Mat-2.108 Sovelletun matematiikan erikoistyö
Spatiaalisen autokorrelaation testaaminen

Esa-Pekka Horttanainen
41867M

18. syyskuuta 2003

Sisältö

1 Johdanto

Spatiaalinen tilastoanalyysi on yksi tilastotieteen pitkälle erikoistunut osa-alue. Siinä tutkitaan muun muassa kappaleiden sijainnin spatiaalisia riippuvuuksia, eri tyyppisten kappaleiden sijaintien välistä korrelaatiota tai jonkin ilmiön spatiaalista autokorrelaatiota. Spatiaalinen autokorrelaatio tarkoittaa, että toisiaan lähellä olevat alueet ovat mitatun ilmiön suhteen samankaltaisempia kuin kauempana sijaitsevat

Spatiaalisen tilastoanalyysin sovellusalueita ovat muun muassa terveydenhuolto, ympäristötaloustiede ja paikkatietojärjestelmät. Sovelluksia ovat esimerkiksi terveysilmiöiden tai sosiaalisten ongelmien alueellinen tutkiminen, liikeyritysten markkina-alueiden tutkiminen sekä spatiaalisesti jakautuneiden ilmiöiden ajallinen vertailu.

Spatiaalista autokorrelaatiota voidaan käyttää hyväksi esimerkiksi jatkuvan ilmiön (korkeusmalli, kasvillisuus, meritutkimukset) interpolointiin (kriging). (Virrantaus 2001)

Tässä työssä esitellään yksi tapa testata (Mantel-testi), tulkita (variogrammi) ja hyödyntää (kriging) spatiaalista autokorrelaatiota. Menetelmien esittely pohjautuu Manlyn (2001) kirjan esitykseen.

Työssä käy ilmi, että spatiaalisen autokorrelaation testaaminen ei aina suju ongelmitta.

2 Spatiaalinen autokorrelaatio

Spatiaalisen tilastoanalyysin yksi sovellus on mitatun ilmiön spatiaalisen autokorrelaation tutkiminen. Spatiaalinen autokorrelaatio tarkoittaa, että toisiaan lähellä olevat alueet ovat jonkin ilmiön suhteen samankaltaisempia kuin kauempana sijaitsevat (Virkkala ym. 2000, s. 11). Nyqvistin (2002, s. 17) mukaan monien spatiaalisten aineistojen tapauksissa käytettävissä olevat muuttujat eivät selitä vasteen jakauman havaittua maantieteellistä yhtenäisyyttä, jolloin autokorrelaation huomioiva spatiaalinen malli voi tuottaa parempia ennusteita, vaikka todellista autokorrelaatiota ei olisikaan.

2.1 Mantel-testi autokorrelaatiolle

Kun jotakin ilmiötä mitataan eri paikoissa, ollaan usein kiinnostuneita ilmiön spatiaalisen autokorrelaation olemassaolosta. Spatiaalinen autokorrelaatio on yleensä positiivista, eli lähellä toisiaan olevat mittaukset tuottavat samankaltaisia havaintoja. Spatiaalista autokorrelaatiota voidaan tutkia vertaamalla jotain havaintoerojen mittaa havaintopaikkojen etäisyyden mittaan. Yksi tapa testata autokorrelaation merkitsevyyttä on käyttää Mantelin satunnaistamistestiä. (Manly 2001, s. 229-230)

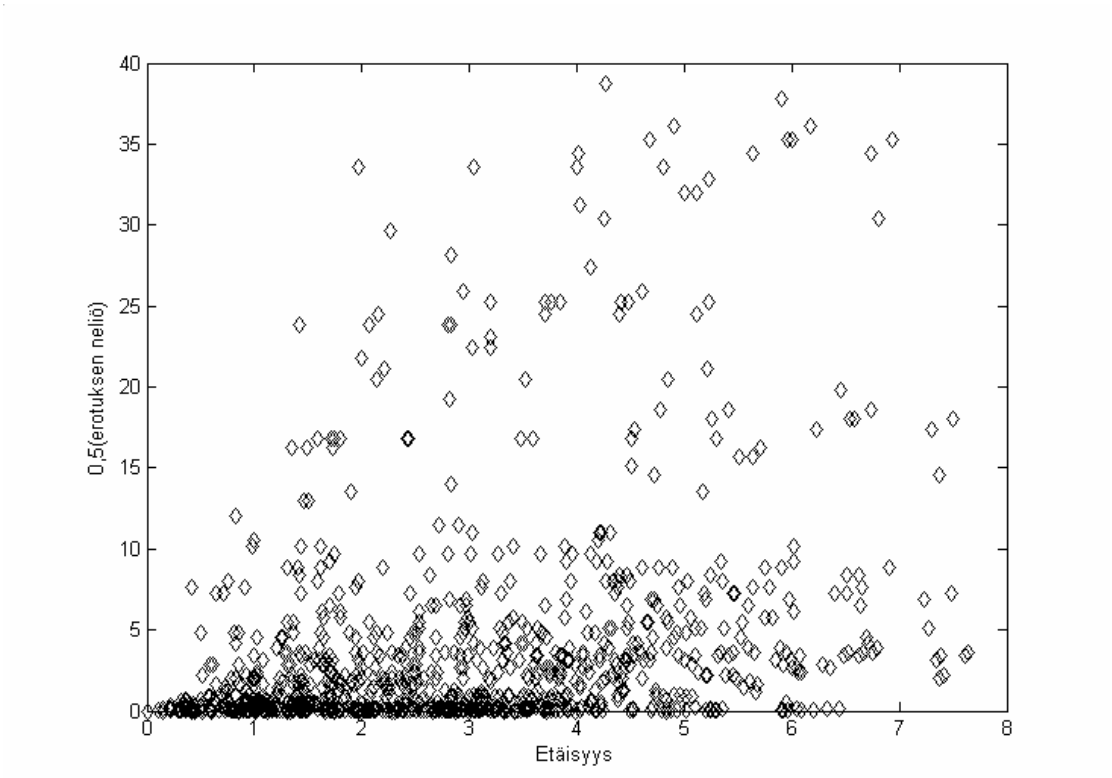
Oletetaan, että havaintopaikkoja $\mathbf{x_i}$ on n kappaletta. Olkoon kahden havaintopaikan $\mathbf{x_i}$ ja $\mathbf{x_j}$ välisen etäisyyden mitta $d_{i,j}$. havaintopaikkaparien etäisyydet saadaan koottua etäisyysmatriisiin

