HELSINGIN YLIOPISTO

Jälkiositus

Tilastotiede
Maisterintutkielma
Yhteiskuntatilastotieteen linja

Laatija:
Antti Mattila

Ohjaajat:
Dos. Maria Valaste
Dos. Kimmo Vehkalahti

12.12.2022
Helsinki

Tiedekunta: Valtiotieteellinen

Koulutusohjelma: Tilastotiede

Opintosuunta: Yhteiskuntatilastotieteen linja

Tekijä: Antti Mattila

Työn nimi: Jälkiositus

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: Joulukuu 2022

Sivumäärä: 40+5

Avainsanat: Survey, jälkiositus, painotus, katokorjaus

Ohjaajat: Maria Valaste, Kimmo Vehkalahti

Säilytyspaikka:

Muita tietoja:

Tiivistelmä:

Survey-tutkimuksilla kerätään tietoa ympäröivästä maailmasta. Ne perustuvat usein otantaan eli siihen, että tietoa poimitaan valituilta tutkimuskohteilta. Vastaajia ja muita tietolähteitä poimitaan satunnaisesti tai jollain muulla tavalla valitsemalla. Otanta säästää tutkimuksen tekijältä aikaa ja kustannuksia, mutta tuottaa tutkimukselle epävarmuustekijöitä. Epävarmuus johtuu siitä, että poimittu otos ei vastaa tutkittavaa perusjoukkoa. Osa tutkimuskohteista jää tavoittamatta, tutkimuskohteet eivät vastaa tai vastaukset jäävät joiltain osin puutteellisiksi. Näin ollen otantatutkimus tuottaa aineistoja, joiden edustavuus ei ole täydellistä. Edustavuutta pyritään korjaamaan tilastollisilla menetelmillä, kuten painottamalla saatuja vastauksia sekä imputoimalla (paikkaamalla) tyhjiksi jääneitä vastauksia.

Tässä opinnäytteessä keskitytään otannan korjaamiseen painottamalla. Opinnäytteessä luodaan katsaus katokorjausmenetelmään nimeltään jälkiositus. Menetelmä esitellään ja siihen liittyvä konkreettinen esimerkki käydään lävitse. Konkreettinen esimerkki liittyy Risto Lehtosen ja Erkki Pahkisen oppikirjaan [Lehtonen ja Pahkinen, 2004]. Siinä ennustetaan työttömien lukumääriä ja työttömien suhteellista osuutta entisen Keski-Suomen läänin alueella. Oppikirjan esimerkki toistettiin pääpiirteissään käyttäen eri otosta kuin oppikirjassa oli käytetty. Otokseen valikoituneiden kuntien takia tulokset poikkesivat jonkin verran oppikirjassa esitetyistä tuloksista.

Jälkiosittaminen on loogista aineiston painottamista, joka tapahtuu aineiston keräämisen jälkeen. Siinä käytetään apuna lisätietoa esimerkiksi rekistereistä tai aiemmista tutkimuksista. Jälkiosittamisen historia ulottuu saman nimisenä 1970-luvulle ja eri nimisenä vähintäänkin 1940-luvulle. Jälkiosittaminen perustuu aineiston jakamiseen homogeenisiin soluihin sen jälkeen, kun aineisto on kerätty. Homogeenisten solujen käsittely vähentää epävarmuutta, jota tutkimuksen tekemiseen sisältyy. Epävarmuuden vähentyminen näkyy estimaattorien varianssien pienentymisenä.

1 Johdanto

Survey-tutkimukset perustuvat tutkimusyksiköiltä kerättyyn dataan ja muihin tietolähteisiin, kuten rekisteriaineistoihin. Datan kerääminen tutkimusyksiköiltä edellyttää, että tutkimusyksiköt voidaan tavoittaa ja että niiltä saadaan tutkimusongelmien ratkaisemissa tarvittavaa informaatiota. Hyvän aineiston luomisessa tarvitaan hyvää otanta-asetelmaa. Otanta-asetelmasta syntyvä aineisto ei kuitenkaan aina vastaa otanta-asetelmaa. Tämä johtuu toisaalta ongelmista otantakehikossa ja toisaalta vastaajien valikoitumisessa [Kish, 1992]. Otantakehikon ongelmia ovat sen vanhentuneisuus ja sen luokittelu- ja muut virheet. Vastaustapahtuma on myös herkkä lähde ongelmille, koska kysymyslomakkeet ja tavat kysyä kysymyksiä vaikuttavat vastauksiin. Lisäksi vastauskato on keskeinen aineistoa vääristävä ongelmalähde, erityisesti jos se on valikoitunutta ja suurta [Laaksonen, 2013].

Tässä tutkielmassa paneudutaan erityisesti kahteen ongelmaan saadussa vastaajien aineistossa: kehikon puutteisiin ja vastauskatoon. Näiden ongelmien ratkaisemiseen on monia lähestymistapoja, joten kaikkia niistä ei voida käsitellä tässä tutkielmassa. Keskitymmekin yhteen menetelmään, jossa vastanneiden aineiston harhaa pyritään oikaisemaan ns. jälkiosituspainoilla. Vastaavasti menetelmäaluetta kutsutaan jälkiosittamiseksi. Jälkiosittaminen edellyttää tavoiteperusjoukkotason tiedon olemassaoloa ja sen tarkoituksenmukaista soveltamista. Mikäli otanta-asetelmassa on käytetty ositteita, voi tulla sekaannusta ellemme erottele niitä käsitteellisesti jälkiositteista. Kutsumme otanta-asetelman ositteita jatkossa esiositteiksi. On hyvä huomata, että jälkiosituksessa on otettava esiositus huomioon, jos esiosittaminen on tehty. Tällöin jälkiositteet perustuvat esiositteisiin [Laaksonen, 2013].

Jälkiositus on ehkä ensimmäinen yleisesti käytetty uudelleenpainotusmenetelmä. Sitä lienee käytetty eri nimikkeillä pitkään, mutta menetelmän pioneerityöksi usein luonnehdittu Holtin ja Smithin artikkeli loi sille selvät puitteet [Holt ja Smith, 1979]. Tutkielman alkuosassa Holtin ja Smithin artikkeli on valittu lähteeksi, jota täydennetään muilla artikkeleilla, kuten Smithin artikkelilla vuodelta 1991 [Smith, 1991].

Luvussa 2 esitellään survey-tutkimuksen perusteet. Tutkielma jatkuu jälkiosituksen perusteiden esittelyllä (luku 3). Sen jälkeen käydään läpi menetelmäperhettä laajentava esittely (luku 4). Tällöin tulee esille muun muassa, että jälkiositus on myös kalibrointimenetelmä. Kalibroinnin käsite tuli laajemmin julkisuuteen 1990-luvulla [Deville ja Särndal, 1992, Zhang, 2000]. Kalibrointimenetelmät hyödyntävät tavoiteperusjoukkotason aggregaattitietoa, niin sanottuja reunajakaumia, joiden korkea laatu on kalibroinnin onnistumisen perusedellytys. On myös muita uudelleenpainotusmenetelmiä, kuten sellaisia, joissa käytetään mikrotason dataa lisäinformaationa. Tällöin lisäinformaatio on otosyksiköiden tasolta kerättyä [Kalton ja Flores-Cervantes, 2003]. Luvussa 5 esitellään jälkiositusta käyttäneitä tutkimuksia. Lukuun 6 on kerätty kritiikkiä, jota jälkiosittamista kohtaan on esitetty. Luku 7 sisältää empiirisen esimerkin jälkiosittamisesta. Esimerkki perustuu Lehtosen ja Pahkisen oppikirjaan [Lehtonen ja Pahkinen, 2004]. Luku 8 sisältää opinnäytteen yhteenvedon.

2 Survey-tutkimuksen piirteitä

Survey on empiiris-kvantitatiivinen tutkimusmenetelmä. Surveyllä pyritään tyypillisesti tuottamaan tietoa valitsemalla vastaajia, joiden ajatellaan edustavan suurempaa tutkimusyksiköiden (alkioiden) kokonaisuutta. Survey-tutkimus sisältää satunnaisuutta. Otos, joka poimitaan, säästää rahaa ja nopeuttaa tutkimuksen tekemistä, koska kaikkia mahdollisia tietolähteitä ei tutkita tiedon löytämiseksi. Otos voi poimittaessa perustua satunnaispoimintaan, mutta vastaajien muodostama joukko ei kuitenkaan tyypillisesti ole täysin satunnainen [Laaksonen, 2013, Lehtonen ja Pahkinen, 2004, Särndal, Swensson ja Wretman, 1992].

Survey-tutkimuksessa on pääsääntöisesti vastauskatoa, joka ei ole satunnaista. Vastaajiksi valikoituu tietynlaisia vastaajia, joten vastaukset painottuvat tietynlaisiin vastaajiin. Kaikki tavoitellut henkilöt eivät vastaa, tai jos vastaavatkin, osa kysytyistä kohdista jää tyhjiksi tai sisältää kelvottomia vastauksia. Näin ollen, surveyn tekijän on tarkkailtava aineistonsa laatua ja suoritettava aineistolleen korjaustoimenpiteitä, jotta aineiston tuottama tieto tutkittavana olevasta ilmiöstä on todenmukaista [Särndal, Swensson ja Wretman, 1992].

Aineiston painottaminen siten, että aliedustettuna olevat vastaajat saavat enemmän painoa vastauksilleen on keskeinen aineiston edustavuuden korjaustapa. Painotus perustuu vastaajien vastaamistodennäköisyyteen, jota korjataan käyttäen vastaajien lukumääriä, vastauksia ja rekisteritietoa vastaajista. Jälkiosittaminen on aineiston painottamista [Laaksonen, 2013].

Painottamisen ohella aineistolle voidaan tehdä *imputointeja*. Imputointi tarkoittaa aineiston puuttuvien kohtien paikkaamista loogisilla arvauksilla ja päätelmillä. Imputoimalla korjataan yleensä eräkatoa eli sitä, että vastaajan vastauksista puuttuu erä vastauksia. Teknisesti on mahdollista myös imputoida kokonaan tyhjiksi jääneitä vastauksia, jolloin imputoimalla korjataan yksikkökatoa eli kokonaisten yksiköiden puuttumista aineistosta. Imputointi perustuu joko aineiston matemaattiseen mallintamiseen tai aineiston osien kopioimiseen sopivaksi todetulta luovuttajalta, jonka vastaukset ovat sopivia puuttuvan tietokohdan korjaamiseen. Imputointi kasvattaa aineiston kokoa, koska sen avulla saadaan käyttöön aineistoa, joka olisi jäänyt käyttämättä ilman imputointia. Imputoidut arvot täytyy merkitä koodaamalla

aineistoon, jotta niitä voidaan tarvittaessa käsitellä muista arvoista erillään. Imputointia ei käsitellä tässä opinnäytteessä tämän enempää. Aineiston imputointia käsitellään laajemmin esimerkiksi Laaksosen kirjoissa [Laaksonen, 2013, Laaksonen, 2018].

2.1 Harha, varianssi ja puuttuneisuusmekanismit

Vastauskato tuottaa harhaa ja lisää estimaattorien varianssia eli epävarmuutta. Siksi vastauskadon korjaamiselle on olemassa motiivi: estimaattien epävarmuuden vähentäminen. Vastauskato voi olla yksikkötasoista, jolloin yksikön koko vastaus puuttuu. Se voi olla myös erävastauskatoa, jolloin osa yksikön vastauksesta jää puuttumaan. Erävastauskatoa syntyy esimerkiksi, kun vastaaja lopettaa kyselyyn vastaamisen kesken tai jättää vastaamatta joihinkin kysymyksiin [Laaksonen, 2013].

Vastauskato voidaan luokitella sen mukaisesti, millainen puuttuneisuusmekanismi sen tuottaa. Seppo Laaksonen esittää kirjassaan [Laaksonen, 2013] puuttuneisuusmekanismit, jotka tuottavat vastauskatoa. Perusajatus on, että puuttuneisuus on valikoitunutta. Puuttuneisuus perustuu vastaajan ominaisuuksiin, kuten hänen varallisuuteensa, ikäänsä, perhekokoonsa tai asuinpaikkaansa. Esimerkiksi köyhät ja rikkaat vastaavat kyselyihin huonommin kuin keskituloiset, joten heidän vastauksiaan jää aineistosta puuttumaan. Puuttuneisuutta kannattaa korjata, jotta aineisto vastaa paremmin tutkimuksen tavoiteperusjoukkoa.

