591

Sairauden alueellisen ryvästymisen tutkiminen spatiaalisella tapaus-verrokki -asetelmalla

Arto Haikonen

4.9. 1998

Pro gradu -Tutkielma
Jyväskylän yliopisto
Tilastotieteen laitos

Tiivistelmä

Arto Haikonen

Sairauden alueellisen ryvästymisen tutkiminen spatiaalisella tapaus-verrokki -asetelmalla.
Tilastotieteen Pro gradu-tutkielma, Jyväskylän yliopisto, 1998
53 sivua + 7 sivua liitteitä.

Pro gradu-tutkielmassa esitetään yksityiskohtaisesti tilastollinen analyysimenetelmä epidemitapausten spatiaaliselle ryvästymiselle. Lähestymistapana on spatiaalinen tapaus-verrokki -asetelma, missä on käytössä tapausten ja kontrollien asuinpaikan koordinaatit. Tilastotieteellisesti keskeinen ongelma on tutkia tapausten ryvästymistä, kun populaatio jakautuu alueelle epätasaisesti. Sovellettava menetelmä perustuu kahden lajin pisteprosessiin ja erityisesti sen toisen kertaluvun ominaisuuteen, jonka avulla pyritään johtamaan testisuure ja jakaumaapproksimaatiot. Pisteprosessilla pyritään mallintamaan sairaustapausten ja verrokin heterogeenisuutta. Aiheeseen on perehdytty kirjallisuuden sekä simuloidun pisteprosessin avulla.

Tutkimus tehdään osana Kansanterveyslaitoksen projektia, jossa kehitetään menetelmää ympäristön ihmisen terveydelle aiheuttamien riskien mallintamiseksi. Menetelmän tavoitteena on valmistuttuaan olla reaaliaikaisena päätöstukijärjestelmänä, kun mietitään alueellisia toimenpiteitä sairastuvuusriskiä kohottavien ympäristötekijöiden vaikutusalueella.

Sovellusaineistona on vuosina 1981-92 keuhkosyöpään sairastuneet miehet Pohjois-Savossa olevalla alueella, jonka koko on 2500 km^2. Tapausten asuinpaikan koordinaatit on ollut käytössä ja niitä on kaikkiaan 122. Kontrolleiksi on tilattu väestörekisterikeskukselta 400 samalla alueella vuonna 1980 asuneita, iän suhteen samoin jakautuneita miehiä.

Menetelmä toimi hyvin simuloidulle, klusteroituneelle, aineistolle. Analysoinnissa havaittiin, että sekä tapaukset että kontrollit ovat klusteroituneet. Tapaukset eivät kuitenkaan olleet klusteroituneet ympäristöefektiä enempää.

Yllättävää oli, että tutkimusalueelta ei löytynyt sairastuvuusylimäärää niin kuin etukäteen odotettiin. Keuhkosyöpäaineiston tapaukset ja kontrollit olivat molemmat klusteroituneet ja tapausten spatiaalinen jakautuminen vastasi populaation jakaumaa. On huomattavaa, että tutkimuksessa ei ole suoritettu vakiointia sosioekonomisen taustan mukaan, mikä saattaa olla vahva sekoittava tekijä.

Sisältö

1 Johdanto

Viime aikoina on lisääntynyt tietoisuus ympäristön mahdollisista terveysvaikutuksista, jolloin on herännyt kysymys: aiheuttaako tietyllä alueella sijaitseva päästölähde tai muu ympäristön epäpuhtaus terveyshaittaa alueen asukkaille?

Kehitys tilastomenetelmissä sekä karttakoordinaattien linkittäminen Suomessa väestörekisterin tietoihin mahdollistavat kuitenkin huomattavan tarkat alueittaiset analyysit. Vaikka alueittaisessa analyysissä havaittaisiinkin sairastuvuusylimäärä, on johtopäätösten oltava varovaisia. Havaitut tapausmäärät pienalueella ovat yleensä pieniä, jolloin satunnaisvaihtelu peittää mahdollisen alueellisen vaihtelun. Sekoittavien tekijöiden hallinta on alueittaisessa analyysissä ja puutteellisilla tiedoilla usein epätäydellistä, jolloin jokin muu tekijä kuin epäilty ympäristötekijä voi olla sairausryvästyksen syy. Tarkemmat paikalliset tutkimukset, erityisesti tarkempi altistuksen määrittäminen, ovat tarpeen spesifien hypoteesien testaamiseksi.

Tilastollisia menetelmiä on kehitetty määrittämään harvinaisten sairauksien klusteroitumisen esiintymisen tietyllä pienalueella ja ajanjaksolla. Tarkoituksena on tutkia onko sairastapauksien klusteroituminen tilastollisesti merkitsevää vai onko kyse vain populaatiossa esiintyvästä satunnaisesta vaihtelusta. Testit jaetaan kahteen luokkaan: yleisiin ja spesifeihin testeihin. Yleiset testit keskittyvät nimensä mukaisesti sairauden yleiseen jakautumiseen isossa populaatiossa, kun taas spesifit testit keskittyvät yhteen tai useampaan pienempään joukkoon, jotka on valittu jollain tietyllä perusteella esimerkiksi suhtessa oletettuun haittalähteeseen. Testien tarkoituksena on määrittää lisätutkimuksen tarve. Koska henkilötason epidemiologisten tutkimusten suorittaminen suurissa joukoissa on epätarkoituksenmukaista ja usein mahdotonta, voidaan alueellisilla tarkasteluilla määrittää ne pienalueet, joissa tarkat tutkimukset ovat tarpeen. Tällä tavoin voidaan myös luoda ja tarkentaa tutkimushypoteeseja.

Spatiaalisen pistekuvion analysointi aloitetaan yleensä piirtämällä tapahtumat kartalle. Jos tapahtumia on paljon, voidaan ne kuvata pienalueittain. Sitten voidaan tutkia aineiston satunnaisuutta määrittelemällä empiirisiä kertymäfunktioita kuvaamaan aineistoa. Tällaisia funktioita voivat olla esimerkiksi kaikkien pisteiden välisten etäisyyksien jakauma, lähimmän naapurin etäisyydet sekä neliölukumäärät. Lopuksi voidaan yrittää kuvata aineistoa mahdollisimman hyvin sopivalla mallilla.

Prosessit voidaan laajentaa kuvaamaan kahta tai useampaa eri lajia (tässä tapauksessa tapauksia ja kontrolleja), jolloin saadaan ns. kahden lajin prosessit. Toinen mahdollinen laajennus on ajanvaikutuksien huomioon ottaminen. Esimerkiksi sairauden spatiaalista jakautumista voidaan tutkia vertaamalla

tapauksien ja kontrollien jakautumista alueeseen. Jos sairaustapauksiin lisätään vielä diagnoosipäivä, saadaan huomioitua ajanvaikutukset.

