriteeri on vain perusjoukkotaso ja toisessa eli potenssikiintiöinnissä optimointikriteeri on yksistään aluetaso. Optimointikriteerit voidaan asettaa myös niin, että ne ovat voimassa sekä perusjoukon että aluetason estimoinnille. Näistä on kaksi esimerkkiä: alueiden tärkeyskertoimiin perustuva kiintiöinti on alalu-

vussa 4.2.3 ja vaihtelukerroinrajoitteista johdettuun minimiotoskokoon perustuva kiintiöinti alaluvussa 4.2.4. Muita aluekiintiöintiratkaisuja on koottu alalukuun 4.2.5.

4.2.1 Alueiden ja tilastoyksiköiden lukumääriin perustuvat kiintiöinnit

Perusjoukossa olevien tilastoyksikköjen lukumäärällä tarkoitetaan alueiden lukumäärää *D* ja aluetason kokoja N_d. Otoskoko *n* oletetaan annetuksi. Tällöin yksinkertaisin kiintiöintimenetelmä on tasakiintiöinti, jossa jokaisen alueen otoskoko on sama

$$n_{d,tas} = n/D. \tag{4.1}$$

Kyseessä on eräänlainen mekaaninen kiintiöinti, jonka soveltaminen johtaa harvoin hyviin estimointituloksiin, koska alueiden koot ja sisäiset vaihtelut ovat todennäköisesti aina erilaisia. Erityisesti suuret alueet, joissa on suuri vaihtelu, kärsivät estimoinnin tehokkuuden ja tarkkuuden osalta. Aluetason otosvarianssien estimoimiseksi kokonaisotoskoko pitää olla vähintään $n \geq 2D$.

Sitä kannattaa käyttää silloin, kun suuremmilla alueilla (ositteissa) on odotettavissa suurempi varianssi kuin pienillä alueilla. Kiintiöintiä voidaan pitää eräänlaisena odotettavissa olevana kiintiöintinä, jos käytetään puhdasta SRS-otantaa. Alueen *d* otoskoko n_d on suoraan verrannollinen alueen suhteelliseen osuuteen koko perusjoukossa ja on SRSWOR-otannassa alueen *d* odotettavissa oleva otoskoko:

$$n_{d,suh} = f_d\, n = (N_h / N) n. \tag{4.2}$$

Suhteellinen kiintiöinti takaa alueille niiden suhteellista kokoa vastaavan osuuden otoksessa, mutta ei välttämättä johda tehokkaaseen alue-estimointiin. Tämä koskee erityisesti alueilla, joissa vastemuuttujalla on suuri varianssi. Samoin kooltaan pienten alueiden kohdalta ehto $n_{d,suh} > 1$ ei välttämättä toteudu, jolloin otosvarianssia ei pystytä lainkaan estimoimaan.

Kummassakaan kiintiöintimenetelmässä ei ole kriteeriä, minkä suhteen aluetason otoskoko valitaan. Näiden kahden kiintiöinnin välimuodon ovat esittäneet Costa ym. (2004). Lauseke on

$$n_{d,ta\_su} = k\, n(N_d / N) + (1-k) n/D, \tag{4.3}$$

jossa kerrointa *k* ($0 \leq k \leq 1$) voidaan vaihdella. Sen ääriarvolla (*k = 0*) tuloksena on tasakiintiöinti ja toisella (*k = 1*) suhteellinen kiintiöinti. Kertoimen valinnasta ei ole yksikäsitteistä ohjetta.

4.2.2 Alueparametreihin nojautuvat kiintiöinnit

Alueparametreilla tarkoitetaan vaste- tai apumuuttujan aluetunnuslukuja kuten keskiarvoja ($\overline{Y}_d$ tai $\overline{X}_d$), keskihajontoja ($S_d\,(y)$ tai $S_d\,(x)$) tai vaihtelukertoimia ($CV_d\,(y)$ tai $CV_d\,(x)$). Nämä alueparametrit ovat käytössä niissä aluekiintiöintiratkaisuissa, joista esittelyt ovat seuraavassa. Käyttö aluekiintiöintiin ehdollistuu tietenkin siihen, miten kyseiset tiedot ovat saatavilla. Tietolähteinä voivat olla koko perusjoukon kattavat rekisterit, joista ainakin apumuuttujan *x* tiedot on saatavissa. Vastemuuttujan *y* tietoja ei tietenkään ole, koska ne ovat niitä tuntemattomia suureita, joita aiotaan estimoida poimittavasta otoksesta. Yksi mahdollisuus on etsiä vastemuuttujalle *y* sijaismuuttuja *y**, jonka arvellaan korreloivan vastemuuttujan kanssa. Tällainen voi olla toistotutkimuksissa tekeillä olevaa edeltävältä ajankohdalta oleva tutkimustulos. Eräs vaihtoehto on poimia pieni esiotos ja estimoida siitä aluekiintiöintiin tarvittavat vastemuuttujan parametrit. Tällöin astuvat kuvaan myös luvussa 3 esitetyt aluemallitukset ja niihin kytkeytyvät estimointitekniikat. Alueparametreihin ehdollistuvissa kiintiöintimenetelmissä on optimointikriteerit.

Optimaalinen kiintiöinti perustuu alueiden kokoon, apumuuttujan *x* aluekohtaisiin keskihajontoihin ja tilastoyksiköiden mittauskustannuksiin aluetasolla. Särndal ym. (1992) ovat kuvanneet kiintiöintimenetelmän perusteellisesti. Apumuuttujan kaikki tilastoyksikkökohtaiset arvot tai ainakin aluetason keskihajonnat on tunnettava koko perusjoukossa. Kiintiöinnin erikoistapaus on Neyman-kiintiöinti, jossa oletetaan mittauskustannukset yhtä suuriksi kaikilla alueilla. Tällöin aluekohtaisen otoskoon lauseke on

$$n_{d,opt} = (N_d S_d / \sum_{d=1}^{D} N_d S_d)n, \tag{4.4}$$

missä S_d on apumuuttujan keskihajonta ja N_d tilastoyksiköiden lukumäärä alueella *d*.

Otoskoko tulee suureksi alueelle, jossa on paljon tilastoyksiköitä tai jossa apumuuttujan keskihajonta on muita alueita suurempi. Erot otoskokojen välillä voivat olla huomattavat, kun niitä verrataan esimerkiksi suhteellisen kiintiöinnin vastaaviin. Aluetasolla tämä kiintiöinti ei kuulu parhaimpiin, koska kiintiöinnin perustana on optimointi perusjoukon tasolla.

Aluetaso-optimaalisista kiintiöintiratkaisuista eräs on potenssikiintiöinti, jonka esitti Bankier (1988). Potenssikiintiöinti perustuu alueiden sisäisiin ominaisuuksiin, joista mukana ovat apumuuttujan vaihtelukerroin ja kokonaismäärä. Alueiden otoskoot lasketaan kaavalla

$$n_d = (X_d^a CV_d(x) / \sum_{d=1}^{D} X_d^a CV_d(x))n, \tag{4.5}$$

missä X_d tarkoittaa apumuuttujan *x* arvojen kokonaismäärää ja $CV_d(x)$ *x*:n vaihtelukerrointa alueella *d*. Eksponentti *a* on eräänlainen voimakkuusluku, jolla voidaan säädellä apumuuttujan merkitystä. Normaalisti *a*:lle käytetään arvoa ½ tai ⅓. Jos olisi kyseessä sellainen erikoistapaus, että alueiden CV-arvot olisivat samat, riippuisivat otoskoot ainoastaan apumuuttujan *x* kokonaismääristä X_d. Sitä suositellaan käytettäväksi kyselytutkimuksessa, johon liittyvässä perusjoukossa on monta pientä aluetta ja joille myös on saatava luotettavat estimaatit. Tämäkin kiintiöinti perustuu siihen, että apumuuttujan *x* tilastoyksikkökohtaiset arvot tunnetaan koko perusjoukossa.

Edellä esitetyt kaksi aluekiintiöintiä poikkeavat optimointikriteereiltä toisistaan. Neyman-kiintiöinti ottaa huomioon vain perusjoukkotason optimoinnin. Potenssikiintiöinnissä alueiden tärkeys ilmaistaan kokonaismäärien potensseista johdettujen painokertoimien muodossa. Aluekiintiöinti on mahdollista johtaa myös siten, että kiintiöintiratkaisussa otetaan samanaikaisesti huomioon sekä aluetason että perusjoukkotason estimaattorien keskivirheet. Näistä seuraavassa on kaksi alalukua, joista toisessa (4.2.3) mukana ovat alueille annetut tärkeyskertoimet. Toisessa (4.2.4) ratkaistaan kokonaisotoskoko ja otoskoot alueille, kun ennakkoon on annettu yläraja sekä perusjoukkotason estimaatin että aluetason estimaattien keskihajontojen ylärajoille.

4.2.3 Tärkeyskertoimien käyttö aluekiintiöinnissä

Longfordin tutkimus (2006) voidaan lukea ensimmäisten merkittävien tutkimusten joukkoon optimaalisten otoskokojen analyyttisessa johtamisessa. Nyt herää kysymys, voisiko optimaalisten otoskokojen analyyttinen laskeminen eri alueille olla mahdollista mallipohjaisessa estimoinnissa samanlaisen periaatteen mukaan kuin Longfordin artikkelissa on esitelty suoran estimoinnin tapauksessa. Ongelmana voisi olla esim. MSE:n aluekeskiarvon minimointi otoskokojen funktiona. Eri alueille määrätään painot (tärkeyskertoimet), joiden avulla muodostetaan optimoitava lauseke ja lasketaan alueiden otoskoot. Ongelmaksi muodostuu laskelmien monimutkaisuus, mutta simuloinnin avulla voidaan tutkia otoskokojen $\boldsymbol{n} = (n_1, n_2, \ldots, n_D)'$ vaikutusta alueiden estimointituloksiin.

Artikkelissa esitetty otoskokojen johtaminen on varteenotettava lähtökohta mallipohjaiseen estimointiin liittyvän tehokkaan otanta-asetelman tutkimiseen, mutta Longfordin työ perustuu suoraan estimointiin, mikä rajoittaa sen käyttökelpoisuutta. Joka tapauksessa sen ansioksi luettava, että analyyttinen ratkaisu on johdettu, vaikkakin varsin yksinkertaisten oletusten vallitessa.

Longford esittelee menetelmän, jonka avulla voidaan laskea eri alueille kohdistettava otoskoko estimoitaessa alueiden keskiarvoja tai jonkin ominaisuuden suhteellisia esiintyvyyksiä, kun estimaattori on suora, yhdistetty tai empiirinen Bayes-estimaattori. Tämä menetelmä edellyttää tietynlaisten aluepainotusten määräämistä eri alueille. Tehokkuuskriteerinä on otanta-asetelman ja perusjoukon ominaisuusparametrin θ yhdistelmän optimaalinen kombinaatio. Optimilla tarkoitetaan mahdollisimman pientä keskineliövirhettä (MSE), vaikka esiteltävä optimointimenetelmä voidaan liittää muuhunkin kriteeriin kuin MSE:n optimointiin. Käytettävissä olevat resurssit, joihin liittyy tavallisesti perusjoukosta poimittu kiinteänkokoinen otos, rajoittavat aina otanta-asetelmien käyttöä. Otantakustannuksia pidetään kuitenkin vakiona tilastoyksikköä kohti. Esimerkkitapauksena on käytetty Sveitsin Kantoneihin (maksimi 1,23 miljoonaa ja minimi 15 000 asukasta) liittyvää alue-estimointia.

Periaatteessa kysymys on perusjoukon tiettyä ominaisuutta kuvaavan määrällisen suureen tehokkaasta estimoinnista sovitettuna otanta-asetelmaan. Ongelma on siis optimointitehtävä tiettyjen rajoitusten vallitessa (vrt. lineaarinen optimointi). Monen tekijän (parametrin) samanaikainen optimointi on huomattavasti monimutkaisempaa, ja konfliktitilanteita varmasti tulee; usein joudutaan kasvattamaan tietylle alueelle suunnattavaa otoskokoa yhden tai useamman muun alueen kustannuksella.

Yksi vaikeus otannan kohdistamisessa on sovittaa yhteen koko perusjoukon (esim. valtion) ja alueiden (esim. maakunnat, kunnat) tehokas estimointi saman otanta-asetelman puitteissa. Longford osoittaa, että näiden kahden optimoinnin yhteensovittaminen on mahdollista tiettyjen yksinkertaistettujen oletusten ollessa voimassa.

Longford viittaa myös Singhin ym. tutkimukseen (1994). Yhtä suuret alueotoskoot antavat luotettavat tulokset silloin, kun alueiden sisäiset varianssit ovat yhtä suuret, äärellisyyskorjaus voidaan jättää huomiotta, otantakustannukset ovat samat tilastoyksikköä kohti ja kun kyseessä on aluekeskiarvojen estimointi. Estimoitaessa alueiden kokonaismääriä tasakiintiöinti ei anna hyviä tuloksia, koska suurten alueiden estimointi kärsii. Sama koskee myös suhteellisten osuuksien ja suhteiden (esim. työttömyys- %) estimointia alueilla, vaikka haitta ei ole niin suuri silloin, kun suhteellinen osuus tai suhdeluku ei ole lähellä nollaa eikä ykköstä.

Longfordin perusideana on ollut minimoida alueiden otosvarianssien painotettu keskiarvo ensin aluetasolla ja sitten koko perusjoukon tasolla. Tavoitteena on ollut kehittää optimointi, joka ottaa kokonaisuuden huomioon mahdollisimman hyvin.

Ensin määritellään seuraavat merkintätavat: aluetasoisten (*D* kpl) määrällisten ominaisuuksien θ_d estimaattorit ovat $\hat{\theta}_d$ (*d = 1,2,..., D*), ja niiden estimoidut keskineliövirheet ovat $\hat{v}_d$ (alue-estimaattoreiden MSE:t), jotka puolestaan ovat alueiden otoskokojen n_d funktioita: $\hat{v}_d = \hat{v}_d(n_d)$. Tilastoyksiköiden aluekohtaiset määrät ovat N_d (*d* = 1,2,…, *D*), perusjoukon kokonaismäärä $N = \sum_{d=1}^{D} N_d$ ja otoksen koko on $n = \sum_{d=1}^{D} n_d$. Estimoitavat perusjoukon määrälliset ominaisuudet θ ovat yksittäisen muuttujan funktioita (keskiarvo, kokonaismäärä yms.). Muuttuja voi olla myös monen muuttujan yhdistelmä. Alue-estimaattori $\hat{\theta}_d$ on suora estimaattori, jos se on vain alueen *d* vastemuuttuja-arvojen funktio. Suora estimaattori oletetaan lisäksi harhattomaksi.

Aluetasoisten määrällisten suureiden θ_d tehokas estimointi pohjautuu otanta-asetelmaan, joka minimoi otosvarianssien (MSE-arvojen) painotetun summan

$$min_n \sum_{d=1}^{D} P_d \hat{v}_d \ , \tag{4.6}$$

jonka rajoitteena on kokonaisotoskoko $n = \sum_{d=1}^{D} n_d$. Kertoimilla P_d voidaan säädellä alueille tulevia otoskokoja. P_d:n suurempi arvo verrattuna toiseen kertoimeen $P_{d'}$ ilmaisee alueen *d* varianssin v_d suurempaa pienentämistarvetta, koska tällöin alueen *d* vaikutus summan (4.6) pienentämiseen on merkittävämpi. Longford ei kuitenkaan määrittele P_d-kertoimille sellaisia arvoja, joiden summa olisi vakio.

Optimointiongelma (3.1) voidaan ratkaista Lagrangen kerroinmenetelmällä tai kirjoittamalla $n_1 = n - n_2 - ... - n_D$, jolloin ongelma sisältää *D - 1* funktionaalisesti riippumatonta muuttujaa. Ratkaisu täyttää ehdon

$$P_d(\partial v_d / \partial n_d) = vakio. \tag{4.7}$$

Aluekohtaisten otoskokojen lausekkeita ei voida ratkaista yleisessä tapauksessa analyyttisesti, mutta kun $v_d = \sigma_d^2 / n_d$, kuten SRS-otannassa alueiden sisällä, ratkaisu on suoraan verrannollinen lausekkeen $\sigma_d \sqrt{P_d}$ kanssa, toisin sanoen

$$n_{d opt} = n \frac{\sigma_d \sqrt{P_d}}{\sigma_1 \sqrt{P_1} + ... + \sigma_D \sqrt{P_D}} \ . \tag{4.8}$$

Kun alueiden sisäiset varianssit σ_d^2 ovat yhtä suuret (σ^2), edellinen lauseke yksinkertaistuu: alueiden optimaaliset otoskoot ovat suoraan verrannollisia termin $\sqrt{P_d}$ kanssa eivätkä riipu varianssista σ^2.

Parhaiten sopivien kertoimien P_d löytäminen on vaikeaa, joten on järkevämpää määritellä ko. prioriteeteille käyttökelpoinen tyyppiluokitus ja kuvata niiden vaikutusta otoskokojen määräytymiseen. Longford esittää tyyppiä $P_d = N_d^q$ **,** missä $0 \leq q \leq 2$. Jos $q = 0$, tulee kaikille alueille yhtä suuri otoskoko, mikä tarkoittaa yksinkertaisesti tasaista kiintiöintiä, ja kun q kasvaa, otanta suosii isompia alueita. Kun $v_d = \sigma_d^2 / n_d$, optimaalinen otoskoko q:n arvolla 2 on $n \times (N_d / N)$, eli kyseessä on suhteellinen kiintiöinti. Jos q:lle annetaan suurempi arvo kuin 2, suosii kiintiöinti entistä enemmän suuria alueita, jolloin lähestytään optimaalista Neyman-kiintiöintiä. Negatiivinen q:n arvo merkitsisi pienimpien alueiden suosimista, eli kiintiöinti muistuttaisi jossain määrin alueoptimaalista potenssikiintiöintiä, mutta tämä hankaloittaisi perusjoukon tason estimointia, varsinkin silloin, kun alueiden kokoerot ovat suuret.

Aluepainotukset P_d voidaan määrätä muullakin tavalla kuin sitomalla ne alueiden kokoon N_d. Muitakin lukumääriä voidaan käyttää pohjana, esimerkiksi tiettyjen pienalueiden, kuten eri etnisten ryhmien, kokoja. P_d voidaan määritellä eri tavalla eri alueille tai kaava voidaan jättää käyttämättä tietyille alueille.

Joissakin julkaistuissa kyselytutkimuksissa ilmoitetaan estimaatit vain niissä tapauksissa, että alueen otoskoko on riittävän suuri tai estimaatin vaihtelukerroin (CV) alittaa ennalta asetetun rajan (vrt. potenssikiintiöinti). Jos tällaisia rajoja asetetaan, ne voitaisiin liittää päätösprioriteettien määrittelyyn. Tällöin saattaa tulla eteen sellainen vaikeus, että optimoitava funktio (1) on epäjatkuva, jolloin standardiratkaisut eivät ole sovellettavissa. Estimaattien ilmoittamisen rajat on määriteltävä huolellisen harkinnan perusteella. Jos ne ovat liian matalia, ne ovat tehottomia, ja liian korkeat rajat voivat aiheuttaa turhan monelle alueelle osin huonot estimaatit.

Longford esittelee aluepainotusluokan $P_d = N_d^q$ perusteella Sveitsin 26 Kantonin otoskokojen määräytymistä, kun kokonaisotoskoko on n = 10 000 ja kun q:n arvo vaihtelee välillä 0 – 2. Yksittäisen alueen otoskoko voi vaihdella suurestikin riippuen q:n arvosta. Jos $q = 0$, tulee eri Kantoneista yhtä monta havaintoa otokseen (10000/26 = 385), ja jos $q = 1$, on Kantonin otoskoko

suoraan verrannollinen Kantonin asukasluvun neliöjuureen. Tapauksessa $q = 2$ on otoskoko suoraan verrannollinen Kantonin asukaslukuun. 1,23 miljoonan asukkaan Kantonin otoskoko voi vaihdella välillä 385–1694 ja 500 000 asukkaan Kantonin otoskoko välillä 385–681. Huomionarvoinen seikka on vielä, että kun on kyseessä suhteellisen pieni Kantoni (n. 250 000 asukasta), vaihtelee Kantonille tuleva otoskoko varsin vähän (n_d = 344–385).

Longford on käsitellyt esityksessään myös kahta muuta aluekiintiöintiä. Ensimmäisessä kiintiöinnissä optimoitava funktio ottaa huomioon sekä aluetason että perusjoukon tarpeet, ja toisessa kiintiöinnissä on kysymys asetelma- ja malliperusteisen estimaattorin yhdistelmästä, johon liittyvä optimoitavaan funktioon sisältyy sekä aluetason että perusjoukon tason estimointikriteerit. Näitä tapauksia ei esitellä tässä tarkemmin.

Longford kuvaa optimaalisen otoskoon määräytymistä keinotekoisessa tilanteessa, jossa oli kyseessä ositettu SRS-otanta, ja alueiden (ositteiden) varianssit olivat samat. Menetelmän kriittinen kohta on tärkeyspainotusten asettaminen alueille sekä koko populaatiolle. Niitä ei välttämättä ole helppo päättää. Lisäksi eräät muut oletukset, kuten alueiden sisäiset yhtä suuret varianssit, eivät pidä paikkaansa todellisuudessa. Menetelmä voidaan ulottaa myös monimutkaisempiin tilanteisiin, mutta siinä tapauksessa vaaditaan lisäparametreja. Todellisuus asettaa omat haasteensa optimaaliselle otannalle; esimerkiksi kadon epätasainen jakautuminen alueilla voi muodostua ongelmaksi. Joka tapauksessa tällaista lähestymistapaa voidaan käyttää periaatteessa kaikkiin pienalue-estimaattoreihin, joiden MSE voidaan määrätä analyyttisesti, joko tarkkana tai approksimaatioarvona. Tällöin voidaan tehdä laskelmia erilaisilla lähtöarvoilla ja –oletuksilla.

Herkkyysanalyysi on olennaista tutkittaessa otanta-asetelman muuttumista siihen vaikuttavien estimoitavien parametrien ja niihin liittyvän epävarmuuden vaihdellessa. Tutkimuskohdetta ei myöskään tule tehdä turhan monimutkaiseksi, jolloin sen hallinta ja tutkiminen vaikeutuvat.

Oikeaa otoskokoa kannattaa aina yrittää etsiä laskelmien avulla. Sen merkitystä voidaan kuvata Longfordin mukaan niin, että yhden tilastoyksikön pienennys otoskoossa johtaa suurempaan tarkkuuden menetykseen kuin yhden tilastoyksikön lisäys tarkkuuden paranemiseen, mikä johtuu estimaattorien varianssien tai MSE:n lausekkeista. Otanta-asetelmat, joissa alueotoskokojen n_d (d on kiinnitetty) vaihtelu on pienempää, sopivat paremmin pienalue-estimointiin. Pienemmät ositteet tai vastaavat, joista otokset poimitaan, ovat parempia kuin suuret, jos vain resurssit antavat myöten.

Longfordin idea otosvarianssien summan minimoinnista alueiden otoskokojen funktiona on ollut lähtökohtana myös tässä tutkimuksessa, jossa etsitään ratkaisua alueiden MSE-keskiarvojen minimointiin otoskokojen funktiona. Valitettavasti analyyttinen ratkaisu ei onnistu monimutkaisen matemaattisen taustan vuoksi, mutta approksimatiivinen ratkaisu on mahdollista johtaa, kuten myöhemmin esitetään. Alueiden välisen vaihtelun voimakkuus on päätösprioriteettien tilalla.

4.2.4 Vaihtelukerroinrajoitteista johdettuun minimiotoskokoon perustuva kiintiöinti

Choudry ym. (2012) ovat kehittäneet suoraan estimointiin epälineaarista ohjelmointia (NLP) käyttävän menetelmän, jossa ensin asetetaan vastemuuttujan *y* estimoitavien parametrien (ositteiden keskiarvot sekä koko perusjoukon keskiarvo) asetelmaperusteisille vaihtelukertoimille (CV) ylärajat (tutkimuksessa 15 % ositteille ja 6 % perusjoukolle) ja sen jälkeen etsitään ohjelman avulla ositteille sellainen otoskiintiöinti, että sen avulla voidaan määrätä minimiotoskoko *n* ($n = n_1 + ... + n_D$), jonka mukaan edellä mainitut CV-rajoitteet ovat vielä voimassa. Minimoitavana on lauseke

$$g(\mathbf{f}) = \sum_{d=1}^{D} f_h N_h \ ,$$

missä f_d = alueen *d* otantasuhde ($f_h = n_d / N_d$ ja $0 < f_d \leq 1$), $\mathbf{f} = (f_1, ..., f_D)$ ja N_d = alueen *d* tilastoyksiköiden lukumäärä. Rajoitteet saadaan epäyhtälöistä, joissa asetelmaperusteisille CV-lausekkeille on määritelty ylärajat.

Tutkimusaineistona käytettiin Kanadan tilastoviraston keräämää MRTS-aineistoa. Siten kysymyksessä ei ole uuteen otantaan perustuva tutkimus, vaan siinä oli käytetty vastemuuttujan *y* tilalla sijaismuuttujaa. Aineisto oli kerätty Kanadan 10 provinssista, jotka olivat samalla ositteita. Sijaismuuttujan keskiarvo, hajonta ja vaihtelukerroin tunnettiin jokaisesta provinssista. Tutkimus ei perustunut otossimulointeihin, vaan erilaisiin ositekohtaisiin sijaismuuttujan tunnuslukuarvoihin perustuviin laskelmiin. Kaikki CV-rajoitteet saavutettiin vasta, kun kokonaisotoskoko *n* oli varsin suuri eli 3 446, ja ositteiden otoskoot vaihtelivat välillä 104–1 056. Alueiden otoskokojen perusteella ei ole kysymys pienalue-estimoinnista.

Tutkimuksessa kehitetyn menetelmän tehokkuutta verrattiin eräisiin aikaisemmin kehitettyihin kiintiöintimenetelmiin, mm. Longfordin (2006) kehittämään, kun otoskoko (*n*) oli niissäkin 3 446. Vertailu on tulosten perusteella hieman hankalaa. Missään aiemmassa menetelmässä eivät kaikkien ositteiden CV-arvot jääneet korkeintaan 15 %:iin, mutta toisaalta löytyi matalia (alle 10 %), joskin varsin korkeitakin ositteiden CV-arvoja (jopa 85 %). Koko perusjoukon CV-arvoissa

ei ollut menetelmien välillä kovin suuria eroja (5–9 %). Kaiken kaikkiaan esitelty menetelmä lisää kiintiöintivalikoimaa, mutta ongelmana on, miten sitä sovelletaan tilanteeseen, jossa on kerättävä tuntemattomasta vastemuuttujasta oma otanta-aineisto. Vastemuuttujan aluekohtaisista (osite-) keskiarvoista ja hajonnoista on oltava jonkinlainen ennakkokäsitys, jotta laskelmia voidaan tehdä.

Edellä kuvatun kiintiöinnin periaatteita on käytetty johdettaessa tähän tutkimukseen uutta testattavaa kiintiöintiä. Poikkeus on se, että ositekohtaisten keskiarvojen sekä perusjoukon keskiarvon sijasta on käytetty kokonaismääriä. Vastemuuttujan tilalla on pienen esiotoksen pohjalta muodostettu sijaismuuttuja y^*, jonka johtaminen on kuvattu tarkemmin alaluvussa 5.2. Kiintiöintiä kutsutaan tässä NLP-kiintiöinniksi.

Sijaismuuttujan y^* arvot tunnetaan täydellisesti, joten sille voidaan laskea aluekohtaiset kokonaismäärät Y_d^*, keskiarvot $\bar{Y}_d^*$ sekä varianssit $S_d^2(y^*)$ tai hajonnat $S_d(y^*)$ (d = 1,2,…, D). Vastaavat arvot perusjoukolle ovat Y^*, $\bar{Y}^*$, $S^2(y^*)$ ja $S(y^*)$.

Sijaismuuttujan y^* vaihtelukerroin alueella d on $C_d(y^*) = S_d(y^*)/\bar{Y}_d^*$ sekä perusjoukossa $C(y^*) = S(y^*)/\bar{Y}^*$. Arvot on esitetty liitetaulukoissa B.1 ja B.2.

Oletetaan, että perusjoukon sijaismuuttuja-arvoista poimitaan SRS-otos, jonka koko on n. Alueiden otoskoot ovat n_d, ja $n = \sum_{d=1}^{D} n_d$. Edelleen määritellään aluekohtaiset otantasuhdeluvut $f_d = n_d / N_d$, joiden käänteisluvut ovat $k_d = f_d^{-1} = N_d / n_d$. Sijaismuuttujan kokonaismäärän aluekohtainen estimaattori on $\hat{Y}_d^* = N_d \bar{y}_d^*$, missä $\bar{y}_d^*$ on sijaismuuttujan otoskeskiarvo. Koko perusjoukon kokonaismäärän estimaattori on $\hat{Y}^* = \sum_{d=1}^{D} \hat{Y}_d^*$.

Alueen d kokonaismäärän estimaattorin varianssi on

$$V(\hat{Y}_d^*) = N_d^2 V(\bar{y}_d^*) = N_d^2 (1/n_d - 1/N_d) S_d^2(y^*) = N_d (k_d - 1) S_d^2(y^*) \tag{4.9}$$

ja suhteellinen varianssi (vaihtelukertoimen neliö) on varianssin ja kokonaismäärän neliön suhde

$$\begin{aligned} RV(\hat{Y}_d^*) &= V(\hat{Y}_d^*)/(Y_d^*)^2 = V(\hat{Y}_d^*)/N_d^2(\bar{Y}_d^*)^2 = [(k_d - 1)/N_d]\,[S_d^2(y^*)/(\bar{Y}_d^*)^2] \\ &= [(k_d - 1)/N_d]\, C_d^2(y^*). \end{aligned} \tag{4.10}$$

Perusjoukon kokonaismäärän estimaattorin varianssi on vastaavien alue-estimaattoreiden varianssien summa $V(\hat{Y}^*) = \sum_{d=1}^{D} V(\hat{Y}_d^*)$, ja suhteellinen varianssi on

$$RV(\hat{Y}^*) = V(\hat{Y}^*)/(Y^*)^2 = \sum_{d=1}^{D} V(\hat{Y}_d^*)/(Y^*)^2 = \sum_{d=1}^{D} N_d(k_d - 1)S_d^2(y^*)/(Y^*)^2 . \quad (4.11)$$

Tavoitteena on löytää pienin kokonaisotoskoko $n = \sum_{d=1}^{D} n_d$, jonka voimassa ollessa alueiden kokonaismäärien sekä perusjoukon kokonaismäärän estimaattorien CV-arvot eivät ylitä ennalta asetettuja rajoja. Samalla voidaan ratkaista myös aluekohtaiset otoskoot n_d. Ongelma muotoillaan seuraavaksi matemaattisesti. Ensin määritellään otantasuhdelukujen vektori $\boldsymbol{f} = (f_1,..., f_D)^T$. Minimoitavana on muuttujien f_d funktio

$$g(\boldsymbol{f}) = \sum_{d=1}^{D} f_d N_d , \quad (4.12)$$

kun rajoitteet ovat seuraavia:

$$CV(\hat{Y}_d^*) \le CV_{0d} , d = 1,\ldots, D \quad (4.13)$$

$$CV(\hat{Y}^*) \le CV_0 \quad (4.14)$$

$$0 < f_d \le 1 , d = 1,\ldots, D. \quad (4.15)$$

Arvot CV_{0d} ja CV_0 ovat ennalta asetettuja rajoja alueen d kokonaissumman sekä perusjoukon kokonaissumman CV-arvolle. Erisuuruusmerkkejä käytetään sen vuoksi, että CV-arvon yläraja voi alittua joillakin alueilla tai perusjoukossa.

Jos otantasuhdelukujen f_d tilalla käytetään niiden käänteislukuja k_d, muuttuu minimoitava funktio (5.13) muuttujien k_d konveksiksi funktioksi

$$\tilde{g}(\boldsymbol{k}) = \sum_{d=1}^{D} N_d k_d^{-1} . \quad (4.16)$$

Rajoitteet (4.13) ja (4.14) määritellään uudelleen suhteellisten varianssien (4.10) ja (4.11) avulla, niin että rajoitteet ovat lineaarisia muuttujien k_d suhteen:

$$RV(\hat{Y}_d^*) \le RV_{0d} , d = 1,\ldots, D \quad (4.17)$$

$$RV(\hat{Y}^*) \le RV_0 \quad (4.18)$$

$$k_d \ge 1, d = 1,\ldots, D. \quad (4.19)$$

Suhteellisten varianssien rajoitteet ovat seuraavia: $RV_{0d} = CV_{0d}^2$ ja $RV_0 = CV_0^2$. Rajoitteiden lineaarisuus muuttujien k_d suhteen sekä optimoitavan funktion konveksisuus johtavat nopeammin saavutettavaan ratkaisuun. Rajoite-epäyhtälöä (4.19) voidaan muokata sellaiseksi, että jokaisen alueen otoskooksi tulee vähintään kaksi, mikä mahdollistaa varianssien harhattoman estimoinnin. Kokonaisotoskoon *n* minimiksi ja sitä vastaavien alueiden otoskokojen arvoiksi tulee todennäköisesti desimaalilukuja, joten ne on lopuksi pyöristettävä kokonaisluvuiksi.

Optimoinnissa asetetaan ensin vektorille $\boldsymbol{k^0} = (k_1^0,..., k_D^0)^T$ sopivat alkuarvot (esim. 2), minkä jälkeen haetaan minimi funktiolle (4.16) rajoitteiden (4.17) – (4.19) vallitessa. Otoskoot tallennetaan vektoriin $\boldsymbol{n^0} = (n_1^0,..., n_D^0)^T$, jonka komponentit lasketaan kaavalla $n_d^0 = N_d / k_d^0$.

Tässä tutkimuksessa 34 alueen NLP-kiintiöinti on ratkaistu Excel-taulukkolaskentaohjelman Ratkaisin-apuohjelman GRG Nonlinear –moduulin avulla. Tähän tutkimukseen sovelletut otoskokojen johtaminen sekä lopulliset otoskoot 34 ja 14 alueelle on kuvattu tarkemmin alaluvussa 6.4.

4.2.5 Monivaiheinen otanta-asetelma aluekiintiöinnissä

Aluekiintiöinti voi liittyä myös tilanteeseen, jossa koko otoksen poiminta voi jakaantua useampaan vaiheeseen. Tavoitteena voi olla esimerkiksi estimoinnin tehostaminen pienten alueiden tasolla, joskus suurempien alueiden tai perusjoukon estimointitehokkuuden kustannuksella. Klassinen esimerkki tästä on kohdassa alaluvussa 4.3 (Rao) kuvattu Kanadan työvoimatutkimuksessa käytettävä 2-vaiheinen otanta, jossa 10 provinssista poimitaan ensi vaiheessa 42 000 kotitaloutta ja sitten loput 19 000 taloutta provinssia pienemmiltä UIR-alueilta.

