Matti Kurkela

Paikkatiedoista hyötyä suunnitteluun ja päätöksentekoon – SPATIO-TILASTOLLISEN ANALYYSIN MENETELMIN

Kirsi Virrantaus

Tiedon tuottamisen problematiikan rinnalla olisi jo pian keskityttävä myös tiedosta saatavan hyödyn luomiseen. Kirjoituksessa pyritään avaamaan paikkatiedon roolia tutkimuksen, suunnittelun ja päätöksenteon tukena.

INSPIRE-DIREKTIIVIN mukaisten paikkatietojen tuottaminen työllistää eurooppalaisia karttalaitoksia ja muita alan toimijoita. Erityisesti metadatan kerääminen ja sitä kautta paikkatiedon laadun varmistaminen vaatii osaamista ja resursseja. Tavoittelemme eurooppalaista yhtenäistä paikkatietovarantoa, jota voitaisiin käyttää luotettavasti ja sujuvasti yli hallinnollisten rajojen. Tiedon tuottamisen problematiikan rinnalla olisi jo pian keskityttävä enemmän tiedosta saatavan hyödyn luomiseen. Miksi paikkatietoja pitää kerätä niin ajantasaisina ja laadukkaina? Ketä ne hyödyttävät ja miten? Ei ainoastaan paikannuspalveluja ja navigointia maalla ja merellä. Ei ainoastaan realististen 3D-mallien muodostamista lähiympäristöstämme tai mahdollisuutta virtuaalimatkailuun. Valtavalla paikkatietomassalla on myös muuta potentiaalia. Tässä kirjoituksessa pyrin avaamaan toisenlaisen näkymän paikkatietoon ja sen rooliin *tutkimuksen, suunnittelun ja päätöksenteon tukena.*

PAIKKATIEDON MALLINTAMISESTA ILMIÖIDEN YMMÄRTÄMISEEN

Mitä paikkatieto on – sen opimme jo 80-luvulla. Nyt on aika ajatella asiaa uudestaan ja antaa paikkatiedolle lisää merkitystä ja arvoa. Paikkatieto määriteltiin Maanmittauslaitoksen LIS-projektin työn aikana tiedoksi, joka koostuu koordinaateista, geometriasta ja topologiasta sekä sijaintikohteisiin liitetystä ominaisuustiedosta. Myöhemmin Tekniikan Sanastokeskus on antanut ISO-standardin mukaisen, väljemmän määritelmän: "Paikkatieto on tieto kohteista, joiden paikka maan suhteen tunnetaan." (TSK 32, 432, 2005). Nämä määritelmät, vaikkakaan eivät rajaa pois tiedon sijainnillista jatkuvuutta, korostavat kovasti paikkatiedon luonnetta yksittäisten kohteiden kuvauksina ja geometrisinä malleina. Paikkatietoaineisto -termi täydentää hyvin paikkatieto-sanan määritelmää ja siitä lausutaankin "Paikkatietoaineisto on paikkatiedoista koostuva tietoaineisto." (TSK 432 2005) Tämän kirjoituksen tarkoituksena ei ole keskittyä paikkatiedon määritelmiin, mutta tietynlainen laaja-alaisempi ymmärrys paikkatiedon roolista muussakin mielessä,

kuin vain todellisuuden konkreettisten kohteiden mallintajana, on artikkelin pohjalla. Haluan korostaa paikkatietoa sisältävien aineistojen merkitystä maailmassa tapahtuvien ilmiöiden kuvaajina.

Otetaan esimerkiksi tavanomainen kuntarekisteri, joka sisältää paikkatietoa rakennuksista. Useimmiten kunnan karttajärjestelmä sisältää rakennusten nurkkapisteiden koordinaatit ja pisteitä yhdistävät janat tai murtoviivat; rakennukseen liittyvät ominaisuustiedot löytyvät yhteyksillä rakennusvalvonta- ja kiinteistöosiin. Tämä järjestelmä tukee niin kartantuotantoa kuin erilaisia kyselysovelluksiakin. Mutta ajan tasalla pidettävä ja historiatiedot tallentava rekisterikokonaisuus sisältää aikadimensiossa tarkasteltuna kuvauksen koko rakennuskannan kehittymisestä. Kartat, joita tuotetaan, ovat rakennuskanta -ilmiön aikasidonnaisia kuvauksia. Karttoja ja tietokantahaluja tarvitaan päivittäisessä toiminnassa, mutta ilmiötä kuvaava tietoaineistojen joukko muodostaa kokonaisuuden, josta saamme irti informaatiota historiasta ja tukea tulevaisuuden ennakointiin sekä myös selityksiä moniin kiinnostaviin tapahtumiin yhdyskunnassa. Tätä informaatiota voidaan sitten käyttää suunnittelun ja päätöksenteon tukena.

Samalla tapaa esimerkiksi sademäärää tai lumen syvyyttä kuvaavia sääkarttoja voidaan kutsua sade- tai lumi-ilmiöiden kuvauksiksi. Olennaista tässä on, että yksittäinen sääkartta tai lumitilannetta kuvaava kartta ei suinkaan kuvaa koko ilmiötä, vaan ainoastaan yhden tietyn ajanhetken tilanteen. Asioiden sijainti yhdellä hetkellä tietyssä paikassa on seurausta kokonaisilmiöstä, jonka periaatteiden tunteminen yleisellä tasolla on tutkimusnäkökulmasta hyödyllisempää kuin yksittäisen tilannetiedon tunteminen. Käyttäessämme paikkatietoaineistoja suunnittelussa meidän tulisi pystyä näkemään "metsä puilta" eli ymmärtämään ilmiö yhden kartan tai paikkatietoaineiston tulkinnan sijaan. Tähän tehtävään saadaan hyvä teoreettinen ja menetelmällinen tuki *spatio-tilastollisesta analyysistä*.

