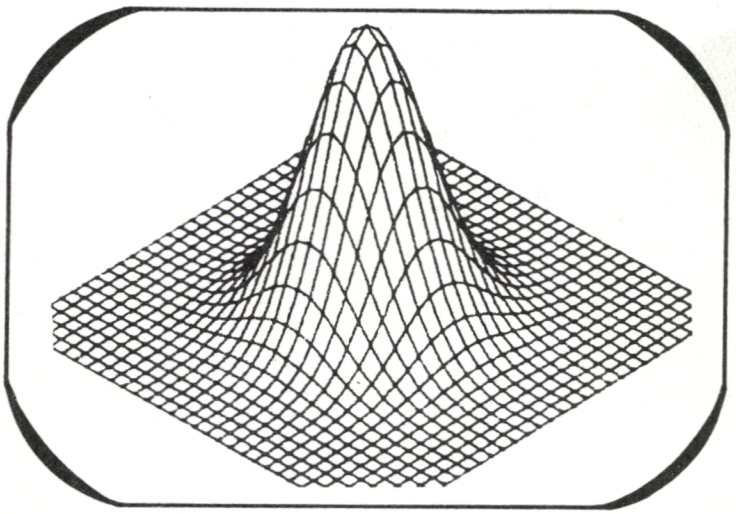
METSÄNTUTKIMUSLAITOKSEN
TIEDONANTOJA 86
Matemaattinen osasto

LEIMIKON PUUSTON TILAVUUDEN ARVIOINTI REGRESSIOENNUSTINTA KÄYTTÄEN

Timo Pekkonen



Helsinki 1983

Metsäntutkimuslaitoksen tiedonantoja 86

Timo Pekkonen

LEIMIKON PUUSTON TILAVUUDEN ARVIOINTI REGRESSIOENNUSTINTA KÄYTTÄEN

HELSINKI 1983

Helsinki 1983
ISSN 0358-4283
Valtion painatuskeskus

Alkusanat

Leimikoiden pystymittausmenetelmällä arvioidaan tietyllä alueella pystyssä olevasta puustosta, leimikosta, kauppahinnan ja hakkuupalkkojen määräämiseksi tarvittavat tunnukset. Menetelmä on kehitetty 1960-luvun lopulla. Sitä käyttäen mitataan nykyisin noin 14 miljoonaa kiintokuutiometriä runkopuuta vuodessa. Määrä vastaa runsasta kolmannesta teollisuuden käyttämästä kotimaisesta raakapuusta. Tässä tutkimuksessa esitetään uusi puuston kokonaistilavuuden laskentamenetelmä sekä sen tilastolliset ominaisuudet. Tutkimus on syntynyt osana pystymittausmenetelmän kehittämistyötä.

Tutkimus on tehty Metsäntutkimuslaitoksessa, osana metsänarvioimistieteen tutkimusosaston ja matemaattisen osaston yhteisprojektia. Esitän parhaat kiitokseni työtovereilleni, jotka ovat lukeneet aikaisemmat menetelmää koskevat käsikirjoitukseni ja kommenteillaan edesauttaneet työn täsmentymistä. Erityisesti haluan kiittää Jouko Laasasenahoa siitä luottamuksesta ja tuesta, jota hän työtäni kohtaan on osoittanut sekä Jaakko Heinosta niistä lukuisista keskusteluista, jotka ovat auttaneet yli ajattelun salakuoppien ja jotka ovat vaikuttaneet asioiden painotukseen tutkimuksessa.

Käsikirjoituksen ovat lukeneet myös Esa Läärä, Seppo Mustonen, Erkki Nenonen, Risto Seppälä ja Timo Teräsvirta. Teräsvirran esittämät lukua 3 koskevat kommentit ovat olleet erityisen hyödyllisiä. Ne ovat selventäneet ennakkoinformaation hyödyntämisessä mahdollisten lähestymistapojen eroja ja niiden perusteella olen voinut täsmentää omaa tutkimusongelmasta lähtevää ajattelutapaani. Kiitän kaikkia käsikirjoituksen lukeneita heidän esittämistään arvokkaista huomautuksista. Olen pyrkinyt ottamaan ne huomioon lopullisessa käsikirjoituksessa.

Helsingissä 29.9.1982

Timo Pekkonen

Sisällys

Tiivistelmä

Tutkimuksessa tarkastellaan pystymitatun leimikon runkopuun kokonaistilavuuden määräämistä. Lähtökohtana on rungon tilavuuden ja läpimitan välistä riippuvuutta kuvaava polynomin muotoinen superpopulaatiomalli, jonka virhetermin varianssin oletetaan olevan verrannollisen läpimitan tunnettuun potenssiin. Tilavuuden arvioinnissa käytetään regressioennustinta. Ennustimen ominaisuuksia on tarkasteltu sekä teoreettisesti että simulointikokein. Ennustimelle ja sen mallin suhteen lasketulle keskineliövirheelle on johdettu yksinkertaistuskaavat kahden läpimitan suhteen kiintiöidyn otoksen tapauksessa. Otokset ovat analogisia ns. suhteellisesti ja Neymannin mukaisesti kiintiöidyille otoksille. Yksinkertaistukset edellyttävät, että ennustimessa on mukana termit, jotka ovat verrannolliset virhetermin hajontaan ja varianssiin. Tällöin jälkimmäinen otos minimoi ennustimen mallin suhteen lasketun keskineliövirheen. Molemmissa otoksissa ennustin on harhaton riippumatta mallin mahdollisesta virheellisestä määrittelystä. Tuloksiin perustuen käytännön otantamenetelmäksi silloin, kun kiintiöintiä ei voida suorittaa, suositellaan menetelmää, jossa poimintatodennäköisyydet ovat verrannolliset mallin virhetermin hajontaan. Tällainen menetelmä tuottaa usein otoksia, jotka ovat lähes optimaalisia.

Tutkimuksen toisessa osassa tarkastellaan inventoitiaineistoihin sisältyvän keskimääräisen ennakkoinformaation hyödyntämistä leimikon tilavuuden määräämisessä. Tätä varten määritellään ennustin, jossa ennakkoinformaatio yhdistetään otosinformaation kanssa tietyllä painolla. Ennustimen käytön ongelmana on määrätä, miten ennakkoinformaatiolla täydennetään otosinformaation mahdollisia puutteita ja mikä on ennakkoinformaation paino. Ongelmaan esitetään osaksi teoreettisiin tarkasteluihin ja osaksi intuitioon perustuva käytännön ratkaisu. Sen mukaisesti paino riippuu ennakkoinformaation tarkkuudesta sekä otoksen läpimittajakaumasta eli otoksen onnistumisesta mallin suhteen lasketulla keskineliövirheellä mitattuna. Menetelmää havainnollistetaan simulointikokeilla. Niiden mukaisesti ennakkoinformaation käytön hyöty on suurin pienillä otoksilla. Menettely suojaa poikkeavilta ennusteilta, jotka johtuvat otoksen epäonnistuneesta läpimittajakaumasta.

Tutkimuksen ensimmäinen osa pohjaa pääosin Royallin tekemään työhön (kts. esim. JASA 1973, 68. 880-889). Toisen osan pääasiallisena lähteenä on Teräsvirran artikkeli (Scand.J.Stat. 1982, 8. 33-38).

1. Johdanto

1.1. Nykyinen pystymittausmenetelmä

Pystymittauksen tavoitteena on arvioida leimikon runkopuun kokonaistilavuus sekä sen jakautuminen puutavaralajeihin. Menetelmä on kuvattu tutkimuksessa Nousiainen ym. (1972). Arviointia varten leimikon rungot jaetaan luonnollisiin ositteisiin, runkolajeihin, puulajin ja rungoista pääasiassa syntyvien puutavaralajien mukaisesti. Syntyviin puutavaralajeihin, joita ovat tyypillisesti sahatukit, pylväät, kuitupuu ja hukkapuu, vaikuttavat puulaji, rungon koko, muoto ja mahdollinen vikaisuus.

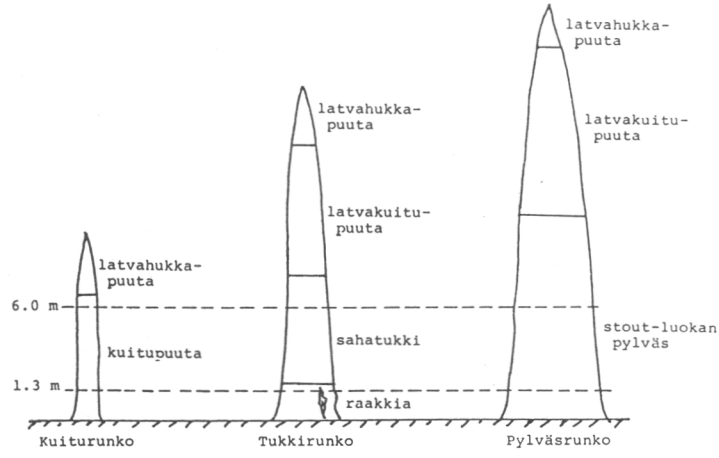
Arviointia varten luetaan kaikki leimikon rungot ja mitataan koepuut. Puiden luvun yhteydessä mitataan jokaisen leimikon rungon läpimitta 1.3 metrin korkeudelta, ns. rinnankorkeusläpimitta. Näin saadaan runkojen jakautuminen 2 cm:n läpimittaluokkiin runkolajeittain eriteltynä.

Koepuiden avulla estimoidaan eri läpimittaluokkien runkojen keskitilavuus ja rungoista syntyvät puutavaralajiosuudet. Koepuiden poimintaa varten leimikko jaetaan 1-4 koepuualueeseen siten, että koepuualueen mitattava puusto on silmämääräisesti arvioiden mahdollisimman tasasuhtainen. Kultakin koepuualueelta poimitaan otannalla omat koepuut, joiden tietoja käytetään eri koepuualueilla toisistaan riippumatta. Otanta tehdään joko ryhmittelemällä läpimittaluokat ja määräämällä otantasuhteet ryhmittäin tai ns. relaskooppia hyväksi käyttäen. Otantasuhteet läpimittaluokkaryhmissä määrätään siten, että ne kasvavat läpimittojen suuretessa. Relaskooppia käytettäessä koepuiden poimintatodennäköisyydet ovat verrannolliset läpimitan neliöön.

Jokaisesta koepuusta mitataan rinnankorkeusläpimitta, läpi-

mitta 6 metrin korkeudelta ja pituus sekä määrätään runkolaji ja saatavat puutavaralajit. Näiden tietojen avulla saadaan taulukoista koepuun runkotilavuus sekä puutavaralajien määrät. Kuvassa 1 on esimerkki eri mäntyrunkolajien koepuista syntyvistä puutavaralajeista.

Koepuiden tiedot yleistetään koepuualueen rungoille ositekeskiarvojen perusteella. Koepuiden otantaa ei kustannussyistä voida toteuttaa siten, että jokaiseen läpimittaluokkaan välttämättä tulisi koepuita. Mikäli koepuita ei ole sattunut johonkin läpimittaluokkaan, tarvittavina keskiarvoina käytetään saman puulajin toisen runkolajin vastaavia keskiarvoja tai, mikäli niitäkään ei ole voitu laskea, niin taulukkoarvoja.



Kuva 1: Esimerkki eri mäntyrunkolajien koepuista syntyneistä puutavaralajeista.

12. Menetelmän kehittämistavoitteet

Kun pystymittausmenetelmää lähdettiin kehittämään edelleen, tavoitteena oli toisaalta käytettyjen taulukoiden korvaaminen niitä joustavammilla funktioilla (vrt. Laasasenaho 1982). Toisaalta tavoitteena oli nykyistä tehokkaamman koepuutietojen yleistysmenetelmän kehittäminen, jolloin koepuumääriä vähentämällä saataisiin kustannussäästöjä. Mittausmenetelmän perusteita, kuten mitattavia tunnuksia, ei siinä vaiheessa analysoitu, vaan niiden osalta kehitystyö jätettiin jatkotutkimusten varaan. Tästä syystä tässä tutkimuksessa keskitytään pelkästään koepuiden otantamenetelmän ja koepuutietojen yleistysmenetelmän tehostamiseen.

Mittausmenetelmän kokonaiskustannuksista muodostavat puidenluvun ja koepuiden mittauskustannukset valtaosan. Metsähallituksen selvityksen (Koeseloste, 1979) mukaan puidenluvun kustannukset ovat keskimäärin 1,51 mk ja koepuiden mittauskustannukset keskimäärin 0.52 mk mitattavaa kuutiometriä kohden. Koepuiden mittauskustannukset olivat siten keskimäärin noin 34 % varsinaisista mittauskustannuksista. Kustannukset vaihtelivat huomattavasti leimikoittain lähinnä leimikoiden järeydestä johtuen. Kun vuosittain menetelmällä mitataan puuta noin 14 milj. m^3, niin esitettyjä arvoja käyttäen saadaan koepuiden mittauskustannuksiksi vuosittain noin 7,3 milj. markkaa. Arvio on karkea, mutta antaa kuvan niiden kustannusten suuruusluokasta, joita tutkimuksessa esitettävin menetelmin on tarkoitus vähentää.

Edellisessä kappaleessa mainittiin, että koepuita ei voida taata jokaiseen runkojen läpimittaluokkaan. Tämä tyhjien luokkien ongelma koskee erityisesti läpimittajakauman laidoilla olevia luokkia, joissa runkoja on vähän, sekä vähämerkityksisiä sivurunkolajeja. Uudelta menetelmältä toivottiin ongelmaan entistä parempaa ratkaisua. Yleisesti ottaen menetelmän tulee toimia koepuumäärästä riippumatta. Sen

tulee antaa järkeviä tuloksia pienillä koepuumäärillä tai jopa silloin, kun koepuita ei ole lainkaan. Toisaalta menetelmän tulee tehokkaasti käyttää hyväkseen koepuiden sisältämä informaatio.

Menetelmää kehitettäessä tavoitteeksi asetettiin myös menetelmän tilastollisten ominaisuuksien, erityisesti tulosten luotettavuuden selvittäminen. Näin saadaan perusteet tarvittaville koepuumäärille. Nykyisen menetelmän tilastomatemaattisia perusteita ei ole tarkoin selvitetty, vaan ohjeet koepuumääristä on laadittu 'varman päälle'. Käytännössä tämä johtaa 'yliluotettavuuteen', josta tinkimällä on mahdollista saada kustannussäästöjä.

Tilastolliselta kannalta tarkasteltuna pystymittaustoiminta voidaan nähdä eri leimikoilla tehtävinä otantatutkimuksina. Näin ajatellen saadaan menetelmän kehittämiselle tietyt lähtökohdat. Ensinnäkin on selvitettävä millaisia ovat perusjoukot, joissa otantatutkimuksia suoritetaan. Toisin sanoen on selvitettävä ne leimikoiden ominaisuudet, jotka vaikuttavat otantaan ja tulosten luotettavuuteen. Tämä johtaa leimikoiden yleisten ominaisuuksien kuvaamiseen ns. superpopulaatiomallilla. Superpopulaatiomalli muodostaa tutkimuksen peruslähtökohdan. Toiseksi leimikoiden perusjoukosta on olemassa tiettyä määrällistä ennakkoinformaatiota. Tutkimuksessa selvitetään, miten tätä tietoa voidaan hyödyntää tietyn leimikon puustotunnuksia määritettäessä sekä pohditaan ennakkoinformaation käytön merkitystä. Kolmas lähtökohta liittyy mittausmenetelmän hyvyyskriteeriin. On luonnollista vaatia, että menetelmä toimii keskimäärin hyvin eri leimikoissa. Tämän mukaisesti voidaan tietylle leimikolle sallia harhaisia arvioita, kunhan harha todennäköisesti on riittävän pieni.

Tutkimuksessa selvitetään polynomiseen regressiomalliin perustuvan menetelmän ominaisuuksia ja käyttömahdollisuuksia

pystymittauksessa. Menetelmää verrataan myös nykyisin käytössä olevaan ositekeskiarvoihin perustuvaan menetelmään. Menetelmien vertailu selventää osaltaan uuden menetelmän teoreettisia ominaisuuksia ja antaa pohjaa menetelmän ominaisuuksien tulkinnalle. Superopulaatiomalliin liittyvien oletusten oikeellisuutta on testattu käytännön leimikkoaineistossa. Näin on saatu kuva menetelmän odotettavissa olevasta toimivuudesta käytännössä.

Tutkimuksessa tarkastellaan yksinkertaistettua ongelmaa. Rajoitutaan tarkastelemaan yhden runkolajin runkojen kokonaistilavuuden ennustamista. Rajoittuminen on luonnollinen, sillä puuston kokonaistilavuuden arviointi voidaan nähdä ensisijaisena ongelmana, jonka jälkeen esimerkiksi puutavaralajimäärien arviointi yksinkertaistuu niiden osuuksien määräämiseksi.