$$\mathbf{D} = \begin{pmatrix} 0 & d_{1,2} & d_{1,3} & \cdots & d_{1,n} \\ d_{2,1} & 0 & d_{2,3} & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n-1,1} & d_{n-1,2} & d_{n-1,3} & \cdots & d_{n-1,n} \\ d_{n,1} & d_{n,2} & d_{n,3} & \cdots & 0 \end{pmatrix}, \qquad (1)$$

joka on symmetrinen eli $d_{i,j} = d_{j,i}$. Korrelaation laskemiseen tarvitaan vielä toinen matriisi, erotusmatriisi

$$\mathbf{C} = \begin{pmatrix} 0 & c_{1,2} & c_{1,3} & \cdots & c_{1,n} \\ c_{2,1} & 0 & c_{2,3} & \cdots & c_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{n-1,1} & c_{n-1,2} & c_{n-1,3} & \cdots & c_{n-1,n} \\ c_{n,1} & c_{n,2} & c_{n,3} & \cdots & 0 \end{pmatrix}, \qquad (2)$$

joka on myös symmetrinen ja jonka alkiot $c_{i,j}$ ovat havaintojen c_i ja c_j erotuksen itseisarvoja eli $c_{i,j} = |c_i - c_j|$. Alkioina voidaan käyttää myös erotuksen neliön puolikasta ($c_{i,j} = 0{,}5(c_i - c_j)^2$), jolloin etäisyys-erotus –parien kuvaaja vastaa variogrammipilveä (katso luvut 2.2 ja 3.1). Esimerkki variogrammipilvestä on kuvassa 1.



Kuva 1. Variogrammipilvi Norjan järvien sulfaattipitoisuuksista vuonna 1981.

Matriisien **C** ja **D** ala- tai yläkolmioelementin etäisyys-erotus –pareille ($d_{i,j}$, $c_{i,j}$) saadaan laskettua Pearsonin korrelaatiokerroin. Jos tämä kerroin on epätavallisen suuri verrattuna sellaisen korrelaatiokertoimen jakaumaan, joka saadaan, jos matriisin

C havainnot järjestetään uudelleen sattumanvaraisesti, on kysymyksessä spatiaalisesti autokorreloitunut data.

Mantelin satunnaistestissä jakauma korrelaatiokertoimelle saadaan järjestämällä matriisin **C** alkiot satunnaisesti tarpeeksi monta kertaa. Testin nollahypoteesi on, että havainnoilla ei ole spatiaalista autokorrelaatiota. Jos testin p-arvo alittaa ennalta päätetyn riskitason α, nollahypoteesi hylätään, ja todetaan datan olevan spatiaalisesti autokorreloitunutta.

2.2 Variogrammi

Jos Y_i ja Y_j ovat saman satunnaismuuttujan eri paikoissa mitattuja arvoja, niiden erotuksen neliön odotusarvo on

$$\begin{aligned} &E(Y_i - Y_j)^2 = (Y_i - \mu)^2 - 2(Y_i - \mu)(Y_j - \mu) + (Y_j - \mu)^2 \\ &= Var(Y_i) - 2Cov(Y_i, Y_j) + Var(Y_j). \end{aligned} \tag{3}$$

