Spatiaalinen autokorrelaatio maantieteellisessä mallintamisessa

Juha Oksanen

Kandidaatintyö
Oulun Yliopisto
Maantiede

Oulu, 25. huhtikuuta 2016

Sisältö

1 Johdanto

Spatiaalisen autokorrelaation ongelman ja samalla sen hyötyjen ymmärtäminen vaatii tilastotieteen perustekäsitteiden ja menetelmien tuntemusta. Tilastolllsen korrelaation eli riippuvuuden muuttujien väillä ja autokorrelaation, eli muuttujien riippuvuuden itsestään ero on tärkeä esimerkki. Pearsonin tulomomenttikorrelaatiokerroin kertoo selittävän ja selitettävän muuttujan välisen korrelaation voimakkuuden, autokorrelaatio tilastotieteessä taas viittaa aikasarjan havaintojen rippuvuuteen edeltävistä havainnoista. Spatiaalinen autokorrelaatio on siis havainnon muuttujasta x_i riippuvuus ympäröivistä $x_j, j \neq i$ havainnoista jonkin tilallisen prosessin kautta jollakin välimatkalla (Fotheringham 2009: 399.) Spatiaalinen autokorrelaatio on ongelma perinteiselle tilastotieteelle. Etenkin lineaariselle regressiomallinnuksen tapauksessa koska se perustuu identtisen ja samoin jakautumisen (i.i.d, identical and independent distribution) oletukselle (Griffith 2009: 399.) Tämä tarkoittaa sitä että jokainen havainto x_i on satunnainen realisaatio samasta taustajakaumasta ja riippumaton muista havainnoista. Spatiaalisissa aineistoissa tämä ei yleensä päde, sillä niissä etenkin riippumattomuuus muista lähellä olevista havainnoista on kyseenalaista.

Spatiaalisen autokorrelaation tuominen maantieteilijöiden käyttöön on ollut pitkä ja kivinen tie. Griffith (1992: 266) antaa kunnian spatiaalisen autokorrelaation tunnistamisesta William Seeley Gossetille, joka käytti kirjailijanimeä Student. Samalla hän tuo esille sen, että spatiaalinen autokorrelaatio ei ole helposti selitettävissä maantieteilijälle, joka ei ole perehtynyt tilastotieteeseen. Goodchild, Griffith ja Odland yrittivät 1980-luvun lopulla tätä aukkoa, mutta saivat osakseen kritiikkiä. (Griffith 1992: 266) Cliff ja Ord (1973: 1981) tekivät töitä sen eteen, että maantieteilijöillä olisi keinoja tunnistaa spatiaalinen autokorrelaatio, ja laskea sen tilastollinen merkitsevyys. Heidän julkaisunsa ova hyvin matemaattisia ja jäävät helposti ymmärtämättä jos lukija ei omaa vahvaa tilastollista pohjaa. Spatiaalinen autokorrelaatio tarkoittaa eri taustalta tuleville tutkijoille eri asioita, ja niiden yhteen nivominen ei ole helppo tehtävä.

Tilastollinen mallintaminen on monelle tieteenalalle yhteistä, ja monet niistä käyttävät paikkaan sidottuja aineistoja. Esimerkiksi ekologia ja taloustiede,ja taloustieteen alalla erityisesti ekonometria, ovat hyötyneet menetelmistä joita on kehitetty spatiaalisen autokorrelaation huomioonottamiseksi.

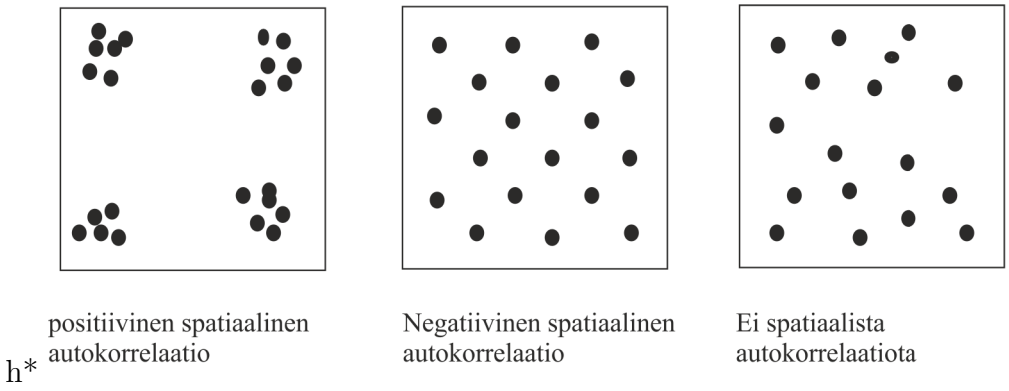
Tämän tutkielman tarkoitus on avata spatiaalisen autokorrelaation käsitettä sekä sen tutkimusta ja hyväksikäyttöä monitieteellisestä ja tilastotieteellisestä näkökulmasta. Käsitteen määrittelyn ja historian lisäksi pyrin avaamaan sen tilastotieteellistä merkitystä ja sen luomia mahdollisuuksia spatiaalisen aineiston esittämisessä ja erilaisten mallien luomisessa. Ensin luon katsauksen spatiaalisen autokorrelaation määritelmään ja historiaan ja toiseksi avaan erilaisia tapoja kvantifioida spatiaalista autokorrelaatiota. Seuraavaksi perehdyn spatiaalisen autokorrelaation ja regressiomallien ongelmaan ja Kriging-interpolointiin. Lisäksi käyn läpi erästä tutkimusta jossa spatiaalista aukorrelaatiota on käytetty parantamaan arvioita metsäpalojen vaikutuksista käyttäen kaukokartoitusaineistoa.