TILASTOLLISET MENETELMÄT VAATIVAT LAAJENNUKSIA

Ilmiöiden tutkiminen tilastotieteen avulla tarkoittaa todennäköisyysmatematiikkaan pohjautuvien mallien luomista tarkasteltavasta aineistosta. Kerättyjen aineistojen perusteella pyritään tutkimaan kohteena olevan ilmiön käyttäytymistä ja mahdollisia useiden ilmiöiden yhteiskäyttäytymisiä. Paikkatiedon luonteeseen kuuluu, ettei se täytä kaikkia tavanomaisia tilastomatematiikan oletuksia ja siksi näitä menetelmiä ei voida soveltaa suoraviivaisesti paikkatiedon analysointiin. Oletukset satunnaisuudesta ja muuttujien riippumattomuudesta eivät useinkaan tutkittavissa paikkatietoaineistoissa toteudu. Spatio-tilastollisen analyysin peruskäsitteeksi voidaan määritellä *spatiaalinen stokastinen prosessi*. Sillä tarkoitetaan "tilastollista mallia, joka määrittää kiinnostuksen kohteena olevaa paikkasidonnaista ilmiötä kuvaavien todennäköisyysmuuttujien jakauman" (Heikkinen, 2004). Stokastisuus tarkoittaa "satunnaista". Mitä on sitten "spatiaalinen satunnaisuus" ja "spatiaalinen satunnainen prosessi"? Miten spatiaalisuuden erityispiirteet voidaan ottaa huomioon tilastollisten menetelmien yhteydessä?

Paikkatietoaineistoista voidaan tunnistaa tiettyjä, nimenomaan sijaintiin perustuvia ns. *sijaintirelaatioita*, joita tavanomaisessa tiedossa ei esiinny. Perusrelaatioita ovat *etäisyys* ja *suunta*, jotka voidaan laskea paikkatiedon koordinaattitietojen avulla. Muita keskeisiä sijaintirelaatioita ovat *viereisyys (adjacency)* ja *jatkuvuus (connectivity)*. Nämä ovat keskeisiä *topologisia* käsitteitä paikkatietokohteita mallinnettaessa; esimerkiksi kuntarajojen kuvaaminen tehokkaana paikkatietomallina edellyttää viereisten kuntien yhteisen rajan tunnistamisen ja tämän relaation toteuttamisen tavalla tai toisella paikkatiedon hallinnassa. Tieverkon mallinnus edellyttää taas verkon osien *solmu*pisteiden tunnistamista, nämä solmut ilmaisevat osien jatkuvuuden. Vaikka sijaintitietoja hallitaan tavanomaisen relaatiotiedonhallinnan työkaluin, nämä sijaintirelaatiot on tehokkuuteen pyrittäessä huomioitava ohjelmistolaajennuksin. Relaatiot konkretisoituvat nimenomaan mallinnuksen yhteydessä, mutta ne ovat keskeisiä myös spatio-tilastollisen analyysin toteuttamisessa, koska ne ovat työkaluja spatiaalisen tiedon erityisominaisuuksien ja spatiaalisen järjestyksen huomioon ottamiseen.

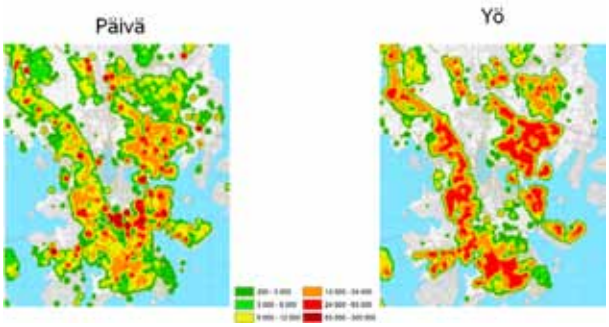
HOTSPOTIT JA SPATIAALINEN AUTOKORRELAATIO

Spatio-tilastolliset menetelmät perustuvat ns. ensimmäisen ja toisen kertaluvun vaikutuksiin. *Ensimmäisen kertaluvun vaikutukset* perustuvat prosessin *odotusarvoon* ja tarkastelumittakaava on pieni. Käytännön analyysimenetelminä käytetään visuaaliseen tulkintaan perustuvia menetelmiä, kuten ns. Kernel-tiheyspintamenetelmä tai ns. koealamenetelmät, joissa tarkasteltavalta alueelta otetaan koealoja ja ilmiön intensiteettiä koealoilla verrataan yleiseen odotusarvoon. *Kernel-tiheyspinta*-menetelmässä havaittujen kohteiden määrättyä ympäristöä painotetaan sopivalla tavalla ja näiden painotettujen alueiden summasta muodostuu ilmiön jakautumista kuvaava pinta. Menetelmällä saadaan yleiskuva ilmiön käyttäytymisestä ja voidaan visuaalisella tarkastelulla tunnistaa selkeitä tihentymiä, *hotspoteja*. Nähdään, että aineisto ei olekaan satunnaisesti jakautunut, vaan jostain syystä keskittynyt ryppäiksi tietyille alueille.