Jos varianssi on molemmissa paikoissa σ^2, saadaan

$$E(Y_i - Y_j)^2 = 2(\sigma^2 - Cov(Y_i, Y_j)). \tag{4}$$

Korrelaatio $\rho(Y_i, Y_j) = Cov(Y_i, Y_j) / \sigma^2$, josta saadaan

$$E(Y_i - Y_j)^2 = 2\sigma^2(1 - \rho(Y_i, Y_j)). \tag{5}$$

Jos Y_i:n ja Y_j:n korrelaatio oletetaan riippuvaiseksi ainoastaan niiden välisestä etäisyydestä h, voidaan yhtälö kirjoittaa muotoon

$$\gamma(h) = \sigma^2(1 - \rho(h)). \tag{6}$$

Yhtälöä (6) sanotaan muuttujan Y variogrammiksi. Yhtälö (5) jaetaan yleensä kahdella, mistä johtuen yhtälöä (6) kutsutaan myös semivariogrammiksi. Yleensä

myös variogrammilla tarkoitetaan juuri semivariogrammia. Tässä työssä käytetään termiä variogrammi.

Variogrammiyhtälön paikkansapitävyyden edellytykset ovat muuttuja Y:n sisäinen stationaarisuus ($E(Y(s+h)-Y(s))=0$ ja $Var(Y(s+h)-Y(s))=2\gamma(h)$), ja että Y_i:n ja Y_j:n korrelaatio on riippuvainen ainoastaan niiden välisestä etäisyydestä h.

Variogrammi ilmaisee sitä suuruutta, jota muuttujan havaintoarvojen erot lähestyvät, kun havaintoparien välimatka kasvaa (Manly 2001, s. 243). Täten variogrammi kuvailee spatiaalisen autokorrelaation laatua. Koska spatiaalinen autokorrelaatio yleensä pienenee etäisyyden kasvaessa, variogrammiyhtälöstä nähdään, että variogrammi lähestyy varianssia etäisyyden kasvaessa.

Variogrammi määritetään havaintoaineistosta empiirisesti tasoittamalla aineisto sopivalla tavalla, esimerkiksi jakamalla aineisto etäisyysluokkiin, ja laskemalla variogrammiestimaatti käyttäen kaavaa

$$\hat{\gamma}(h)=\sum_{i,j}0{,}5(y_i-y_j)^2/N(h) \qquad (7)$$

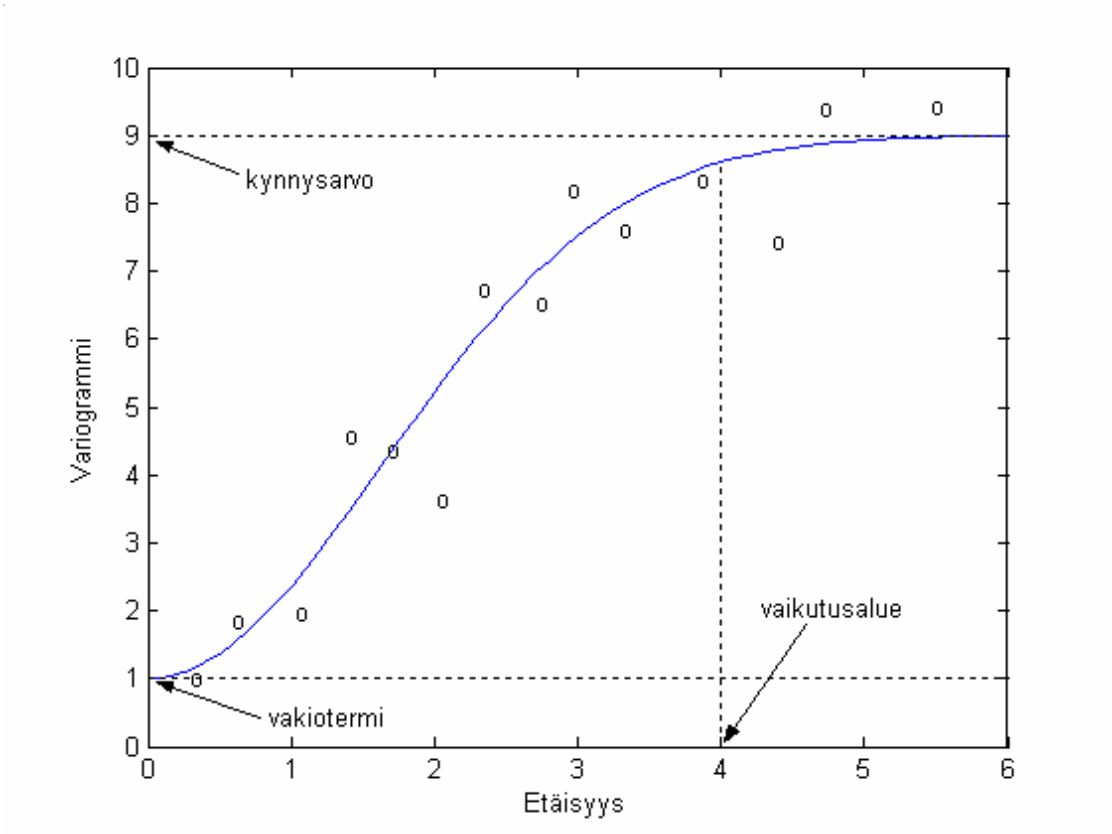
jokaiselle luokalle. Kaavassa h on luokkakeskus, $N(h)$ on luokan havaintojen lukumäärä ja summaus käy läpi kaikki pisteparit, joiden välinen etäisyys kuuluu luokkaan. Saadut estimaatit plotataan luokkakeskuksia vastaan.