2 Spatiaalinen autokorrelaatio

Maantieteessä kaikki on riippuvaista kaikesta, mutta lähellä olevat asiat ovat riippuvaisempia toisistaan kuin kauempana olevat, on Waldo Toblerin kuuluisa Maantieteen ensimmäinen laki, josta myöhemmin käytettiin nimeä Tobler's First Law (TFL) (Miller 2004: 284). Tästä ilmiöstä on käytetty erilaisia termejä, kuten esimerkiksi spatiaalinen riippuvuus, spatiaalinen assosiaatio ja viimein spatiaalinen autokorrelaatio. Spatiaalinen autokorrelaatio on sanana tullut käyttöön vasta kvantitatiivisen vallankumouksen myötä mutta sen olemassaolo on tunnistettu jo 1800- luvun lopulla. Cliff ja Ord (1983: 8) viittaavat Charles Darwinin veljenpoikaan Francis Galtoniin, eräänä ensimmäisistä tutkijoista, joka puhui maantietellisten painokertoimien käytöstä spatiaalisen riippuvuden huomiomiseksi Sana spatiaalinen autokorrelaatio vakiintui kuitenkin tiedeyhteisön käyttöön 1970-luvulla etenkin sellaisten tutkijoiden kuin Andrew Cliff ja Keith Ord julkaisujen myötä. Heidän vuonna 1969 julkaisemansa artikkeli "The Problem of Spatial Autocorrelation"oli läpimurto maailmaan, jossa maantietellisen riippuvuuden tutkimus oli tilastotieteellisesti vielä hyvin alkeellista, koska tilastotieteeseen perehtyneitä maantieteilijöitä ja spatiaalisen datan käsittelystä kiinnostuneita tilastotieteilijöitä ei ollut kovin montaa (Miller 2004: 284-286).

Spatiaalinen autokorrelaatio määrittyy sekä sijainnin samankaltaisuudesta että havaintojen samankaltaisuudesta. Se on erittäin riippuvaista siitä millä tiheydellä otos on tehty tai kaukokartoituksessa sen resoluutiosta eli erottelutarkkuudesta. Suuren tiheyden käyttäminen saattaa aiheuttaa sen, että spatiaalista autokorrelaatiota havaitaan aineistossa, jossa sitä ei ole, koska lähekkäiset havainnot eivät tuo uutta informaatiosisältöä. Toisaalta taas liian harva otos piilottaa hienojakoisemman spatiaalisen rakenteen. Sama pätee rasteriaineistoihin, joissa pikselikoko voi olla niin suuri että se peittää alleen spatiaalisen rakenteen, tai niin pieni että se luo illuusion sellaisen olemassaolosta (Longley ym. 2005: 89-90).

Athur Getis (2008: 298) määrittelee spatiaalisen autokorrelaation ja perinteisten tilastollisten korrelaatiokertoimien välisen eron hyvin tyhjentävästi:



Kuva 1: Spatiaalisen autokorrelaation ilmenemismuodot.

Spatiaalinen autokorrelaatio on muuttujan sisäisen vaihtelun korreloitumista sen sijainnin kanssa, kun taas perinteiset tilastolliset korrelaatiokertoimet mittaavat muuttujien välisen vaihtelun korrelaatiota. Spatiaalinen autokorrelaatio voi ilmentyä kahdella tasolla, globaalilla ja paikallisella, eivätkä ne ole toisiaan poissulkevia. Globaalit spatiaalisen autokorrelaation mittarit, kuten Moranin I ja Gearyn C -statistiikat, ovat paikallisten autokorrelaation mittareiden summia. (Anselin 1995: 95) Paikallisen spatiaalisen autokorrelaation esittämiseen on kehitetty erikseen työkaluja, kuten Getis-Ordin G(i) ja G*(i). Spatiaalisen autokorrelaation havaittu voimakkuus on myös datapisteiden tiheydestä riippuvainen, ja sen takia samalta alueelta tehdyt lisähavainnot eivät välttämättä paranna tulosten tarkkuutta(Viladomat & Mazumder, McInturff , McCauley, Hastie 2014: 410).

Kuvassa 1. esitetään miten positiivinen spatiaalinen autokorrelaatio ilmenee ryvästymisenä, negatiivinen tasaisena hajontana ja spatiaalista autokorrelaation puute satunnaisena jakautumisena. Satunnainen jakauma on siis nollahypoteesi, jota vasten spatiaalisen autokorrelaation olemassaoloa lähdetään tutkimaan. Kuvasarjan ensimmäinen positiivista autokorrelaatiota kuvaava esimerkki ei välttämättä näy positiivisena autokorrelaationa globaaleille testeille, koska se on spatiaalisesti heterogeeninen. Globaalit spatiaalisen autokorrelaation mittarit olettavat taustalla olevan prosessin olevan stationaarinen eli toimivan samalla tavalla koko aineistossa ja kaikkiin suuntiin

yhtä voimakkaana. Negatiivinen spatiaalinen autokorrelaatio näkyy kuvassa hilarakenteena, jossa havainnot ovat yhtä kaukana naapurihavainnoista.

Aikasarja-analyysistä poiketen spatiaalisten aineistojen riippuvuusuhteet ovat monisuuntaisia, joten niiden matemaattinen esittäminen on huomattavasti monimutkaisempaa (Longley & Goodchild, Maguire, Rhind 2005: 87). Se tarkoittaa samalla sitä, että niiden laskenta etenkin suurilla aineistoilla vaatii paljon tietokonekapasiteettia, jota ei ennen 1990-luvun puoliväliä ollut juurikaan käytössä. Siispä 1970 -luvun aikana ja vielä 1980-luvulla laskenta kesti tunteja, suurilla aineistoilla jopa vuorokauden (Bivand 2009: 285). Spatiaalinen autokorrelaatio auttaa representaatioiden luomisessa, mutta merkittävästi vaikeuttaa analyysia ja tilastollista ennustamista. Kun esittämisen ja laskennan vaikeuksiin lisätään se, että spatiaalinen autokorrelaatio on hyvin skaalariippuvaista on oikean skaalan löytäminen ja oikeiden asioiden esittäminen on hyvin haastavaa (Longley ym. 2005: 87).

