Juha Heikkinen

# Tilastolliset mallit ja metsien inventoinnin luotettavuus

teema

Metsätieteellisen Seuran vuosijuhla tarjoaa oivallisen tilaisuuden Suomen valtakunnallisten metsäinventointien alkuvaiheiden muistelemiseen. Seura nimittäin edisti niitä varsin merkittävällä tavalla.

Vuonna 1912 Metsänhoitoyhdistys Tapio pani alkuun tutkimuskokeen, jossa oli tavoitteena selvittää linja-arvioinnin soveltuvuutta ja tarkkuutta Suomen yksityismetsien tuotannon ja kulutuksen inventoinnissa. Koealueiksi valittiin Sahalahden ja Kuhmalahden pitäjät Hämeestä, ja työn suorittaminen annettiin Verner Cajanuksen huoleksi. Cajanuksella ei kuitenkaan koskaan ollut tilaisuutta saattaa tutkimusta loppuun. Hänen kuoltuaan Metsätieteellinen Seura otti vuonna 1920 työn loppuun saattaakseen ja uskoi toteutuksen Yrjö Ilvessalolle (Ilvessalo 1923).

Sahalahden ja Kuhmalahden pilotti-inventoinneissa kehitettyä metodiikkaa sovellettiin pienin muutoksin neljässä ensimmäisessä valtakunnan metsien inventoinnissa (Ilvessalo 1927, 1943, 1956, 1962). Tulosten luotettavuuden arviointimenetelmän kehittämiseksi Ilvessalo kääntyi matematiikan professori J.W. Lindebergin puoleen. Lindebergin sekä hänen ruotsalaisten ja norjalaisten samojen ongelmien kanssa painivien kollegojensa toimesta alkoi – siis metsien inventoinnin motivoimana – tutkimus (esim. Lindeberg 1924, 1926, Langsaeter 1926, 1932, Östling 1932), joka loi pohjaa linja-arviointimenetelmien, systemaattisen otannan ja spatiaalisen tilastotieteen kehitykselle paljon yleisemminkin.

Viidennessä inventoinnissa (1964–70) pätkittiin koko maan yli ulottuneet yhtenäiset arviointilinjat arviolta yhden maastotyöpäivän mittaisiksi lohkoiksi, jotka sijoiteltiin säännöllisesti inventointialueen kattavan neliöhilan muotoon (Kuusela ja Salminen 1969). Tämän muutoksen myötä luotettavuudenarviointiongelma muuttui tilastolliselta kannalta aidosti spatiaaliseksi, ja vahvan todennäköisyysteoreettisen perustan sen ratkaisemiseksi kehitti Bertil Matérn väitöskirjassaan (Matérn 1960). Taaskin inventointiprofessorin – tällä kertaa Manfred Näslund Ruotsista – asettamat ongelmat johtivat klassikkoteokseen tilastotieteen alalla. Matérnin työ on merkittävästi vaikuttanut modernin spatiaalisen tilastotieteen kehitykseen, ja ainakin Suomen ja Ruotsin inventoinneissa Matérnin esittämät luotettavuudenarvointimenetelmät ovat edelleen käytössä (ks. esim. Tomppo ym. 1998, 311–312).

Tämän esityksen tavoitteena on edellämainittuja töitä esitellen valottaa pääkohtia otantaan perustuvien metsäinventointien luotettavuustarkastelujen problematiikasta ja tilastollisten mallien käytön perusteista tässä yhteydessä.

## Linja-arvioinnit

Suomen metsien neljässä linja-arvioinnissa inventoitiin kussakin 50–100 maan yli lounaasta koilliseen ulottuvaa linjaa. Mittausten ja laskennan yksityiskohtiin sen tarkemmin puuttumatta käsitellään tässä esimerkkinä metsämaan osuuden $P$ arviointia. Sitä arvioitiin yksinkertaisesti metsämaan kuvioita leikkaavien linjanosien osuudella $\hat{P}$ linjas-

ton kokonaispituudesta. Mutta, kuten Ilvessalo (1923) kirjoitti Sahalahti-Kuhmalahti-raportissaan: ”Tunnettaessa linja-arvioimisen mukaan ainoastaan tiluslajin suhteellista esiintymistä osottava keskiarvoprosentti on tulos vähäarvoinen; on aivan välttämätöntä, että lisäksi tunnetaan tämän keskiarvoprosentin luotettavuus l. tarkkuus, s.o. se mahdollinen virhe, mikä keskiarvoon liittyy.”