Tämän jälkeen empiiriseen variogrammiin yritetään sovittaa sopiva funktio. Tavallisia variogrammimalleja ovat Gaussin malli, pallofunktio- sekä eksponenttifunktiomalli. Mallit ovat muotoa

$$\gamma(h)=c+(S-c)(1-\exp(-3h^2/a^2)) \qquad \text{(Gaussin malli)}$$

$$\gamma(h)=\begin{cases}c+(S-c)(1{,}5(h/a)-0{,}5(h/a)^3), \text{ kun } h\le a\\ c, \text{ muulloin}\end{cases} \qquad \text{(pallofunktiomalli)}$$

$$\gamma(h)=c+(S-c)(1-\exp(-3h/a)) \qquad \text{(eksponenttifunktiomalli)}$$

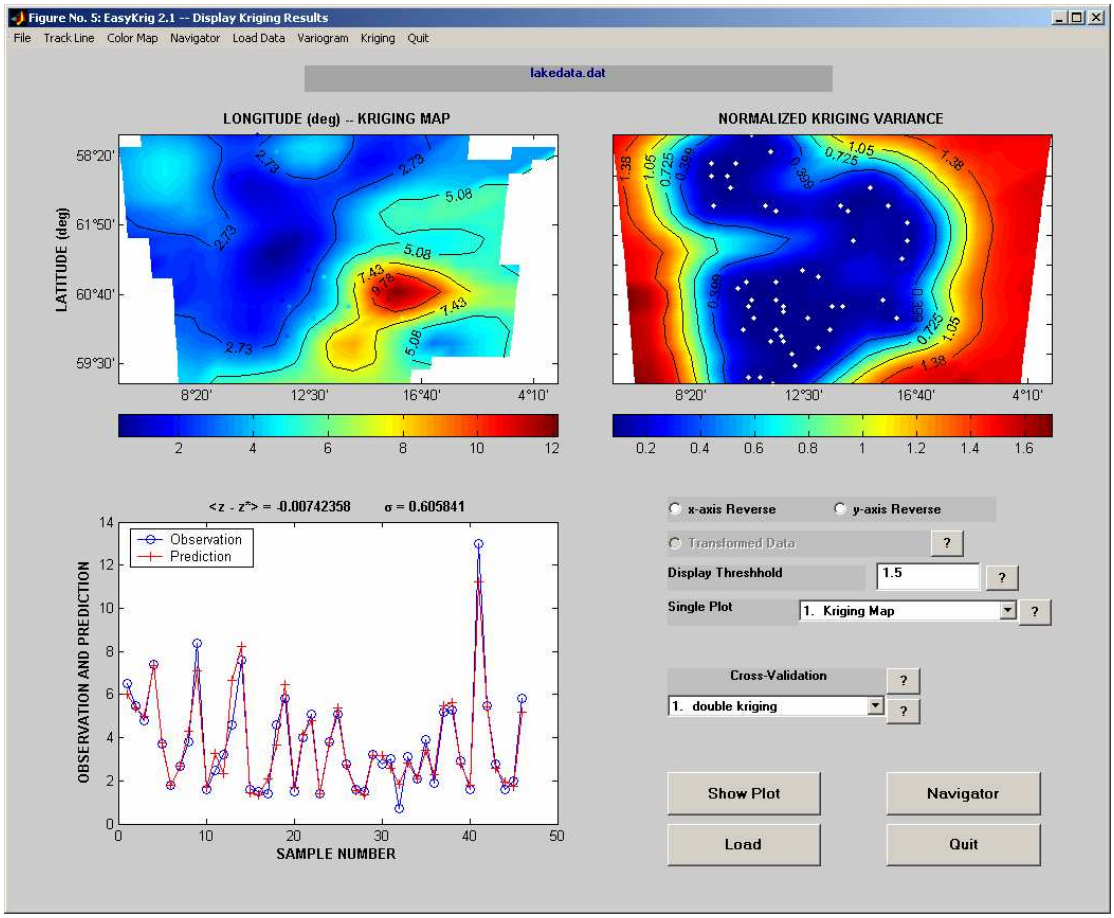


Kuva 2. Esimerkki empiirisestä (pallot) ja teoreettisesta (jatkuva viiva) variogrammista.