2. Puuston tilavuuden ennustaminen

21. Superpopulaatiomalli

Tarkastellaan yhden runkolajin leimikon kokonaistilavuuden määräämistä. Merkitään puiden läpimittaa symbolilla x ja runkotilavuutta symbolilla y. Oletetaan, että jokaisesta leimikon puusta tunnetaan läpimitat x_i, i=1,...,N ja että n:stä koepuusta, otoksesta s, tunnetaan myös runkotilavuus y_i, $i \in s$. Yksinkertaisuuden vuoksi oletetaan, että koepuiden tilavuus ei sisällä virhettä. Näitä merkintöjä käyttäen ongelmana on kokonaistilavuuden

$$T = \sum_{1}^{N} y_i \qquad (1)$$

arviointi. Oletetaan edelleen, että runkotilavuudet y_i ovat satunnaisprosessin tuottamia eli ovat satunnaismuuttujia. Tällöin myös kokonaistilavuus T on satunnaismuuttuja, ongelmana on sen arvon ennustaminen. Puhetapaa ennustamisesta käytetään tutkimuksessa erotuksena estimoinnista, jossa tavoitteena on arvion määrääminen ei-satunnaismuuttujalle.

Puun tilavuuden y_i tuottava satunnaisprosessi H oletetaan määritellyksi k. asteen polynomilla

$$H: \quad y_i = \beta_o+\beta_1 x_i+\ldots+\beta_k x_i^k+\varepsilon_i, \quad i=1,\ldots,N \qquad (2)$$

Parametrit β_i, i=1,...,k oletetaan leimikkokohtaisiksi tuntemattomiksi vakioiksi, virhetermit ε_i riippumattomiksi tuntemattomiksi satunnaismuuttujiksi, joiden odotusarvo on 0 ja varianssi $\sigma^2 x_i^{2g}$, merkitään lyhyesti $\varepsilon_i \sim (0,\sigma^2 x_i^{2g})$. Varianssin kerroin σ^2 oletetaan tuntemattomaksi vakioksi, eksponentin 2g arvo sensijaan oletetaan tunnetuksi.

Superpopulaatiomallin muodon valinta perustuu havaintoaineistoihin. Käytännössä käytetään mallia, jossa rungon ns.

muotokorkeuden, $y/(0.25 \cdot \pi \cdot x^2)$, riippuvuus läpimitasta on kuvattavissa parabelilla. Tällöin polynomissa on 2., 3. ja 4. asteen termit. Polynominmuotoinen malli tarjoaa myös mallin riittävän joustavuuden ja toisaalta teoreettisten tarkastelujen yksinkertaisuuden. Virhetermin hajonnan määräävänä vakiona g käytetään arvoa 2. Tämä on sopusoinnussa Laasasenahon (1982) tuloksen kanssa, jonka mukaan runkotilavuuden ehdollinen variaatiokerroin on läpimitasta lähes riippumaton vakio, eli käytännössä g:n arvo on välillä 2...3.

Mallia (2) kutsutaan leimikoiden superpopulaatiomalliksi. Se määrittelee kuvitellun perusjoukon, josta poimittuja otoksia todellisten leimikoiden ajatellaan olevan. Superpopulaatiomallien käyttäytymistä eri leimikoilla, esimerkiksi kerroinvektorin β jakaumaa ei yleisesti ole tutkittu. Sen sijaan on laadittu yleisiä tilavuusfunktioita (vrt. Laasasenaho, 1982), joita voidaan pitää leimikkokohtaisten mallien (2) odotusarvona. Luvussa 3 esitetään, miten näiden sisältämää ennakkoinformaatiota voidaan hyödyntää leimikkokohtaisia malleja estimoitaessa.

22. Runkotilavuuden ennustin

Jaetaan kokonaistilavuus T komponentteihin Royall'in (1970) esittämällä tavalla.

$$T = \sum_s y_i + \sum_{\bar{s}} y_i \qquad (3)$$

jossa symbolilla Σ_s merkitään summausta yli koepuiden ja symbolilla $\Sigma_{\bar{s}}$ summausta yli ei-koepuiden. Otoksen poimimisen jälkeen hajoitelman ensimmäinen komponentti tunnetaan. Ongelmaksi jää näin ollen toisen komponentin eli ei-koepuiden kokonaistilavuuden ennustaminen. Merkitsemällä symbolilla $\tilde{T}(s)$ otokseen s perustuvaa ei-koepuiden tilavuuden ennustinta (engl. predictor), saadaan kokonaistilavuuden T ennustimeksi

$$\hat{T}(s) = \sum_s y_i + \tilde{T}(s). \qquad (4)$$

Ennustimen argumentti on kirjoitettu näkyviin korostamaan ennustimen riippuvuutta otoksesta s. Ennustimen arvoa tietyllä otoksella ja tietyillä satunnaismuuttujien y_i, $i \in s$, arvoilla kutsutaan ennusteeksi.

Superpopulaatiomalli H määrittelee tilavuuksien y_i yhteisjakauman. Koska ennustin $\hat{T}(s)$ on satunnaismuuttujien y_i, $i \in s$, funktio, sen ominaisuuksia voidaan tarkastella viittaamatta lainkaan otantajakaumaan. Merkitään symbolilla E_H yhteisjakauman suhteen määriteltyä odotusarvoa. Keskeinen käsite myöhemmissä tarkasteluissa on mallin H suhteen laskettu keskineliövirhe, joka määritellään seuraavasti (vrt. Kolehmainen, 1977):

$$MSE_H[\hat{T}(s)] = E_H[\hat{T}(s) - T]^2. \qquad (5)$$

Mallin suhteen laskettua keskineliövirhettä kutsutaan lyhyesti keskineliövirheeksi. Keskineliövirhe MSE_H on ennus-

timen ja todellisen tilavuuden neliöpoikkeaman odotusarvo mallin suhteen laskettuna. Se riippuu otoksesta ja kuvaa ennustimen luotettavuutta otoksen poimimisen jälkeen. Tämä kiinnostaa tutkijaa silloin, kun hän tarkastelee laskemiensa ennusteiden luotettavuutta. Ennen otoksen poimimista, kun hän valitsee otantamenetelmän ja määrää tarvittavan otoskoon, tutkija tarvitsee mitan, joka kuvaa ennustimen luotettavuutta ennen otoksen poimimista. Tällaisena otantamenetelmästä riippuvana mittana käytetään keskineliövirheen odotusarvoa otantajakauman suhteen laskettuna eli arvoa

$$E_p\{MSE_H[\hat{T}(s)]\}, \qquad (6)$$

jossa alaindeksi p viittaa otantajakaumaan.

Myöhemmin tarkastellaan pelkästään ennustimia, jotka riippuvat otoksesta, mutta eivät otantamenetelmästä. Tällöin mitan (6) mielessä optimaalinen otantamenetelmä on yksinkertaisesti sellainen, joka todennäköisyydellä yksi tuottaa jonkin keskineliövirheen $MSE_H[\hat{T}(s)]$ minimoivan otoksen. Tällainen 'harkintaotannan' käyttäminen on herkkä superpopulaatiomallin määrittelyvirheille. Cassel ym. (1977) pohtivat pääosin artikkeleihin Royall (1970) ja Royall ja Herson (1973a ja b) viitaten virheellisistä oletuksista johtuvia ongelmia ja keinoja niiden välttämiseksi. Myös tässä tutkimuksessa tarkastellaan käytettävien ennustimien robustisuutta. Huomattakoon lisäksi, että tarkasteltavan pystymittausongelman luonteeseen kuuluu, ettei harkintaotantaa voida käyttää, vaan käytännössä joudutaan turvautumaan satunnaisotantaan. Tavoitteena olevan otantamenetelmän tulee suurella todennäköisyydellä tuottaa otoksia, joissa keskineliövirhe on mahdollisimman pieni.

Tutkijaa kiinnostaa myös kysymys, miten keskineliövirhe MSE_H suhtautuu vastaavaan klassiseen mittaan, otantavariansiin. Tätä kysymystä valaistaan seuraavassa luvussa tarkas-

teltavien ennustimien yhteydessä.

Esitetään vielä keskineliövirheen komponentteihin jako, jota käytetään myöhemmissä tarkasteluissa (vrt. liite).

$$MSE_H[\hat{T}(s)] = V_H[\hat{T}(s)] + \{B_H[\hat{T}(s)]\}^2 + V_H(\sum_{\tilde{s}} y_i). \quad (7)$$

Ensimmäinen komponentti on ei-koepuiden tilavuuden ennustimen mallivarianssi mallin H suhteen laskettuna, toinen komponentti ennustimen $\hat{T}(s)$ otokseen s liittyvän malliharhan neliö. Malliharha määritellään odotusarvona (vrt. Kolehmainen 1977)

$$B_H[\hat{T}(s)] = E_H[\hat{T}(s) - T]. \quad (8)$$

Kolmas keskineliövirheen komponentti on ei-koepuiden tilavuuden mallivarianssi.

23. Polynominen regressioennustin

Tarkastellaan tässä kappaleessa polynomista regressioennustinta, joka saadaan estimoimalla otoksesta superpopulaatiomalli (2) ja ennustamalla estimoidulla mallilla ei-koepuiden kokonaistilavuus. Otokseen s perustuva mallin (2) kerroinvektorin painotetun pns-menetelmän estimaattori on

$$\hat{\beta} = (X_s' W_s X_s)^{-1} X_s' W_s y_s, \qquad (9)$$

jossa X_s muodostuu rivivektoreista $X_i = [1\ x_i \ \ldots\ x_i^k]$, $i \in s$, diagonaalinen painomatriisi W_s alkioista x_i^{-2g} ja vektori y_s vastaavasti alkioista y_i , $i \in s$.

Olkoon $\tilde{x} = [N-n\ \Sigma_{\tilde{s}} x_i \ \ldots\ \Sigma_{\tilde{s}} x_i^k]'$. Tällöin ennustin (4) voidaan kirjoittaa muodossa

$$\hat{T}(s) = \sum_s y_i + \tilde{x}'\hat{\beta}. \qquad (10)$$

Lineaaristen mallien yleisen teorian mukaisesti on $E_H(\hat{\beta}) = \beta$ kaikilla otoksilla s. Näin ollen on

$$E_H[\hat{T}(s)] = E_H(\sum_s y_i) + E_H(\tilde{x}'\hat{\beta}) = E_H(T) \qquad (11)$$

kaikilla otoksilla s. Ennustimen sanotaan tällöin olevan malliharhattoman. Ennustimen keskineliövirheeksi saadaan hajotelmaa (7) sekä lineaaristen mallien teoriaa soveltamalla

$$\begin{aligned} MSE_H[\hat{T}(s)] &= V_H[\tilde{x}'\hat{\beta}] + V_H(\sum_{\tilde{s}} y_i) \\ &= \sigma^2 \tilde{x}' (X_s' W_s X_s)^{-1} \tilde{x} + \sigma^2 \sum_{\tilde{s}} x_i^{2g} \end{aligned} \qquad (12)$$

Tarkastellaan seuraavaksi ennustinta ja sen keskineliövirhettä kahdessa erityistapauksessa.

Olkoot

$$M_j(s) = \frac{\Sigma_s v_i x_i^j}{\Sigma_s v_i}, \quad (v_i = x_i^{-g}) \tag{13}$$

ja

$$m_j(s) = \frac{\Sigma_s x_i^j}{n} \tag{14}$$

otoksen s painotettu ja painottamaton j. origomomentti, ja olkoot M_j ja m_j vastaavat momentit perusjoukossa. Määritellään Royallia ja Hersonia (1973a) mukaellen tasapainoinen otos s_b siten, että polynomin termien astelukuja vastaavat origomomentit ovat yhtäsuuret otoksessa ja perusjoukossa, eli

$$m_j(s_b) = m_j, \quad j=1,\ldots,k. \tag{15}$$

Vastaavasti määritellään painottaen tasapainoinen otos s_B siten, että sille on voimassa

$$M_j(s_B) = m_j, \quad j=1,\ldots,k. \tag{16}$$

Ositetusta otannan käsitteistössä suhteellinen ja Neymannin kiintiöinti ovat analogisia tasapainoisen ja painottaen tasapainoisen otoksen käsitteille.

Ennustin $\hat{T}(s)$ ja sen keskineliövirhe MSE_H yksinkertaistuvat määriteltyjen otosten tapauksessa huomattavasti. Royall ja Herson (1973a) ovat osoittaneet, että kun ennustimessa on mukana termi x^{2g}, niin (vrt. myös liite).

$$\hat{T}(s_b) = N \frac{\Sigma_{s_b} y_i}{n} \tag{17}$$

Liitteessä on osoitettu, että keskineliövirhe supistuu tällöin muotoon

$$MSE_H[\hat{T}(s_b)] = N^2 \frac{\sigma^2}{n}(1-\frac{n}{N})m_{2g} \tag{18}$$

Tämä kaava on approksimatiivisesti voimassa, vaikka termi x^{2g} ei olisikaan mukana ennustimessa.

Vastaavat yksinkertaistukset painottaen tasapainoisen otoksen tapauksessa edellyttävät, että termin x^{2g} lisäksi myös termi x^g on mukana ennustimessa. Tällöin on liitteen mukaan

$$\hat{T}(s_B) = N \frac{\Sigma_{s_B} v_i y_i}{\Sigma_{s_B} v_i}, \quad (v_i = x_i^{-g}) \qquad (19)$$

ja

$$MSE_H[\hat{T}(s_B)] = N^2 \frac{\sigma^2}{n} (m_g^2 - \frac{n}{N} m_{2g}). \qquad (20)$$

Kaava (18) saadaan (20):stä korvaamalla m_g^2 momentilla m_{2g}. Koska $m_{2g}-m_g^2$ on kerrointa N/(N-1) vaille x^g:n varianssi perusjoukossa, nähdään, että painottaen tasapainoinen otos on keskineliövirheen suhteen sitä edullisempi, mitä suurempi on x^g:n ja siten myös läpimittojen varianssi perusjoukossa. Huomattakoon, että kaavojen (17) ja (19) johdossa on oleellista, että ennustimessa on mukana termi, joka on kääntäen verrannollinen kertoimien estimoinnissa käytettyihin painoihin, ja kaavassa (19) lisäksi niiden neliöjuureen. Ennustimen ja sen estimoinnissa käytetyn painotuksen tulee näin ollen olla keskenään sopusoinnussa.

Tarkastellaan seuraavaksi keskineliövirheen ja klassisen, otantajakauman suhteen määritellyn otantavarianssin suhdetta. Oletetaan Sukhatmen ja Sukhatmen (1970 ss. 197-201) mukaisesti, että x:n jakauma on diskreetti. Olkoon perusjoukon puiden tilavuudet ja läpimitat

$$(y_{ij}, x_j),\ i=1,\ldots,N_j',\ j=1,\ldots,h \text{ ja } \Sigma_j N_j = N.$$

Oletetaan, että ennustimessa on termi x^{2g}, jolloin estimoidun regressiomallin residuaalien summa otoksessa on

nolla. Tällöin ennustin $\hat{T}(s)$ voidaan kirjoittaa muodossa

$$\hat{T}(s) = N \cdot m' (X' \bar{W} X)^{-1} X' \bar{W} \bar{y}^*, \qquad (21)$$

jossa $m = [1\ m_1 \ldots m_k]'$, matriisin X riveinä ovat vektorit X_j, $j=1,\ldots,h$, $\bar{W} = \lceil n_1/x_1^{2g}, \ldots, n_h/x_h^{2g} \rfloor$. Vektorin $\bar{y}^* = [\bar{y}_1^*, \ldots, \bar{y}_h^*]'$ alkiot ovat

$$\bar{y}_j^* = \begin{cases} \bar{y}_j = \frac{1}{n_j} \sum_i^{n_j} y_{ij}, & \text{jos } n_j > 0; \\ 0, & \text{jos } n_j = 0. \end{cases} \qquad (22)$$

Lausekkeen (21) yksinkertaistamista varten tarkastellaan sen osamatriisia

$$(X' \bar{W} X)^{-1} X' \bar{W} = (X' \bar{W} X)^{-1} X' \bar{W}\ I \doteq C \qquad (23)$$

Tämä voidaan muodollisesti tulkita regressiokerroinmatriisiksi, jossa j. sarakkeella c^j on otoksesta estimoidun mallin

$$\delta_{ij} = c_{oj} + c_{1j} x_i + \ldots + c_{kj} x_i^k \qquad (24)$$

kertoimet. Merkintä δ_{ij} tarkoittaa Kroneckerin δ-symbolia. Mallin selittäjänä on osoitinmuuttuja, joka saa arvon yksi, kun läpimitta on x_j ja arvon nolla muutoin. Näin tulkittuna voidaan kirjoittaa

$$N\, m' c^j = N_j + d_j, \qquad (25)$$

jossa d_j on mallin (24) residuaalien summa perusjoukossa. Tällöin (21) voidaan kirjoittaa muodossa

$$\hat{T}(s) = [N_1 + d_1 \ \ldots\ N_h + d_h] \bar{y}^*. \qquad (26)$$

Rajoitutaan tarkastelemaan ennustimen eli klassisen otantateorian mielessä estimaattorin ehdollista otantavarianssia $V_p(\hat{T}|n_1,\ldots,n_h)$. Tämä voidaan lausua normaalilla tavalla riippumattomien keskiarvojen $\bar{y}_j$, $(n_j>0)$ otantavarianssien avulla eli kaavan (26) mukaisesti

$$V_p(\hat{T}|n_1,\ldots,n_h) = \sum_{n_j>0}(N_j+ d_j)^2(1-\frac{n_j}{N_j})\frac{S_j^2}{n_j}, \qquad (27)$$

jossa S_j^2 on arvojen y_{ij}, $i=1,\ldots,N_j$ varianssi.