2.1 Miten spatiaalista autokorrelaatiota mitataan

Yksinkertaisin esimerkki spatiaalisesta aineistosta on ruudukko, josta ilmenee binäärisen tai luokitellun vasteen tila, esimerkiksi esiintyykö jossakin ruudussa laji X, jos esiintyy, niin ruutu saa arvon 1, jos ei, ruutu saa arvon 0. Jos ruuduilla on yhteistä rajaa, lasketaan ne naapureiksi. Naapuruus voi syntyä kahden 1 ruudun, 1 ja 0 ruudun tai kahden 0 ruudun välille. (Cliff & Ord 1981: 11). Positiivisen spatiaalisen autokorrelaation tilanteessa 1,1 naapuruuksia on enemmän kuin tilanteessa, jossa spatiaalista autokorrelaatiota ei ole havaittu, ja negatiivisen spatiaalisen autokorrelaation tapauksessa 1,0 naapuruuksia on enemmän kuin jos ruudukko olisi satunnaisen prosessin tulos (Cliff & Ord 1981: 11 - 13). Naapurusluku (join count statistic) on tällaiselle aineistolle luonnollinen tapa selvittää globaalin spatiaalisen autokorrelaation olemassa olo. Lasketaan jokaisen ruudun naapuruudet, ja verrataan niiden suhteita tilanteeseen jossa ruudukko on täysin satunnainen. Satunnainen ruudukko on asymptoottisesti normaalisti jakautunut, kun ruutujen määrä on riittävän suuri. (Cliff & Ord 1973: 5). Tällaisessa aineistossa normaalijakauma on

nähtävissä naapuruusparien jakautumisessa.

Naapurusluvusta edelleen kehittyneempiä ja monikäyttöisempiä globaalin autokorrelaation mittareita ovat Moranin I ja Gearyn C, jotka perustuvat Pearsonin tulomomenttikorrelaatiokertoimeen ja painokerroinmatriisin luontiin jolla jokaisen ruudun tai pisteen vaikutus kaikkiin muihin pisteisiin ilmaistaan (Wong & Lee 2005: 367). Moranin I lasketaan kaavalla

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{W \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Jossa W on painokerroinmatriisin alkioiden summa, w_{ij} on painokerroin kyseisten yksiköiden välillä ja x_i on yksikön i arvo. Moranin I:n arvot vaihtelevat välillä [-1,1], ja nollahypoteesin mukainen odotusarvo on $E_(I) = \frac{-1}{n-1}$ (Wong & Lee 2005: 367).

Gearyn C lasketaan puolestaan kaavalla:

$$C = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{2W \sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Gearyn C:n vaihteluväli on [0,2], jossa 0 on täydellinen spatiaalinen autokorrelaatio, 2 on täydellinen negatiivinen autokorrelaatio ja 1 on odotusarvo kaikilla n. (Wong & Lee 2005: 373-374). Edellä kuvatut menetelmät mittaavat vain koko alueen eli globaalia spatiaalista autokorrelaatiota. Nämä menetelmät eivät kerro mitään ryvästymisestä tai spatiaalisesta heterogeenisyydestä, vaan olettavat spatiaalisten prosessien olevan stationäärisiä (Fotheringham 2009: 401) Globaalit mittarit avasivat kuitenkin tien spatiaalisten riippuvussuhteiden mittaamiselle ja saivat aikaan kehityskaaren, joka edelleen tuo uusia näkökulmia tähän aiheeseen (Fotheringham 2009: 402) Globaalien mittareiden mukanaan tuomat oletukset spatiaalisesta stationäärisyydestä ovat useimmiten epärealistisia, varsinkin jos otoskoko kasvaa suureksi (Anselin 1995: 94-95) Spatiaalisesta autokorrelaatiosta on tullut tärkeä työkalu mallinnukselle, ja sen käyttökohteet vain laajenevat sitä mukaa kun tietokoneiden laskentateho kasvaa. Sen myötä suurempia datamääriä voidaan käsitellä lyhyemmässä ajassa, ja kehittyneet työkalut tulevat käyttöön yhä kasvavalle

joukolle maantieteilijöitä. Mielestäni on tärkeää perehtyä käytettyjen menetelmien mahdollisiin heikkouksiin, siten analyysistä saadaan tarkempaa.

Painokerroinmatriisin rakentuminen vaikuttaa vahvasti siihen, millaisia tuloksia nämä mittarit antavat. Matriisi W voidaan rakentaa hyvin monella tavalla, joista seuraavaksi muutama esimerkki. Yksinkertaisin tapa matriisin luontiin on binäärinen matriisi, jossa naapuriksi lasketaan kaikki ne joilla on yhteistä rajaa havaintoalueen x_i kanssa, kaikki muut yksiköt saavat arvon 0. Muita vaihtoehtoja ovat esimerkiksi keskipiste-etäisyyden mukaan tehtävä rajaus jonkin hyvin perustellun vaikutusetäisyyden mukaan, etäisyyden mukaan pienenevä painotus eli (inverse distance weights, IDW), tehtyjen havaintojen pohjalta luotu vaikutusaluematriisi ja monia muita tapoja. Lisäksi näitä tapoja voidaan yhdistellä ja standardisoida, Row-standardised IDW on yksi hyvin suosittu tapa luoda painokerroinmatriisi, jossa rivin alkiot jaetaan niiden summalla.

Tutkijat huomasivat pian että niin hyödyllisiä kuin globaalin spatiaalisen autokorrelaation mittarit olivatkin, ne peittivät usein spatiaalista vaihtelua ja antoivat virheellisen kuvan tutkittavan ilmiön käyttäytymisestä. Getis ja Ord tekivät uraauurtavaa työtä paikallisen järjestymisen ja etäisyyden mukaan heikkenevän järjestäytymisen saralla (Anselin 1995: 94). Tarvittiin uusia menetelmiä ja mittareita näiden ilmiöiden voimakkuuden mittaamiseksi. Tähän kutsuun vastasivat Getis ja Ord kehittämällä G*(i) ja G(i) indeksit, lisäksi Anselin popularisoi Moran hajontakuvion ja termin LISA:n (Local Indicator of Spatial Autocorrelation) (Anselin 1995: 94-95). LISA on mikä tahansa mittari joka täyttää kaksi ehtoa:

1. Jokaisen havainnon LISA:n tulee antaa kuva siitä, miten voimakasta ryvästyminen on havainnon ympärillä

2. LISA:n summa yli koko aineiston on suhteessa globaaliin spatiaalisen autokorrelaation mittariin,

Parantaakseen laskennan luotettavuutta Anselin (1995: 95) ehdottaa että raakojen arvojen sijasta käytettään jotain stardardisoivaa menetelmää,

esimerkiksi kunkin havainnon erotusta keskiarvosta.