Kaikissa malleissa *c* tarkoittaa vakiotermiä (nugget effect), joka saattaa esiintyä, jos hyvin lähekkäin olevat havaintoarvot eroavat toisistaan, *S* on kynnysarvo (sill), joka on kuvaajan maksimiarvo, ja *a* on vaikutusalue (range of influence), joka määritellään usein kohdaksi, jossa kuvaajan arvo on 95% kynnysarvon ja vakiotermin erotuksesta. Vakiotermi voi sisältää myös varsinaista mittausvirhettä (Haining 1990, s. 29), ja osa geostatistisista ohjelmistoista olettaa vakiotermin nollaksi, koska muuten variogrammia vastaava kovarianssifunktio olisi origossa epäjatkuva (Upton ym. 1985, s. 368). Esimerkki empiirisestä variogrammista ja siihen sovitetusta Gaussin mallista on kuvassa 2.

2.3 Kriging

Variogrammi kuvailee spatiaalisen autokorrelaation laatua (Manly 2001, s.245). Variogrammin avulla laskettua mallia käytetään hyväksi muun muassa eri tyyppisissä geostatistiikan analyyseissa. Yksi yleisimmistä on kriging, joka on nimetty menetelmän uranuurtajan D. G. Krigen mukaan (Manly 2001, s. 248).



Kuva 3. Esimerkki Matlabin kriging-toolboxista.

Kriging on eräänlainen interpolointiprosessi, jonka avulla estimoidaan muuttujan arvoa mittauspisteiden välillä. Estimoinnissa lasketaan lineaarikombinaatio kaikista havainnoista, ja ongelmaksi muodostuu painokertoimien määrittäminen. Kriging-menetelmiä on eri tyyppisiä, ja Manly (2001, s. 248-249) esittelee tavallisen krigingin vaiheet:

1. Empiirisen variogrammin laskeminen.
2. Useiden teoreettisten variogrammimallien sovittaminen empiiriseen variogrammiin, sopivimman mallin valinta.
3. Varsinainen kriging-estimointi.

Tässä työssä ei tehdä erikseen kriging-estimointia johtuen työn rajauksesta sekä vaiheessa 2 esiin tulleista ongelmista. Esimerkki yhdestä Internetistä ladatusta Matlabin kriging-toolboxista on kuvassa 3.

3 Spatiaalisen autokorrelaation testaaminen

Vuonna 1972 aloitettiin norjalainen tutkimusohjelma, joka tutki happosateiden vaikutuksia Skandinaviassa. Tutkimukseen kuului muun muassa happamuuden, sekä sulfaatti-, nitraatti- ja kalsiumpitoisuuksien mittaaminen eräistä Norjan järvistä. (Manly 2001, s. 7)

Tässä työssä käytetään hyväksi tutkimuksen tuloksista järvien sulfaattipitoisuusdataa. Työssä yritetään tutkia, onko järvien sulfaattipitoisuus spatiaalisesti korreloitunutta, eli muistuttavatko toisiaan lähellä olevien järvien sulfaattipitoisuudet enemmän toisiaan, kuin toisistaan kauempana olevien.

3.1 Autokorrelaation testaaminen Mantel-testillä

Manlyn (2001, s. 8-9) kirjassa on taulukko Norjan järvien sulfaattipitoisuuksista vuosina 1976, 1977, 1978 ja 1981. Vuoden 1977 data jätettiin heti tarkastelun ulkopuolelle selvästi vähimpine havaintoineen. Tarkoitus oli ensiksi testata, minä vuonna sulfaattipitoisuuden spatiaalinen autokorrelaatio oli voimakkainta, ja tutkia tämän vuoden autokorrelaatiota variogrammin avulla.

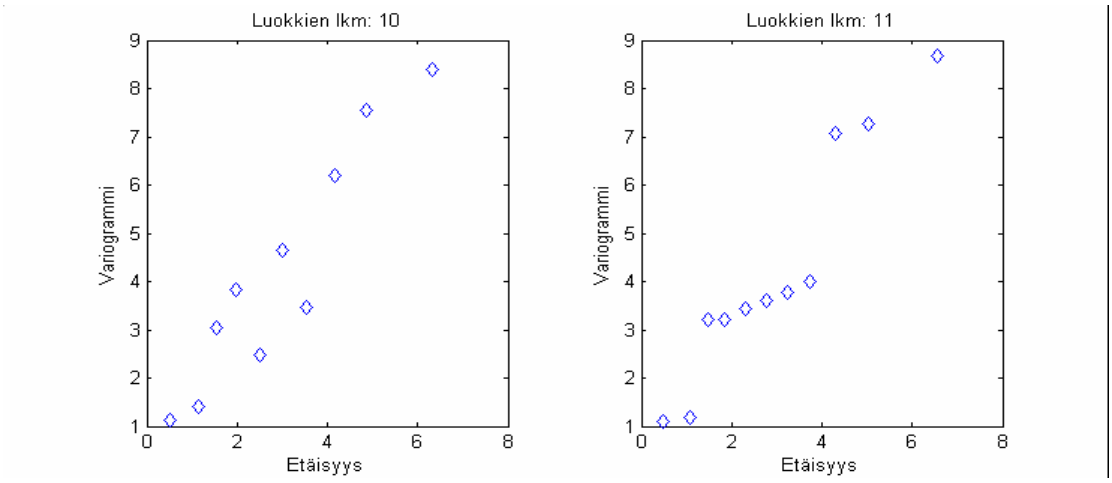
Datalle tehtiin Mantel-testi Matlabilla käyttäen hyväksi Internetistä löytyviä valmiiksi ohjelmoituja funktioita. Testi käyttää Matlabin Randperm-funktiota matriisin sekoittamiseen, ja testeissä matriisit sekoitettiin 5000 kertaa. Etäisyysmatriisina käytettiin havaintopisteiden euklidista etäisyyttä. Havaintopisteiden sijainnit oli annettu datassa pituus- ja leveysasteina, mutta tästä johtuva virhe on muutaman asteen kokoisella alueella huomaamaton. Yksi pituus- tai leveysaste vastaa kuuden desimaalin tarkkuudella 111,12 kilometriä. Toinen vaihtoehto olisi käyttää etäisyyden mittana etäisyyksien käänteislukuja (Manly 2001, s. 230), varsinkin, jos korrelaatiota ei havaita etäisyyksien perusteella. Tässä työssä korrelaatio oli kuitenkin havaittavissa jo normaaleista etäisyyksistä. Erotusmatriisina käytettiin sulfaattipitoisuuksien erotusta.