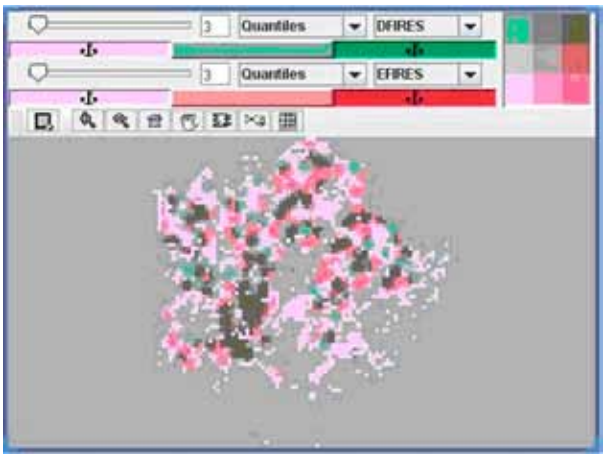
Toisen kertaluvun vaikutusten tarkastelu on yksityiskohtaisempaa ja perustuu kohteiden välisiin *etäisyyksiin*. Paikkatiedon

$$I = \frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}$$

KUVA 1. Moranin indeksin laskukaava (O'Sullivan & Unwin, 2003). Kaavassa oikealla ylhäällä on analyysin keskeinen termi, kovarianssitermi. Muuttuja *y* tarkoittaa tutkittavan aineiston yhden tietokohteen ominaisuusarvoa, $\bar{y}$ tarkoittaa aineistosta laskettua keskiarvoa, joten termissä kerrotaan keskenään kahden kohteen (*i* ja *j*)laskettu erotus keskiarvosta. Kaksi summalauseketta aikaansaavat sen, että tietokohteet käydään läpi tietyssä numerojärjestyksessä. Keskeinen tekijä lausekkeessa on *w*, ns. *painomatriisi* (weights matrix), jota voidaan soveltaa viereisyyden kuvaamiseen. Käytettäessä painomatriisia *viereisyysmatriisin* tapaan matriisin *w* alkio kohdassa (*i, j*) on 1, jos *i* ja *j* jakavat yhteisen rajaviivan (viereisyys-relaatio) , alkion arvo on 0, jos näin ei ole. Näin lauseke summaa kohteiden kerrotut ominaisuustietoerotukset aina, jos kohteet ovat käytetyn määritelmän mukaan viereisiä, muutoin tuota tuloa ei oteta huomioon. Yleisemmin painomatriisilla voidaan ottaa huomioon laajempi naapurusto ja ykkösillä ja nollilla hallita käsiteltävän alueen laajuutta. Muut termit kaavassa normeeraavat tulosta ja aikaansaavat sen, että positiivisen autokorrelaation tilanteessa indeksi saa positiivisen arvon ja negatiivisen autokorrelaation tilanteessa negatiivisen arvon. Jos indeksin arvo on 0, kohteet ovat aineistossa satunnaisesti sijoittuneet.



KUVA 2. Kernel-tiheyslaskennan tulos Helsingin eräästä onnettomuusaineistosta. Erikseen on laskettu päivä- ja yöonnettomuudet. Visuaalinen tulos paljastaa, että onnettomuudet käyttäytyvät yöllä eri tavoin kuin päivällä. Tutkimuksesta on julkaistu artikkeli (Krisp et al., 2008).



KUVA 3. Kahden muuttujan spatiaalista jakaumaa kuvaava kaksimuuttujakartta (bivariate map), joka paljastaa tarpeen tutkia ilmiötä eri aikakategorioissa. Kuvassa esitetään päiväonnettomuuksien jakauman suhde iltaonnettomuuksien jakaumaan. Tummat sävyt kuvaavat alueita, joissa kummassakin kategoriassa on korkea tiheys. Punainen kuvaa alueita joissa onnettomuuksia tapahtuu paljon iltaisin, mutta vain vähän päivisin. Vihreillä alueilla onnettomuuksien jakaumien suhde on päinvastoin. Esimerkiksi tällaista visuaalista työkalua kokenut analysoija käyttää aineistoon perehtyessään ja lopullista analyysimenetelmää suunnitellessaan. Tulos olisi helpommin tulkittavissa yksinkertaisen pohjakartan avulla. (Spatenkova, 2009).

ominaispiirteistä keskeisintä, *spatiaalista autokorrelaatiota* analysoidaan useimmiten juuri etäisyyteen pohjautuvilla menetelmillä. Spatiaalinen autokorrelaatio tarkoittaa, että ilmiö pyrkii lähiympäristön samanlaisuuteen – toisiaan lähellä olevissa sijainneissa mitatut arvot pyrkivät olemaan samanlaisempia kuin toisistaan kauempana olevissa sijainneissa. Ilmiö on analoginen aikasarjoista tutun autokorrelaation kanssa; esimerkiksi lämpötila muuttuu siten, että lämpökäyrässä viereisinä ajanhetkinä lämpötila on selkeästi samanlaisempi kuin satunnaisesti otettuina ajanhetkinä. Spatiaalisen autokorrelaation laki näyttää toimivan hyvin mm. luonnon *spatiaalisesti jatkuvissa* ilmiöissä kuten maaperä, kasvillisuus, sademäärä, mutta ei kuitenkaan kaikissa esimerkiksi ihmisen rakentamissa rakenteissa, jotka ovat luonteeltaan *diskreettejä* kohteita. Spatiaalinen autokorrelaatio vaikuttaa siten, että emme voi ottaa satunnaista otantaa paikkatietoaineistosta analyysiä varten, koska näin saatu *otos* ei ole koko *populaatiota* oikein kuvaava. Ennen otoksen ottamista ilmiön käyttäytyminen tulee tutkia ja mahdollinen spatiaalinen autokorrelaatio tunnistaa. Sen jälkeen otantaan voidaan soveltaa spatiaalisia otantamenetelmiä, jotka ottavat huomioon autokorrelaation tai muun spatiaalisen järjestyksen.