Pahkinen (2012) kuvaa esimerkeillä 2-vaiheista otantaa, jonka tavoitteena on estimoinnin tehostaminen. Ensin poimitaan suhteellisen iso SRS-otos, jos se voidaan tehdä kohtalaisen pienin kustannuksin. Saatu (alustava) otos jaetaan ositteisiin, joiden sisältä voidaan poimia pienempi, lopullinen tutkittava otos esim. ositetulla otannalla sopivaa kiintiöintiä käyttäen. Menetelmän etuna on se, että otanta on tietyssä mielessä kattavampi, vaikka lopullinen otoskoko on etukäteen kiinnitetty. Samansuuntainen ajatus on Falorsin ja Righin (2008) artikkelissa, jossa tehdään päällekkäisiä ositusjakoja eri kriteerein.

4.2.6 Optimaalisen aluekiintiöinnin etsiminen simulointien tai esiotoksen avulla

Monissa otanta-alan tutkimuksissa asetelman optimointi edellyttää tarkkaa tietoa vastemuuttujan arvoista. Koska sellaisia ei ole käytettävissä, eräs käyttökelpoinen menetelmä on poimia perusjoukosta pienehkö esiotos ja laskea sen havaintoarvoista jonkin sopivan mallin avulla puuttuvien arvojen tilalle sijaismuuttujan arvot. Esimerkiksi Fabrizio ja Trivisano (2007) ovat käyttäneet tätä menetelmää ja myös otossimulointia johtopäätöstensä perustelemiseksi. Vastaavaa ratkaisua ovat käyttäneet myös Choudry ym. (2012). He ovat poimineet sijaismuuttujan arvot toistuvien tutkimusten aikaisemmista havaintoaineistoista,

Otossimulointien avulla voidaan hakea kokeellisesti sellaisia kiintiöintejä, jotka mahdollistavat tehokkaiden ja luotettavien estimointitulosten saavuttamisen, kuten esimerkiksi mahdollisimman pienen MSE:n, CV:n, ennustevirheen tai harhan. Optimien etsiminen voi kohdistua yksittäisiin alueisiin tai niiden kokonaisuuteen. Kokeellisen kiintiöinnin taustalla on se, että analyyttinen ratkaisu (esim. MSE:n minimointi kaikille alueille yhtä aikaa tai alueiden MSE-keskiarvon minimointi) on monien mallien kohdalla käytännössä erittäin monimutkaista, joten simulointi saattaa olla ainut keino saada edes arvioiduksi alueiden mahdolliset otoskoot. Ongelmana on kuitenkin sen seikan todistaminen, että saadut ratkaisut todella ovat parhaita tai ainakin lähellä niitä.

Tämä tutkimus käynnistyi aikanaan siltä pohjalta, että simuloiduista SRS-otoksista etsittiin ne, joista lasketut tehokkuutta ja luotettavuutta ilmaisevat mittarit (alueiden MSE-ja CV-keskiarvot, keskimääräinen suhteellinen virhe ja harha yms.) olivat parhaita. Tämän jälkeen tutkittiin, millaisia nämä otokset olivat apumuuttujan aluekohtaisten ominaisuuksien kannalta tarkasteltuina. Tavoitteena oli hakea näille ”parhaille” otoksille yhteisiä piirteitä ja lainalaisuuksia, mutta niiden löytäminen osoittautui hyvin hankalaksi varsinkin silloin, kun useiden estimointitavoitteiden piti toteutua yhtä aikaa.

Simulointitulosten perusteella voitiin kuitenkin tehdä eräitä johtopäätöksiä: 1) estimointitulokset eivät olleet välttämättä huonoja, vaikka joiltakin alueilta ei ollut lainkaan havaintoja, 2) pientä MSE-keskiarvoa ja pientä keskimääräistä suhteellista virhettä on vaikea saavuttaa samanaikaisesti, vaan on tehtävä kompromisseja ja 3) pelkät apumuuttujan aluekohtaiset varianssit eivät näyttäneet olevan merkittävä tekijä otoskokojen määräytymisessä. Merkittävä puute oli tietysti se, että simuloiduista otoksista ei pystytty havaitsemaan systemaattisesti piilevänä esiintyvää alueiden välistä vaihtelua, joka on käytetyn estimointimallin keskeinen ominaisuus.

Ratkaisun etsiminen otossimulointien avulla ei ole hyödytöntä, vaan niiden antamat tulokset voivat näyttää oikean suunnan myöhemmälle analyyttiselle ratkaisulle tai ne voivat tukea sitä. Jos ongelma on niin monimutkainen, ettei analyyttista ratkaisua voida johtaa, ovat otossimuloinnit ainoa keino saada edes jonkinlainen käsitys ongelmaan liittyvistä tekijöistä ja niiden keskinäisistä vuorovaikutuksista, mutta analyyttinen todistaminen jää tietysti puuttumaan.

4.2.7 Muita aluekiintiöintiin liittyviä tutkimuksia

Falorsin ja Righin (2008) tutkimus pohjautuu ositusperiaatteeseen ja tasapainotettuun otantaan. Otanta-asetelma on kehitetty tilanteeseen, jossa *N* tilastoyksikköä sisältävä perusjoukko on jaettu monella eri tavalla (ositukset *1,…,B*) toisensa poissulkeviin pienalueisiin. Näitä on M_b kappaletta osituksessa *b*, ja N_{bd} tarkoittaa ko. osituksen alueen d tilastoyksiköiden määrä perusjoukossa, jolloin on voimassa $\sum_{d=1}^{M_b} N_{bd} = N$. Merkintä U_{bd} tarkoittaa jälkiositteen *b* alueelle *d* kuuluvia perusjoukon tilastoyksiköitä. Kaikkiaan eri osituksista saadaan $\sum_{b=1}^{B} M_b = Q$ erilaista pienaluetta. Otoskoko on kiinteä (*n*). Tavoitteena on tuottaa alue-estimaatit vastemuuttujien aluekohtaisille kokonaismäärille, joiden otantavirheet (varianssit) jäävät ennalta asetettujen rajojen alapuolelle. Tekijät ovat käyttäneet suoraa mallitehosteista modifioitua GREG-estimaattoria, ja otantakiintiöinti perustuu alueilta saatavan lisäinformaation (apumuuttujien) käyttöön. Käsite tasapainotettu otanta tarkoittaa seuraavaa: 1) asetelmaperusteisessa malliavusteisessa estimoinnissa otoksessa lasketut apumuuttujien Horvitz-Thompson -tyyppiset keskiarvo- tai kokonaismääräestimaatit ovat yhtä suuret kuin niiden tunnetut vastineet koko perusjoukossa ja 2) malliperusteisessa estimoinnissa apumuuttujien otoskeskiarvot ovat yhtä suuret kuin niiden tunnetut vastineet perusjoukossa.

Otoskokojen n_{bd} (*b* = jälkiosite, *d* = alue) johtaminen perustuu sisältymistodennäköisyyksiin π_k. Otoskoko johdetaan kaavalla $n_{bd} = \sum_{k \in U_{bd}} \pi_k$ kahdessa vaiheessa: 1) ensin määritellään alustavat sisältymistodennäköisyydet π_k' ratkaisemalla asetettu minimointiongelma vastemuuttujan kokonaismääräestimaatin varianssin asettamissa rajoissa, ja 2) kalibrointivaiheessa tarkennetaan vaiheessa 1 saatuja sisältymistodennäköisyyksiä, jotta voidaan varmistua siitä, että otoskoot n_{bd} ovat kokonaislukuja.

Joskus esiintyy tilanteita, joissa kokonaisotoskoko *n* on kiinnitetty, mutta ei ole riittävästi informaatiota johtaa otoskokoja edellä kuvatulla tavalla. Tutkimuksen tekijät ehdottavat tällöin käytettäväksi yksinkertaista kiintiöintikaavaa pienalueiden otoskokojen laskemisen pohjaksi:

$$n_{bd} = \alpha_b \, n(N_{bd} / N) + (1 - \alpha_b) n / M_b ,$$

missä termi α_b on välillä 0–1 oleva vakio ja joka on määriteltävä estimoinnin painotustarpeen mukaan. Kaava on välimuoto suhteellisen ja tasakiintiöinnin välillä, joten sen perusteella ei otoskoko tule koskaan nollaksi. Lopulliset otoskoot määräytyvät käytettävien apumuuttujien alue-estimaattien varianssien optimoinnin ja kalibroinnin perusteella. Costa ym. (2004) ovat kehittäneet vastaavanlaisen tasa- ja suhteellisen kiintiöinnin välimuodon yhden ositussäännön mukaisesti. Tavoite on tasata todennäköisyysotannasta aiheutuva satunnaisuus pienten ja suurten alueiden välillä.

Falorsin ja Righin kuvaama kiintiöinti on mielenkiintoinen ja sovelluskelpoinen myös mallipe-rusteiseen estimointiin. Tekijät ovat tehneet simulointikokeita, joissa he ovat muodostaneet erilaisia osa-aluejakoja jopa 44 apumuuttujan avulla ja verranneet keskenään seuraavien estimaattorien tehokkuutta: Horvitz-Thompson, synteettinen estimaattori ja modifioitu GREG-estimaattori. Vertailu, joka perustui suhteellisen harhan (ARB) ja MSE:n analysointiin, osoitti modifioidun GREG-estimaattorin hyvää suorituskykyä.

Aluekiintiöinti perustuu analyyttiseen tarkasteluun, mutta alueiden välisen vaihtelun huomioon ottaminen ei nouse esille mitenkään. Otoskokojen laskennassa käytettävä painotus johtaa kysymykseen, mihin se voi perustua. Kun lopputulos sijoittuu aina suhteellisen ja tasaisen kiintiöinnin välille, tuntuvat mahdollisuudet rajoitetuilta.

Khan ym. (2010) ovat johtaneet epälineaarista ohjelmointia (AIMNLPP) käyttävään, usean optimointikriteerin kombinaatioon perustuvan optimaalisen otoskiintiöinnin seuraavaan tilanteeseen: jokaisesta perusjoukon tilastoyksiköstä mitataan *p* eri ominaisuutta kuvaavien muuttujien Y_1, Y_2,…, Y_p arvot, ja estimoitavina ovat em. muuttujien keskiarvot. Muuttujat voivat korreloida keskenään. Käytettävissä on lisäksi *p* kpl apumuuttujia X_1, X_2, …, X_p, yksi vastaavaa Y-muuttujaa kohti. Apumuuttujat vaikuttavat erikseen ja yhdessä. Havaintoja kerätään alueilta, joita on *D* kappaletta. Ositekohtaiset otoskoot ($n_1, n_2, ..., n_D$) ovat satunnaismuuttujia. Käytössä on ollut malliavusteiden suhde-estimointi ja regressioestimointi.

Optimoinnin tavoitteena on ollut minimoida Y-muuttujien keskiarvojen varianssien lisäysten painotettu keskiarvo, missä lisäykset johtuvat siitä, ettei käytetä yksittäisten keskiarvojen varianssien optimointia sellaisenaan, vaan optimointi tähtää estimoinnin kokonaistehokkuuden vähenemisen minimointiin. Syyt, jotka johtavat tällaiseen periaatteeseen, johtuvat käytettävissä olevien resurssien (budjetti, ositteiden koko, minimiotoskoot yms.) rajallisuudesta. Artikkelissa on johdettu optimaalisten otoskokojen lausekkeet ositteiden otosmäärille $n_1, n_2, ...n_D$ annettujen rajoitteiden vallitessa. Yksi rajoite oli se, että alueen *d* otoskoon n_d tuli olla välillä 2 - N_d, missä N_d = alueen *d* tilastoyksiköiden määrä. Laskennalliset otoskoot pyöristettiin mahdollisuuksien mukaan lähimpään kokonaislukuun. LINGO-ohjelmistoa (kehitetty lineaaristen ja epälineaaristen optimointiongelmien ratkaisuun) sovellettiin mm. aineistoon, joka sisälsi kahden eri viljalajin sadon (Y_1 ja Y_2) ennustamisen kahden aikaisemman vuoden sadon (X_1 ja X_2) avulla, ja tuloksia verrattiin eräillä muilla menetelmillä saatuihin vastaaviin tuloksiin. Loppupäätelmissä todettiin, että tässä kuvattu menetelmä on parempi kuin aiemmin kehitetyt, mutta käyttökelpoisuus edellyttää riittävän isoa kokonaisotoskokoa, jotta mm. ositteiden laskennalliset otoskoot saadaan järkeviksi.

Artikkelista ei sinänsä ole hyötyä tämän tutkimuksen ongelman ratkaisemisessa, mutta siitä voisi johtaa yhden vastemuuttujan ja apumuuttujan tapaukseen oman sovelluksen: minimoidaan vastemuuttujan alueellisten kokonaismäärien estimoinnissa kokonaistehokkuuden häviö tiettyjen rajoitteiden vallitessa. Myös se on ansiokasta, että jälleen on kehitetty yksi uusi optimaalisen aluekiintiöinnin analyyttinen ratkaisu.

Keto ja Pahkinen (2010) ovat lähestyneet optimaalista aluekiintiöintiä kokeellisesti hierarkkiseen lineaariseen sekamalliin perustuvassa EBLUP-estimointia käyttävässä tutkimuksessaan, jossa estimoitavina olivat vastemuuttujan aluekohtaiset kokonaismäärät, ja käytössä oli yksi apumuuttuja. Ensimmäisessä vaiheessa on poimittu 19 aluetta (ja samalla 19 ositetta) sisältävästä havaintoaineistosta (400 kunnan aineisto 2007) 1500 SRS-otosta ja valittu sitten niistä otokset, joista lasketut estimoinnin tehokkuusmittarien arvot olivat pienimmät. Näiden otosten perusteella määräytyivät kokeellisen kiintiöinnin otoskoot. Toisessa vaiheessa tekijät ovat poimineet 1500 aluekiintiöityjä otoksia a) uudenlaisen kokeellisen kiintiöinnin ja sille vertailuksi valittujen b) suhteellisen ja c) tasakiintiöinnin mukaisesti, minkä jälkeen he ovat vertailleet kiintiöintejä estimoinnin eri laatumittarien perusteella.

Kokeellinen kiintiöinti osoittautui monessa suhteessa tehokkaammaksi kuin kaksi tavanomaista kiintiöintiä. Erityinen havainto oli se, että otoksen puuttuminen ($n_d = 0$) joiltakin alueilta ei välttämättä heikentänyt estimointituloksia. Valitun kokeellisen kiintiöinnin tehokkuuden analyyttinen todistaminen kuitenkin puuttui, joten kokeellisella kiintiöinnillä ei ole yleistä todistusvoimaa. Joka tapauksessa tutkimus antoi arvokasta tietoa siitä, mihin jatkotutkimusta kannattaa suunnata.

4.3 Tehokkaan alue-estimoinnin kokonaisstrategian luominen

Onnistuneen alue-estimoinnin ensimmäisiä edellytyksiä on sellaisen kokonaisstrategian kehittäminen, joka ottaa alueet huomioon tutkimuskohteina jo otanta-asetelmassa. Eräät kirjoittajat ovat käsitelleet aihetta varsin perusteellisesti ja ehdottaneet, millaisia asioita tällaiseen monipuoliseen strategiaan tulisi kuulua.

Singh ym. (1994) ovat esittäneet ensimmäisiä merkittäviä mielipiteitä pienalue-estimointiin liittyvän havaintoaineiston hankintaan ja otanta-asetelman suunnitteluun erityisesti laajoissa kyselytutkimuksissa. Estimoinnin kannalta on merkittävää, onko otettava huomioon ensisijaisesti kokonaisuus (esim. koko maa), aluetaso (esim. maakunnat) vai molemmat edellä mainitut tasot yhtä aikaa, ikään kuin tasavertaisina. Kirjoittajat toteavat ensin, että olisi luotava kokonaisstrategia, jonka mukaan otanta ym. tutkimuksen toimeenpanoon liittyvät asiat suunnitellaan. Strategiaan kuuluu mm. seuraavia asioita: 1) päätös siitä, millaisia estimointimenetelmiä (suorat, epäsuorat) on järkevää käyttää, 2) estimoinnin painopisteen (aluetaso, kokonaistaso) määrittely, 3) otanta-asetelman suunnittelu sekä 4) otoskiintiöinnin suunnittelu aluetasolla.

Ryvästys tulisi minimoida ja ositetun otannan tulisi perustua mieluummin pienempiin ositteisiin kuin suuriin. Alueiden rakenteiden tunteminen edistää kyselytutkimuksen suunnittelua. Tässä viitataan esimerkkiin, jonka myös Rao (2003) on maininnut. Estimointiosassa käsiteltiin varsin vähän malliperusteisia estimaattoreita. Niiden käyttö on perusteltua silloin, kun estimaattorin keskineliövirhe (MSE) on olennaisesti pienempi kuin asetelmaperusteisen estimaattorin varianssi. Kirjoittajat eivät kuitenkaan määrittele, kuinka suuri tämä riittävä ero on.

Rao (2003) pohtii pienalue-estimointiin liittyviä kiintiöintikysymyksiä lähinnä suoran estimoinnin näkökulmasta. Ideaalitavoitteena on hänen mukaansa löytää optimaalinen otanta-asetelma, joka minimoi vastemuuttujan estimaattorin varianssin tai MSE:n annettujen kokonaiskustannus-

ten määräämissä rajoissa. Lopputuloksena on kuitenkin todennäköisesti vain jonkinlainen kompromissi, ei optimaalinen otanta-asetelma, eli ainoastaan mahdollisimman lähellä optimia oleva.

Rao nostaa kiintiöinnistä esille muutamia asioita. Suurten ositteiden sijaan Rao suosittelee ositteiden koon tasaamista esimerkiksi pienentämällä suurten alueiden kokoja leikkauksin tai muiden vastaavien operaatioiden tuloksena. Oikea otoksen kohdentaminen ja toteutus useammassa vaiheessa voi täyttää luotettavuusvaatimukset sekä pienille että suurille alueille. Koko otosta ei kohdenneta yhdellä kertaa, vaan ensin poimitaan riittävä otos suuremmilta alueilta ja loppuotos kohdennetaan yhdessä tai useammassa vaiheessa järkevästi pienemmille alueille, tai päinvastoin. Lopputuloksena voi olla luotettavat estimaatit kaikenkokoisille alueille. Rao mainitsee esimerkin Kanadan työvoimatutkimuksesta, jossa tällaisella asteittaisella otannalla voitiin pienentää vastemuuttujan CV-arvoja huomattavasti pienillä alueilla, jopa puoleen alkuperäisestä. Tämän tutkimuksen ovat kuvanneet tarkemmin Singh ym. (1994). Kokonaisotoskoko oli 59 000 kotitaloutta, joista 42 000 poimittiin ensi vaiheessa 10 provinssista (läänejä vastaavat) ja loput 17 000 kotitaloutta poimittiin provinsseja pienemmiltä 61 UIR-alueelta. Ensimmäisen vaiheen tavoitteena oli saada hyvät estimaatit koko maan ja provinssien tasolla. Toisen vaiheen yliotannalla saatiin luotettavat ennusteet UIR-alueisiin valtakunnan tason kustannuksella, mikä tarkoitti sillä tasolla pientä, mutta ei merkittävää CV-arvojen nousua.

Rao keskittyy erilaisten alue-estimointimenetelmien esittelyyn. Hän ei käsittele aluekiintiöintiä erityisen paljon, eikä varsinkaan analyyttisessä mielessä. Alueiden välinen vaihtelu ja sen vaikutus estimointimenetelmään tai otoskiintiöintiin ei ole systemaattisesti esillä. Kirjan anti ei ole kovinkaan merkittävä tälle tutkimukselle.

Marker (2001) on sitä mieltä, että koska tutkimusympäristö on sangen vaihteleva aluemielessä, on aina tarvetta malliperusteiseen estimointiin. Sopivalla otanta-asetelmalla ja otannan kiintiöinnillä voidaan kuitenkin luoda edellytykset suoran estimoinnin käytölle myös pienalue-estimoinnissa. Keinoja ovat osittaminen ja ylisuuret otantamäärät joillakin tärkeillä alueilla. Tasapaino on löydettävissä kokonaistehokkuuden kannalta. Monesti on estimoitavina populaatio- ja aluetunnuslukuja samanaikaisesti. Jos yhdelle tai useammalle alueelle lisätään otantaa, ei kokonaisuus saa kärsiä liikaa. Ositteiden kokoa on syytä harkita tarkkaan. Marker ei määrittele täsmällisesti, mitä asioita tässä tilanteessa on otettava huomioon.
Marker on esittänyt varteenotettavia periaatteita otoskiintiöinnin toteuttamiseen, mutta ei ole pohtinut asiaa analyyttisesti, esimerkiksi sitä, miten alueiden ominaisuudet vaikuttavat alueiden

otantamääriin. Lisäksi artikkelissa on keskitytty vain CV:n pienentämiseen tai tasaamiseen alueiden kesken.

4.4 Tiivistelmä aiemmista tutkimuksista

Edellä referoitujen tutkimusten tuloksista voidaan tehdä seuraava yhteenveto:

1) Otoskiintiöintiä käsittelevä tutkimus, kuten muukin estimointia käsittelevä otantatutkimus, näyttää liittyvän enemmän suoraan estimointiin kuin mallipohjaiseen.
2) Uudenlaisia ratkaisuja tehokkaaseen otoskiintiöintiin kehitetään jatkuvasti, kuten nähdään uusimmasta kirjallisuudesta.
3) Optimaaliseen aluekiintiöintiin liittyvät kysymykset ovat tutkimuksissa esillä lähes aina, vaikka niitä etsitään analyyttisin keinoin vain harvoissa tapauksissa. Merkittävin syy tähän niukkuuteen on optimoitavien lausekkeiden monimutkaisuus.
4) Mallipohjaisia estimaattoreita voidaan soveltaa monipuolisemmin alue-estimoinnissa suoriin estimaattoreihin verrattuina. Mikä tahansa malli ei kuitenkaan sovellu havaintoaineiston estimointiin, joten tämän valinnan pitää perustua huolelliseen harkintaan. Lisäksi liian monimutkainen malli voi vaikeuttaa estimointia.
5) Tehokas alue-estimointi edellyttää asetelman mukauttamista alueiden sisäiseen ja väliseen vaihtelurakenteeseen tutkittavien muuttujien osalta. Kuitenkin käytetyt mallit sisältävät aluevaikutuskomponentin vain harvoissa tutkimuksissa.
6) Pitkäaikainen ja usealla taholla tehty aluetilastojen tuotanto osoittaa, että pelkkä otanta-asetelman tunteminen ei riitä otannassa, vaan lisäksi on tarvittu aluemallista saatavaa tukea laskentavaiheisiin.
5) Aluekiintiöinti otanta-asetelman sisällä on saanut useita ratkaisuja, joissa on käytetty alueisiin liittyviä ennakkotietoja. Näissä ei kuitenkaan ole käytetty estimoinnin tukena sovellettuja malleja. Tässä työssä on kokeiltu aluekiintiöintiä, jossa tavanomaisten kiintiöintivälineiden joukkoa on täydennetty lisäämällä mukaan yleisesti käytetyn EBLUP-mallin informaatio.

4.5. Tehokkaan otoskiintiöinnin analyyttisen ratkaisun mahdollisuus malliperusteisessa pienalue-estimoinnissa

Alue-estimointiin liittyvät valinnat lähtevät seuraavista peruskysymyksistä: 1) mitä tunnuslukuja estimoidaan 2) millaista aluekohtaista vastemuuttuja- sekä aputietoa on mahdollista käyttää estimoinnin pohjaksi 3) miten voimakas yhteys niiden välillä on ja 4) mitä tiedetään alueiden rakenteista ja niiden samankaltaisuudesta tai erilaisuudesta. Jos estimointi perustuu vain vastemuuttujatietoon, on sitä kerättävä kaikilta alueilta, jotta jokaiselle alueelle saadaan estimaatit tai ennusteet, ja estimointimenetelmä on tällöin suora. Jos on käytettävissä kovariaattitietoa (apumuuttujatietoa) otosyksiköistä tai koko perusjoukon tilastoyksiköistä, voidaan vastemuuttujan aluekohtaista vaihtelua selittää kovariaattien vastaavalla vaihtelulla. Tällöin voidaan nähdä esimerkiksi vaihtelun voimakkuus, jolloin otantaa voidaan kohdistaa alueille voimakkuuden mukaisesti. Edelleen, jos löydetään keskenään samankaltaisia alueita, ne voidaan yhdistää yhdeksi isommaksi alueeksi ja poimia siitä ikään kuin yksi otos, tai sitten voidaan jättää jotkin samankaltaisista alueista kokonaan ilman havaintoja.

Aluekiintiöinti päätetään ennen otoksen poimintaa, ja siinä vaiheessa on käytettävissä yleensä vain vähän estimoinnissa tarvittavaa tietoa. Myös valittu malli ja siihen liittyvä estimointimenetelmä olisi otettava kiintiöinnissä huomioon, koska estimaatit tai ennusteet sekä niiden varianssit tai MSE:t lasketaan valinnasta riippuen asiaankuuluvilla kaavoilla, jotka sisältävät alueiden otoskoot. Suorassa estimoinnissa voi kiintiöinti perustua vain hyvin likimääräiseen arviointitietoon vastemuuttujan arvojen vaihtelusta, kun taas epäsuorassa estimoinnissa on käytettävissä usein varsin tarkkoja kovariaattitietoja. Valinnan soveltuvuuden arviointiin on usein käytettävissä tapauskohtaisesti omat menetelmänsä, mutta käytännön tilanteessa ei ensimmäisenä tulisi kiinnittää mallia, johon sitten yritetään sovittaa havainnot enemmän tai vähemmän onnistuneesti. Ennen valintaa kannattaa perehtyä käytettävissä olevaan havaintoaineistoon sekä alueisiin liittyvään informaatioon sekä ottaa huomioon estimoinnin tavoitteet.

Tutkimuskäytäntö aluetilastojen tuotannossa on johtanut mallitehosteiseen tai –perusteiseen estimointiin. Yleisesti käytettyjä menetelmiä ovat mm. regressiotehosteinen estimointi, yleistetty regressioestimointi (GREG), asetelma- ja malliperusteinen suhde-estimointi, jälkiositusestimointi sekä EBLUP-estimointi. GREG- ja EBLUP-estimointi voivat perustua erilaisiin malleihin. Viimeksi mainitun estimoinnin kohdalla voidaan erottaa aluetason mallit ja yksikkötason mallit, joissa niissäkin on mukana myös aluevaikutus. Tässä tutkimuksessa otetaan aluekiintiöinnin työkaluiksi 1) alueiden sisäiset ominaisuudet ja 2) alueiden väliset yhteydet, mikä kiintiöinnissä jää

yleensä huomioon ottamatta. Tarkoitus on selvittää yhteydet apumuuttujien avulla, ja aluekiintiöinti ehdollistetaan malliin, joka kuvaa tulosmuuttujien riippuvuutta apumuuttujista ja aluevaikutuksista.

Määrätyissä tilanteissa voidaan tehokas aluekiintiöinti johtaa analyyttisesti, jos tehokkuutta mittaavat kriteerit ovat riittävän yksinkertaisia, olipa kyseessä suora, suora malliavusteinen tai malliperusteinen estimointi. Tehokkuuden kriteerinä käytetään yleensä vastemuuttujan aluekohtaisten varianssien tai keskineliövirheiden (MSE) keskiarvon tai painotetun keskiarvon minimointia, johon alueiden optimaalisten otoskokojen lausekkeiden johtaminen perustuu. Joissakin tapauksissa otoskokojen optimointi perustuu alueiden CV-arvoille asetettuihin ylärajoihin. Sen sijaan estimaattien tai ennusteiden muuta tehokkuutta ja tarkkuutta (suhteellinen virhe tai harha yms.) ei ole käytetty kriteereinä analyyttisissa ratkaisuissa.

Monet malliavusteiset tai –perusteiset estimointimenetelmät ovat sellaisia, joissa otoskiintiöintiongelmaan ei ole täsmällistä analyyttistä ratkaisua liian monimutkaisten lausekkeiden vuoksi, minkä mm. Longford (2006) toteaa. Ratkaisuksi voidaan esittää korkeintaan approksimaatioita tai periaatteita, jotka perustuvat kaavojen tutkimiseen, intuitioon, asiantuntija-arvioon tai kokeelliseen lähestymiseen. Tässä tutkimuksessa on johdettu analyyttisesti kolme eri aluekiintiöintiä tilanteessa, johon on valittu aluemalli ja siihen soveltuva EBLUP-estimointi. Malli sisältää vaste- ja apumuuttujan lisäksi erityisen piilotetun aluevaikutuskomponentin. Apuna käytetään myös vastemuuttujan korvaavaa sijaismuuttujaa, joka johdetaan perusjoukosta poimitusta pienestä esiotoksesta. Kiintiöintien johtamisessa hyödynnetään tietoa, joka on käytettävissä mallista ja estimointimenetelmästä, tilastoyksiköistä sekä alueiden sisäisestä ja välisestä vaihtelusta. Kiintiöintien johtaminen kuvataan tarkemmin luvussa 5.

Aluekiintiöinti on mahdollista tehdä malliperusteiseen estimointiin analyyttisin perustein. Kolmen johdetun kiintiöinnin tehokkuutta suhteessa viiteen muuhun kiintiöintiin mitataan vertailussa, jonka tulokset esitellään luvussa 6.

5 TUTKIMUKSESSA JOHDETUT JA TESTATUT KIINTIÖINNIT

5.1 Malliperusteiseen estimointiin soveltuvan kiintiöinnin tarve

Luvussa 4 esitellyistä kiintiöinneistä parametriperusteiset nojautuvat alueiden sisäisiin ominaisuuksiin, mutta nekään eivät hyödynnä estimointimenetelmissä käytettyjä malleja. Tässä tutkimuksessa on alue-estimoinnin perustana alaluvussa 3.5.3 esitelty aluemalli (3.13). Mallin avulla ei sellaisenaan voida kuvata vastemuuttujan *y* vaihtelua, koska lausekkeen (3.14) perusteella vastemuuttujan *y* varianssi on mallin varianssikomponenttien summa. Mallia on kuitenkin käytetty apuna uusien kiintiöintien kehittämisessä. Kiintiöinnit perustuvat malliin tai sijaismuuttujaan tai molempiin. Lähtökohtana on apumuuttujasta *x* käytössä oleva rekisteritieto sekä kahden kiintiöinnin kohdalla lisäksi perusjoukosta poimittava pieni esiotos. Vastemuuttujan aluekohtaista vaihtelua pystytään mittamaan jossain määrin, kun apumuuttujan alueiden välinen vaihtelu tunnetaan, ja aluekiintiöinti kehitetään tältä pohjalta.

5.2 Vastemuuttujaa korvaavan sijaismuuttujan johtaminen esiotoksesta

Tehokas otoskiintiöinti vaatisi täydellisen informaation siitä, mitä arvoja vastemuuttuja *y* saa ja miten sen arvot vaihtelevat alueittain, mutta vain otosarvot tunnetaan. Koska käytössä on täydellinen apumuuttujainformaatio (arvo tunnetaan jokaisesta tilastoyksiköstä), voidaan sitä käyttää apuna kiintiöinnissä, joka perustuu vastemuuttujaa korvaavan sijaismuuttujan y^* käyttöön. Sijaismuuttujalla tarkoitetaan muuttujaa, joka korvaa alkuperäisen vastemuuttujan esimerkiksi siten, että sen arvot on laskettu jollakin tavalla (apumuuttujista) tai sitten ne ovat todellisen vastemuuttujan aikaisempia arvoja, kun kyseessä on toistuva tutkimus. Sijaismuuttujia ovat käyttäneet mm. Choudry ym. (2012) sekä Fabrizio ym. (2007), jälkimmäiset myös esiotosta.

Sijaismuuttujan tietojen keruussa on käytetty kaksiasteista otantaa. Alue on tulkittu myynnissä olevien huoneistojen rypääksi. Sijaismuuttujan muodostamisen päävaiheet ovat seuraavat:

1) Rypäät (alueet) lajitellaan nousevaan järjestykseen apumuuttujan vaihtelukertoimen (CV) arvon mukaan.
2) Ensimmäinen poiminta-aste on systemaattinen otanta, jossa poimitaan kolme ryvästä tasavälein. Kukin poimittu ryväs edustaa yhtä ryväsryhmää vaihtelukertoimen tason mukaan.
3) Jokaiselta kolmesta rypäästä poimitaan viiden huoneiston SRSWOR-otos. Näin saatavaa 15 huoneiston otosta nimitetään esiotokseksi, jonka poimintaan käytetään osa otantaresurssista.

4) Esiotoksen jokaisen kolmen alueen otoksista johdetaan tavallinen regressiomalli vastemuuttujan *y* ja apumuuttujan *x* välillä. Kolmen mallin avulla saadaan alueryhmien erilaisia ominaisuuksia ainakin jossain määrin esille.

5) Kutakin regressiomallia sovelletaan siihen ryhmään, josta ao. alue on peräisin. Ryhmän jokaiselle tilastoyksikölle lasketaan sijaismuuttujan y^* arvo (merkintä: y^*_{dk}) seuraavasti:

$$y^*_{dk} = \alpha_r + \beta_r x_{dk} \ (d = 1,2,\ldots, D;\ k = 1,2,\ldots, N_d\ ;\ r = 1,2,3),$$

missä *r* = CV-ryhmän numero. Poikkeuksena tähän sääntöön ovat esiotokseen kuuluvat 15 tilastoyksikköä, joiden sijaismuuttujan arvo on sama kuin todellinen *y*-arvo. Tämän jälkeen on käytössä *N* kappaletta (koko perusjoukon määrä) sijaismuuttujan arvoja. Nyt niistä voidaan laskea jokaiselle alueelle sijaismuuttujan tunnusluvut, mm. keskiarvot, varianssit ja vaihtelukertoimet (CV).

Havaintoaineistoihin sovellettu sijaismuuttujan laskentatekniikka esitellään alaluvuissa 6.1.6 ja 6.1.7.

5.3 Sijaismuuttujan suhteellisiin variansseihin perustuva kiintiöinti

Alueiden otoskoot johdetaan sen oletuksen pohjalta, että ensin poimitaan kultakin alueelta *d* SRSWOR-otos, jonka otoskoko on n_d, minkä jälkeen voidaan laskea alueille sijaismuuttujan otoskeskiarvojen varianssit ja vaihtelukertoimet. Itse kiintiöinti perustuu aluekohtaisten CV-arvojen neliöiden eli suhteellisten varianssien painotetun keskiarvon minimointiin alueiden otoskokojen n_d funktiona. Otoskokojen johtaminen on kuvattu yksityiskohtaisesti seuraavaksi.