Voimakkain autokorrelaatio oli vuoden 1981 datalla korrelaatiokertoimen arvon ollessa 0,38 (vuoden 1976 datalla kertoimen arvo oli 0,30 ja vuoden 1978 datalla 0,36). Kyseisen korrelaatiokertoimen merkitsevyystestin p-arvo oli myös pienin (0,0002), eli testin tulos oli tilastollisesti merkitsevin. Mantel-testi tehtiin myös variogrammipilveä vastaavalle erotusmatriisille, eli matriisille, jonka alkiot $c_{i,j} = 0{,}5(c_i - c_j)^2$. Myös tämän testin korrelaatiokerroin (0,32) oli suurin ja testin tulos tilastollisesti merkitsevin (p-arvo 0,0002). Jatkoanalyysi päätettiin siis tehdä vuoden 1981 sulfaattipitoisuudelle. Vuoden 1981 sulfaattipitoisuuksien variogrammipilvi on kuvassa 1.

3.2 Empiirinen ja teoreettinen variogrammi

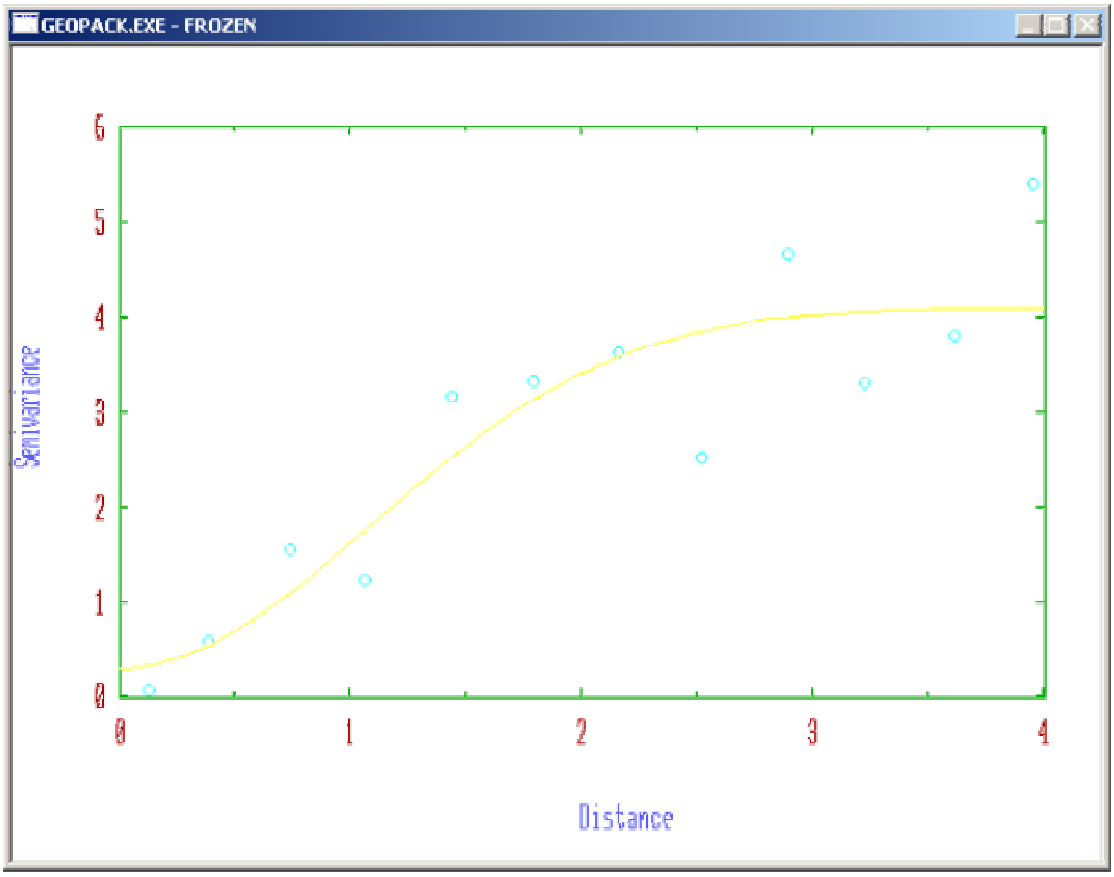
Empiiristä variogrammia varten aineisto jaettiin etäisyysluokkiin siten, että jokaiseen luokkaan tuli yhtä paljon havaintoja, jolloin luokkavälin pituus vaihteli. Toinen vaihtoehto olisi ollut käyttää tasavälistä luokitusta, mutta useimmat valmiit variogrammiohjelmat käyttävät tasavälistä luokitusta, ja tässä työssä haluttiin testata, mikä vaikutus muuttuvapituuksisella luokkavälillä on variogrammin estimoinnissa. Estimointi suoritettiin Matlabissa.