Taulukko 1: Puutteneisuusmekanismit [Laaksonen, 2013]

Lyhenne	Kuvaus suomeksi	Kuvaus englanniksi
MNAR	Puuttuneisuus ei satunnaista	Missing not At Random
MAR	Puuttuneisuus satunnaista ehdollisesti	Missing At Random
MARS	Puuttuneisuus satunnaista otanta-asetelman puitteissa	Missing At Random Under Sampling Design
MCAR	Puuttuneisuus täysin satunnaista	Missing Completely At Random

Taulukossa 1 esitetyistä puuttuneisuusmekanismeista ainoastaan MCAR tuottaa aineiston, jolle ei tarvitse tehdä painotusta vastaajien valikoitumisen korjaamiseksi.

MCAR-tilanteessa ajatellaan, että puuttuneisuus on täysin satunnaista, eli se ei riipu mistään muuttujasta (selittäjästä tai selitettävästä) tilastollisesti merkittävissä määrin. Yleensä puuttuneisuus johtuu joko selitettävästä tai selittävästä muuttuja, joten käytännössä surveyn tekijän on reagoitava puuttuneisuuteen [Laaksonen, 2013].

MARS-tilanteessa ajatellaan, että puuttuneisuus on satunnaista esimerkiksi otantaositteen sisäisesti. Tällöin puuttuneisuutta syntyviä ongelmia voidaan korjata ja hallita painottamalla [Laaksonen, 2013].

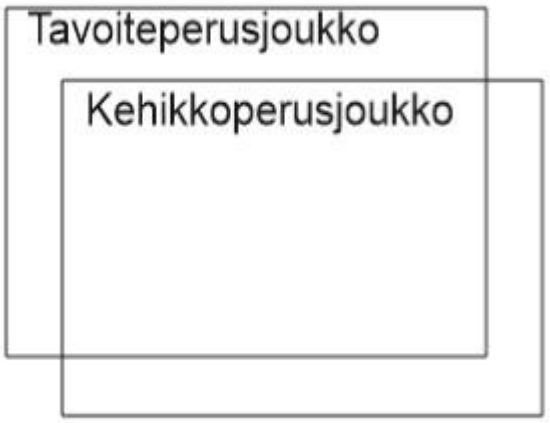
Survey perustuu otantaan, mikä tuottaa harhaa estimaatteihin. Alkioiden valitsemisen ohella muita harhanlähteitä ovat otantakehikon alipeitosta tuleva harha, aineiston keräämisen mittausvirheet, havaintojen puuttuminen vastauskadon takia sekä datan syöttöön ja käsittelyyn liittyvät virheet [Särndal, Swensson ja Wretman, 1992]. Tässä opinnäytteessä keskitytään otannasta aiheutuvaan estimaattorien harhaan.

2.2 Surveyn perusjoukot

Surveyn tekijällä on mielessään joukko ihmisiä tai muita tutkimuskohteita, joista hän on kiinnostunut. Tätä joukkoa kutsutaan *kiinnostusperusjoukoksi*. Mikäli survey perustuu vastaajien valitsemiseen, sen tulokset eivät yleisty koko kiinnostusperusjoukkoon. *Tavoiteperusjoukolla* tarkoitetaan sitä osaa kiinnostusperusjoukosta, jota pyritään tutkimaan. Tavoiteperusjoukon tutkimiseen tarvitaan *kehikkoperusjoukkoa*, eli tietoa mahdollisten vastaajien yhteystiedoista. Yhteystiedot voivat olla poimintayksiköitä (ryväs) tai suoraan keruuyksiköitä [Laaksonen, 2013, Laaksonen, 2018].

Kehikkoperusjoukko sisältää alipeittoa ja ylipeittoa. Alipeitto ja ylipeitto kuvaavat kehikkoperusjoukon ongelmia. Niitä syntyy esimerkiksi siitä, että ihmiset muuttavat, kuolevat ja syntyvät ja toisaalta siitä, että yrityksiä perustetaan ja niiden toimintoja lopetetaan. Alipeitolla tarkoitetaan esimerkiksi lapsia, joita ei ole lisätty kehikkoperusjoukkoon ja ylipeitolla puolestaan esimerkiksi yrityksiä, jotka ovat lopettaneet toimintansa. Koska kehikkoperusjoukko ei pysy ajan tasalla, sitä päivitetään. Kehikkoperusjoukkoa päivittämällä syntyy *päivitetty*

kehikkoperusjoukko. Aineisto, jonka surveyn tekijä kerää, muodostaa *tutkimusperusjoukon* [Laaksonen, 2013].



Kuva 1 Tavoiteperusjoukko ja kehikkoperusjoukko

Kuvassa 1 esitetään tavoiteperusjoukon ja kehikkoperusjoukon käsitteet. Olennaista on huomata, että käsitteet ovat osaksi päällekkäisiä ja osaksi toisistaan irrallisia. Tavoiteperusjoukon osa, jota kehikkoperusjoukko ei kata, muodostaa alipeiton. Vastaavasti osa kehikkoperusjoukkoa, joka ei sisälly tavoiteperusjoukkoon, muodostaa kehikkoperusjoukon ylipeiton [Särndal, Swensson ja Wretman, 1992]. Otos, joka surveytä varten poimitaan käyttäen kehikkoperusjoukkoa, osuu tavoiteperusjoukkoon ja ylipeiton alueelle. Alipeittoon kuuluvia alkioita ei pystytä poimimaan otokseen, koska alipeitto ei sisälly kehikkoperusjoukkoon. Alipeittoon kuuluvia alkioita ei kyetä tunnistamaan aineistoa poimittaessa [Laaksonen, 2013]. Alipeittoa ja vastauskatoa ei pystytä aina erottamaan toisistaan. Surveyn vastausprosentti voi olla korkea, vaikka alipeittoa esiintyisi runsaasti. Simuloimalla on osoitettu, että kattava kehikkoperusjoukko voi laskea kyselyn vastausprosenttia, mutta samalla kyselyn harha vähentyy [Eckman ja Kreuter, 2017].

2.3 Asetelmakerroin (deff) kuvaa otannan tehokkuutta

Asetelmakerroin (deff) määritellään kahden varianssin suhteena. Estimaattorin deff-arvo kuvastaa otannan tehokkuutta suhteessa satunnaisotantaan SRS. Deff on estimaattorikohtainen tunnusluku. Deff määritetään suhteena

$$\text{Deff}=\text{Var (Survey)} / \text{Var (SRS)} \quad (1)$$

Kaavassa 1 merkintä Var (Survey) tarkoittaa käytetyn otanta-asetelman varianssiestimaattia ja Var (SRS) puolestaan kuvastaa verrokkina olevaa satunnaisotannan varianssiestimaattia. Molemmat varianssiestimaatit on arvioitu samasta datasta [Laaksonen, 2018].

Kun deff saa arvon 1, käytetty otanta-asetelma on yhtä tehokas kuin satunnaisotanta. Arvoa 1 pienempi deff-arvo osoittaa, että käytetty otanta-asetelma on tehokkaampi kuin satunnaisotanta. Tällöin otanta-asetelma tuottaa tarkkoja estimaatteja pienemmällä otoskoolla kuin satunnaisotanta. Tyypillisesti deff saa kuitenkin arvokseen yli 1 olevia arvoja, koska otanta-asetelmaan liittyy havaintojen valikoitumista ja ryvästymistä [Kish 1965, Laaksonen, 2013, Lehtonen ja Pahkinen, 2004]. Kun deff on arvoltaan yli yhden, estimaattorin varianssi on suurempi kuin mitä se olisi satunnaisotantaa käyttämällä ollut.

Otanta-asetelman tehokkuuteen vaikuttaa se, miten aineisto jakautuu. Otannan tehokkuutta mitataan efektiivisellä otoskoolla. Efektiivinen otoskoko lasketaan jakamalla otoskoko n estimaattorin deff-arvolla:

$$n_{eff} = n \ / \ deff \tag{2}$$

Efektiivinen otoskoko osoittaa, kuinka suurella satunnaisotannan otoskoolla päästään samaan estimaattorin tarkkuuteen kuin käytetyssä otanta-asetelma päästiin [Laaksonen, 2013].

3 Jälkiosituksen perusteet

Jälkiositukseksi voidaan kutsua menetelmiä, joissa aineistoa painotetaan jälkiosituspainoilla sen jälkeen, kun aineisto on kerätty. Olkoon otos S, joka poimitaan N alkion perusjoukosta. Otos S kokoa merkataan symbolilla n. Otoksen painotusta varten selvitetään, mikä on vastaajan todennäköisyys (sisältymistodennäköisyys) tulla poimituksi otokseen. Sisältymistodennäköisyyden käänteislukuna saadaan asetelmapaino alkiolle p [Laaksonen, 2013]:

$$w_p = 1/\pi_p \tag{3}$$

missä $0 < \pi_p \leq 1$. Asetelmapaino osoittaa, kuinka montaa perusjoukon alkiota otokseen valittu alkio vastaa. Asetelmapainojen summa vastaa tavoiteperusjoukon kokoa N. Kun asetelmapainot on laskettu ja otanta on tehty, asetelmapainoja muokataan lisäinformaatiota käyttäen. Vastanneiden perusteella muokattuja painoja kutsutaan peruspainoiksi. Jälkiosittamista tehdään, jotta estimaattien laatua saadaan parannettua harhan ja varianssin vähentämisen mielessä [Little, 1993]. Jälkiosittamisen keskeinen ajatus on aineiston jakaminen tutkittavan asian kannalta homogeenisiin, toisensa poissulkeviin, ryhmiin. Jako esitetään alla matemaattisesti kaavassa (4) [Särndal, Swensson ja Wretman, 1992]. Kaavassa (4) g on ryhmän tunniste, joka käy läpi arvot 1, …, G. Olennaista on huomata, että perusjoukko jaetaan g ryhmään, jotka ovat toisensa poissulkevia ja yhdessä kattavat perusjoukon.

$$\cup_{g=1}^{G} U_g = U; \sum_{g=1}^{G} N_g = N \tag{4}$$

Ryhmiin jakamisen jälkeen voidaan tuottaa estimaatteja suhteellisista osuuksista perusjoukossa. Kun ryhmien koko ja ryhmän koon suhde otoksen kokoon tiedetään, voidaan tehdä päätelmiä ilmiön esiintymisen suhteellisista osuuksista tavoiteperusjoukosta [Smith,1991]. Jälkiosittamalla rakennetut estimaattorit ovat siis totaali- ja suhde-estimaattoreita, jotka estimoivat perusjoukon totaaleja ja suhdelukuja [Holt ja Smith, 1979].

Päätelmien tekemistä tavoiteperusjoukosta tukee tieto siitä, millaista valikoituneisuutta otokseen sisältyy. Esimerkiksi tutkittaessa kulutustottumuksia on hyvä tietää, millaiset ominaisuudet vaikuttavat ihmisten vastaamistodennäköisyyteen ja näin ollen siihen, millainen aineisto saadaan kerättyä. Kyseinen lisäinformaatio auttaa nostamaan otanta-aineiston laatua jo otantavaiheessa, koska lisäinformaation avulla otokseen voidaan etsiä ja poimia vastaajia, joiden koetaan edustavan tavoiteperusjoukkoa mahdollisimman hyvin. Otokseen poimittavia ihmisiä voidaan jakaa ja kiintiöidä ositteisiin (stratum), joista poimimalla vastaajien edustavuutta saadaan parannettua. Aineistoa analysoitaessa vastanneita voidaan painottaa sen mukaisesti, kuinka hyvin saatu otos vastaa aiemmissa tutkimuksissa ja rekisteritietolähteissä julkaistua tietoa [Laaksonen, 2013].

Tutkittaessa esimerkiksi alkoholin käyttöä koko maan tasolla on vaikeaa osittaa poimittavia ihmisiä homogeenisiin ryhmiin ennen kuin otosta on poimittu, koska ihmisten kulutustottumuksista ei ole olemassa kattavaa aineistoa, jota voitaisiin käyttää otannan lisäinformaationa [Little, 1993]. Joistain vastaajaryhmistä lisäinformaatiota voi olla, mutta se on sirpaleista, vanhentunutta, vaikeasti hankittavaa tai liian kallista käyttää. On siis mahdollista, ja jopa todennäköistä, että aineiston esiositus ei vastaa ihmisten kulutustottumuksia. Kun esiositus ei tuota laadukasta aineistoa, on aineistolle tehtävä jälkiosituksia [Laaksonen, 2013]. Jälkiositusten tekemiseen on kulutustutkimuksessa käytettävissä lisäinformaatiota, mutta mikäli lisäinformaatiota ei ole, jälkiositus voidaan tehdä käyttäen pelkkää otanta-aineistoa. Tällöin jälkiositusta kutsutaan *endogeeniseksi jälkiosituksiksi* [Breidt ja Opsomer, 2008].