Ennustimen keskineliövirhe mallin suhteen laskettuna voidaan vastaavasti kirjoittaa muodossa

$$MSE_H[\hat{T}(s)] = \sum_{n_j>0}\{[(N_j+d_j)-n_j]^2\frac{\sigma_j^2}{n_j} + (N_j-n_j)\sigma_j^2\} \qquad (28)$$

Kun superpopulaatiomallin mukaisesti oletetaan, että $S_j^2=\sigma_j^2$, $j=1,\ldots,h$, nähdään lausekkeista (27) ja (28) suoraviivaisesti kehittelemällä, että kiintiöinnin $(n_1,\ldots n_h)$ mukaisille otoksille s on voimassa

$$MSE_H[\hat{T}(s)] = V_p(\hat{T}|n_1,\ldots,n_h) + \sum_j \frac{d_j^2}{N_j}\sigma_j^2. \qquad (29)$$

Toisin sanoen mallin suhteen laskettu keskineliövirhe on aina vähintään yhtäsuuri, kuin vastaavan kiintiöinnin mukainen ehdollinen otantavarianssi. Yhtäsuuruus on voimassa silloin, kun $d_j=0$, $j=1,\ldots,h$. Liitteen mukaisesti näin on, kun otos on joko suhteellisen tai Neymannin kiintiöinnin mukainen. Kappaleessa 25. esitettävien simulointikokeiden mukaisesti ero ei muissakaan tapauksissa ole kovin suuri. Näin ollen mallin suhteen laskettua keskineliövirhettä voidaan käyttää numeerisesti yhtäläisenä epävarmuuden mittana kuin ehdollista otantavarianssia. Esimerkiksi molempien mittojen mukaiset luottamusvälit ovat approksimatiivisesti yhtä leveät.

Tarkastellaan vielä kysymystä optimaalisesta otantamenetelmästä. Kokeensuunnitteluun liittyvissä tutkimuksissa on selvitetty suhteellisen paljon havaintojen optimaalista jakaumaa, kun estimoidaan regressiomallia. Esimerkkeinä mainittakoon tutkimukset Elfving (1952), De La Garza (1954), Hoel (1958) sekä Demaershalk ja Kozak (1974 ja 1975). Royall ja Herson (1973a) toteavat, että lähtökohta kokeensuunnitteluun liittyvissä tutkimuksissa poikkeaa käsillä olevasta ongelmasta. Ensinnäkin estimoitava suure on riippumaton havaintopisteiden jakaumasta ja toiseksi minimoitavana suureena on yleensä ollut estimoitavan suureen varianssi. Ennustimessa (10) sitä vastoin ennustettavana arvona on eikoepuiden kokonaistilavuus, joka riippuu otoksesta, toiseksi keskineliövirhe on varianssin ja otoksesta riippuvan suureen $\sigma^2 \Sigma_s x_i^{2g}$ summa.

Kysymykseen optimaalisesta otantamenetelmästä liittyy käytännössä myös ennustimen robustisuusominaisuudet. Mikäli optimaalinen otantamenetelmä on herkkä mallin määrittelyvirheille, ei sitä voida käytännössä soveltaa. Royall (1970) on näyttänyt, että mikäli superpopulaatiomalli on muotoa

$$H: \quad y_i = \beta_1 x_i + \varepsilon_i, \ \varepsilon_i \sim (0, \sigma^2 x_i^2) \qquad (30)$$

niin vastaavan ennustimen keskineliövirheen minimoi otos, jossa ovat n x-arvoltaan suurinta perusjoukon alkiota. Ennustin on luonnollisesti herkkä perusjoukon poikkeamille oletetusta mallista. Malliin (30) liittyen erilaisten suhde-ennustimien robustisuutta eri otantamenetelmien yhteydessä ovat tutkineet mm. Royall ja Herson (1973a ja b) sekä Brewer (1979). Koska yhtälön (17) mukaan pns-menetelmällä estimoitu suhde-ennustin supistuu tasapainoisessa otoksessa yksinkertaiseen malliharhattomaan muotoon, Royall ja Herson suosittelevat yksinkertaista satunnaisotantaa, joka suurilla otoksilla tuottaa likimain tasapainoisia otoksia. Brewer käyttää mallin (30) kertoimille asymptoottisesti otantahar-

hatonta estimaattoria. Suurille perusjoukoille hän suosittelee otantamenetelmää, jossa poimintatodennäköisyydet ovat suoraan verrannolliset hajontoihin σx_i eli likimain painottaen tasapainoisia otoksia tuottavaa Neymannin kiintiöinnille analogista menetelmää.

Tässä tutkimuksessa käytetyn superpopulaatiomallin tapauksessa näyttää Brewerin suositus sopivalta, edellyttäen, että painotetulla pns-menetelmällä estimoidussa ennustimessa on mukana termit x^g ja x^{2g}. Painottaen tasapainoisessa otoksessa ennustin supistuu yksinkertaiseen muotoon (19), joka on malliharhaton riippumatta todellisen mallin muodosta. Vaikka tässä tutkimuksessa ei ole osoitettu, että painottaen tasapainoinen otos minimoi keskineliövirheen, tuntuu tällainen hypoteesi tulosten (20) ja (29) valossa uskottavalta. Myös kappaleen 25. simulointikokeet tukevat hypoteesia. Näin ollen Brewerin suosittelema otantamenetelmä olisi omiaan tuottamaan keskineliövirheeltään minimaalisia otoksia.

24. Ositekeskiarvoihin perustuva ennustin

Tutkimuksen tavoitteena on ollut tutkia polynomisen regressioennustimen ominaisuuksia käytössä olevan ositekeskiarvoihin perustuvan ennustimen korvaajana. Tällöin joudutaan pohtimaan regressioestimoinnin ja ositettuun otantaan perustuvan estimoinnin välistä suhdetta yleisestikin. On tunnettua, että tarkastellun kaltaisessa tilanteessa ositettuun otantaan perustuvan estimoinnin hyvänä puolena on, että estimaattori on harhaton riippumatta y:n ja x:n välisen riippuvuuden muodosta. Toisaalta ositetun otannan ongelmana on luokkaleveydestä aiheutuva komponentti ositteen varianssissa, joka kasvattaa estimaattorin varianssia. Seuraavassa unohdetaan nämä molemmat puolet. Toisaalta tarkasteluissa oletetaan superpopulaatiomalli (2), jolloin regressiomalli ei sisällä malliharhaa. Toisaalta oletetaan edellisen kappaleen lopun mukaisesti, että läpimittajakauma on diskreetti, jolloin ositteen leveys voidaan pitää nollana. Idealisoinnin tavoitteena on selvittää vähemmän tutkittua asiaa, miten menetelmien keskinäiseen suhteeseen vaikuttaa se, että regressioennustin hyödyntää kaikkien ositteiden otosyksiköiden sisältämää informaatiota toisistaan riippuvina eikä erillisinä kuten ositekeskiarvoihin perustuva ennustin.

Tarkastellaan sivulla 13 määriteltyä perusjoukkoa, jossa läpimittajakauma oletettiin diskreetiksi. Kun oletetaan, että jokaisesta ositteesta on otosyksiköitä eli että $n_j>0$, $j=1,\ldots,h$, voidaan ennustin (4) määritellä ositekeskiarvojen avulla muodossa

$$\hat{T}(s) = \Sigma_s y_i + \sum_j^h (N_j - n_j)\bar{y}_j \qquad (31)$$

Ennustinta $\hat{T}(s)$ kutsutaan lyhyesti osite-ennustimeksi. Välittömästi voidaan todeta, että se on malliharhaton.

Keskineliövirheeksi saadaan helposti lauseke

$$MSE_H[\dot{\hat{T}}(s)] = \sum_1^h N_j^2 (1-\frac{n_j}{N_j}) \frac{\sigma^2 x_j^{2g}}{n_j} \qquad (32)$$

eli normaalin ositetun otannan summaestimaattorin otantavarianssin kaltainen lauseke.

Merkitään suhteellisesti kiintiöityä otosta symbolilla s_b' ja Neymannin kiintiöinnin mukaista symbolilla s_B'. Liitteessä on näytetty, että näissä tapauksissa on

$$MSE_H[\dot{\hat{T}}(s_b')] = MSE_H[\hat{\hat{T}}(s_b)] \qquad (33)$$

ja

$$MSE_H[\dot{\hat{T}}(s_B')] = MSE_H[\hat{\hat{T}}(s_B)]. \qquad (34)$$

Toisin sanoen molempien otosten tapauksessa sekä osite-ennustin että regressioennustin ovat yhtä luotettavat. Yleisesti keskineliövirheiden yhtäsuuruus ei ole voimassa. Mikäli edellisessä kappaleessa esitetty hypoteesi otoksen s_B optimaalisuudesta pitää paikkansa, niin tuloksen (34) mukaan kumpikin ennustin on optimissaan yhtä luotettava. Selvitettäväksi jää näin ollen, kumpi on herkempi ja kuinka paljon herkempi poikkeamille optimaalisesta otoksesta. Kysymys on olennainen silloin, kun ei voida käyttää optimaalisen otoksen takaavaa otantamenetelmää.

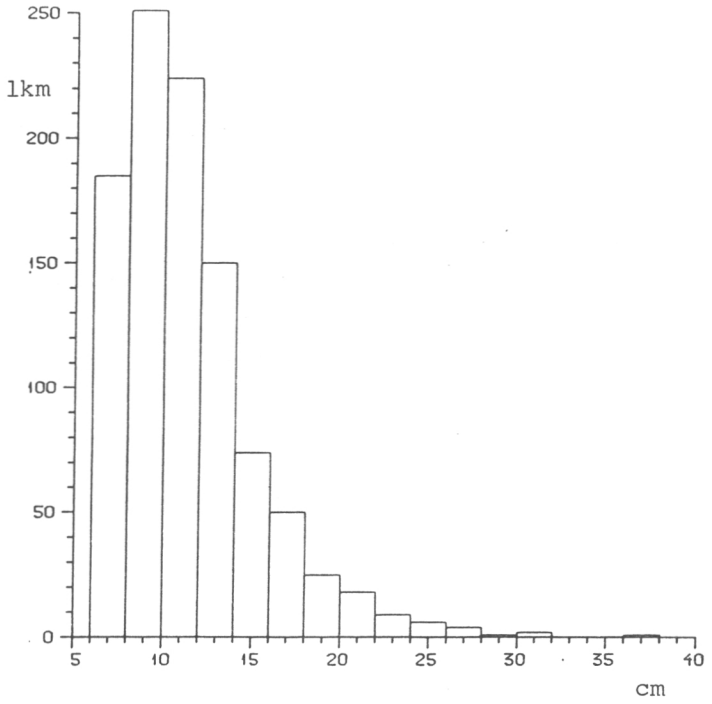
25. Simulointikokeisiin perustuvia tarkasteluja

Simulointitarkasteluja varten generoidaan 1000 rungon teoreettinen metsikkö. Metsikön läpimittajakuma on esitetty kuvassa 1. Se on ollut diskreetti siten, että puiden läpimitat ovat olleet 7,9,...,37 cm. Jokaiselle puulle on generoitu runkotilavuus superpopulaatiomallilla

$$Y_i = 0.26655 \cdot x_i^2 + 0.03138 \cdot x_i^3 - 0.00042134 \cdot x_i^4 + \varepsilon_i; \quad (35)$$

$$\varepsilon_i \sim n(0, 0.0081 \cdot x_1^4).$$

Generoidun metsikön kokonaistilavuus on 90 m^3. Metsikkö vastaa käytännössä pientä kuitupuuksi myytävää leimikkoa.



Kuva 2: Simulointikokeita varten generoidun metsikön läpimittajakuma.

Kokeissa perusjoukosta poimittiin 1000 riippumatonta otosta käyttäen sekä yksinkertaista satunnaisotantaa (SRS) että poimintaa vaihtelevin todennäköisyyksin (PPX). Käytetyt poimintatodennäköisyydet olivat verrannolliset läpimitan neliöön. PPX-otanta toteutettiin käytännössä systemaattisena, kumuloimalla satunnaisjärjestyksessä puiden läpimitan neliöitä ja käyttämällä vakiopituista otantaväliä. Näin menetellen otoskoko saatiin pysymään haluttuna vakiona. Mainittakoon, että käytetyt poimintatodennäköisyydet vastaavat käytännön pystymittauksessa yleistä ns. relaskooppiotantaa.

Tarkastellaan aluksi polynomiennustinta. Taulukkoihin 1 ja 2 on koottu ennustimen otoksien perusteella laskettu harha $B_p(\hat{T})$, estimoitu otantavarianssi V_p, mallin suhteen lasketun keskineliövirheen MSE_H keskiarvo, $E_p(MSE_H)$ sekä sen minimi- ja maksimiarvo otoksissa. Vertailun vuoksi on laskettu myös otosten kiintiöintiä vastaava ehdollisen otantavarianssin keskiarvo, $E(V_p(|))$, sekä kaavoilla (18) ja (20) lasketut keskineliövirheet otoksille s_b ja s_B. Kaikissa käytetyissä kaavoissa on käytetty teoreettista arvoa $\sigma^2=0.0081$.

Taulukko 1: SRS-menetelmällä poimituista otoksista lasketut tunnukset, prosenttia tilavuudesta.

n	B_p	$\sqrt{V_p}$	$\sqrt{E[V_p(\|)]}$	$\sqrt{E(MSE_H)}$	$\min\sqrt{MSE_H}$	$\max\sqrt{MSE_H}$	$\sqrt{MSE_H \mid s_b}$
10	-1.51	37.25	33.61	33.84	4.55	174.53	5.93
25	-0.45	8.29	7.73	7.86	2.85	46.21	3.72
50	-0.31	3.71	3.68	3.80	2.02	15.52	2.60
100	-0.05	2.06	2.05	2.15	1.45	4.90	1.79
200	-0.02	1.25	1.22	1.30	1.00	3.56	1.19
500	-0.03	0.57	0.57	0.62	0.49	0.90	0.60

Taulukko 2: PPX-menetelmällä poimituista otoksista lasketut tunnukset, prosenttia tilavuudesta.

n	B_p	$\sqrt{V_p}$	$\sqrt{E[V_p(I)]}$	$\sqrt{E(MSE_H)}$	Min$\sqrt{MSE_H}$	Max$\sqrt{MSE_H}$	$\sqrt{MSE_H \mid s_B}$
10	-0.33	10.59	10.22	10.34	4.55	122.04	4.55
25	0.05	3.04	3.13	3.20	2.84	14.94	2.84
50	0.04	1.99	1.99	2.03	1.96	3.58	1.96
100	0.00	1.35	1.31	1.33	1.32	1.40	1.32
200	0.00	0.82	0.83	0.84	0.83	0.87	0.83
500	-0.04	0.33	0.34	0.34	0.34	0.35	0.26

Taulukoista havaitaan ensinnäkin, että keskineliövirhe MSE_H on otoksissa ollut keskimäärin vain hieman suurempi kuin ehdollinen otantavarianssi $V_p(I)$. Tämän mukaisesti kaavassa (29) residuaalit sisältävä termi

$$MSE_H - V_p(I) = \sum_j \frac{d_j^2}{N_j}\sigma_j^2 \qquad (36)$$

on ollut suhteellisesti katsoen pieni. Näin ollen, mikäli perusjoukko on superpopulaatiomallin mukainen, mittaavat keskineliövirhe MSE_H sekä ehdollinen otantavarianssi $V_p(I)$ yhtäläisesti ennusteiden luotettavuutta.

Taulukoista nähdään myös, että eri otoksista lasketun keskineliövirheen vaihtelu on suuri. Koska MSE_H mittaa otoksen poimimisen jälkeistä ennustimen epävarmuutta, on se huomattavasti tarkempi mitta kuin otoksen poimimista edeltävää epävarmuutta mittaava otantavarianssi V_p. Tämän mukaisesti, kun käytetään oheismuuttujia, otantatilanne tulisi pyrkiä mallittamaan siten, että oheismuuttujiin ja otokseen perustuen voitaisiin laatia ennustimelle tai estimaattorille otannan jälkeistä epävarmuutta kuvaava mitta. Huomattakoon, että tulosten mukaan tällaiseksi käy klassinen ehdollinen otantavarianssi yhtä hyvin kuin keskineliövirhe MSE_H.