Empiirisen variogrammin estimoinnissa havaittiin hieman yllättäen selvä lineaarinen trendi. Myös luokkakoon muuttaminen vaikutti huomattavasti variogrammin muotoon, mikä näkyy kuvasta 4.



Kuva 4. Empiiriset variogrammit 10 ja 11 etäisyysluokalla.

Manly (2001, s. 247) käyttää kirjassaan esimerkkiä samalle datalle, ja hän on saanut estimoitua GEOPACK-ohjelmalla kuvan 5 kaltaisen hyvin käyttäytyvän Gaussin mallin.

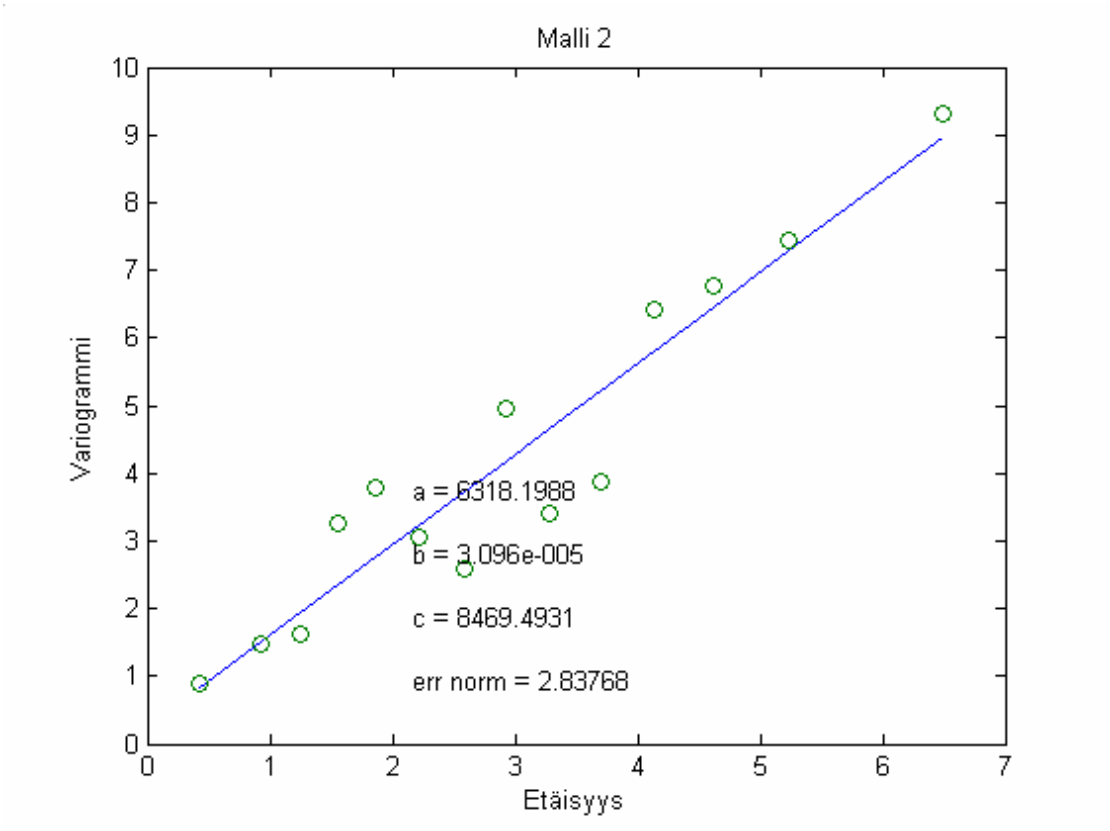


Kuva 5. GEOPACK-ohjelmalla estimoidut empiirinen variogrammi ja Gaussin malli.

Kuvan asteikkoa tarkastelemalla huomaa, että malli jättää käyttämättä havainnot lähes puolesta pisteiden maksimietäisyydestä. GEOPACK käyttää myös kiinteäpituuksisia etäisyysluokkia. Myös Matlabilla saatiin samanlaisia estimointituloksia, kun aineistosta poistettiin kauimpana toisistaan olevat havainnot.