Jälkiosittaminen on malliavusteista estimointia, jonka taustalla on yksisuuntainen ANOVA-malli [Särndal, Swensson ja Wretman, 1992]. Jälkiosittaminen perustuu saatuun aineistoon ja sitä tukevaan lisäinformaatioon. Saatu aineisto ehdollistaa jälkiosituksen, eli jälkiositusta tehtäessä on käytettävissä enemmän informaatiota kuin otantaa suunniteltaessa oli käytettävissä [Smith, 1991]. Informaation määrän kasvattaminen parantaa estimoinnin laatua, koska informaatio vähentää epävarmuutta.

Jälkiosittamisen pääajatuksena on aineiston jakaminen osiin, joita painotetaan siten, että painotettu aineisto vastaa tunnettuja marginaalijakaumia. Marginaalijakaumien, eli aineiston reunajakaumien, tasolle painottamisen taustalla on ajatus siitä, että

marginaalijakauma, joka tunnetaan lisäinformaation avulla, vaikuttaa estimoitavan muuttujan arvoihin. Esimerkiksi jos tiedetään, että vähemmistön suhteellinen osuus perusjoukossa on 0,20, mutta otoksessa vähemmistön edustajien suhteellinen osuus on vain 0,05, muodostetaan vähemmistön edustajille painon korjauskerroin

$w_{mi} = \frac{0,20}{0,05} = 4.$

Vastaavasti muiden kuin korjatun vähemmistön edustajien painotusta säädetään alaspäin korjauskertoimella w_p, 0 <w_p< 1. Painotusta säädettäessä minkään vastaajan lopullinen paino ei saa jäädä pienemmäksi kuin 1, koska vastaaja edustaa perusjoukossa vähintään itseään [Laaksonen, 2013]. Aineiston analyysissä käytettävät analyysipainot skaalataan siten, että niiden keskiarvoksi saadaan arvo yksi [Lehtonen ja Pahkinen, 2004].

Kun yli- tai aliedustetun vähemmistön osuus otoksessa on painottamalla korjattu marginaalijakauman tasolle, vähemmistöön kuuluvien vastaajien vaikutus estimoitaviin muuttujiin asettuu aiempaa oikeammalle tasolle kadon tuottaman harhan oikaisemisen ansiosta. Edellä esitetty vähemmistön suhteellisen osuuden korjaus voidaan tehdä myös silloin, kun vähemmistön suhteellisesta osuudesta on saatu uutta tietoa otannan toteuttamisen jälkeen. Tällöin menetelmää käytetään otantakehikossa olevan virheen oikaisemiseen [Laaksonen, 2013].

Painotuksen on todettu nostavan painotetun estimaattorin varianssia, mutta painottamista kannattaa tehdä, koska se vähentää estimaattorin harhaa [Kish, 1992]. Marginaalijakaumien toteutumisen ohella jälkiosituksella pyritään siihen, että taustamuuttujien tunnetut yhteisjakaumat toteutuvat siinä laajuudessa kuin on mahdollista ja järkevää. Jälkiositusta tehtäessä luodaan taustamuuttujan arvojen suhteen homogeeniset painotussolut ja lasketaan niihin osuvien havaintojen lukumäärät. Tyhjät painotussolut yhdistetään muihin painotussoluihin, jotta koko perusjoukko saadaan katettua painotuksella. Mikäli tyhjiä painotusssoluja ei yhdistetä muihin vastaajasoluihin, perusjoukkoon jää havaintoja, joita ei vastaa mikään otokseen poimittu alkio. Tällöin otoksesta lasketut totaalien estimaatit eivät vastaa tavoiteperusjoukon totaaleja, jollei totaaleja korjata estimoinnin jälkeen [Fuller, 1966, Gelman, 2007, Kalton ja Flores-Cervantes, 2003, Lazzeroni ja Little, 1998, Kim, Li ja Valiant, 2007]. Totaaliestimaattien korjaaminen ei ole aina

mahdollista, koska totaalien oikeita arvoja ei tiedetä, vaan ne ovat estimoitavia tunnuslukuja [Ekholm ja Laaksonen, 1991, Gelman, 2007].

Jälkiosittamisella korjataan harhaa (bias) ja suurta varianssia, joka syntyy vastauskadosta ja otoksen huonosta edustavuudesta. Vastauskato ei käytännössä koskaan ole täysin satunnaista, joten vastaamattomuus vääristää otoksesta tehtäviä päätelmiä johonkin suuntaan, jollei vastaamattomuuden vaikutuksia oikaista painottamalla vastanneita. Tyhjien solujen ohella liian suuret havaintojen painot ovat toinen ongelmalähde, jota voidaan korjata jälkiosittamalla [Smith, 1991]. Liian suuret painot johtuvat siitä, että lisäinformaation muuttujat, joiden perusteella painot muodostetaan, eivät huomioi tutkittavan muuttujan poikkeuksellisia arvoja yhtä hyvin kuin keskimääräisiä arvoja [Laaksonen, 2013].

Siirrytään seuraavaksi tarkastelemaan jälkiosittamiseen liittyviä kaavoja. SRS-otoksen (simple random sampling) ajatellaan olevan täysin satunnaisesti poimittu otos ilman palauttamista. Jälkiositusta tehtäessä SRS-otoksen ajatellaan muodostavan esiosituksen, jossa kukin alkio kuuluu omaan osittеeseensa. Seuraavat kaavat pätevät SRS-otoksen jälkiosittamiseen.

Poimitaan satunnaisotos otos s, jonka koko on n. Otoksen alkiot y indeksoidaan luvuilla 1, …, n. Alkioiden odotusarvoksi $\overline{y_s}$ saadaan [Smith, 1991]

$$\overline{y_s} = \frac{1}{n} \sum_{n=1}^{s} y_n \tag{5}$$

Otoksen alkioiden odotusarvo on populaation odotusarvon

$$\overline{Y} = \frac{1}{n} \sum_{n=1}^{N} y_n \tag{6}$$

harhaton estimaattori [Holt ja Smith, 1979].

Otoksen alkiot voidaan jakaa H ryhmään, jotka vastaavat alkioiden ryhmittymistä lisäinformaation mukaisiin homogeenisiin ryhmiin tavoiteperusjoukossa. Ryhmät voivat olla esimerkiksi ikään ja sukupuoleen perustuvia. Kullekin ryhmälle on määritelty painokerroin, joka kertoo, kuinka montaa tavoiteperusjoukon alkiota kunkin otosryhmän yksi alkio vastaa. Otosalkioiden luokittelua H ryhmään kutsutaan jälkiositukseksi [Smith, 1991]. Koska otokseen poimitaan vain osa perusjoukosta,

otoksen alkioiden keskiarvo voi vaihdella otoksittain. Vaihtelua kuvaa otantavarianssin kaava [Smith, 1991]

$$V(y)_s = \left(\frac{1}{n} - \frac{1}{N}\right) S^2, \tag{7}$$

jossa S^2 kuvaa perusjoukon varianssia

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2. \tag{8}$$

Kun otos on jälkiositettu H ositteeseen, ositteita voidaan käsitellä toisistaan riippumattomina kokonaisuuksina. Laskemalla yhteen ositekohtaisia, toisistaan riippumattomia tunnuslukuja yli ositteiden, saadaan estimaatteja perusjoukon tunnusluvuille [Holt ja Smith, 1979]. Estimoinnin kannalta on olennaista, että käytetty estimaattori on harhaton ja tarkka, eli että estimaattorin tuottama arvo vastaa perusjoukon todellista arvoa ja että estimaatin varianssi on riittävän pieni. Edellä esitetyssä otantavarianssin kaavasta nähdään, että varianssiin vaikuttavat sekä perusjoukon varianssi S^2 että sen edessä oleva kerrointermi, joka koostuu otoskoon käänteisluvun ja perusjoukon koon käänteisluvun erotuksesta [Smith, 1991].

Määritellään ositteelle h odotusarvon estimaatti [Smith, 1991]

$$\overline{y_h} = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}. \tag{9}$$

Määritellään estimaatti perusjoukon odotusarvolle, eli jälkiositettu estimaattori y_{ps}. Se saadaan laskemalla yhteen ositekohtaiset estimaatit. Aloitetaan määrittelemällä perusjoukon odotusarvo

$$\overline{Y} = \sum_{h=1}^{H} \sum_{j=1}^{N_h} \frac{y_{hj}}{N} = \sum_{h=1}^{H} W_h \overline{Y_h}, \tag{10}$$

jossa oikea puoli esittää painotettua ositteiden summaa. Kaavan oikealla puolella esiintyvä $\overline{Y_h}$ kuvaa ositteen odotusarvoa perusjoukossa ja se määritellään alkioiden summan ja ositteen koon suhteena

$$\overline{Y_h} = \sum_{j=1}^{N_h} \frac{y_{hj}}{N_h}. \tag{11}$$

Jälkiositus tuottaa otanta-aineistosta estimaattorin $\hat{y}_{ps}$, joka määritellään [Smith, 1991]

$$\hat{y}_{ps} = \sum_{h=1}^{H} w_h \, \overline{y}_h. \tag{12}$$

Näin ollen, jälkiositettu estimaattori $\hat{y}_{ps}$ on jälkiositteiden otoskeskiarvojen $\overline{y}_h$ ositekohtaisella painolla w_h painotettu summa.

Jälkiositus perustuu painojen muokkaamiseen sen perusteella, miten saatujen havaintojen lukumäärät vastaavat tavoiteperusjoukon lukumääriä. Jälkiosituksessa tuotetaan g-paino, joka saadaan jakamalla perusjoukon ositteen koko sillä lukumäärällä, mitä saatujen havaintojen painot vastaavat. G-paino lasketaan seuraavasti: g=N/$\widehat{N}$. Vastaava jälkiosituspaino w_p^* saadaan kertomalla muunnettava painomuuttuja w g-painolla

$$w_p^* = g_k * w_k. \tag{13}$$

3.1 Endogeeninen jälkiositus

Ennen aineiston keräämistä ei voida aina tietää, millaisia ryhmiä perusjoukossa on ja millä tavalla otos tulee jakautumaan ryhmiin. Jälkiositusta kutsutaan *endogeeniseksi* (sisäsyntyinen) jälkiositukseksi, jos käytetyt jälkiositteet on luotu otanta-aineistosta siten, että tietoa perusjoukon vastaavasta jaosta ei ole käytetty [Breidt ja Opsomer, 2008].

Esimerkiksi tehtäessä metsien inventointia ei voida tietää, minkä kokoisia jälkiositettua otosta vastaavat ryhmät ovat tavoiteperusjoukossa [Breidt ja Opsomer, 2008]. Näin ollen ei tiedetä havaintojen sisältymistodennäköisyyttä, mikä vaikeuttaa havaintojen painottamista tunnettujen perusjoukon marginaalijakaumien tasolle. Perusjoukon kokoa voidaan kuitenkin mallintaa, samoin kuin voidaan mallintaa havaintojen lukumäärien muutosta eri otosten välillä [Breidt ja Opsomer, 2008]. Perusjoukon tasolla kerättyä lisäinformaatiota voidaan käyttää mallintamaan otokseen saatuja havaintoja. Mallinnuksessa syntyneiden ennusteiden perusteella

otoksen havainnot luokitellaan ja luokituksesta tuotetaan endogeeninen jälkiositus [Tipton, Opsomer ja Moisen, 2013].

Endogeenisen jälkiosituksen varianssin on todettu olevan asymptoottisesti sama kuin tehtäessä jälkiositus etukäteen päätetyllä kiinteällä jälkiositteiden lukumäärällä [Breidt ja Opsomer, 2008]. Tuloksen on todistettu pätevän myös ei-parametrisessa tilanteessa, eli se pätee, vaikka muuttujien todennäköisyysjakaumasta ei tehdä oletuksia [Dahlke et al., 2013].