Tutkimus tehdään osana Kansanterveyslaitoksen projektia, jossa kehitetään menetelmää ympäristön ihmisen terveydelle aiheuttamien riskien mallintamiselle. Menetelmän tavoitteena on valmistuttuaan olla reaaliaikaisena päätöstukijärjestelmänä, kun mietitään alueellisia toimenpiteitä sairastuvuusriskiä kohottavien ympäristötekijöiden vaikutusalueella.

Tutkielman sisältö on seuraava. Luvussa 2 esitellään tutkimusongelma, aineisto ja käsitellään tietosuojakysymyksiä. Luvuissa 3-6 esitellään pisteprosessien teoriaa siinä laajudessa, kuin se on työssä tarpeen. Erityisesti keskitytään kahden lajin prosessien toisen kertaluvun ominaisuuteen. Lisäksi pohditaan spatiaalista tapaus-kontrolli -asetelmaa ja siihen liittyvää teoriaa. Luvuissa 4 ja 6 esitellään sairauksien klusteroitumisen tutkimiseen liittyvä tilastollinen teoria. Luvussa 7 kuvataan laskennassa käytettävä S-plus-ohjelmisto, simuloidaan edellä esitetyn teorian mukaisesti klusteroitunut riskipopulaatio, josta arvotaan satunnaisesti sekä tapaukset että kontrollit. Luvussa 8 analysoidaan simuloituaineisto ja tehdään johtopäätökset tuloksien perusteella. Viimeisessä luvussa 9 käsitellään sovellusaineistona olevaa keuhkosyöpäaineistoa ja tulkitaan tulokset luvun 8 tyyliin.

Teoriaosuus, joka käsittää luvut 3-6, perustuu, jos ei toisin mainita, P.J. Digglen kirjaan "Statistical Analysis of Spatial Point Patterns" (1983) sekä hänen artikkeliinsa "Point Prosess Modelling in Environmental Epidemiology" (1993), joka on julkaistu kokoomateoksessa "Statistics for the Environment", toim. V. Barnett ja F. Turkman, s.89-110. Wiley. Lisäksi lähteenä ja pohjatietojen hankinnassa on käytetty N. Cressie, "Statistics for Spatial Data" (1991) New York: Wiley sekä B.D. Ripley "Spatial Statistics" (1981) New York: Wiley.

2 Tutkimusaineiston ja -ongelmien esittely

2.1 Tutkimusongelma

Tutkimuksen sisällöllinen ongelma on selvittää keuhkosyöpään sairastuneiden henkilöiden lukumäärän alueellinen vaihtelu suhteessa riskipopulaation vaihteluun. Tarkoituksena on löytää vastaus kysymykseen: onko sairastuneita henkilöitä jollain osa-alueella enemmän, kuin riskipopulaatio antaisi olettaa? Jos vastaus edelliseen kysymykseen on kyllä, niin seuraava kysymys, johon yritetään saada vastausta on: onko alueella syöpäriskiä kohottavia ympäristötekijöitä vai onko sairastuvuusylimäärä peräisin pelkästä satunnaisvaihtelusta?

Tapausten ja kontrollien vaihtelua kuvataan intensiteetin ja erilaisten etäisyyskertymäfunktioiden avulla. Vaihtelua mallinnetaan stokastisen prosessin avulla.

2.2 Tietosuoja

Tietosuojakysymykset ovat vakavasti otettava ongelma. Periaatteena on, että yhden henkilön identifioiminen raportista on mahdotonta. Tietosuojakysymykset on otettu tutkimuksessa huomioon siten, että aineisto on ollut tutkimuksen aikana salattuna ja vain tutkijan saatavilla. Aineistosta lasketaan vain tunnuslukuja, joista asuinpaikka ei paljastu. Raportissa käytetään suhteellisia koordinaatteja sekä isoja merkkejä kuvissa häivyttämään yksilön tarkka asuinpaikka. Ennen raportin julkaisemista Kansanterveyslaitoksen eettinen toimikunta tarkistaa tietosuojakysymysten huomioonottamisen. Tutkimuksen päätyttyä kaikki aineistoon liittyvä tuhotaan.

2.3 Tutkimusalue ja syöpätapaukset

Empiirinen aineisto on kerätty 50×50 kilometriä olevalta alueelta Pohjois-Savosta. Tapauksiksi on valittu vuosina 1981-92 keuhkosyöpään sairastuneet miehet (122 kpl). Naisia ei aineistoon ole otettu, koska heitä on vain 15 kpl. Kontrolleiksi on tilattu Väestörekisterikeskukselta 400 samalla alueella vuonna 1980 asunutta miestä. Ikävakiointi on otettu huomioon kontrolleja tilattaessa siten, että molempien ryhmien ikäjakaumat ovat samanlaiset. Miesten approksimatiivinen ikäjakauma on esitetty taulukossa 1.

Taulukko 1: Sairastuneiden miesten jakauma

ikä v. 80	%
35 - 54	30
55 - 59	20
60 - 64	20
65 - 80	30
yht.	100

2.4 Aineiston kuvaus

Taulukossa 2 ja 3 on esitetty aineiston suhteelliset "koordinaattijakaumat". Koordinaatit on skaalattu alkuperäisestä s.e. kuvan vasemman alakulman koordinaatit on (0,0) tietosuojan vuoksi. Taulukkoja vertaamalla huomataan, että aineistojen marginaalijakaumat ovat hyvin samanlaiset. Taulukon 2 aineisto on hieman oikealle vino pituuskoordinaattien suhteen: mediaani=20, keskiarvo=26 (keskiarvon ja mediaanin ero on siis 6 km). Havaintojen lisääntyessä 122:sta 400:aan aineisto normalisoituu eli keskiarvon ja mediaanin ero on enää 1 km (x:n) ja 3 km (y:n suhteen).

Taulukko 2: Sairastuneiden miesten koordinaattitiedot

km	x	y
Min.	0	0
1st Qu.	19	17
Median	20	34
Mean	26	34
3rd Qu.	34	43
Max.	50	50

Kuvassa 1 on esitetty aineiston hajontakuvio, jonka sivun pituus 50 km. Siitä on jätetty pois koordinaatit henkilöiden identifioimisen estämiseksi. Kuvan perusteella on mahdotonta löytää vastausta tämän luvun alussa esitettyihin kysymyksiin, joten aineiston tarkempi analysointi on tarpeen. Aineiston tilastollinen analysointi suoritetaan luvussa 9.