Alueelta *d* poimitaan SRSWOR-otos, joka sisältää sijaismuuttujan arvoja n_d kappaletta (otoskoko), ja alueiden otoskokojen summa $\sum_{d=1}^{D} n_d = n$. Sijaismuuttujan otoskeskiarvon $\bar{y}^*_d$ varianssin lauseke on

$$V(\bar{y}^*_d) = (1 - n_d / N_d) S^2_d(y^*) / n_d = (1/n_d - 1/N_d) S^2_d(y^*), \qquad (5.1)$$

missä oikean puolen varianssi on sijaismuuttujan varianssi alueella *d*, ja lauseke $1 - n_d / N_d$ on äärellisyyskorjaus. Otoskeskiarvon $\bar{y}^*_d$ hajonta on varianssin (5.1) neliöjuuri, ja otoskeskiarvon CV saadaan, kun hajonta jaetaan sijaismuuttujan aluekeskiarvolla $\bar{Y}^*_d$. Alueiden otosten CV-

arvot ovat samat, vaikka otoskeskiarvojen tilalla käytettäisiin kokonaismäärien estimaattoreita. Otoskeskiarvon $\bar{y}_d^*$ suhteellinen varianssi RV on CV:n neliö, jonka lauseke on

$$RV(\bar{y}_d^*) = V(\bar{y}_d^*)/(\bar{Y}_d^*)^2 . \tag{5.2}$$

Kiintiöintiin liittyvien otoskokojen n_d johtaminen ei perustu alueiden otoskeskiarvojen CV-arvoihin, vaan alueiden suhteellisiin variansseihin. Syy on se, että suorien CV-arvojen käyttö johtaisi 4. asteen yhtälöön, jota on vaikea ratkaista analyyttisesti. Minimoitavana on sijaismuuttujan suhteellisten varianssien painotettu keskiarvo

$$\sum_{d=1}^{D} w_d RV(\bar{y}_d^*)$$

otoskokojen n_d (d = 1,…,D) funktiona Lagrangen menetelmän avulla, kun voimassa on lisäksi oletus $\sum_{d=1}^{D} n_d = n$. Painokertoimien w_d arvoille ei aseteta ennakko-oletuksia.

Minimointiongelman ratkaisu on esitetty liitteessä H. Alueen d otoskoon n_d lauseke on

$$n_{d,sij} = \frac{\sqrt{w_d}\,C_d(y^*)}{\sum_{d=1}^{D}\sqrt{w_d}\,C_d(y^*)}\,n , \tag{5.3}$$

eli otoskoko on suoraan verrannollinen painon w_d neliöjuuren ja sijaismuuttujan y^* CV-arvon C_d tuloon. Kaava (5.3) vastaa luvussa 4 esitetyn Longfordin (2006) johtamaa kaavaa (4.8), joka perustuu aluevarianssien painotetun keskiarvon minimointiin. Painoilla w_d voidaan säädellä alueelle d tulevaa otoskokoa. Niiden summalle ei ole määritelty kiinteää arvoa. Jos painot ovat samat, on alueen otoskoko suoraan verrannollinen alueen CV-arvoon. Tällainen painotus suosii alueita, joiden CV-arvo on suuri, riippumatta alueen muista ominaisuuksista, mutta CV-arvoltaan pienille alueille tulee pieni otoskoko. Jos painojen summaksi määritellään tasan yksi, voidaan painoiksi määritellä esimerkiksi alueiden suhteelliset koot $w_d = N_d / N$. Laskelmissa käytetään painokertoimien w_d arvoina edellä kuvattuja alueiden suhteellisia kokoja. Suuret alueet, joilla on korkea CV-arvo, saavat suuren otoskoon, ja vastaavasti pienet alueet, joiden CV-arvo on alhainen, voivat saada hyvinkin pienen otoskoon. Nämä painot eivät sisällä mitään tietoa aluevaikutuksista. Lausekkeesta (5.3) voidaan vielä havaita, että niissä esiintyvät painojen w_d neliöjuuret, mikä tarkoittaa sitä, että painot eivät sellaisenaan vaikuta kiintiöintiin.

Otoskoon lauseke (5.3) sisältää vaihtelukertoimiin perustuvaa tietoa alueiden sisäisestä vaihtelusta. Sen sijaan on vaikeaa kehittää sellaiset painokertoimet w_d, joihin sisältyisi alueiden väli-

nen vaihtelu ja jotka voitaisiin ilmaista matemaattisesti. Kertoimiin pitäisi joka tapauksessa kuulua alueiden havaintoyksiköiden määrät N_d, koska MSE:n lauseke sisältää ne. Kerroin voi olla esimerkiksi muotoa $w_d = P_d N_d$, jossa osakerroin P_d edustaa alueen *d* tärkeyttä kiintiöinnissä. Koska kiintiöinti perustuu sijaismuuttujaan y^*, sen alueominaisuudet kuten keskiarvo ja vaihteluväli ovat tärkeitä aluevaikutusten mittaamisessa ja osakertoimien P_d määräämisessä. Täsmällisen matemaattisen kaavan kehittäminen vaikutuksen mittaamiseen on kuitenkin erittäin vaikeaa. Kiintiöinti pitää suunnitella EBLUP-estimointiin, jossa MSE-approksimaation (3.23) komponentit (3.24) sekä alue-ennuste (3.21) ovat tunnetusti monimutkaisia. Yksi lisäongelma on riittävän pitkän vaihteluvälin asettaminen kertoimille P_d, koska lausekkeen (5.3) perusteella P_d:n neliöjuuri pienentää kertoimien välisiä eroja huomattavasti.

Tähän tutkimukseen lasketut otoskoot 34 ja 14 alueen aineistoille on esitelty alaluvussa 6.4.

5.4 Sijaismuuttujan, apumuuttujan ja mallin käyttöön perustuva kiintiöinti

Toinen johdettava kiintiöintiratkaisu perustuu otossimuloinnin ja aluemallin yhteiskäyttöön. Alaluvussa 5.2 kuvatun sijaismuuttujan y^* arvot voidaan laskea perusjoukon jokaiselle havaintoyksikölle, joista tunnetaan myös apumuuttujien arvot. Näistä muuttujista on siis käytössä täydellinen rekisteriaineisto. Tällöin on mahdollista simuloida rekisteristä otoksia ja suorittaa niille EBLUP-estimointi, josta saadaan lasketuksi alueittain tarvittavat ennusteet ja niille tunnusluvut. Tässä johdettava kiintiöinti ehdollistuu ennalta asetetun optimointikriteerin saavuttamiseen, joka liittyy alueiden MSE-arvojen keskiarvoon.

Kiintiöinti toteutetaan seuraavien yleisperiaatteiden mukaan:

1) Kiinnitetään otoskoko *n*.
2) Edellä kuvatusta rekisteristä simuloidaan M kpl SRSWOR-otoksia, jolloin nämä kiintiöityvät satunnaisesti alueille.
3) Jokaiselle otokselle suoritetaan EBLUP-estimointi, minkä tuloksena saadaan otoksille lasketuiksi varianssikomponentit, sijaismuuttujan kokonaismäärien alue-ennusteet, niiden MSE-estimaatit ym. estimoinnin tuottamia suureita. Alueiden otoskoot ja estimoinnin tulokset tallennetaan otoskohtaisesti.

4) Rakennetaan optimointikriteeri, jonka mukaan simuloidut otokset voidaan asettaa järjestykseen. Kriteerinä on tässä otoskohtainen alueiden MSE-arvojen keskiarvo ja otosten järjestys nouseva, jolloin pienimmän MSE-keskiarvon sisältävä otos on ensimmäisenä.
5) Asetetun kriteerin mukaisista otoksista voidaan valita *m* kpl (*m* < *M*) alku- tai loppupäästä ja muodostaa kiintiöityneiden alueiden otoskokojen jakaumat.
6) Otoskokojen jakaumia tutkimalla voidaan muodostaa käsitys alueiden otoskokojen tasosta optimointikriteerin saavuttamiseksi.
7) Aluekohtaisten otoskokojen jakaumista johdetaan kiintiöinti, jossa alueiden otoskokojen summa on *n*. Kiintiöinnin lähtökohtana on jokin jakauman tunnusluku kuten mediaani. Jos tunnuslukujen summa ei ole tasan *n*, muutetaan joidenkin alueiden otoskokoja harkinnan mukaan ylös- tai alaspäin siten, että summaksi tulee *n*.

Otoskokojen johtaminen 34 ja 14 alueen aineistoille edellä kuvattujen periaatteiden mukaan sekä lopullisten otoskokojen määräytyminen on kuvattu tarkemmin alaluvussa 6.4.

5.5 MSE:n tärkeimpään komponenttiin ja apumuuttujaan perustuva *g1*-kiintiöinti

Kuten alaluvussa 3.5.4 todettiin, keskineliövirheen MSE:n ensimmäisen komponentin g_{1d} osuus on MSE:n kokonaisarvosta usein 85–90 % jopa yli 90 %. Näin on sitä todennäköisemmin, mitä voimakkaampi alueiden välinen vaihtelu on. Koska MSE:n approksimaatiokaava (3.23 ja 3.24) on hyvin monimutkainen, ei alueiden MSE-keskiarvojen minimointi ole mahdollinen, minkä vuoksi keskitytään vain tähän komponenttiin, jonka lauseke on

$$g_{1d}(\sigma_v^2,\sigma_e^2) = (N_d - n_d)^2(1-\gamma_d)\sigma_v^2 \ , \tag{5.5}$$

missä termi γ_d määritellään seuraavasti:

$$\gamma_d = \sigma_v^2/(\sigma_v^2 + \sigma_e^2 n_d^{-1}) = n_d\sigma_v^2/(n_d\sigma_v^2 + \sigma_e^2) \ .$$

Komponentin g_{1d} lauseke voidaan esittää myös seuraavassa muodossa:

$$g_{1d}(\sigma_v^2,\sigma_e^2) = (N_d - n_d)^2(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-1} \ .$$

Kiintiöinti perustuu siihen, että haetaan minimi alueiden g_{1d} -arvojen keskiarvolle, joka on määritelty otoskokojen funktiona. Ko. keskiarvon lauseke on

$$1/D\sum_{d=1}^{D} g_{1d}(\sigma_v^2,\sigma_e^2) = 1/D\sum_{d=1}^{D}(N_d - n_d)^2(n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} \ . \tag{5.6}$$

Varianssikomponenttien suhdetta merkitään symbolilla $\delta = \sigma_e^2/\sigma_v^2$, ja sen estimaattia symbolilla $\hat{\delta} = \hat{\sigma}_e^2/\hat{\sigma}_v^2$. Tämä suhde voidaan esittää myös alueiden välisen sisäkorrelaation φ (lauseke 3.19) avulla seuraavasti:

$$\varphi = \sigma_v^2/(\sigma_v^2 + \sigma_e^2) = 1/(1 + \sigma_e^2/\sigma_v^2) = 1/(1+\delta) \Rightarrow \delta = 1/\varphi - 1 \,. \tag{5.7}$$

Keskiarvolausekkeen (5.6) minimointi ja alueen *d* otoskoon n_d (d = 1,…, D) lausekkeen johtaminen sekä keskiarvolausekkeen minimi on esitetty liitteessä F. Otoskoon n_d lauseke voidaan esittää varianssisuhteen δ ja sisäkorrelaation φ avulla seuraavasti:

$$\begin{aligned} n_d &= \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} - \delta = \frac{N_d n - (N - N_d D - n)\delta}{N + D\delta} \\ &= \frac{N_d n - (N - N_d D - n)(1/\varphi - 1)}{N + D(1/\varphi - 1)} \,. \end{aligned} \tag{5.8}$$

Koska $n+\delta D < N+\delta D$, on helppo päätellä, että $n_d < N_d$. Lausekkeesta (5.8) nähdään myös, että sen arvo riippuu varianssisuhteesta δ, mutta ei suoraan varianssikomponenttien arvosta. Jos vaihtelu on pelkästään alueiden välistä (δ = 0 ja φ = 1), on lopputulos suhteellinen kiintiöinti.

Tutkitaan vielä, voiko lausekkeen (5.8) avulla laskettava otoskoko saada negatiivisen arvon. Ratkaistaan epäyhtälö

$$\frac{(N_d + \delta)(n + \delta D)}{N + \delta D} - \delta < 0 \,,$$

josta saadaan pienen sievennyksen jälkeen kaksi vaihtoehtoista epäyhtälöä:

$$\delta > N_d n / (N - N_d D - n) \;\text{ tai }\; n < (N\delta - N_d \delta D)/(N_d + \delta).$$

Jos alue on pieni (N_d pieni) tai otoskoko n on pieni tai molemmat ovat pieniä, on mahdollista, että alueen *d* otoskoko saa negatiivisen arvon. Tällainen tilanne voi syntyä myös silloin, kun varianssisuhde δ on riittävän iso eli kun vaihtelu on lähes kokonaan alueiden sisäistä. Riittävän iso otoskoko estää ongelman syntymisen. Laskennalliset otoskoot on luonnollisesti pyöristettävä kokonaisluvuiksi. Jos erikoistapauksissa tulee negatiivisia arvoja, ne muutetaan nolliksi.

Koska otoskokojen lauseke (5.8) sisältää malliin liittyvän sisäkorrelaation, joka riippuu otoksesta ja on tuntematon ennen kiintiöintiä ja otantaa, korvataan tämä ennen otoskokojen laskentaa osit-

teiden välisellä apumuuttujan *x* sisäkorrelaation muunnoksella, jota kutsutaan homogeenisuusmitaksi (luvun 6 lauseke 6.1). Perustelu tälle menettelylle on se, että koska apumuuttuja *x* vaikuttaa vastemuuttujaan *y*, siirtyy myös alueiden välinen vaihtelu apumuuttujan kautta vastemuuttujaan. Särndal (1992) on osoittanut, että homogeenisuusmitalla voidaan ilmaista em. vaihtelun osuus muuttujan kokonaisvaihtelusta, kun alueiden koot ovat erilaiset.

6 HAVAINTOAINEISTO JA SIMULOINTIKOKEET

6.1 Havaintoaineisto

6.1.1 Perusjoukko ja alueet

Tutkimuksessa on käytetty todellista havaintoaineistoa, jonka sisältämät tiedot on kerätty Alma Mediapartners Oy:n ylläpitämästä rekisteristä. Rekisteri sisältää koko maata koskevia tietoja kiinteistövälitysten kautta myytävistä kiinteistöistä (asunnot, loma-asunnot, tontit sekä maa- ja metsätilat), joten rekisterin ulkopuolelle jäävät suoraan myynnissä olleet kiinteistöt. Tietoja voi hakea ”Etuovi.com” -verkkopalvelun kautta. Tutkimusaineisto sisältää tietoja 34 kunnan kerros- ja rivitaloasunnoista, jotka olivat myynnissä kiinteistövälitysten kautta vuoden 2011 maaliskuussa. Kunta on määritelty tässä tutkimuksessa alueeksi, koska sillä on selkeät maantieteelliset rajat. Aineistosta on tehty kaksi perusjoukkoversiota, joista ensimmäisessä jokainen kunta on oma alueensa (34 aluetta). Toisessa perusjoukkoversiossa osa kunnista on yhdistetty lähinaapuruuden tai muun yhteisen sijainnin mukaan, jolloin osa alueista on eräänlaisia kuntayhtymiä ja osa yksittäisiä kuntia. Tämän jaotuksen tuloksena alueiden lukumäärä supistuu arvoon 14. Alueet ovat samalla ositetun otannan perustana olevia ositteita.

Tutkimusaineiston lähteenä ollut rekisteri sisälsi alun perin tiedot yli 16 000 tilastoyksiköstä (asunnosta), joista tutkimusaineistoon valittiin 34 kunnasta yhteensä 9 815 kpl. Noin sadan asunnon tiedoissa oli ilmiselviä virheitä tai puutteellisuuksia, minkä johdosta ne hylättiin suoraan. Toinen valintaperuste on kuvattu alaluvussa 6.1.2. Perusjoukkoversion 1 (alkuperäiset 34 aluetta) asuntojen määrä vaihtelee välillä 111–833 ja on keskimäärin 289. Alueiden väliset kokoerot eivät ole kovin suuret. Perusjoukkoversion 2 (14 aluetta) tilastoyksiköiden määrä vaihtelee välillä 111–1 333 ja on keskimäärin 701. Tässä versiossa on kaksi selvästi muita pienempää aluetta (alle 150 asuntoa) ja kolme yli 1 000 asuntoa sisältävää aluetta.

6.1.2 Vaste- ja apumuuttuja

Asuntoaineiston vastemuuttujana (y) on asunnon pyydetty myyntihinta (1 000 €) ja apumuuttujana (x) asunnon koko eli pohjapinta-ala (m^2). Tilastoyksiköiden (asuntojen) valinnassa on ollut periaatteena, että käytettyyn apumuuttujaan (pinta-ala) tulee riittävää vaihtelua alueiden välillä, jolloin valitun mallin käytölle on edellytyksiä. Tästä syystä 15 kunnan pienimmät asunnot ja neljän kunnan suurimmat asunnot on jätetty pois. Muiden kuntien myytävät asunnot ovat muka-

na sellaisenaan. Aineisto ei tästä syystä edusta täysin kuntien koko asuntovalikoimaa eikä myöskään pinta-ala- ja hintajakaumaa.

Liitetaulukot A.1–A.4 sisältävät perusjoukkoversioiden 1 ja 2 alueiden vaste- ja apumuuttujan tunnuslukutiedot.

6.1.3 Perusjoukkoversio 1: 34 aluetta

Kaikkien 34 kunnan vastemuuttujan *y* tunnuslukujakauma on liitetaulukossa A.1. Asunnon hinta vaihtelee erittäin voimakkaasti. Halvimman yksittäisen asunnon hinta on 4 900 € ja kalleimman peräti 3 896 170 €. Myös keskiarvot ja hajonnat vaihtelevat suuresti. Suurin keskiarvo (707 000 €) on luonnollisesti Helsingissä. Muun pääkaupunkiseudun keskiarvo on pääasiassa välillä 200 000–300 000 €. Pienimmät aluekeskiarvot ovat hieman yli 60 000 €. Hajonta on suurin Helsingissä (562 000 €) ja Porvoossa (208 000 €), ja pienimmät hajonnat ovat alle 40 000 €. CV-arvot vaihtelevat välillä 0.230–0.918. Vastemuuttujan kokonaismäärä vaihtelee suuresti alueittain (minimi 6 872 550 € ja maksimi 437 901 680 €)

Asunnon koon tunnuslukuyhteenveto ilmenee liitetaulukosta A.2. Vaihteluväli on koko perusjoukossa erittäin suuri eli 17–508.5 m^2. Alueittaisissa vaihteluväleissä on selviä eroja. Kolmen alueen maksimikoko on alle 100 m^2. Aluekeskiarvot ovat välillä 51.4–123.9 m^2. Hajontojen vaihteluväli on 11–58 m^2. CV-arvot vaihtelevat arvojen 0.194 ja 0.635 välillä. Apumuuttujan korrelaatio vastemuuttujan kanssa on 0.674 koko perusjoukossa ja on hyvin erilainen alueittain, sillä se vaihtelee välillä 0.146–0.877.

6.1.4 Perusjoukkoversio 2: 14 aluetta

Perusjoukko versiossa 2 on 14 aluetta. Tämän aluejaon mukainen vastemuuttujan *y* tunnuslukujakauma on liitetaulukossa A.3. Koko aineiston minimi ja maksimi ovat samat kuin 34 paikkakunnan aineistossa (4 900 € ja 3 896 170 €). Suurin keskiarvo (707 000 €) on Helsingissä. Keskiarvo on yli 200 000 € Pirkkalassa, Porvoossa ja pääkaupunkiseudulla. Pienin aluekeskiarvo on Lapissa (111 970 €). Hajonta on suurin Helsingissä (562 000 €) ja Porvoossa (208 000 €), ja pienimmät hajonnat ovat Lapissa ja Oulun seudulla (noin 50 000 €). Pienin CV-arvo on 0.358 (Radanvarsikunnat) ja suurin 0.916 (Porvoo). Vastemuuttujan summa vaihtelee 25 408 770 eurosta (Porvoo) 437 901 680 euroon (Helsinki).

Liitetaulukko A.4 sisältää asunnon koon tunnuslukuyhteenvedon. Vaihteluväli on sama kuin 34 paikkakunnan kohdeperusjoukossa eli 17–508.5 m^2. Alueittaisissa vaihteluväleissä on selviä eroja. Kahden alueen (Lappi ja Oulun seutu) maksimikoko on alle 200 m^2 ja viiden alueen maksimikoko yli 400 m^2. Aluekeskiarvon minimi on 55.2 (Oulun seutu) ja maksimi 123.9 m^2 (Helsinki). Hajontojen vaihteluväli on 16–58 m^2. CV-arvot vaihtelevat arvojen 0.218 (Jyväskylä) ja 0.635 (Porvoo) välillä. Apumuuttujan korrelaatio vastemuuttujan kanssa on koko kohdeperusjoukossa 0.674 ja vaihtelee alueittain voimakkaasti eli on välillä 0.207–0.877.

6.1.5 Alueiden välinen vaihtelu havaintoaineistoissa

Aluekiintiöinnin kannalta olennainen ennakkotieto on se, missä määrin alueiden välistä vaihtelua yleensäkin esiintyy. Sen selvittämiseksi tutkimusaineistosta on tehty kaksi yksisuuntaista varianssianalyysiä, joissa luokittelevana muuttujana on aluejako. Ensimmäisessä analyysissä luokittelu perustuu perusjoukkoversion 1 34 alueeseen ja toisessa perusjoukkoversion 2 14 alueeseen (taulukot 6.1 ja 6.2).

TAULUKKO 6.1. Asunnon velattoman hinnan ja asunnon koon varianssitaulut: 34 aluetta.

Muuttujat	Vaihtelulähde	Neliö-summa (SS)	Vapaus-asteet	Keskineliö-summa (MS)
Vastemuuttuja *y*:	Alueiden välinen	190 397 178	33	5 769 611
Asunnon hinta (€)	Alueiden sisäinen	265 448 625	9 781	27 139
	Yhteensä	455 845 803	9 814	
Apumuuttuja *x*:	Alueiden välinen	3 305 407	33	100 164
Asunnon koko (m^2)	Alueiden sisäinen	6 593 889	9 781	674
	Yhteensä	9 899 296	9 814	

Vastemuuttujan kokonaisvaihtelusta on n. 40 % alueiden välistä ja apumuuttujan kokonaisvaihtelusta n. 33 % alueiden välistä. Aluemallin käytölle on edellytyksiä.

14 paikkakunnan kohdeperusjoukon muuttujien varianssitaulu on seuraavanlainen:

TAULUKKO 6.2. Asunnon velattoman hinnan ja asunnon koon varianssitaulut: 14 aluetta.

Muuttujat	Vaihtelulähde	Neliö-summa (SS)	Vapaus-asteet	Keskineliö-summa (MS)
Vastemuuttuja y:	Alueiden välinen	177 668 769	13	13 666 828
Asunnon hinta (€)	Alueiden sisäinen	278 177 034	9 801	28 383
	Yhteensä	455 845 803	9 814	
Apumuuttuja x:	Alueiden välinen	2 296 482	13	176 652
Asunnon koko (m^2)	Alueiden sisäinen	7 602 814	9 801	776
	Yhteensä	9 899 296	9 814	

Vastemuuttujan kokonaisvaihtelusta on n. 39 % alueiden välistä ja apumuuttujan kokonaisvaihtelusta n. 23 % alueiden välistä. Aluemallin käytölle on edellytyksiä.

Kohdeperusjoukon apumuuttujalle lasketaan vielä vastine ryväsotannasta tutulle käsitteelle sisäkorrelaatio *ICC*, jonka teorian on esitellyt mm. Pahkinen (2012). Peruskaava on tarkoitettu yhtä suurien rypäiden tapaukseen. Jos rypäät ovat erisuuria, käytetään regressioanalyysista tuttua selitysastetta R^2 korvaamaan sisäkorrelaatiota *ICC*, ja muunnoksella saadaan homogeenisuusmitta

$$R_a^2 = 1 - R^2 = 1 - MSW / S^2 , \tag{6.1}$$

missä *MSW* tarkoittaa rypäiden (tässä tapauksessa ositteiden) sisäistä keskineliösummaa ja S^2 muuttujan varianssia. Kun kaavaa (6.1) sovelletaan 34 ja 14 alueen kohdeperusjoukkoon, saadaan apumuuttujan sisäkorrelaatioiden arvoksi seuraavat:

TAULUKKO 6.3. Apumuuttujan homogeenisuusmitta aluejaotuksittain.

Aluejaotus	Homogeenisuusmitta muuttujana apumuuttuja *x* (asunnon koko)
34 aluetta	0.332
14 aluetta	0.231

Homogeenisuusmitta on hyvin lähellä suhdelukua, joka saadaan, kun ryhmien välinen neliösumma jaetaan kokonaisneliösummalla. Laskettuja mitan arvoja on käytetty hyväksi kiintiöinnissä, jossa alueiden otoskoot johdettiin EBLUP-estimointiin liittyvän MSE:n g_{1d} -komponentin pohjalta.

6.1.6 Sijaismuuttujan laskentatekniikka 34 alueelle

Tässä tutkimuksessa käytetään kolmea esiotokseen ja siitä johdettuun sijaismuuttujaan (alaluku 5.2) perustuvaa otoskiintiöintiä. Ensin 34 aluetta järjestetään apumuuttujan *x* CV-arvon (arvot ovat liitetaulukossa A.2) mukaan nousevaan järjestykseen. Kolme systemaattisella otannalla poimittua aluetta ovat Riihimäki (ryhmä 1), Raisio (ryhmä 2) ja Hämeenlinna (ryhmä 3). Näistä poimituista viiden asunnon SRSWOR-otoksista saadaan oikean vastemuuttujan *y* ja apumuuttujan *x* välille seuraavat regressioyhtälöt:

Ryhmä 1	y = -16.396 + 3.288x
Ryhmä 2	y = -8.664 + 2.435x
Ryhmä 3	y = -139.420 + 4.365x

Perusjoukon jokaiselle tilastoyksikölle (asunnolle) on laskettu regressioyhtälön avulla sijaismuuttujan arvo y_{dk}^{*} sen mukaan, mihin alueryhmään tilastoyksikkö kuuluu, lukuun ottamatta esiotokseen kuuluvia tilastoyksiköitä. Sijaismuuttujan aluekohtainen tunnuslukuyhteenveto on liitetaulukossa B.1.

6.1.7 Sijaismuuttujan laskentatekniikka 14 alueelle

Esiotoksen ja sijaismuuttujan käyttöön perustuvaa otoskiintiöintiä sovelletaan 14 alueeseen samojen periaatteiden mukaan kuin alaluvussa 6.1.6 on kuvattu. Ensin alueet järjestetään apumuuttujan *x* CV-arvojen (liitetaulukko A.4) mukaan nousevaan järjestykseen. Kolme systemaattisella otannalla poimittua aluetta ovat Lappi (ryhmä 1), Radanvarsikunnat (ryhmä 2) ja Satakunta-Pohjanmaa (ryhmä 3). Näistä poimituista viiden asunnon SRSWOR-otoksista saadaan oikean vastemuuttujan *y* ja apumuuttujan *x* välille seuraavat regressioyhtälöt:

Ryhmä 1	y = -58.282 + 3.221x
Ryhmä 2	y = 55.696 + 1.853x
Ryhmä 3	y = -148.372 + 4.960x

Jokaiselle perusjoukon tilastoyksikölle paitsi esitokseen kuuluville lasketaan sijaismuuttujan arvot y_{dk}^{*}. Sen aluekohtainen tunnuslukuyhteenveto on liitetaulukossa B.2.

6.2 Aluekiintiöintien otoksista laskettavat laatumittarit

6.2.1 Otoskohtaiset laatumittarit

Otoskiintiöintiä käsittelevä kirjallisuus on keskittynyt aikaisemmin pääasiassa vain keskineliövirheen (MSE) minimointiin, mutta nykyään käytetään yleisesti myös kiintiöinnin laatua mittaavia muita suureita tai tunnuslukuja. Niillä mitataan estimaattorin tarkkuutta ja harhaa. Estimoitava suure on vastemuuttujan *y* kokonaismäärä koko perusjoukossa eli $Y_d = \sum_{k \in U_d} y_{dk}$. Otoksia simuloidaan kiintiöintiä kohti *r* kpl. Aluksi määritellään yksittäiseen otokseen liittyvät laatumittarit.

Alueiden MSE-arvojen keskiarvon (*Average MSE, AMSE*) lauseke otoksessa *i* (*i* = 1,…, *r*) on

$$AMSE_i = 1/D \sum_{d=1}^{D} mse(\hat{Y}_{di,EBLUP}) . \tag{6.2}$$

MSE-keskiarvon lisäksi voidaan käyttää myös alue-ennusteiden vaihtelukertoimien (CV) keskiarvoa: CV-keskiarvo (*Average CV, ACV*) otoksessa *i* on

$$ACV_i\,\% = 100 \times 1/D \sum_{d=1}^{D} (\sqrt{mse(\hat{Y}_{di,EBLUP})} / \hat{Y}_{di,EBLUP}) . \tag{6.3}$$

MSE- ja CV-keskiarvot voidaan laskea myös painotettuina keskiarvoina, jos halutaan korostaa esim. alueiden kokoja N_d (*d* = 1,…, *D*).

Yhden otoksen alue-ennusteista ja todellisista arvoista lasketut tarkkuutta ja harhaa mittaavat laatuluvut (mm. Rao 2003) ovat prosentteina ilmaistavat keskimääräinen suhteellinen ennustevirhe *ARE* (*Average Absolute Relative Error*) ja keskimääräinen suhteellinen harha *ARB* (*Average Absolute Relative Bias*). Otoskohtainen *ARE%* lasketaan seuraavasti:

$$ARE_i\,\% = 100 \times 1/D (\sum_{d=1}^{D} \left| \hat{Y}_{di,EBLUP} - Y_d \right| / Y_d) . \tag{6.4}$$

Otoskohtaisen keskimääräisen harhan *ARB%* lauseke poikkeaa *ARE%*-lausekkeesta vain siinä, että itseisarvomerkit ovat otoskohtaisen summan ympärillä. Arvo on ei-negatiivinen.

$$ARB_i\,\% = 100 \times 1/D \left| \sum_{d=1}^{D} (\hat{Y}_{di,EBLUP} - Y_d) / Y_d \right| . \tag{6.5}$$

Kaavoissa (6.2) - (6.5) muuttuja *D* tarkoittaa alueiden määrää. Kaavasta (6.4) nähdään, että virheen etumerkin vaikutus häviää, mutta kaavan (6.5) perusteella itseisarvomerkkien sisällä oleva summalauseke voi olla positiivinen, negatiivinen tai nolla. Lopputulos on ei-negatiivinen.

6.2.2 Aluekohtaiset laatumittarit

Toistuviin otossimulointeihin liittyvät laatuluvut, jotka lasketaan otoskiintiöinnin mukaan simuloiduista otoksista, voidaan laskea myös aluekohtaisesti mittaamaan tarkkuutta, harhaa ja tehokkuutta (Rao 2003). Olkoon simuloituja otoksia yhdessä aluekiintiöinnissä *r* kappaletta ja alueita *D* kappaletta. Vastemuuttujan *y* alueeseen *d* liittyvän kokonaismäärän Y_d EBLUP-ennuste on $\hat{Y}_{di,EBLUP}$ otoksessa *i* (*i = 1,2,..., r).*

Kiintiöinnin simuloiduista otoksista alueiden yli laskettu keskimääräinen suhteellinen virhe *MARE* prosentteina määritellään seuraavasti:

$$MARE\% = 100 \times 1/D \sum_{d=1}^{D} (1/r \sum_{i=1}^{r} \left| \hat{Y}_{di,EBLUP} - Y_d \right| / Y_d) . \quad (6.6)$$

Ensin lasketaan otoksista aluekohtaiset keskimääräiset suhteelliset virheet (sulkulauseke) ja sitten niiden keskiarvo. *MARE%* on laskentatavasta johtuen aina ei-negatiivinen.

Kiintiöinnin simuloiduista otoksista alueiden yli laskettu keskimääräinen absoluuttinen suhteellinen harha *MARB%* määritellään seuraavasti:

$$MARB\% = 100 \times 1/D \sum_{d=1}^{D} \left| 1/r \sum_{i=1}^{r} (\hat{Y}_{di,EBLUP} - Y_d) / Y_d \right| . \quad (6.7)$$

Ensin lasketaan aluekohtaiset absoluuttiset keskimääräiset harhat (itseisarvolauseke) ja sitten niiden keskiarvo. Myös *MARB%* on laskentatavasta johtuen aina ei-negatiivinen.

Ennustevirheen neliöintiin perustuva laatumittari on mm. Raon (2003) esittelemä *RRMSE* (*Relative Root Mean Square Error*), joka on eri kuin alue-ennusteen keskineliövirhe MSE. Jatkossa siitä käytetään nimeä suhteellinen keskivirhe.

Kaikista otossimuloinneista (*r* kpl) lasketun alueen *d RRMSE*:n lauseke on seuraava:

$$RRMSE_d\,\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} (\hat{Y}_{di,EBLUP} - Y_d)^2} \,/ Y_d . \quad (6.8)$$

Alueen *d* keskimääräisen neliöidyn ennustevirheen neliöjuuri suhteutetaan oikeaan arvoon $\bar{Y}_d$. *RRMSE* on vastine alue-ennusteen vaihtelukertoimelle (CV). Ero on siinä, että MSE:n tilalla on alue-ennusteen keskimääräinen neliöity ennustevirhe, ja nimittäjässä on alueen oikea kokonaismäärä. Kiintiöinnin otossimuloinneista lasketaan alueiden yli keskiarvo *MRRMSE%*:

$$MRRMSE\% = 1/D \sum_{d=1}^{D} RRMSE_d \ \%. \tag{6.9}$$

Kun yhdestä otoksesta saatavat kokonaismäärien alue-ennusteet lasketaan yhteen, saadaan estimaatti perusjoukon kokonaismäärälle *Y*:

$$\hat{Y}_{EBLUP} = \sum_{d=1}^{D} \hat{Y}_{d,EBLUP} \ .$$

Kiintiöinnin otossimuloinneista (*r* kpl) voidaan laskea perusjoukon kokonaismäärälle oma RRMSE-arvo:

$$RRMSE\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} (\hat{Y}_{i,EBLUP} - Y)^2} \ / Y \ , \tag{6.10}$$

missä *i* = otossimuloinnin numero, $\hat{Y}_{i,EBLUP}$ on otossimuloinnista *i* saatava kokonaismäärän estimaatti ja *Y* on perusjoukon kokonaismäärän oikea arvo.

Rao (2003) esittelee tutkimusta, jossa on käytetty myös laatumittaria *EFF* (*Average Relative Efficiency*). Se on prosenttiluku, jonka avulla voidaan verrata käytetyn estimaattorin tehokkuutta suhteessa jälkiositusestimaattoriin. *EFF*-luvun laskennan lähtökohtana on neliöity ennustevirhe. Laatumittaria voidaan soveltaa minkä tahansa kahden estimaattorin tehokkuuden vertailuun. Mittari ei kuitenkaan esiinny yleisesti alan kirjallisuudessa. Tässä tutkimuksessa ei ole käytetty myöskään *ARE*- eikä *MARE*-laatumittareita, vaan *RRMSE*- ja *MRRMSE*-mittareita.

6.3 Optimaalisen aluekiintiöinnin kriteerit

Luotettavien alue-ennusteiden tuottamiseksi asetetaan tehokkaalle aluekiintiöinnille erilaisia tavoitteita tai kriteerejä, joiden samanaikainen saavuttaminen on tuskin mahdollista. Tavoitteeksi onkin järkevää asettaa se, että kaikille alueille saadaan ainakin kohtuullisen hyvät tulokset, ja että myös perusjoukon estimointitulos on hyvä. Aluekiintiöinnin tuottamia tuloksia voidaan arvioida ainakin seuraavien kriteerien pohjalta:

1) Alue-ennusteiden MSE-approksimaatiot ja CV-arvot ja niiden keskiarvot
2) Alueiden suhteellinen ennustevirhe (*ARE* tai *RRMSE*) ja niiden keskiarvo
3) Alueiden suhteellinen harha (*ARB*) ja niiden keskiarvo

4) Estimointimenetelmän suhteellinen tehokkuus verrattuna johonkin toiseen estimointimenetelmään (*EFF*)

5) Vastemuuttujan aluekohtaisten kokonaismäärien luottamusvälien peittoprosentit.