3.2 Arvioihin perustuva jälkiositus

Arvioihin perustuva jälkiositus (Judgment post-stratification (JPS)) on biologian alojen mittaustarpeisiin kehitetty menetelmä, joka laajentaa jälkiosituksen käyttöä järjestystietoa sisältäville aineistoille [MacEachern, Stasny ja Wolfe, 2004]. Teknisessä mielessä JPS laajentaa G.A. McIntyren esittelemää Ranked Set Sampling -menetelmää [McIntyre, 1952] jälkiosituksen aihepiirissä. Alkuperäisessä RSS -menetelmässä järjestystietoa tuotetaan poimittaviin alkioihin. JPS-menetelmän erona on se, että järjestystietoa tuotetaan poimittuihin alkioihin [Chen, 2013]. JPS-menetelmää ei pidä sekoittaa judgment sampling -menetelmään, jota käytettäessä poimitaan alkioita niistä tehtyjen arvioiden perusteella [Kish, 1992].
JPS-menetelmän tarkoituksena on tuottaa estimaatteja otanta-aineistosta, johon on lisätty arvioinnilla tai aineiston mallintamisella tuotettua järjestystietoa. Järjestystietoa voidaan tuottaa monella tavalla, esimerkiksi eri piirteitä arvioimalla, jolloin yksittäisellä havainnolla voi olla useita havainnon olemusta kuvaavia arvioita. Järjestystieto voi olla monen arvioitsijan tai eri arviointimenetelmien tuottamaa ja sellaisenaan siis epäyhtenäistä arvosteluperiaatteiltaan [MacEachern, Stasny ja Wolfe, 2004].

Järjestystiedon tuottaminen on usein halvempaa ja nopeampaa kuin tarkkojen mittausten tekeminen [MacEachern, Stasny ja Wolfe, 2004]. Esimerkiksi kasvien ja eläinten kokoja on halvempaa verrata karkeasti laadituilla järjestysluvuilla kuin mittaamalla tarkasti ja vertaamalla tarkkoja mittauksia toisiinsa [Ozturk, 2013]. JPS-menetelmien kehitystyössä on pyritty takaamaan se, että järjestystietoon perustuvalla osittamisella päästään parempiin tuloksiin kuin käyttämättä järjestystietoa, vaikka järjestystiedoissa esiintyisi tasatuloksia ja mittausvirheitä [MacEachern, Stasny ja Wolfe, 2004].

Otannan tehokkuus on mitta otannan laadulle. Tehokkuudella mitataan estimaattorien varianssien suhdetta. JPS-menetelmään perustuvan estimoinnin on todettu olevan asymptoottisesti vähintään yhtä tehokasta kuin estimointi ilman ositetietoa, eli yksinkertaiseen satunnaisotantaan (SRS-otanta), perustuva estimointi. Tätä tulosta on konkretisoitu toteamalla, että aineiston analyysi palautuu SRS-otannan analyysiksi, kun JPS-menetelmällä tuotettuja jälkiositteita ei käytetä painotuksen osana [Frey ja Feeman, 2013, MacEachern, Stasny ja Wolfe, 2004]. Sen lisäksi Frey ja Feeman ovat osoittaneet, että odotusarvon ja varianssin estimoinnin tehokkuus suhteessa SRS-otantaan kasvaa käytettäessä JPS-menetelmillä tuotettua jälkiositusta. Jälkiosituksen on perustuttava riittävän laadukkaasti tehtyyn aineiston luokitteluun tai mallinnukseen, jotta sen käytöstä on hyötyä [Frey ja Feeman, 2011]. Freyn ja Feemanin tulosta on kritisoitu. Teoreettisilla tarkasteluilla ja simulaatiokokeilla on esitetty, että JPS-estimointi voi olla SRS-estimointia tehottomampaa tilanteissa, joissa ositteet jäävät havaintojen lukumääriltään pieniksi. Kritiikki perustuu kysymykseen siitä, mikä on sopiva määrä jälkiositteita saadulle otoskoolle ja sille jakaumalle, jota tarkasteltava ilmiö noudattaa [Dastbaravarde et al., 2012].

JPS- ja RSS-menetelmien on todistettu matemaattisesti olevan erikoistapauksia eräästä menetelmien luokasta, jota Matthews ja Wolfe kutsuvat englanninkielisellä nimellä Unified Ranked Sampling (URS). URS-luokkaan on matemaattisesti todistettu kuuluvan muitakin menetelmiä, joiden pohjalta havaintojen keskinäisiin arviointeihin perustuvia jälkiosittamisen menetelmiä voidaan kehittää [Matthews ja Wolfe, 2017]. Tätä aihepiiriä ei käsitellä tässä opinnäytteessä tämän enempää.

3.3 Ositteiden luonti

Jälkiosittaminen tehdään esiositteiden sisälle käyttäen tietoa, joka on saatu aineistoa poimittaessa. Ositteita tehtäessä pyritään siihen, että ositteiden sisällä vastaajat ovat toistensa kaltaisia (homogeenisia) ja ositteiden välillä toisistaan erilaisia (heterogeenisia). Ositteiden sisäinen homogeenisuus vähentää variaatiota. Tämä nähdään kokonaisvarianssin kaavasta, jonka mukaan kokonaisvariaatio on kahden komponentin summa:

$$SST=SSB + SSW, \tag{14}$$

missä SST tarkoittaa kokonaisvarianssia, SSB ositteiden välistä varianssia ja SSW ositteiden sisäistä varianssia. Ositetun otannan käyttö vähentää variaatiota. Jälkiosittamalla pyritään vähentämään variaatiota aineiston analysointivaiheessa.

Jälkiosittaminen edellyttää vähintään yhden muuttujan käyttöä. Tyypillisesti jälkiosittavana muuttujana voidaan käyttää vastaajien ikää, sukupuolta tai asuinaluetta kuvaavaa muuttujaa [Särndal, Swensson ja Wretman, 1992]. Mikäli jälkiositus perustetaan monen muuttujan ristikkäisluokittelulle, vastaajasolujen määrä kasvaa suureksi. Tällöin on olemassa riski siitä, että painotussolujen joukkoon jää tyhjiä soluja ja soluja, joissa on vain muutamia vastaajia. Vastaajamäärältään tyhjiä tai pieniä painotussoluja pyritään välttämään. Ne vääristävät varianssia kuvaavia tunnuslukuja alaspäin ja toisaalta nostavat ennusteisiin liittyvää epävarmuutta ja voivat lisätä ennusteisiin harhaa [Fuller, 1966, Kim, Li ja Valiant, 2007, Särndal, Swensson ja Wretman, 1992].

Sopiva jälkiositteiden lukumäärä vaihtelee tilannekohtaisesti. Yleisellä tasolla on esitetty, että jälkiositteita tulisi luoda alle 10, jotta kussakin jälkiositteessa olisi riittävästi havaintoja. Jälkiosituksen osalta on todettu, että kuhunkin jälkiositteeseen pitäisi päätyä vähintään 20 havaintoa [Särndal, Swensson ja Wretman, 1992]. On todettu, että liian suuri ositteiden lukumäärä vastaa regressiopuumallinnuksen ylisovitettua mallia, jossa on liian monta parametria [Pulkkinen et al., 2018].

4 Jälkiositus suhteessa muihin menetelmiin

Jälkiosituksen kaltaisia menetelmiä ovat raking ratio- ja kalibrointimenetelmät. Kuten johdannossa todettiin, jälkiositus on terminä melko tuore. Terminologian tuoreudesta riippumatta, painojen kalibrointia on harjoitettu ainakin 1940-luvulta lähtien [Deming ja Stephan, 1940]. Raking ratio -menetelmä on kalibroinnin laajennus homogeenisille soluille [Hidiroglou ja Patak, 2006]. Jälkiosituksen suhde luvussa esitettäviin menetelmiin on se, että esiteltävät menetelmät ovat jälkiosituksen erikoistapauksia [Laaksonen, 2013].

4.1 Jälkiositus ja kalibrointi

Kalibroinnilla tarkoitetaan painotuksen korjaamista vastaamaan tiedossa olevia muuttujien marginaalijakaumia. Kalibroinnin toteuttamista varten on kehitetty esimerkiksi Calmar2 -makropaketti (CALibration on MARgins), jolla voidaan kalibroida otanta-aineisto marginaalijakaumien tasolle [Deville ja Särndal, 1992, Deville, Särndal ja Sautory, 1993]. Kalibroinnilla pidetään muutettavan painomuuttujan kokonaismassa samana, eli kalibrointi ei lisää eikä vähennä havaintojen yhteistä painomassaa. Tilanne esitetään matemaattisesti seuraavasti [Laaksonen, 2013]: Olkoon r kalibroimalla tuotetun painomuuttujan indikaattori, r= {1, …, r, …, n}, ja olkoon x perusjoukon alkioille tunnettu lisämuuttujan arvo. Tällöin

$$\sum_{r=1}^{n} w_r(cal)x_r \ = \sum_{i=1}^{N} x_i. \qquad (15)$$

Kullekin vastanneiden painotukselle on mahdollista luoda toisistaan poikkeavia kalibrointeja, kunhan kalibroinnissa tuotettu painotus toteuttaa lisäinformaation mukaiset marginaalijakaumat. Kalibrointi ei edellytä tietoa muuttujien yhteisjakaumasta, mikä lisää kalibroinnin käytettävyyttä [Laaksonen, 2013].

Kalibroinnissa käytetään etäisyysmittaa painojen lähtötason ja uusien painojen välillä. Etäisyysmittoja käyttämällä pyritään siihen, että alkuperäisiä painoja muutetaan kalibroimalla mahdollisimman vähän. Etäisyysmitat esitetään Devillen ja Särndalin artikkelissa [Deville ja Särndal, 1992].

4.2 Iteratiiviset haravointimenetelmät

Iteratiiviset haravointimenetelmät (engl. raking ja raking-ratio) ovat painotuksen adjustointimenetelmiä, joilla haravoidaan painojen kokoa askel kerrallaan. Hidiroglou ja Patak [Hidiroglou ja Patak, 2006] esittävät periaatteen, jolla haravointi (raking) toimii: Aineisto jaetaan soluihin, jotka ovat toisensa poissulkevia. Kunkin solun painoja säädetään iteratiivisesti, kunnes kaikkien solujen painojen summa vastaa lisäinformaation mukaisia marginaalijakaumia riittävän täsmällisesti. Koska menetelmä on iteratiivinen, on tärkeää, että ratkaisua haravoiva algoritmi konvergoituu. Käytännössä konvergoitumisen sujuvuuteen vaikuttavia tekijöitä ovat muuttujien lukumäärä, muuttujien luokkien lukumäärä [Battaglia et al., 2009] ja tyhjien painotussolujen esiintymisen laajuus [Fuller, 2009].

Raking-menetelmän käyttämä aineiston jakaminen soluihin on jälkiosittamista. Raking-menetelmä perustuu kalibrointiin [Deming ja Stephan, 1940]. Raking on numeerinen menetelmä, joka voidaan toteuttaa esimerkiksi SAS-ohjelmiston makrolla [Izrael et al., 2004]. Raking-menetelmää käsitellään kadonkorjauksen kannalta Bethlehemin ja Kellerin julkaisussa [Bethlehem ja Keller, 1987].

5 Jälkiositusta käyttäneitä tutkimuksia

Jälkiositusta on käytetty sekä tilastotieteen alalla että tilastotiedettä soveltavilla tutkimusaloilla. Soveltavia tutkimuksia on julkaistu esimerkiksi luonnonvaratutkimuksesta [Strunk et al., 2016], terveystutkimuksesta [Beltrán-Sánchez et al., 2013, Van der Heyden et al., 2014], politiikan tutkimuksesta [Lax ja Phillips, 2009] ja kuluttajatutkimuksesta [Yang et al., 2015]. Jälkiositusta käyttäneitä tutkimuksia on julkaistu 2000-luvulla aiempaa enemmän.

Luonnonvarojen tutkimisen osalta mainitaan jälkiosituksen sovelluskohteina erityisesti metsän inventoinnin ja metsätalouden suunnittelun menetelmät. Tyypillistä näille sovellusalueille tehdyissä jälkiosittamisissa on ollut laserkeilausaineiston (LiDAR) käyttö. LiDAR-aineistoa on mallinnettu tilastollisia menetelmiä ja maastosta kerättyä aineistoa hyödyntäen luokkamuotoon (kuten ’puustoa’, ’niittyä’) ja sen jälkeen saatua luokkajakoa on visualisoitu ruutukarttoina [Dahlke et al., 2013, Tipton, Opsomer ja Moisen, 2013].