Otosta s vastaava sarake taulukossa 2 vahvistaa esittyä hypoteesia, jonka mukaan painottaen tasapainoinen otos minimoi ennustimen $\hat{T}(s)$ keskineliövirheen. Mikäli näin ei olisi, olisi mitä todennäköisimmin jostakin otoksesta laskettu keskineliövirhe ollut alle sarakkeen arvon. Näin ei kuitenkaan ole käynyt, vaan otoksista laskettu minimiarvo on sarakkeen mukainen suurinta otoskokoa lukuunottamatta. Poikkeuksen syynä on se, että otoskoolla n=500 läpimitan neliöön verrannolliset poimintatodennäköisyydet suurimmilla läpimitoilla kasvaisivat yli yhden. Tästä johtuen poimitut otokset eivät ole voineet olla painottaen tasapainoisia.

Käytännön kannalta on tarpeellista tutkia ennustimen robustisuusominaisuuksia. Poikkeamat määritellystä superpopulaatiomallista voivat johtua joko virheellisestä mallin muodosta tai väärän varianssirakenteen olettamisesta. Tarkastellaan aluksi edellistä tapausta. Taulukon 3 tuloksia laskettaessa on ennustimesta pudotettu pois neljännen asteen termi. Tämä vastaa oletusta, että rungon muotokorkeuden riippuvuutta läpimitan arvosta kuvaa suora. Tuloksia laskettaessa on MSE_H:n kaavoissa σ^2:n asemasta käytetty otoksesta laskettua jäännösvarianssia.

Taulukko 3: Otoksista lasketut tunnukset, kun ennustimena on ollut $\hat{T}(s) = \Sigma_S y + \Sigma_{\bar{S}}(\beta_2 x_i^2 + \beta_3 x_i^3)$, prosenttia tilavuudesta.

	SRS-menetelmä			PPX-menetelmä		
n	B_p	$\sqrt{V_p}$	$\sqrt{E(MSE_H)}$	B_p	$\sqrt{V_p}$	$\sqrt{E(MSE_H)}$
10	2.17	8.20	8.20	0.58	5.04	5.13
25	1.18	4.43	4.14	0.26	3.00	3.01
50	0.97	2.95	2.79	0.09	2.02	2.06
100	0.78	2.08	1.88	-0.13	1.37	1.38
200	0.56	1.40	1.24	-0.05	0.83	0.86
500	0.30	0.68	0.62	0.14	0.35	0.35

Taulukosta nähdään, että ennustin sisältää otantaharhaa. SRS-menetelmän tapauksessa harha on käytännössä merkitsevä, PPX-menetelmää käytettäessä se sen sijaan on merkityksetön. Pääsyynä harhaan on, että ennustin ei ole yksinkertaistuskaavojen (17) ja (19) mukaisessa sopusoinnussa estimoinnissa käytettyjen painojen kanssa, joina ovat olleet arvot x_i^{-4}. Tästä johtuen ennustin on malliharhainen otoksissa s_b ja s_B.

Keskineliövirhe MSE_H kuvaa taulukon 3 mukaisesti edelleen hyvin ennustimeen liittyvää epävarmuutta. Tosin SRS-menetelmällä se keskimäärin aliarvioi estimoitua otantavarianssia V_p. Kun verrataan taulukkoa 3 taulukoihin 1 ja 2 huomataan, että termin x^4 pudottaminen ennustimesta on vähentänyt huomattavasti otantavarianssia pienillä otoksilla. Väheneminen on merkittävä vaikka syntynyt harhakin otettaisiin huomioon. Termin pudottaminen mallista vastaa itseasiassa termin kertoimen β_4 estimoimista harhaisesti nollalla. Pienillä otoksilla tämä harhainen estimointi on tulosten mukaan kannattanut. Suurilla otoksilla sen sijaan otantavarianssi on kasvanut. Tämä johtuu ennustimen suurentuneesta jäännösvarianssista, joka isoilla otoksilla tulee suuremmaksi varianssin lähteeksi kuin läpimittajakauman vaihtelu otoksissa.

Taulukossa 4 on tarkasteltu, miten virheellisen varianssirakenteen olettaminen vaikuttaa ennustimen ominaisuuksiin. Perusjoukkoa generoitaessa on satunnaiskomponentin jakaumana ollut

$$\varepsilon_j \sim n(0,\ 0.000324 \cdot x_i^5). \qquad (37)$$

Taulukon tuloksia laskettaessa on MSE_H:n kaavassa jälleen käytetty otoksista laskettua jäännösvarianssia.

Taulukko 4: Otoksista lasketut tunnukset, kun perusjoukko on generoitu mallilla (35), jossa $\varepsilon_i \sim n(0,0.000324x^5)$, prosenttia tilavuudesta.

	SRS-menetelmä			PPX-menetelmä		
n	B_p	$\sqrt{V_p}$	$\sqrt{E(MSE_H)}$	B_p	$\sqrt{V_p}$	$\sqrt{E(MSE_H)}$
10	-0.65	25.03	22.42	-0.16	7.73	7.29
25	-0.20	6.17	5.09	0.06	2.31	2.38
50	-0.14	3.05	2.55	0.03	1.46	1.52
100	0.02	1.75	1.45	0.00	0.97	0.99
200	0.02	1.06	0.88	0.00	0.58	0.63
500	-0.02	0.50	0.42	-0.02	0.22	0.25

Odotetusti tuloksissa ei ole malliharhaa. Samoin odotettavissa oli, että SRS-menetelmässä MSE_H aliarvioi otantavarianssia ja PPX-menetelmässä yliarvioi sitä. Tämä johtuu siitä, että jäännösvarianssi on SRS-otoksissa keskimäärin pienempi ja PPX-otoksissa keskimäärin suurempi kuin estimoitavan mallin mukainen jäännösvarianssi koko perusjoukossa. SRS-menetelmässä aliarvio on ollut huomattava, PPX-menetelmässä yliarviota voidaan sen sijaan pitää merkityksettömänä. Royall ja Cumberland (1978) ovat esittäneet ennustimen mallivarianssille robustisen estimointimenetelmän, jossa edellä esitetystä syystä aiheutuvat virhearviot on minimoitu.

Taulukossa 5 on verrattu regressioennustinta sekä osite-ennustinta keskenään. Taulukon tuloksia laskettaessa on käytetty otantamenetelmää, jossa otokseen on poimittu aluksi yksinkertaisella satunnaisotannalla yksi koepuu kustakin ositteesta. Loput koepuut on tämän jälkeen poimittu normaalisti käyttäen joko SRS- tai PPX-menetelmää. Näin menetellen on vältetty osite-ennustimeen liittyvä tyhjien luokkien ongelma. Osite-ennustimen keskineliövirhe $MSE_H(\hat{\bar{Y}})$ on laskettu kaavalla (32). Keskineliövirheiden laskennassa on

käytetty teoreettista arvoa $\sigma^2 = 0.0081$.

Taulukko 5: Otoksista lasketut regressio- ja osite-ennustimen keskimääräiset keskineliövirheet, kun on käytetty otantamenetelmää, joka takaa vähitään yhden koepuun jokaiseen ositteeseen, prosenttia tilavuudesta.

	SRS-menetelmä		PPX-menetelmä	
n	$\sqrt{E[MSE_H(\hat{T})]}$	$\sqrt{E[MSE_H(\hat{\bar{T}})]}$	$\sqrt{E[MSE_H(\hat{T})]}$	$\sqrt{E[MSE_H(\hat{\bar{T}})]}$
25	3.04	3.48	3.08	3.46
50	2.14	2.35	2.01	2.17
100	1.52	1.64	1.33	1.39
200	1.04	1.11	0.84	0.85
500	0.54	0.57	0.34	0.35

Tulosten mukaan osite-ennustimen keskineliövirhe on pienillä otoksilla ollut keskimäärin noin 10 % suurempi kuin regressioennustimella. Suurilla otoksilla ero luonnollisesti vähenee otosten läpimittajakaumien stabiloituessa. Koska osite-ennustimen keskineliövirhe ei sisällä ositteen leveydestä aiheutuvaa komponenttia ja regressioennustin on malliharhaton, aiheutuu ero pelkästään siitä, että luokkakeskiarvot 'korjaavat toisiaan' regressioennustimen tapauksessa, kun taas osite-ennustimessa keskiarvot ovat riippumattomia eivätkä vaikuta toisiinsa.

3. Yleisten tilavuusfunktioiden hyödyntäminen ennustamisessa

31. Yleisperiaate

Tutkimuksessa käsitellään stokastisen ennakkoinformaation hyödyntämistä ennustimen parametrien estimoinnissa. Tavoitteena on keskineliövirheen MSE_H pienentäminen. Teräsvirta (1981) toteaa, että valittavana on ainakin kaksi lähestymistapaa, nimittäin bayesilainen tai otantateoreettinen lähestymistapa, jota hän tutkimuksessaan käyttää. Tässä tutkimuksessa käytetty lähestymistapa ei ole puhtaasti kumpaakaan. Otantateoreettisesta se poikkeaa siinä, että superpopulaatiomallin parametreja ei oleteta kiinteiksi, vaan stokastisiksi. Puhtaasti bayesilaisesta lähtökohta taas poikkeaa siinä, että parametrien priorijakaumasta oletetaan tunnetuksi korkeintaan odotusarvo ja varianssi. Esiteltävään ajattelutapaan on vaikuttanut Väliaho (1969), jossa tarkastellaan regressiomallin parametreille asetettavia lineaarisia side-ehtoja ja ns. toiveita, jotka ovat approksimatiivisia side-ehtoja. Toiveille määritellään paino, josta riippuu missä määrin ne vaikuttavat parametrien estimaatteihin. Kun toiveen paino on nolla perustuu estimointi pelkästään havaintoihin, kun paino kasvaa äärettömyyksiin, muuttuu toive side-ehdoksi.

Kun ajatellaan ennustimen kertoimia koskeva ennakkoinformaatio määritellyksi approksimatiivisin side-ehdoin eli toivein, niin toiveiden paino nolla vastaa tilannetta, jolloin ennakkoinformaatiota ei hyödynnetä lainkaan. Toiveiden painon kasvaessa lähenevät ennusteet ennakkoinformaation mukaisia ennusteita, jolloin käytännössä myös ennustimen mallivarianssi pienenee. Varianssin pieneminen tapahtuu mahdollisen harhan kustannuksella. Käytännössä ongelmana on, miten ennakkoinformaatio puetaan toiveiksi ja kuinka suureksi valitaan toiveiden paino. Lähtökohdalla on läheiset

yhteydet ns. sekaestimointiin (vrt. Teräsvirta, 1981).

Lähtökohdan hyvänä puolena on, että käytettyjen ennustimien kiinnostavia ominaisuuksia, kuten odotusarvoa ja keskineliövirhettä, voidaan tarkastella olettaen tunnetuksi ainoastaan tarkasteltavien jakaumien odotusarvo ja varianssi.

Lähtökohdan selventämiseksi tarkastellaan yksinkertaista esimerkkiä. Oletetaan, että superpopulaatiomalli on muotoa

$$H: \quad y_i = \mu + \varepsilon_i, \quad i=1,\ldots,N, \tag{38}$$

jossa virhekomponentit ovat riippumattomia ja $\varepsilon_i \sim (0,\sigma^2)$. Oletetaan lisäksi, että perusjoukoissa, joita tarkastellaan, μ on stokastinen. Merkitään

$$G: \quad \mu \sim (\mu_o, \sigma_o^2). \tag{39}$$

Leimikkokohtaista parametria μ estimoitaessa voidaan toiveena pitää likimääräistä yhtälöä $\mu \approx \mu_o$. Estimoinissa toive voidaan huomioida määrittelemällä μ:n estimaattori otoskeskiarvon ja toiveena olevan arvon μ_o painotettuna keskiarvona. Toisin sanoen μ:n estimattoriksi saadaan lauseke

$$\hat{\mu}_R = \frac{1}{n+\kappa} \left(\Sigma_s y_i + \kappa \cdot \mu_o \right). \tag{40}$$

jossa tuntematon suure κ määrää toiveen painon. Estimaattori $\hat{\mu}_R$ on harhainen, muodoltaan se vastaa μ:n sekaestimaattoria. Vastaava ennustin on

$$\hat{T}_R(s) = \Sigma_s y_i + (N-n)\hat{\mu}_R \tag{41}$$

ja sen keskineliövirhe on liitteen mukaan

$$MSE_H[\hat{T}_R(s)] = (N-n)^2 \left[\frac{n\sigma^2}{(n+\kappa)^2} + \frac{\kappa^2}{(n+\sigma)^2}(\mu_o-\mu)^2 + \frac{\sigma^2}{(N-n)} \right]. \tag{42}$$

Tässä tutkimuksessa κ määrätään siten, että se minimoi keskineliövirheen odotusarvon E_G. Vaatimus on luonnollinen, koska sen mukaan ennustimen keskineliövirheen odotusarvo eri leimikoilla minimoituu. Tämän mukaisesti κ määrätään minimointitehtävän

$$\min_{\kappa} E_G\{MSE_H[\hat{T}_R(s)]\} \qquad (43)$$

ratkaisuna. Sijoittamalla (42) lausekkeeseen (43) ja jättämällä pois κ:sta riippumattomat termit saadaan yhtäpitävä minimointitehtävä

$$\min_{\kappa} \left[\frac{n\sigma^2}{(n+\kappa)^2} + \frac{\kappa^2\sigma_o^2}{(n+\kappa)^2} \right], \qquad (44)$$

jonka ratkaisu κ_o on

$$\kappa_o = \frac{\sigma^2}{\sigma_o^2} . \qquad (45)$$

Tällöin μ:n estimaattori $\hat{\mu}_R$ on

$$\hat{\mu}_R = \frac{\sigma_o^2}{n\sigma_o^2+\sigma^2} \sum_s y_i + \frac{\sigma^2}{n\sigma_o^2+\sigma^2} \mu_o . \qquad (46)$$

Ennustimen $\hat{T}_R$ keskineliövirhe arvolla κ_o on

$$MSE_H[\hat{T}_R(s)] = N^2\left(1-\frac{n}{N}\right)\left(1+\frac{\kappa_o}{N}\right)\frac{\sigma^2}{n+\kappa_o} . \qquad (47)$$

Estimaattori $\hat{\mu}_R$ on sama kuin Bayesilaisella päättelyllä saadun posteriorijakauman keskiarvo, kun jakaumat H ja G oletetaan normaaleiksi (vrt. Zellner 1971 ss. 14-15). Estimaattori voidaan tulkita painotetuksi keskiarvoksi, joka on laskettu aineistosta, jossa otokseen s on yhdistetty teoreettinen havainto μ_o painolla κ. Tällöin ajatellaan μ:n olevan kiinteän ja μ_o:n stokastisen. Tulkinta on tällöin otantateoreettinen. Olettamalla μ_o:n varianssiksi σ_o^2 ja, että teoreettinen havainto ei korreloi otoshavaintojen kanssa, voidaan painotetun pns-menetelmän mukaisesti

johtaa μ:n estimaattoriksi $\hat{\mu}_R$:n lauseke, jossa κ:n arvoksi saadaan κ_o (Teräsvirta, keskustelu). Vastaavalla ennustimella on näin ollen haluttu ominaisuus, että se minimoi (43):n odotusarvolausekkeen.

Eri lähestymistavat johtavat siten samaan lopputulokseen. Bayesilaisen lähestymistavan heikkoutena voidaan pitää vaatimusta jakaumien muodon olettamisesta. Otantateoreettiseen ajatteluun taas liittyy tulkintavaikeus, sillä oletus μ_o:n stokastisuudesta ja μ:n kiinteydestä on nurinkurinen tutkimusongelman asettelun kannalta. Tässä tutkimuksessa käytetyn lähestymistavan eräänä hyvänä puolena voidaan pitää sitä, että siinä samoin kuin bayesilaisessa ajattelussa pysyy selkeästi erillään toisaalta superpopulaatiomallin virhetermin jakauma, leimikon sisäinen vaihtelu, ja mallin parametrien jakauma, leimikkojen välinen vaihtelu.

32. Polynominen regressioennustin

Tarkastellaan seuraavaksi, miten kappaleessa 2.1 mainittujen yleisten tilavuusfunktioiden sisältämää informaatiota voidaan hyödyntää polynomiennustinta estimoitaessa. Esitetään aluksi kuitenkin kokonaistilavuuden Bayes-ennustin Kolehmaisen (1977) mukaisesti. Oletetaan superpopulaatiomallin virhetermit normaalisti jakautuneiksi ja kerroinvektorin β priorijakaumaksi

$$G': \ \beta \sim N(\beta_o, \ \sigma_o^2 V_o). \tag{48}$$

Tällöin kokonaistilavuuden T Bayes-ennustin on

$$\hat{T}_B(s) = \sum_s y_i + \tilde{x}' \hat{\beta}_B, \tag{49}$$

jossa

$$\hat{\beta}_B = (X_s' W_s X_s + \frac{\sigma^2}{\sigma_o^2} V_o^{-1})^{-1} (X_s' W_s y_s + \frac{\sigma^2}{\sigma_o^2} V_o^{-1} \beta_o) \tag{50}$$

on β:n posteriorijakauman keskiarvovektori.