G(i) indeksin laskeminen:

$$G_i(d) = \frac{\sum_{i=1}^{n} w_{ij}(d)x_j - W_i(d)\bar{x}(i)}{s_i\sqrt{[((n-1)S_{1i}) - W_i^2]/(n-2)}}$$

Jossa $w_ij(d)$ on matriisin W alkio joka on etäisyyden d funktio tai mahdollisesti 1/0-matriisi jossa 1 on linkki etäisyyden d sisällä x_i:stä pois lukien x_i (Ord & Getis 1995: 289). Spatiaalisen painokerroinmatriisin W rakentamisesta on tehty lukuisia tutkimuksia, ja sen luomiseen on useita eri mahdollisuuksia, useimmin käytetään euklidista etäisyyttä, eli lineaarista etäisyyttä pisteiden väillä, mutta pienimmän hinnan (hinta on tässä tapauksessa mikä tahansa mittari jolla etäisyyden voittamiseen käytettävää energiaa, työtä tai aikaa voidaan esittää luotettavasti) etäisyys on myös paljon käytetty metodi. Edelleen $s(i)$ on otoskeskipoikkeama ja $S_{1i} = \sum_j w_{ij}^2, (j \neq i)$ (Ord & Getis 1995: 289).

$G_i(d)$:n tulkinta on huomattavasti Moranin I:n tulkintaa haastavampaa mutta siitä saadaan huomattavasti enemmän informaatiota. $G_i(d)$:n tulkintaa varten se on hyödyllistä esittää tutkittavan muuttujan funktiona hajontakuviossa, jolloin nähdään miten $G_i(d)$:n arvot käyttäytyvät aineiston jakauman mukaan ja niille voidaan laskea korrelaatio. Tämän kuvion avulla voidaan nähdä mitkä $G_i(d)$:n arvot ovat suhteellisesti pieniä ja mitkä suuria. Suuret arvot kertovat kuumista pisteistä jossa suuret x_i arvot ovat ryvästyneet, ja pienet G_i arvot siitä mihin pienet x_i arvot ryvästyvät. Moranin I indeksissä pienet arvot tarkoittavat negatiivista spatiaalista autokorrelaatiota. (Lanorte & Danese, Lasaponara, Murgante 2013: 45).

$G_i(d)$ antaa meille siis huomattavasti enemmän informaatiota X:n käyttäytymisestä kuin globaalin autokorrelaation indeksit C ja I. Se kertoo positiivisen spatiaalisen autokorrelaation laajuuden ja voimakkuuden, ja sen onko kyseessä suurten vai pienten arvojen ryvästymisestä. $G_i(d)$ on erittäin hyödyllinen esimerkiksi kaukokartoitusaineiston analyysissä josta enemmän esimerkkitapauksessa. $G_i(d)$:n lisäksi on monia muita paikallisen spatiaalisen autokorrelaation mittareita, mutta niiden toimintaperiaate ja tilastollinen

perusta on hyvin samantapainen.

3 Spatiaalinen autokorrelaatio ja regressiomallinnus

Regressiomallinnus perustuu jonkin muuttujan Y keskimääräisen käyttäytymisen populaatiossa tai havaintojoukossa arvioinnista muuttujista $X_1...X_n$. Perinteiseen lineaariseen regressiomallinnukseen liittyy monia rajoitteita ja oletuksia. Esimerkiksi virhetermien tulee olla riippumattomia toisistaan sekä havaintojen muuttujista x_i, ja vasteiden y_i oletetaan olevan riippumattomia ja samoin jakautuneita (independent, identically distributed, i.i.d.). Regressioyhtälön yleinen muoto $Y = \beta * X + \varepsilon$, jossa Y on mallinnettava muuttuja, X on kovariaatti jonka, tai joiden perusteella Y:n arvoja ennustetaan, β on kerroin ja ε on virhetermi. Lineaarinen regressio perustuu pienimmän neliösumman menetelmään, jolla estimoidaan β:n arvot minimoimalla virhetermin neliösumma. Regressiomallinnuksen ongelmana maantieteessä on se, että data on kerätty paikoista ja paikat eivät ole riippumattomia ympärillä olevista paikoista.

On kolme erilaista tapaa miten spatiaalinen autokorrelaatio ilmenee residuaaleissa:

- Etäisyyteen perustuva riippuvuus muuttujien välillä, kuten esimerkiksi eliön dispersaalikyky jätetään huomioimatta.
- Epälineaariset yhteydet ympäristön ja muuttujien välillä mallinnetaan lineaarisesti.
- Mallissa ei oteta huomioon jotakin ympäristömuuttujaa, mikä johtaa residuaalien spatiaaliseen järjestyneisyyteen.

(Dormann & McPherson, Araujo, Bivand, Bolliger, Carl, Davies, Hirzel, Jetz, Kissling, Kühn, Ohlemüller, Peres-Neto, Reineking, Schröder, Schurr ja Wilson 2007: 610).

Spatiaalinen epästationaarisuus ja spatiaalinen riippuvuus voidaan ottaa huomioon estimoimalla β_i jokaiselle x_i erikseen jolloin kaava näyttää seuraavalta

$$Y = \alpha + \beta_i * x_i + \epsilon$$

Jolloin jokainen piste tai ruutu saa oman regressiokertoimensa. Globaalin regressiomallin virhetermien spatiaalista riippuvuussuhdetta voidaan näin tarkastella helposti. Jos $\hat{\beta}$ estimaattia verrataan $\hat{\beta}_i$ estimaatteihin havaitaan, että virhetermit ovat positiivisia niissä yksiköissä y_i joissa $\hat{\beta}$ on suurempi kuin $\hat{\beta}_i$ ja negatiivisia kun $\hat{\beta} < \hat{\beta}_i$, ja jos $\hat{\beta}_i$ on spatiaalisesti riippuvainen, ovat residuaalit myös spatiaalisesti riippuvaisia. (Fotheringham 2009: 402).