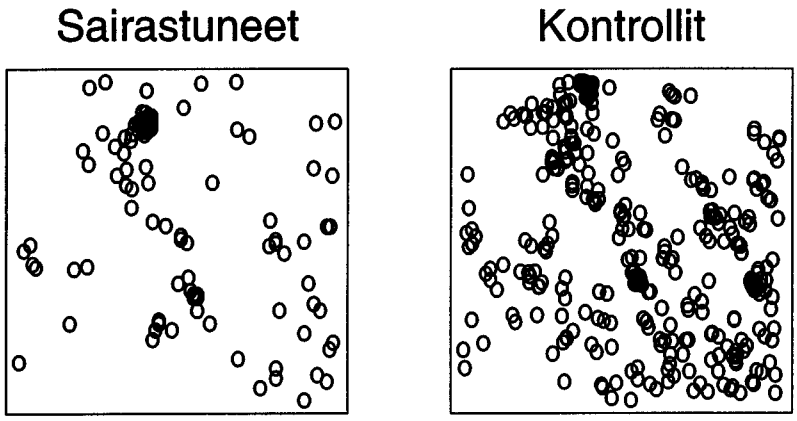
Taulukko 3: Kontrollien koordinaattitiedot

km	x	y
Min.	0	0
1st Qu.	19	15
Median	28	19
Mean	29	22
3rd Qu.	42	34
Max.	51	43

2.5 Aineiston tilastollinen käsittely

Spatiaalisessa tapaus-verrokkitutkimuksessa pyritään siis kontrolloimaan riskipopulaation heterogeenisuutta suhteessa tapauksiin. Keuhkosyöpätapaukset ja kontrollit tulkitaan kahden lajin pisteprosessiksi. Tässä tutkimuksessa ensimmäisen lajin pisteet tulkitaan keuhkosyöpätapauksiksi ja toisen lajin kontrolleiksi (verrokiksi). Tapaukset ja kontrollit oletetaan satunnaisotokseksi riskipopulaatiosta.

Sovellusaineistona oleva kehkosyöpäaineisto analysoidaan luvussa 9. Tutkimustuloksiin liittyy paitsi tilastollisia mutta myös tulkinnallisia ongelmia, joita käsitellään tarkemmin luvun 9 lopussa.



Kuva 1: Keuhkosyöpäaineisto Pohjois-Savosta. Vasemmalla on sairastuneiden henkilöiden ja oikealla kontrollihenkilöiden sijainnit. Alueen sivujen pituudet ovat 50 km.

3 Pisteprosessilähestymistapa

3.1 Pisteprosessitausta

Spatiaaliseksi pisteprosessiksi sanotaan kokoelmaa satunnaiskoordinaatteja $N = \{x_n\}$ R^d:ssä tai sen osajoukossa A. Tässä tutkimuksessa käsitellään tason pisteprosesseja ($d = 2$). Pisteet voivat kuvata esimerkiksi sairaustapausten tai kontrollien maantieteellistä sijaintia. Sijaintiin voidaan yhdistää kovariaattitietoa kuten sukupuoli, pituus ja ikä, jolloin kyseessä on merkkinen pisteprosessi. Prosessin pisteitä sanotaan tapahtumiksi tai objekteiksi.

Matemaattinen määritelmä R^d:n pisteprosessille N on seuraava. Se on satunnaismuuttuja, joka saa arvoja mitallisessa avaruudessa $[E, \mathcal{E}]$, missä E on niiden R^d:n alkioiden jonojen $\varphi = \{x_1, x_2, ...\}$ joukko, jotka toteuttavat seuraavat kaksi säännöllisyysehtoa (Stoyan et. al.1987):

- φ on lokaalisti äärellinen, ts. jokaisessa rajoitetussa R^d:n osajoukossa on enintään äärellinen määrä φ:n alkioita.
- Jono $\varphi = \{x_1, x_2, ...\}$ on yksinkertainen, ts. jos $i \neq j$, niin $x_i \neq x_j$.

σ-algebra $\mathcal{E}$ määritellään suppeimpana σ-algebrana E:ssä, jolle kuvaukset $\varphi \to \varphi(B)$ ovat mitalliset. Tässä $\varphi(B)$ tarkoittaa pisteiden lukumäärää borel-joukossa B. Prosessin jakaumaa $[E, \mathcal{E}]$:ssä merkitään P:llä.

Jatkossa prosessia merkitään N:llä ($N = \{x_1, x_2, ...\}$). Se samaistetaan lukumäärämittaan, jolloin $N(B)$ on niiden x_i:den lukumäärä, jotka kuuluvat B:hen.

Pisteprosessi on *stationaarinen* eli *siirtoinvariantti*, jos sen jakauma säilyy siirrossa. Lisäksi, jos stationaarisen prosessin jakauma säilyy kierrossa, sitä sanotaan *isotrooppiseksi.*

Pisteprosessi on *täydellisesti satunnainen* (CSR), jos mielivaltaisen joukon A, jonka pinta-alaa merkitään $|A|$:lla, pisteet $\{x_i\}$ ovat riippumaton satunnaisotos tasajakaumasta A:ssa. Prosessin tiheyttä eli keskimääräistä pisteiden lukumäärää yksikköalaa kohti kuvaa parametri λ. Täydellistä satunnaisuutta vastaa *Poisson-prosessi.*

Tässä jakauman määrittely esitellään stationaariselle Poisson-prosessille. Stationaarisella Poisson-prosessilla on kolme perusominaisuutta.

1. *Pisteiden lukumäärän Poisson-jakauma* N(A). Prosessin N pisteiden lukumäärä joukossa A noudattaa Poisson-jakaumaa parametrein $\lambda\,|A|$ (λ määritellään seuraavassa kappaleessa).

2. *Riippumaton jakautuminen.* N:n pisteet $x_1, x_2, ..., x_k$ muodostaa k riippumatonta A:ssa tasajakautunutta satunnaismuuttujaa jokaiselle $k \geq 0$.