LiDAR-aineistojen osalta ratkaistavana ongelmana on ollut kooltaan laajojen tai muulla tavalla vaikeasti läpikäytävien alueiden kartoitus päätöksentekoa varten [Næsset et al., 2013, Roberge et al., 2016, Strand ja Aune-Lundberg, 2012, Strunk et al., 2016, Tomppo et al., 2008]. LiDAR-aineistoon perustuvan jälkiosituksen on todettu parantavan estimaattorien tarkkuutta. Tarkkuuden parantuminen johtuu siitä, että tietoa ositteeseen, kuten alueeseen, kuulumisesta voidaan käyttää estimaattorin arvon ennustamisessa [Dahlke et al., 2013, Næsset et al., 2013].

Metsän inventointia on tutkittu jälkiosituksen kontekstissa myös simuloiduilla aineistoilla [Ene et al., 2013, Magnussen, Andersen ja Mundhenk, 2015]. Simuloimalla on tutkittu esimerkiksi sitä, kuinka tapa, jolla jälkiositteet luodaan, vaikuttaa jälkiositukseen perustuvien estimaattoreiden variansseihin [Magnussen, Andersen ja Mundhenk, 2015]. Ositteiden luontimenetelmien laadukkuuden tärkeyttä käsitellään esimerkiksi Crosbyn ja kumppanien artikkelissa. Kyseinen artikkeli käsittelee kehikkotiedosta johtuvien luokitteluvirheiden vaikutusta estimaattorien harhaan ja varianssiin [Crosby et al., 2017].

Muilla kuin luonnonvarojen tutkimusalueella jälkiositusta on käytetty selvemmin korjaamaan vastauskadon tuottamia ongelmia. Belgialaisessa artikkelissa, jossa

arvioidaan kadonkorjausmenetelmiä terveystutkimuksen haastatteluaineistolla, todetaan että erityisesti jälkiosittaminen vaikutti sairauksien esiintymisen yleisyydestä laadittuihin estimaatteihin [Van der Heyden et al., 2014]. Lääketieteen alalla jälkiosittamista on käytetty myös esimerkiksi arvioitaessa metabolisen oireyhtymän esiintymistä yhdysvaltalaisessa aikuisväestössä. Kyseisessä tutkimuksessa jälkiosituksella korjattiin vastauskatoa [Beltrán-Sánchez et al., 2013].

Kari Djerf on tutkinut työttömyyden tilastointia Suomessa ja siinä käytettyä jälkiositusta artikkelissaan [Djerf, 1997]. Djerf vertailee artikkelissaan jälkiosituksia toisiinsa. Vertaillut jälkiositukset perustuvat eri muuttujiin ja muuttujien toisistaan eroaviin käyttötapoihin. Djerf osoittaa, että estimaattoreiden tarkkuutta ja tehokkuutta voitiin nostaa jälkiosituksella, esimerkiksi koska jälkiositus homogenisoi vastaajia vastaajasoluihin [Djerf, 1997]. Jälkiositusta on käytetty työvoimatutkimukseen myös Norjassa [Zhang, 2000]. Phillip S.Kottin julkaisussa [Kott, 1994] jälkiositusta käytetään yritysten omistajien tutkimuksessa. Tutkimuksen otanta oli tehty ositettuna, mutta esiositusta muutettiin tekemällä jälkiositus [Kott, 1994].

Vaaliennusteiden teko on keskeinen yhteiskuntatieteiden sovelluskohde. Otantoihin perustuviin vaaliennusteisiin liittyy virhelähteitä, joita pyritään välttämään ja joiden vaikutusta pyritään korjaamaan. Yhdysvaltalaisessa tutkimuksessa osoitettiin, että jälkiosittamalla pystyttiin laatimaan kelvollisia vaaliennusteita Barack Obaman valinnalle vuoden 2012 presidentinvaaleissa. Jälkiosittamisella korjattiin Xbox-pelikonsolia käyttäneiden vastaajien valikoitumisesta aiheutuvaa harhaa siinä määrin, että aineistosta laadittu ennuste Obaman kannatukselle vastasi muilla surveymetodiikan menetelmillä laadittuja kannatusennusteita riittävän tarkasti. Tämä osoittaa, että jälkiosittamalla edullisesti kerättyä aineistoa päästiin yhtä hyviin tuloksiin kuin käyttämällä enemmän aikaa ja rahaa vaativia muita otantamenetelmiä [Wang et al., 2015].

6 Kritiikkiä jälkiosittamisesta

Jälkiosittamiselle on esitetty kritiikkiä esimerkiksi tilastollisen estimoinnin käytännön työtilanteista. Esimerkiksi metsävarojen inventointia tutkittaessa oli havaittu, että estimointi ei hyötynyt millään tavalla aineiston jälkiosittamisesta. Tämän havainnon perusteluksi esitettiin, että metsän luokittelussa käytetty muuttuja ei ollut jälkiosituksen tekemiseen sopiva [Strunk et al., 2016]. Kuten Holtin ja Smithin artikkelissa todetaan, jälkiositus ei aina tuota etua osittamattomaan otokseen verrattuna, mutta se tarjoaa mahdollisuuden nostaa surveyn laatua käyttäen informaatiota otokseen poimittujen havaintojen jakautumisesta ositteisiin [Holt ja Smith, 1979].

Kun jälkiositus tehdään käyttäen painoja, joiden koon vaihtelu (variaatio) on suurta, jälkiosittamisella syntyy pidempiä luottamusvälejä kuin muilla verratuilla menetelmillä. Syy pidempiin luottamusväleihin on se, että luottamusvälien päätepisteet riippuvat painojen variaation määrästä. Näin ollen, jos jälkiosituspainot vaihtelevat liikaa, niitä on vähintäänkin trimmattava sopivalle vaihteluvälille. Painoista pois trimmattu painomassa on siirrettävä muille havainnolle, jotta painomassan kokonaissumma pysyy ennallaan [Kish, 1992, Lazzeroni ja Little, 1998, Little, 1993]. Myös muita painojen korjausmenetelmiä on tutkittu [Kalton ja Flores-Cervantes, 2003, Lazzeroni ja Little, 1998, Vandendijck, Faes ja Hens, 2016].

Jälkiosittamisen on todettu toimivan hyvin niissä tilanteissa, joissa painotettavia havaintoja on riittävän paljon painotussoluissa. Mikäli havaintoja on vähän, jokin muu painotusmenetelmä voi olla tilanteeseen paremmin sopiva. Aineiston eri osia voidaan lisäksi painottaa eri menetelmiä käyttäen ja menetelmiä yhdistämällä. Esimerkiksi solutason painotuksen (engl. cell weighting) on todettu toimivan pienillä painotussoluilla jälkiosittamista tehokkaammin estimaattorien varianssin mielessä [Kalton ja Flores-Cervantes, 2003, Kish, 1992].

Malliperusteista (model-based) ja malliavusteista (model-assisted) estimointia on verrattu ositettuun otantaan esimerkiksi artikkelissa [McRoberts, Næsset ja Gobakken, 2013]. Ositteiden lukumäärä ja tapa, jolla ositteita rajataan voivat vaikuttaa ositetun otannan luottamusvälien pituuteen. Ositettu otanta tuotti LiDAR-aineiston kontekstissa lyhyempiä luottamusvälejä kuin SRS-otanta, mutta ositetulla otannalla luodut luottamusvälit olivat pidempiä kuin malliperusteisella ja

malliavusteisella estimoinnilla luodut luottamusvälit [McRoberts, Næsset ja Gobakken, 2013]. Tulosta luottamusvälien pituudesta on vaikeaa arvioida, koska jälkiosittamisen teko voi perustua myös niihin menetelmiin, joihin sitä oli verrattu [Dahlke et al., 2013]. Joka tapauksessa, sopiva jälkiositteiden lukumäärä vähentää estimaattoreiden varianssia, mutta liian suuren ositteiden lukumäärän on osoitettu kasvattavan varianssia [Miratrix, Sekhon ja Yu, 2013]. Varianssia minimoitaessa on pyrittävä harhattomiin tai harhaltaan riittävän vähäisiin estimaattoreihin [Tipton, Opsomer ja Moisen, 2013].

Yleisellä tasolla painotusmenetelmien valinnasta on esitetty, että lisäinformaation valinta on tärkeämpää kuin painotusmenetelmän valinta. Ajatusta on perusteltu sillä, että painotusmenetelmien tuotokset yleensä korreloivat vahvasti keskenään [Kalton ja Flores-Cervantes, 2003]. Myös Holt ja Smith esittivät, että estimointitavan on oltava sopiva siihen aineistoon, josta estimaatteja tehdään [Holt ja Smith, 1979].

Jälkiosittamista, vastausalttiusmallinnusta ja kalibrointia on vertailtu Kirchnerin ja Felderin kirjoittamassa kirjanluvussa [Kirchner ja Felderer, 2017]. Tutkimuksessa tarkasteltiin jälkiositusta, vastausalttiusmallinnusta ja kalibrointia vastauskadon oikaisun ja harhan korjaamisen menetelminä. Puhelinhaastatteluilla ja internetin kautta kerättyä palkka-aineistoa painotettiin ja verrattiin kokonaisaineistoon. Kokonaisaineistona käytettiin palkkarekisteriä. Vertailutapana oli regressiokertoimien tarkastelu. Regressiomalleja sovitettiin aineiston eri osiin, kuten luokiteltuihin ikäryhmiin. Vertailun tulos oli, että jälkiosittaminen sopii joihinkin aineiston osiin paremmin kuin muut menetelmät, mutta osassa aineistoa kalibrointi ja vastausalttiusmallinnus tuottivat lähempänä vertailuaineistona olleet ennusteet regressiokertoimille [Kirchner ja Felderer, 2017]. Ilman vertailuaineistoa, kuten rekisteriin kerättyä aineistoa, ei voida tietää, mikä painotusmenetelmä toimii parhaiten aineiston eri osissa.

Raking-menetelmää on kritisoitu esimerkiksi Amazon Mechanical Turk -sivuston käyttöä tutkivassa artikkelissa [Yang et al., 2015]. Amazon Mechanical Turk on sivusto, jolla internetin käyttäjät voivat esimerkiksi laatia survey-kyselyitä vastattavaksi tai vastata survey-kyselyihin ja saada vastaamisesta palkkioita. Artikkelissa todetaan, että raking-menetelmällä tuotetut painot voivat tuottaa huonompia estimointituloksia kuin estimointi painottamattomalla datalla. Havaintoa

perusteltiin sillä, että painotuksessa käytettävän muuttujan on korreloitava ennustettavan muuttujan kanssa riittävän vahvasti.

Lisäksi todettiin, että raking-menetelmällä tuotetut painot voivat olla hyviä oikaisemaan vastausharhaa jonkin muuttujan suhteen (esimerkiksi ikä-muuttuja), mutta samojen painojen käyttö voi lisätä harhaa jonkin toisen muuttujan suhteen (esimerkiksi rotu-muuttuja) [Yang et al., 2015].

7 Empiirinen esimerkki

Jälkiositusta kokeiltiin toistamalla esimerkki 3.1 Risto Lehtosen ja Erkki Pahkisen kirjasta [Lehtonen ja Pahkinen, 2004]. Empiirisessä esimerkissä käytetään kirjassa [Lehtonen ja Pahkinen, 2004] julkaistua Province'91 -aineistoa, jonka alkuperäinen lähde on Tilastokeskus. Aineisto koostuu 32 suomalaisesta kunnasta entisen Keski-Suomen läänin alueelta. Province'91-aineisto muodostaa opinnäytteessä käytetyn otantakehikon. Kirjassa [Lehtonen ja Pahkinen, 2004] esitetty kokonaisaineisto esitetään liitteessä 1. Province'91-aineistossa esitetään kuntien väkiluvut (POP91), työvoimaan kuuluvien lukumäärä (LAB91), työttömien lukumäärät (UE91) ja kotitalouksien lukumäärä kunnissa (HOU85). Aineiston julkaisija on tuottanut otannan jälkeen muuttujan kuntamuotoa kuvaavan muuttujan URB85, jonka arvot kuvastavat sitä, että onko kunta kaupunkimainen vai maaseutumainen [Lehtonen ja Pahkinen, 2004].