Analyyttinen ratkaisu optimaaliseen kiintiöintiin perustuu yhdestä tai useammasta muuttujasta riippuvan otoksesta saatavan lausekkeen optimointiin, ja tässä tapauksessa muuttujia olisivat eri alueiden otoskoot. MSE- ja CV-keskiarvon sekä keskimääräisen suhteellisen virheen tai harhan minimointi otoskokojen funktiona ei kuitenkaan onnistu monimutkaisten ja osin tuntemattomienkin lausekkeiden vuoksi. Analyyttisen ratkaisun mahdottomuuteen on viitannut esimerkiksi Longford (2006).

On kuitenkin mahdollista kehittää ja kokeilla kiintiöintejä, joiden taustalla on tietynlaisen optimin etsiminen. Aikaisemmissa luvuissa on kuvattu kolmen kiintiöinnin johtaminen. Näiden kiintiöintien suorituskykyä on verrattu viiden muun kiintiöinnin vastaaviin sillä tavalla, että ensin on 34 alueen ja 14 alueen havaintoaineistosta simuloitu satunnaisotoksia kaikkien kahdeksan kiintiöintimenetelmän pohjalta, minkä jälkeen on vertailtu kiintiöinneittäin simuloitujen otosten estimointituloksia ja tässä luvussa esiteltyjen laatumittarien saamia arvoja. Analyysien antamia tuloksia esitellään ja vertaillaan seuraavaksi.

6.4 Simuloidut otokset

Kahdeksan eri kiintiöinnin toimivuuden vertailua varten on simuloitu ositetun otannan mukaisesti sekä 34 alueen että 14 alueen asuntoaineistosta 1500 satunnaisotosta jokaista kiintiöintiperiaatetta kohti. Otoksia on yhteensä 2 × 8 × 1500 = 24 000 kappaletta. Kiintiöintikohtaiset tiedot ovat seuraavia:

1) Tasakiintiöinti ja suhteellinen kiintiöinti (alaluku 4.2.1), joihin otoskokojen laskenta edellyttää vain numerotietoja. Koska otoskoko 170 ei ole tasan jaollinen 14:llä, on Porvoon ja Helsingin otoskokoja lisätty yhdellä tasakiintiöinnissä, koska niiden apumuuttujan CV-arvot ovat korkeimmat.

2) Apumuuttujaan x perustuvat optimaalinen (Neyman) ja alueoptimaalinen potenssikiintiöinti (alaluku 4.2.2), joihin tarvittavat apumuuttujan aluekohtaiset tunnusluvut (hajonta, summa ja CV) ovat liitetaulukoissa A.2 ja A.4.

3) Sijaismuuttujaan y^* perustuva NLP-kiintiöinti (johtaminen alaluvussa 4.2.4).

Lähtötiedot ja otoskoot sisältävät ratkaisut on esitetty liitetaulukoissa J.1–J.4. Alueille ja perusjoukolle asetetaan sellaiset CV-rajat, että minimiotoskooksi tulisi mahdollisimman lähellä arvoa 170 oleva luku. Kun 34 alueen CV-arvojen yläraja on 20.45 % ja perusjoukon vastaava 8 %, saadaan otoskoolle n ratkaisuksi desimaaliarvo 171.86, mutta alueiden pyöristettyjen otoskokojen summaksi tulee 170. Vastaavasti 14 alueella saavutetaan otokoko 169.92, kun asetetut CV-ylärajat ovat 13.25 % alueille ja 6 % perusjoukolle, eli selvästi matalammat kuin 34 alueen aineistossa. Pyöristettyjen otoskokojen summaksi tulee 170.

4) Sijaismuuttujan y^* aluekohtaisiin suhteellisiin variansseihin perustuva kiintiöinti (johtaminen alaluvussa 5.3), josta käytetään jatkossa nimeä ”RV-optimaalinen kiintiöinti”.

Laskennalliset ja lopulliset otoskoot ovat liitetaulukoissa C.1 ja C.2. Kahden alueen laskennallinen otoskoko on pyöristetty ylöspäin lähimpään kokonaislukuun 34 alueen aineistossa, jotta otoskokojen summaksi tulee 170. Vastaavasti 14 alueen aineistossa yhden ison alueen laskennallinen otoskoko on pyöristetty lähinnä pienempään kokonaislukuun summan 170 saavuttamiseksi.

5) MSE-komponenttiin g_{1d} ja apumuuttujaan perustuva kiintiöinti (johtaminen alaluvussa 5.5).

Komponentti g_{1d} merkitään jatkossa ”*g1*” ja kiintiöinnistä käytetään nimeä ”*g1*-kiintiöinti”. Liitetaulukot G.1 ja G.2 sisältävät lausekkeen (5.8) avulla lasketut 34 ja 14 alueen otoskoot, kun varianssikomponenttien suhde (ja sisäkorrelaatio) vaihtelee. Laskelmissa on käytetty sisäkorrelaation paikalla apumuuttujasta saatavaa homogeenisuusmittaa (taulukko 6.3 alaluvussa 6.1.5). Taulukoista voidaan nähdä muun muassa, että kun otoksen sisäkorrelaatio on vähintään sama kuin apumuuttujan homogeenisuusmitta, eivät laskennalliset otoskoot muutu kovinkaan paljon, ja että negatiivisia otoskokoja esiintyy. Otoskoko on muutettu nollaksi kolmelle pienimmälle alueelle 34 alueen aineistossa, koska näiden laskennalliset otoskoot jäävät alle yhden. Muiden alueiden otoskoot on pyöristetty vähintään arvoon 2. Laskennalliset otoskoot on pyöristetty mahdollisuuksien mukaan lähimpään kokonaislukuun. Kaksi 14 alueen aineiston pienintä aluetta saavat otoskoon nolla, koska niiden laskennallinen otoskoko jää negatiiviseksi. Muiden alueiden laskennalliset otoskoot on pyöristetty mahdollisuuksien mukaan lähimpään kokonaislukuun, lukuun ottamatta suurinta aluetta, jonka otoskokoa pienennettiin yhdellä, jotta kokonaisotoskooksi tulee 170.

6) Sijaismuuttujan y^*, apumuuttujan x sekä mallin käyttöön perustuva kiintiöinti (johtaminen kuvattu alaluvussa 5.4), josta käytetään jatkossa nimeä ”Simu-optimaalinen kiintiöinti”.

Sijais- ja apumuuttujan rekisteristä simuloidaan 1 500 SRSWOR-otosta, joissa alueiden otoskoot kiintiöityvät satunnaisesti. Jokaiselle otokselle suoritetaan EBLUP-estimointi, joka tuottaa tarvittavat arvot estimoitaville suureille alueittain. Jokaisesta otoksesta lasketaan lisäksi alueiden MSE-arvojen keskiarvot, jotka tallennetaan otoskokojen ja estimointitulosten ohella yhteen tiedostoon, joka sisältää 1 500 otoksen tiedot.

Otokset lajitellaan nousevaan järjestykseen niiden alueiden MSE-keskiarvojen mukaan. Tämän jälkeen otoksista poimitaan 30 ensimmäistä, joiden MSE-keskiarvo on siis pienin. Näistä otoksista muodostetaan alueittain otoskokojen jakauma. Perusajatuksena on selvittää, millaisilla otoskokojen arvoilla saavutetaan pienimmät MSE-keskiarvot.

Johdettavan kiintiöinnin lähtökohtana ovat alueiden jakaumien mediaanit. Jos mediaanien summa jää alle kokonaisotoskoon (n), lisätään pienimpien alueiden otoskokoja yhdellä siten, että kokonaisotoskoko n tulee täyteen. Jos mediaanien summa ylittää kokonaisotoskoon (n), pienennetään suurimpien alueiden otoskokoja tarpeen mukaan.

Otoskokojen jakaumat 34 alueen otoksista ovat liitekuviossa K.1. ja K.2. Otoskoot vaihtelevat varsin voimakkaasti, mutta jakaumista on kuitenkin nähtävissä, että niiden taso nousee alueiden koon kasvun myötä. Jakaumista johdettava kiintiöinti lähestyy tässä mielessä suhteellista kiintiöintiä, mutta poikkeaa siitä kuitenkin jonkin verran. Otoskokojen mediaanien summa 34 alueen aineistossa on 159, minkä johdosta 11 pienimmän alueen mediaaniperusteista otoskokoa on kasvatettu yhdellä. Vastaavasti 14 alueen otoskokojen mediaanien summa on 166, minkä vuoksi kahden pienimmän alueen mediaanipohjaista otoskokoa on kasvatettu kahdella.

Kiintiöintien lyhenteet ovat tuloksia esittävissä kuvioissa ja taulukoissa seuraavia:

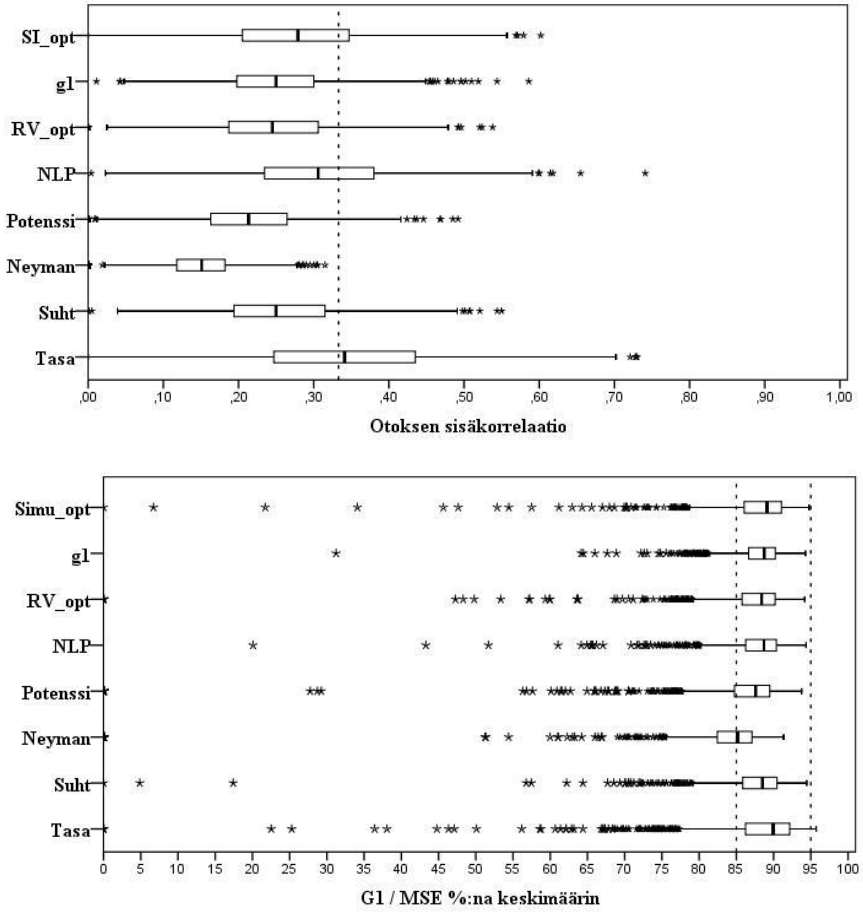
Kiintiöinti	Lyhenne
Tasakiintiöinti	Tasa
Suhteellinen kiintiöinti	Suht
Optimaalinen kiintiöinti (Neyman-)	Neyman
Alueoptimaalinen potenssikiintiöinti	Potenssi
NLP-kiintiöinti	NLP
RV-optimaalinen kiintiöinti	RV_{opt}
g1 -kiintiöinti	*g1*
Simu-optimaalinen kiintiöinti	SI_{opt}

Otoskoot ovat olleet aineistokohtaisesti samat, eli sekä 34 alueen että 14 alueen otoksissa 170 (keskimäärin 5 ja keskimäärin noin 12 aluetta kohti). Otantasuhteeksi (n / N) tulee 1.73 %. Kiintiöintikohtaiset alueiden otoskoot on esitetty liitetaulukoissa D.1 ja D.2.

SAS-ohjelman SURVEYSELECT–proseduurin avulla on poimittu alueilta kiintiöinnin mukaan SRS-otokset. Käytetyt siemenluvut ovat liitetaulukossa E. EBLUP-estimoinnin varianssikomponenttien, regressiokertoimien ja aluevaikutusten estimoinnissa (alaluvut 3.5.2–3.5.4) on käytetty MIXED–proseduuria. SAS-ohjelma on laskenut otoksittain varianssikomponenttien estimaatit, vastemuuttujalle alue-ennusteet ja niiden MSE-approksimaation kaavan (3.23) mukaisesti, MSE-komponentit ja vaihtelukertoimet (CV). Jatkossa MSE-approksimaatiosta käytetään lyhyttä nimeä MSE. Edelleen SAS-ohjelma on tallentanut poimitut otokset yhteen tiedostoon ja estimointien tulokset toiseen. Jälkimmäistä tiedostoa on muokattu SPSS-ohjelman avulla tarvittavien uusien suureiden (sisäkorrelaatio ym.) ja laatumittareiden tuottamiseksi (alaluku 6.2). Otokset ja niistä saadut estimointitulokset on numeroitu, jotta tulokset voidaan liittää oikeisiin otoksiin.

6.5 Sisäkorrelaatio ja *g1* –komponentin prosenttiosuus

Kuten alaluvussa 3.5.5 todetaan, tulisi poimitun otoksen olla sellainen, ettei estimoitu aluevarianssi tulisi nollaksi. Tässä tutkimuksessa esiintyi muutama nollavarianssitapaus sekä 34 että 14 alueen otoskiintiöinneissä. Otoksesta estimoitu alueiden välinen vaihtelu (aluevarianssi) vaikuttaa sisäkorrelaatioon (φ) ja myös siihen, kuinka suureksi MSE:n komponentin *g1* %-osuus MSE:n kokonaisarvosta muodostuu. Kuviossa 6.1 esitetään sisäkorrelaation ja *g1*:n %-osuuden jakaumat otoskiintiöinneittäin 34 alueen simuloinneissa. Sisäkorrelaatiokuvioon on merkitty katkoviiva arvon 0.332 kohdalle. Ko. arvo on apumuuttujan homogeenisuusmitta eli aluevaihtelun osuus kokonaisvaihtelusta.



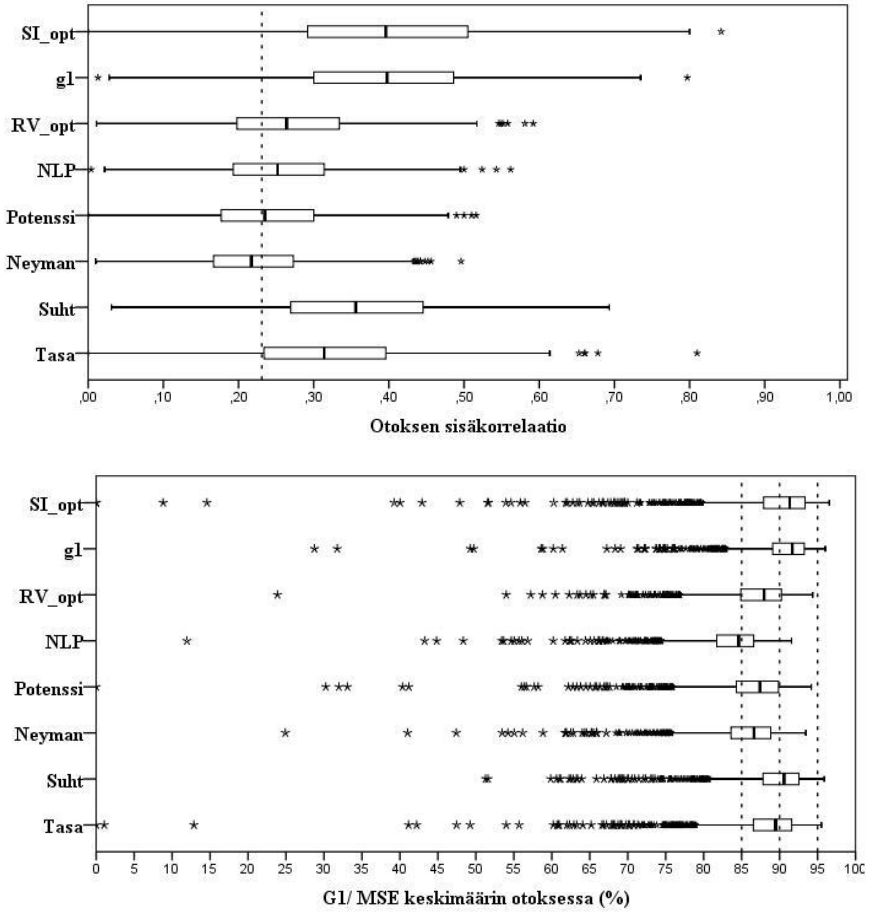
Kuvio 6.1. Sisäkorrelaation ja *g1* -komponentin %-osuuden vaihtelu 34 alueen kiintiöintien otoksissa. Ylemmän kuvion katkoviiva on apumuuttujan homogeenisuusmitan arvon kohdalla.

Neyman-kiintiöinti erottuu molemmissa jakaumissa varsin selvästi muista kiintiöinneistä. Sen saamat arvot ovat alimmalla tasolla. Tasakiintiöinnin ja NLP-kiintiöinnin sisäkorrelaation vaihteluvälit ovat pisimmät. Näiden jakaumien mediaanit ovat lähellä arvoa 0.33, kun taas muiden kiintiöintien mediaanit ovat selvästi alle ko. arvon. MSE-komponentin *g1* %-osuuden jakaumien kvartiilivälit vaihtelevat Neyman-kiintiöintiä lukuun ottamatta välillä 85–92 prosenttia. Nissisen

(2009) toteama suuri %-osuus saa vahvistusta. Kaiken kaikkiaan voidaan sisäkorrelaatiojakaumista todeta, että eri kiintiöintien otoksien koostumuksissa esiintyy suurta vaihtelua.

Kuvio 6.2 sisältää 14 alueen otoksien sisäkorrelaation ja *g1*-komponentin %-osuuden jakaumat. Katkoviiva on arvon 0.231 kohdalla (apumuuttujan homogeenisuusmitta aineistossa).

Malliin perustuvissa kiintiöinneissä (SI_{opt} ja *g1*) sekä tasa- ja suhteellisessa kiintiöinnissä on sisäkorrelaation vaihteluväli huomattavasti suurempi verrattuna parametriperusteisiin kiintiöinteihin, joiden mediaanit ovat lähellä apumuuttujan homogeenisuusmittaa. MSE:n komponentin *g1* prosenttiosuuden jakaumissa on selvä tasoero samojen kiintiöintiryhmien välillä kuin sisäkorrelaation jakaumissa. Näyttää siltä, että aluekohtaisiin ominaisuuksiin nojautuvat parametriperusteiset kiintiöinnit tuottavat otoksiin vähemmän vaihtelua kuin lukumääriin tai malliin perustuvat kiintiöinnit, ainakin jos vertaillaan sisäkorrelaatio- ja *g1*:n prosenttiosuusjakaumia.



Kuvio 6.2. Sisäkorrelaation ja *g1*-komponentin %-osuuden vaihtelu 14 alueen kiintiöintien otoksissa. Ylemmän kuvion katkoviiva on apumuuttujan homogeenisuusmitan arvon kohdalla.

6.6 Estimoinnin tulokset ja laatumittarit kiintiöinneittäin 34 alueen otoksissa

Eri kiintiöintien mukaisista 34 alueen simuloiduista otoksista laskettuja tunnuslukuja, suureita sekä laatumittareita (alaluku 6.2) esitellään ja vertaillaan aluekohtaisesti ja otoskohtaisesti.

6.6.1 Alueiden MSE- ja CV- keskiarvojen analysointi

Taulukkoon 6.4 on koottu kiintiöinneittäin 1500 otoksesta lasketut alueiden (niiden koko on suluissa) MSE-keskiarvojen neliöjuuret sekä CV-keskiarvot. Alimmilla riveillä ovat alueiden keskiarvoista lasketut aritmeettiset ja koolla painotetut kiintiöintikohtaiset keskiarvot.

Taulukko 6.4. Alueiden MSE-keskiarvot sekä alueiden aritmeettiset ja painotetut keskiarvot 34 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi (Kunta)	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Pieksämäki	111	4 167	6 523	8 184	6 313	6 757	6 754	8 921	5 665
Porvoo	112	4 223	6 589	6 807	4 980	2 605	4 312	8 978	5 733
Iisalmi	118	4 434	6 927	8 084	6 705	7 180	6 536	9 459	6 021
Varkaus	139	5 238	8 161	8 932	7 351	4 472	5 915	8 637	7 101
Kirkkonummi	140	5 284	8 231	9 000	7 410	8 527	7 768	8 709	7 162
Raisio	144	5 426	7 678	9 836	7 610	8 760	7 976	8 938	6 750
Pirkkala	148	5 579	7 893	9 495	7 822	6 023	7 076	9 186	7 556
Siilinjärvi	160	6 036	8 539	10 941	9 095	9 737	8 868	9 940	8 173
Salo	161	6 075	8 591	10 331	8 514	5 652	6 865	9 997	7 554
Savonlinna	167	6 303	8 909	10 716	8 833	6 451	7 523	10 369	7 835
Hyvinkää	171	6 464	9 137	10 978	9 053	10 412	9 488	10 630	8 037
Kokkola	173	6 543	9 242	10 495	8 131	5 615	7 035	10 748	9 787
Vihti	177	6 683	9 444	11 357	9 362	10 769	9 811	9 029	9 040
Kaarina	182	6 875	9 712	11 666	9 625	11 075	10 085	11 297	9 299
Kemi	199	7 524	10 629	12 778	10 540	7 020	8 510	11 258	10 176
Mikkeli	215	8 133	9 280	13 775	11 375	11 751	11 025	12 150	10 982
Riihimäki	225	8 506	11 076	14 404	11 897	13 683	12 461	11 732	10 562
Pori	233	8 865	11 532	11 882	9 600	6 693	8 757	12 213	11 002
Kempele	239	9 038	11 765	15 301	13 563	14 533	13 236	12 464	11 220
Nurmijärvi	245	9 277	12 073	14 854	12 187	9 070	9 998	12 789	11 515
Seinäjoki	249	9 423	12 265	15 945	13 174	15 145	12 771	12 992	10 880
Hämeenlinna	255	9 684	12 605	13 511	11 450	10 474	10 952	13 350	13 076
Kouvola	274	10 374	12 593	16 589	14 491	16 662	15 175	13 339	12 872
Lappeenranta	311	11 797	14 317	17 933	15 482	18 923	15 966	14 262	14 631
Rovaniemi	356	13 517	15 445	21 660	18 898	21 687	18 314	15 486	14 658
Espoo	365	13 867	15 837	20 163	18 203	22 229	20 265	15 877	15 034
Lahti	428	16 246	17 575	25 984	22 688	26 050	23 744	17 713	15 831
Kuopio	454	17 246	17 750	23 176	21 419	27 614	21 820	17 965	16 807
Turku	471	17 907	18 437	23 265	21 194	28 673	22 665	18 659	18 360
Jyväskylä	494	18 754	18 457	27 241	24 596	30 035	25 359	18 745	18 279
Vantaa	595	22 624	21 379	32 933	29 706	36 227	28 646	21 009	19 432
Helsinki	621	24 022	21 858	22 495	22 237	23 423	22 192	21 596	22 404
Tampere	650	24 710	22 480	32 128	30 706	39 542	33 405	21 506	21 233
Oulu	833	31 724	26 130	40 076	37 605	45 674	35 969	24 768	25 438
Alueiden keskiarvo *MMSE*		12 886	13 609	18 112	16 188	18 931	16 243	14 067	13 091
Painotettu keskiarvo		17 183	16 739	23 015	21 036	25 262	21 144	16 767	16 159

MSE-keskiarvot ovat oikeiden arvojen neliöjuuria, joten keskiarvojen erot ovat todellisuudessa paljon suurempia. Kun verrataan eri kiintiöintiryhmiä, havaitaan, että parametriperusteisten kiintiöintien keskiarvot ovat selvästi korkeammalla tasolla kahteen muuhun ryhmään verrattuina. Kolmen kiintiöinnin (tasa-, suhteellinen ja *g1*-kiintiöinti) MSE-keskiarvot kasvavat alueen koon kasvun myötä, kun taas muiden kiintiöintien osalta keskiarvot eivät noudata samaa kehitystä, koska niiden otoskoot eivät välttämättä ole verrannollisia alueiden kokoon N_d. Tasakiintiöinnin MSE-keskiarvot ovat alhaisimmat pienimpien alueiden osalta, mutta toisaalta kiintiöinnissä esiintyy yksi hyvin suuri MSE-keskiarvo (neliöjuuri on 31 724) suurimman alueen (Oulu) kohdalla, millä on kasvattava vaikutus mm. alue-ennusteen luottamusväliin ja CV-keskiarvoon. Kolmen kiintiöinnin (suhteellinen, *g1*- ja Simu-optimaalinen kiintiöinti) suurinkin MSE-keskiarvon neliöjuuri on selvästi alle 30 000. Pienin kaikkien alueiden MSE-keskiarvojen keskiarvo on tasakiintiöinnissä, joskin Simu-optimaalisen kiintiöinnin vastaava keskiarvo on vain hieman suurempi. Toisaalta jälkimmäisen kiintiöinnin painotettu MSE-keskiarvo on pienin. Neyman- ja NLP-kiintiöintien keskiarvot ovat selvästi korkeimmat.

Alueiden CV-keskiarvoissa (taulukko 6.5) esiintyy hyvin suurta vaihtelua, ja niiden suuruudella ei ole selvää yhteyttä alueiden kokoihin. Suurin yksittäinen aluekeskiarvo on 75.46 %, ja yli 50 %:n arvot liittyvät pääasiassa pieniin alueisiin, mutta suuremmista alueista on Rovaniemellä ja Oulussa varsin korkea keskiarvon taso otoskoosta riippumatta. Yksittäisen alueen CV-arvojen taso on useimmiten korkea silloin, kun alueen apumuuttuja poikkeaa ominaisuuksiltaan selvästi muista alueista tai perusjoukosta. Poikkeus tästä on Helsinki, jonka CV-arvot ovat joka kiintiöinnissä hyvin alhaiset (vaihteluväli 5.37–6.60 %), vaikka sekin on hyvin poikkeava alue, mutta sillä tavalla, että sen apumuuttujan arvojen taso on huomattavan korkea ja vaihteluväli pitkä. Neljän alueen keskiarvoissa on varsin vähän vaihtelua, joskin niiden taso on Helsinkiä huomattavasti korkeampi. Tasakiintiöinnillä on selvästi alhaisin CV-keskiarvojen taso ja Simu-optimaalisella kiintiöinnillä toiseksi alhaisin. Muiden kiintiöintien keskiarvot ovat selvästi korkeammat.

Taulukko 6.5. Aluekohtaiset CV-keskiarvot prosentteina sekä alueiden aritmeettiset ja painotetut keskiarvot 34 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi (Kunta)	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Pieksämäki	111	45.93	75.46	68.71	61.97	59.74	58.30	59.67	55.64
Porvoo	112	17.30	27.38	27.68	20.14	10.20	17.21	37.02	23.98
Iisalmi	118	27.12	40.08	46.08	39.36	41.38	38.21	48.42	35.47
Varkaus	139	45.39	71.41	60.86	53.96	41.99	50.48	68.85	53.35
Kirkkonummi	140	15.55	23.71	24.93	21.38	25.59	22.62	24.74	20.74
Raisio	144	27.79	35.25	40.39	35.35	40.73	36.59	39.36	32.69
Pirkkala	148	18.70	26.65	32.45	26.60	19.96	23.79	31.24	25.56
Siilinjärvi	160	25.70	38.28	52.23	40.72	41.80	39.11	44.88	36.50
Salo	161	30.99	40.71	46.73	40.33	29.71	34.66	46.19	37.75
Savonlinna	167	27.60	38.66	46.57	37.95	27.65	32.45	44.06	33.37
Hyvinkää	171	17.12	23.50	26.93	23.24	27.23	24.63	27.02	20.88
Kokkola	173	22.09	35.48	33.13	26.59	18.99	23.53	53.60	32.81
Vihti	177	23.12	55.49	43.80	33.83	38.31	36.55	41.42	32.13
Kaarina	182	19.05	25.59	29.16	25.51	30.25	26.98	29.04	24.83
Kemi	199	37.07	52.80	63.14	57.65	34.91	46.57	61.90	49.72
Mikkeli	215	25.72	27.24	44.52	33.17	33.39	31.82	36.92	31.44
Riihimäki	225	22.18	28.89	37.68	31.04	36.04	32.47	30.50	27.56
Pori	233	20.07	26.12	24.82	21.11	15.18	19.39	28.81	25.92
Kempele	239	25.42	32.68	42.16	37.39	39.43	36.39	34.58	31.25
Nurmijärvi	245	21.45	29.00	37.35	29.34	20.59	23.19	30.39	27.38
Seinäjoki	249	23.81	31.56	42.09	34.15	38.47	32.76	33.58	27.77
Hämeenlinna	255	17.23	21.71	22.55	20.01	18.87	19.09	22.87	22.85
Kouvola	274	25.14	28.79	35.05	32.36	37.02	33.97	30.47	29.67
Lappeenranta	311	18.77	21.75	26.05	23.68	29.48	24.35	21.88	22.38
Rovaniemi	356	32.83	40.49	75.00	51.74	55.43	47.93	40.28	37.14
Espoo	365	12.30	13.88	17.67	16.38	21.13	18.71	13.74	13.23
Lahti	428	30.20	33.82	56.27	45.11	49.17	46.06	33.80	30.09
Kuopio	454	20.66	20.96	26.72	24.87	31.75	25.41	21.37	19.91
Turku	471	18.03	17.84	21.44	20.45	28.04	21.93	18.03	17.93
Jyväskylä	494	19.89	19.29	27.28	25.35	31.49	26.25	19.61	19.22
Vantaa	595	23.28	21.65	40.81	33.98	41.52	31.32	20.99	19.33
Helsinki	621	6.60	5.55	5.28	5.58	6.02	5.63	5.37	5.79
Tampere	650	16.06	14.16	19.84	19.49	26.22	21.38	13.59	13.45
Oulu	833	34.19	27.77	51.00	43.89	53.08	40.02	26.17	26.80
Alueiden keskiarvo *MCV%*		23.95	31.58	38.13	32.17	32.37	30.88	33.54	28.37
Painotettu keskiarvo		22.84	26.92	35.78	30.40	32.86	29.43	28.05	24.84

6.6.2 Alueiden suhteellisen keskivirheen ja keskimääräisen suhteellisen harhan analysointi

Taulukko 6.6 sisältää eri kiintiöintien otoksista laskettujen suhteellisen keskivirheen (*RRMSE%*) aluekohtaiset keskiarvot ja keskiarvojen aritmeettiset ja painotetut keskiarvot sekä koko perusjoukon suhteellisen keskivirheen (alaluku 6.2). Eri alueiden välillä esiintyy hyvin selviä eroja. Kahden pienen alueen eli Varkauden ja erityisesti Pieksämäen alue-ennusteiden suhteelliset virheet ovat erittäin suuria ja vaihtelevat väleillä 34–97 % ja 43–122 %. Neljän alueen tasoa voidaan pitää varsin korkeana (aluekohtaiset keskiarvot 18–42 %). Kolmen alueen suhteellisen virheen taso jää lähelle 10 prosenttia. Tasakiintiöinnin suhteellisen virheen tasoa voidaan pitää alhaisimpana ja NLP-kiintiöinnin toiseksi alhaisimpana, kun taas Neyman- ja g1-kiintiöintien vastaava taso on korkein. Viimeisen kiintiöinnin osalta on huomattava, että kolmen alueen otoskoko

on nolla, mutta niistä yhden keskimääräinen suhteellinen virhe on erittäin alhainen. Toisaalta Neyman on perusjoukkotason tehokkain kiintiöinti, koska sen *RRMSE%*-arvo on 4.19 ja tasakiintiöinti tehottomin (arvo 6.96). Muiden kiintiöintien vastaavat luvut ovat välillä 4.67–6.34.

Taulukko 6.6. Alueiden suhteelliset keskivirheet $RRMSE_d$% sekä niiden keskiarvot ja perusjoukon tason suhteellinen keskivirhe *RRMSE%* 34 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi (Kunta)	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Pieksämäki	111	42.88	75.21	90.58	66.51	70.61	75.59	122.01	62.29
Porvoo	112	15.32	15.89	13.35	13.31	8.86	12.10	8.84	16.45
Iisalmi	118	14.59	20.83	21.05	18.10	19.99	18.26	28.33	18.45
Varkaus	139	53.90	94.88	95.26	75.51	33.56	54.01	96.72	78.38
Kirkkonummi	140	12.19	13.77	13.02	11.61	14.52	12.36	14.80	13.57
Raisio	144	26.99	34.48	42.33	31.69	34.59	33.53	40.62	31.16
Pirkkala	148	9.71	10.38	9.40	9.27	8.07	8.68	10.45	10.20
Siilinjärvi	160	13.50	16.50	18.42	15.57	15.33	15.16	17.57	16.36
Salo	161	23.99	35.29	38.86	31.14	18.53	23.67	41.09	30.50
Savonlinna	167	18.78	21.19	18.97	19.25	16.11	18.12	23.00	19.48
Hyvinkää	171	13.09	18.17	21.20	16.08	15.40	16.54	20.95	16.08
Kokkola	173	16.82	23.56	25.25	18.95	12.96	16.94	26.10	23.26
Vihti	177	12.73	15.46	17.05	14.41	14.64	15.67	16.69	14.50
Kaarina	182	12.60	17.21	21.01	15.45	16.31	15.23	19.93	16.14
Kemi	199	27.21	32.26	30.17	30.20	20.63	25.46	34.12	32.42
Mikkeli	215	13.90	13.40	13.74	13.65	13.88	13.59	15.59	14.25
Riihimäki	225	12.26	12.85	11.37	11.66	13.72	11.93	13.04	13.16
Pori	233	24.25	31.96	33.58	26.09	17.77	23.19	34.49	30.38
Kempele	239	9.78	12.49	13.86	12.66	12.98	12.35	13.54	12.10
Nurmijärvi	245	11.28	13.72	15.79	12.90	9.16	10.66	13.00	12.64
Seinäjoki	249	11.90	12.80	12.40	12.66	12.77	11.86	13.60	12.29
Hämeenlinna	255	15.33	20.93	23.55	17.78	15.20	16.39	22.28	21.04
Kouvola	274	23.30	28.15	35.18	29.41	31.32	30.29	28.56	28.13
Lappeenranta	311	15.94	15.53	17.49	14.72	17.23	15.55	15.72	16.32
Rovaniemi	356	15.88	17.76	30.02	21.84	20.55	19.24	18.41	17.28
Espoo	365	13.14	11.14	9.73	12.31	18.85	14.71	9.99	11.64
Lahti	428	16.30	15.61	21.98	18.38	19.02	18.05	14.46	15.51
Kuopio	454	14.47	14.72	17.10	14.98	17.62	16.03	15.03	13.72
Turku	471	20.86	18.80	19.30	18.32	24.25	20.72	18.31	19.10
Jyväskylä	494	12.27	11.31	15.37	12.90	14.56	13.63	12.22	11.28
Vantaa	595	16.11	13.91	27.82	22.93	23.66	19.85	12.64	12.50
Helsinki	621	27.97	19.40	12.62	17.49	21.65	18.72	16.44	21.00
Tampere	650	13.87	9.87	10.41	11.26	16.28	12.69	9.25	10.08
Oulu	833	19.02	15.10	25.11	21.51	22.36	19.24	14.47	15.10
Alueiden keskiarvo *MRRMSE %*		18.30	22.19	24.77	20.90	19.50	20.00	24.48	20.79
Painotettu keskiarvo		17.67	18.71	21.87	19.05	19.25	18.46	19.53	18.03
Perusjoukon *RRMSE %*		6.96	4.93	4.19	4.67	6.34	5.01	4.97	5.16

Taulukko 6.7 sisältää eri kiintiöintien otoksista laskettujen absoluuttisen suhteellisen harhan aluekohtaiset keskiarvot *ARB%* (alaluku 6.2) sekä alueiden aritmeettiset ja painotetut keskiarvot. Keskimääräinen suhteellinen harhan arvoissa on suurta vaihtelua. Varkauden ja Pieksämäen alue-ennusteet ovat voimakkaasti harhaisia, ja vaihteluvälit ovat 23–86 % ja 35–118 %. Myös kahden muun pienen alueen ennusteet ovat selvästi harhaisia. Kolmen suuren alueen (Vantaa, Helsinki ja Oulu) keskimääräiset suhteelliset harhat ovat suuria. Matalia suhteellisen virheen ja harhan tasoja löytyy sekä pieniltä että suurilta alueilta, joten otoskoko ei yksin selitä näiden ar-

voja, vaan alueiden apumuuttujan *x* ominaisuuksilla on oma vaikutuksensa. Pienen otoskoon vuoksi alue-ennusteen regressio-osalla on voimakas vaikutus ennusteen arvoon (lauseke 3.20). Sillä, miten paljon alueet poikkeavat apumuuttujan ominaisuuksien osalta muista alueista, on yhteys suhteelliseen virheeseen ja harhaan, jotka ovat yleensä suuria pienimpien alueiden kohdalla. Toisaalta harha ja virhe voivat olla pieniäkin, vaikka otoskoko on nolla (Porvoo *g1*-kiintiöinnissä). Tasa- ja NLP-kiintiöinnin suhteellisen harhan tasot ovat alhaisimmat, kun taas Neyman-ja *g1*-kiintiöinnin tasot ovat korkeimmat. Jälkimmäistä korkeaa tasoa selittää ainakin osittain se, että kolmen pienimmän alueen otoskoko on nolla. Tästä huolimatta näistä yksi on varsin vähän harhainen.