3. Jos $A \cap B = \emptyset$, niin $N(A)$ ja $N(B)$ ovat toisistaan riippumattomia.

Jos N on stationaarinen Poisson-prosessi tiheydellä λ, niin eo. ominaisuuksien perusteella voidaan määritellä pisteprosessin jakauma, kun λ tunnetaan. Ja jos $A_1, ..., A_k$ ovat erillisiä rajoitettuja joukkoja, niin $N(A_1), ..., N(A_k)$ ovat riippumattomia Poisson-jakautuneita satunnaismuuttujia parametrein $\lambda |A_1|, ..., \lambda |A_k|$. Silloin vektorin $(N(A_1), ..., N(A_k))$ yhteisjakauma on muotoa:

$$\begin{aligned} & \mathrm{P}(N(A_1) = n_1, ..., N(A_k) = n_k) \\ = \; & \frac{\lambda^{n_1+...+n_k} |A_1|^{n_1} \cdot ... \cdot |A_k|^{n_k}}{n_1! \cdot ... \cdot n_k!} \exp\left(-\lambda \sum_{i=1}^{k} |A_i|\right) \end{aligned} \tag{1}$$

Poisson-prosessi on täydellisesti satunnainen referenssiprosessi. Se on klusteri- ja säännöllisen prosessin välissä, koska se on täysin satunnainen prosessi, jossa pisteiden sijainnit ovat toisistaan riippumattomia. Sillä jos ajatellaan pisteiden välisiä vuorovaikutuksia, niin klusteriprosessissa pisteet ovat keskimääräistä lähempänä toisiaan. Vastaavasti täysin säännöllisessä prosessissa pisteet ovat keskimääräistä kauempana toisistaan eli ne hylkivät toisiaan. Poisson-prosessia käytetään muiden prosessien konstruoinnissa (ks. luku 3.6).

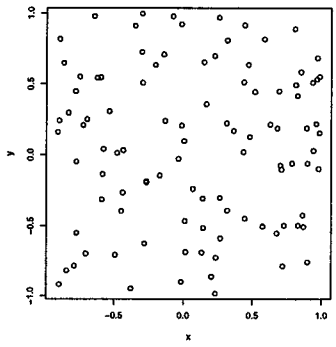
Kuvassa 2 on esitetty simuloimalla saatu homogeenisen Poisson-prosessin realisaatio. Prosessi on täysin satunnainen eikä sisällä muuta informaatiota generoivasta prosessista (vrt. kuva 3).

Kuvassa 3 on esitetty simuloimalla saatu klusteroitunut esimerkkiaineisto. Simulointiin ja simuloituun aineistoon palataan luvuissa 7-8.

3.2 Yhden lajin pisteprosessit

3.2.1 Ensimmäisen kertaluvun ominaisuudet

Kuvatkoon $\{x_i\}$ prosessin pisteitä R^2:ssa. Epidemiologiassa tapahtumiksi kutsutaan R^2:n niitä pisteitä, jotka edustavat populaation tai sairaustapahtumien alueellista jakautumista. Merkitään A:n pinta-alaa $|A|$:lla ja $N(A)$:lla tapahtumien lukumäärää joukossa A. Pisteen x sisältävää infinidesimaalisen pientä pinta-alaa merkitään dx:llä ja sen alaa $|dx|$:llä. Oletetaan vielä, että tapahtumat ovat yksinkertaisia eli $P(N(dx) > 1) = o(dx)$.



Kuva 2: Poisson-prosessin realisaatio

Ensimmäisen kertaluvun intensiteetti määritellään intesiteettifunktion avulla seuraavasti:

$$\lambda(x) = \lim_{|dx| \to 0} \left\{ \frac{\mathrm{E}\left[N(dx)\right]}{|dx|} \right\}.$$

Intensiteettifunktio $\lambda(x)$ sallii tapahtumien keskimääräisen lukumäärän vaihtelun alueen A sisällä. Koska $\mathrm{E}\left[N(A)\right] = \int_A \lambda(x)\, dx$, niin stationaarisessa tilanteessa $\lambda(x) = \lambda$ ja $\mathrm{E}\left[N(A)\right] = \lambda\, |A|$.

3.2.2 Toisen kertaluvun ominaisuus

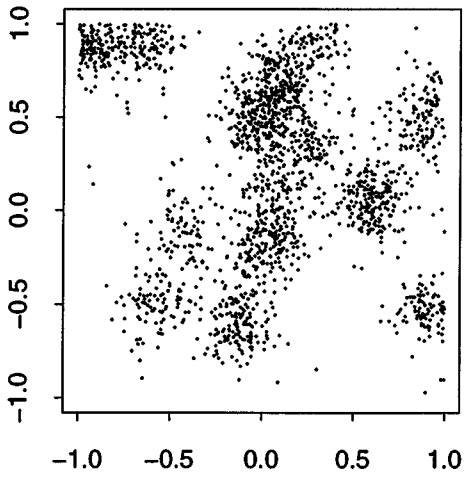
Toisen kertaluvun intensiteettifunktio $\lambda_2(x, y)$ määritellään seuraavasti:

$$\lambda_2(x, y) = \lim_{|dx||dy| \to 0} \left\{ \frac{\mathrm{E}\left[N(dx)N(dy)\right]}{|dx||dy|} \right\}.$$

Jos $N(dx)$ ja $N(dy)$ ovat korreloimattomia, niin $\lambda_2(x, y) = \lambda(x)\, \lambda(y)$. Prosessin kovarianssitiheys määritellään seuraavasti:

$$\gamma(x, y) = \lambda_2(x, y) - \lambda(x)\, \lambda(y).$$

Jos prosessi on stationaarinen, niin $\lambda_2(x, y) = \lambda_2(x - y)$ ja jos se on isotrooppinen, niin $\lambda_2(x, y) = \lambda(\|x - y\|)$, missä $\|x\| = \sqrt{x^2 + y^2}$ tarkoittaa Euklidista etäisyyttä.



Kuva 3: Klusteriprosessin realisaatio

Määritellään seuraavaksi K-funktio isotrooppisille prosesseille seuraavasti:

> $K(r) = \lambda^{-1}$ E [satunnaisesti valitusta tapahtumasta enintään etäisyydellä r olevien muiden tapahtumien lukumäärä].

K-funktio saadaan integroimalla $\lambda_2(u)$ r-säteisen ympyrän yli ja jakamalla λ^2:lla. Koska $\lambda_2(\|x\|)\,\lambda^{-1}$ vastaa ehdollista ensimmäisen kertaluvun intensiteettiä siten, että annettu x on tapahtuma alkuperäisessä prosessissa, saadaan

$$K(r) = 2\,\pi\,\lambda^{-2} \int_0^r \lambda_2(u)\,u\,du. \tag{2}$$

Vastaavasti

$$\lambda_2(r) = \lambda^2 (2\,\pi\,r)^{-1} K'(r) \tag{3}$$

edellyttäen, että K on derivoituva. Parikorrelaatiofunktio $g(r)$ määritellään seuraavasti:

$$g(r) = \frac{\lambda_2(r)}{\lambda^2} \tag{4}$$

Funktion $K(r)$ käytännön etu on, että se voidaan estimoida aineistosta yksinkertaisesti, kun taas λ_2 ja g ovat tiheyksiä ja niiden estimointiin tarvitaan

jotain tasoitusmenetelmää.