Spatiaalisten aineistojen analysointiin regressiomenetelmillä on kehitetty useita eri lähestymistapoja. Monet niistä keskittyvät spatiaalisen autokorrelaation vaikutuksen poistoon mallin toiminnassa, eivät niinkään sen hyväksikäyttämiseen osana mallin parametreja (Dormann ym. 2007: 610). Autokovariaattimalli on muunnelma perinteisestä OLS-regressiomallista, jossa jokaiselle havainnolle lasketaan niiden painotettu riippuvuus ympäröivistä arvoista, ja sille oma regressiokerroin. Autokovariaattimallit toimivat paremmin, jos naapurusto voidaan määrittää esimerkiksi jonkin ekologisen taustatiedon perusteella, esimerkiksi lajin disperisaalin eli leviämiskyvyn kautta (Dormann ym. 2007: 611). CAR (Conditional Autoregression) ja SAR (Simultaneous Autoregression) puolestaan hyödyntävät naapuruusmatriisia W, joka on spatiaalinen painokerroinmatriisi jokaisen havainnon x_i suhteesta ympäröiviin havaintoihin x_j. CAR mallit pyrkivät ottamaan spatiaalisen autokorrelaation huomioon vasteen arvoissa lisäämällä termin $\rho W(Y - X\beta)$ regressiomalliin (Dormann ym. 2007: 613-614). Tässä lisätermissä ρ on regressiokerroin, W on spatiaalinen painokerroinmatriisi ja $Y - X\beta$ on mallin residuaalit.

Regressiomallinnuksen taustalla oleva lineaarialgebra ja sen lainalaisuudet saavat aikaan sen että sovitteiden βX ja vasteen Y erotus eli residuaalivektorit ovat ortogonaalisia, eli kohtisuorassa malliavaruuteen nähden. Siksi niiden perusteella voidaan tehdä päätelmiä mallin toiminnasta ja hyvyydestä. Kun tämä viedään maantieteelliseen kontekstiin, ja erityisesti spatiaalisen autokor-

relaation maailmaan, residuaalit käyttäytyvät eri tavoin sen mukaan miten voimakasta spatiaalinen autokorrelaatio on. Jos residuaalit ovat spatiaalisesti ryhmittyneitä on todennäköistä että spatiaalinen autokorrelaatio joissakin mallin muuttujista on vastuussa tästä, ja mallia tulee muokata vastaavasti. (Dormann ym. 2007: 610).

4 Kriging ja interpolaatio

Kriging on nimetty Eteläafrikkalaisen kaivosinsinöörin, Daniel Krigen, mukaan. Se perustuu osittain hänen pyrkimykselleen yleistää koekairausten tuloksia saadakseen selville kuinka suuri ja rikas jokin mineraaliesiintymä on. Optimaalisten lineaaristen prediktorien yhdistäminen spatiaaliseen aineistoon on kuitenkin laajemman tutkijajoukon työtä, esimerkiksi Matheron, Gandin ja Kolmogorov olivat metodin kehittämisessä etualalla (Cressie 1993: 106). Kriging-menetelmiä on useita, mutta niiden taustalla on pyrkimys sovittaa optimaalinen prediktori datapisteiden välisen vaihtelun selittämiseksi ja käyttää tätä mallia mittauspisteiden välisen alueen tai tilan arvojen estimointiin mahdollisimman tarkasti (Cressie 1993: 106.)

Kriging-menetelmillä on paljon yhteistä regressiomallien kanssa, mutta menetelmien käyttökohteet ovat hyvin erilaisia. Kriging tähtää estimointiin, ei ennustamiseen, koska pyrkimyksenä on selvittää jonkin pysyvän, mutta tuntemattoman parametrin arvo.(Chiles & Delfiner 1999: 151.) Niiden kytkökset spatiaaliseen autokorrelaatioon eivät ole niin selkeitä kuin regressiomalleissa, mutta aivan yhtä tärkeitä. (Semi)variogrammi kuvaa vaihtelua datapisteparien välillä yhdellä akselilla ja datapisteparin välistä etäisyyttä toisella (Longley ym. 2005: 336.). Semi viittaa tässä yhteydessa varianssin puolittamiseen, jolloin saadusta kuvasta tulee helpommin luettava. Siitä voidaan havaita miten kauas spatiaalinen autokorrelaatio vaikuttaa (etäisyys), ja miten paljon vaihtelua aineistossa on suurimmillaan (kynnys). Etäisyyttä ja kynnysarvoja käytetään naapuruusetäisyyden määrittämiseen, jolloin mukaan saadaan kaikki ne havainnot, joilla on merkitystä.(Chiles & Delfiner 1999:

157). Hippuvaikutus (nugget effect) taas on se vaihtelun osa joka ei häviä vaikka etäisyys mittauspisteiden välillä supistuu lähes nollaan, se on siis etäisyydestä rippumatonta vaihtelua, joka on olennainen osa tutkittavaa ilmiötä ja sen jakaumaa (Longley ym. 2005: 336).

Semivariogrammeja voidaan myös luoda anisotrooppisesti, eli ottamalla huomioon suunnan merkitys siinä miten käsiteltävä ilmiö käyttäytyy. Tällöin pisteistä lasketaan esimerkiksi 90^o lohkoissa varianssit muihin pisteisiin ja tarkastellaan suunnan vaikutusta varianssin käyttäytymiseen (Longley ym. 2005: 336). Näemme tästä, että Kriging perustuu suoraan ajatuksella kahden pisteen välisen etäisyyden vaikutuksesta niiden arvojen samankaltaisuuteen, mutta esittää saman asian toisin päin, eli etäisyyden kasvaessa erot pisteiden arvoissa kasvavat. Kriging vaatii käyttäjältä huomattavasti enemmän perehtymistä käsisteltävän ilmiön käyttäytymiseen kuin esimerkiksi IWD-interpolointi, mutta antaa huomattavasti tarkempia tuloksia. Krigin-menetelmillä tehdyt interpolaatiot pyrkivät suurimpaan mahdolliseen tarkkuuteen, joten ne eivät ole sileitä, koska tavoitteena ei ole tasainen käyrä kuten splini-menetelmissä vaan raaka lukujoukko, joka kuvaa ilmiön käyttäytymistä tilassa(Chiles & Delfiner 1999: 160).