Autokorrelaation tunnistaminen auttaa ilmiön tutkimisessa ja yleensä se tutkitaankin ensimmäisenä tehtävänä analyysiin ryhdyttäessä. Autokorrelaation tunnistamiseksi on monia laskennallisia ja visuaalisia menetelmiä, jotka siis perustuvat useimmiten kohteiden etäisyyksien tarkasteluun. Tunnetuin on *Moranin indeksi (Moran's I)*, joka laskee sijaintien keskinäistä kovarianssia ja tuottaa siitä autokorrelaation tunnistamiseksi yksinkertaisen arvon.

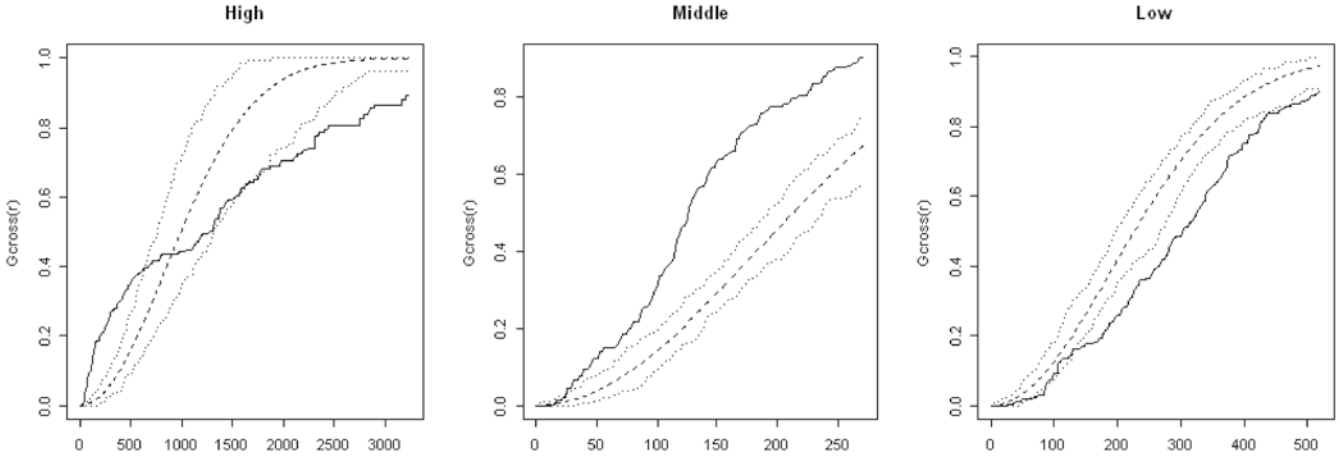
Kuvan 1 tekstissä on kuvattu kuinka viereisyys-relaatiota hyödynnetään spatiaalisen autokorrelaation laskemisessa. Spatiaalisen autokorrelaation ymmärtämistä helpottaa esimerkki pelilaudasta. Tavanomainen "shakkilauta"-järjestys, jossa mustat ja valkoiset ruudut sijaitsevat säännöllisesti joka toisessa ruudussa, on erinomainen esimerkki negatiivisesta spatiaalisesta autokorrelaatiosta. Mustat ja valkoiset ovat järjestyneet niin, ettei synny mustia tai valkoisia yhtä ruutua suurempia alueita. Jos mustat ja valkoiset olisivat laudalla sijoittuneet siten, että kaikki mustat olisivat toisella puoliskolla ja kaikki valkoiset toisella puoliskolla, vallitsisi laudalla positiivinen spatiaalinen autokorrelaatio.

Spatiaalinen autokorrelaatio voi olla suuntautunutta ja paikkatiedot voivat ilmentää tutkittavassa prosessissa trendiä, eli pyrkimystä nouseviin tai laskeviin arvoihin. Konkreettinen esimerkki tästä on korkeusmalli harjualueelta, jossa voidaan havaita selkeä nouseva-laskeva-trendi sekä suuntariippuva spatiaalinen autokorrelaatio. Tämän käyttäytymisen voi tunnistaa laskennallisesti käyttämällä ns. *variogrammipilvi*-kuvaajaa. Variogrammipilvi syntyy kaikkien pisteparien välisen mitatun ominaisuuden eron ja keskinäisen spatiaalisen etäisyyden avulla (samaan tapaan kuin Moranin indeksissä paitsi nyt verrataan kahden pisteen ominaisuutta, eikä eroa keskiarvosta) ja se kuvaa ilmiön käyttäytymistä jopa eri ilmansuunnissa. Variogrammipilvi piirretään koordinaatistoon kuvaajana, jossa x-akselilla on kohteiden välinen etäisyys ja y-akselilla pisteparin ominaisuuksien erotuksesta laskettu luku. Tuloksen tulkinta perustuu visuaaliseen tarkasteluun, tutkitaan miten pisteet sijoittuvat kuvaajaan. Jos pisteitä on paljon lähellä origoa, se tarkoittaa että aineistossa on paljon pisteitä, jotka ovat lähellä toisiaan sekä sijainnillisesti, että ominaisuutensa arvon puolesta ja siis aineistossa ilmenee spatiaalinen autokorrelaatio. Variogrammi on keskeinen työkalu myös interpoloinnista tutussa kriging-menetelmässä, jossa sitä käytetään juuri spatiaalisen autokorrelaation alueellisen laajuuden tunnistamiseen.