Province'91-aineistosta luotiin opinnäytettä varten 10 kuntaa sisältävä otos. Tiedot otokseen valituista kunnista esitetään taulukossa 2. SAS-kielinen lähdekoodi, jolla otos luotiin, esitetään liitteessä 2.

Taulukko 2 Otokseen valitut kunnat

Jälkiosite	Kunta	Väkiluku 1991	Työvoima 1991	Työttömät 1991	Kotitaloudet 1985
1	Jämsä	12907	6016	666	4663
1	Jämsänkoski	8118	3818	528	3019
1	Suolahti	6159	3022	457	2389
2	Joutsa	4594	2069	194	1823
2	Kannonkoski	1919	821	153	726
2	Konginkangas	1636	675	142	556
2	Konnevesi	3453	1557	201	1215
2	Korpilahti	5181	2144	239	1793
2	Laukaa	16042	7218	874	4952
2	Luhanka	1153	522	54	435

Taulukon 2 otos on jälkiositettu kahteen jälkiositteeseen, joita kuvaa jälkiosite-muuttuja. Jälkiositteet ovat aineiston tuottajan määrittelemiä. Jälkiositus perustuu

muuttujan URB85 sisältämään tietoon siitä, onko kunta kaupunkimainen vai maaseutumainen. Kaupunkimaista kuntaa kuvataan arvolla ’1’ ja maaseutumaista kuntaa arvolla ’2’ [Lehtonen ja Pahkinen, 2004].

Otos poimittiin SAS-ohjelmiston proseduurilla Proc Surveyselect. Otos toteutettiin satunnaisotantana käyttäen siemenlukuna numerosarjaa 090909. Otokseen valikoitui kolme kaupunkimaista kuntaa ja seitsemän maaseutumaista kuntaa. Kokonaisaineistossa kaupunkimaisia kuntia on seitsemän ja maaseutumaisia kuntia kaksikymmentäviisi. Asukasluvultaan suurin otoksen kunnista oli Laukaa, joka on maaseutumainen kunta. Kahta suurinta kuntaa, Jyväskylää ja Jyväskylän maalaiskuntaa, ei valittu otokseen. Otokseen valikoituneista kunnista pienin oli Luhanka 1153 asukkaallaan, joista työttöminä oli 54. Otoksessa ei ole puuttuneisuutta, koska se perustuu rekistereistä kerättyyn aineistoon. Aineistoa kuitenkin painotetaan, jotta siitä saatavat tulokset vastaisivat tavoiteperusjoukkkoa.

Työttömien suhteelliselle osuudelle laskettiin deff-arvo. Laskenta perustuu oppikirjassa esitettyihin kaavoihin. Laskenta suoritettiin käyttäen lähdekoodina kirjan [Lehtonen ja Pahkinen, 2004] web-laajennuksen training key 63:a [Lehtonen ja Pahkinen, 2004b]. Laskelman tuloksena saatiin työttömien suhteelliselle osuudelle deff-arvo 0,83. Deff on suurempi kuin mitä oppikirjaan esimerkin otokselle oli saatu käyttäen otoskokoa n=8. Oppikirjan esimerkin otoksella deff sai arvokseen 0,70. Alle yhden oleva deff-arvot kuvastavat sitä, että jälkiosittaminen tehosti otantaa suhteessa SRS-otokseen.

7.1 Kokonaisaineiston tunnuslukuja

Taulukko 3 Kokonaisaineiston tunnuslukuja

Kokonaisaineiston tunnuslukuja kunnille						
Variable	N	Odotusarvo	Hajonta	Summa	Minimi	Maksimi
Työttömät	32	471.8125	743.4029	15098	54.00000	4123
Työvoima	32	3729	6124	119325	522.00000	33786
Kotitaloudet	32	2867	4772	91753	435.00000	26881
Kuntamuoto	32	1.78125	0.42001	57.00000	1.00000	2.00000

Taulukko 3 perustuu kokonaisaineistoon. Siitä nähdään, että työttömien kokonaismäärä aineistossa on 15098 ja työvoiman kokonaismäärä on 119325. Estimaattoreita luotaessa pyritään luomaan harhattomia estimaattoreita, joiden arvot ovat mahdollisimman lähellä todellisia kokonaismääriä (taulukko 3).

Taulukko 4 Korrelaatiokertoimia

Pearson Correlation Coefficients, N = 32 Prob > |r| under H0: Rho=0				
	Työttömät	Työvoima	Kotitaloudet	Kuntamuoto
Työttömät	1.00000	0.99807	0.99667	-0.48756
		<.0001	<.0001	0.0046
Työvoima	0.99807	1.00000	0.99749	-0.46666
	<.0001		<.0001	0.0071
Kotitaloudet	0.99667	0.99749	1.00000	-0.47911
	<.0001	<.0001		0.0055
Kuntamuoto	-0.48756	-0.46666	-0.47911	1.00000
	0.0046	0.0071	0.0055	

Taulukossa 4 esitetään korrelaatioita aineiston keskeisille muuttujille. Taulukosta 4 nähdään, että kotitalouksien lukumäärä korreloi vahvasti työttömien lukumäärän ja työvoiman lukumäärän kanssa. Näin ollen, kotitalouksien lukumäärä on hyvä

muuttuja sen kanssa korreloivien muuttujien arvojen ennustamiseen. Taulukko 4 tuotettiin SAS-kielisellä lähdekoodilla, joka esitetään liitteessä 4.

7.2 Otosperusteisia tunnuslukuja

Käyttäen oppikirjan otosta, n=8, saadaan alla olevat taulukot 5 ja 6. Taulukoista nähdään, että työttömien lukumäärän ennuste oppikirjan esimerkin aineistolla on 18106 ja vastaavasti työttömyysprosentiksi on ennustettu 12,97 % (0,1297*100 % = 12,97 %).

Taulukko 5 Työttömien ja työvoiman lukumäärät [Lehtonen ja Pahkinen, 2004]

Muuttuja	Painojen summa	Muuttujan summa	Muuttujan summan hajonta std.err	Variaatiokerroin CV
Työttömiä	32	18106	7945	0.44
Työvoima	32	139548	66663	0.48

Taulukko 6 Työttömyysprosentti [Lehtonen ja Pahkinen, 2004]

Ratio Analysis: Työttömyysaste 1991							
Osoittaja	Nimittäjä	N	Painojen summa	v.asteet	Osuus työttömiä	Std Err	Varianssi
Työttömiä	Työvoima	8	32	6	0.129747	0.005697	0.000032452

Käyttäen omaa otosta, n=10, saadaan alla olevat ennusteet työttömien ja työvoiman lukumäärille (taulukko 7) sekä työttömien suhteelliselle osuudelle (taulukko 8):

Taulukko 7 Työttömien ja työvoiman lukumäärien estimaatit omassa otoksessa

Tunnuslukuja				
Muuttuja	Painojen summa	Muuttujan summa	Muuttujan summan hajonta Std err	Variaatiokerroin CV
Työttömiä	32	10484	2631	0.25
Työvoima	32	83590	22911	0.27

Taulukko 8 Työttömyysprosentin estimaatti omalla otoksella

Ratio Analysis: Työttömyysaste 1991							
Osoittaja	Nimittäjä	N	Painojen summa	v.asteet	Osuus työttömiä	Hajonta Std err	Varianssi
Työttömiä	Työvoima	10	32	8	0,125427	0,006644	0,000044139

Taulukkoja 5 ja 7 vertaamalla nähdään, että työttömien ja työvoiman lukumäärät on ennustettu opinnäytettä varten tehdyllä otoksella pienemmiksi kuin oppikirjan esimerkissä. Eroa ennusteiden välillä on työttömien lukumäärän osalta noin 42 prosenttia ja työvoiman lukumäärän osalta noin 40 prosenttia. Oppikirjan esimerkissä työttömien lukumäärä ja työvoiman lukumäärä on estimoitu todellisia arvojaan suuremmiksi. Opinnäytettä varten tehdyssä tilastoajossa vastaavat suureet estimoitiin todellisia arvojaan pienemmiksi. Tilastoajot tehtiin SAS-kielisellä lähdekoodilla, joka esitetään liitteessä 3.

Taulukoissa 6 ja 8 esitetään estimaatit työttömyysprosentille kahdessa eri otoksessa. Taulukko 6 perustuu oppikirjan otokseen ja taulukko 8 opinnäytettä varten tehtyyn otokseen. Opinnäytettä varten tehdyssä otoksessa työttömyysprosentin estimaatti jää pienemmäksi kuin oppikirjan esimerkissä. Proc Surveymeans –tulosteen perusteella työttömyysprosentiksi arvioidaan noin 12,6. Työttömyysprosentin 95 prosentin luottamusväliksi saadaan Proc Surveymeansilla väli [11,0; 14,1]. Oppikirjan esimerkissä työttömyysprosentiksi arvioitiin 12,9, mikä osuu omasta otoksesta lasketulle työttömyysprosentin luottamusvälille. Oppikirjan esimerkistä lasketulle otokselle laskettiin työttömyysprosentin luottamusväli, joksi saatiin [11,5;14,4]. Oppikirjan otoksen ja oman otoksen perusteella lasketut luottamusvälit ovat osittain päällekkäiset.

Työttömyysprosentin estimaatin jäämistä alle oppikirjan esimerkin arvojen selittää se, että Jyväskylä ja Jyväskylän maalaiskunta puuttuvat omasta otoksesta, mutta ne sisältyvät oppikirjassa käytettyyn otokseen. Jyväskylä ja Jyväskylän maalaiskunta dominoivat aineistoa, koska ne ovat aineiston suurimpia kaupunkeja. Työttömien suhteelliset osuudet olivat kaupunkimaisissa kunnissa suurempia kuin maaseutumaisissa kunnissa.

Omasta otoksesta lasketut vaihtelukertoimet jäivät pienemmiksi kuin oppikirjan esimerkissä, mikä kuvaa sitä, että omassa aineistossa kuntien työttömien lukumäärät ja työvoiman määrät vaihtelivat vähemmän kuin oppikirjan esimerkissä.

Siirrytään tarkastelemaan jälkiosituspainoja. Kaupunkikuntien g-painoiksi (katso luku 3) saadaan 7/(3*3,2)=0,73 ja maaseutukuntien g-painoiksi 25/(7*3,2)=1,12, missä 3,2 on otoskokoon 10 ja perusjoukon kokoon 32 liittyvä peruspaino, joka saadaan jakamalla perusjoukon koko otoksen koolla: 32/10=3,2. G-painojen summa vastaa otoskokoa n=10. Vastaavasti jälkiosituspainoiksi saadaan otoksen kaupunkikunnille 3,2*0,73=2,34 ja maaseutumaisille kunnille 3,2*1,12=3,58. Jälkiosituspainot summautuvat perusjoukon kooksi N=32. Tilasto-ohjelmilla laskettaessa käytetään painojen tarkempia likiarvoja, jotka on siis esitetty useammalla desimaalilla (taulukko 9).

Taulukko 9 Oman otoksen jälkiosituspainot

Kunta	Työttömät 91	Jälkiosite	Perus paino	G_POST	wstar_Post	t_y_Post	t_z_Post
Jämsä	666	1	3.2	.7292	2.333	1554.000	2.333
Jämsänkoski	528	1	3.2	.7292	2.333	1232.000	2.333
Suolahti	457	1	3.2	.7292	2.333	1066.333	2.333
Joutsa	194	2	3.2	1.116	3.571	692.857	.0000
Kannonkoski	153	2	3.2	1.116	3.571	546.429	.0000
Konginkangas	142	2	3.2	1.116	3.571	507.143	.0000
Konnevesi	201	2	3.2	1.116	3.571	717.857	.0000
Korpilahti	239	2	3.2	1.116	3.571	853.571	.0000
Laukaa	874	2	3.2	1.116	3.571	3121.429	.0000
Luhanka	54	2	3.2	1.116	3.571	192.857	.0000
Summa	3508		32	10.00	32.00	10484.476	7.000

Taulukossa 9 esitetään SAS-laskennan tulosteena g-painot ja jälkiosituspainot omalle otokselle. Jälkiosituspainoilla kerrottujen työttömien lukumäärät esitetään sarakkeessa t_y_post. Kyseisen sarakkeen sarakesumma on estimaatti työttömien lukumäärälle perusjoukossa eli entisen Keski-Suomen läänin alueella. Estimaatti jäi tasolle 10484, mikä on noin 30 % pienempi kuin työttömien todeksi tiedetty lukumäärä 15098. Tätä havaintoa selittää edelleen se, että suurimmat kaupungit Jyväskylä ja Jyväskylän maalaiskunta puuttuvat otoksesta, minkä takia työttömien lukumäärän ennuste jää omassa otoksessa todellista arvoa selvästi pienemmäksi.