Käytännössä bayesilaista lähestymistapaa vaikeuttaa se, että yleiset tilavuusfunktiot eivät ole polynomin muotoisia. Vaikka ne voitaisiin sellaisina estimoidakin, on β:n kovarianssimatriisin estimointi hankalaa. Tämä johtuu siitä, että tilavuusfunktioiden perustana olevat aineistot on kerätty eri metsiköistä, korkeintaan muutama koepuu metsikköä kohden, mikä tekee metsiköiden välisen kovarianssimatriisin $\sigma_o^2 V$ estimoinnin epävarmaksi. Tämä on eräs syy, joka on vaikuttanut tutkimuksessa käytettävään lähestymistapaan ennakkoinformaation hyödyntämiseksi.

Yleinen, esimerkiksi inventointiaineistosta estimoitu tilavuusfunktio f(x) estimoi tietyn läpimittaisten runkojen tilavuuden odotusarvoa kaikissa Suomen metsissä. Toisin sa-

noen, kun unohdetaan funktion f(x) estimointivirhe, niin

$$f(x) = E_G[E_H(Y|x)] \qquad (51)$$

jossa E_G tarkoittaa yli metsiköiden määriteltyä odotusarvoa. Funktio f(x) voidaan kirjoittaa muodossa

$$f(x) = \mathbf{x}'\beta + t \qquad (52)$$

jossa $\mathbf{x}=[1\ x\ \ldots\ x^k]'$ ja $t=f(x)-\mathbf{x}'\beta$ on funktion f läpimittaan x liittyvä metsikkökohtainen harha. Kun tarkastellaan yhtälöä (52) tietyssä metsikössä, yhtälön molemmat puolet ovat kiinteitä, ei-stokastisia. Kun taas tarkastellaan yhtälöä eri metsiköissä eli jakauman G kannalta, niin sekä β että t ovat stokastisia. Tarkastellaan aluksi ennakkoinformaation hyödyntämistä tietyn metsikön kannalta.

Yhtälön (52) ennakkoinformaatio kohdistuu yhteen läpimittaan. Tarkastellaan tilannetta, jossa ennakkoinformaatiota käytetään läpimitoilla x_i, i=1,...,l. Läpimittojen valintaan palataan myöhemmin. Olkoon $r=[f(x_1)\ \ldots\ f(x_l)]'$, $t=[t_1\ldots\ t_l]'$ ja R rivivektoreista $X_i=[1\ x_i\ldots\ x_i^k]$, i=1,...,l, muodostettu matriisi. Näin merkittynä ennakkoinformaatio voidaan kirjoittaa muodossa

$$r = R\beta + t. \qquad (53)$$

Tässä muodossa ennakkoinformaatio on Teräsvirran (1981) tarkasteleman stokastisen ennakkoinformaation erikoistapaus, joka saadaan jättämällä pois stokastinen komponentti. Teräsvirran tarkastelema sekaestimointi johtaa tässä tapauksessa sidotun pns-menetelmän käyttämiseen, jossa estimoitaville parametreille asetetaan side-ehto

$$r = R\beta \qquad (54)$$

Käytännössä otosinformaatio tulee tällöin hylätyksi ja ennusteet ovat suoraan ennakkoinformaation mukaisia. Tästä syystä tarkastellaan ongelmaa lähtien Väliahon (1969) lineaarisia toiveita koskevista tarkasteluista. Yhtälö (54) tulkitaan parametrivektoriin kohdistuviksi toiveiksi ja sitä vastaten määritellään diagonaalinen painomatriisi $\kappa \cdot W_r$, jossa W_r:n alkiot ovat x_i^{-2g}. Kerroin κ määrää näin ollen toiveiden yleisen voimakkuden ja W_r:n alkiot toiveiden painojen keskinäisen suhteen. Parametreja estimoitaessa toiveita voidaan teknisesti käsitellä havaintoina (vrt. Väliaho, 1969). Tällöin saadaan β:n estimaattoriksi lauseke

$$\hat{\beta}_R = (X_s' W_s X_s + \kappa \cdot R' W_r R)^{-1} (X_s' W_s y_s + \kappa \cdot R' W_r r), \qquad (56)$$

Tämä on muodollisesti sama kuin vastaava sekaestimaattorin lauseke, joka on johdettu olettaen t stokastiseksi (vrt. Teräsvirta, 1981). Estimaattorin yhteys Bayes-estimaattoriin (50) on myös ilmeinen.

Estimaattoria $\hat{\beta}_R$ vastava ennustin on

$$\hat{T}_R(s) = \sum_s y_i + x' \hat{\beta}_R \qquad (57)$$

Tämä ei yleisesti ottaen ole malliharhaton. Sen malliharha on (vrt. liite)

$$B_H[\hat{T}_R(s)] = \tilde{x}' E_H(\hat{\beta}_R - \beta) = \kappa \cdot \tilde{x}' U R' W_r t, \qquad (58)$$

jossa $U = (X_s' W_s X_s + R' W_r R)^{-1}$. Ennustimen keskineliövirhe MSE_H on liitteen mukaan

$$\begin{aligned} MSE_H[\hat{T}_R(s)] &= \tilde{x}' MSE_H(\hat{\beta}_R) \tilde{x} + \sigma^2 \sum_{\tilde{s}} x_i^{2g} \\ &= \sigma^2 \tilde{x}' U X_s' W_s X_s U \tilde{x} \\ &+ \kappa^2 (\tilde{x}' U R' W_r t)^2 \\ &+ \sigma^2 \sum_{\tilde{s}} x_i^{2g}. \end{aligned} \qquad (59)$$

Keskineliövirhe kuvaa ennustimen luotettavuutta yhdellä leimikolla. Tarkastellaan tilannetta nyt eri leimikkojen kannalta. Yhtälön (53) harhavektori t on nyt stokastinen. Merkitään sen jakaumaa symbolilla G. Yhtälön (51) mukaisesti $E_G(t)=0$. Kuten edellisen kappaleen esimerkissä on tavoitteena nytkin minimoida keskineliövirheen (59) odotusarvo E_G. Erona aikaisempaan on, että optimointitehtävän muuttujana on nyt parametrin κ lisäksi ennakkoinformaation läpimittajakauma. Tästä syystä minimointitehtävä monimutkaistuu eikä sitä sen vuoksi voida ratkaista analyyttisesti. Seuraavassa esitetty ongelman ratkaisu perustuu intuitioon.

Tarkastellaan aluksi optimaalisen κ:n arvon määräämistä. Oletetaan, että ennakkoinformaatiossa läpimittajakauma on sama kuin otoksessa, jolloin $R=X_s$ ja $W_r=W_s$ ja että vektori r sisältää koepuiden ennakkoinformaation mukaiset tilavuudet. Helposti nähdään, että ennustimen $\hat{T}_R$ keskineliövirhe supistuu tällöin approksimatiiviseen muotoon

$$MSE_H[\hat{T}_R(s)] \approx \frac{\sigma^2}{(1+\kappa)^2}\tilde{x}'(X_s'W_sX_s)^{-1}\tilde{x} + \frac{\kappa^2}{(1+\kappa)^2}(\textstyle\sum_{\tilde{s}} t_i)^2 + \sigma^2\textstyle\sum_{\tilde{s}} x_i^{2g}, \tag{60}$$

Kaava on approksimatiivinen, koska toisessa komponentissa ei-koepuiden tilavuuden kokonaisharhan estimaattori on korvattu todellisella arvollaan. Optimaalisen κ:n arvon määräämiseksi saadaan näin minimointitehtävä

$$\min_{\kappa}\left[\frac{1}{(1+\kappa)^2}\sigma^2\tilde{x}'(X_s'W_sX_s)^{-1}\tilde{x} + \frac{\kappa^2}{(1+\kappa)^2}E_G(\textstyle\sum_{\tilde{s}} t_i)^2\right] \tag{61}$$

Ratkaisuksi saadaan edellisen kappaleen mukaisesti

$$\kappa_o = \frac{\sigma^2\tilde{x}'(X_s'W_sX_s)^{-1}\tilde{x}}{E_G(\textstyle\sum_{\tilde{s}} t_i)^2} \tag{62}$$

Kaavan osoittaja on ei-koepuiden tilavuuden ennustimen mallivarianssi ja nimittäjässä ei-koepuiden kokonaistilavuuden

harhan neliön odotusarvo eri leimikoilla eli jakauman G suhteen laskettu harhan varianssi. Näin ollen kaava on operatiivinen, sillä osoittaja voidaan estimoida otoksen avulla ja nimittäjä estimoidaan käytännössä tilavuusfunktioiden metsikön välisen varianssikomponentin avulla.

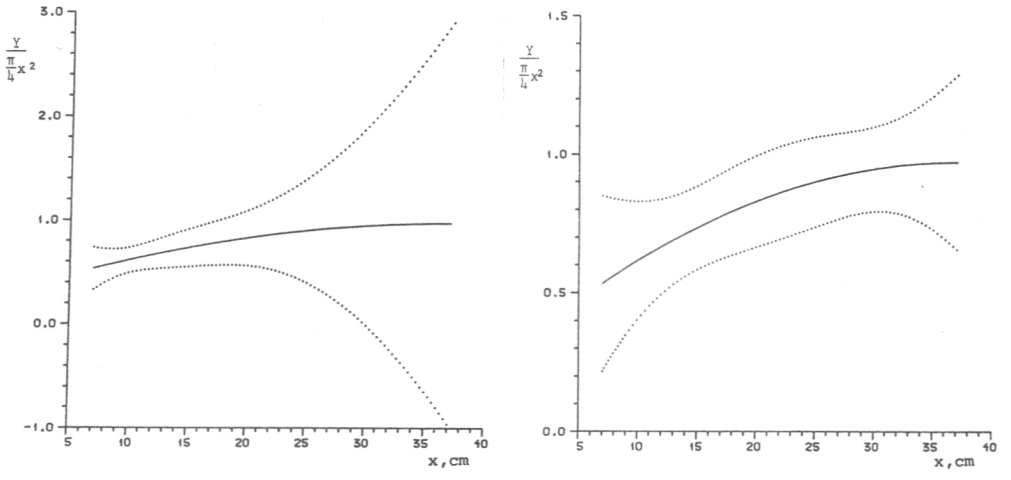
Edellisen kappaleen esimerkin valossa tulos (62) on luonnollinen. Esimerkistä poiketen κ:n arvo riippuu nyt otoksesta. Mitä suurempi on otoksesta laskettu keskineliövirhe, sitä suurempi on optimaalinen κ:n arvo. Toisin sanoen huonoissa otoksissa ennakkoinformaatiota kannattaa hyödyntää voimakkaammin. Käytännön kannalta tuloksessa (62) on myös huomattavaa, että nimittäjässä oleva harhan varianssi eri metsiköissä on huomattavasti helpompi estimoida kuin bayesilaisen lähestymistavan edellytyksenä oleva β:n kovarianssimatriisi $\sigma_o^2 V_o$.

Tarkastellaan seuraavaksi ennakkoinformation läpimittajakauman valintaa. Sovelletaan Teräsvirran (1981) sekaestimaattoria koskevia tuloksia. Niiden mukaan ennakkoinformaatiota kannattaa hyödyntää, kun

(i) 'virhe-kohina'-suhde $t't/\sigma^2$ ei ole kovin suuri'
(ii) ennakkoinformaatiota on suunnilla, joilla mahdollista otosinformaatiota on vähän,
(iii) ennakkoinformaatio ei ole liian harhaista suunnilla, joilla otosinformaatiota on vähän.

Tämän tutkimuksen kannalta tarkasteltuna Teräsvirran ehdot voidaan tulkita suosituksiksi sille, millä alueilla ennakkoinformaatiota milläkin painolla kannattaa käyttää. Optimaalinen ennakkoinformaation läpimittajakauma riippuu ilmeisesti otoksesta. Edellä tarkasteltu tapaus, jossa ennakkoinformaatiolla ja otoksella on saman muotoinen läpimittajakauma, on ehdon (ii) suositusta vastaan.

Seuraavassa lähestytään ongelmaa ennakkoinformaation mukaisen tilavuuden luottamusvyön avulla. Luottamusvyö voidaan määrätä, kun tarkastellaan ennakkoinformaatiota teoreettisena aineistona. Kun oletetaan, että siitä olisi estimoitu ennustimen kerroinvektori, olisi matriisi $(R'W_rR)^{-1}$ kerrointa vaille pns-estimaattorin kovarianssimatriisi. Näin ollen matriisin $(R'W_rR)^{-1}$ avulla voidaan määrittää rungon tilavuusennustimen luottamusvyön muoto. Kuvaan 3a on piirretty ennakkoinformaation mukaisen muotokorkeuden luottamusvyön muoto silloin, kun läpimittajakauma on ollut sama kuin simulointikoikeissa käytetyssä perusjoukossa.



a) b)

Kuva 3: Ennakkoinformaation luottamusvyön muoto, kun läpimittajakauma on (a) kuvan 1 mukainen, (b) tasainen.

Kuten kuvasta nähdään, on ennakkoinformaatio voimakkainta eli luottamusvyö kapea, siellä, missä on eniten havaintoja ja todennäköisesti myös otosinformaatiota. Tämä on vastoin Teräsvirran suositusta (ii). On luonnollista vaatia, että

ennakkoinformaation luottamusvyö olisi tasalevyinen. Tasalevyinen muotokorkeuden luottamusvyö merkitsee kutakuinkin vakiona pysyvää tilavuuden suhteellista luotettavuutta, toisin sanoen ennakkoinformaationa käytettyjen yleisten tilavuusfunktioiden ajatellaan olevan suhteellisesti yhtä luotettavia eri läpimitoilla. Lähes tasalevyinen luottamusvyö saadaan käyttämällä ennakkoinformaatiossa tasaista läpimittajakaumaa. Kuvassa 3b on luottamusvyön muoto, kun läpimitat ovat tasaisesti jakutuneet 7 cm:stä 37 cm:iin.

Kun ennakkoinformaation läpimittajakaumana käytetään tasaista jakaumaa ja sen painona kaavan (62) mukaista arvoa, poiketaan optimaalisesta menettelystä siinä, että käytetty läpimittajakauma on otoksesta riippumaton, ja että κ ei ole käytetyn jakauman mukainen optimiarvo.

33. Ositekeskiarvoihin perustuva ennustin

Ennakkoinformaatiota käytettäessä osite-ennustimen estimoinnissa voidaan suoraan soveltaa kappaleen 31. esimerkin tuloksia. Kukin osite on toisistaan riippumattomaton ja ennakkoinformaatio voidaan huomioida ositekohtaisella optimipainolla. Otoksen tyhjät luokat eivät aiheuta ongelmia, koska niiden keskiarvona voidaan käyttää ennakkoinformaation mukaista arvoa. Tällainen menettely vastaa puhtaasti ositetun otannan periaatetta. Käsillä olevassa ongelmassa on kuitenkin luonnollista käyttää hyväksi sitä, että eri luokkien poikkeamat ennakkoinformaatiosta korreloivat keskenään. Tuntuu luonnolliselta, että ennakkoinformaatioon tehdään ensin yleinen tasokorjaus koko otoksen perusteella ja vasta sen jälkeen ennakkoinformaatiota käytetään eri ositteissa. Näin menetellen voidaan todennäköisesti pienentää ennakkoinformaatiota käyttävän ennustimen harhaa ositteissa, joissa koepuita on vähän tai ei lainkaan. Koska tällainen menettely on ensin mainittua yleisempi, esitetään teoreettiset tulokset sen mukaisina.