Tilastotieteen keinoin luotettavuuden arviointia voi lähestyä esimerkiksi seuraavalla tavalla (ks. esim. Salminen 1973, 13–14). Kuvitellaan, että linja-arviointia toisettaisiin useita kertoja koko linjastoa satunnaisesti siirrelleen linjojen suunnan ja välimatkan pysyessä samana. Tällöin saataisiin metsämaan osuudelle yleensä hieman erilaisia arvioita

$$\hat{P}_1, \hat{P}_2, \ldots \tag{1}$$

Kun toistojen määrää $K$ kasvatetaan rajatta, lähestyy arvioiden keskiarvo

$$\frac{1}{K}\sum_{k=1}^{K}\hat{P}_k \tag{2}$$

jotain tiettyä lukua. Tätä käytännössä yleensä tuntematonta lukua sanotaan osuusestimaattorin $\hat{P}$ *odotusarvoksi* $E(\hat{P})$. Jos mittauksiin ei liity minkäänlaista systemaattista virhettä, sanotaan estimaattoria $\hat{P}$ *harhattomaksi*, ja sen odotusarvo on metsämaan todellinen osuus. Hypoteettisista toistoista saatavien estimaattien $\hat{P}_k$ keskihajonta

$$\sqrt{\frac{1}{K}\sum_{k=1}^{K}\left\{\hat{P}_k - E(\hat{P})\right\}^2} \tag{3}$$

kuvaa osuusestimaattorin luotettavuutta. Sen rajaarvoa toistojen määrän kasvaessa sanotaan estimaattorin $\hat{P}$ *keskivirheeksi* $s(\hat{P})$.

Keskivirhettä käyttäen voidaan yhdellä luvulla kuvata inventointitulokseen liittyvää epävarmuutta, minkä takia se on käytännöllinen käsite esimerkiksi erilaisia otanta-asetelmia vertaillessa. Sinänsä sillä ei kuitenkaan ole selvää suoraa tulkintaa, mutta sen avulla voidaan konstruoida ns. *luottamusvälejä*. Keskeisen raja-arvolauseen nojalla – jonka kehityksessä Lindebergillä oli muuten myöskin keskeinen rooli (Lindeberg 1922) – keskiarvotyyppiset estimaattorit ovat hyvin yleisin ehdoin asymptoottisesti, eli suurilla havaintomäärillä likimain, normaalisti jakautuneita. Tällöin 95 %:n luottamusväli

$$\left[\hat{P}_k - 2s(\hat{P}),\ \hat{P}_k - 2s(\hat{P})\right] \tag{4}$$

sisältää odotusarvon $E(\hat{P})$ 95 % toistoista. Toisin sanoen harhattoman asetelman tapauksessa on 95 %:n todennäköisyydellä valittu sellainen otos, josta laskettu luottamusväli sisältää metsämaan todellisen osuuden.

Ongelma on tietysti se, että keskivirhettäkään ei tunneta, kun otoksia on käytännössä vain yksi. Täydellisesti satunnaistetun otannan tapauksessa keskivirheeseen päästään kuitenkin helposti käsiksi yksittäisten havaintojen vaihtelua tarkastelemalla. Jos arvio $\hat{P}$ on keskenään riippumattomien havaintojen $p_n$, $n = 1,...,N$, keskiarvo, niin sen keskivirhettä voidaan arvioida yksittäisten havaintojen keskihajonnan