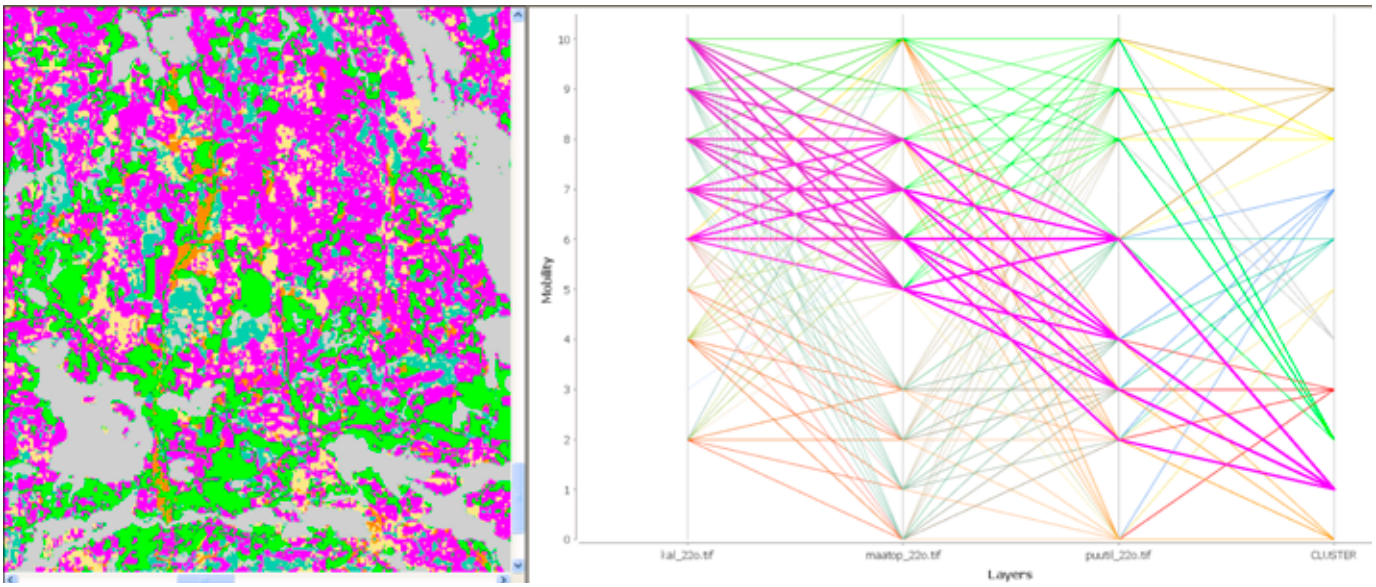
Spatiaalisesta autokorrelaatiosta ja sen tutkimisesta löytyy esimerkkejä ja teoriaa spatio-tilastollisen analyysin perusoppikirjoista ja oppimateriaaleista (Heikkinen, 2004; O'Sullivan & Unwin, 2003).

SPATIO-TEMPORAALISET ILMIÖT

Spatiotilastollinen analyysi tarjoaa menetelmiä myös ilmiöiden ajallisen käyttäytymisen paljastamiseen. Tutkiminen voi perustua



KUVA 4. Esimerkki väestötiheyden ja rakennuspalojen keskinäisen korrelaation analyysistä. Kuva on Olga Spatenkovan väitöskirjatutkimuksesta, jossa hän käytti etäisyyteen perustuvaa G-funktiota. Kuvan käyrät kuvaavat G-funktiota (= ehyt viiva), käytetyn teoreettisen mallin mukaista satunnaista jakaumaa (= katkoviiva) ja sen ympärillä simuloinnin tuloksena saadut käyrät. Analyysiä varten alueet on jaettu väestön tiheydeltään korkean tiheyden, keskitiheyden ja matalan tiheyden alueisiin. G-funktiot on laskettu erikseen näille alueille. Tuloksena todetaan, että korkean ja keskitiheyden alueilla rakennuspaloja tapahtuu enemmän kuin satunnaisuuden pohjalta voidaan olettaa. (Spatenkova, 2009).