Tarkastellaan sivulla 13 määriteltyä perusjoukkoa, jossa läpimittajakauma on diskreetti ja käytetään kappaleen 25. merkintöjä. Oletetaan perusjoukon olevan superpopulaatiomallin (2) mukainen. Merkitään symboleilla

$$\mu_j = X_j\beta \quad \text{ja} \quad \sigma_j^2 = \sigma^2 x_j^{2g}, \quad j=1,\ldots,h \qquad (63)$$

tilavuuden odotusarvoa ja varianssia eri ositteissa. Tilavuusfunktion f(x) mukainen ennakkoinformaatio voidaan merkintää (53) vastaten kirjoittaa muodossa

$$r_j = f(x_j) = \mu_j + t_j, \quad j=1,\ldots,h, \qquad (2^6)$$

jossa t_j on ennakkoinformaation harha j:nnessä ositteessa. Suoritetaan ennakkoinformaatiolle tasokorjaus otoksen perus-

teella seuraavasti:

$$\hat{r}_j = r_j + a x_j^g \tag{65}$$

jossa

$$a = \frac{1}{n} \sum_j n_j (\bar{y}_j - r_j) \cdot x_j^{-g}. \tag{66}$$

Käyttäen tasokorjattua ennakkoinformaatiota saadaan j:nnen ositteen keskiarvon estimaattoriksi (40):n mukaisesti

$$\hat{\mu}_{R_j} = \frac{1}{n_j+\kappa_j} \left(\sum_i^{n_j} y_{ij} + \kappa_j \hat{r}_j \right) \tag{67}$$

Osite-ennustimeksi $\mathring{T}_R$ saadaan vastaavasti

$$\mathring{T}_R(s) = \sum_j^h \left[\sum_i^{n_j} y_{ij} + (N_j - n_j) \hat{\mu}_{R_j} \right] \tag{68}$$

Liitteen mukaisesti ennustin voidaan edelleen kehittää muotoon

$$\mathring{T}_R(s) = \sum_j \left[\sum_i y_{ij} + (A_j + B_j) \sum_i y_{ij} + (A_j \kappa_j + B_j n_j) r_j \right] \tag{69}$$

jossa

$$A_j = \frac{N_j - n_j}{n_j + \kappa_j} \quad \text{ja} \quad B_j = \frac{1}{x_j^g} \sum_i A_i \frac{\kappa_i x_i^g}{n} \tag{70}$$

Asettamalla $B_j = 0$, $j = 1, \ldots, h$ saadaan erikoistapauksena osite-ennustin, jossa tasokorjausta (65) ei ole tehty. Ennustin $\mathring{T}_R(s)$ on harhainen. Sen malliharha on liitteen mukaisesti

$$B_H(\mathring{T}_R) = \sum_j (A_j \kappa_j - B_j n_j) t_j \tag{71}$$

ja keskineliövirhe

$$MSE_H(\mathring{T}_R) = \sum_j (A_j+B_j)^2 n_j \sigma_j^2 + \left[\sum_j (A_j\kappa_j - B_j n_j) t_j\right]^2 + \sum_j (N_j - n_j)\sigma_j^2 \qquad (72)$$

Keskineliövirheen komponentit ovat yleisen hajoitelman (7) mukaiset. Kaavasta nähdään, että tasokorjaus kasvattaa ennustimen mallivarianssia, mutta pienentää harhaa.

Tutkimuksessa ei puututa lähemmin ongelmaan arvojen κ_j optimaalisesta määräämisestä. Mainittakoon kuitenkin, että mikäli tasokorjausta (65) ei tehdä, voidaan suoraan soveltaa kappaleen 31. esimerkin tuloksia. Mikäli tasokorjaus tehdään mutkistaa kertoimien B_j rakenne optimointia.

34. Simulointikokeisiin perustuvia tarkasteluja

Tarkastelut tehdään kappaleessa 25. esitellyllä teoreettisella aineistolla samoja otantamenetelmiä ja otosten lukumääriä käyttäen. Ennakkoinformaationa on ollut kertoimella 0.9 korjattu superpopulaatiomalli (35). Yleistä tilavuusfunktiota ei käytetä sen tähden, että sen suhde superpopulaatiomalliin on sattumanvarainen, mikä puolestaan vaikeuttaa tulosten tulkintaa. Käytetyn ennakkoinformaation 10 %:n aliarvio vastaa seuraavan luvun mukaisesti kutakuinkin ennakkoinformaation keskimääräistä harhaa käytännössä.

Taulukossa 6 tarkastellaan ennakkoinformaation vaikutusta, kun käytetään regressio-ennustinta (57). Ennakkoinformaation läpimittajakauma on ollut tasainen ja sen paino on määrätty otoskohtaisesti kaavalla (62), jossa on käytetty todellisia harhan t_i arvoja. Keskineliövirhe MSE_H on laskettu kaavalla (59) teoreettista arvoa $\sigma^2=0.0081$ käyttäen. Taulukossa on otantajakauman suhteen laskettu keskineliövirhe MSE_p, joka on saatu otoksista estimoidun otantavarianssin V_p ja -harhan B_p neliön summana, sekä ennakkoinformaation otoksissa käytetty keskimääräinen paino $E(\kappa)$.

Taulukko 6: Otoksista lasketut tunnukset, kun on käytetty regressioennustinta $\hat{T}_R$, prosenttia tilavuudesta.

	SRS-menetelmä				PPX-menetelmä			
n	B_p	$\sqrt{MSE_p}$	$\sqrt{E(MSE_H)}$	$E(\kappa)$	B_p	$\sqrt{MSE_p}$	$\sqrt{E(MSE_H)}$	$E(\kappa)$
10	-3.01	5.51	5.52	11.44	-0.70	4.68	4.72	1.06
25	-0.72	3.44	3.49	0.61	-0.03	2.85	2.97	0.10
50	-0.28	2.61	2.69	0.14	0.03	1.97	2.02	0.04
100	-0.07	1.94	2.00	0.11	0.00	1.35	1.33	0.02
200	-0.02	1.24	1.29	0.01	0.00	0.82	0.84	0.01
500	-0.03	0.57	0.62	0.00	-0.04	0.34	0.34	0.00

Taulukosta nähdään, että sekä mallin että otantajakauman suhteen lasketut keskineliövirheet ovat samansuuruiset. Taulukkoihin 1 ja 2 verrattuna havaitaan keskineliövirheen voimakas pienentyminen pienillä otoksilla. SRS-menetelmää käytettäessä pieneneminen on ollut voimakkaampaa kuin PPX-menetelmää käytettäessä. Otoskoolla n=10 pieneneminen on ollut suurempi kuin voisi päätellä kaavan (47) ja keskimääräisen κ:n arvon perusteella. Tämä johtuu siitä, että ennakkoinformaatio 'korjaa' voimakkaimmin otoksia, joissa läpimittajakauma on ollut luotettavuuden kannalta epäedullinen. Ennakkoinformaation voidaan siten ajatella korjaavan havaintojen kiintiöinnissä tapahtunutta epäonnistumista. Ennustimen otantaharha pienenee nopeasti otoskoon kasvaessa. Käytännössä otosharhaa voidaan pitää merkittävänä vain kun on käytetty SRS-menetelmää ja otoskoko on ollut n=10.

Taulukossa 7 tarkastellaan, kuinka herkkä keskineliövirheen MSE_p pieneneminen on ennakkoinformaation painon valinnalle.

Taulukko 7: Otoksista estimoitu $\sqrt{MSE}_p$, kun kaavan (62) mukaista κ:n arvoa on muutettu kertoimella b, prosenttia tilavuudesta

	SRS-menetelmä		PPX-menetelmä	
b	n=10	n=25	n=10	n=25
10	6.83	3.67	4.82	2.77
5	6.33	3.47	4.68	2.79
2	5.80	3.40	4.65	2.82
1	5.51	3.44	4.68	2.85
0.5	5.32	3.55	4.73	2.88
0.25	5.22	3.72	4.79	2.91
0.1	5.29	4.05	4.93	2.95

Taulukosta nähdään, että κ voidaan vapaasti valita varsin

laajoissa rajoissa, etenkin, kun käytetään PPX-menetelmää. Taulukosta näkyy kuitenkin suuntaus, että otoskoon ja otantamenetelmän mukaisen keskineliövirheen pienetessä kannattaa κ:n arvoa kasvattaa.

Taulukon 8 tulokset on laskettu käyttäen osite-ennustinta ja sekä tasokorjaamatonta että -korjattua ennakkoinformaatiota. Ennustin on kaavan (69) mukainen, kun kaavaa (45) vastaten on ollut $\kappa_j=\sigma^2 x_j^4/t_j^2$. Keskineliövirhe MSE_H on laskettu kaavalla (72). Harhan t_j, j=1,...,l arvoina on käytetty todellisia arvoja.

Taulukko 8: Otoksista lasketut tunnukset osite-ennustinta $\mathring{T}_R$ käyttäen, prosenttia tilavuudesta.

a) tasokorjaamaton ennakkoinformaatio

	SRS-menetelmä			PPX-menetelmä		
n	B_p	$\sqrt{MSE_p}$	$\sqrt{E(MSE_H)}$	B_p	$\sqrt{MSE_p}$	$\sqrt{E(MSE_H)}$
10	-7.55	7.67	7.61	-7.18	7.32	7.36
25	-5.64	5.85	5.80	-4.88	5.12	5.21
50	-4.16	4.40	4.35	-3.05	3.37	3.47
100	-2.76	3.04	3.02	-1.54	1.93	1.98
200	-1.56	1.91	1.92	-0.57	0.96	1.00
500	-0.58	0.80	0.82	-0.09	0.39	0.36

b) Kaavalla (65) tasokorjattu ennakkoinformaatio

n	B_p	$\sqrt{MSE_p}$	$\sqrt{E(MSE_H)}$	B_p	$\sqrt{MSE_p}$	$\sqrt{E(MSE_H)}$
10	-1.55	4.75	4.71	-0.09	4.58	4.60
25	-1.18	3.20	3.06	0.01	2.86	2.91
50	-1.01	2.34	2.23	0.06	2.01	2.02
100	-0.70	1.70	1.61	0.08	1.38	1.36
200	-0.44	1.16	1.12	0.07	0.84	0.85
500	-0.17	0.59	0.59	0.02	0.33	0.35

Taulukoista havaitaan, että tasokorjaus on pienetänyt huomattavasti ennustimen harhaa ja siten myös keskineliövirhettä MSE_H. Kun verrataan taulukon 8b tuloksia taulukkoon 6, nähdään, että osite-ennustin on ollut SRS-menetelmän tapauksessa regressioennustinta tehokkaampi ja PPX-menetelmää käytettäessä samanveroinen. Vertailussa on kuitenkin huomattava, että käytetty ennakkoinformaatio suosii osite-ennustinta, koska sen muodosta johtuen tasokorjaus (65) on ollut tehokkaimmillaan.

4. Superpopulaatiomallin ja todellisuuden välinen yhteensopivuus

Oletettuun superpopulaatiomalliin saattaa liittyä kahdenlaisia virheitä. Toisaalta voi olla, että mallin muoto on virheellinen, tai toisaalta virhetermin kovarianssirakenne ei vastaa todellisuutta. Tässä kappaleessa selvitetään, miten simulointitesteissä käytetty superpopulaatiomalli

$$H: \quad Y_i = \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \varepsilon_i \qquad (73)$$

vastaa todellisuutta.

Testiaineistona on käytetty kolmasosaa syys-lokakuussa 1979 lasketuista viiden suurimman pystymittaajan leimikoista. Aineistossa oli 774 koepuualuetta. Siitä poimittiin tarkasteluihin ne pääpuulajien kuitu- ja tukkirunkolajit, joista koepuita oli mitattu yli 50 kpl. Kunkin koepuualueen jokaiselle runkolajille estimoitiin malli (73). Malleja estimoitiin kaikkiaan taulukon 9 mukaiset määrät.

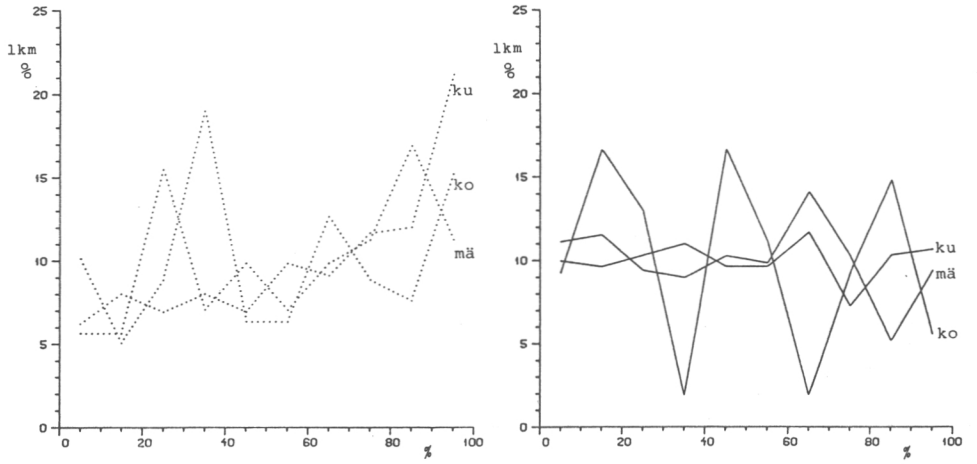
Taulukko 9: Estimoitujen mallien lukumäärä testiaineistossa

	kuitu	tukki
Mänty (mä)	71	234
Kuusi (ku)	274	291
Koivu (ko)	79	59

Koepuiden lukumäärä estimoinneissa vaihteli välillä 50...407.

Mallin oletusten testaamisessa käytettiin runkolajien koepuutietoja. Koepuut on mitattu siten, että niiden läpimitta saattaa saada 57 eri arvoa. Saman läpimitan arvon saaneiden koepuiden avulla voitiin arvioida puhtaan virheen (pure

error) osuus mallin jäännösvaihtelusta ja testata mallin yhteensopivuutta (lack of fit). Kullekin runkolajille laskettiin koepuualueittain F-testisuure (vrt. Draper ja Smith, 1967) ja taulukoitiin F-arvoja vastaavat kertymäfunktion arvot. Nollahypoteesin vallitessa eli, kun malli ja aineisto ovat yhteensopivat ja kun virhetermin jakauma on normaalinen, tulisi kertymäfunktion arvojen jakauma olla tasainen. Kuvaan 5 on piirretty runkolajeja vastaavien jakaumien kuvaajat.

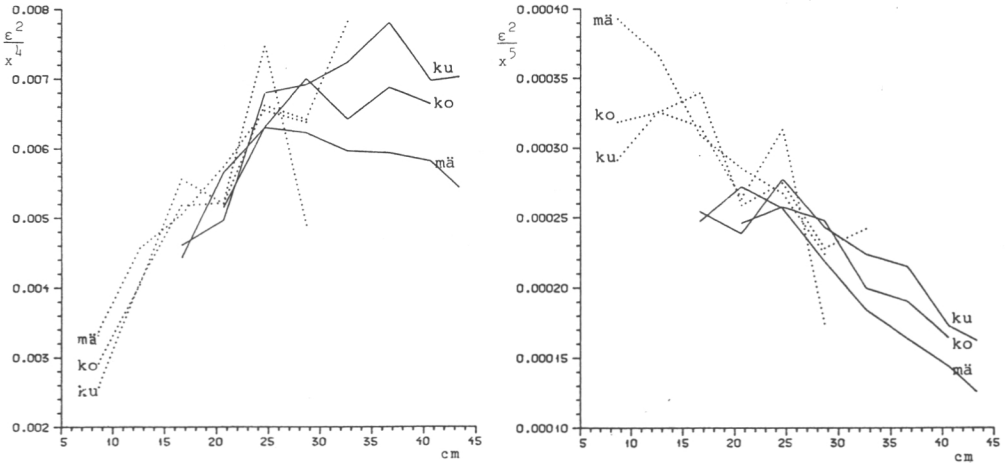


kuiturunkolajit tukkirunkolajit

Kuva 5: Superpopulaatiomallin yhteensopivuustestien F-arvoja vastaavien kertymäfunktion arvojen jakauma.

Kun otetaan huomioon koivun estimoitujen mallien pienestä lukumäärästä johtuva vaihtelu, on yhteensopivuutta pidettävä hyvänä. Ainoastaan kuiturunkolajeilla on ollut jonkin verran odotettua enemmän suuria kertymäfunktion arvoja, mikä osoittaa lievää yhteensopivuuden puutetta.

Virhetermien jakauma testattiin runkolajeittain ja koepuualueittain estimoitujen mallien avulla. Mallien residuaalien neliöt jaettiin vastaavan läpimitan potensseilla 2g=4 ja 5. Kuvassa 6 on runkolajikohtaiset keskiarvot kaikilta koepuualueilta yhteensä. Mikäli keskiarvoja yhdistävät viivat olisivat vaakasuoria, olisi virhetermi verrannollinen kuvaa vastaavaan läpimitan potenssiin. Kuvien perusteella voidaan päätellä eksponentin 2g arvon olevan 4:n ja 5:n välillä, tukkirunkolajeilla lähempänä neljää kuin kuiturunkolajeilla.



2g=4 2g=5

Kuva 6: Estimoiduista malleista laskettujen muunnettujen residuaalien keskiarvot kuiturunkolajeista (···) ja tukkirunkolajeista (——).

Koska puut vaikuttavat toistensa kasvuun ja kasvupaikan laatu vaihtelee alueittain metsikön sisällä, on oletettavissa, että vierekkäisten puiden virhetermit ε_i superpopulaatiomallissa ovat keskenään korreloituneita. Testiaineistosta riippuvuuden voimakkuutta ei ole voitu kuitenkaan estimoida, koska koepuiden sijainti ei ole riittävällä tarkkuudella tiedossa. Kysymystä ei myöskään yleisesti ole tutkittu sillä tarkkuudella, että olisi perusteltua syytä

käyttää muuta kuin diagonaalista painomatriisia ennustimen estimoinnissa.