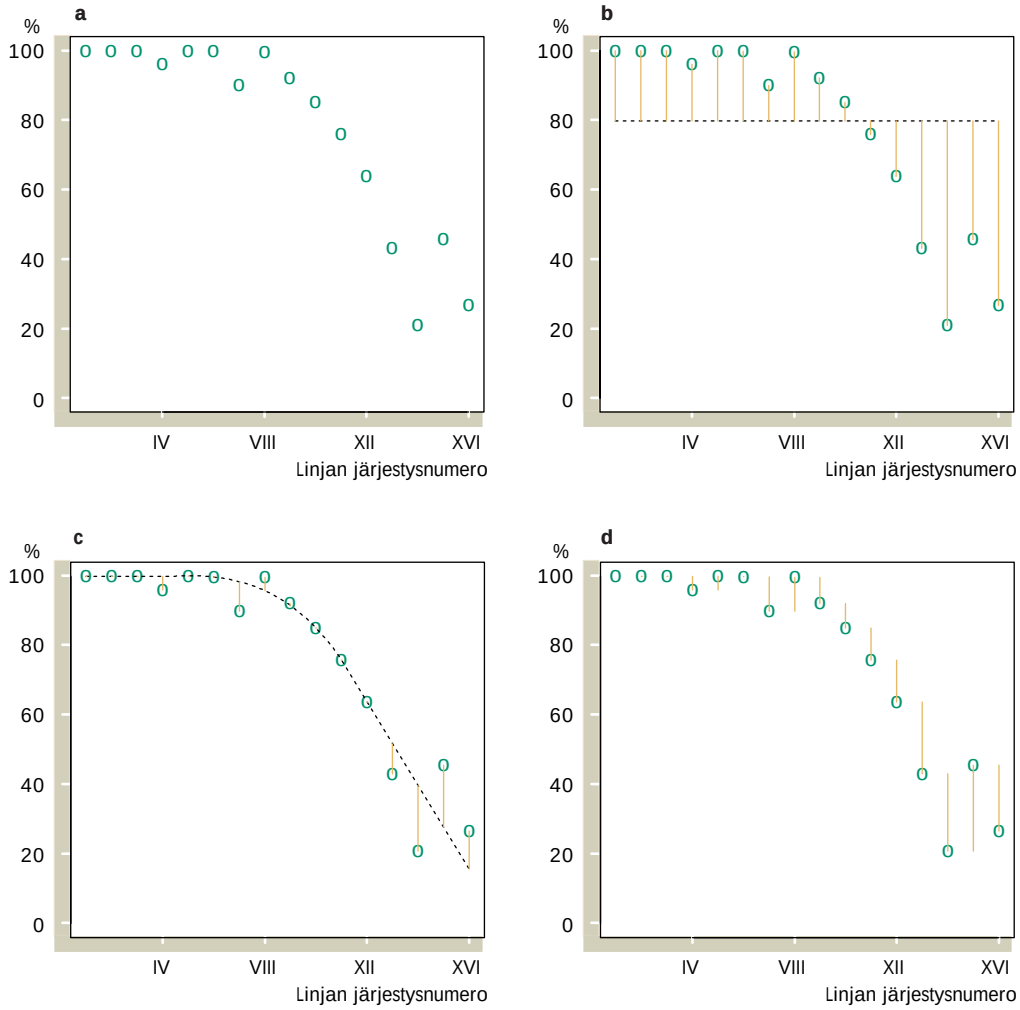
$$s(p) = \sqrt{\frac{1}{N}\sum_{k=1}^{K}\left\{p_n - \hat{P}\right\}^2} \tag{5}$$

avulla:

$$\hat{s}(\hat{P}) = \frac{s(p)}{\sqrt{N}} \tag{6}$$

Tässä klassisessa tapauksessa luotettavuuden arviointi perustuu siis olennaisesti ottaen yksittäisten havaintojen $p_n$ poikkeamiin otoskeskiarvosta $\hat{P}$.

Linja-arvioinnin tapauksessa ”yksittäinen havainto” $p_n$ voisi olla esimerkiksi metsämaan osuus yhden linjan pituudesta (jolloin $\hat{P}$ olisi luonnollisesti linjan pituuksilla *painotettu* keskiarvo). Riippumattomuusoletuksen realistisuuden suhteen metsien inventointi eroaa kuitenkin radikaalisti tavallisesta otantatutkimuksesta. Useimmilla metsämuuttujilla on voimakas (positiivinen) *spatiaalinen autokorrelaatio*, t.s. lähekkäin sijaitsevat kohteet ovat keskimäärin ”samanlaisempia” kuin kauempana toisistaan sijaitsevat. Sen takia systemaattinen otanta – s.o. tasavälein sijoitetut linjat – on huomattavasti tehokkaampi kuin täysin satunnaistettu, ja satunnaisotannan oletukseen perustuvat keskivirhearviot ovat liian pessimistisiä.