5 Esimerkkitapaus: G(i) statistiikka ja metsäpalojen laajuuden sekä voimakkuuden arviointi kaukokartoitusaineistosta

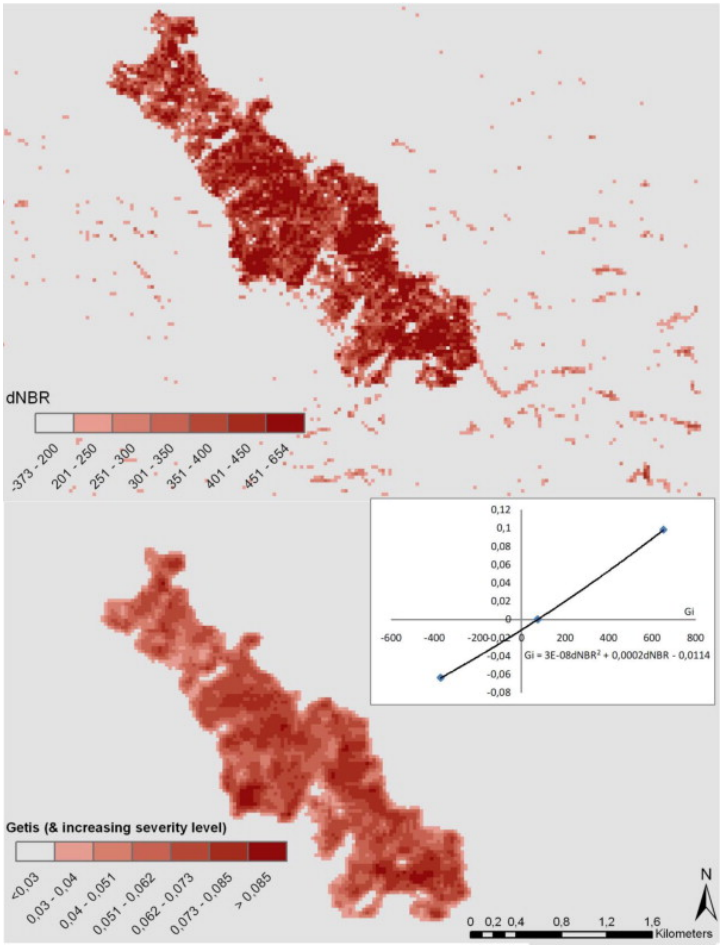
Esimerkkiartikkeli käsittelee metsäpalalojen voimakkuuden ja laajuuden arviointia kaukokartoitusaineistosta automaattisesti, ilman kallista kenttätutkimusta (Lanorte ym. 2013). Tutkimuksen aineistona oli kahden eri kaukokartoitusmoduulin kuvaama aineisto. Molemmat anturit ovat Terra satelliitissa. Terra kuvaa saman alueen aina 16 vuorokauden välein, joten se soveltuu erinomaisesti kasvillisuuden muutoksen seurantaan. MODIS (Moderate Resolution Imaging Spectroradiometer) toimii 36 eri aallonpituusalueella joista tässä tutkimuksessa käytettiin 7 ensimmäistä infrapuna aluetta ja niiden erottelutarkkuus on 250 metriä ja 500 metriä. ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) taas kuvaa 15 metrin resoluutiolla infrapunasäteilyn eri aallonpituuksia VNIR(Visible/Near Infrared), SWIR(Short Wave Infrared) ja TIR(Thermal Infrared). Näistä aallonpituuskaistoista voidaan laskea erilaisia kasvillisuusindeksejä ja muutosindeksejä, kuten NBR (Normalized Burn Difference) ja NDVI (Normalized Vegetation Difference Index).

NBR voidaan laskea helposti kaavalla $NBR = \frac{NIR-SWIR}{NIR+SWIR}$, ASTER ainestosta ja $NBR = \frac{MODIS_2-MODIS_7}{MODIS_2+MODIS_7}$ MODIS ainestosta, sillä näiden radiometrinen kuva on suoraan infrapunasäteilyn heijastuvuusmatriisi kyseisillä aallonpituusalueilla. ΔNBR saadaan vähentämällä ennen paloa lasketusta NBR-indeksistä palon jälkeinen NBR. Tällä tavalla paloalue saadaan helposti määritettyä, mutta kuten kaikessa satelliittiaineistossa, tähän voi jäädä virheitä esimerkiksi ihmistoiminnan, pilvien ja tiedonsiirtohävikin takia. ΔNBR ja sen suhteellistettua RdNBR (Relative delta Normalized Burn Difference) arvoa voidaan käyttää, jos halutaan arvioida palon intensiteettiä, sen vapauttaman hiilidioksidin määrää ja muita vaikutuksia. (Miller & Thode 2007: 70).

Lanorten ja kumppaneiden tutkimus keskittyy kuitenkin määrittelemään

paloalueen laajuutta ja palon intensiteettiä. Jos paloaluetta pyrittäisiin määrittämään pelkästään NBR indeksin avulla suoraan kaukokartoitusaineistosta riski yliarviointiin on suuri, ja paloalue näyttää helposti hyvin pirstaleiselta, mikä ei todennäköisesti vastaa todellisuutta. $G_i(d)$ indeksin käyttäminen apuvälineenä parantaa arvioinnin luotettavuutta, koska se poistaa hajahavaintoja ja virhepikseleitä, jotka johtuvat esimerkiksi ihmisen toiminnasta. Kun $G_i(d)$:n laskennassa käytetään intensiteettinä dNBR-arvoa ja tarkastellaan korrelaatiota $G_i(d)$:n ja dNBR:n välillä, voidaan $G_i(d)$ luokitella ja esittää kartalla. Tämän kartan avulla voidaan tehdä arviointia siitä miten laaja palo on oikeasti ollut, koska se poistaa hajanaisuutta ja selventää alueen rajoja. Tutkimuksessa käytettiin Moranin I indeksin laskua erilaisilla välimatkoilla, jotta saatiin selville mikä välimatka olisi kaikkein hyödyllisin $G_i(d)$:n laskemiseen. Tämän jälkeen saadulla naapurustolla laskettiin paikalliset autokorrelaatioindeksit.