Jotta saataisiin kuva ennakkoinformaation sopivasta painosta κ, laskettiin testiaineistosta yleisten tilavuusfunktioiden koepuualueiden välinen hajontakomponentti. Tulokset ovat taulukossa 10.

Taulukko 10: Läpimittaan perustuvien puulajikohtaisten tilavuusfunktioiden (Laasasenaho, 1982) koepuualueiden välinen hajontakomponentti, prosenttia tilavuudesta.

	Kuiturunkolajit	Tukkirunkolajit
Mänty	14 %	10 %
Kuusi	10 %	9 %
Koivu	19 %	8 %

Taulukon mukaan harhan määrä on vaihdellut suhteellisen voimakkaasti eri runkolajeilla. Tämä johtunee siitä, että testiaineistossa eri tyyppiset leimikot eivät ole painottuneet keskimääräisellä tavalla. Simulointikokeiden mukaisesti yhdistetyn ennustimen keskineliövirhe ei ole kuitenkaan herkkä painon valinnalle. Näin ollen riittää tuntea vain keskimääräisen harhan suuruusluokka, esimerkiksi taulukon 10 tarkkuudella.

5. Mitä se on? - Pohdintaa

Tutkimuksen teoreettisten tarkastelujen lähtökohtana on käytetty rungon läpimitan ja tilavuuden välisen riippuvuuden määrittelevää polynomista superpopulaatiomallia. Lähtökohta sopii hyvin tarkasteltavaan ongelmaan, jossa estimointia toistetaan leimikolta toiselle. Superpopulaatiomallin voidaan ajatella määrittelevän leimikoiden muodostaman perusjoukon. Tärkein etu tällä lähtökohdalla on, että sen mukaisesti saadaan otantateoria nivottua muuhun tilastollisen päättelyn teoriaan, jolloin voidaan soveltaa yleisiä tilastollisia periaateita ja tuloksia.

Superpopulaatiomallin mukaisesti tutkimuksessa ajatellaan leimikon tilavuuden olevan satunnaismuuttujan ja ongelmana olevan sen arvon ennustaminen. Tästä johtuen käytetään termejä ennustaminen, ennustin ja ennuste termien estimointi, estimaattori ja estimaatti asemesta, joita käytetään totunnaisesti otantateoriassa kiinteäksi ajateltua lukua arvioitaessa. Käytäntö on yleinen superpopulaatiomallioletuksiin perustuvissa teoreettisissa tarkasteluissa.

Ennusteiden luotettavuuden mittana on käytetty superpopulaatiomallioletuksiin perustuvaa keskineliövirhettä MSE_H. Mitta riippuu otoksesta. Tästä syystä mitan mielessä hyvä otantamentelmä on sellainen, joka suurella todennäköisyydellä tuottaa otoksia, joista laskettuna keskineliövirhe on pieni. Tällaisen menetelmän konstruoimiseksi on kiinnostavaa tietää, millainen otos minimoi keskineliövirheen. Simulointikokeiden mukaisesti tällainen otos on kappaleessa 23. määritelty painottaen tasapainoinen otos, s_B, jossa läpimitan tilavuuden hajonnan käänteisluvulla painotetut origomomentit ovat yhtä suuret kuin perusjoukon painottamattomat momentit, ensimmäisestä momentista superpopulaatiomallin asteluvun mukaiseen momenttiin asti. Tällöin en-

nustin supistuu vastaavalla tavalla painotetun otoskeskiarvon kerrannaiseksi ja sen keskineliövirhe supistuu yksinkertaiseen muotoon (20). Supistuksen edellytyksenä on, että ennustimessa on mukana termit, jotka ovat suoraan verrannolliset tilavuuden läpimitan mukaiseen ehdolliseen hajontaan ja varianssiin. Tällainen ennustimen muoto on suositeltava käytännössä.

Edellisen mukaisesti otos tulisi kiintiöidä eri läpimitan arvoilla, koska näin voidaan taata optimaalinen otos, s_B. Silloin, kun kiintiöintiä ei voida tehdä tulisi poimintatodennäköisyydet olla verrannolliset tilavuuden ehdolliseen hajontaan. Tällainen otantamenetelmä tuottaa usein otoksia, jotka ovat lähes optimaalisia. Äskettäin Isaki ja Fuller (1982) ovat osoittaneet, että tällaisen otantamenetelmän yhteydessä edellä suositeltu ennustin on lievin oletuksin näytettävissä parhaaksi otokseen perustuvaksi lineaariseksi harhattomaksi ennustimeksi, kun hyvyyden mittana käytetään sekä otantamenetelmän että superpopulaatiomallin suhteen määriteltyä ennustimen varianssia.

Isaki ja Fuller ovat tutkineet mainitun otantamentelmän ja ennustimen asymptoottisia ominaisuuksia. He tarkastelevat kooltaan kasvavien perusjoukkojen jonoa ja niistä poimittujen otosten jonoa, jossa otoskoko n kasvaa. Heidän tulostensa mukaan tietyin jonoja koskevin säännöllisyysoletuksin on voimassa kappaleen 22. merkintöjä käyttäen

$$E_p[MSE_H(\hat{T})] = MSE_H[\hat{T}(s_B)] + O(n^{-2}) \qquad (74)$$

Toisin sanoen keskineliövirheen otantamenetelmän suhteen laskettu odotusarvo lähenee minimiään vähintään samalla nopeudella kuin n^{-2} pienenee, kun sekä perusjoukon että otoksen koko kasvaa. Tämän mukaisesti voidaan odottaa, että suositeltu otantamenetelmä ja ennustin eli suositeltu otantastrategia toimii hyvin suurilla otoksilla.

Painottaen tasapainoiseen otokseen liittyvien tulosten kautta tässä tutkimuksessa tarkastellulla ennustimella on mielenkiintoinen yhteys ositetun otannan teoriaan. Jos oletetaan, että

1) perusjoukon läpimittajakauma on diskreetti,
2) jokainen läpimitta määrää oman ositteen, jolloin luokkaleveydestä ei aiheudu lisäkomponenttia ositteen varianssiin ja
3) ositevarianssit, klassisen otantateorian mielessä määriteltyinä, ovat verrannolliset superpopulaatiomallin mukaiseen läpimitan potenssiin,

niin tällöin Neymannin allokointi eli allokointi, jossa otatasuhteet ovat verrannolliset ositehajontoihin, tuottaa painottaen tasapainoisen otoksen. Silloin on voimassa:

1) sekä ositekeskiarvoihin perustuva ennustin $\mathring{T}$ että regressioennustin $\hat{T}$ supistuvat samaksi lausekkeeksi, ehdollisten hajontojen käänteisluvulla painotettuun otoskeskiarvoon perustuvaksi ennustimeksi, ja
2) molempien ennustimien keskineliövirheet MSE_H sekä estimaattoreiksi tulkittuina niiden otantavarianssit V_p saavat minimiarvonsa, jotka ovat yhtäsuuret.

Tämän mukaisesti optimissaan molemmat ennustimet ovat yhtä tehokkaat ja yhtä robustit sekä superpopulaatiomallin että otantajakauman suhteen tarkasteltuna. Kun otos ei ole optimaalinen on regressioennustin jonkin verran (n. 10 %) tehokkaampi, kun otoskoko on pieni (alle 50).

Koska mallin suhteen määritelty keskineliövirhe ja klassinen otantavarianssi poikkeavat määrittelyiltään toisistaan, on tutkimuksessa vertailtu niitä keskenään regressioennustimen luotettavuuden mittaajana. Keskineliövirheen MSE_H arvo riippuu otoksen läpimittajakaumasta eli diskreetin jakauman

tapauksessa se on sama samoin kiintiöidyillä otoksilla. Sen luonnollinen vastine on myös kiintiöinnistä riippuva estimaattoriksi tulkitun ennustimen ehdollinen otantavarianssi. Kaavan (29) mukaisesti keskineliövirhe on aina vähintään yhtäsuuri kuin vastaava ehdollinen otantavarianssi. Simulointikokeiden (taulukot 1 ja 2) mukaisesti ero ei kuitenkaan keskimäärin ole kovin suuri, vain muutamia prosentteja. Näin mittoja voidaan käyttää toistensa approksimoinnissa.

Simulointikokeissa (taulukot 1 ja 2) korostuvat myös kaksi erilaista ennusteiden luotettavuuden mittaa, nimittäin keskineliövirhe MSE_H, joka riippuu otoksesta ja kuvaa siten ennustimen luotettavuutta otoksen poimimisen jälkeen sekä keskineliövirheen otantajakauman suhteen laskettu odotusarvo, joka puolestaan kuvaa ennustimen luotettavuutta ennen otantaa. Tulosten mukaisesti nämä saattavat pienillä otoksilla poiketa erittäinkin paljon toisistaan. Tämän mukaisesti, kun ennustamisessa tai estimoinnissa käytetään oheismuuttujia, otantatilanne tulisi pyrkiä mallittamaan siten, että voitaisiin tarkastella myös tulosten otannan jälkeistä luotettavuutta.

Tutkimuksen keskeisenä aiheena ovat olleet tarkastelut, joissa selvitetään, miten esimerkiksi inventoititutkimuksista saadun, laajahkoa aluetta koskevaa ennakkoinformaatiota voidaan hyödyntää yhden leimikon tunnuksia ennustettaesssa. Ongelman asettelusta päädytään ennustimiin, joiden parametrien estimoinnissa käytetään muodollisesti ns. sekaestimaattoreiden kaltaisia lausekkeita. Ennakkoinformaation hyödyntämisessä ilmenee kaksi ongelmaa,

1) millä voimalla ennakkoinformaatio huomioidaan leimikon tunnuksia ennustettaessa ja
2) miten ennakkoinformaatiolla täydennetään otosinformaatiota, sen mahdollisia puutteita.

Ongelmiin annetaan intuitiivisesti perustellut ratkaisut. Keskimääräisten tilavuusfunktioiden tapauksessa, ennakkoinformaatiota voidaan hyödyntää sitä tehokkaammin, mitä suurempi on tilavuuden leimikon sisäisen varianssin suhde leimikoiden väliseen varianssiin. Varianssikomponentteja ei tähän mennessä ole tutkittu. Vasta aivan äskettäin, osittain tämänkin tutkimuksen tuloksiin liittyen, on metsäntutkimuksessa esitetty estimaattoreihin liittyvän virhevarianssin jakoa eri komponentteihin (vrt. Kilkki, 1982). Otantasuunnitelmissa komponentteihin jako vaatii nykyistä enemmän otosyksiköitä eli toistoja leimikoiden sisällä.

Pystymittausongelmassa ennakkoinformaation hyöty käytännössä on, että se suojaa järjettömiltä ennusteilta silloin, kun koepuita on vähän ja läpimittajakauma otoksessa on epäedullinen. Tällainen suoja on tärkeä pystymittauksen kaltaisessa massalaskennassa, jossa on mahdoton käsin tarkistaa kaikkien tulosten mielekkyyttä. Tämän lisäksi ennakkoinformaatiota hyödyntävä ennustin täyttää johdannossa asetetun tavoitteen, että menetelmä toimii koepuumääristä riippumatta ja hyödyntää otosinformaatiota tehokkaasti jo pienilläkin otoksilla.

Superpopulaatiomalleihin perustuvaa otantateoriaa on kritikoitu tulosten herkkyydestä virheellisille mallioletuksille. Eräs keskeisimpiä tehtäviä teorian kehittelyssä onkin ollut löytää tehokas ja robusti otantamenetelmä ja ennustin. Tässä tutkimuksessa on toisaalta selvitetty suositellun otantastrategian herkkyyttä mallioletuksille, ja toisaalta on testattu mallioletusten paikansapitävyyttä käytännössä. Käytännössä sovellettavassa superpopulaatiomallissa oli testeissä havaittavissa lievää yhteensopivuuden puutetta, mutta simulointikokeiden valossa sen voidaan katsoa olevan siksi vähäistä, ettei käytännössä ole odotettavissa syntyvän ongelmia.

LÄHDELUETTELO

BREWER, K.R.W. 1979. A Class of Robust Sampling Designs for Large-Scale Surveys. Journal of the American Statistical Association. Vol. 74. 911-915.

CASSEL, C-M., SÄRNDAHL, G-E. and WRETMAN, J. 1977. Foundations of inference in survey sampling. Wiley. New York.

DE LA GARZA, A. 1954. Spacing of information in polynomial regression. Annals of Mathematical Statistics. Vol. 25. 123-130.

DEMAERSCHALK, J.P. and KOZAK, A. 1974. Suggestions and criteria for more effective regression sampling. Canadian Journal of Forest Research. Vol. 4. 341-348.

DEMAERSCHALK, J.P. and KOZAK, A. 1975. Suggestions and criteria for more effective regression sampling. 2. Canadian Journal of Forest Research. Vol. 5. 496-497.

DRAPER, N. and SMITH, H. 1966. Applied Regression Analysis. Wiley. New York.

ELFVING, G. 1952. Optimum allocation in linear regression theory. Annals of Mathematical Statistics. Vol. 23. 255-262.

ISAKI, C.T. and FULLER, W.A. 1982. Survey Design Under the Regression Superpopulation Model. Journal of the American Statistical Association. Vol. 77. 89-96.

HOEL, P.G. 1958. Efficiency problems in polynomial estimation. Annals of Mathematical Statistics. Vol. 29. 1134-1150.

KOESELOSTUS No. 134. 1979. Pystymittauksen kustannukset ja tuotokset metsähallinnon leimikoilla v. 1978. Metsähallitus, Kehittämisjaosto. Hirvas.

KILKKI, P. 1982. Sample Trees in Volume Estimation. Manuscript.

KOLEHMAINEN, O. 1977. Superpopulaatiomalleihin perustuvasta äärellisen populaation estimointiteoriasta. Vaasan Kauppakorkeakoulun julkaisuja. No 40. Vaasa.

LAASASENAHO, J. 1982. Taper Curves and Volume Functions of Scots Pine, Norway Spruce and Birch. Manuscript.

NOUSIAINEN, J., RANTANEN, V. ja TIIHONEN, P. 1972. Kiintokuutiometrin käyttöön perustuva kuitu- ja tukkipuiden kuutioimismenetelmä. Metsäntutkimuslaitoksen julkaisuja. Vol. 77.2.

ROYALL, R. 1970. On finite population sampling theory under certain linear regression models. Biometrika. Vol. 57. 377-387.

ROYALL, R. and CUMBERLAND, W.G. 1978. Variance Estimation in Finite Population Sampling. Journal of the American Statistical Association. Vol 73. 351-358.

ROYALL, R. and HERSON, J. 1973a. Robust Estimation in Finite Populations I. Journal of the American Statistical Association. Vol. 68. 880-889.

ROYALL, R. and HERSON, J. 1973b. Robust Estimation in Finite Populations II: Stratification on a Size Variable. Journal of the American Statistical Association. Vol. 68. 890-893.

SUKHATME, P.V. and SUKHATME, B.V. 1970. Sampling Theory of Surveys with Applications. Iowa State University Press. Iowa.

TERÄSVIRTA, T. 1981. Some Results on Improving the Least Squares Estimation of Linear Models by Mixed Estimation. Scandinavian Journal of Statistics. Vol 8. 33-38.

VÄLIAHO, H. 1969. A Synthetic Approach to Stepwise Regression Analysis. Societas Scientiarum Fennica. Commentationes Physico-Mathematicae.

Todistusliite

Apulause: $[1\ a']\begin{bmatrix}1 & a' \\ a & A\end{bmatrix}^{-1} = [1\ 0]$

mielivaltaiselle vektorille a ja matriisille A, joille käänteismatriisi on olemassa.

Todistus: Yhtälö nähdään oikeaksi sijoittamalla siihen identiteetti $\begin{bmatrix}1 & a' \\ a & A\end{bmatrix}^{-1} = \begin{bmatrix}1+a'Ba & -a'B \\ -Ba & B\end{bmatrix}$, jossa $B=(A-aa')^{-1}$.

□

Yhtälö (7) s. 10

$$MSE_H[\hat{T}(s)] = V_H[\tilde{T}(s)] + B_H[\hat{T}(s)] + V_H(\Sigma_{\tilde{s}} y_i)$$

Todistus: Jätetään ennustimen argumentti merkitsemättä näkyviin. Tällöin

$$MSE_H(\hat{T}) = E_H(\hat{T}-T)^2 = E_H(\Sigma_s y_i + \tilde{T} - \sum_i^N y_i)^2 =$$

$$E_H\{[\tilde{T} - E_H(\tilde{T})] + E_H[\tilde{T} - \Sigma_{\tilde{s}} y_i] + [E_H(\Sigma_{\tilde{s}} y_i) - \Sigma_{\tilde{s}} y_i]\}^2.$$

Tulos saadaan tämän jälkeen ottamalla huomioon, että virhetermit ε_i ovat riippumattomia ja että $E_H(\tilde{T}-\Sigma_{\tilde{s}} y_i) = E_H(\hat{T}-\sum_1^N y_i)$.