**Kuva 1.** a) Metsämaan osuus Ilvesvuoren alueen linjoilla Sahalahden ja Kuhmalahden pitäjien inventoinnissa 1912 (Ilvessalo 1923, Taulukko 1 ja kuva 23), sekä b) näiden poikkeamat koko linjaston pituudesta lasketusta metsämaan osuudesta, c) systemaattista muutosta kuvaavasta ”keskiarvoviivasta” ja d) edellisen linjan osuudesta.

Sahalahti-Kuhmalahti-raportissa tätä asiaa selvitettiin ja havainnollistettiin oivallisesti ja seikkaperäisesti (Ilvessalo 1923, 24–30). Siinä julkaistun aineiston pohjalta on uudelleenpiirretty kuva 1a, jossa esitetään metsämaan osuudet 16 Ilvesvuoren osa-alueen linjalta Kuhmalahden pitäjästä. Satunnaisotantaan pohjautuvan keskivirhearvion käyttö tarkoittaisi tässä tapauksessa sitä, että otannasta johtuvan satunnaisvaihtelun mitta perustettaisiin linjakohtaisten osuuksien poikkeamiin koko alueelta laskettua metsämaan osuutta kuvaavasta vaakasuorasta viivasta (kuva 1b).

Ilvesvuoren alueella metsämaan osuus näyttää kuitenkin selvästi systemaattisesti vähenevän tutkimusalueen toista päätä kohti. Tällainen trendi on spatiaalisen autokorrelaation tyypillinen ilmenemismuoto, ja se saadaan luonnollisesti systemaattisella otannalla tarkemmin kuvattua kuin jos linjojen paikat olisi satunnaisesti valittu. Sen takia poikkeamat keskiarvosta antavat linja-arvioinnin tarkkuudesta

liian pessimistisen kuvan, ja Ilvessalo – ilmeisesti Cajanuksen ajatusten mukaisesti – esitti, että vaakasuoran viivan sijasta pyrittäisiin hakemaan ”tiluslajin esiintymistä osottava eksaktinen viiva, joka syntyy, jos linjojen väli on äärettömän pieni” (Ilvessalo 1923, 26). Kuvassa 1c on pyritty mahdollisimman hyvin toistamaan Ilvessalon esitys (Ilvessalo 1923, kuva 23) tällaiseksi viivaksi kuvan 1a aineistoon sovitettuna, seuraten ”vain sellaisia vaihteluita, jotka näyttävät todella systemaattisilta” (Ilvessalo 1923, 27). Sahalahti-Kuhmalahti-inventoinnin luotettavuusarviot perustuivat linjoittaisiin poikkeamiin tällaisista ”keskiarvoviivoista” (kuva 1c). Ilvessalo (1923, 26) perusteli: ”Näiden poikkeuksien suuruutta saattaa pitää epävarmuuden mittana; ne […] osottavat sitä tilapäistä vaihtelua suhteellisessa esiintymisessä, joka riippuu käytettyjen linjojen asemasta. Näin menetellen saadaan siis eliminoitua pois eri linjojen välillä olevat systemaattiset eroavaisuudet ja sitä tietä keskivirhe oikein lasketuksi.”

Ilvessalo myönsi kuitenkin itsekin, että keskiarvoviivan määrääminen ”on luonnollisesti enemmän tai vähemmän mielivaltainen” (Ilvessalo 1923, 27). Toisaalta hän painotti, että ”jos vaihtelut eivät selvästi ole systemaattisia, ei niitä tasoteta […] Tällä tavalla menetellen, siis määräten funktio ylimalkaisesti, saadaan keskivirhe, joka ei ilmota tarkkuutta ainakaan liian suureksi” (Ilvessalo 1923, 27). Tämä konservatiivisuusperiaate on ollut tärkeänä prinsiippinä myöhemmässäkin kehityksessä.

Lindeberg esitti elegantin vaihtoehdon, jossa keskiarvoviivaa ei tarvittu ja mielivaltaisuutta saatiin ainakin vähennettyä (Lindeberg 1924, 1926). Poikkeamat keskiarvoviivasta korvattiin yksinkertaisesti kahden vierekkäisen linjan välisellä erotuksella (kuva 1d). Tätä *differensointitekniikkaa* käytettiin Suomen kaikkien neljän valtakunnallisen linjainventoinnin luotettavuuden arvioinnissa, ja se on varsin laajasti käytössä nykyaikaisessakin tilastotieteessä. Esimerkiksi aikasarja-analyysin ehkä suosituimmassa työkalussa, ARIMA-malleissa (Box ja Jenkins 1976), se on olennainen komponentti.