Jos estimoinnissa pakotti vakiotermin nollaan, saatiin koko aineistoon sovitettua myös näennäisesti sopiva Gaussin malli, varsinkin, kun luokkakoko valittiin sopivasti. Vakiotermin mukaan ottaminen näytti kuitenkin, ettei Gaussin malli ole sopivin. Parhaiten aineistoon sopi eksponenttimalli (kuva 6), joka näyttää lähes suoralta.

Tämän takia ei olekaan ihme, että kynnysarvo (*c*) ja vaikutusalue (*a*) ovat erittäin suuret. Vakiotermi (*b*) taas on lähes olematon.



Kuva 6. Sulfaattipitoisuusaineistoon sovitettu eksponenttimalli 14 etäisyysluokalla.

Kriging-estimointia ei tässä työssä tehty työn rajauksen takia. Variogrammin estimoinnissa esiin tulleet ongelmat vahvistavat kuitenkin geostatistiikan asiantuntijoidenkin mielipidettä siitä, että kriging on erittäin paljon ”käsityötä” vaativa menetelmä.

4 Tulosten tarkastelu

Tässä työssä käytiin läpi spatiaalisen autokorrelaation testaamista. Spatiaalinen autokorrelaatio on havaittavissa aineistosta esimerkiksi Mantel-testin avulla. Korrelaation luonnetta voi yrittää tarkastella variogrammin avulla. Variogrammia taas voi käyttää hyväksi kriging-estimoinnissa.

Työssä tutkittu aineisto vaikutti olevan selvästi spatiaalisesti autokorreloitunut. Lähekkäin olevien järvien sulfaattipitoisuudet vaikuttaisivat siis muistuttavan toisiaan. Aineistosta estimoitu variogrammi ei kuitenkaan käyttäytynyt odotetusti. Etäisyyden kasvaessa spatiaalisen autokorrelaation tulisi vähetä ja variogrammin pitäisi lähestyä varianssia. Tässä variogrammi vaikutti lineaariselta, mikä viittaisi ei-stationaarisuuteen (Haining 1990, s. 97). Tämä rikkoo yhden variogrammiyhtälön paikkansapitävyyden perusoletuksen, sisäisen stationaarisuuden. Näin ollen aineiston spatiaalista autokorrelaatiota tulisi ehkä yrittää tulkita jollain muulla tavoin.

Variogrammin määrittäminen ei ole siis niin yksinkertaista, kuin siihen liittyvä teoria antaa ymmärtää. Jos variogrammia käytetään hyväksi esimerkiksi krigingissä, voivat tulokset vaihdella suurestikin riippuen siitä, miten empiirinen ja teoreettinen variogrammi määritellään. Empiiristä variogrammia määritellessä kannattaa kokeilla useita eri etäisyysluokkia, sekä kiinteäpituuksisille että muuttuvapituuksisille etäisyysluokille.

Geostatistiikassa variogrammi- ja kriging-menetelmiä on kuitenkin käytetty paljon, ja parhaimmillaan niistä voi olla suurtakin hyötyä. Tutkimusten tuloksiin kannattaa siis suhtautua varauksella. Kuitenkin, jos heikkokin malli toimii käytännössä, kannattaa sitä tietenkin käyttää apuna. Käytännön tutkimuksen vaativin ongelma onkin terveen järjen käyttäminen, sillä lähes kaikkiin ongelmiin on jo kehitetty valmiita malleja, mutta niiden käyttökelpoisuudesta varmistuminen ja oikea soveltaminen on oma ongelmansa.

Lähdeluettelo

Haining, R. 1990. *Spatial data analysis in the social and environmental sciences*. Cambridge, Cambridge University Press. 409 s.

Manly, B. F. J. 2001. *Statistics for Environmental Science and Management*. Boca Raton, Chapman & Hall. 326 s.

Nyqvist, T. 2002. *Atlastyyppisen aineiston luokittelu ja luokittelumenetelmien vertailu*. Pro gradu –tutkielma. Helsinki, Helsingin yliopisto, Tietojenkäsittelytieteen laitos. 55 s.

Upton, G. & Fingleton, B. 1985. *Spatial Data Analysis by Example, Volume 1: Point Pattern and Quantitative Data.* Norwich, John Wiley & Sons Ltd. 410 s.

Virkkala, R., Korhonen, K. T., Haapanen, R. & Aapala, K. 2000. *Metsien ja soiden suojelutilanne metsä- ja suokasvillisuusvyöhykkeittäin valtakunnan metsien 8. inventoinnin perusteella*. Helsinki, Suomen ympäristökeskus & Metsäntutkimuslaitos, Suomen ympäristö, luonto ja luonnonvarat 395. 52 s.

Virrantaus, K. 2001. *Johdantoa GIS Analysis –opintojaksolle*.
http://www.hut.fi/Units/Cartography/courses/fall2001/maa-123420.htm.
Otettu 17.9.2003.