Kuvassa 2. on dNBR-indeksin arvot yläpuolella ja siitä näkee hyvin, miten paljon havaintoja on ympäri kuvaa, joiden mukaan paloalue olisi hyvin pirstaleinen ja paljon todellista laajempi. Alemmassa osassa on RdNBR-indeksin perusteella laskettu $G_i(d)$ jossa naapuruusetäisyys on 2 pikseliä. Tulokset puhuvat puolestaan, kuva on huomattavan tarkkarajainen ja antaa erinomaisen kuvan palon intensiteetistä ja levinneisyydestä. Kuvassa on myös esitetty $G_i(d)$-indeksin lähes lineaarinen suhde δNBR-indeksin arvoihin. Lisäksi kuvasta ilmenee se millä tavalla luokittelu on tehty, jotta esitys olisi mahdollisimman selkeä. Kun tutkimuksen tuloksia verrattiin maastossa tehtyihin mittauksiin, olivat tulokset erittäin tyydyttäviä. Paikallisen spatiaalisen autokorrelaation mittareiden käyttäminen työkaluna pelkkien indeksien lisäksi paransi analyysin tarkkuutta ja luotettavuutta. Tutkimuksen tuloksia voi suoraan hyödyntää myös muualla maailmassa, sillä molempien antureiden aineisto kattaa koko maailman, ja niiden antamaa informaatiota voidaan käyttää suoraan kuvailluilla metodeilla.



Kuva 2: G(i)-indeksin käyttö metsäpaloalueen määrittämiseksi (Lanorte ym. 2013).

6 Yhteenveto ja pohdinta

Spatiaalisen autokorrelaation käsite ei siis ole aivan yksikertainen, eikä sen mittareiden käyttö ole aivan ongelmatonta. Tilastollisten analyysien tarkkuus paranee ja tutkittavasta ilmiön käyttäytymisestä tilassa saadaan paljon parempi kuva. Nykypäivän tilanteessa, jossa paikkatieto on läsnä jokapäiväisessä elämässä ja laskentakapasiteetti on halvempaa kuin koskaan ennen, ei ole mitään syytä jättää spatiaalista autokorrelaatiota huomiotta mallien suunnittelussa. Spatiaalisten mallien suunnitteluun ja käyttöön on saatavissa

hyvää referenssimateriaalia (esim. Wong & Lee 2005). Mallit, jotka ottavat tilalliset riippuvuudet huomioon ja käyttävät niitä hyväksi mallin suunnittelussa ja osana sen rakennetta ovat todennäköisesti tarkempia ja vähemmän virhealttiita kuin perinteiset tilastolliset mallit (Dormann ym. 2007: 619.)

Spatiaalisen autokorrelaation käsite on tarpeellinen niin luonnonmaantieteessä kuin kulttuurimaantieteessäkin, mutta näiden lisäksi sitä käytetään ekologiassa, biomaantieteessä, epidemiologiassa ja ekonometriassa(Griffith 2009: 344). Se on eräs tärkeimpiä maantieteilijöiden metodologisia saavutuksia, mutta se on jäänyt vähälle huomiolle monen maantieteilijän tutkimuksissa todennäköisesti sen matemaattisen ja tilastollisen luonteen vuoksi. Mallit jotka onnistuneesti ottavat huomioon spatiaalisen autokorrelaation, ovat usein laskennallisesti intensiivisiä ja vaativat syvempää tilastollista osaamista (Dormann ym. 2007: 617).

Legendre (1993: 1671) tarjoaa ekologeille ja ekologisen maantieteen tutkijoille työkaluja spatiaalisen autokorrelaation huomiomiseksi tutkimuksissa, painottaen spatiaalisen rakenteen huomioonottamista kaikissa analyyseissä. Vaikka artikkeli on julkaistu jo vuonna 1993, eivät kaikki biografian ja ekologian tutkijat vieläkään käytä spatiaalisen autokorrelaation huomioonottavia tilastollisia menetelmiä, koska niitä ei aina tunneta. Kuten voidaan huomata Dormann ym. (2007: 609) tutkimuksesta joka perehtyi juuri tähän kysymykseen 14 vuotta myöhemmin.

Kuten aiemmin todettiin spatiaalinen autokorrelaatio on erinomainen työkalu kun halutaan luoda esityksiä ja amalysoida rasteriaineistoja, mutta se on eräs suurimmista kompastuskivistä kun tarkoituksena on mallintaa ja ennustaa ilmiöiden käyttäytymistä tilassa (Longley ym. 2005: 87). Aineiston keräysvaiheessa on syytä perehtyä sen mahdollisiin autokorrelaatio-ominaisuuksiin, jotta otanta on mahdollisimman tehokasta. Spatiaalisen autokorrelaation ongelmaa voidaan lähestyä kahdesta suunnasta. Induktiivinen lähestymistapa perustuu aineiston pohjalta tehtyihin päätelmiin ilmiön luonteesta esimerkiksi autokorrelaatioindeksien laskentaa hyväksi käyttäen. Toinen vaihtoehto ilmenee hyvin regressiomallien kohdalla, jossa teoriaa ja tietämystä ilmiöstä käytetään mallin ja siihen mukaan otettavien parametrien valinnassa ennen analyysiä. (Longley ym. 2005: 95).