## Lohkoinventoinnit

Edellä kuvattua lähestymistapaa inventoinnin luotettavuusanalyysiin voidaan luonnehtia lähinnä (otanta-)asetelmaperusteiseksi. Matérnin työ (Matérn 1960) on puolestaan selkeästi mallipohjainen. Tarkastellaan esimerkkinä jälleen metsämaan osuuden $P(A)$ arviointia inventointialueella $A$, ja määritellään funktio $M$ siten, että $M(x)$ saa arvon 1, jos piste $x \in A$ on metsämaalla ja arvon 0, jos ei ole. Funktion $M$ kuva on siis inventointialueen metsämaakartta, ja koealainventoinnilla saatava luonnollinen keskiarvoestimaatti metsämaan osuudelle voidaan ilmaista muodossa

$$\hat{P}(A) = \frac{1}{N}\sum_{n=1}^{N} M(x_n) \qquad (7)$$

missä $x_1,...,x_N$ ovat koealakeskipisteitä.

Matérnin tarkastelut perustuivat hyvin yleiseksi määriteltyyn funktion $M$ tilastolliseen malliin. Täsmällisemmin sanoen hän oletti, että tämä funktio on realisaatio stationaarisesta ja isotrooppisesta stokastisesta prosessista, jolla on laskeva korrelaatiofunktio (Matérn 1960, 70). Tämän määrittelyn yksityiskohtiin takertumatta olennaista on, että malli kuvaa inventoinnin kohteena olevan ilmiön satunnaisuutta, kun taas asetelmaperusteisessa lähestymistavassa keskivirheen arviointi perustuu otoksen poimintaan liittyvän satunnaismekanismin malliin. Mallipohjaisella tavalla voidaan osittain välttää systemaattisen otannan rajoitetusta satunnaisuudesta johtuvat ongelmat.

Konkreettisemman kuvan saamiseksi lähestymistapojen erosta kuvitellaan, että haluttaisiin tehdä simulointikoe keskivirhe-estimaattorin ominaisuuksien tutkimiseksi. Asetelmapohjaisessa lähestymistavassa pitäisi ensin valita yksi metsämaakartta, josta sitten poimittaisiin otoksia koealaverkkoa satunnaisesti siirrellen. Mallipohjaisessa lähestymistavassa riittää puolestaan varsin väljästi määritelty metsämaakartan oletettuja ominaisuuksia kuvaava malli, jonka pohjalta voidaan tuottaa erilaisia karttoja ja tutkia miten näistä saadut arviot vaihtelevat. Jälkimmäinen tapa on selvästi helpommin yleistettävissä. Tämän tyyppisiä kokeita Matérn toteuttikin saaden varsin vahvoja, yleispäteviä ja hyödyllisiä tuloksia esimerkiksi otanta-asetelman valintaan, lohkojen muodon ja koon vaikutukseen ja erilaisten keskivirhe-estimaattorien ominaisuuksiin liittyen (Matérn 1960, kappaleet 5, 6.6, 6.8 ja 6.10).

Suomen nykyisen maastoaineistoon perustuvan

inventointinnin luotettavuusarviot perustuvat yhteen Matérnin esittämistä keskivirhe-estimaattoreista (Salminen 1973, Tomppo ym. 1998, 311–312). Se on Lindebergin differenssimenetelmän suhteellisen yksinkertainen yleistys aidosti spatiaaliseen tilanteeseen. Linja-arvioinnin kahden peräkkäisen linjan välisen poikkeaman sijasta käytetään neljän vierekkäisen lohkon

$$\begin{matrix} c3 & c4 \\ c1 & c2 \end{matrix} \tag{8}$$

ryhmässä laskettua erotusta

$$\frac{1}{2}(p_{c1} + p_{c4}) - \frac{1}{2}(p_{c2} + p_{c3}) \tag{9}$$

joka voidaan tulkita poikkeamaksi kahden neliöryhmän keskipisteen havaintoon kohdistuvan lineaarisen ennusteen välillä. Edellä kuvattua mallikehikkoa hyväksikäyttäen Matérn osoitti tämäntyyppisten neliömuotoestimaattorien tuottavan hyvin yleisillä oletuksilla yliarvion keskivirheelle (Matérn 1960, 111, 116).