8 Yhteenveto

Jälkiosituksella täsmäytetään saatuja estimaatteja vastaamaan ajan tasalla olevaa lisäinformaatiota, kuten tietoa perusjoukon marginaalijakaumista. Sillä korjataan vastauskadosta aiheutuvaa harhaa ja parannetaan estimaattoreiden tilastollisia ominaisuuksia [Holt ja Smith, 1979].

Jälkiositus voi perustua esiositukseen, mutta sen luominen on mahdollista myös pelkkää otanta-aineistoa käyttäen [Hidiroglou ja Patak, 2006, Smith, 1991].

Jälkiositusta on sovellettu moniin painotus- ja kadonkorjaustilanteisiin. Yleensä jälkiositus on toiminut hyvin, mutta painotusmenetelmiä on syytä arvioida niiden käyttötilanteiden mukaisesti. Jälkiosituksen on todettu sopivan tietynlaisiin otanta-asetelmiin paremmin kuin muihin otanta-asetelmiin [Kalton ja Flores-Cervantes, 2003].

Jälkiosituksen aktiivisena tutkimusalueena on ollut kadonkorjaus sekä estimaattorien varianssi ja harhattomuus [Hidiroglou ja Patak, 2006, Smith, 1991, Rao, Yang ja Hidiroglou, 2002]. 2000-luvulla jälkiosituksen menetelmiä on kehitetty erityisesti JPS-menetelmien osalta ja erityisesti biologisten ilmiöiden kontekstissa [MacEachern, Stasny ja Wolfe, 2004].

Jälkiositus voidaan toteuttaa esimerkiksi taulukkolaskentaohjelmalla sen jälkeen, kun peruspainot on muodostettu ja tarvittava lisäinformaatio on hankittu esimerkiksi tilastoviranomaiselta. Jälkiosituksen toteuttamiseen on olemassa ohjelmia, joilla painoja kalibroidaan marginaalijakaumien tasolle ja joilla painotusta voidaan korjata esimerkiksi numeerisia raking-menetelmiä käyttäen [Deville ja Särndal, 1992, Deville, Särndal ja Sautory, 1993, Izrael et al., 2004].

Jälkiosittamisesta annettu kritiikki on syytä huomioida valittaessa painotusmenetelmiä, koska jälkiosituksen toteuttamisesta aiheutuu kustannuksia. Kustannuksia aiheutuu esimerkiksi lisäinformaation hankkimisesta ja tilastollisten menetelmien kouluttamisesta. Jälkiosittamisesta saatavien hyötyjen täytyy olla suurempia kuin menetelmän käytöstä aiheutuvat kustannukset ovat. Muussa tapauksessa jälkiosituksen käyttämiselle ei ole olemassa taloudellisia perusteita [Kish, 1992, Lax ja Phillips, 2009]. Useissa tutkimuksissa on kuitenkin todettu, että jälkiosituksen tekeminen on nostanut sitä käyttäneen tutkimuksen tasoa.

8.1 Yhteenveto empiirisestä kokeesta

Opinnäytettä varten toistettiin oppikirjan [Lehtonen ja Pahkinen, 2004] esimerkki jälkiosittamisesta käyttäen alkuperäisestä poikkeavaa otoskokoa. Oppikirjassa esitetyt tulokset kyettiin pääosiltaan toistamaan käyttäen eri otoskokoa. Koska otokseen valikoitui eri kuntia kuin oppikirjan esimerkissä, saadut tulokset poikkesivat oppikirjassa esitetyistä tuloksista jonkin verran. Erityisesti kahden väestöltään suurimman kaupungin puuttuminen otoksesta näkyi siinä, että työttömien lukumäärän ja työttömyysprosentin ennusteet jäivät opinnäytteen otanta-aineistosta laskettuina todellista arvoaan pienemmiksi. Tätä selittää se, että entisen Keski-Suomen läänin suurimmissa kaupungeissa työttömiä oli suhteellisesti enemmän kuin maaseutumaisissa kunnissa.

Tutkimusta voidaan laajentaa jatkamalla empiiristä koetta kattamaan erilaisia otoksia. Tällöin saataisiin parempi käsitys siitä, miten otokseen valikoituneet alkiot vaikuttavat estimaattorien arvoihin.

Lähteet

[Battaglia et al., 2009] Battaglia, M., Izrael, D., Hoaglin, D., ja Frankel, M. (2009). *Practical Considerations in Raking Survey Data*. Survey Practice, 2(5).

[Beltrán-Sánchez et al., 2013] Beltrán-Sánchez, H., Harhay, M., Harhay, M., ja McElligott, S. (2013). *Prevalence and Trends of Metabolic Syndrome in the Adult U.S. Population*, 1999–2010. Journal of the American College of Cardiology, 62(8):697–703.

[Bethlehem ja Keller, 1987] Bethlehem, J. ja Keller, W. (1987). *Linear Weighting of Sample Survey Data*. Journal of Official Statistics, 3(2):141–153.

[Breidt ja Opsomer, 2008] Breidt, J. ja Opsomer, J. (2008). *Endogenous Post-Stratification in Surveys: Classifying with a Sample-Fitted Model*. The Annals of Statistics, 36(1):403–427.

[Chen, 2013] Chen, T. (2013). *Judgment Post-Stratification with Machine Learning Techniques: Adjusting for Missing Data in Surveys and Data Mining*. Väitöskirja, The Ohio State University.

[Crosby et al., 2017] Crosby, M., Matney, T., Schultz, E., Evans, D., Grebner, D., Londo, H., Rodgers, J., ja Collins, C. (2017). *Consequences of Landsat Image Strata Classification Errors on Bias and Variance of Inventory Estimates: A Forest Inventory Case Study*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(1):243–251.

[Dahlke et al., 2013] Dahlke, M., Breidt, F., Opsomer, J., ja Van Keilegom, I. (2013). *Nonparametric Endogenous Post-Stratification Estimation*. Statistica Sinica, 23(1):189–211.

[Dastbaravarde et al., 2012] Dastbaravarde, A., Arghami, N., ja Sarmad, M. (2012). *Some Theoretical Results Concerning non-Parametric Estimation by Using a Judgment Post-stratification Sample*. arXiv, 1211.4040(stat.ME).

[Deming ja Stephan, 1940] Deming, W. ja Stephan, F. (1940). O*n a Least Squares Adjustment of Sampled Frequency Table When the Expected Marginal Totals are Known*. Annals of Mathematical Statistics, 35:615–630.

[Deville ja Särndal, 1992] Deville, J.-C. ja Särndal, C.-E. (1992). *Calibration Estimators in Survey Sampling*. Journal of the American Statistical Association, 87(418):376–382.

[Deville, Särndal ja Sautory, 1993] Deville, J.-C., Särndal, C.-E., ja Sautory, O. (1993). *Generalized Raking Procedures in Survey Sampling*. Journal of the American Statistical Association, 88(423):1013–1020.

[Djerf, 1997] Djerf, K. (1997). *Effects of Post-Stratification on the Estimates of the Finnish Labour Force Survey*. Journal of Official Statistics, 13(1):29–39.

[Eckman ja Kreuter, 2017] Eckman, S. ja Kreuter, F. (2017) *The Undercoverage-Nonresponse Tradeoff*. Teoksessa Biemer, P.T. et al.: Total Survey Error in Practice, Wiley 2017, USA, p.97-113.

[Ekholm ja Laaksonen, 1991] Ekholm, A. ja Laaksonen, S. (1991). *Weighting via Response Modeling in the Finnish Household Budget Survey*. Journal of Official Statistics, 7(3):325–337.

[Ene et al., 2013] Ene, L., Næsset, E., Gobakken, T., Gregoire, T., Ståhl, G., ja Holm, S. (2013). *A simulation approach for accuracy assessment of two-phase post-stratified estimation in large-area LiDAR biomass surveys*. Remote Sensing of Environment, 133:210–224.

[Frey ja Feeman, 2011] Frey, J. ja Feeman, T. (2011). *An improved mean estimator for judgment post-stratification*. Computational Statistics & Data Analysis, 56(2):418–426.

[Frey ja Feeman, 2013] Frey, J. ja Feeman, T. (2013). *Variance estimation using judgment post-stratification*. Annals of the Institute of Statistical Mathematics, 65(3):551–569.

[Fuller, 2009] Fuller, W. (2009). *Sampling Statistics*. Wiley, USA.

[Fuller, 1966] Fuller, W. A. (1966). *Estimation Employing Post Strata*. Journal of the American Statistical Association, 61(316):1172–1183.

[Gelman, 2007] Gelman, A. (2007). *Struggles with Survey Weighting and Regression Modeling*. Statistical Science, 22(2):153–164.

[Gelman ja Little, 1997] Gelman, A. ja Little, T. (1997). *Poststratication Into Many Categories Using Hierarchical Logistic Regression*. Survey Methodology, 23(2):127–135.

[Hidiroglou ja Patak, 2006] Hidiroglou, M. ja Patak, Z. (2006). *Raking Ratio Estimation: An Application to the Canadian Retail Trade Survey*. Journal of Official Statistics, 22(1):71–80.

[Holt ja Smith, 1979] Holt, D. ja Smith, T. (1979). *Post Stratification*. Journal of the Royal Statistical Society. Series A (General). http://www.jstor.org/stable/2344652, käyty 1.11.2022.

[Izrael et al., 2004] Izrael, D., Hoaglin, D., ja Battaglia, M. (2004). *To Rake or Not To Rake Is Not the Question Anymore with the Enhanced Raking Macro*. In Proceedings of the Twenty-Ninth Annual SAS Users Group International Conference. SAS Institute, Montreal, Canada.

[Kalton ja Flores-Cervantes, 2003] Kalton, G. ja Flores-Cervantes, I. (2003). *Weighting Methods*. Journal of Official Statistics, 19(2):81–97.

[Kim, Li ja Valiant, 2007] Kim J.J., Li, J., ja Valiant R. (2007), *Cell collapsing in poststratification*. Survey Methodology 33(2):139-150, 2007.

[Kirchner ja Felderer, 2017] Kirchner, A., Felderer, B. (2017). *Wage Regression across Survey Modes: A Validation Study*. Teoksessa Biemer, P.T. et al.: Total Survey Error in Practice, Wiley 2017, USA, p.531-556.

[Kish, 1965] Kish, L. (1965) *Survey Sampling*, Wiley, USA.

[Kish, 1992] Kish, L. (1992). *Weighting for Unequal Pi*. Journal of Official Statistics, 8(2):183–200.

[Kott, 1994] Kott, P. (1994). *Reweighting and Variance Estimation for the Characteristics of Business Owners Survey*. Journal of Official Statistics, 10(4):407–418.

[Laaksonen, 2013] Laaksonen, S. (2013). *Surveymetodiikka: Aineiston kokoamisesta puhdistamisen kautta analyysiin*. Toinen painos. Ventus Publishing Aps. www.bookboon.com, käyty 1.11.2022.

[Laaksonen, 2018] Laaksonen, S (2018) Survey Methodology and Missing Data: Tools and Techniques for Practitioners, ISBN 978-3-319-79011-4, Springer.

[Lax ja Phillips, 2009] Lax, J. ja Phillips, J. (2009). *How Should We Estimate Public Opinion in the States*. American Journal of Political Science, 53(1):107–121.

[Lazzeroni ja Little, 1998] Lazzeroni, L. ja Little, R. (1998). *Random-Effects Models for Smoothing Post-Stratification Weights*. Journal of Official Statistics, 14(1):61–78.

[Lehtonen ja Pahkinen, 2004] Lehtonen, R. ja Pahkinen, E. (2004). *Practical methods for design and analysis of complex surveys*, Wiley, 2004, ISBN: 978–0-470-09163-0.

[Lehtonen ja Pahkinen, 2004b] Lehtonen, R. ja Pahkinen, E. (2004) oppikirjaan liittyvä VLISS-laboratorio, training key 63, https://vliss-trainingkeys.github.io/LehtonenPahkinen2004/Chapter3.html#training_key_63:_design_effect_and_allocation_under_stratified_sampling, käyty 1.11.2022.