Oletetaan, että tapahtumat on generoinut Poisson-prosessi, jonka intensiteetti on λ, joten $N(A)$ noudattaa Poisson-jakaumaa parametrilla $\lambda\,|A|$. Silloin kahden erillisen joukon A ja B satunnaismuuttujat $N(A)$ ja $N(B)$ ovat riippumattomia. Edellisestä seuraa, että tapahtumien odotettu lukumäärä r-säteisessä ympyrässä on $\lambda\,\pi\,r^2$. Riippumattomuuden perusteella se on samalla enintään etäisyydellä r olevien seuraavien tapahtumien odotettu lukumäärä ympyrän keskipisteestä, jolloin $K(r) = \pi\,r^2$.

Prosessista voidaan muodostaa uusi prosessi harventamalla. Harvennusmenetelmistä yksinkertaisin on riippumaton harvennus. Siinä prosessin $N = \{x_1, x_2, ...\}$ kukin piste x_i säilyy todennäköisyydellä $p(x_i)$. Jos N on stationaarinen ja $p(x_i) \equiv p$, niin harventamalla saatu prosessi $N' = \{x_1, x_2, ...\}$ on myös stationaarinen. Sen tiheys on ilmeisestikin $p\,\lambda$ ja K-funktio pysyy muuttumattomana. Tämä seuraa siitä, että prosessien tapahtumat havaitaan todennäköisyydellä p ja jäävät havaitsematta todennäköisyydellä $1-p$, silloin havaittujen tapahtumien intensiteetti on $p\,\lambda$. Harvennetulle prosessille

$$\begin{aligned}\lambda_2^{th}(x-y) &= \lim_{|dx||dy|\to 0} \frac{p^2 \mathrm{E}\,N(dx)\,N(dy)}{|dx|\;|dy|} \\ &= p^2\,\lambda_2(x-y),\end{aligned}$$

joten

$$\begin{aligned}K_{th}(r) &= 2\pi\lambda_{th}^{-2}\int_0^r \lambda_2^{th}(u)\;u\;du \\ &= 2\pi\frac{1}{p^2\;\lambda^2}\int_0^r p^2\;\lambda_2(u)\;u\;du \\ &= 2\pi\lambda\int_0^r \lambda_2(u)\;u\;du \\ &= K(r)\end{aligned}$$

säilyy muuttumattomana satunnaisessa harvennuksessa. Tämä toisen kertaluvun ominaisuuden invarianssi riippumattomassa harvennuksessa on tärkeä spatiaalisessa tapaus-verrokkitutkimuksessa.

Koska $\mathrm{E}\,[N(A)] = \int_A \lambda(x)\,dx$, voidaan lukumäärämitalle $N(A)$ määritellä varianssi ja kovarianssi. Jos prosessi on stationaarinen, niin $\int_A \lambda\,dx = \lambda\,|A|$. Edelleen

$$\begin{aligned}\mathrm{E}[N(A)^2] &= \mathrm{E}\left[\left\{\int_A N(dx)\right\}^2\right] \\ &= \mathrm{E}\left[\int_A N(dx)\int_A N(dy)\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}\left[\int_A N(dx) + \int_A \int_A N(dx)\, N(dy)\right] \\
&= \int_A \lambda\, dx + \int_A \int_A \lambda_2(x,y)\, dx\, dy \\
&= \lambda|A| + \int_A \int_A \lambda_2(x-y)\, dx\, dy,
\end{aligned}
$$

jolloin

$$
\mathrm{Var}\{N(A)\} = \int_A \int_A \lambda_2(x-y)\, dx\, dy + \lambda\,|A|\,(1-\lambda\,|A|) \tag{5}
$$

ja

$$
\mathrm{Cov}\{N(A), N(B)\} = \int_A \int_B \lambda_2(x-y)\, dx\, dy + \lambda\,|A \cap B| - \lambda^2\,|A|\,|B|, \tag{6}
$$

missä $|A \cap B|$ kuvaa joukkojen A ja B leikkauksen alaa.

3.3 Muita prosessin karakterisointeja

Määritellään vielä kaksi K-funktiota lokaalimpaa mittaa eli kertymäfunktiot $F(x)$ ja $G(y)$, joiden avulla voidaan testata prosessin satunnaisuutta:

1. $F(x) = \mathrm{P}\{X \leq x\} = 1 - \mathrm{P}\{X > x\} = \mathrm{P}\{$etäisyys mielivaltaisesta otantapisteestä lähimpään tapahtumaan on korkeintaan $x\}$.

2. $G(y) = \mathrm{P}\{$etäisyys mielivaltaisesta tapahtumasta sen lähimpään naapuriin on korkeintaan $y\}$

$\mathrm{P}\{X > x\}$ on todennäköisyys, että x-säteisen ympyrän sisällä ei ole prosessin pisteitä (tyhjätila). Funktio $F(x)$, jota sanotaan tyhjäntilan tunnusluvuksi, estimoidaan siten, että ensiksi valitaan useita otantapisteitä, sitten lasketaan niistä etäisyys lähimpään tapahtumaan ja lopuksi muodostetaan kertymäfunktio. Tunnusluvun $G(y)$ estimaattorina käytetään sen empiiristä kertymäfunktiota. Näihin tehdään yleensä reunakorjaus.

F- ja G-funktioita käytetään aineiston kuvaamiseen. Niiden avulla voidaan muodostaa käsitys aineiston pisteidenvälisistä etäisyyksistä. Niille voidaan myös simuloida luottamusvälit tulkintojen helpottamiseksi.

3.4 Kahden lajin prosessit

Oletetaan, että data koostuu kahta tyyppiä olevista tapahtumista joukossa A. Oletetaan vielä, että tapahtumia on kaiken kaikkiaan n, joista 1-tyyppiä olevat on numeroitu kuten $1, ..., n_1$ ja 2-tyyppiä $n_1 + 1, ..., n$ eli 1-lajia on n_1 ja $n_2 = n - n_1$ kpl. Lisäksi $N_j(A)$ kuvaa j-tapahtumien (j=1,2) lukumäärää A:ssa. Jatkossa oletetaan, jos ei toisin mainita, että prosessit ovat sekä stationaarisia että isotrooppisia.