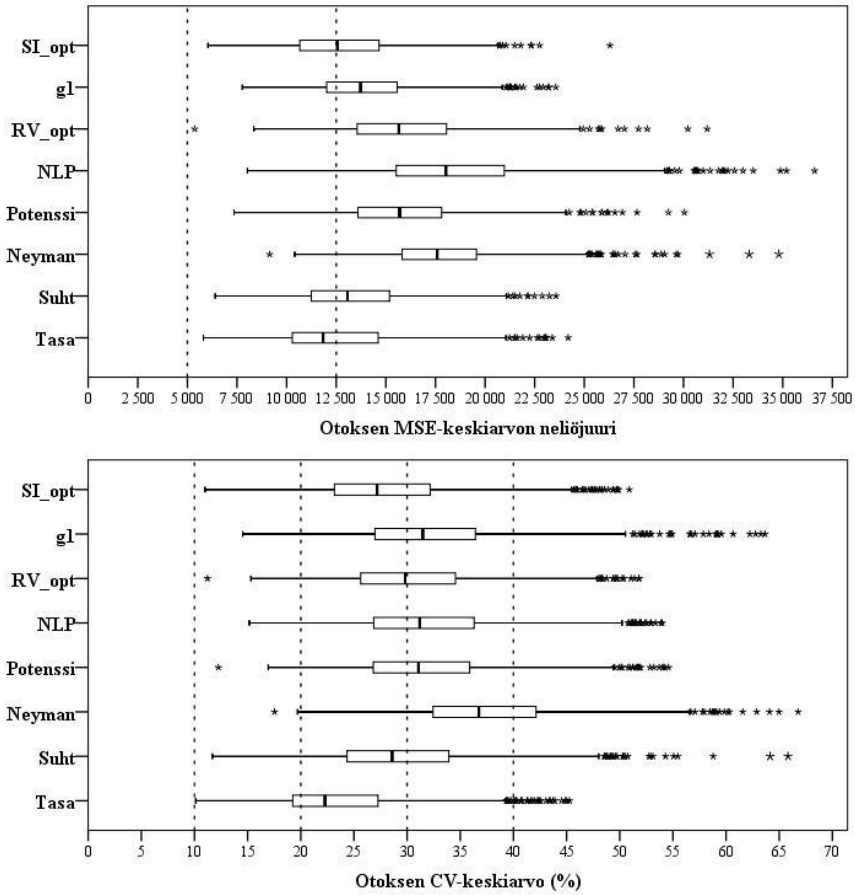
Taulukko 6.7. Keskimääräisen absoluuttisen suhteellisen harhan *ARB%* aluekohtaiset arvot sekä alueiden aritmeettiset ja painotetut keskiarvot 34 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi (Kunta)	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Pieksämäki	111	35.30	68.76	84.22	60.70	66.04	71.15	118.06	55.01
Porvoo	112	4.48	4.52	2.13	2.61	1.83	2.30	6.40	4.98
Iisalmi	118	7.09	14.87	16.35	12.55	13.94	12.76	27.06	11.91
Varkaus	139	42.11	85.22	86.48	65.09	22.72	43.38	85.96	67.16
Kirkkonummi	140	2.19	0.87	5.36	0.54	3.00	0.46	2.89	0.08
Raisio	144	12.96	24.95	37.55	22.81	23.94	24.35	31.45	19.68
Pirkkala	148	3.39	3.30	3.82	4.07	2.46	3.27	3.47	3.80
Siilinjärvi	160	1.31	3.52	9.42	4.52	1.50	3.30	3.64	2.93
Salo	161	15.40	27.34	32.87	24.83	10.50	16.47	33.45	21.89
Savonlinna	167	4.17	7.50	7.27	6.30	4.29	5.18	10.06	6.84
Hyvinkää	171	2.38	8.32	14.99	7.20	4.48	6.41	11.89	5.74
Kokkola	173	6.74	13.95	17.47	10.70	4.43	7.55	17.21	13.67
Vihti	177	2.65	4.95	9.18	5.52	4.49	5.59	5.78	4.02
Kaarina	182	3.64	10.04	16.75	9.32	6.73	8.32	13.58	8.54
Kemi	199	11.98	15.80	14.14	13.42	7.09	10.91	15.82	16.15
Mikkeli	215	1.16	1.01	3.53	2.47	1.24	1.71	1.25	1.32
Riihimäki	225	0.20	1.18	0.75	0.35	0.33	0.53	1.92	0.55
Pori	233	11.33	20.26	23.78	14.92	7.80	12.45	21.97	17.30
Kempele	239	4.70	7.09	8.75	7.93	8.90	8.04	7.81	6.67
Nurmijärvi	245	4.52	6.57	10.42	7.24	3.11	4.52	5.80	6.33
Seinäjoki	249	0.56	0.58	3.31	1.50	0.53	1.23	0.84	0.82
Hämeenlinna	255	5.25	11.75	16.98	9.55	4.81	7.96	13.87	11.43
Kouvola	274	12.53	19.67	30.03	22.39	23.28	23.17	20.20	18.97
Lappeenranta	311	1.40	6.31	11.95	5.95	4.19	6.05	6.75	5.79
Rovaniemi	356	2.67	7.01	20.35	10.59	4.48	7.50	6.94	5.96
Espoo	365	7.59	6.08	5.28	8.13	12.76	10.32	4.53	6.46
Lahti	428	0.21	2.86	11.22	4.88	0.26	2.92	2.41	2.08
Kuopio	454	4.57	6.34	9.98	8.09	9.21	8.08	6.21	5.56
Turku	471	3.38	6.59	11.91	6.91	7.37	7.33	6.85	5.58
Jyväskylä	494	3.20	4.85	10.68	6.88	5.52	6.45	5.31	4.19
Vantaa	595	11.05	9.71	24.57	18.56	19.08	15.58	8.72	8.70
Helsinki	621	16.57	11.05	7.51	10.70	12.40	11.55	9.18	12.37
Tampere	650	1.94	0.80	3.65	0.68	2.51	0.07	0.79	0.21
Oulu	833	3.82	4.78	15.39	9.77	5.24	7.09	4.50	4.00
Alueiden keskiarvo *MARB%*		7.42	12.60	17.30	11.99	9.13	10.70	15.37	10.79
Painotettu keskiarvo		6.43	9.40	14.61	10.16	8.16	8.99	10.51	8.29

Kaikkien edellä tutkittujen tunnuslukujen ja laatumittarien tarkastelun perusteella voidaan tehdä johtopäätös, että aluetasolla tasakiintiöinti on paras ja Simu-optimaalinen kiintiöinti toiseksi paras sekä Neyman-kiintiöinti selvästi huonoin. Toisaalta perusjoukon tason laatumittarin *RRMSE%* perusteella Neyman-kiintiöinti on paras ja tasakiintiöinti huonoin.

6.6.3 Otoskohtaiset laatumittarit

Eri kiintiöintien otoskohtaisten tunnuslukujen ja laatumittarien (alaluku 6.2) jakaumat esitetään graafisesti. Seuraava kuvio sisältää *AMSE*- ja *ACV%*-keskiarvot. Kiintiöintien *ACV%*- jakaumissa on mukana 99 % otoksista muutaman hyvin suuren CV-arvon vuoksi, mutta *AMSE*-jakaumat sisältävät kaikkien otosten arvot. MSE-keskiarvot ovat todellisten keskiarvojen neliöjuuria. Kaiken kaikkiaan otoskohtaisissa keskiarvoissa esiintyy suurta vaihtelua.

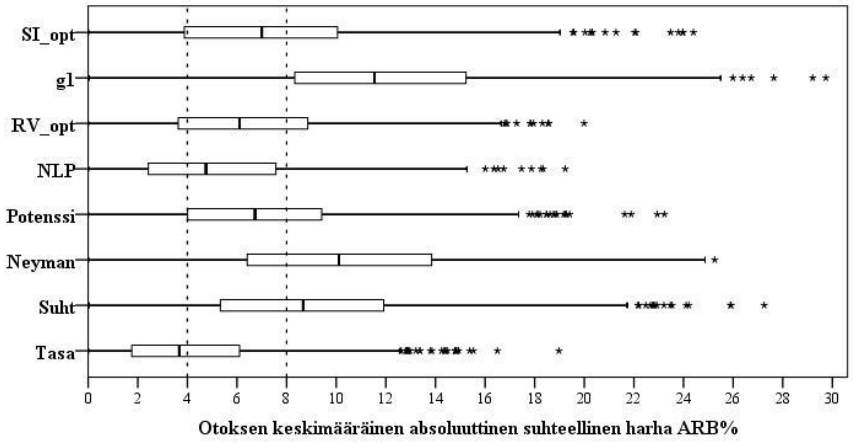


Kuvio 6.3. Eri kiintiöintien otosten *AMSE*- ja *ACV%*-jakaumat: 34 aluetta.

Lukumääriin tai malliin perustuvien kiintiöintien *AMSE*-jakaumat ovat varsin lähellä toisiaan ja selvästi alhaisemmalla tasolla parametriperusteisiin kiintiöinteihin verrattuina. Tasakiintiöinnin *ACV*-jakauma on alimmalla tasolla ja Simu-optimaalisen kiintiöinnin *ACV%*-jakauma toiseksi alimmalla tasolla. Malliperusteisen *g1*-kiintiöinnin *ACV%*-arvot ovat korkealla tasolla, mutta

pitää ottaa huomioon, että otoksista puuttuvat havainnot kolmelta alueelta. Kun *AMSE* ja *ACV%*-jakaumia tarkastellaan yhdessä, voidaan päätellä, että tasakiintiöinti on paras, kun taas Neyman- ja NLP-kiintiöinti ovat huonoimmat.

Kuviossa 6.4 esitetään kiintiöintikohtaisten otosten keskimääräisen absoluuttisen suhteellisen harhan jakaumat. Tasakiintiöinnissä on selvästi alhaisin taso. Sen jakauman yläkvartiili on noin 6 %, kun taas korkeimman harhan tason saaneen g1-kiintiöinnin yläkvartiili on noin 16 %. Myös Neyman- ja suhteellisen kiintiöinnin harhan taso on korkea. Parametriperusteiset kiintiöinnit eivät ryhmänä erotu selvällä tavalla muista kiintiöinneistä. Malliperusteisen *g1*-kiintiöinnin huonointa tulosta selittänee ainakin osittain se, että kolmen alueen otoskoko on nolla.



Kuvio 6.4. Eri kiintiöintien otosten keskimääräisen absoluuttisen suhteellisen harhan *ARB%* jakaumat: 34 aluetta.

Johtopäätös 34 alueen kiintiöintikohtaisten otosten tunnuslukujen ja laatumittarien analyysista on, että tasakiintiöintiä voidaan pitää parhaana. Malliperusteista Simu-optimaalista kiintiöintiä voidaan pitää toiseksi parhaana, joskin kiintiöinnin otoksissa esiintyy korkeahko suhteellisen harhan taso. Malliperusteisen *g1* -kiintiöinnin varsin heikkojen tulosten taustalla on kolmen alueen otoskoko nolla.

6.7 Estimoinnin tulokset ja laatumittarit kiintiöinneittäin 14 alueen otoksissa

Koska alkuperäisten alueiden kokoerot eivät ole erityisen suuret (runsaasti pieniä samankokoisia alueita ja vain muutama suurempi alue), tällä seikalla saattaa olla vaikutuksensa eri kiintiöintien toimivuuteen. Jotta kiintiöintejä voidaan verrata muunkinlaisen aluerakenteen pohjalta, on 34 alueesta muodostettu 14 aluetta yhdistämällä maantieteellisesti lähekkäin olevia alueita (alaluku 6.1), minkä jälkeen näiltä 14 alueelta on simuloitu otokset ja laskettu samat tunnusluvut, muut suureet ja laatumittarit niiden periaatteiden mukaan kuin alaluvussa 6.4 on esitetty. 14 alueen

joukossa on vain kaksi pientä aluetta, ja alueiden kokoerot ovat huomattavasti suuremmat kuin 34 alueen joukossa (kolmen alueen koko on yli 1 000). Analyysien kiintiöintikohtaisia tuloksia ja vertailuja esitellään seuraavaksi.

6.7.1 Alueiden MSE- ja CV- keskiarvojen analysointi

Taulukko 6.8 sisältää alueiden MSE- keskiarvot sekä keskiarvojen aritmeettiset ja painotetut keskiarvot (neliöjuuriarvoina). Tasa- ja suhteellisen kiintiöinnin sekä *g1*-kiintiöinnin kohdalla on havaittavissa, että alueiden MSE-keskiarvot kasvavat alueen koon myötä, kun taas muissa kiintiöinneissä ei näin suoraa yhteyttä ole, koska niiden otoskokoihin vaikuttavat merkittävästi muutkin tekijät kuin pelkästään alueiden koot. Tässäkin on otettava huomioon, että arvot ovat todellisten arvojen neliöjuuria.

Taulukko 6.8. Alueiden MSE-keskiarvot sekä niiden aritmeettiset ja painotetut keskiarvot (todellisten arvojen neliöjuuria) 14 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Porvoo	112	3 780	7 638	7 321	5 227	2 440	4 418	12 163	5 612
Pirkkala	148	5 242	8 895	11 390	8 635	10 569	9 869	16 049	7 409
Etelä-Savo	493	17 923	19 514	27 104	23 510	23 473	20 893	20 253	17 782
Jyväskylä	494	17 949	19 543	28 608	25 864	32 596	25 071	20 282	17 811
Lappi	555	20 205	21 000	32 217	27 730	25 442	22 700	21 658	19 116
Kaakkois-Suomi	585	21 303	22 143	30 675	27 923	32 030	25 834	22 840	20 161
Helsinki	621	22 095	22 821	20 721	20 884	21 526	21 450	23 402	23 919
Satakunta-Pohjanmaa	655	23 918	23 835	26 040	24 037	18 731	20 000	23 431	21 690
Radanvarsikunnat	818	29 869	26 678	35 543	33 734	58 522	39 561	25 321	24 239
Kuopion seutu	871	31 817	27 533	36 706	35 933	62 319	42 139	26 170	25 008
Turun seutu	958	35 022	29 415	38 172	38 338	68 555	46 365	27 240	26 719
Oulun seutu	1 072	39 215	31 158	48 257	45 811	47 588	37 819	28 985	26 291
Pääkaupunkiseutu	1 100	40 249	31 199	40 694	41 600	38 132	33 791	29 089	30 717
Häme-Pirkanmaa	1 333	48 804	34 559	45 301	49 150	95 396	59 117	32 435	31 381
Alueiden keskiarvo *MMSE*		28 383	24 486	32 653	31 657	45 549	32 494	24 085	22 489
Painotettu keskiarvo		33 302	27 452	36 791	36 375	56 676	37 903	26 304	25 192

Taulukon luvuista näkyy sama asia kuin 34 alueen MSE-keskiarvoista. Parametriperusteisten kiintiöintien MSE-keskiarvot ovat huomattavasti korkeammalla tasolla kuin muiden kiintiöintien. Kiintiöinneistä erottuu muista selvästi malliperusteinen Simu-optimaalinen kiintiöinti. Sen MSE-keskiarvot ovat alhaisimmat kahdeksalla alueella eivätkä erityisen korkeita millään kuudella muulla alueellakaan. Lisäksi alueiden aritmeettinen keskiarvo (*MMSE*) ja painotettu keskiarvo ovat selvästi alhaisimmat. Kiintiöintiä voidaan pitää MSE-keskiarvojen osalta selkeästi parhaana. Malliperusteinen *g1*-kiintiöinti ja suhteellinen kiintiöinti ovat seuraavaksi parhaita alueiden keskiarvon perusteella, joskin edellisessä kaksi pienintä aluetta saavat suuren MSE-keskiarvon,

mikä johtuu otoskoosta nolla. Tasakiintiöintiä voidaan pitää kolmeen em. kiintiöintiin verrattuna huomattavasti heikompana. Sen MSE-keskiarvot ovat neljällä suurimmalla alueella erittäin suuret.

Taulukko 6.9 sisältää aluekohtaiset CV- keskiarvot sekä alueiden aritmeettiset ja painotetut CV-keskiarvot. Aluekohtaisissa CV-keskiarvoissa esiintyy suurta vaihtelua, joskaan ei niin suurta kuin 34 alueen CV-arvoissa, koska alueiden otoskoot ovat suuremmat. Niiden suuruudella ei ole selvää yhteyttä alueiden kokoihin. Korkeaa keskiarvon tasoa esiintyy sekä pienillä että suurilla alueilla ja otoskoosta riippumatta. Alueiden keskiarvojen vertailun perusteella voidaan havaita samanlainen, joskaan ei yhtä selvä tasoero parametriperusteisten ja muiden kiintiöintien välillä kuin MSE-keskiarvojen kohdalla. Muista alueista ominaisuuksiltaan poikkeavan Helsingin CV-arvot ovat joka kiintiöinnissä jälleen hyvin alhaiset (5–6 %). Kolmen alueen keskiarvot vaihtelevat vähän, joskin niiden taso on Helsinkiä huomattavasti korkeampi. Malliperusteisen Simu-optimaalisen kiintiöinnin aluekeskiarvojen keskiarvo (*MCV%*) on alhaisin. Tasa- ja suhteellisen kiintiöinnin vastaavat keskiarvot ovat sitä hieman korkeampia. Neyman- ja NLP-kiintiöinnin alueiden keskiarvot ovat selvästi korkeimmat. Kahden pienen alueen korkeaa CV-arvoa *g1*-kiintiöinnissä selittää alueiden otoskoko nolla, mikä aiheuttaa niihin korkeat MSE-arvot.

Taulukko 6.9. Aluekohtaiset CV- keskiarvot prosentteina sekä niiden aritmeettiset ja painotetut keskiarvot 14 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Porvoo	112	14.85	30.93	29.03	20.38	9.42	17.20	46.65	22.45
Pirkkala	148	17.07	29.12	37.11	28.22	33.87	32.15	50.14	24.11
Etelä-Savo	493	26.63	29.47	38.97	34.47	34.00	30.75	30.44	26.42
Jyväskylä	494	18.88	20.49	28.52	26.23	31.79	25.54	21.35	18.76
Lappi	555	32.52	33.90	56.76	47.53	40.91	37.11	34.85	30.56
Kaakkois-Suomi	585	20.30	21.09	27.58	25.60	28.54	23.95	21.85	19.31
Helsinki	621	5.39	5.57	4.94	5.02	5.15	5.17	5.70	5.96
Satakunta-Pohjanmaa	655	20.85	20.99	22.85	20.93	16.45	17.49	20.78	18.94
Radanvarsikunnat	818	19.73	17.56	23.58	22.38	37.38	26.11	16.62	15.90
Kuopion seutu	871	23.56	20.74	26.98	26.42	41.03	30.42	19.69	18.78
Turun seutu	958	19.79	16.81	21.17	21.30	34.69	25.5	15.67	15.29
Oulun seutu	1 072	29.90	23.43	38.47	36.10	36.10	28.86	21.60	19.48
Pääkaupunkiseutu	1 100	15.70	11.89	15.84	16.24	14.65	12.97	11.03	11.71
Häme-Pirkanmaa	1 333	18.03	12.78	16.66	18.10	33.29	21.96	12.00	11.60
Alueiden keskiarvo *MCV%*		20.23	21.06	27.75	24.92	28.38	23.94	23.45	18.52
Painotettu keskiarvo		20.64	18.80	25.64	24.15	29.61	23.48	18.80	16.91

6.7.2 Alueiden suhteellisen keskivirheen ja keskimääräisen suhteellisen harhan analysointi

Taulukko 6.10 sisältää eri kiintiöintien otoksista lasketut aluekohtaiset suhteelliset keskivirheet (*RRMSE$_d$%)* sekä alueiden keskiarvot ja perusjoukon *RRMSE%*-arvot kiintiöinneittäin. Laatumittarin arvot ovat huomattavasti pienempiä verrattuina 34 alueen kiintiöintien vastaaviin. Valtaosa niistä on alle 20 %, ja alle 10 %:n arvoja on viidellä alueella.

Taulukko 6.10. Alueiden suhteelliset keskivirheet *RRMSE$_d$*% sekä niiden keskiarvot ja perusjoukon tason suhteellinen keskivirhe *RRMSE%* 14 alueen kiintiöintien otoksissa.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Porvoo	112	11.12	20.25	15.18	12.72	6.99	12.00	7.03	17.66
Pirkkala	148	7.37	11.15	9.81	9.33	10.07	10.06	5.52	10.29
Etelä-Savo	493	15.73	17.87	19.98	18.40	18.32	16.93	18.79	17.06
Jyväskylä	494	10.86	12.23	17.17	15.16	19.72	14.60	12.50	11.42
Lappi	555	16.12	17.43	21.95	20.64	16.60	17.46	17.96	16.85
Kaakkois-Suomi	585	15.16	16.28	20.58	18.99	22.28	17.58	16.22	15.65
Helsinki	621	16.84	18.42	11.94	13.01	13.05	14.40	18.93	22.06
Satakunta-Pohjanmaa	655	17.42	17.80	19.03	17.25	13.52	14.58	17.34	15.82
Radanvarsikunnat	818	10.84	10.21	11.58	11.22	14.84	12.62	9.82	9.36
Kuopion seutu	871	15.41	14.04	16.60	16.61	28.64	19.71	13.39	12.98
Turun seutu	958	16.43	14.97	16.88	17.91	30.41	20.77	14.13	14.36
Oulun seutu	1 072	14.26	12.20	16.65	16.12	14.56	13.31	11.00	10.65
Pääkaupunkiseutu	1 100	9.65	7.78	8.18	8.66	7.36	7.33	7.33	8.43
Häme-Pirkanmaa	1 333	11.84	9.15	9.78	10.38	18.30	13.65	8.62	8.37
Alueiden keskiarvo *MRRMSE%*		13.50	14.27	15.38	14.74	16.76	14.64	12.76	13.64
Painotettu keskiarvo		13.74	13.23	14.83	14.54	17.79	14.75	12.65	12.70
Perusjoukon *RRMSE%*		4.82	5.01	4.73	4.62	7.91	5.32	5.05	5.40

Näistäkin luvuista voidaan havaita, että parametriperusteisten kiintiöintien suhteellisen virheen taso on korkeampi kuin muiden kiintiöintien vastaava taso, joskin ero ei ole erityisen suuri. Muista alueista poikkeavalla Helsingillä on varsin korkea suhteellisen virheen taso lähes kauttaaltaan. Yhdellä pienellä, keskisuurella ja suurella alueella on vakaa matala suhteellisen virheen taso. Vielä on huomattava, että *g1*-kiintiöinnissä kahden pienen alueen suhteellinen virhe on alhainen, vaikka niiden otoskoko on nolla. Lisäksi kiintiöinnillä on alhaisin alueiden keskiarvo. Muilta alueilta tuleva otosinformaatio vastaa hyvin alueiden ominaisuuksia. Koko perusjoukon tasolla potenssikiintiöinti on tehokkain. Sen *RRMSE%*-arvo on 4.62. Neyman-kiintiöinti on toiseksi tehokkain (4.73) ja NLP-kiintiöinti varsin selvästi tehottomin (7.91).

Taulukko 6.11 sisältää kiintiöintien otoksista lasketut aluekohtaiset absoluuttisen suhteellisen harhan keskiarvot *ARB%* , alueiden aritmeettiset keskiarvot *MARB%* ja painotetut keskiarvot.

Taulukko 6.11. Keskimääräisen absoluuttisen suhteellisen harhan *ARB%* aluekohtaiset arvot sekä alueiden aritmeettiset ja painotetut keskiarvot 14 alueen kiintiöintien otoksissa.

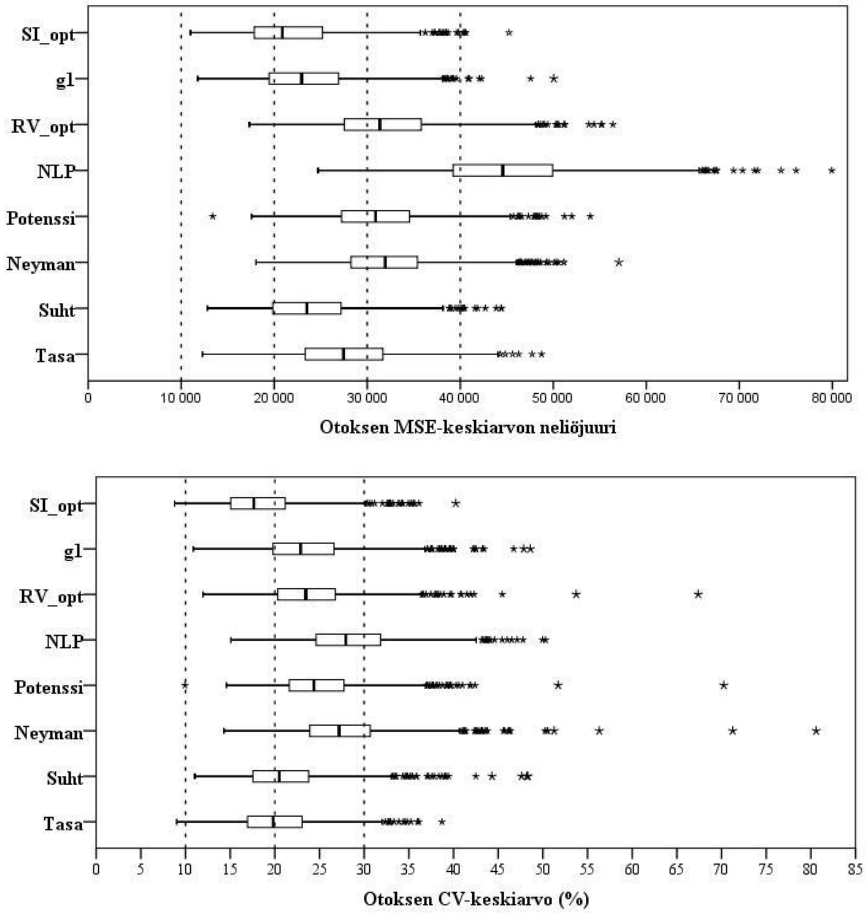
Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Porvoo	112	0.67	0.65	0.80	1.07	0.58	0.70	2.74	1.14
Pirkkala	148	0.17	0.21	0.61	0.32	2.22	0.51	0.31	0.41
Etelä-Savo	493	4.31	4.15	9.20	6.88	7.86	5.65	4.31	4.51
Jyväskylä	494	4.24	4.70	11.45	9.31	14.04	8.40	4.39	3.75
Lappi	555	1.24	1.63	1.65	0.30	2.10	1.13	2.00	1.97
Kaakkois-Suomi	585	5.98	6.15	13.54	11.10	14.53	9.40	5.73	5.07
Helsinki	621	7.73	7.79	5.35	6.26	5.93	6.57	7.46	9.20
Satakunta-Pohjanmaa	655	7.13	6.77	8.37	7.75	5.21	5.89	5.84	5.92
Radanvarsikunnat	818	1.11	0.94	1.29	1.23	5.67	1.82	0.90	0.74
Kuopion seutu	871	6.19	4.04	7.87	7.87	21.43	10.11	3.77	3.74
Turun seutu	958	6.12	4.19	8.43	8.51	21.46	10.36	3.26	3.72
Oulun seutu	1 072	1.12	0.67	4.03	3.27	0.27	1.32	0.40	0.09
Pääkaupunkiseutu	1 100	3.68	2.11	3.32	3.49	2.23	2.34	1.87	2.34
Häme-Pirkanmaa	1 333	2.50	1.37	2.89	2.86	9.93	3.55	1.09	1.00
Alueiden keskiarvo *MARB%*		3.73	3.24	5.63	5.02	8.10	4.84	3.15	3.11
Painotettu keskiarvo		3.94	3.20	5.69	5.19	8.95	5.10	2.92	2.99

Näistäkin luvuista voidaan havaita, että parametriperusteisten kiintiöintien suhteellisen harhan taso on korkeampi kuin muiden kiintiöintien vastaava taso, ja ero on selvempi kuin suhteellisen virheen kohdalla. Alueittainen keskimääräinen suhteellinen harha vaihtelee lähes nollasta prosentista yli 21 prosenttiin. Lähes poikkeuksetta on nähtävissä, että alueilla, joissa on voimakas suhteellinen virhe (taulukko 6.10), on myös voimakas suhteellinen harha, ja toisaalta on alueita, joissa sekä suhteellinen virhe että harha ovat matalia. Edelleen havaitaan, että kaksi pienintä aluetta, joiden suhteellinen virhe on pieni *g1*-kiintiöinnissä, ovat samassa kiintiöinnissä vain lievästi harhaisia, vaikka niiden otoskoko on nolla. Helsingin ennusteiden suhteelliset harhat ovat kaikissa kiintiöinneissä varsin suuria. Kahden ison alueen ennusteiden harhat ovat kauttaaltaan alhaisia. Apumuuttujan alueominaisuuksilla on yhteys suhteelliseen virheeseen ja harhaan, kuten 34 alueen kiintiöinneissä. NLP-kiintiöinnillä on selvästi korkein suhteellisen harhan taso. Malliperusteisella $Simu_{opt}$ -kiintiöinnillä ja *g1*-kiintiöinnillä on alhaisimmat alueiden keskiarvot, mutta ero muihin ei-parametriperusteisiin kiintiöinteihin on pieni.

Malliperusteista SI_{opt} -kiintiöintiä voidaan tunnuslukujen ja laatumittarien alueellisten keskiarvojen perusteella pitää kokonaisuutena parhaana, koska sen aluekohtaiset keskiarvot ja niiden keskiarvot ovat alhaisimmat tai lähes alhaisimmat muihin kiintiöinteihin verrattuina. Myös *g1*-kiintiöintiä voidaan pitää hyvänä.

6.7.3 Otoskohtaiset laatumittarit

14 alueen aineiston otoksista on laskettu otoskohtaiset laatumittarit samalla tavalla kuin 34 alueen kohdalla (alaluku 6.6.3). Kuvio 6.5 sisältää *AMSE*- ja *ACV*-keskiarvot. Kiintiöintien *ACV*-jakaumat sisältävät kaikki otokset (ei ollut hyvin suuria otosten CV-keskiarvoja). Jakaumista voidaan havaita suuri vaihtelu, kuten 34 alueen otoksissa. MSE-keskiarvoista täytyy edelleen ottaa huomioon, että ne ovat todellisten keskiarvojen neliöjuuria. Erot ovat suurempia kuin kuviosta voi päätellä.

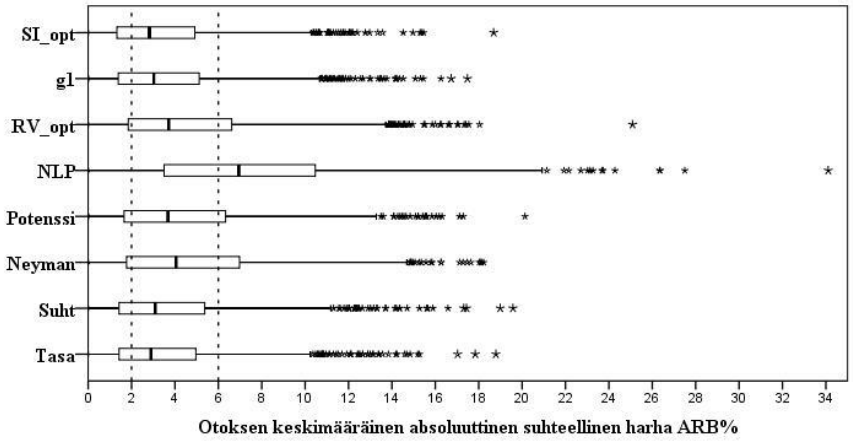


Kuvio 6.5. Eri kiintiöintien otosten *AMSE*- ja *ACV%*-jakaumat: 14 aluetta.

Parametriperusteisten kiintiöintien MSE-keskiarvojen jakaumat ovat selvästi korkemmalla tasolla muihin kiintiöinteihin verrattuina. Jälkimmäisiin kuuluvista on malliperusteisen Simu-optimaalisen kiintiöinnin MSE-jakauman taso alhaisin ja vaihteluväli lyhin. NLP-kiintiöinnin MSE-jakaumien tasot ovat selvästi korkeimmat.

Simu-optimaalisen kiintiöinnin CV-jakauman taso on alhaisin. Tasa- ja suhteellinen kiintiöinti ovat hieman korkeammalla tasolla. NLP- ja Neyman-kiintiöinti ovat huonoimpia, ja *g1*-kiintiöinnin tasoon vaikuttaa nostavasti se, että kahden alueen otoskoko on nolla.