[Little, 1993] Little, R. (1993). *Post-Stratification: A Modeler's Perspective*. Journal of the American Statistical Association, 88(423):1001–1012.

[MacEachern, Stasny ja Wolfe, 2004] MacEachern, S., Stasny, E., ja Wolfe, D. (2004). *Judgement Post-Stratification with Imprecise Rankings*. Biometrics, 60(1):207–215.

[Magnussen, Andersen ja Mundhenk, 2015] Magnussen, S., Andersen, H.-E., ja Mundhenk, P. (2015). *A Second Look at Endogenous Poststratification*. Forest Science, 61(4):624–634.

[Matthews ja Wolfe, 2017] Matthews, M. ja Wolfe, D. (2017). *Unified ranked sampling*. Statistics and Probability Letters, 122(C):173–178.

[McIntyre, 1952] McIntyre, G. (1952). *A Method for Unbiased Selective Sampling Using Ranked Sets*. Australian Journal of Agricultural Research, 3(4):385–390.

[McRoberts, Næsset ja Gobakken, 2013] McRoberts, R., Næsset, E., ja Gobakken, T. (2013). *Inference for lidar-assisted estimation of forest growing stock volume*. Remote Sensing of Environment, 128:268–275.

[Miratrix, Sekhon ja Yu, 2013] Miratrix, L., Sekhon, J., ja Yu, B. (2013). *Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments*. Journal of the Royal Statistical Society: Series B, 75(2):369–396.

[Næsset et al., 2013] Næsset, E., Bollandsås, O., Gobakken, T., Gregoire, T., ja Ståhl, G. (2013). *Model-assisted estimation of change in forest biomass over an 11-year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide activity data*. Remote Sensing of Environment, 128:299–314.

[Ozturk, 2013] Ozturk, O. (2013). *Combining multi-observer information in partially rank-ordered judgment post-stratified and ranked set samples*. The Canadian Journal of Statistics, 41(2):304–324.

[Pulkkinen et al. 2018] Pulkkinen, M., Ginzler, C., Traub, B., ja Lanz, A. (2018). *Stereo-imagery-based post-stratification by regression-tree modelling in Swiss National Forest Inventory*, Remote Sensing of Environment, Vol. 213, p. 182–194.

[Rao, Yang ja Hidiroglou, 2002] Rao, J., Yang, W., ja Hidiroglou, M. (2002). *Estimating Equations for the Analysis of Survey Data Using Poststratification Information*. Sankhya: The Indian Journal of Statistics, San Antonio Conference: Selected articles, 64:364–378.

[Roberge et al., 2016] Roberge, C., Wulff, S., Reese, H., ja Ståhl, G. (2016). *Improving the precision of sample-based forest damage inventories through two-phase sampling and post-stratification using remotely sensed auxiliary information*. Environmental Monitoring and Assessment, 188(4):213–.

[Smith, 1991] Smith, T. (1991). *Post-Stratification*. Journal of the Royal Statistical Society. Series D (The Statistician), 40(3):315–323.

[Strand ja Aune-Lundberg, 2012] Strand, G.-H. ja Aune-Lundberg, L. (2012). *Small-area estimation of land cover statistics by post-stratification of a national area frame survey*. Applied Geography, 32:545–555.

[Strunk et al., 2016] Strunk, J., Mills, J., Ries, P., Temesgen, H., ja Jeroue, L. (2016). *An urban forest-inventory-and-analysis investigation in Oregon and Washington*. Urban Forestry & Urban Greening, 18:100–109.

[Särndal, Swensson ja Wretman, 1992] Särndal, C.-E., Swensson, B., ja Wretman, J. (1992). *Model-Assisted Survey Sampling*. Springer, New York, USA.

[Tipton, Opsomer ja Moisen, 2013] Tipton, J., Opsomer, J., ja Moisen, G. (2013). *Properties of Endogenous Post-stratified estimation using remote sensing data*. Remote Sensing of Environment, 139:130–137.

[Tomppo et al., 2008] Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., ja Katila, M. (2008). *Combining national forest inventory field plots and remote sensing data for forest databases*. Remote Sensing of Environment, 112(5): 1982-1999.

[Van der Heyden et al., 2014] Van der Heyden, J., Demarest, S., Van Herck, K., De Bacquer, D., Tafforeau, J., ja Van Oyen, H. (2014). *Association between variables used in the field substitution and post-stratification adjustment in the Belgian health interview survey and non-response*. International Journal of Public Health, 59(1):197–206.

[Vandendijck, Faes ja Hens, 2016] Vandendijck, Y., Faes, C., ja Hens, N. (2016). *Prevalence and trend estimation from observational data with highly variable post-stratification weights*. The Annals of Applied Statistics, 10(1):94–117.

[Wang et al., 2015] Wang, W., Rothschild, D., Goel, S., ja Gelman, A. (2015). *Forecasting elections with non-representative polls*. International Journal of Forecasting, 31(3):980–991.

[Yang et al., 2015] Yang, H.-C., Donovan, S., Young, S., Greenblatt, J., ja Desroches, L.-B. (2015). *Assessment of household appliance surveys collected with Amazon Mechanical Turk*. Energy Efficiency, 8(6):1063–1075.

[Zhang, 2000] Zhang, L.-C. (2000). *Post-Stratification and Calibration - A Synthesis*. The American Statistician, 54(3):178–184.

Liitteet

Liite 1: Province'91 aineiston kuvaus ja aineisto [Lehtonen, Pahkinen, 2004].

Osite= 1=kaupunki, 2=muu kuin kaupunki

Ryväs = tieto rypääseen kuulumisesta

Havainto = Havainnon numero

Kunta = Kunnan nimi

Väestö 91= Väestö 1991 (Tilastokeskus)

Työvoima 91 = Työvoima 1991 (Tilastokeskus)

Työttömät 91 =Työttömät 1991 (Tilastokeskus)

Kotitaloudet 85 = Kotitaloudet 1985 (Tilastokeskus)

Kuntamuoto 85 = Kuntamuoto 1985, missä 1= Kaupunki, 2 maaseutu

PROVINCE'91 AINEISTO

Osite	Ryväs	Havainto	Kunta	Väestö 91	Työvoima 91	Työttömät 91	Kotittalou det 85	Kuntamuo to 85
1	1	1	Jyväskylä	67200	33786	4123	26881	1
1	2	2	Jämsä	12907	6016	666	4663	1
1	2	3	Jämsänkoski	8118	3818	528	3019	1
1	2	4	Keuruu	12707	5919	760	4896	1
1	3	5	Saarijärvi	10774	4930	721	3730	1
1	5	6	Suolahti	6159	3022	457	2389	1
1	3	7	Äänekoski	11595	5823	767	4264	1
2	5	8	Hankasalmi	6080	2594	391	2179	2
2	6	9	Joutsa	4594	2069	194	1823	2
2	7	10	Jyväskmlk	29349	13727	1623	9230	2
2	4	11	Kannonkoski	1919	821	153	726	2
2	4	12	Karstula	5594	2521	341	1868	2
2	8	13	Kinnula	2324	927	129	675	2
2	8	14	Kivijärvi	1972	819	128	634	2
2	3	15	Konginkangas	1636	675	142	556	2
2	5	16	Konnevesi	3453	1557	201	1215	2
2	1	17	Korpilahti	5181	2144	239	1793	2
2	2	18	Kuhmoinen	3357	1448	187	1463	2
2	4	19	Kyyjärvi	1977	831	94	672	2
2	5	20	Laukaa	16042	7218	874	4952	2
2	6	21	Leivonmäki	1370	573	61	545	2
2	6	22	Luhanka	1153	522	54	435	2
2	7	23	Multia	2375	1059	119	925	2
2	1	24	Muurame	6830	3024	296	1853	2
2	7	25	Petäjävesi	3800	1737	262	1352	2
2	8	26	Pihtipudas	5654	2543	331	1946	2
2	4	27	Pylkönmäki	1266	545	98	473	2
2	3	28	Sumiainen	1426	617	79	485	2
2	1	29	Säynätsalo	3628	1615	166	1226	2
2	6	30	Toivakka	2499	1084	127	834	2
2	7	31	Uurainen	3004	1330	219	932	2
2	8	32	Viitasaari	8641	4011	568	3119	2

Liite 2: Sas-kielinen lähdekoodi oman otoksen luontia varten

```
libname gradu 'z:/gradu';

proc surveyselect data=gradu.Province91 out=gradu.OtosProvince91 n=10
method=SRS seed=090909 noprint;
title1 "Otos province91:sta";
run;
```

Liite 3: Sas-kielinen lähdekoodi Proc Surveymeans –ajoja varten

```
libname gradu 'z:/gradu';

data gradu.Province91;
input Str Clu Id LABEL $ 10-22 POP91 LAB91 UE91 HOU85 URB85;
datalines;
1 1  1 Jyväskylä    67200 33786  4123 26881  1
1 2  2 Jämsä       12907  6016   666  4663  1
1 2  3 Jämsänkoski  8118  3818   528  3019  1
1 2  4 Keuruu      12707  5919   760  4896  1
1 3  5 Saarijärvi  10774  4930   721  3730  1
1 5  6 Suolahti     6159  3022   457  2389  1
1 3  7 Äänekoski   11595  5823   767  4264  1
2 5  8 Hankasalmi   6080  2594   391  2179  2
2 6  9 Joutsa       4594  2069   194  1823  2
2 7 10 Jyväskmlk   29349 13727  1623  9230  2
2 4 11 Kannonkoski  1919   821   153   726  2
2 4 12 Karstula     5594  2521   341  1868  2
2 8 13 Kinnula      2324   927   129   675  2
2 8 14 Kivijärvi    1972   819   128   634  2
2 3 15 Konginkangas 1636   675   142   556  2
2 5 16 Konnevesi    3453  1557   201  1215  2
2 1 17 Korpilahti   5181  2144   239  1793  2
```

```
2  2  18  Kuhmoinen    3357  1448   187  1463  2
2  4  19  Kyyjärvi     1977   831    94   672  2
2  5  20  Laukaa      16042  7218   874  4952  2
2  6  21  Leivonmäki   1370   573    61   545  2
2  6  22  Luhanka      1153   522    54   435  2
2  7  23  Multia       2375  1059   119   925  2
2  1  24  Muurame      6830  3024   296  1853  2
2  7  25  Petäjävesi   3800  1737   262  1352  2
2  8  26  Pihtipudas   5654  2543   331  1946  2
2  4  27  Pylkönmäki   1266   545    98   473  2
2  3  28  Sumiainen    1426   617    79   485  2
2  1  29  Säynätsalo   3628  1615   166  1226  2
2  6  30  Toivakka     2499  1084   127   834  2
2  7  31  Uurainen     3004  1330   219   932  2
2  8  32  Viitasaari   8641  4011   568  3119  2
;
run;


* Tehdään n=10 kunnan otos gradu.province91 -datasta;

proc surveyselect data=gradu.Province91 out=gradu.OtosProvince91 n=10
method=SRS seed=090909 noprint;
title1 "Otos province91:sta";
run;

*Lisätään otokseen asetelmapaino SW: Perusjoukon koko / otoksen koko;

data gradu.OtosProvince91;
set gradu.OtosProvince91;
SW=32/10;
* Asetelmapaino ;
run;
```

```
* Ositteiden koodit ovat: 1: kaupunkimainen kunta 2: muu kuin kaupunkimainen
kunta;


data poststrata;
input URB85 _PSTOTAL_ ;
datalines;
1 7
2 25
;
run;


proc surveymeans data=gradu.otosprovince91 mean  sum all sumwgt cv cvsum std
varsum; *Sum laskee totaalin;
strata URB85;
weight SW; *Tämän on oltava: perusjoukon koko 32 / otoskoko;
var UE91 LAB91;

ratio 'Työttömyysaste 1991' UE91 / LAB91;
poststrata URB85 / pstotal=poststrata out=gradu.psweight;
* jälkiosituspainot viedään tiedostoon gradu.psweight;
run;
```

Liite 4: SAS-kielinen lähdekoodi korrelaatioiden laskemista varten

```
* Muuttujien korrelaatioista taulukko;

proc corr data=gradu.Province91;
title "Proc Corr korrelaatioita";
var ue91 lab91 hou85 urb85;
run;
```