3.4.1 Ensimmäisen kertaluvun ominaisuus

Kahden lajin prosesseille määritellään intensiteettifunktio samaan tapaan kuin yhden lajin prosesseille. Intensiteetti λ_j= E [tyyppiä j olevien pisteiden lukumäärä yksikköalaa kohti]:

$$\lambda_j(x) = \lim_{|dx|\to 0} \left\{ \frac{\mathrm{E}\,[N_j(dx)]}{|dx|} \right\}, \qquad j = 1, 2.$$

Ensimmäisen kertaluvun intensiteettifunktio yleistyy suoraan toisen (tai useamman) kertaluvun intensiteettifunktioksi.

3.4.2 Toisen kertaluvun ominaisuus

Toisen kertaluvun intensiteettifunktio:

$$\lambda_{12}(x, y) = \lim_{|dx||dy|\to 0} \left\{ \frac{\mathrm{E}\,[N_1(dx)N_2(dy)]}{|dx||dy|} \right\}.$$

Siirtoinvariantissa tilanteessa $\lambda_j(x) = \lambda_j, j = 1, 2$ ja $\lambda_{12}(x, y) = \lambda_{12}(\|x - y\|)$. K-funktio määritellään vastaavasti: $K_{ij}(r) = \lambda_i^{-1}$ E [enintään etäisyydellä r olevien j-tyypin pisteiden lukumäärä satunnaisesti valitusta i-tyypin pisteestä] eli

$$K_{12}(r) = 2\,\pi(\lambda_1\,\lambda_2)^{-1} \int_0^r \lambda_{12}(u)\,u\,du. \tag{7}$$

Oletetaan, että jos kaksi toisistaan riippumatonta siirtoinvarianttia pisteprosessia on generoinut tyyppiä 1 ja 2 olevat tapahtumat, niin $K_{12}(r) = \pi r^2$. Jos taas oletetaan, että tyyppiä 1 ja 2 olevat tapahtumat on generoitu satunnaisotannalla stationaarisesta prosessista, niin $K_{11}(r) = K_{12}(r) = K_{22}(r)$. Ja jos $K_{12}(r) \geq \pi r^2$ ($K_{12}(r) \leq \pi r^2$), niin tyyppien välillä on positiivinen (negatiivinen) assosiaatio argumentin ilmoittamassa mittakaavassa.

K_{12}-funktio voidaan skaalata, jolloin saadaan L_{12}-funktio:

$$L_{12}(r) = \sqrt{\frac{K_{12}(r)}{\pi}}, \; r > 0. \tag{8}$$

Eo. skaalaus stabiloi estimaattorin $\hat{K}_{12}(r)$:n varianssin, kun prosessit ovat riippumattomia Poisson-prosesseja. Lisäksi kahdelle riippumattomalle isotrooppiselle prosessille $L_{12}(r) = r, \; r > 0$.

Ristikorrelaatiofunktio $g_{12}(r)$ määritellään seuraavasti:

$$g_{12}(r) \;=\; \frac{\frac{d}{dr}K_{12}(r)}{2\,\pi\,r}, \; r > 0 \tag{9}$$

tai

$$K_{12}(r) = \int_0^r 2\pi r\, g_{12}(u)\, du,\ r > 0. \tag{10}$$

Niin L_{12}- kuin ristikorrelaatiofunktiota voidaan käyttää data-analyyttisiin tarkasteluihin.

3.5 Ensimmäisen- ja toisen kertaluvun ominaisuuksien estimointi siirtoinvariantille pisteprosessille

Intensiteetin λ_j luonnolliset estimaattorit ovat

$$\hat{\lambda}_j = \frac{n_j}{|A|},\ j = 1, 2 \tag{11}$$

Vastaavasti saadaan estimaattori

$$\widehat{\lambda_1 K}_{11}(r) = n_1^{-1} \sum_{i=1}^{n_1} \sum_{\substack{j=1 \\ j \neq i}}^{n_1} I_{[0,r]}(d_{ij}),$$

missä

$$I_A(x) = \begin{cases} 1, & \text{jos } x \in A \\ 0, & \text{muuten} \end{cases}$$

ja $d_{ij} = \|x_i - x_j\|$. Lauseen 11 estimaattori on harhainen reunavaikutusten vuoksi, jotka voivat olla huomattavia. Koska tapahtumia A:n ulkopuolella ei havaita, tarvitaan reunakorjausta. Eräs menetelmä on painottaa tapahtumat siten, että kun $i \neq j$, niin ω_{ij}^{-1} on alueen A sisään jäävä suhteellinen osuus d_{ij}-säteisen ympyrän ympärysmitasta, jonka keskipiste on x_i, kun $i = j$, niin $\omega_{ij} = 0$. Silloin

$$\widehat{\lambda_1 K}_{11}(r) = n_1^{-1} \sum_{i=1}^{n_1} \sum_{\substack{j=1 \\ j \neq i}}^{n_1} \omega_{ij} I_{[0,r]}(d_{ij}).$$

Kun $\tilde{\lambda}_1 = (n_1 - 1)/|A|$, niin silloin K_{11}, K_{22} ja K_{12} estimaattorit ovat muotoa:

$$\hat{K}_{11}(r) = |A|\{n_1(n_1 - 1)\}^{-1} \sum_{i=1}^{n_1} \sum_{\substack{j=1 \\ j \neq i}}^{n_1} \omega_{ij} I_{[0,r]}(d_{ij}), \tag{12}$$

$$\hat{K}_{22}(r) = |A|\{n_2(n_2 - 1)\}^{-1} \sum_{i=n_1+1}^{n} \sum_{\substack{j=n_1+1 \\ j \neq i}}^{n} \omega_{ij} I_{[0,r]}(d_{ij}) \tag{13}$$

sekä

$$\hat{K}_{12}(r) = \{n_1 \tilde{K}_{12}(r) + n_2 \tilde{K}_{21}(r)\}/n \tag{14}$$

missä

$$\tilde{K}_{12}(r) = |A|\{(n_1 - 1)(n_2 - 1)\}^{-1} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \omega_{ij} I_{[0,r]}(d_{ij})$$

ja

$$\tilde{K}_{21}(r) = |A|\{(n_1 - 1)(n_2 - 1)\}^{-1} \sum_{j=n_1+1}^{n} \sum_{i=1}^{n_1} \omega_{ij} I_{[0,r]}(d_{ij}).$$

Konvekseille joukoille A saadaan harhattomat estimaattorit vain, jos $r < \frac{h}{2}$, missä h on joukon A halkaisija.