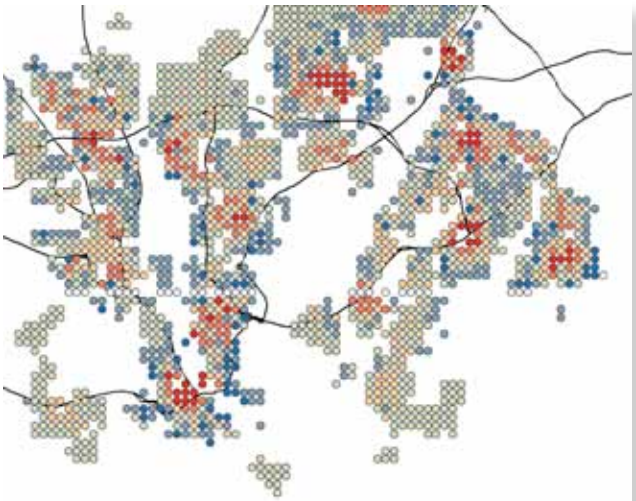


KUVA 5. Esimerkki tutkimuksessa kehitetystä sovelluksesta, jossa k-means-menetelmällä (vasen puoli) saatua klusterointia (samanlaisten pikseleiden muodostamia luokkia, joissa samanlaisuus lasketaan kolmen ominaisuuden avulla) tarkastellaan PCP-visualisointina. Tässä yksi murtoviiva kuvaa yhtä datavektoria ja voi kuvata jopa yli 10 000 pikselin ryhmää. Oikealla puolella kuvassa koordinaatistoon kuvautuvasta murtoviivajoukosta on havainnollistettu yksi tietty klusteri, sama klusteri on kuvattu myös kartalla violetilla sävyllä (Nikander et al., 2010).

aluksi visualisointeihin. Kuvassa 2 on esimerkki tutkimuksesta, jossa tarkasteltiin onnettomuuksien esiintymistä Helsingissä; onnettomuudet on jaettu hyvin karkealla temporaalisella resoluutiolla päivä- ja yöonnettomuuksiin. Tämä kuvapari osoittaa tutkijalle sen, että ilmiö käyttäytyy eri tavoin päivällä ja yöllä. Voidaan olettaa, että onnettomuuksia selittävät syyt ovat myös erilaiset päivällä ja yöllä. Kuvasta 3 voidaan havaita, että tarkastellut päiväonnettomuudet ja yöonnettomuudet sijaitsevat lähtökohtaisesti eri paikoissa ja myös, että alueella on tiettyjä sijainteja, joissa onnettomuuksia tapahtuu sekä päivällä että yöllä. Tämä visualisointi kertoo tutkijalle, että onnettomuuksien syitä tulee hakea erikseen erityyppisten alueiden kohdalla. Kuvan 2 kartat on tuotettu Kernel-tiheysmenetelmällä ja kuvassa 3 on käytetty ns. *kaksimuuttujakartta*-menetelmää.

SPATIAALISEN KORRELAATION PALJASTAMINEN

Edellä kuvattu tutkimus voi jatkua siten, että tutkija tarkastelee valittuja ilmiöitä ja laskee *spatiaalista korrelaatiota* eri asioiden välillä. Spatiaalisen korrelaation tutkiminen voi perustua kahden



KUVA 6. Kaavamainen esimerkki GWR-analyysistä, jossa pyritään selittämään onnettomuuksia eri sosio-ekonomisten muuttujien avulla. Kuvassa eri värein on havainnollistettu alueellisesti vaihtelevat parhaat selittäjät. Kuva esittää Helsingin aluetta, ja olisi havainnollisempi pohjakartan kanssa. Kuva on opiskelijoiden harjoitustyöstä v. 2008 järjestetyllä GWR-kurssilla, jolla menetelmän kehittäjät prof. Stewart Fotheringham ja Martin Charlton Irlannista (National University of Ireland, Center of Geocomputation) luennoivat Teknillisessä korkeakoulussa. Kyseinen GWR-ohjelmisto on vapaasti ladattavissa Internetistä ja menetelmä löytyy myös uusimmasta ArcGIS-versiosta.

eri ilmiön havaittujen ja mitattujen kohteiden välisten etäisyyksien laskemiseen. Täysin satunnaisessa järjestyksessä kohteiden väliset etäisyydet käyttäytyvät tietyllä, odotetulla tavalla. Tätä tapaa verrataan tarkasteltavasta aineistosta laskettuun dataan. Menetelmä voi perustua esimerkiksi kumulatiiviseen frekvenssikäyrään. Ns. G-funktio kuvaa kahden aineiston lähimpien pisteparien etäisyyden käyttäytymistä. Kuvaajassa esitetään y-akselilla kullakin x-akselilla kuvatulla etäisyydellä esiintyvien pisteparien määrä. Kuvaajan muodosta voidaan päätellä ilmiöiden spatiaalinen korrelaatio, eli pyrkivätkö ilmiöt sijaitsemaan maantieteellisesti lähellä toisiaan. Jos käyrä nousee jyrkästi lähellä origoa, ja sitten tasaantuu, kahdessa aineistossa on paljon toisiaan lähellä olevia pistepareja. Empiirisestä aineistosta laskettua käyrää voidaan verrata satunnaisuutta kuvaavaan teoreettiseen käyrään.

Menetelmällä voidaan tarkastella esimerkiksi saatuja poliisihälytyksiä ja alkoholin anniskelupaikkojen sijaintia. Tutkimus*hypoteesina* on, että poliisihälytyksiä tapahtuu keskimääräistä enemmän anniskelupaikkojen läheisyydessä. Kerätyn datan perusteella voidaan laskea jokaisen poliisihälytyksen etäisyys lähimpään alkoholin anniskelupaikkaan. Lasketuista etäisyyksistä muodostetaan kumulatiivinen frekvenssikäyrä, jonka muodosta voidaan visuaalisesti tulkita onko näiden kahden asian välillä korrelaatiota. Korrelaatio on, jos muodostuu jyrkästi nouseva käyrä, eli aineistossa on paljon hyvin lyhyitä etäisyyksiä.