Kuviossa 6.6 esitetään kiintiöintikohtaisten otosten keskimääräisen absoluuttisen suhteellisen harhan jakaumat. Ei-parametriperusteisten kiintiöintien harhan tasot ovat alhaisimmat ja lähellä toisiaan. Niiden jakaumien yläkvartiilit ovat noin 5 %, kun taas parametriperusteisten kinntiöintien harhan jakaumien yläkvartiilit ovat yli 6 %. NLP-kiintiöinnin harhan taso on selvästi korkein. Malliperusteisen *g1*-kiintiöinnin osalta on vielä mainittava kahden alueen otoskoko nolla.



Kuvio 6.6. Eri kiintiöintien otosten keskimääräisen absoluuttisen suhteellisen harhan *ARB%* jakaumat: 14 aluetta.

14 alueen otoksista laskettujen tunnuslukujen ja laatumittarien analyysin perusteella voidaan malliperusteista Simu_{opt} -kiintiöintiä ja *g1*-kiintiöintiä pitää parhaiten toimivina kiintiöinteinä, kun otetaan huomioon myös perusjoukon taso. Myös suhteellinen kiintiöinti toimii hyvin.

6.8 Alue-ennusteiden 95 %:n luottamusvälit ja niiden peitto

Lopuksi tarkastellaan alueiden oikeiden kokonaismäärien osumista EBLUP-ennusteista laskettujen luottamusvälien sisälle (lauseke 3.24). Taulukko 6.12 sisältää kiintiöinneittäin 34 alueen 95 %:n luottamusvälien peittoprosentit. Helsinkiä lukuun ottamatta muiden alueiden peittoprosenteista vain kymmenkunta jää alle 95 %:n, ja loput ovat yli 95 %. Muutamien alueiden jopa 100 %:n peitto selittynee joko sillä, että suurista MSE-arvoista seuraavat pitkät luottamusvälit tai sillä, että alueiden normaalisuus ei toteudu.

Taulukko 6.12. Vastemuuttujan oikeiden kokonaismäärien osuminen 95 %:n luottamusväleille 34 alueen kiintiöinneissä. Prosenttiluvut on laskettu kiintiöintien otoksista.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi (Kunta)	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Pieksämäki	111	98.80	99.07	100.00	99.87	99.87	99.67	99.27	98.93
Porvoo	112	94.67	99.60	99.80	98.67	95.80	98.67	100.00	97.93
Iisalmi	118	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Varkaus	139	95.73	96.13	99.00	98.00	99.33	98.73	98.07	96.13
Kirkkonummi	140	96.07	99.47	99.87	99.67	98.80	99.27	99.53	98.73
Raisio	144	96.80	98.87	99.93	99.33	99.87	99.33	99.67	98.00
Pirkkala	148	99.60	99.87	99.93	100.00	99.93	99.93	100.00	99.93
Siilinjärvi	160	99.73	99.80	99.93	100.00	100.00	99.87	100.00	99.87
Salo	161	99.40	99.60	100.00	99.93	99.87	99.93	99.93	99.47
Savonlinna	167	98.87	99.93	100.00	100.00	99.73	99.80	99.93	99.80
Hyvinkää	171	98.47	99.20	99.73	99.53	99.80	99.87	99.47	99.00
Kokkola	173	99.27	99.40	99.93	99.87	99.73	99.67	99.60	99.40
Vihti	177	99.67	99.53	99.73	100.00	99.87	99.67	99.73	99.80
Kaarina	182	99.47	99.87	99.93	99.93	99.93	99.93	100.00	99.87
Kemi	199	99.47	99.93	100.00	100.00	99.80	99.93	99.93	99.93
Mikkeli	215	99.00	99.73	99.80	99.73	99.67	99.80	99.60	99.73
Riihimäki	225	99.87	99.93	100.00	100.00	100.00	100.00	99.93	99.93
Pori	233	92.33	94.47	92.67	93.13	93.27	93.87	94.40	93.53
Kempele	239	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Nurmijärvi	245	99.13	99.73	100.00	99.80	99.93	99.87	99.87	99.93
Seinäjoki	249	99.87	100.00	100.00	99.93	100.00	100.00	100.00	99.93
Hämeenlinna	255	98.33	98.00	97.87	98.47	98.47	98.73	98.27	98.47
Kouvola	274	97.27	98.13	99.67	99.20	99.53	99.47	99.20	98.27
Lappeenranta	311	96.87	99.20	99.93	99.53	99.40	99.47	99.20	98.33
Rovaniemi	356	99.53	99.93	99.67	100.00	99.93	99.87	99.80	99.87
Espoo	365	86.73	95.80	99.93	96.53	92.60	96.33	97.73	94.40
Lahti	428	99.73	99.93	99.93	100.00	100.00	99.93	99.93	99.67
Kuopio	454	99.27	99.47	99.87	99.87	99.93	99.87	99.67	99.33
Turku	471	88.80	92.93	97.93	96.60	96.93	95.60	95.33	99.20
Jyväskylä	494	99.40	99.93	100.00	99.93	100.00	99.93	99.87	99.73
Vantaa	595	97.87	99.27	97.53	97.47	98.33	98.53	99.87	99.20
Helsinki	621	24.60	35.87	53.20	39.80	33.40	38.07	42.00	34.33
Tampere	650	95.27	98.53	100.00	99.73	98.67	99.00	98.53	98.07
Oulu	833	99.33	99.80	99.00	99.80	99.47	99.87	99.33	99.40
Alueiden keskiarvo		95.57	97.09	98.08	97.48	97.11	97.43	97.58	97.00

Helsinki on muista alueista selvä poikkeus, sillä sen peittoprosentit vaihtelevat arvosta 24.6 % (tasakiintiöinti, pienin otoskoko) arvoon 53.2 % (Neyman-kiintiöinti, suurin otoskoko). Nämä prosenttiluvut ovat selvästi yhteydessä Helsingin kiintiöintikohtaisiin otoskokoihin, ja nähtävästi ainut keino saada myös tämän paikkakunnan peittoprosentti muiden alueiden tasolle on korottaa sen otoskoko moninkertaiseksi, mutta tämä johtaisi todennäköisesti muiden 33 alueen ennusteiden luotettavuuden heikkenemiseen. Helsinki poikkeaa ominaisuuksiltaan selkeästi muista alueista (sekä vaste- että apumuuttujan osalta), mikä vaikeuttaa tehokasta otoskiintiöintiä. Alueen keskimääräinen suhteellinen virhe ja harha ovat lähes joka kiintiöinnissä huomattavan korkeita, mutta toisaalta MSE-keskiarvot eivät ole erityisen suuria, ja CV-keskiarvot ovat jopa huomattavan alhaisia (taulukot 6.4 ja 6.5).

Taulukko 6.13 sisältää 14 alueen kiintiöintikohtaiset peittoprosentit, joista havaitaan sama säännönmukaisuus kuin 34 alueen kiintiöinneistä: Helsingin prosenttiluvut jäävät huomattavan alhaiselle tasolle (34–54 %) verrattuina muihin alueisiin, joiden vastaavat luvut ovat välillä 96–100 %. Kuten 34 alueen osalta, näkyy myös 14 alueen tapauksessa se, että verrattuina muihin alueisiin Helsingin MSE- keskiarvot eivät ole erityisen korkeita missään kiintöinnissä, ja toisaalta CV-keskiarvot ovat jopa yllättävän pieniä, kun taas keskimääräinen suhteellinen virhe ja harha ovat huomattavan suuria kiintiöinnistä riippumatta (taulukot 6.8–6.11).

Taulukko 6.13. Vastemuuttujan oikeiden kokonaismäärien osuminen 95 %:n luottamusväleille 14 alueen kiintiöinneissä. Prosenttiluvut on laskettu kiintiöintien otoksista.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Porvoo	112	98.20	98.80	99.80	99.73	98.73	99.40	100.00	97.13
Pirkkala	148	99.80	100.00	100.00	99.93	100.00	100.00	100.00	99.93
Etelä-Savo	493	99.67	99.73	100.00	99.93	99.93	100.00	99.73	99.73
Jyväskylä	494	99.67	100.00	99.93	99.87	99.80	99.93	99.87	99.80
Lappi	555	100.00	99.93	99.93	99.93	99.93	99.87	100.00	99.87
Kaakkois-Suomi	585	98.53	98.60	99.73	99.33	99.07	99.20	98.60	98.00
Helsinki	621	40.73	39.67	53.80	49.20	54.33	45.47	41.13	33.60
Satakunta-Pohjanmaa	655	99.07	98.87	99.47	99.53	99.13	99.13	98.87	99.00
Radanvarsikunnat	818	99.73	99.60	100.00	100.00	100.00	99.87	99.67	99.60
Kuopion seutu	871	99.33	99.20	99.73	99.73	99.87	99.80	99.67	99.13
Turun seutu	958	97.93	96.27	98.93	98.00	98.53	98.53	96.67	95.87
Oulun seutu	1 072	100.00	99.80	99.87	99.93	99.93	99.80	99.93	99.73
Pääkaupunkiseutu	1 100	99.13	99.07	99.73	99.53	99.87	99.53	98.87	97.87
Häme-Pirkanmaa	1 333	99.47	98.33	99.87	99.87	99.80	99.67	98.80	98.80
Alueiden keskiarvo		95.09	94.85	96.49	96.04	96.35	95.73	95.13	94.15

Korkea peittoprosentti ei kuitenkaan ole välttämättä merkki kiintiöinnin paremmasta tarkkuudesta. Mitä suurempia ovat alueiden MSE-arvot, sitä pitempiä ovat myös luottamusvälit ja sitä todennäköisemmin luottamusvälit sisältävät vastemuuttujan oikeat arvot. Neyman-kiintiöinnissä on saatu Helsingille selvästi korkeimmat peittoprosentit, mutta Helsingillä on myös suurimmat otoskoot samassa kiintiöinnissä.

Luottamusvälien peittoprosenttien tarkastelun perusteella voidaan päätellä, että jos yksittäisen alueen ominaisuudet poikkeavat huomattavasti muiden alueiden vastaavista, kuten Helsingin tapauksessa, on hyvin todennäköistä, että alueen luottamusvälin peittoprosentti jää alhaiseksi, kiintiöinnistä riippumatta. Myös muu alue-estimoinnin luotettavuus heikkenee samalla. Alueen tuloksia voidaan parantaa sen otoskokoa lisäämällä, mutta suurena riskinä on muiden alueiden estimointitulosten heikkeneminen.

7 JOHTOPÄÄTÖKSET

Tässä tutkimuksessa tarkasteltiin aluetilastojen laskentaa ja erityisesti sitä, miten tilastojen laatuun voidaan vaikuttaa suunnittelemalla otoskiintiöinti aluetilastojen näkökulmasta. Tehdystä kirjallisuuskatsauksesta ilmeni, että metodista kehitystyötä on tehty erityisesti estimointiongelmien ratkaisemiseksi. Syynä on se, että otanta-asetelmin kerätyissä aineistoissa otoskoot muotoutuvat alueittain satunnaisesti, jolloin mukana voi olla alueita ilman havaintoja. Niihin estimaatit ovat laskettavissa vain joko malliavusteisesti tai -perusteisesti lainaamalla tietoja muilta havaintoja sisältäviltä alueilta. Tämä oli perussyy, miksi tutkimuksen luvussa 3 on yksinomaan tarkasteltu aluetilastojen laskennassa käytettyjä mallitusratkaisuja, ja sen jälkeen on pohdittu, miten mallitus olisi lisättävissä aluekiintiöintiin.

Otoksen satunnaisesta kiintiöitymisestä aluetasolla päästään eroon tilanteessa, jossa otanta-asetelman suunnitteluvaiheessa tiedetään sovellettava aluejako. Kun tässä tilanteessa alue määrittellään otanta-asetelman ositteeksi, ongelma muuttuu alueoptimaaliseksi kiintiöinniksi. Aluekiintiöinti sinällään on monitahoinen ongelma, josta osoituksena ovat kirjallisuudesta poimitut useat ehdotukset. Niissä optimointikriteerit ovat erilaisia. Eräissä sellaisia ei ole lainkaan, toisissa ne on asetettu aluetasolle, toisissa perusjoukkotasolle ja eräissä samanaikaisesti sekä perusjoukko- että aluetasolle. Yhteistä niille kaikille on kuitenkin se, että niissä ei ole estimointia tehostavaa mallia lainkaan mukana. Edellisistä poiketen tässä työssä on kehitetty kolme aluekiintiöintiä, joista yksi perustuu perusjoukosta poimittavaan esiotokseen ja regressiomalliin ja kaksi perustuvat varsinaisessa estimointivaiheessa käyttöön otettavaan aluemalliin ja siihen soveltuvaan estimointimenetelmään.

Aluekiintiöinti edellyttää yleensä esitietoja perusjoukosta. Parhaimmillaan ne voisivat olla vastemuuttujan aluetason tiedot kuten keskihajonta ja vaihtelukerroin. Näin optimaalinen tilanne ei kuitenkaan ole, koska vastemuuttuja on juuri se muuttuja, jonka tiedot vasta kerätään otoksella. Tällöin on turvauduttava vastemuuttujan korvaavaan sijaismuuttujaan, joksi tässä tutkimuksessa valittiin malliperusteisesti pienestä esiotoksesta laskettu regressioennuste. Toinen aluekiintiöintiratkaisun mahdollistava esitieto on apumuuttuja, joka rekisteritiedon tapaan kattaa perusjoukon kaikki tilastoyksiköt. Kiintiöintiratkaisujen toimivuus testattiin otossimulointien avulla. Simulointi on hyvin yleinen menetelmä tilastomenetelmien kehitystyössä varsinkin uusien ratkaisujen tapauksissa. Koealustana tässä työssä on reaalinen perusjoukko, joka on osa maamme asuntomyyntirekisteriä vuodelta 2011. Asunnoista koostuva perusjoukko on jaettu 34 alueeseen, joiden välinen vaihtelu on kohtalaisen voimakasta. Tästä perusjoukosta on tehty toinen versio siten, että

naapurialueiden yhdistelyn jälkeen alueiden lukumääräksi on tullut vain 14. Alueiden väliset kokoerot ovat suuremmat kuin 34 alueen tapauksessa, ja vastemuuttujan alueiden välisen vaihtelun osuus kokonaisvaihtelusta on vähentynyt. Näin on saatu sopiva vertailuperusjoukko, jossa aluekiintiöiden toimivuuksia vertaillaan eri perusjoukkojen tapauksessa. Tällä ratkaisulla oli tarkoitus kokeilla, missä määrin simulointitulokset riippuvat valitun perusjoukon ominaisuuksista. Perusjoukkojen eroavuutta toisistaan on mitattu alueiden välistä vaihtelua kuvaavalla homogeenisuusmitalla.

Varsinaisessa alue-estimoinnissa käytettäväksi malliksi on valittu hierarkkinen, yhtä apumuuttujaa käyttävä yksikkötason sekamalli, joka sisältää kiinteän vaikutuksen (synteettisen regressio-osan) ja erilliset aluevaikutukset. Estimointimenetelmänä on valittuun malliin perustuva paras lineaarinen ennustin (EBLUP). Ensisijaisena tavoitteena on ollut löytää sellainen otoskiintiöinti, joka tuottaisi estimoitujen alue-ennusteiden keskineliövirheiden (MSE) keskiarvolle mahdollisimman pienen arvon. Lisäksi on kiinnitetty huomiota alue-ennusteiden vaihtelukertoimiin (CV) sekä muuhun alue-ennusteiden tarkkuuteen ja tehokkuuteen, joiden mittaamiseen on käytettävissä useita niille kehitettyjä laatumittareita. Koska MSE-keskiarvon minimointi alueiden otoskokojen funktiona on käytännössä mahdotonta MSE:n monimutkaisen rakenteen vuoksi, tehokasta kiintiöintiä on etsitty kolmella muulla tavalla.

Kokeilussa olivat mukana kolme tässä työssä kehitettyä kiintiöintiä: parametriperusteinen RV_{opt} –kiintiöinti sekä malliin perustuvat *g1*- ja SI_{opt} -kiintiöinti. Näistä g1-kiintiöinti käyttää vain apumuuttujatietoa, kun taas muut kaksi edellyttävät apumuuttujan lisäksi sijaismuuttujan käyttöä samaan tapaan kuin eräät kirjallisuudesta poimitut aluekiintiöintiratkaisut sisältävät. Näiden vertailukohteiksi otettiin seuraavat kirjallisuudesta poimitut viisi kiintiöintiä: tasakiintiöinti, suhteellinen kiintiöinti, perusjoukon suhteen optimaalinen Neyman-kiintiöinti, alueoptimaalinen potenssikiintiöinti ja epälineaarinen optimikiintiöinti (NLP).

Ensimmäisen eli RV_{opt} –kiintiöinnin lähtökohtana on ollut apumuuttuja ja perusjoukosta poimittu pieni esiotos. Esiotoksesta on muodostettu kolme regressiomallia, joiden mukaisesti on laskettu vastemuuttujaa korvaavalle sijaismuuttujalle arvot apumuuttujan avulla. Ajatus on se, että apumuuttujassa esiintyvä vaihtelu siirtyy sijaismuuttujaan. Optimointikriteeri, johon aluekiintiöinti on perustunut, on sijaismuuttujan aluekohtaisten suhteellisten varianssien minimointi otoskokojen funktiona. Kiintiöinti suosii ensisijaisesti alueita, joiden sisäinen suhteellinen vaihtelu on suuri, riippumatta alueen koosta, ja otoskokojen väliset erot ovat varsin suuria. Kiintiöinnissä ei voida ottaa huomioon alueiden välistä vaihtelua.

Toinen kiintiöinti on johdettu analyyttisesti MSE:n tärkeimmän komponentin *g1*:n avulla. Perustelu vain yhden komponentin käyttöön on, että jos alueiden välistä vaihtelua esiintyy riittävästi, on hyvin mahdollista, että *g1*:n osuus MSE:n kokonaisarvosta on 85–90 %, jopa yli 90 %, ja komponentin lauseke on varsin yksinkertainen. Kiintiöinti on sikäli mielenkiintoinen, että se perustuu täysin apumuuttujan arvoihin, joiden avulla mallin komponentti *g1* on estimoitavissa ja sen avulla otoskoot ratkaistavissa alueittain. Laskennalliset otoskoot kehittyvät suhteellisen kiintiöinnin suuntaan, mutta erojakin on.

Myös kolmannen eli SI_{opt} -kiintiöinnin perustana on ollut apumuuttuja, esiotos ja niiden avulla johdettu sijaismuuttuja. Tällöin on mahdollista muodostaa samanlainen lineaarinen sekamalli apu- ja sijaismuuttujan välillä kuin tässä tutkimuksessa on käytetty. Rekisteristä simuloidaan SRSWOR-otoksia, joissa alueiden otoskoot määräytyvät satunnaisesti. Otoksille suoritetaan EBLUP-estimointi, jonka tuloksista voidaan tutkia, millaisilla otoskokojakaumilla voidaan otoksissa päästä ennalta asetettuun optimaaliseen tavoitteeseen eli otoskohtaisesti alueiden MSE-arvojen mahdollisimman pieniin keskiarvoihin. Kiintiöinti on johdettu otoskokojakaumien mediaaneista. Alueiden otoskoot kasvavat pääsääntöisesti alueiden koon kasvun myötä, mutta alueet, joiden ominaisuudet poikkeavat selvästi perusjoukon vastaavista, saavat otoskoon, joka ei vastaa niiden suhteellista kokoa. Tässä mielessä alueiden välinen vaikutus näkyy kiintiöinnissä.

Estimoinnin kohteena on ollut vastemuuttujan kokonaismäärä aluekohtaisesti, mutta osittain myös perusjoukon tasolla. Simuloinnilla on tuotettu kiintiöintimenetelmittäin 1500 ositettua yksinkertaista satunnaisotosta ja laskettu valittuun aluemalliin sovellettavan EBLUP-estimoinnin avulla kaikista otoksista samat malliperusteiset estimaatit ja tunnusluvut sekä niihin perustuvat estimoinnin tehokkuutta ja luotettavuutta kuvaavat laatumittarit. Tulokset ovat etupäässä alue- ja otoskohtaisia, mutta myös perusjoukon tason tuloksia. Laatumittareista keskeinen on aluekohtaisesti laskettava alueen ennustevirheen neliöön perustuva keskimääräinen suhteellinen virhe ($RRMSE_d$ %). Tälle mittarille on laskettu aluekohtaiset arvot ja perusjoukkotason arvo jokaisessa kiintiöinnissä.

Simulointituloksista käy ilmi, miten kiintiöintimenetelmien tulokset ovat erilaiset, kun perusjoukko on jaettuna 34 alueeseen ja 14 alueeseen. Koska kokonaisotoskoko on molemmissa rakenteissa sama eli 170, on selvää, että 34 alueen otoskoot ovat huomattavasti pienemmät kuin 14 alueen rakenteessa. Alueiden välinen vaihtelu on jälkimmäisessä rakenteessa pienempi, mutta alueiden väliset kokoerot ovat suurempia. Toisaalta tuloksissa on eroa sen mukaan, tarkastellaanko niitä aluekohtaisesti ja otoskohtaisesti vai perusjoukon tasolla. Kun kiintiöinnit jaetaan

kahteen ryhmään muodostamisperiaatteen mukaan ja vertaillaan niiden tuloksia, löytyy selviä eroja myös näin tarkasteltuna.

Tuloksia tarkastellaan aluksi alue- ja otoskohtaisella tasolla. Alustava tulosten tarkastelu paljastaa, että parametriperusteisten (Neyman, potenssi, NLP ja RV_{opt}) kiintiöintien tulokset ovat useimmissa tapauksissa selvästi huonommat kuin muiden kiintiöintien. Kaksi ensimmäistä kiintiöintiä käyttävät apumuuttujatietoa. Niistä molemmat suosivat ensisijaisesti suuria alueita, joiden sisäinen vaihtelu on voimakasta, mutta potenssikiintiöinnissä voi myös pieni alue saada suuren otoskoon. NLP- ja RV_{opt} –kiintiöinti käyttävät sijaismuuttujatietoa. Ne suosivat ensisijaisesti alueita, joiden sisäinen suhteellinen vaihtelu on voimakasta, alueen koosta riippumatta. Kaikissa neljässä kiintiöinnissä alueiden otoskokojen väliset erot ovat suuret, erityisesti NLP-kiintiöinnissä, eikä niissä voida ottaa huomioon alueiden välistä vaihtelua. Ne kaikki soveltuvat hyvin huonosti valitun aluemallin pohjalta käytettävään EBLUP-estimointiin.

Tasakiintiöinnissä sama otoskoko takaa kaikkien alueiden tasaisen edustuksen. Otos on siinä mielessä tasapainoinen. Kiintiöinti on yleisesti tuskin koskaan huonoin ratkaisu, mutta se ei ota käytettyä aluemallia eikä estimointimenetelmää huomioon. Kun perusjoukko on jaettu 34 alueeseen, on tasakiintiöinti paras aluetasolla ja otoskohtaisesti eri laatumittareilla mitattuna. Toisaalta suurimpien alueiden MSE-arvot ovat korkeita, mikä johtaa vastemuuttujan alue-estimaattien pitkiin luottamusväleihin ko. alueilla.

Kun perusjoukko muodostuu 14 alueesta, ovat tasakiintiöinnissä monen suuren alueen MSE- ja CV-arvot korkeat ja luottamusvälit pitkät. Alueiden suhteellisen keskivirheen ($RRMSE_d$ %) ja harhan (ARB_d %) tasot ovat muihin kiintiöinteihin verrattuna alhaisimpien joukossa. Kiintiöinnin otoskohtainen MSE-keskiarvojakauma ei ole tasoltaan alhaisimpien joukossa, mutta otoskohtainen CV-keskiarvojakauma ja keskimääräisen suhteellisen harhan jakauma vastaavasti ovat tasoltaan alhaisia, joskaan eivät parhaita. Suurten alueiden korkeat MSE- ja CV-arvot ovat kiintiöinnin suurimpia ongelmia.

Suhteellinen kiintiöinti takaa alueille niiden suhteellista kokoa vastaavan edustuksen, mutta ei ota käytettyä mallia eikä estimointimenetelmää huomioon. Toisaalta kun on kysymys vastemuuttujan aluekohtaisten kokonaismäärien estimoinnista, parantaa suhteellinen kiintiöinti estimoinnin luotettavuutta. Kiintiöinnin alue- ja otoskohtaiset tulokset kahdessa eri aluerakenteessa eivät ole koskaan heikoimpia, joskaan eivät parhaimpiakaan. Puutteista voidaan mainita se, että joidenkin pienien alueiden suhteellisen keskivirheen ja harhan arvot ovat varsin korkeita. Kaiken kaikkiaan

kiintiöinti on siinä mielessä turvallinen vaihtoehto, että estimointitulokset tuskin ovat kovin huonoja. Toisaalta aivan parhaiden tulosten saavuttaminen on epätodennäköistä.

g1-kiintiöinti ei toimi hyvin 34 alueen perusjoukossa. Havainto koskee sekä alue- että otoskohtaisia tuloksia. Merkittävä syy tähän on se, että otoskoot ovat nollia kolmella alueella, mikä aiheuttaa niistä kaikille suuret MSE- ja CV-arvot sekä kahdelle korkean suhteellisen keskivirheen ja harhan tason. Toisaalta kolmannessa näistä alueista on matala suhteellisen keskivirheen ja harhan taso. Sen sijaan 14 alueen perusjoukossa kiintiöinnin MSE-arvojen taso on alhaisimpien joukossa sekä alue- että otoskohtaisella tasolla ja huomattavasti alhaisempi kuin esimerkiksi tasakiintiöinnissä. Aluekohtaisten CV-arvojen taso on muuten alhainen lukuun ottamatta kahta otoskoon nolla saanutta aluetta. Alueiden suhteellisen keskivirheen ja harhan taso on kokonaisuudessaan alhaisin eri kiintiöinneistä. Tämä on huomionarvoinen seikka, kun otetaan huomioon kahden alueen otoskoko nolla.

SI_{opt} –kiintiöinti toimii varsin hyvin 34 alueen perusjoukossa, sekä alue- että otoskohtaisten tulosten perusteella. MSE-arvojen taso on tasakiintiöintiin verrattuna lähes yhtä alhainen, mutta CV-arvojen taso on jonkin verran korkeampi. Myös suhteellisen keskivirheen taso on korkeampi ja suhteellisen harhan taso vielä korkeampi tasakiintiöintiin verrattuna. Kiintiöinti toimii paremmin 14 alueen perusjoukossa alue- ja otoskohtaisten tulosten mukaan. MSE- ja CV-arvojen taso on selvästi alhaisin. Myös suhteellisen keskivirheen ja harhan tasot ovat alhaisimpien joukossa.

Vastemuuttujan oikeiden aluekohtaisten osuminen estimaateista laskettujen 95 %:n luottamusväleihin on muuten hyvä paitsi Helsingin osalta. Peittoprosentti on 95–100 % valtaosassa alueita, mutta Helsingin peittoprosentti on 34 alueen perusjoukon kiintiöinneissä aluekohtaisesti vain 25–53 % ja 14 alueen perusjoukon kiintiöinneissä vastaavasti välillä 34–54 %. Helsinki on ongelmallinen alue kaikissa kiintiöinneissä.

Kiintiöintien toimivuutta on verrattu myös perusjoukon tasolla siten, että kiintiöintikohtaisesti on laskettu perusjoukon kokonaismäärän estimaatin (alue-estimaattien summa) suhteellinen keskivirhe *RRMSE%*. Neyman-kiintiöinnin tulisi antaa paras arvo tälle laatumittarille. Tämä toteutuu vain 34 alueen perusjoukossa, jossa sille laskettu *RRMSE%*-arvo on 4.19 %. Vastaavasti paras tässä tutkimuksessa kehitetyistä uusista aluekiintiöintimenetelmistä on *g1*-kiintiöinti, jonka arvo on 4.97 %. Esimerkiksi tasakiintiöinnin vastaava arvo on jopa niinkin suuri kuin 6.96 %. Vastaavasti 14 alueen perusjoukossa potenssikiintiöinnillä on alhaisin perusjoukon tason suhteellinen virhe (4.62 %) ja Neyman-kiintiöinnillä lähes yhtä alhainen (4.73 %). Paras tässä tutkimuk-

sessa kehitetyistä kiintiöinneistä on jälleen *g1*-kiintiöinti (5.05 %). Edellä olevasta selviää, että perusjoukon rakenteella on vahva vaikutus yksittäisen kiintiöintimenetelmän toimivuuteen.

Loppupäätelmäksi voidaan tiivistää, että optimaalinen aluekiintiöinti on monitahoinen ongelma useastakin syystä. Kiintiöinti ehdollistuu kolmeen seikkaan, jotka ovat asetettu optimointikriteeri, perusjoukon rakenne ja perusjoukosta käytettävissä olevat tiedot. Osoituksena tästä olivat erilaiset tässä tutkimuksessa mukana olevat aluekiintiöintimenetelmät, joiden keskinäinen vertailu tehtiin simulointikokein. Niiden joukosta on vaikea arvioida, mikä niistä on paras. Tässä tutkimuksessa kehitetyistä kolmesta uudesta menetelmästä *g1*-kiintiöinti näyttäisi kehityskelpoiselta ratkaisulta varsinkin siksi, että se ei tarvitse tietoa vastemuuttujasta tai sen korvaavasta sijaismuuttujasta, vaan ainoastaan apumuuttujasta. Pieni alue voi saada otoskooksi nolla tässä kiintiöinnissä, mutta jos alueen ominaisuudet ovat lähellä perusjoukon vastaavia, on tällaiselle alueelle silti mahdollista saavuttaa hyviä estimointituloksia. Myös SI_{opt} –kiintiöinti, jonka kehittämisessä yhdistettiin esiotoksen avulla laskettu sijaismuuttuja, aluemalli ja EBLUP-estimointi, toimi hyvin ja erityisesti 14 alueen perusjoukossa. Tästä syystä on perusteltua tutkia tällaisenkin kiintiöinnin soveltuvuutta mallipohjaiseen alue-estimointiin erilaisissa aineistoissa. Koska kehittäminen alkaa esiotoksesta, sen poiminta tulee suunnitella siten, että otos edustaa alueiden ominaisuuksia ja niiden välistä vaihtelua mahdollisimman hyvin.

LÄHDEKIRJALLISUUS

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician* **42** 174–177.

Boonstra, H.J.B., van den Brakel, J., Buelens, B., Krieg, S. and Smeets, M. (2008). Towards small area estimation at Statistics Netherlands. *METRON- International Journal of Statistics* **66** 21–49.

Brackstone, G.J. (2002). Strategies and approaches for small area statistics. *Survey Methodology* **28** 117–123.

Costa, A., Satorra, A and Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT* **28** (1) 69–86**.**

Choudry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology* **38** 23-29.

Datta, G.S. and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10** 613–627.

Fabrizi, E. and Trivisano, C. (2007). Efficient stratification based on non-parametric regression methods. *Journal of Official Statistics* **23** 35–50.

Falorsi, P.D. and Righi, P. (2008). A balanced sampling approach for multi-way stratification for small area estimation. *Survey Methodology* **34** 223–234.

Gelman, A. (2007). Struggles with Survey Weighting and Regression Modelling. *Statistical Science* **22** 153–164.

Keto, M. and Pahkinen, E. (2009). On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation". In: J. Wywial and W. Gamrot (eds.). (2010). *Survey Sampling Methods in Economic and Social Research*. Katowice: Katowice University of Economics.

Keto, M. (2012). On sample allocation for effective EBLUP estimation of small area totals**.** *Contributed paper, BNU Workshop*, Valmiera, Latvia, August 2012.

Khan, M.G.M., Maiti, T. and Ahsan, M.J. (2010). An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach. *Journal of Official Statistics* **26** 695–708.

Kish, L. (1990). Rolling samples and censuses. *Survey Methodology* **20** 3–22.

Lehtonen, R. and Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *IASS Satellite Conference on Small Area Estimation*, Riga: Latvian Council of Science, 121–128.

Lehtonen, R. and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester; John Wiley & Sons. 2nd Edition.

Lehtonen, R., Myrskylä, M., Särndal, C.-E. and Veijanen, A. (2006). The role of models in model-assisted and model-dependent estimation for domains and small areas. *Working paper, BNU Workshop,* Ventspils, Latvia, August 2006.

Lehtonen, R., Myrskylä, M., Särndal, C.-E. and Veijanen, A. (2007). Estimation for domains and small areas under unequal probability sampling. *Pisa: The SAE2007 Conference,* September 2007.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The Effect of Model Choice in Estimation for Domains, Including Small Domains. *Survey Methodology* **29** 33–44.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* **7** 649–673.

Lohr, S. L. and Prasad, N.G.N. (2003). Small area estimation with auxiliary survey data. *The Canadian Journal of Statistics* **31** 383–396.

Longford, N. T. (2006). Sample Size Calculation for Small-Area Estimation. *Survey Methodology* **32** 87–96.

Longford, N. T. (2007). On Standard Errors of Model-Based Small-Area Estimators. *Survey Methodology* **33** 69–79.

Malec, D., Davis, W. W. and Cao, X. (1999). Model-Based Small-Area Estimates of Overweight Prevalence Using Sample Selection Adjustment. *Statistics in Medicine* **18**, 3189–3200.

Marker, D. A. (1999). Organization of Small Area Estimators Using a Generalized Linear Regression Framework. *Journal of Official Statistics* **15** 1–24.

Marker, D. A. (2001). Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators. *Survey Methodology* **27** 183–188.

Meza, J. L. and Lahiri, P. (2005). A note on the C_P statistic under the nested error regression model. *Statistics Canada* **31** 4–8.

Nissinen, K. (2009). *Small Area Estimation With Linear Mixed Models From Unit-Level Panel and Rotating Panel Data*. University of Jyväskylä, Department of Mathematics and Statistics, Report **117**. (Dissertation).

Pahkinen, E. (2012). *Kyselytutkimusten otantamenetelmät ja aineistoanalyysi*. Jyväskylä: Jyväskylä University Library Publishing Unit.

Pfeffermann, D. and Sverchkov, M. (2007). Small-Area Estimation Under Informative Probability Sampling of Areas and Within the Selected Areas. *Journal of the American Statistical Association* **102** 1427–1439.

Pfeffermann, D. and Sverchkov, M. (2004). Prediction of Finite Population Totals Based on the Sample Distribution. *Survey Methodology* **30** 79–92.

Prasad, N.G.N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association* **85** 163–171.

Rao, J. N. K. (2003). *Small Area Estimation*. Hobogen, New Jersey: Wiley.

Rao, J.N.K. and Choudry, G. H. (1999). Small Area Estimation: Overview and Empirical Studies in Cox ym. (eds) *Business Survey Methods.* New York: John Wiley & Sons, Inc.

Rao, J. N. K. and Ghosh, M. (1994). Small Area Estimation: An Appraisal. *Journal of Statistical Science* **9** 55–93.

Rao, J. N. K. and You, Y. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics* **3** 431–439.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology* **20** 3–22.

Särndal, C-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Thompson, S.K. (2002). *Sampling*. New York: Wiley.

Torabi, M. and Rao, J.N.K. (2005). Mean squared error estimators of small area means using survey weights. *SSC Annual Meeting, June 2005, Proceedings of the Survey Methods Section.*

Torabi, M. and Rao, J.N.K. (2008). Small area estimation under a two-level model. *Survey Methodology* **34** 223–234.

You, Y. and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology* **32** 97–103.