Vaihtoehtoisen reunakorjausmenetelmän on esitellyt Stoyan (1987 s.124). Jos A on ympyrä, jonka halkaisia on h ja tapahtumien välinen Euklidinen etäisyys on d_{ij}, niin reunakorjauspaino on:

$$s(r) = \begin{cases} \frac{h^2}{2}\left[\arccos\left(\frac{r}{h}\right) - \frac{r}{h}\left(1 - \left(\frac{r}{h}\right)^2\right)^{\frac{1}{2}}\right] & , \text{jos } 0 < r \leq h \\ 0 & , \text{jos } r > h. \end{cases} \tag{15}$$

Jos taas A on suorakulmio, jonka sivujen pituudet ovat a ja b, $a<b$, niin

$$s(r) = a\,b - \frac{r\,(2\,a\,+\,2\,b\,-\,r)}{\pi}, \quad 0 < r < a. \tag{16}$$

Silloin

$$\hat{K}(r) = \hat{K}_{ij}(r) = \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n} \frac{I_{[0,r]}(d_{ij})}{\hat{\lambda}^2\, s\,(\|d_{ij}\|)},$$

kun $r > 0$ on pieni.

3.6 Eräitä pisteprosesseja

3.6.1 Poisson-prosessi

Luvussa 3.1 määritelty homogeeninen Poisson-prosessi kuvaa yksinkertaisinta mahdollista, täysin satunnaista, stokastista rakennetta. Sen intensiteetti on

$$\lambda = \lim_{|dx| \to 0} \frac{E\,N(dx)}{|dx|}.$$

Riippumattomuusominaisuuden perusteella voidaan päätellä, että

$$\lambda_2(r) = \lambda^2. \tag{17}$$

Tästä seuraa edelleen, että

$$K(r) = \pi r^2, \tag{18}$$

$$EN(A) = \mathrm{Var}\{N(A)\} = \lambda |A|. \tag{19}$$

Edelleen, koska satunnainen objekti ja satunnainen otantapiste voidaan samaistaa, niin

$$\begin{aligned} F(x) = G(x) &= P\{X \leq x\} \\ &= P\{N(b(0,x)) > 0\} \\ &= 1 - \exp\{-\pi \lambda x^2\}, \end{aligned} \tag{20}$$

kun $x > 0$ ja $b(0,x)$ on origokeskeinen ympyrä ja x sen säde. Ympyrän keskipiste voidaan siirtää origoon, koska kyseessä on stationaarinen prosessi.

3.6.2 Epähomogeeninen Poisson-prosessi

Jos Poisson-prosessin vakio tiheys λ korvataan intensiteettifunktiolla $\lambda(x)$, saadaan epästationaarinen pisteprosessi. Kyseessä on *epähomogeeninen Poisson-prosessi*, jos seuraavat kolme postulaattia ovat voimassa.

1. jos $\lambda > 0$ ja A rajoitettu joukko, niin $N(A) \sim$ Poisson $(\int_A \lambda(x)\, dx)$
2. ehdolla, että $N(A) = n$, niin nämä n tapahtumaa on jakautunut $\lambda(x)$:n mukaan joukkoon A.
3. $N(A) \perp N(B)$, kun $A \cap B = \emptyset$.

Epähomogeenisen pisteprosessin tiheys määriteltiin aiemmin raja-arvona:

$$\lambda(x) = \lim_{|dx| \to 0} \frac{E\, N(dx)}{|dx|}.$$

Kovariaatit, kuten väkiluku voidaan ottaa mukaan asettamalla $\lambda(x) = f(Z_1(x), \ldots, Z_p(x))$, missä f on linkkifunktio ja β_j:t ovat z_j:tä vastaavat kertoimet. Tyypillisesti

$$\lambda(x) = \exp\{\sum_{j=1}^{p} \beta_j Z_j(x)\}.$$

3.6.3 Klusteriprosessi

Monissa tilanteissa ollaan kiinnostuneita termeistä: leviäminen ja levinneisyys. Esimerkiksi kasvit saavat alkunsa, kun "äitikasvi" levittää ympärilleen siemeniä, jolloin kehittyy heterogeenisia populaatioita. Stokastista heterogeenisuutta kuvaamaan on kehitetty kahdesti stokastisia prosesseja, joissa ensin simuloidaan äitipisteet ja sitten simuloidaan äitipisteiden ympärille lapsipisteitä. Eräs tällainen prosessi on *Poisson-klusteriprosessi.*

Algoritmi, jolla Poisson-klusteriprosessi simuloidaan, on seuraava:

1. Äitipisteet ovat Poisson jakautuneet intensiteetillä ρ.
2. Jokaisen äitipisteen klusteriin liitetään riippumattomasti satunnainen määrä S lapsipisteitä todennäköisyydellä $\{p_s, s = 0, 1, ...\}$.
3. Lapsipisteiden sijainnit suhteessa äitipisteeseen ovat riippumattomasti jakautuneet kaksiulotteisen tiheysfunktion $h(\cdot)$ mukaan.

Yleensä, jos ei toisin mainita, lopullisessa kuviossa on ainoastaan lapsipisteitä. Poisson-klusteriprosessi on stationaarinen intensiteetillä $\lambda = \rho E[S] = \rho\,\mu$. Se on isotrooppinen, jos $h(\cdot)$ on isotrooppinen. Toisen kertaluvun ominaisuuksien kuvaamista varten määritellään kohtien 1-3 perusteella konvoluutiojakauma,

$$h_2(z) = \int h(x)\, h(x - z) dx,$$

saman äitipisteen kahden lapsipisteen vektorierotuksen jakauma. Sellaisten lapsipisteiden lukumäärän jakauman toinen kumulantti on $E\,[S(S-1)]$, jolloin

$$\lambda_2(x - y) = \lambda^2 + \rho\, E\,[S(S-1)]\, h_2(x - y), \tag{21}$$

missä ensimmäinen termi kuvaa kahta eri äitipisteestä olevaa lapsipistettä ja toisessa molemmat lapsipisteet ovat samasta äitipisteestä. Kun oletetaan, että $h(\cdot)$ ja $h_2(\cdot)$ ovat isotrooppisia, kaavojen (2) ja (21) perusteella saadaan

$$K(r) = \pi\, r^2 + \frac{E\,[S(S-1)]\, H_2(r)}{\rho\, \mu^2}, \tag{22}$$

missä $H_2(r)$ on $h_2(\cdot)$:n kertymäfunkito. Tässä oikean puolen termi πr^2 vastaa Poisson-prosessin K-funktiota, ja toinen termi on ei-negatiivinen. Siten $K(r) \geq \pi r^2$, $r > 0$, mikä tarkoittaa klusteroitumista kaikissa mittakaavoissa.