□

Yhtälöiden (17) - (20) todistuksessa käytetään hyväksi apulausetta sekä sitä, että matriisissa $X_s'W_sX_s$ termiä x^{2g} vastaavalla rivillä ovat summat $\Sigma_s x_i^j$, j=1,...,k ja termiä x^g vastaavalla rivillä ovat summat $\Sigma_s x_i^j/x_i^g$, j=1,...,k. Yhtälöiden todistuksessa käytetään merkintää $m = [1\ m_1 \ldots m_k]'$ ja $m(s) = [1\ m_1(s) \ \ldots \ m_k(s)]'$.

Yhtälö (17) s. 12

$\hat{T}(s_b) = N \frac{\Sigma_{s_b} y_i}{n}$, kun ennustimessa on termi x^{2g}.

Todistus: Permutoidaan $\hat{\beta}$:n ja m(s):n elementit niin, että ensimmäinen elementti vastaa termiä x^{2g}. Tällöin

$$m(s_b)'\hat{\beta} = \frac{1}{n}[m_{2g}(s_b)\ 1 \ldots m_k(s_b)] \begin{bmatrix} m_{2g}(s_b) & 1 \ldots m_k(s_b) \\ 1 & \\ \vdots & A \\ m_k(s_b) & \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{s_b} y_i \\ a \\ \end{bmatrix}$$

Erottamalla $m_{2g}(s_b)$ yhteiseksi tekijäksi momenttivektorista ja -matriisista voidaan soveltaa apulausetta, jolloin saadaan

$$m(s_b)'\hat{\beta} = \frac{1}{n} \Sigma_{s_b} y_i .$$

Ennustin $\hat{T}(s_b)$ voidaan tällöin kirjoittaa muodossa

$$\hat{T}(s_b) = (n \cdot m(s_b) + \tilde{x})'\hat{\beta} = N[1\ m_1 \ldots m_k]\hat{\beta} = N \cdot m(s_b)'\hat{\beta} = N \frac{\Sigma_{s_b} y_i}{n} .$$

□

Yhtälö (18) s. 12

$MSE_H[\hat{T}(s_b)] = N^2 \frac{\sigma^2}{n} (1 - \frac{n}{N})\ m_{2g}$, kun ennustimessa on termi x^{2g}. Yhtälö on approksimatiivisesti voimassa, vaikka termi x^{2g} ei olisikaan ennustimessa.

Todistus: Yhtälön (17) johdon kanssa analogisella tavalla nähdään, että

$$\tilde{x}'(X_s'W_sX_s)^{-1}\tilde{x} = (N-n)^2 m'(X_s'W_sX_s)^{-1}m = \frac{(N-n)^2}{n} m_{2g}.$$

Sijoittamalla tämä tulos keskineliövirheen kaavaan (12) saadaan yhtälö (18).

Approksimatiivinen tulos nähdään kirjoittamalla

$$\tilde{x}'(X_s'W_sX_s)^{-1}\tilde{x} = \frac{(N-n)}{n}\,\tilde{x}'b,$$

jossa b on sellaisen mallin kertoimien pns-estimaattori, jossa termiä x^{2g} selitetään termeillä 1, x, ..., x^k. Mitä suurempi on tämän mallin selitysaste, sitä paremmin neliömuoto approksimoi suuretta

$$\frac{N-n}{n}\,\Sigma_{\tilde{s}}x_i^{2g} = \frac{(N-n)^2}{n}\,m_{2g}.$$

□

Yhtälö (19) s. 13

$\hat{T}(s_B) = N\,\dfrac{\Sigma_{s_B} v_i y_i}{\Sigma_{s_B} v_i}$, $(v_i = x_i^{-g})$, kun ennustimessa on termit x^g sekä x^{2g}.

Todistus: Yhtälön (17) johdon mukaisesti voidaan kirjoittaa

$$\hat{T}(s_B) = N\cdot m'\hat{\beta}.$$

Permutoidaan vektorit $\hat{\beta}$ ja m siten, että ensimmäinen elementti vastaa termiä x^g. Tällöin ennustin voidaan kirjoittaa muodossa

$$\hat{T}(s_B) = \frac{N}{\Sigma_{s_B} v_i}\,[m_g\ 1 \ldots m_k]\begin{bmatrix} M_g(s_B) & 1 \ldots M_k(s_B) \\ 1 & \\ \vdots & A \\ M_k(s_B) & \end{bmatrix}^{-1}\begin{bmatrix} \Sigma_{s_B} v_i y_i \\ \\ a \\ \\ \end{bmatrix}$$

Koska $M_j(s_B) = m_j$, $j=1,\ldots,k$, voidaan soveltaa aputulosta ja näin saada yhtälö (19)

□

Yhtälö (20) s 13

$MSE_H[\hat{T}(s_B)] = N^2 \frac{\sigma^2}{n} (m_g^2 - \frac{n}{N} m_{2g})$, kun ennustimessa on termit x^g ja x^{2g}.

Todistus: Merkitään $U=(X_s' W_s X_s)^{-1}$. Tällöin

$$\tilde{x}'U\tilde{x} = [N\cdot m - n\cdot m(s_B)]'U[N\cdot m - n^2\cdot m(s_B)]$$
$$= N^2 m'Um - 2Nn\, m(s_B)'U\, m + n^2 m(s_B)'Um(s_B)$$

Soveltamalla oletusta $M_j(s_B)=m_j$, $j=1,\ldots,k$ sekä apulausetta saadaan

$$\tilde{x}'U\tilde{x} = \frac{N^2 m_g}{\Sigma_{s_B} 1/x_i^g} - 2\,N\,m_{2g} + n\,m_{2g}(s_B).$$

Sijoittamalla lauseke keskineliövirheen kaavaan (12) ja ottamalla huomioon, että

$$\Sigma_{s_B} \frac{1}{x_i^g} = \frac{n}{M_g(s_B)} = \frac{n}{m_g},$$

saadaan

$$MSE_H[\hat{T}(s_B)] = \sigma^2\left[\frac{N^2 m_g^2}{n} - 2\sum_1^N x_i^{2g} + \Sigma_{s_B} x_i^{2g} + \Sigma_{\tilde{s}_B} x_i^{2g}\right]$$
$$= N^2 \frac{\sigma^2}{n} (m_g^2 - \frac{n}{N} m_{2g}).$$

□

Tulos s. 15 (sivun 14 merkinnöin)

Suhteellisesti kiintiöidylle tai Neymannin kiintiöinnin mukaiselle otokselle on voimassa

$$N\, m'(X'\bar{W}X)^{-1}X'\bar{W} = [N_1\ N_2\ \ldots\ N_h].$$

Todistus: 1° Tarkastellaan suhteellisesti kiintiöityä otosta, jossa $m_j(s)=m_j$, $j=1,\ldots,k$. Matriisin $X'\bar{W}$ termiä x^{2g} vastaava rivi on $[n_1\ n_2\ \ldots\ n_h]$. Permutoimalla matriisilauseke siten, että termiä x^{2g} vastaavat alkiot ja rivit tulevat ensimmäisiksi ja soveltamalla apulausetta saadaan

$$N\, m'(X'\bar{W}X)^{-1}X'\bar{W} = \frac{N}{n}[n_1\ n_2\ \ldots\ n_h] = [N_1\ N_2\ \ldots\ N_h].$$

2° Neymannin mukaisesti kiintiöidyssä otoksessa on $M_j(s)=m_j$, $j=1,\ldots,k$. Matriisin $X'\bar{W}$ termiä x^g vastaava rivi on $[n_1/x_1^g\ n_2/x_2^g\ \ldots\ n_h/x_h^g]$. Permutoimalla matriisit sopivasti ja soveltamalla apulausetta saadaan

$$N\cdot m'(X'\bar{W}X)^{-1}X'\bar{W} = \frac{N}{\Sigma_s 1/x_i^g}\left[\frac{n_1}{x_1^g}\ \frac{n_2}{x_2^g}\ \ldots\ \frac{n_h}{x_h^g}\right] = \left[N_1\ N_2\ \ldots\ N_h\right].$$

Jälkimmäinen yhtälö seuraa siitä, että Neymannin kiintiöinnin mukaisessa otoksessa $n_i/N_i \propto x_i^g$.

□

Yhtälöt (32) ja (33) s. 19

$$MSE_H[\hat{\bar{T}}(s_b')] = MSE_H[\hat{\bar{T}}(s_b)] \quad \text{ja} \quad MSE_H[\hat{\bar{T}}(s_B')] = MSE_H[\hat{\bar{T}}(s_B)].$$

Todistus: Osite-ennustimen keskineliövirhe (32) voidaan esittää muodossa

$$MSE_H\,\hat{\bar{T}}(s) = \sigma^2\left[\sum_1^h N_j\frac{N_j}{n_j}x_j^{2g} - \sum_1^h N_j x_j^{2g}\right].$$

1° Sijoittamalla lausekkeeseen $n_j/N_j = n/N$ saadaan yhtälö (18).

2° Sijoittamalla lausekkeeseen $n_j/N_j = nx_j^g/\sum_1^h N_i x_i^g$ saadaan yhtälö (20).

□

Yhtälö (42) s. 28

$$MSE_H[\hat{T}_R(s)] = (N-n)^2 \left[\frac{n\sigma^2}{(n+\kappa)^2} + \frac{\kappa^2}{(n+\sigma)^2}(\mu_o-\mu)^2 + \frac{\sigma^2}{(N-n)} \right] .$$

Todistus:

$$\begin{aligned} MSE_H[\hat{T}_R(s)] &= E_H[\Sigma_s y_i - (N-n)\hat{\mu}_R - \sum_1^N y_i]^2 \\ &= E_H[(N-n)\hat{\mu}_R - (N-n)\mu + \Sigma_{\bar{s}}\varepsilon_i]^2 \\ &= (N-n)^2 MSE_H(\hat{\mu}_R) + (N-n)\sigma^2 \end{aligned}$$

$$\begin{aligned} MSE_H(\hat{\mu}_R) &= E_H\left(\frac{1}{n+\kappa} \Sigma_{s\ i} + \frac{\kappa}{n+\kappa} \mu_o - \mu \right)^2 \\ &= E_H\left[\frac{1}{n+\kappa} \Sigma_s \varepsilon_i + \frac{\kappa}{n+\kappa} (\mu_o-\mu) \right]^2 \\ &= \frac{n}{(n+\kappa)^2} \sigma^2 + \frac{\kappa^2}{(n+\kappa)^2} (\mu_o-\mu)^2 \end{aligned}$$

Yhtälö (42) saadaan sijoittamalla $MSE_H(\hat{\mu}_R)$:n lauseke $MSE_H[\hat{T}_R(s)]$:n kaavaan.

□

Yhtälö (58) s. 33

$$B_H[\hat{T}_R(s)] = \tilde{x}' E_H(\hat{\beta}_R-\beta) = \kappa \tilde{x}' U R' w_r t.$$

Todistus: Merkitään

$$Z = \begin{bmatrix} X_s \\ R \end{bmatrix}, \quad u = \begin{bmatrix} y_s \\ r \end{bmatrix} \text{ ja } W_Z = \begin{bmatrix} W_s & 0 \\ 0 & \kappa W_r \end{bmatrix}$$

Tällöin $\hat{\beta}_R=(Z'W_Z Z)^{-1}Z'W_Z u$ ja (53):n mukaan $u=Z\beta+\begin{bmatrix} \varepsilon \\ t \end{bmatrix}$, jossa $E_H[\varepsilon\ t]' = [0\ t]'$. Saadaan

$$\begin{aligned} E_H(\hat{\beta}_R-\beta) &= E_H[(Z'W_Z Z)^{-1}Z'W_Z(Z\beta+\begin{bmatrix} \varepsilon \\ t \end{bmatrix})-\beta] = \kappa(Z'W_Z Z)^{-1}R'W_r t \\ &= \kappa U R' W_r t. \end{aligned}$$

□

Yhtälö (59) s. 33

$$\begin{aligned} MSE_H[\hat{T}_R(s)] &= \tilde{x}'MSE_H(\hat{\beta}_R)\tilde{x} + \sigma^2 \sum_{\tilde{s}} x_i^{2g} \\ &= \sigma^2 \tilde{x}' U X_s' W_s X_s U \tilde{x} \\ &+ \kappa^2 (\tilde{x}' U R' W_r t)^2 \\ &+ \sigma^2 \sum_{\tilde{s}} x_i^{2g}. \end{aligned}$$

Todistus: Edellisen yhtälön todistuksen merkintöjä käyttäen

$$\begin{aligned} MSE_H(\hat{\beta}_R) &= E_H(\hat{\beta}_R - \beta)(\hat{\beta}_R - \beta)' \\ &= U X_s' W_s \, E_H(\varepsilon\varepsilon') \, W_s X_s U + U R' W_r t t' W_r R U. \end{aligned}$$

Yhtälö (59) saadaan ottamalla huomioon, että $E_H(\varepsilon\varepsilon') = \sigma^2 W_s^{-1}$ sekä käyttämällä yhtälöä (58)

□

Yhtälö (69) s. 39

$$\hat{\bar{T}}_R(s) = \sum_j \left[\sum_i y_{ij} + (A_j + B_j) \sum_i y_{ij} + (A_j \kappa_j + B_j n_j) r_j \right]$$

jossa

$$A_j = \frac{N_j - n_j}{n_j + \kappa_j} \quad \text{ja} \quad B_j = \frac{1}{x_j^g} \sum_i A_i \frac{\kappa_i x_i^g}{n}$$

Todistus: Sijoitetaan yhdistetty estimaattori (67) ennustimen $\hat{\bar{T}}_R$ lausekkeeseen (68). Tällöin saadaan

$$\hat{\bar{T}}_R(s) = \sum_j \sum_i^{n_j} y_{ij} + \sum_j \frac{N_j - n_j}{n_j + \kappa_j} \left[\sum_i^{n_j} y_{ij} + \kappa_j r_j + \frac{\kappa_j x_j^g}{n} \sum_l n_l \frac{\bar{y}_l - r_l}{x_l^g} \right]$$

Vaihtamalla summausjärjestystä jälkimmäisessä termissä saadaan

$$\mathring{T}_R(s) = \sum_j \sum_i y_{ij} + \sum_j \left[\frac{N_j - n_j}{n_j + \kappa_j} \sum_i y_{ij} + \frac{N_j - n_j}{n_j + \kappa_j} \kappa_j r_j \right.$$

$$\left. + n_j \frac{\bar{y}_j - r_j}{x_j^g} \sum_l \frac{N_l - n_l}{n_l + \kappa_l} \frac{\kappa_l x_l^g}{n} \right]$$

$$= \sum_j [\sum_i y_{ij} + (A_j + B_j) \sum_i y_{ij} + (A_j \kappa_j + B_j n_j) r_j].$$

□

Yhtälö (71) ja (72) s. 39,40

$$B_H(\mathring{T}_R) = \sum_j (A_j \kappa_j - B_j n_j) t_j$$

ja

$$MSE_H(\mathring{T}_R) = \sum_j (A_j + B_j)^2 n_j \sigma_j^2$$

$$+ \left[\sum_j (A_j \kappa_j - B_j n_j) t_j \right]^2$$

$$+ \sum_j (N_j - n_j) \sigma_j^2$$

Todistus: Ennustimen termien järjestystä vaihtamalla nähdään, että

$$\mathring{T} - T = \sum_j (A_j \sum_i y_{ij} + A_j \kappa_j r_j) + \sum_j B_j (\sum_i y_{ij} - n_j r_j) + \sum_j \sum_{\tilde{s}_j} (\mu_j + \varepsilon_i)$$

jossa symbolilla $\tilde{s}_j$ on merkitty j. ositteen ei-koepuiden joukkoa. Edelleen kehittelemällä saadaan

$$\mathring{T} - T = \sum_j A_j [\sum_i y_{ij} + \kappa_j r_j - (n_j + \kappa_j) \mu_j] + \sum_j B_j (\sum_i y_{ij} - n_j r_j) + \sum_j \sum_{\tilde{s}_j} \varepsilon_i$$

$$= \sum_j A_j [\kappa_j (r_j - \mu_j) + \sum_{s_j} \varepsilon_i] + \sum_j B_j [n_j (\mu_j - r_j) + \sum_{s_j} \varepsilon_i] + \sum_j \sum_{\tilde{s}_j} \varepsilon_i$$

$$= \sum_j [A_j \kappa_j - B_j n_j] (r_j - \mu_j) + \sum_j (A_j + B_j) \sum_{s_j} \varepsilon_i + \sum_j \sum_{\tilde{s}_j} \varepsilon_i .$$

Yhtälöt (71) ja (72) saadaan kehitetyn lausekkeen ja sen neliön odotusarvona mallin suhteen määriteltynä

□

ISSN 0358-4283