MONIMUUTTUJAMENETELMÄT

Voidaan myös käyttää ns. *monimuuttujamenetelmiä*, jotka mahdollistavat monien ilmiöiden yhtäaikaisen tarkastelun. Nämä menetelmät eivät ole ainoastaan spatio-tilastollisia, vaan monet niistä perustuvat laskennalliseen käsittelyyn ja/tai visualisointiin ja tulkintaan. Tilastollinen ja laskennallisen sekä visuaalinen käsittely muodostavatkin todellisessa analyysiprosessissa ketjun, jossa edellisen analyysin tuloksia käytetään seuraavassa. Tässä artikkelissa keskitytään tilastollisiin menetelmiin, vaikkakin raja tilastollisten ja ns. *datan louhinta*menetelmien välillä ei ole aivan täsmällinen. Seuraavassa mainitaan nimeltä myös muutama laskennallinen ja visuaalinen louhintamenetelmä, koska ne ovat käyttökelpoisia yhdessä tilastollisen analyysin kanssa.

Monimuuttuja-analyysimenetelmissä tarkastellaan useaa asiaan vaikuttavaa ilmiötä/muuttujaa yhtä aikaa. Tehokas monimuuttujamenetelmä on *PCP* (parallell coordinate plot, rinnakkaiskoordinaatisto), jossa murtoviivat kuvaavat yhtä sijaintia ja vertikaaliset akselit ominaisuuksia. Aineistoa tarkastellaan permutoimalla (vaihtelemalla) akseleiden järjestystä ja hakemalla *klustereita*, eli samanlaisia sijaintikohderyhmiä. Tulkinnassa käytetään myös karttaa, jossa tulos näkyy samanlaisten alueiden hahmottumisena. Sijainnin ja monimuuttujaista samanlaisuutta voidaan myös laskea esimerkiksi *k-means*-klusterointimenetelmällä. K-means klusteroinnissa aineistosta haetaan samanlaisia kohteita ja samanlaisuus lasketaan n-ulotteisessa avaruudessa jollain etäisyydenlaskentaperiaatteella. Menetelmä tuottaa halutun määrän "samanlaisia" kohteita sisältäviä luokkia. Kuvassa 5 on esimerkki edellä mainittujen menetelmien käytöstä yhdessä (Nikander et al., 2010). Muita monimuuttujaanalyysiin käytettyjä menetelmiä ovat *itseorganisoituvat kartat* (SOM; self-organizing maps). SOM on havainnollinen ja tehokas menetelmä eikä vaadi välttämättä teorian syvällistä ymmärtämistä. SOM on periaatteessa monimuuttujaklusterointia, jossa laskenta perustuu opetusaineistosta tuotettuihin sääntöihin ja niiden soveltamiseen varsinaiseen tutkimusaineistoon. Visuaaliset ja laskennalliset menetelmät toimivat analyysiprosessissa hyvin aineiston tarkastelun ensi vaiheissa. Niillä pystytään saamaan käsitystä aineistojen luonteesta ja käyttäytymisestä ilman hypoteesiksi asetettuja malleja ja teorioita. Kun vihjeitä aineiston autokorrelaatioista ja spatiaalisesta korrelaatiosta saadaan, voidaan luoda tilastollisesti testattavia hypoteeseja ja edetä tarkempaan tarkasteluun.

LÄHTEET

Ahola, T., Virrantaus, K., Krisp, J. M. & Hunter, G. (2007) A Spatio-Temporal Population Model To Support Risk Assessment And Damage Analysis For Decision-Making. International Journal Of Geographical Information Science, 21, 935–953.

Heikkinen,J., (2004), Luentomoniste "Spatiaalinen tilastotiede", Helsingin yliopisto, Matematiikan ja tilastotieteen laitos, www.rni.helsinki.fi/~jmh/ss04/spat04handout.pdf.

Nikander,J., Kantola,T., Virrantaus,K., Transparent exploratory analysis of spatial data, julkaisematon käsikirjoitus, 2010.

O'Sullivan,D, Unwin,D., (2003) Geographic Information Analysis, John Wiley & Sons.

Spatenkova, O., (2009) Discovering Spatio-Temporal Relationships: A Case Study of Risk Modelling of Domestic Fires, Aalto-yliopiston teknillinen korkeakoulu, Maanmittaustieteiden laitos, Väitöskirja, Teknillisen korkeakoulun geoinformatiikan ja kartografian julkaisuja TKK-GC-9.

TSK 32, (2005), Geoinformatiikan sanasto, Sanastokeskus TSK, Helsinki, 2005.

GWR (geographically weighted regression, maantieteellisesti painotettu regressio) on monimuuttujamenetelmistä tilastolliseksi selkeimmin luokiteltava menetelmä. GWR on tarjolla joissain kaupallisissa GIS-ohjelmistoissa ja se tarjoaa tehokkaan menetelmän tutkia globaalin regressiomallin sijaan lokaaleja malleja, eli tunnistaa eri alueilla eri muuttujat ilmiön selittäjiksi. GWR:n pohjana on tavanomainen lineaarinen regressio, jossa pyritään mallintamaan suhdetta selitettävän muuttujan ja usean riippumattoman selittäjän välille. Perusmuodossa sama malli pätee koko tarkastelualueelle. Maantieteellisesti painotetussa muodossa mallia rakennetaan pienemmille alueille olettaen spatiaalinen autokorreloituvuus. Tuloksena saadaan useita malleja, joissa parhaiden selittävien muuttujien joukko on erilainen eri alueilla. Esimerkiksi Spatenkovan tutkimuksessa todettiin, että rakennuspaloja näyttävät selittävän Helsingin eri alueilla eri sosio-ekonomiset muuttujat (Kuva 6).