## Yhteenveto

Lindebergin ja Matérnin luotettavuudenarviointimenetelmillä on kaksi tärkeää yhteistä piirrettä. Ensinnäkin itse keskivirhearvioiden laskemiseksi käsillä olevasta aineistosta ei tarvita minkäänlaista tilastollista mallinnusta. Näin ollen ne ovat erityisen hyvin soveltuvia lukuisia tunnuksia käsittäviin operatiivisiin inventointisysteemeihin.

Toisaalta keskivirhe-estimaattorien ominaisuuksien tarkasteluun tarvitaan varsin syvällistä tilastollista analyysiä. Sen avulla on voitu erityisesti osoittaa että inventointitulosten liian tarkaksi väittämisen vaaraa ei yleensä ole, jos lähtöaineiston mahdolliset systemaattiset virheet voidaan olettaa mitättömiksi.

## Viitteet


Box, G.E. & Jenkins, G.M. 1976. Time series analysis: forecasting and control. Holden Day, San Francisco.

Ilvessalo, Y. 1923. Tutkimuksia yksityismetsien tilasta Hämeen läänin keskiosissa. Sahalahden ja Kuhmalahden pitäjien metsät. Acta Forestalia Fennica 26.

— 1927. Suomen metsät. Tulokset vuosina 1921–1924 suoritetusta valtakunnan metsien arvioimisesta. Metsätieteellisen koelaitoksen julkaisuja 11.

— 1943. Suomen metsävarat ja metsien tila. II valtakunnan metsien arviointi. Metsätieteellisen tutkimuslaitoksen julkaisuja 30.

— 1956. Suomen metsät vuosista 1921–24 vuosiin 1951–53. Kolmeen valtakunnan metsien inventointiin perustuva tutkimus. Metsäntutkimuslaitoksen julkaisuja 47(1).

— 1962. IV valtakunnan metsien inventointi. 1. Maan eteläpuoliskon vesistöalueryhmät. Metsäntutkimuslaitoksen julkaisuja 56(1).

Kuusela, K. & Salminen, S. 1969. The 5th national forest inventory of Finland. General design, instructions for field work and data processing. Metsäntutkimuslaitoksen julkaisuja 69(4).

Langsaeter, A. 1926. Om beregning av middelfeilen ved regelmessige linjetaksering. Meddel. fra det norske Skogforsøksvesen 2 h. 7: 5–47.

— 1932. Nøiaktigheten ved linjetaksering av skog, I. Meddel. fra det norske Skogforsøksvesen 4: 431–563.

Lindeberg, J.W. 1922. Eine neue Herleitung des Exponential Gesetzes in der Wahrshceinlichkeitsrechnung. Matematische Zeitschrift 15: 211–225.

— 1924. Über die Berechnung des Mittelfehlers des Resultates einer Linientaxierung. Acta Forestalia Fennica 25.

— 1926. Zür Theorie der Linientaxierung. Acta Forestalia Fennica 31.

Matérn, B. 1960. Spatial variation. Meddelanden från Statens Skogsforskningsinstitut 49(5).

Salminen, S. 1973. Tulosten luotettavuus ja karttatulostus valtakunnan metsien V inventoinnissa. Metsäntutkimuslaitoksen julkaisuja 78(6).

Tomppo, E., Henttonen, H., Korhonen, K.T., Aarnio, A., Ahola, A., Heikkinen, J., Ihalainen, A., Mikkelä, H., Tonteri, T. & Tuomainen, T. 1998. Etelä-Pohjanmaan metsäkeskuksen alueen metsävarat ja niiden kehitys 1968–97. Metsätieteen aikakauskirja – Folia Forestalia 2B/1998: 293–374.

Östling, J. 1932. Erforderlig taxeringsprocent vid linjetaxering av skog. Sveriges Skogsvårdsföreningens Tidskrift 30.


FT Juha Heikkinen (juha.heikkinen@metla.fi) toimii tutkijana Metlan Helsingin tutkimuskeskuksessa.