Havaintoaineiston lähde

Alma Mediapartners Oy (2011). Rekisteri kiinteistövälitysten toimeksiantoihin perustuvista myytävistä kiinteistöistä ”Etuovi.com” –verkkopalvelun kautta.

Liite A: Havaintoaineiston vaste- ja apumuuttujan tunnuslukuja.

Taulukko A.1. Vastemuuttujan *y* (asunnon hinta 1 000 €) tunnuslukuja: 34 aluetta.

Alue (kunta)	N_d	Min	Max	Summa	Keskiarvo	Hajonta	CV
Pieksämäki	111	26.00	195.00	6 872.6	61.91	26.07	0.421
Porvoo	112	4.90	1 950.00	25 408.8	226.86	207.82	0.916
Iisalmi	118	39.80	299.90	15 097.0	127.94	66.57	0.520
Varkaus	139	13.50	176.00	8 874.0	63.84	30.13	0.472
Kirkkonummi	140	12.32	452.40	34 311.1	245.08	119.34	0.487
Raisio	144	11.72	342.00	17 943.0	124.60	83.19	0.668
Pirkkala	148	15.79	705.50	30 322.6	204.88	87.82	0.429
Siilinjärvi	160	35.00	247.20	23 772.1	148.58	45.15	0.304
Salo	161	6.66	430.00	17 102.1	106.22	56.60	0.533
Savonlinna	167	25.90	403.28	22 311.5	133.60	76.86	0.575
Hyvinkää	171	20.89	550.00	36 503.3	213.47	72.22	0.338
Kokkola	173	35.00	650.00	28 003.5	161.87	82.34	0.509
Vihti	177	21.61	318.00	29 582.7	167.13	59.30	0.355
Kaarina	182	14.13	540.00	34 292.6	188.42	83.01	0.441
Kemi	199	15.70	256.00	19 072.5	95.84	53.60	0.559
Mikkeli	215	43.00	389.00	35 678.6	165.95	59.84	0.361
Riihimäki	225	10.18	465.00	38 002.0	168.90	62.77	0.372
Pori	233	29.90	450.00	41 001.7	175.97	85.42	0.485
Kempele	239	18.75	209.00	33 303.6	139.35	31.99	0.230
Nurmijärvi	245	22.36	306.00	44 757.1	182.68	58.16	0.318
Seinäjoki	249	59.50	385.05	39 333.3	157.97	58.94	0.373
Hämeenlinna	255	10.78	534.76	53 042.2	208.01	89.93	0.432
Kouvola	274	17.77	399.69	36 931.0	134.78	77.02	0.571
Lappeenranta	311	10.67	1 290.00	61 573.2	197.98	119.91	0.606
Rovaniemi	356	5.71	230.00	43 070.9	120.99	45.77	0.378
Espoo	365	18.38	1 598.00	119 649.0	327.81	136.91	0.418
Lahti	428	7.31	314.20	54 340.1	126.96	56.85	0.448
Kuopio	454	13.57	599.00	79 123.3	174.28	77.04	0.442
Turku	471	10.83	1 685.00	97 275.1	206.53	161.14	0.780
Jyväskylä	494	8.98	529.00	89 940.6	182.07	69.65	0.383
Vantaa	595	19.34	335.20	109 333.3	183.75	54.43	0.296
Helsinki	621	23.04	3 896.17	437 901.7	705.16	562.38	0.798
Tampere	650	9.70	1 160.00	155 017.4	238.49	121.99	0.512
Oulu	833	7.46	383.80	100 287.5	120.39	53.57	0.445
Perusjoukko	9 815	4.90	3 896.17	2 019 031.0	205.71	215.52	1.048

Taulukko A.2. Apumuuttujan *x* (asunnon koko m^2) tunnuslukuja sekä vaste- ja apumuuttujan välinen korrelaatio: 34 aluetta.

Alue (kunta)	N_d	Min	Max	Summa	Keskiarvo	Hajonta	CV	xy-korrelaatio
Pieksämäki	111	26.0	120.0	6 858.1	61.8	16.28	0.263	0.801
Porvoo	112	32.5	500.0	8 940.2	79.8	50.67	0.635	0.877
Iisalmi	118	27.0	133.4	8 049.1	68.2	20.24	0.297	0.760
Varkaus	139	22.0	180.0	9 301.0	66.9	25.36	0.379	0.828
Kirkkonummi	140	37.0	144.0	12 563.9	89.7	27.00	0.301	0.712
Raisio	144	27.0	130.0	10 763.5	74.7	22.19	0.297	0.430
Pirkkala	148	49.0	208.0	11 149.0	75.3	23.78	0.316	0.823
Siilinjärvi	160	26.5	129.0	10 063.9	62.9	18.12	0.288	0.336
Salo	161	17.0	196.0	11 119.2	69.1	24.47	0.354	0.714
Savonlinna	167	26.5	178.0	11 152.6	66.8	21.07	0.316	0.472
Hyvinkää	171	62.5	308.0	15 230.2	89.1	26.99	0.303	0.365
Kokkola	173	39.0	331.0	13 597.2	78.6	34.82	0.443	0.653
Vihti	177	29.0	205.0	11 632.9	65.7	20.08	0.306	0.443
Kaarina	182	55.0	206.5	15 308.5	84.1	21.43	0.255	0.646
Kemi	199	30.5	193.0	11 972.5	60.2	19.00	0.316	0.146
Mikkeli	215	26.0	230.0	14 633.3	68.1	21.18	0.311	0.388
Riihimäki	225	45.5	143.0	15 985.7	71.0	17.07	0.240	0.432
Pori	233	61.5	508.5	20 655.5	88.7	47.59	0.537	0.309
Kempele	239	38.0	125.0	16 408.5	68.7	14.21	0.207	0.637
Nurmijärvi	245	31.5	128.5	16 371.5	66.8	22.09	0.331	0.678
Seinäjoki	249	27.5	136.0	16 650.2	66.9	17.39	0.260	0.524
Hämeenlinna	255	67.0	454.0	23 574.4	92.4	34.25	0.370	0.429
Kouvola	274	53.5	195.0	21 327.6	77.8	19.00	0.244	0.387
Lappeenranta	311	62.0	245.0	26 422.0	85.0	23.31	0.274	0.683
Rovaniemi	356	20.0	78.0	18 832.0	52.9	13.79	0.261	0.389
Espoo	365	75.0	259.0	34 494.8	94.5	20.67	0.219	0.694
Lahti	428	27.5	76.5	24 640.3	57.6	11.18	0.194	0.223
Kuopio	454	53.0	204.5	36 688.6	80.8	23.68	0.293	0.587
Turku	471	68.0	449.0	42 778.7	90.8	25.82	0.284	0.635
Jyväskylä	494	61.0	200.0	40 000.3	81.0	17.62	0.218	0.509
Vantaa	595	24.0	78.0	32 975.7	55.4	13.95	0.252	0.558
Helsinki	621	77.0	455.0	76 931.1	123.9	57.98	0.468	0.753
Tampere	650	70.0	242.0	57 588.9	88.6	18.20	0.205	0.649
Oulu	833	21.0	82.0	42 801.5	51.4	15.62	0.304	0.326
Perusjoukko	9 815	17.0	508.5	747 462.4	76.2	31.76	0.417	0.674

Taulukko A.3. Vastemuuttujan (asunnon hinta 1 000 €) tunnuslukuja: 14 aluetta.

Alue	N_d	Min	Max	Summa	Keskiarvo	Hajonta	CV
Porvoo	112	4.90	1 950.00	25 408.8	226.86	207.82	0.916
Pirkkala	148	15.79	705.50	30 322.6	204.88	87.82	0.429
Etelä-Savo	493	25.90	403.28	64 862.7	131.57	72.90	0.554
Jyväskylä	494	8.98	529.00	89 940.6	182.07	69.65	0.383
Lappi	555	5.71	256.00	62 143.4	111.97	50.15	0.448
Kaakkois-Suomi	585	10.67	1 290.00	98 504.3	168.38	106.78	0.634
Helsinki	621	23.04	3 896.17	437 901.7	705.16	562.38	0.798
Satakunta-Pohjanmaa	655	29.90	650.00	108 338.5	165.40	75.85	0.459
Radanvarsikunnat	818	10.18	550.00	148 845.2	181.69	65.08	0.358
Kuopion seutu	871	13.50	599.00	126 866.5	145.66	75.79	0.520
Turun seutu	958	6.66	1 685.00	166 612.8	173.92	131.62	0.757
Oulun seutu	1072	7.46	383.80	133 591.1	124.62	50.19	0.403
Pääkaupunkiseutu	1100	12.32	1 598.00	263 293.4	239.36	117.84	0.492
Häme-Pirkanmaa	1333	7.31	1 160.00	262 399.6	196.85	110.76	0.563
Perusjoukko	9 815	4.90	3 896.17	2 019 031.0	205.71	215.52	1.048

Taulukko A.4. Apumuuttujan (asunnon koko m^2) tunnuslukuja sekä vaste- ja apumuuttujan välinen korrelaatio: 14 aluetta.

Alue	N_d	Min	Max	Summa	Keskiarvo	Hajonta	CV	xy-kor-relaatio
Porvoo	112	32.5	500.0	8 940.2	79.82	50.67	0.635	0.877
Pirkkala	148	49.0	208.0	11 149.0	75.33	23.78	0.316	0.823
Etelä-Savo	493	26.5	230.0	32 644.0	66.22	20.25	0.306	0.437
Jyväskylä	494	61.0	200.0	40 000.3	80.97	17.62	0.218	0.509
Lappi	555	30.5	193.0	30 804.5	55.50	16.22	0.292	0.207
Kaakkois-Suomi	585	53.5	245.0	47 749.6	81.62	21.68	0.266	0.601
Helsinki	621	77.0	455.0	76 931.1	123.88	57.98	0.468	0.753
Satakunta-Pohjanmaa	655	27.5	508.5	50 902.9	77.71	36.39	0.468	0.439
Radanvarsikunnat	818	29.0	308.0	59 220.3	72.40	23.84	0.321	0.517
Kuopion seutu	871	22.0	204.5	64 102.6	73.60	23.27	0.324	0.580
Turun seutu	958	17.0	449.0	79 969.9	83.48	25.71	0.308	0.635
Oulun seutu	1 072	21.0	125.0	59 210.0	55.23	16.92	0.306	0.392
Pääkaupunkiseutu	1 100	24.0	259.0	80 034.4	72.76	26.37	0.362	0.754
Häme-Pirkanmaa	1 333	27.5	454.0	105 803.6	79.37	25.54	0.322	0.602
Perusjoukko	9 815	17.0	508.5	747 462.4	76.16	31.76	0.417	0.674

Etelä-Savo: Mikkeli, Pieksämäki ja Savonlinna
Kaakkois-Suomi: Kouvola ja Lappeenranta
Turun seutu: Kaarina, Raisio, Salo ja Turku
Pääkaupunkiseutu: Espoo, Kirkkonummi ja Vantaa
Radanvarsikunnat: Hyvinkää, Nurmijärvi, Riihimäki ja Vihti
Kuopion seutu: Iisalmi, Kuopio, Siilinjärvi ja Varkaus
Helsinki, Jyväskylä, Pirkkala ja Porvoo muodostavat kukin oman alueensa.
Lappi: Kemi ja Rovaniemi
Satakunta-Pohjanmaa: Kokkola, Pori ja Seinäjoki
Oulun seutu: Kempele ja Oulu
Häme-Pirkanmaa: Hämeenlinna, Lahti ja Tampere

Liite B: Havaintoaineiston sijaismuuttujan y^* tunnuslukuja.

Taulukko B.1. Sijaismuuttujan tunnusluvut 34 alueella.

Alue (kunta)	N_d	Min	Max	Summa	Keskiarvo	Hajonta	CV
Pieksämäki	111	69.09	378.16	20 729.5	186.75	53.53	0.287
Porvoo	112	2.44	2043.08	23 408.9	209.01	221.18	1.058
Iisalmi	118	57.08	316.17	18 577.2	157.43	49.28	0.313
Varkaus	139	-43.39	646.28	21 219.5	152.66	110.72	0.725
Kirkkonummi	140	81.43	341.98	29 380.1	209.86	65.75	0.313
Raisio	144	57.08	342.00	24 961.7	173.35	54.82	0.316
Pirkkala	148	74.47	768.50	28 031.2	189.40	103.80	0.548
Siilinjärvi	160	55.86	305.45	23 119.4	144.50	44.13	0.305
Salo	161	-65.21	716.12	26 088.7	162.04	106.80	0.659
Savonlinna	167	-23.75	637.55	25 398.0	152.08	91.98	0.605
Hyvinkää	171	143.52	741.32	35 604.0	208.21	65.72	0.316
Kokkola	173	30.82	1305.40	35 232.1	203.65	151.98	0.746
Vihti	177	61.95	490.51	26 792.6	151.37	48.90	0.323
Kaarina	182	164.44	662.58	47 350.3	260.17	70.48	0.271
Kemi	199	-6.29	703.03	24 515.4	123.19	82.94	0.673
Mikkeli	215	54.65	551.39	33 769.3	157.07	51.57	0.328
Riihimäki	225	96.00	453.79	48 872.0	217.21	56.34	0.259
Pori	233	129.03	2080.18	57 676.4	247.54	207.73	0.839
Kempele	239	108.55	394.60	50 032.5	209.34	46.73	0.223
Nurmijärvi	245	-1.92	421.48	37 303.7	152.26	96.41	0.633
Seinäjoki	249	74.02	430.77	50 663.3	203.47	57.18	0.281
Hämeenlinna	255	142.00	1842.29	67 350.1	264.12	149.90	0.568
Kouvola	274	159.51	624.76	65 632.6	239.54	62.47	0.261
Lappeenranta	311	142.31	587.91	61 643.1	198.21	56.76	0.286
Rovaniemi	356	49.36	240.07	56 082.6	157.54	45.33	0.288
Espoo	365	230.20	835.20	107 434.4	294.34	67.97	0.231
Lahti	428	74.02	235.14	73 999.8	172.90	36.76	0.213
Kuopio	454	120.39	489.29	85 403.3	188.11	57.67	0.307
Turku	471	156.92	1084.65	100 085.4	212.50	62.88	0.296
Jyväskylä	494	184.17	641.20	123 421.4	249.84	57.93	0.232
Vantaa	595	62.52	240.07	98 668.5	165.83	45.88	0.277
Helsinki	621	196.69	1846.66	249 224.4	401.33	253.09	0.631
Tampere	650	213.76	779.30	178 694.9	274.92	59.84	0.218
Oulu	833	42.47	191.01	97 004.5	116.45	38.03	0.327
Perusjoukko	9 815	-65.21	2 080.18	2 053 370.7	209.21	120.96	0.578

Taulukko B.2. Sijaismuuttujan tunnusluvut 14 alueella.

Alue	N_d	Min	Max	Summa	Keskiarvo	Hajonta	CV
Porvoo	112	12.83	2 331.63	27 725.7	247.55	251.33	1.015
Pirkkala	148	146.49	441.12	28 902.1	195.28	44.06	0.226
Etelä-Savo	493	25.46	682.55	76 413.3	155.00	65.24	0.421
Jyväskylä	494	138.2	585.92	100 049.7	202.53	56.74	0.280
Lappi	555	6.14	563.37	66 874.8	120.50	52.23	0.433
Kaakkois-Suomi	585	114.04	730.86	119 706.5	204.63	69.82	0.341
Helsinki	621	233.55	2 108.43	289 439.2	466.09	287.58	0.617
Satakunta-Pohjanmaa	655	-11.97	2 373.79	155 294.7	237.09	180.52	0.761
Radanvarsikunnat	818	109.43	626.42	155 294.5	189.85	43.14	0.227
Kuopion seutu	871	96.46	434.63	167 293.3	192.07	44.18	0.230
Turun seutu	958	87.2	887.69	201 541.0	210.38	47.64	0.226
Oulun seutu	1 072	9.36	344.34	128 237.1	119.62	54.48	0.455
Pääkaupunkiseutu	1 100	-29.33	1 136.27	233 761.4	212.51	130.81	0.616
Häme-Pirkanmaa	1 333	106.65	896.96	270 296.8	202.77	47.32	0.233
Perusjoukko	9 815	-29.33	2 373.79	2 020 830.3	205.89	133.51	0.648

Liite C: Sijaismuuttujan y^* aluekohtaisiin suhteellisiin variansseihin perustuva otoskiintiöinti.

Taulukko C.1. Otoskiintiöinti 34 alueelle.

Paikkakunta (= alue)	N_d	$CV_d(y^*)$	$w_d = N_d / N$	$\sqrt{w_d}\, CV_d(y^*)$	Laskennall. otoskoko	Lopullinen otoskoko
Pieksämäki	111	0.2866	0.0113	0.0305	2.341	2
Porvoo	112	1.0583	0.0114	0.1130	8.682	9
Iisalmi	118	0.3130	0.0120	0.0343	2.636	3
Varkaus	139	0.7253	0.0142	0.0863	6.629	7
Kirkkonummi	140	0.3133	0.0143	0.0374	2.874	3
Raisio	144	0.3162	0.0147	0.0383	2.942	3
Pirkkala	148	0.5480	0.0151	0.0673	5.168	5
Siilinjärvi	160	0.3054	0.0163	0.0390	2.995	3
Salo	161	0.6591	0.0164	0.0844	6.483	7
Savonlinna	167	0.6048	0.0170	0.0789	6.059	6
Hyvinkää	171	0.3156	0.0174	0.0417	3.199	3
Kokkola	173	0.7463	0.0176	0.0991	7.609	8
Vihti	177	0.3230	0.0180	0.0434	3.331	3
Kaarina	182	0.2709	0.0185	0.0369	2.833	3
Kemi	199	0.6733	0.0203	0.0959	7.363	7
Mikkeli	215	0.3283	0.0219	0.0486	3.732	4
Riihimäki	225	0.2594	0.0229	0.0393	3.016	3
Pori	233	0.8392	0.0237	0.1293	9.930	10
Kempele	239	0.2232	0.0244	0.0348	2.675	3
Nurmijärvi	245	0.6332	0.0250	0.1000	7.683	8
Seinäjoki	249	0.2810	0.0254	0.0448	3.438	4
Hämeenlinna	255	0.5675	0.0260	0.0915	7.025	7
Kouvola	274	0.2608	0.0279	0.0436	3.347	3
Lappeenranta	311	0.2864	0.0317	0.0510	3.915	4
Rovaniemi	356	0.2877	0.0363	0.0548	4.208	4
Espoo	365	0.2309	0.0372	0.0445	3.420	3
Lahti	428	0.2126	0.0436	0.0444	3.410	3
Kuopio	454	0.3066	0.0463	0.0659	5.064	5
Turku	471	0.2959	0.0480	0.0648	4.978	5
Jyväskylä	494	0.2318	0.0503	0.0520	3.995	4
Vantaa	595	0.2767	0.0606	0.0681	5.232	5
Helsinki	621	0.6306	0.0633	0.1586	12.182	12
Tampere	650	0.2177	0.0662	0.0560	4.302	4
Oulu	833	0.3266	0.0849	0.0951	7.307	7
Yhteensä	9 815		1.0000	2.2136	170.000	170

Taulukko C.2. Otoskiintiöinti 14 alueelle.

Alue	N_d	$CV_d(y^*)$	$w_d = N_d / N$	$\sqrt{w_d}\, CV_d(y^*)$	Laskennall. otoskoko	Lopullinen otoskoko
Porvoo	112	1.0153	0.0114	0.1085	12.465	12
Pirkkala	148	0.2256	0.0151	0.0277	3.184	3
Etelä-Savo	493	0.4209	0.0502	0.0943	10.842	11
Jyväskylä	494	0.2802	0.0503	0.0629	7.225	7
Lappi	555	0.4335	0.0565	0.1031	11.848	12
Kaakkois-Suomi	585	0.3412	0.0596	0.0833	9.574	10
Helsinki	621	0.6170	0.0633	0.1552	17.838	18
Satakunta-Pohjanmaa	655	0.7614	0.0667	0.1967	22.607	23
Radanvarsikunnat	818	0.2272	0.0833	0.0656	7.540	8
Kuopion seutu	871	0.2300	0.0887	0.0685	7.876	8
Turun seutu	958	0.2264	0.0976	0.0707	8.131	8
Oulun seutu	1 072	0.4555	0.1092	0.1505	17.300	17
Pääkaupunkiseutu	1 100	0.6155	0.1121	0.2061	23.685	23
Häme-Pirkanmaa	1 333	0.2334	0.1358	0.0860	9.885	10
Perusjoukko	9 815	0.6484	1.0000	1.4791	170.000	170

Liite D: Eri kiintiöintien otoskoot simuloinneissa (1500 otosta/kiintiöinti).

Taulukko D.1. Kiintiöintikohtaiset otoskoot: 34 aluetta.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi (kunta)	Koko N_d	Tasa	Suht	Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
Pieksämäki	111	5	2	1	3	2	2	0	3
Porvoo	112	5	2	4	7	22	9	0	3
Iisalmi	118	5	2	2	3	2	3	0	3
Varkaus	139	5	2	3	4	12	7	2	3
Kirkkonummi	140	5	2	3	4	2	3	2	3
Raisio	144	5	3	2	4	2	3	2	4
Pirkkala	148	5	3	3	4	7	5	2	3
Siilinjärvi	160	5	3	2	3	2	3	2	3
Salo	161	5	3	3	4	10	7	2	4
Savonlinna	167	5	3	3	4	8	6	2	4
Hyvinkää	171	5	3	3	4	2	3	2	4
Kokkola	173	5	3	4	6	12	8	2	2
Vihti	177	5	3	3	4	2	3	2	3
Kaarina	182	5	3	3	4	2	3	2	3
Kemi	199	5	4	3	4	10	7	3	3
Mikkeli	215	5	4	3	4	3	4	3	3
Riihimäki	225	5	4	3	4	2	3	4	4
Pori	233	5	4	8	9	16	10	4	4
Kempele	239	5	4	3	3	2	3	4	4
Nurmijärvi	245	5	4	4	5	9	8	4	4
Seinäjoki	249	5	4	3	4	2	4	4	5
Hämeenlinna	255	5	5	7	7	7	7	4	3
Kouvola	274	5	5	4	4	2	3	5	4
Lappeenranta	311	5	5	5	5	2	4	6	4
Rovaniemi	356	5	6	4	4	2	4	7	6
Espoo	365	5	6	6	5	2	3	7	6
Lahti	428	5	7	4	4	2	3	8	8
Kuopio	454	5	8	8	6	2	5	9	8
Turku	471	5	8	9	7	2	5	9	7
Jyväskylä	494	5	9	6	5	2	4	10	8
Vantaa	595	5	10	6	5	2	5	12	11
Helsinki	621	5	11	26	14	9	12	13	9
Tampere	650	5	11	9	6	2	4	14	11
Oulu	833	5	14	10	7	3	7	18	13
Yhteensä	9 815	170	170	170	170	170	170	170	170

Taulukko D.2. Kiintiöintikohtaiset otoskoot: 14 aluetta.

Alue		Lukumääriin perustuvat		Parametriperusteiset				Malliperusteiset	
Nimi	Koko			Neyman	Potenssi	NLP	RV_{opt}	*g1*	SI_{opt}
	N_d	Tasa	Suht						
Porvoo	112	13	2	4	9	38	12	0	4
Pirkkala	148	12	3	2	5	3	3	0	4
Etelä-Savo	493	12	9	7	9	10	11	8	9
Jyväskylä	494	12	9	6	7	4	7	8	9
Lappi	555	12	10	6	8	11	12	9	10
Kaakkois-Suomi	585	12	10	8	9	7	10	9	10
Helsinki	621	13	11	24	21	21	18	10	8
Satakunta-Pohjanmaa	655	12	11	16	17	31	23	11	11
Radanvarsikunnat	818	12	14	13	13	3	8	15	14
Kuopion seutu	871	12	15	14	13	3	8	16	15
Turun seutu	958	12	16	16	14	3	8	18	16
Oulun seutu	1 072	12	18	12	12	12	17	20	21
Pääkaupunkiseutu	1 100	12	19	19	16	21	23	21	16
Häme-Pirkanmaa	1 333	12	23	23	17	3	10	25	23
Yhteensä	9 815	170	170	170	170	170	170	170	170

Liite E: Otossimuloinneissa käytetyt satunnaisluvut.

Kiintiöinti	34 aluetta	14 aluetta
Tasakiintiöinti	657 173 773	922 244 457
Suhteellinen kiintiöinti	88 173 759	966 773 331
Optimaalinen kiintiöinti (Neyman)	865 717 389	266 773 739
Potenssikiintiöinti	999 173 778	166 773 339
NLP-kiintiöinti	856 354 267	462 313 375
RV-optimaalinen kiintiöinti	35 996 549	586 446 879
g1 -kiintiöinti	969 647 391	144 446 883
Simu-optimaalinen kiintiöinti	134 743 459	1 462 783 371

Liite F: Otoskokojen johtaminen EBLUP-estimointia varten *g1*-kiintiöinnissä MSE:n ensimmäisen komponentin avulla.

MSE:n varianssikomponenteista σ_v^2 ja σ_e^2 riippuvan ensimmäisen komponentin g_{1d} lauseke on

$$g_{1d}(\sigma_v^2,\sigma_e^2) = (N_d - n_d)^2(1-\gamma_d)\sigma_v^2 ,$$

missä termi γ_d määritellään seuraavasti:

$$\gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2 n_d^{-1}} = \frac{n_d\sigma_v^2}{n_d\sigma_v^2 + \sigma_e^2} .$$

Kun termin γ_d kaava sijoitetaan g_{1d}:n lausekkeeseen, saadaan sille uusi muoto:

$$g_{1d}(\sigma_v^2,\sigma_e^2) = (N_d - n_d)^2(1-\gamma_d)\sigma_v^2 = (N_d - n_d)^2 \frac{\sigma_e^2}{n_d\sigma_v^2 + \sigma_e^2}\sigma_v^2$$

$$= (N_d - n_d)^2 \frac{\sigma_v^2}{\sigma_v^2/\sigma_e^2 \times n_d + 1} = (N_d - n_d)^2 \frac{1}{(1/\sigma_e^2)n_d + 1/\sigma_v^2}$$

$$= (N_d - n_d)^2 (1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-1} .$$

Alueiden g_{1d} -arvojen keskiarvon lauseke on

$$1/D\sum_{d=1}^{D} g_{1d}(\sigma_v^2,\sigma_e^2) = 1/D\sum_{d=1}^{D}(N_d - n_d)^2(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-1} .$$

Alkuosaa 1/D ei tarvita minimoinnissa, jossa käytetään Lagrangen menetelmää. Muodostetaan funktio

$$F(n_d,\lambda) = \sum_{d=1}^{D}(N_d - n_d)^2(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-1} + \lambda(\sum_{d=1}^{D} n_d - n) ,$$

joka derivoidaan otoskokojen n_d suhteen:

$$\frac{\partial}{\partial n_d}F = -2(N_d - n_d)(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-1} - (N_d - n_d)^2(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-2} \times 1/\sigma_e^2 + \lambda = 0$$

$$\Leftrightarrow -2(N_d - n_d)(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-2}$$

$$-(N_d - n_d)^2(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^{-2} \times 1/\sigma_e^2 + \lambda = 0$$

$$\Leftrightarrow \frac{-2(N_d - n_d)(1/\sigma_e^2 \times n_d + 1/\sigma_v^2) - 1/\sigma_e^2 \times (N_d - n_d)^2}{(1/\sigma_e^2 \times n_d + 1/\sigma_v^2)^2} + \lambda = 0$$

$$\Leftrightarrow \frac{-2(N_d - n_d)(1/\sigma_e^2 \times n_d + 1/\sigma_v^2) - 1/\sigma_e^2 \times (N_d - n_d)^2}{1/\sigma_e^4 \times n_d^2 + 2/(\sigma_e^2\sigma_v^2) \times n_d + 1/\sigma_e^4} + \lambda = 0.$$

Kun viimeisestä yhtälöstä poistetaan sulut ja kerrotaan nimittäjällä, saadaan yhtälö

$$-2N_d n_d/\sigma_e^2 - 2N_d/\sigma_v^2 + 2n_d^2/\sigma_e^2 + 2n_d/\sigma_v^2 - N_d^2/\sigma_e^2 + 2N_d n_d/\sigma_e^2 - n_d^2/\sigma_e^2$$
$$= -\lambda n_d^2/\sigma_e^4 - 2\lambda n_d/(\sigma_e^2\sigma_v^2) - \lambda/\sigma_e^4 .$$

Vasenta puolta voidaan sieventää hieman (vastalukujen poistaminen ja yhdistelyä), jolloin saadaan

$$(1/\sigma_e^2)n_d^2 + (2/\sigma_v^2)n_d - N_d^2/\sigma_e^2 - 2N_d/\sigma_v^2 = -\lambda n_d^2/\sigma_e^4 - 2\lambda n_d/(\sigma_e^2\sigma_v^2) - \lambda/\sigma_v^4$$
$$\Leftrightarrow (1/\sigma_e^2 + \lambda/\sigma_e^4)n_d^2 + (2/\sigma_v^2 + 2\lambda/(\sigma_e^2\sigma_v^2))n_d - N_d^2/\sigma_e^2 - 2N_d/\sigma_v^2 + \lambda/\sigma_v^4 = 0 .$$

Toisen asteen yhtälö ratkaistaan tunnetulla kaavalla:

$$n_d = \frac{-2/\sigma_v^2 - 2\lambda/(\sigma_e^2\sigma_v^2) \pm \sqrt{E}}{2(1/\sigma_e^2 + \lambda/\sigma_e^4)} . \qquad \text{(F.1)}$$

Diskriminantin E lausekkeesta saadaan sievennyksen jälkeen seuraava:

$$E = (2/\sigma_v^2 + 2\lambda/(\sigma_e^2\sigma_v^2))^2 - 4(1/\sigma_e^2 + \lambda/\sigma_e^4)(-N_d^2/\sigma_e^2 - 2N_d/\sigma_v^2 + \lambda/\sigma_v^4)$$
$$= 4/\sigma_v^4 + 8\lambda/(\sigma_e^2\sigma_v^4) + 4\lambda^2/(\sigma_e^4\sigma_v^4) + 4N_d^2/\sigma_e^4 + 8N_d/(\sigma_e^2\sigma_v^2) - 4\lambda/(\sigma_e^2\sigma_v^4)$$
$$+ 4\lambda N_d^2/\sigma_e^6 + 8\lambda N_d/(\sigma_e^4\sigma_v^2) - 4\lambda^2/(\sigma_e^4\sigma_v^4) .$$

Viimeisessä lausekkeesta poistetaan ”vastaluvut” sekä yhdistellään ja järjestellään jäseniä uudelleen, minkä jälkeen saadaan

$$E = 4/\sigma_v^4 + 8N_d/(\sigma_e^2\sigma_v^2) + 4N_d^2/\sigma_e^4 + 8\lambda/(\sigma_e^2\sigma_v^4) - 4\lambda/(\sigma_e^2\sigma_v^4) + 4\lambda N_d^2/\sigma_e^6 + 8\lambda N_d/(\sigma_e^4\sigma_v^2)$$
$$= (2/\sigma_v^2 + 2N_d/\sigma_e^2)^2 + \lambda(4/(\sigma_e^2\sigma_v^4) + 4N_d^2/\sigma_e^6 + 8N_d/(\sigma_e^4\sigma_v^2))$$
$$= 4(1/\sigma_v^2 + N_d/\sigma_e^2)^2 + (4\lambda/\sigma_e^2)(1/\sigma_v^4 + N_d^2/\sigma_e^4 + 2N_d/(\sigma_e^2\sigma_v^2))$$
$$= 4(1/\sigma_v^2 + N_d/\sigma_e^2)^2 + (4\lambda/\sigma_e^2)(1/\sigma_v^2 + N_d/\sigma_e^2)^2 = 4(1+\lambda/\sigma_e^2)(1/\sigma_v^2 + N_d/\sigma_e^2)^2$$
$$\Rightarrow \sqrt{E} = 2(1/\sigma_v^2 + N_d/\sigma_e^2)\sqrt{1+\lambda/\sigma_e^2} .$$

Sijoittamalla E:n neliöjuuri ratkaisukaavaan (F.1) saadaan lauseke

$$n_d = \frac{-2/\sigma_v^2 - 2\lambda/(\sigma_e^2\sigma_v^2) \pm \sqrt{E}}{2(1/\hat{\sigma}_e^2 + \lambda/\hat{\sigma}_e^4)} = \frac{-2/\sigma_v^2 - 2\lambda/(\sigma_e^2\sigma_v^2) \pm 2(1/\sigma_v^2 + N_d/\sigma_e^2)\sqrt{1+\lambda/\sigma_e^2}}{2/\hat{\sigma}_e^2(1+\lambda/\hat{\sigma}_e^2)}$$
$$= \frac{-1/\sigma_v^2 - \lambda/(\sigma_e^2\sigma_v^2) \pm (1/\sigma_v^2 + N_d/\sigma_e^2)\sqrt{1+\lambda/\sigma_e^2}}{1/\hat{\sigma}_e^2(1+\lambda/\hat{\sigma}_e^2)}$$
$$= \frac{(-1/\sigma_v^2)(1+\lambda/\sigma_e^2) \pm (1/\sigma_v^2 + N_d/\sigma_e^2)\sqrt{1+\lambda/\sigma_e^2}}{(1/\sigma_e^2)(1+\lambda/\sigma_e^2)} .$$

Koska negatiivinen juuri ei tule kyseeseen, saadaan sievennyksen tuloksena lauseke

$$n_d = -\sigma_e^2/\sigma_v^2 + (\sigma_e^2/\sigma_v^2 + N_d)/\sqrt{1+\lambda/\sigma_e^2} . \qquad \text{(F.2)}$$

Ratkaistaan λ yhtälön $\sum_{d=1}^{D} n_d = n$ avulla:

$$\sum_{d=1}^{D} n_d = -D\frac{\sigma_e^2}{\hat{\sigma}_v^2} + \frac{D(\sigma_e^2/\sigma_v^2)+N}{\sqrt{1+\lambda/\hat{\sigma}_e^2}} = n \Leftrightarrow \frac{D(\sigma_e^2/\sigma_v^2)+N}{\sqrt{1+\lambda/\hat{\sigma}_e^2}} = n + D(\sigma_e^2/\sigma_v^2)$$

$$\Leftrightarrow \frac{D(\sigma_e^2/\sigma_v^2)+N}{n+D(\sigma_e^2/\sigma_v^2)} = \sqrt{1+\lambda/\sigma_e^2} \Leftrightarrow \frac{(D(\sigma_e^2/\sigma_v^2)+N)^2}{(n+D(\sigma_e^2/\sigma_v^2))^2} = 1+\lambda/\sigma_e^2$$

$$\Rightarrow \lambda = \sigma_e^2\left(\frac{(D(\sigma_e^2/\sigma_v^2)+N)^2}{(n+D(\sigma_e^2/\sigma_v^2))^2} - 1\right).$$

Lauseke $1+\lambda/\sigma_e^2$ saadaan nyt muotoon

$$\begin{aligned} 1+\lambda/\sigma_e^2 &= 1+\sigma_e^2\left(\frac{(D(\sigma_e^2/\sigma_v^2)+N)^2}{(n+D(\sigma_e^2/\sigma_v^2))^2} - 1\right)/\sigma_e^2 = \frac{(D(\sigma_e^2/\sigma_v^2)+N)^2}{(n+D(\sigma_e^2/\sigma_v^2))^2} \\ \Rightarrow \sqrt{1+\lambda/\sigma_e^2} &= \frac{D(\sigma_e^2/\sigma_v^2)+N}{n+D(\sigma_e^2/\sigma_v^2)}. \end{aligned} \qquad \text{(F.3)}$$

Sijoittamalla lausekkeen (F.3) viimeinen muoto lausekkeeseen (F.2) saadaan alueen *d* otoskoolle n_d lopullinen lauseke

$$n_d = -\sigma_e^2/\sigma_v^2 + \frac{\sigma_e^2/\sigma_v^2 + N_d}{\sqrt{1+\lambda/\sigma_e^2}} \Rightarrow n_d = -\sigma_e^2/\sigma_v^2 + \frac{(N_d+\sigma_e^2/\sigma_v^2)(n+D(\sigma_e^2/\sigma_v^2))}{N+D(\sigma_e^2/\sigma_v^2)}. \qquad \text{(F.4)}$$

Kaavasta (G.4) voidaan osoittaa, että $\sum_{d=1}^{D} n_d = n$.