SPATIAALISEN ANALYYSIN POTENTIAALI

Spatiaalinen tilastollinen analyysi tarjoaa teoreettisesti kestävän perustan paikkatietoanalyysille. Menetelmien joukko on kirjava ja kaupalliset ohjelmistot tarjoavat vielä vain vähän kelvollisia työkaluja. Käyttäjän on hallittava menetelmien teoria ja myös osattava käyttää eri ohjelmistoja yhdessä.

Olga Spatenkovan väitöskirjatyössä (Spatenkova, 2009) käytettiin useita yllä mainittuja menetelmiä ja pyrittiin arvioimaan niiden käyttökelpoisuutta saatavuuden, vaativuuden ja soveltuvuuden osalta. Spatio-tilastollisten menetelmien vaatimus käyttäjältään on tilastollisten menetelmien tuntemus. Monet laskennalliset työkalut visuaalisine menetelmineen ovat siinä mielessä tilastollista analyysiä helpompia käyttöön otettavaksi. Spatio-tilastolliset menetelmät ovatkin tällä hetkellä lähinnä tutkijoiden työkaluja, mutta opetuksella ja tuotekehityksellä päästään pian tilanteeseen, jossa menetelmät siirtyvät GIS-käyttäjien arkityökaluiksi. Internetistä löytyy ilmaiseksi käyttöön ladattavia ohjelmistoja kuten R ja GeoVista, sekä yllä mainittu GWR. Yliopistojen täydennyskoulutuksessa spatiaalinen analyysi on myös painopistealueena. Aihealueesta on myös olemassa perusoppikirjaksi erittäin sopiva ja selkeästi kirjoitettu teos Geographical Information Analysis (O'Sullivan & Unwin, 2003), jota itsekin käytän kurssikirjana.

Kirjoittaja on Aalto-yliopiston Insinööritieteiden korkeakoulun professori ja Maanmittaustieteiden laitoksen Geomatiikan koulutusohjelman johtaja. Sähköposti kirsi.virrantaus@tkk.fi.

SELITYKSIÄ ARTIKKELISSA KÄYTETYILLE TERMEILLE

Nämä termien selitykset eivät ole pääosin täsmällisiä määritelmiä, vain joitakin on poimittu Geoinformatiikan sanastosta (TSK 32, 2005). Termit on pyritty avaamaan artikkelin lukemisen helpottamiseksi.

spatiaalinen	sijaintisidonnainen, geometriaan tai topologiaan pohjautuva, avaruudellinen
stokastinen prosessi	matemaattinen prosessi, jolla kuvataan ajassa sattumanvaraisesti etenevää todellisuuden prosessia (ilmiötä)
sijaintirelaatio	suhde, joka ilmaisee miten kohde sijaitsee avaruudessa suhteessa toiseen kohteeseen, sijaintisuhde
viereisyys	adjacency (en), sijaintisidonnaisten kohteiden välinen suhde, jossa kohteet jakavat esimerkiksi yhteisen rajaviivan
jatkuvuus	connectivity (en), verkkoteorian peruskäsite, sijaintisidonnaisten kohteiden yhteydessä suhde, jossa esimerkiksi viivojen päätepisteet yhtyvät
odotusarvo	todennäköisyyksillä painotettu keskiarvo
hotspot	kuuma piste; sijainti, jossa mitattavaa asiaa on merkittävästi enemmän kuin muualla, vastakohta kylmä piste
spatiaalinen autokorrelaatio	ilmiö korreloi itsensä kanssa spatiaalisesti, spatiaalinen itsekorrelaatio, lähellä toisiaan olevat asiat ovat samanlaisempia kuin kaukana toisistaan olevat, vrt. Toblerin maantieteen 1. laki
spatiaalinen jatkuvuus	ilmiö saa arvon jokaisessa tutkimusalueen pisteessä
diskreetti	erillinen, epäjatkuva
otos	näytejoukko
populaatio	kokonaisjoukko
variogrammipilvi	variogram cloud (en), kuvaaja, jossa kaikkien mitattujen kohteiden ominaisuuksia verrataan pareittain toisiinsa ja esitetään ominaisuustiedon ero maantieteellisen etäisyyden funktiona
spatiaalinen korrelaatio	korrelaatio, joka on spatiaalisesti riippuva; sijainnista riippuva korrelaatio
klusteri	rypäs, joukko samanlaisia kohteita
topologia	tieto, joka kuvaa paikkatietokohteiden tai niiden osien välisiä sijaintisuhteita, jotka säilyvät muuttumattomina jatkuvissa muunnoksissa (TSK 32)
solmu	solmupiste, linkin (verkossa) topologinen päätepiste (TSK 32)
viereisyysmatriisi	matriisi, jossa riveinä ja sarakkeina järjestykseen listatut kohteet ja alkioina nollia ja ykkösiä sen mukaan ovatko kyseisen alkion rivillä ja sarakkeessa mainitut kohteet viereisiä
hypoteesi	oletus, jota analyysissä testataan
datan louhinta	pyritään löytämään tietomassasta olennaista informaatiota käyttäen erilaisia laskennallisia, tilastollisia ja visuaalisia menetelmiä