Tutkielman tarkoitus oli perehtyä spatiaalisen autokorrelaation käsitteeseen, ja avata sen monimuotoista taustaa. Lisäksi pyrin avaamaan sen aiheuttamia ongelmia sekä siitä koituvia hyötyjä maantieteellisessä ja muussa spatiaalisiin aineistoihin perustuvassa mallintamisessa. Terminä spatiaalinen autokorrelaatio kiteytyy hyvin Toblerin ensimmäiseen lakiin: kaikki on riippuvaista kaikesta, mutta lähellä olevat asiat ovat riippuvaisempia toisistaan kuin kauempana olevat. Spatiaalinen rakenne ilmiöiden taustalla vaikuttaa niiden ilmenemismuotoihin ja niiden esittämiseen kartalla ja matemaattisesti. Spatiaalinen autokorrelaatio on myös riippuvainen siitä, millä skaalalla ja tarkkuudella aineisto on kerätty, onko sen rakenne itseään toistava, miten otos on rakentunut ja millä menetelmillä sitä halutaan tutkia.

Tutkielmani keskittyi etenkin mallinnuksen osalta hyvin kapeaan siivuun lähinnä pienimmän neliösumman menetelmällä estimoituja malleja. Jos jatkan tämän aiheen parissa pro gradu tutkielman myötä voisi olla hyvä perehtyä probabilistisiin eli bayesiläisen tilastotieteen metodeihin. Esimerkiksi Monte Carlo - Markov Chain menetelmiin perustuviin malleihin, agenttipohjaiseen mallinnukseen ja niin edelleen. Agenttipohjainen mallinnus on eräs simulaatiomentelmä, jossa luodaan joukko itsenäisiä malleja jotka toimivat vuorovaikutuksessa toistensa kannsa. Agenttipohjaiset menetelmät perustuvat myös

yleensä probabilistiseen tilastolliseen traditioon. Tämän tutkielman puitteissa ei ole resursseja perehtyä täysin toisenlaisesta perusajattelusta lähtevään tilastollisen tradition tuomiin mahdollisuuksiin, joten pitäydyin frekventistisessä tilastollisessa lähtökohdassa. Bayesläinen tilastotiede on huomattavasti laskennallisesti intensiivisempää, mutta tarjoaa samalla enemmän tilaa tutkijan omille ennakkokäsityksille ja on monesti taipuvaisempi ottamaan huomioon taustalla olevia näkymättömiä rakenteita.

Spatiaalisten riippuvuuksien ja suhteiden tutkiminen on minusta se ydin, se mitä maantiede minulle tarkoittaa ja mihin haluan keskittyä. Tilastollisen osaamisen kehittäminen ja algoritmien ja mallien rakenteen ymmärtäminen ovat seuraavia kehityspolkuja minun tielläni kohti omaa siivuani tästä tieteenalasta. Kandidaatin tutkielman teko on avannut minulle uusia näkökulmia maantieteelliseen tutkimukseen ja antoi minulle varmistuksen siitä että olen oikealla koulutusalalla.

7 Lähteet

Anselin, L. (1995), Local Indicators of Spatial Association-LISA *Geograp hical Analysis*, 27 (2) Ohio State University Press, s. 94–115

Anselin, L. (2002), Under the hood Issues in the specification and interpretation of spatial regression models *Agricultural Economics* 27 , Elsevier ss. 247–267

Bivand, R. (2009) Applying Measures of Spatial Autocorrelation *Geographical Analysis* 41 Wiley-Blackwell publishing, ss. 375-384

Chiles, J-P. & P. Delfiner (1999) *Geostatistic: modeling spatial uncertainty* John Wiley and Sons, Inc. New York 695 s.

Cliff, A.D. & J.K. Ord,(1973) *Spatial autocorrelation* Pion, Lontoo, 178 s.

Cliff, A.D. & J.K. Ord(1981) *Spatial processes: Models and applications* Pion Lontoo 266 s.

Cressie, N.A. (1993) *Statistics for Spatial Data* John Wiley and Sons Inc. New Jersey 900 s.

Dormann, C.F., J.M. McPherson, M.B. Araujo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, A. Hirzel, W. Jetz, W.D. Kissling, I. Kühn, R. Ohlemüller, P. R. Peres-Neto, B. Reineking, B. Schröder, F.M. Schurr & R. Wilson (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review, *Ecography* 30, 609-628

Fotheringham, A.S. (2009) "The problem of spatial autocorrelation"and local spatial statistics,*Geographical Analysis* 41, 398-403

Getis, A. (2008)A History of the Concept of Spatial Autocorrelation:A Geographer's Perspective *Geographical Analysis* 40, 297-309

Griffith, D. A. (1992) What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics, *Espace géographique*, 21, 265-280.

Haining, R. P. (2003)*Spatial Data Analysis : Theory and Practice* Cambridge University Press, Cambridge 378 s.

Lanorte, A., M. Danese, R Lasaponara & B. Murgante (2013) Multiscale mapping of burn area and severity using multisensor satellite data

and spatial autocorrelation analysis *International Journal of Applied Earth Observation and Geoinformation* 20, 42-51

Legendre, P. (1993) Spatial Autocorrelation: Trouble or New Paradigm? *Ecology*, 74, 1659-1673

Lloyd, C.D. (2010) *Spatial data analysis: An introducition to GIS users* Oxford University Press, Oxford 206 s.

Longley, P.A. M. F. Goodchild, D.J. Maguire & D.W. Rhind (2005) *Geographical Information Systems and Science* John Wiley and Sons, Ltd Chichester, 517 s.

Miller, J.D. & A.E. Thode (2007) .Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR) *RemoteSensing of Environment* 109, 66-80

Miller, H.J. (2004) Tobler's First Law and Spatial Analysis, *Annals of the Association of American Geographers* 94: 284-289

Ord, J.K. & A. Getis (1995) Local Spatial Autocorrelation Statistics: Distributional Issues and an Application *Geographical Analysis* 27 286-306

Viladomat, J., R. Mazumder, A. McInturff , D. J. McCauley & T. Hastie (2014)Assessing the Significance of Global and Local Correlations under Spatial Autocorrelation: A Nonparametric Approach *Biometrics*, 70, 409–418

Wong, D. W. S. & J. Lee (2005) *Statistical analysis of geographic information with ArcView GIS and ArcGIS* John Wiley and Sons, Inc., Hoboken, New Jersey 429 s.