Alueen d otoskoon n_d arvo riippuu paitsi vakioista *N*, *n*, *D* ja N_d, myös aluevarianssista σ_v^2 ja satunnaisvarianssista σ_e^2. Kaava (F.4) ei toimi, jos aluevarianssi σ_v^2 tulee estimoinnissa nollaksi (mikä on mahdollista, joskin epätodennäköistä, jos malli on oikea). Tässä kaavassa esiintyy varianssisuhde σ_e^2/σ_v^2 useassa kohdassa. ja olettamalla ko. suhteelle eri arvoja saadaan erilaisia otoskokoja.

Suhde = 1: $(\sigma_e^2 = \sigma_v^2) \Rightarrow n_d = \frac{(N_d+1)(n+D)}{N+D} - 1$

Suhde = 2: $(\sigma_e^2/\sigma_v^2 = 2) \Rightarrow n_d = \frac{(N_d+2)(n+2D)}{N+2D} - 2$

Suhde = 0.5: $(\sigma_e^2/\sigma_v^2 = 0{,}5) \Rightarrow n_d = \frac{(N_d+0{,}5)(n+0{,}5D)}{N+0{,}5D} - 0{,}5$.

Merkitään varianssikomponenttien suhdetta symbolilla δ : $\delta = \sigma_e^2/\sigma_v^2$, ja sen estimaatti on $\hat{\delta} = \hat{\sigma}_e^2/\hat{\sigma}_v^2$. Tämä suhde voidaan esittää myös alueiden välisen sisäkorrelaation φ (lauseke 3.19) avulla seuraavasti:

$$\varphi = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2} = \frac{1}{1 + \sigma_e^2/\sigma_v^2} = \frac{1}{1+\delta} \Rightarrow \delta = 1/\varphi - 1 \, .$$

Tällöin tulee otoskoon n_d lausekkeeksi yleisessä tapauksessa seuraava:

$$\begin{aligned} n_d &= \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} - \delta = \frac{N_d n - (N - N_d D - n)\delta}{N + D\delta} \\ &= \frac{N_d n - (N - N_d D - n)(1/\varphi - 1)}{N + D(1/\varphi - 1)} \end{aligned} \qquad \text{(F.5)}$$

Mitä suurempi on g_{1d} :n prosenttiosuus MSE:n kokonaisarvosta, sitä parempi approksimaatio se on alueen d MSE:n arvosta. Kun g_{1d} :n lausekkeeseen

$$g_{1d}(\sigma_v^2, \sigma_e^2) = (N_d - n_d)^2 (1 - \gamma_d)\sigma_v^2 = (N_d - n_d)^2 \frac{\sigma_e^2 \sigma_v^2}{n_d \sigma_v^2 + \sigma_e^2}$$

sijoitetaan otoskoon n_d paikalle lauseke (H.5). saadaan sievennyksen jälkeen lauseke

$$g_{1d\rho pt}(\sigma_v^2, \sigma_e^2) = \frac{(N_d + \delta)(N - n)^2}{(N + D\delta)(n + D\delta)} \sigma_e^2 \, . \qquad \text{(F.6)}$$

Tämän lausekkeen arvo puolestaan riippuu suoraan mm. satunnaisvarianssin σ_e^2 arvosta. Nimittäjästä voidaan päätellä, että jos varianssien suhde δ on suuri (alueiden välinen vaihtelu on pieni), on MSE-approksimaatio pienempi kuin päinvastaisessa tapauksessa (alueiden välinen vaihtelu suuri). Alueiden g_{1d}-keskiarvon minimin lauseke on

$$\begin{aligned} &min(1/D \sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = 1/D \sum_{d=1}^{D} \frac{(N_d + \delta)(N - n)^2}{(N + D\delta)(n + D\delta)} \sigma_e^2 \\ &= 1/D \sum_{d=1}^{D} \frac{(N - n)^2 \sigma_e^2}{(N + D\delta)(n + D\delta)} (N_d + \delta) = 1/D \frac{(N - n)^2 \sigma_e^2}{(N + D\delta)(n + D\delta)} (N + D\delta) \, , \end{aligned}$$

joka sievenee vielä hieman. Keskiarvon lauseke voidaan ilmaista joko varianssisuhteen δ tai sisäkorrelaation φ avulla seuraavasti:

$$min(1/D\sum_{d=1}^{D} g_{1d}(\sigma_v^2,\sigma_e^2)) = \frac{(N-n)^2}{D(n+D\delta)}\sigma_e^2 = \frac{(N-n)^2}{D(n+D(1/\varphi-1))}\sigma_e^2. \qquad \text{(F.7)}$$

Tämä keskiarvo riippuu paitsi varianssisuhteesta δ, myös satunnaisvarianssista σ_e^2. Jälkimmäisen suuri arvo kasvattaa keskiarvoa ja pieni arvo pienentää. Toisaalta varianssisuhteen δ suuri arvo (sisäkorrelaatio φ on lähellä nollaa) pienentää keskiarvoa, ja päinvastoin: pieni δ:n arvo (sisäkorrelaatio lähellä ykköstä) kasvattaa keskiarvoa. Kun otoksesta estimoitu aluevarianssi tulee erikoistapauksessa nollaksi, mikä johtaa varianssisuhteen δ arvoon ääretön, on keskiarvolausekkeen (F.7) arvo nolla. Tämä seuraa tietysti myös siitä, että tällöin myös g_{1d} = 0. Nollakeskiarvo ei kuitenkaan ole hyvä asia, koska se johtaa mm. suuriin ennustevirheisiin. Estimoinnin kannalta olisi suositeltavaa saada poimituksi otos, josta laskettu sisäkorrelaatio vastaisi vähintään apumuuttujan alueiden välistä vaihtelua eli homogeenisuusmittaa R_a^2 (lauseke 6.1).

Liite G: Malliperusteisen *g1* -kiintiöinnin otoskokojen riippuvuus sisäkorrelaatiosta sekä alueiden ja perusjoukon ominaisuuksista kaavan (5.8) mukaisesti.

Taulukko G.1. Laskennalliset otoskoot. kun otoskoko n = 170: 34 aluetta.

		N	9 815	(perusjoukon koko)								
		n	170	(kokonaisotoskoko)								
		D	34	(alueiden lukumäärä)								
Varianssisuhde = satunnnaisvarianssi / aluevarianssi												
Sisäkorrelaatio = 1 / (1 + varianssisuhde)												
Varianssisuhde		10.00	8.00	6.00	4.00	2.01	1.00	0.80	0.40	0.20	0.10	0.05
Sisäkorrelaatio		0.091	0.111	0.143	0.200	**0.332**	0.500	0.556	0.714	0.833	0.909	0.952
Alue	N_d	Alueiden laskennallinen otoskoko										
Pieksämäki	111	-3.92	-2.79	-1.63	-0.46	***0.72***	1.32	1.44	1.68	1.80	1.86	1.89
Porvoo	112	-3.87	-2.74	-1.60	-0.43	***0.74***	1.34	1.46	1.70	1.82	1.88	1.91
Iisalmi	118	-3.57	-2.48	-1.37	-0.25	***0.88***	1.46	1.58	1.81	1.93	1.99	2.01
Varkaus	139	-2.52	-1.56	-0.59	0.40	***1.39***	1.90	2.00	2.20	2.31	2.36	2.38
Kirkkonummi	140	-2.47	-1.51	-0.55	0.43	***1.41***	1.92	2.02	2.22	2.32	2.37	2.40
Raisio	144	-2.27	-1.34	-0.40	0.55	***1.51***	2.00	2.10	2.30	2.40	2.44	2.47
Pirkkala	148	-2.06	-1.16	-0.25	0.67	***1.61***	2.09	2.18	2.37	2.47	2.52	2.54
Siilinjärvi	160	-1.46	-0.64	0.20	1.04	***1.90***	2.33	2.42	2.60	2.68	2.73	2.75
Salo	161	-1.41	-0.59	0.23	1.07	***1.92***	2.36	2.44	2.61	2.70	2.75	2.77
Savonlinna	167	-1.11	-0.33	0.46	1.26	***2.07***	2.48	2.56	2.73	2.81	2.85	2.87
Hyvinkää	171	-0.91	-0.16	0.61	1.38	***2.16***	2.56	2.64	2.80	2.88	2.92	2.94
Kokkola	173	-0.81	-0.07	0.68	1.44	***2.21***	2.60	2.68	2.84	2.92	2.96	2.98
Vihti	177	-0.61	0.11	0.83	1.57	***2.31***	2.69	2.76	2.91	2.99	3.03	3.05
Kaarina	182	-0.36	0.33	1.02	1.72	***2.43***	2.79	2.86	3.01	3.08	3.12	3.13
Kemi	199	0.50	1.07	1.65	2.24	***2.84***	3.14	3.20	3.32	3.39	3.42	3.43
Mikkeli	215	1.30	1.77	2.25	2.73	***3.22***	3.47	3.52	3.62	3.67	3.70	3.71
Riihimäki	225	1.80	2.21	2.62	3.04	***3.46***	3.68	3.72	3.81	3.85	3.88	3.89
Pori	233	2.20	2.56	2.92	3.29	***3.66***	3.85	3.88	3.96	4.00	4.02	4.03
Kempele	239	2.51	2.82	3.15	3.47	***3.80***	3.97	4.00	4.07	4.11	4.12	4.13
Nurmijärvi	245	2.81	3.09	3.37	3.66	***3.95***	4.10	4.12	4.18	4.21	4.23	4.24
Seinäjoki	249	3.01	3.26	3.52	3.78	***4.04***	4.18	4.21	4.26	4.29	4.30	4.31
Hämeenlinna	255	3.31	3.52	3.74	3.96	***4.19***	4.30	4.33	4.37	4.39	4.41	4.41
Kouvola	274	4.26	4.36	4.45	4.55	***4.65***	4.70	4.71	4.73	4.74	4.74	4.74
Lappeenranta	311	6.12	5.98	5.83	5.69	***5.54***	5.46	5.45	5.42	5.40	5.39	5.39
Rovaniemi	356	8.38	7.95	7.51	7.07	***6.62***	6.39	6.35	6.26	6.21	6.19	6.18
Espoo	365	8.83	8.34	7.85	7.35	***6.84***	6.58	6.53	6.43	6.37	6.35	6.33
Lahti	428	12.00	11.10	10.20	9.28	***8.36***	7.89	7.79	7.60	7.51	7.46	7.44
Kuopio	454	13.30	12.24	11.17	10.08	***8.99***	8.42	8.31	8.09	7.98	7.92	7.89
Turku	471	14.16	12.99	11.81	10.61	***9.40***	8.78	8.65	8.41	8.28	8.22	8.19
Jyväskylä	494	15.31	14.00	12.66	11.31	***9.95***	9.25	9.11	8.84	8.70	8.63	8.59
Vantaa	595	20.38	18.42	16.43	14.42	***12.39***	11.34	11.14	10.72	10.51	10.41	10.36
Helsinki	621	21.69	19.56	17.41	15.22	***13.01***	11.88	11.66	11.21	10.98	10.87	10.81
Tampere	650	23.15	20.83	18.49	16.11	***13.71***	12.48	12.24	11.75	11.50	11.38	11.32
Oulu	833	32.34	28.85	25.32	21.74	***18.13***	16.27	15.91	15.17	14.80	14.61	14.52
Yhteensä	9 815	170.00	170.00	170.00	170.00	**170.00**	170.00	170.00	170.00	170.00	170.00	170.00

Taulukko G.2. Laskennalliset otoskoot. kun otoskoko *n* = 170: 14 aluetta.

	N	9 815 (perusjoukon koko)									
	n	170 (kokonaisotoskoko)									
	D	14 (alueiden lukumäärä)									
Varianssisuhde = satunnnaisvarianssi / aluevarianssi											
Sisäkorrelaatio = 1 / (1 + varianssisuhde)											
Varianssisuhde		10.00	8.00	6.00	3.33	2.00	1.00	0.75	0.50	0.10	0.05
Sisäkorrelaatio		0.091	0.111	0.143	**0.231**	0.333	0.500	0.571	0.667	0.909	0.952
Alue	N_d	Alueiden laskennallinen otoskoko									
Porvoo	112	-6.20	-4.59	-2.97	**-0.80**	0.29	1.12	1.32	1.53	1.86	1.90
Pirkkala	148	-5.08	-3.57	-2.05	**-0.01**	1.02	1.79	1.98	2.18	2.49	2.52
Etelä-Savo	493	5.66	6.23	6.80	**7.57**	7.96	8.25	8.32	8.39	8.51	8.52
Jyväskylä	494	5.69	6.26	6.83	**7.59**	7.98	8.27	8.34	8.41	8.53	8.54
Lappi	555	7.59	7.99	8.39	**8.93**	9.20	9.41	9.46	9.51	9.59	9.60
Kaakkois-Suomi	585	8.53	8.85	9.16	**9.59**	9.81	9.97	10.01	10.05	10.12	10.12
Helsinki	621	9.65	9.87	10.09	**10.38**	10.53	10.64	10.67	10.70	10.74	10.75
Satakunta-Pohjanmaa	655	10.71	10.83	10.96	**11.13**	11.22	11.28	11.30	11.31	11.34	11.34
Radanvarsikunnat	818	15.78	15.46	15.14	**14.71**	14.49	14.33	14.29	14.25	14.18	14.18
Kuopion seutu	871	17.43	16.97	16.50	**15.88**	15.56	15.32	15.26	15.21	15.11	15.10
Turun seutu	958	20.14	19.44	18.74	**17.79**	17.31	16.95	16.86	16.77	16.63	16.61
Oulun seutu	1 072	23.69	22.68	21.66	**20.29**	19.60	19.09	18.96	18.83	18.62	18.59
Pääkaupunkiseutu	1 100	24.57	23.48	22.38	**20.91**	20.17	19.61	19.47	19.33	19.11	19.08
Häme-Pirkanmaa	1 333	31.82	30.09	28.36	**26.02**	24.85	23.97	23.75	23.53	23.18	23.13
Yhteensä	9 815	170.00	170.00	170.00	**170.00**	170.00	170.00	170.00	170.00	170.00	170.00

Otoksen tuntematon sisäkorrelaatio korvataan apumuuttuja-arvoista laskettavalla alueiden välisen vaihtelun voimakkuutta mittaavalla homogeenisuusmitalla. Tämän jälkeen voidaan *g1* -kiintiöinnin otoskoot laskea. Tutkimuksessa käytettyyn perusjoukkoon liittyvät laskennalliset otoskoot on lihavoitu. Lopulliset otoskoot ovat kokonaislukuja ja joskus harkinnanvaraisia (negatiiviset otoskoot ja pyöristämiset).

Liite H: Kiintiöinti sijaismuuttujan suhteellisten varianssien avulla.

Pienestä esiotoksesta (alaluku 5.2) saadaan regressiokertoimet apumuuttujan *x* ja vastemuuttujan *y* välille. Näiden avulla voidaan laskea apumuuttujan arvoista (täydellinen rekisteri) arvot sijaismuuttujalle y^*. Näitä arvoja on *N* kpl. Poimitaan sijaismuuttujasta *n:*n kokoinen otos, ja alueen *d* otoskoko on n_d.

Sijaismuuttujan y^* alueen *d* otoskeskiarvon $\bar{y}_d^*$ vaihtelukertoimen (CV) neliön (RV) eli suhteellisen varianssin lauseke, joka sisältää äärellisyyskertoimen, on

$$RV(\bar{y}_d^*) = (1/n_d - 1/N_d)S_d^2(y^*)/(\bar{Y}_d^*)^2.$$

Vaihtelukerroin on sama, kun estimoidaan vastemuuttujan kokonaismäärää alueella *d.*

Minimoidaan sijaismuuttujan CV-neliöiden (RV-arvojen) painotettu keskiarvo $\sum_{d=1}^{D} w_d RV(\bar{y}_d^*)$ otoskokojen n_d funktiona käyttämällä Lagrangen menetelmää, kun rajoiteyhtälö on $\sum_{d=1}^{D} n_d = n$. Painokertoimien w_d arvoista ei tehdä ennakko-oletuksia.

Minimoitavana on lauseke $\sum_{d=1}^{D} w_d (1/n_d - 1/N_d)S_d^2(y^*)/(\bar{Y}_d^*)^2$. Muodostetaan funktio

$$F(n_d, w_d, \lambda) = \sum_{d=1}^{D} w_d(1/n_d - 1/N_d)S_d^2(y^*)/(\bar{Y}_d^*) + \lambda(\sum_{d=1}^{D} n_d - n),$$

joka derivoidaan otoskokojen n_d suhteen ja asetetaan nollaksi. Tulos ei riipu termistä $1/N_d$. Sijaismuuttujan y^* varianssi ja keskiarvo tunnetaan joka alueelta. Derivoinnista saadaan yhtälö

$$\frac{\partial}{\partial n_d} F = -\frac{w_d}{n_d^2}\frac{S_d^2(y^*)}{(\bar{Y}_d^*)^2} + \lambda = 0 \Leftrightarrow n_d^2 = \frac{w_d}{\lambda}\frac{S_d^2(y^*)}{(\bar{Y}_d^*)^2} \Rightarrow n_d = \frac{\sqrt{w_d}}{\sqrt{\lambda}}\frac{S_d(y^*)}{\bar{Y}_d^*} = \frac{\sqrt{w_d}}{\sqrt{\lambda}} CV_d(y^*).$$

Ratkaistaan λ yhtälön $\sum_{d=1}^{D} n_d = n$ avulla:

$$\sum_{d=1}^{D} n_d = \sum_{d=1}^{D} \frac{\sqrt{w_d}}{\sqrt{\lambda}} CV_d(y^*) = \frac{1}{\sqrt{\lambda}} \sum_{d=1}^{D} \sqrt{w_d} CV_d(y^*) = n \Rightarrow \frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{d=1}^{D} \sqrt{w_d} CV_d(y^*)}$$

$$\Rightarrow n_d = \frac{\sqrt{w_d}}{\sqrt{\lambda}} CV_d(y^*) = \frac{\sqrt{w_d} CV_d(y^*)}{\sum_{d=1}^{D} \sqrt{w_d} CV_d(y^*)} n.$$

Alueen *d* otoskoko n_d on suoraan verrannollinen tuloon $\sqrt{w_d}CV_d(y^*)$ eli painon w_d neliöjuuren ja sijaismuuttujan y^* alueen *d* CV-arvon tuloon. Painoja w_d voidaan vaihdella alueiden tärkeyden mukaan. Jos painot ovat samoja, on otoskoon lauseke yksinkertaisesti

$$n_d = nCV_d(y^*) / \sum_{d=1}^{D} CV_d(y^*),$$

eli otoskoko on tällöin suoraan verrannollinen alueen CV-arvoon.

Liite J: NLP-kiintiöinnit Excelin Ratkaisin-apuohjelman avulla.

Taulukko J.1. Alueiden perustiedot kiintiöintiä varten: 34 aluetta.

		Kok. määrä	Keskiarvo	Hajonta	CV
Alue (kunta)	N_d	Y_d^*	$\bar{Y}_d^*$	$S_d(y^*)$	$C_d(y^*)$
Pieksämäki	111	20 729.48	186.75	53.53	0.2866
Porvoo	112	23 408.93	209.01	221.18	1.0583
Iisalmi	118	18 577.21	157.43	49.28	0.3130
Varkaus	139	21 219.49	152.66	110.72	0.7253
Kirkkonummi	140	29 380.14	209.86	65.75	0.3133
Raisio	144	24 961.68	173.35	54.82	0.3162
Pirkkala	148	28 031.23	189.40	103.80	0.5480
Siilinjärvi	160	23 119.36	144.50	44.13	0.3054
Salo	161	26 088.69	162.04	106.80	0.6591
Savonlinna	167	25 397.96	152.08	91.98	0.6048
Hyvinkää	171	35 603.99	208.21	65.72	0.3156
Kokkola	173	35 232.12	203.65	151.98	0.7463
Vihti	177	26 792.58	151.37	48.90	0.3230
Kaarina	182	47 350.28	260.17	70.48	0.2709
Kemi	199	24 515.38	123.19	82.94	0.6733
Mikkeli	215	33 769.33	157.07	51.57	0.3283
Riihimäki	225	48 871.99	217.21	56.34	0.2594
Pori	233	57 676.40	247.54	207.73	0.8392
Kempele	239	50 032.50	209.34	46.73	0.2232
Nurmijärvi	245	37 303.70	152.26	96.41	0.6332
Seinäjoki	249	50 663.25	203.47	57.18	0.2810
Hämeenlinna	255	67 350.13	264.12	149.90	0.5675
Kouvola	274	65 632.64	239.54	62.47	0.2608
Lappeenranta	311	61 643.07	198.21	56.76	0.2864
Rovaniemi	356	56 082.64	157.54	45.33	0.2877
Espoo	365	107 434.36	294.34	67.97	0.2309
Lahti	428	73 999.82	172.90	36.76	0.2126
Kuopio	454	85 403.28	188.11	57.67	0.3066
Turku	471	100 085.39	212.50	62.88	0.2959
Jyväskylä	494	123 421.36	249.84	57.93	0.2318
Vantaa	595	98 668.48	165.83	45.88	0.2767
Helsinki	621	249 224.43	401.33	253.09	0.6306
Tampere	650	178 694.90	274.92	59.84	0.2177
Oulu	833	97 004.54	116.45	38.03	0.3266
Perusjoukko	9 815	2 053 370.73	209.21	120.96	0.5782

Taulukko J.2. Minimiotoskoon ratkaisu NLP-kiintiöinnissä: 34 aluetta.

Alue (kunta)	k_d : n yläraja $= N_d / 2$	k_d (≥ 1)	$n_d = N_d / k_d$	$\sqrt{V(\hat{Y}_d^*)}$	$CV(\hat{Y}_d^*)$	$RV(\hat{Y}_d^*)$	RV-raja	Pyöristetty n_d	Lopullinen $CV(\hat{Y}_d^*)$
Pieksämäki	55.5	55.50	2.00	4 163.23	0.2008	0.0403	0.0418	2	0.2008
Porvoo	56	5.18	21.61	4 787.13	0.2045	0.0418	0.0418	22	0.2023
Iisalmi	59	51.37	2.30	3 799.04	0.2045	0.0418	0.0418	2	0.2194
Varkaus	69.5	12.05	11.53	4 339.39	0.2045	0.0418	0.0418	12	0.2001
Kirkkonummi	70	60.64	2.31	6 008.24	0.2045	0.0418	0.0418	2	0.2200
Raisio	72	61.22	2.35	5 104.66	0.2045	0.0418	0.0418	2	0.2220
Pirkkala	74	21.61	6.85	5 732.39	0.2045	0.0418	0.0418	7	0.2022
Siilinjärvi	80	72.73	2.20	4 727.91	0.2045	0.0418	0.0418	2	0.2146
Salo	80.5	16.50	9.76	5 335.14	0.2045	0.0418	0.0418	10	0.2018
Savonlinna	83.5	20.09	8.31	5 193.88	0.2045	0.0418	0.0418	8	0.2086
Hyvinkää	85.5	72.79	2.35	7 281.02	0.2045	0.0418	0.0418	2	0.2219
Kokkola	86.5	13.99	12.36	7 204.97	0.2045	0.0418	0.0418	12	0.2078
Vihti	88.5	71.94	2.46	5 479.08	0.2045	0.0418	0.0418	2	0.2271
Kaarina	91	91.00	2.00	9 019.78	0.1905	0.0363	0.0418	2	0.1905
Kemi	99.5	19.36	10.28	5 013.40	0.2045	0.0418	0.0418	10	0.2075
Mikkeli	107.5	84.41	2.55	6 905.83	0.2045	0.0418	0.0418	3	0.1882
Riihimäki	112.5	112.50	2.00	8 924.20	0.1826	0.0333	0.0418	2	0.1826
Pori	116.5	14.84	15.70	11 794.82	0.2045	0.0418	0.0418	16	0.2025
Kempele	119.5	119.50	2.00	7 863.40	0.1572	0.0247	0.0418	2	0.1572
Nurmijärvi	122.5	26.56	9.23	7 628.61	0.2045	0.0418	0.0418	9	0.2072
Seinäjoki	124.5	124.50	2.00	10 026.79	0.1979	0.0392	0.0418	2	0.1979
Hämeenlinna	127.5	34.11	7.48	13 773.10	0.2045	0.0418	0.0418	7	0.2115
Kouvola	137	137.00	2.00	12 059.31	0.1837	0.0338	0.0418	2	0.1837
Lappeenranta	155.5	155.50	2.00	12 441.68	0.2018	0.0407	0.0418	2	0.2018
Rovaniemi	178	178.00	2.00	11 377.68	0.2029	0.0412	0.0418	2	0.2029
Espoo	182.5	182.50	2.00	17 494.01	0.1628	0.0265	0.0418	2	0.1628
Lahti	214	214.00	2.00	11 099.12	0.1500	0.0225	0.0418	2	0.1500
Kuopio	227	203.02	2.24	17 464.97	0.2045	0.0418	0.0418	2	0.2163
Turku	235.5	225.98	2.08	20 467.46	0.2045	0.0418	0.0418	2	0.2088
Jyväskylä	247	247.00	2.00	20 192.89	0.1636	0.0268	0.0418	2	0.1636
Vantaa	297.5	297.50	2.00	19 270.55	0.1953	0.0381	0.0418	2	0.1953
Helsinki	310.5	66.30	9.37	50 966.40	0.2045	0.0418	0.0418	9	0.2087
Tampere	325	325.00	2.00	27 460.34	0.1537	0.0236	0.0418	2	0.1537
Oulu	416.5	327.61	2.54	19 837.43	0.2045	0.0418	0.0418	3	0.1882
Perusjoukko			**171.86**	85 273.63	0.0415	0.0017	0.0064	**170**	0.0419
			Optimi	$\sqrt{V(\hat{Y}^*)}$	$CV(\hat{Y}^*)$	$RV(\hat{Y}^*)$			

	CV-raja	RV-raja
Alueille asetetut	0.2045	0.0418
Perusjoukolle asetettu	0.0800	0.0064

Taulukko J.3. Alueiden perustiedot kiintiöintiä varten: 14 aluetta.

Alue	N_d	Kok. määrä Y_d^*	Keskiarvo $\bar{Y}_d^*$	Hajonta $S_d(y^*)$	CV $C_d(y^*)$
Porvoo	112	27 725.73	247.55	251.33	1.0153
Pirkkala	148	28 902.11	195.28	44.06	0.2256
Etelä-Savo	493	76 413.30	155.00	65.24	0.4209
Jyväskylä	494	100 049.66	202.53	56.745	0.2802
Lappi	555	66 874.78	120.50	52.23	0.4335
Kaakkois-Suomi	585	119 706.49	204.63	69.82	0.3412
Helsinki	621	289 439.24	466.09	287.58	0.6170
Satakunta-Pohjanmaa	655	155 294.60	237.09	180.52	0.7614
Radanvarsikunnat	818	155 294.40	189.85	43.14	0.2272
Kuopion seutu	871	167 293.33	192.07	44.18	0.2300
Turun seutu	958	201 540.99	210.38	47.64	0.2264
Oulun seutu	1 072	128 237.11	119.62	54.48	0.4555
Pääkaupunkiseutu	1 100	233 761.42	212.51	130.81	0.6155
Häme-Pirkanmaa	1 333	270 296.84	202.77	47.32	0.2334
Perusjoukko	9 815	2 020 830.27	205.89	133.51	0.6484

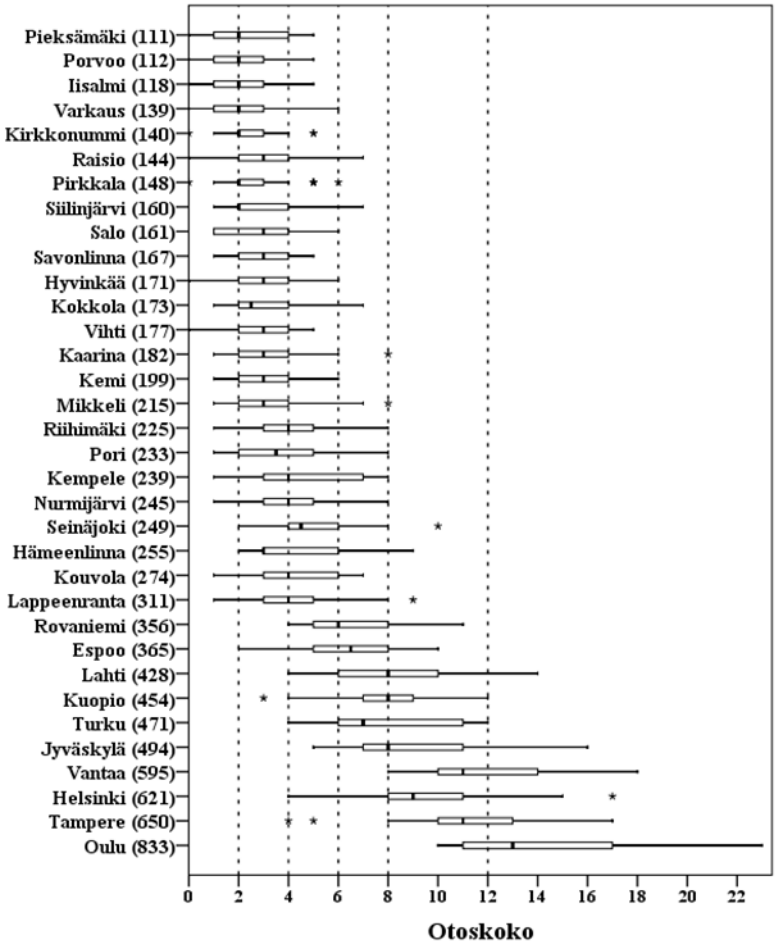
Taulukko J.4. Minimiotoskoon ratkaisu NLP-kiintiöinnissä: 14 aluetta.

Alue	k_d: n yläraja $= N_d / 2$	k_d (≥ 1)	$n_d = N_d / k_d$	$\sqrt{V(\hat{Y}_d^*)}$	$CV(\hat{Y}_d^*)$	$RV(\hat{Y}_d^*)$	RV-raja	Pyöristetty n_d	Lopullinen $CV(\hat{Y}_d^*)$
Porvoo	56	2.91	38.52	3 673.66	0.1325	0.0176	0.0176	38	0.1339
Pirkkala	74	52.04	2.84	3 829.53	0.1325	0.0176	0.0176	3	0.1289
Etelä-Savo	246.5	49.86	9.89	10 124.76	0.1325	0.0176	0.0176	10	0.1317
Jyväskylä	247	111.48	4.43	13 256.58	0.1325	0.0176	0.0176	4	0.1395
Lappi	277.5	52.85	10.50	8 860.91	0.1325	0.0176	0.0176	11	0.1294
Kaakkois-Suomi	292.5	89.23	6.56	15 861.11	0.1325	0.0176	0.0176	7	0.1282
Helsinki	310.5	29.64	20.95	38 350.70	0.1325	0.0176	0.0176	21	0.1323
Satakunta-Pohjanmaa	327.5	20.84	31.44	20 576.53	0.1325	0.0176	0.0176	31	0.1335
Radanvarsikunnat	409	279.12	2.93	20 576.51	0.1325	0.0176	0.0176	3	0.1310
Kuopion seutu	435.5	290.01	3.00	22 166.37	0.1325	0.0176	0.0176	3	0.1325
Turun seutu	479	328.98	2.91	26 704.18	0.1325	0.0176	0.0176	3	0.1305
Oulun seutu	536	91.73	11.69	16 991.42	0.1325	0.0176	0.0176	12	0.1307
Pääkaupunkiseutu	550	51.97	21.17	30 973.39	0.1325	0.0176	0.0176	21	0.1330
Häme-Pirkanmaa	666.5	430.73	3.09	35 814.33	0.1325	0.0176	0.0176	3	0.1346
Perusjoukko			**169.92**	81 787.83	0.0405	0.0016	0.0036	170	0.0405

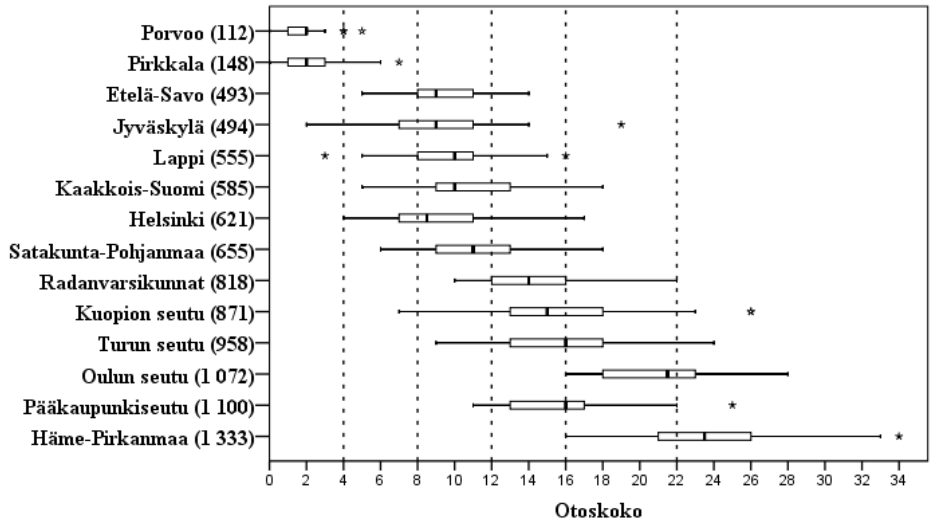
Optimi

	CV-raja	RV-raja
Alueille asetetut	0.1325	0.0176
Perusjoukolle asetettu	0.0600	0.0036

Liite K: Sijais- ja apumuuttujan aineistosta simuloitujen 1 500 otoksen EBLUP-estimointi ja 30 pienimmän MSE-keskiarvon otoksista laaditut otoskokojen jakaumat.



Kuvio K.1. Alueiden otoskokojakaumat 30 pienimmän MSE-keskiarvon mukaisissa 34 alueen otoksissa. Alueiden koot ovat suluissa.



Kuvio K.2. Alueiden otoskokojakaumat 30 pienimmän MSE-keskiarvon mukaisissa 14 alueen otoksissa. Alueiden koot ovat suluissa.