Kaavan (21) perusteella saadaan varianssi

$$\mathrm{Var}\{N(A)\} = \rho\,\mu|A| + \rho\, E\,[S(S-1)] \int_A \int_A h_2(x - y)\, dx\, dy. \tag{23}$$

Jos isotrooppisessa tilanteessa $q(x,y)$ kuvaa todennäköisyyttä, että y-klusterin mielivaltaisesta pisteestä ei ole etäisyydellä x lapsipistettä, niin pisteen lähimmän naapurin etäisyyden kertymäfunktio on muotoa

$$F(x) = 1 - \exp\left\{-2\pi\rho \int_0^\infty \{q(x,y) - 1\}\, y\, dy\right\}. \tag{24}$$

Koska äitipisteiden sijainnit ovat riippumattomia, seuraa lähimmän naapurin etäisyyden tiheysfunktiosta

$$G(y) = 1 - \{1 - f(y)\}q^*(y), \tag{25}$$

missä $q^*(y)$ kuvaa todennäköisyyttä, että saman äitipisteen kaksi lapsipistettä ovat vähintään etäisyydellä y toisistaan.

3.6.4 Cox-prosessi

Kahdesti stokastinen Poisson-prosessi eli *Cox-prosessi* määritellään satunnaisen intensiteetin ja epähomogeenisen Poisson-prosessin avulla seuraavasti:

1. $\{\Lambda(x) : x \in R^2\}$ on ei-negatiivinen stokastinen prosessi (satunnaiskenttä).

2. Prosessi ehdolla $\Lambda(x) = \lambda(x)$ on epähomogeeninen Poisson-prosessi intensiteetillä $\lambda(x)$.

Näin konstruoitu pisteprosessi on stationaarinen (isotrooppinen), jos ja vain jos $\{\Lambda(x)\}$ on stationaarinen (isotrooppinen). Ensimmäisen- ja toisen kertaluvun tiheysfunktiot ovat

$$\lambda = E\left[\Lambda(x)\right]$$

ja

$$\lambda_2(x,y) = E\left[\Lambda(x)\Lambda(y)\right].$$

Kun $\{\Lambda(x)\}$ on isotrooppinen edellinen voidaan kirjoittaa muodossa

$$\lambda_2(r) = \lambda^2 + \gamma(r), \tag{26}$$

missä

$$\gamma(r) = \mathrm{Cov}\{\Lambda(x), \Lambda(y)\}$$

ja r on x:n ja y:n välinen etäisyys.

Coxin prosessin ja klusteriprosessin välillä on yhteys. Itse asiassa klusteriprosessi voidaan tulkita Coxin prosessiksi määritelemällä intensiteettiprosessi seuraavalla tavalla:

$$\Lambda(x) = \mu \sum_{i=1}^{\infty} h(x - X_i), \tag{27}$$

missä $\mu > 0$ ja X_i on Poisson-prosessin pisteet.

Yleiselle Cox-prosessille on vaikea muodostaa tarkkaa lähimmän naapurin etäisyyden jakaumaa. Ehdollistamalla tiheydellä $\{\lambda(x)\}$ todennäköisyys, että vähintään etäisyydellä r ei ole muita tapahtumia on:

$$\exp\left(-\int \lambda(x)\,dx\right), \tag{28}$$

missä integroitava joukko on ympyrä, jonka säde on r.

4 Klusteroitumisen testaus

4.1 Klusteroitumisen hypoteesit

Klusteroitumisen hypoteeseja on kahta tyyppiä. Ensimmäisenä hypoteesina on populaation satunnaisuus (luku 4.1.1). Toisena hypoteesina on tapausten satunnainen jakautuminen mahdollisesti heterogeeniseen populaatioon (luku 4.1.2).

4.1.1 Populaation klusteroituminen

Perusoletuksena populaation klusteroitumista testattaessa on, että populaation on generoinut homogeeninen Poisson-prosessi vakiotiheydellä λ. Toisin sanoen nollahypoteesina on

$$h_0 = \text{populaatio on täysin satunnaisesti jakautunut alueeseen } A.$$

Populaation klusteroitumista testataan siis täydellistä satunnaisuutta vastaan. Jos havaitaan, että F- ja G-funktioiden estimaatien kuvaajat ovat likimain samanlaiset, niin aineisto on satunnainen. Jos taas estimaatien kuvaajista löytyy eroa, on aineisto klusteroitunut (ks. esim. kuva 14). K-funktion estimaatti on simuloitujen luottamusvälien välissä annetulla etäisyydellä r, jos aineisto on täydellisesti satunnainen ja ylläpuolella, jos aineisto on klusteroitunut etäisyyden r määräämässä mittakaavassa. Luottamusvälit on simuloitu ehdolla, että nollahypoteesi on voimassa ja ne ovat ns. pisteittäisiä luottamusvälejä.

Heterogeenisen aineiston sovittamiseen käytettäviä malleja ovat esimerkiksi epähomogeeninen Poisson-, klusteri- ja Cox-prosessi, jotka esitellään luvussa 3.6.

4.1.2 Tapausten klusteroituminen populaatioon

Koska yleensä populaatioiden tiheys epidemiologisessa aineistossa vaihtelee luonnostaan merkittävästi, on vaihteleva tiheys $\lambda(x)$ yleensä parempi lähtökohta, kuin CSR -oletus. Silloin nollahypoteesina h_0 on että,

$$h_0 = \text{tapaukset eivät ole klusteroituneet populaatiota enemmän.}$$

Toisin sanoen tapaukset saadaan joko

- riippumattomalla harvennuksella tai
- riippumattomalla merkkauksella populaatiosta.

Molemmissa tapauksissa $K_{11} \equiv K_{22}$, joten tapausten klusteroitumishypoteesi on $D = K_{11} - K_{22} > 0$ (D määritellään luvussa 5). Vastaavasti, jos $D \leq 0$, niin tapaukset eivät ole klusteroituneet populaatiota enemmän. Tapausten klusteroitumisen testaus käsitellään luvussa 5.

4.2 Monte Carlo -testi

Jopa yksinkertaisimmatkin spatiaalisen pistekuvion stokastiset mallit johtavat hankaliin jakaumateorioihin. Mallin riittävyyden testaamiseen käytetään *Monte Carlo*-testiä, missä havainto ajatellaan yhdeksi (esim. ensimmäiseksi) simulointikokeen tulokseksi. Tämä idea on peräisin Barnardilta, ks. Diggle (1983) s. 7.

Monte Carlo-testin käytön edellytyksenä on hypoteesi, jota voidaan simuloida ja tunnusluku, joka voidaan laskea sekä aineistosta sekä simuloinneista.

Monte Carlo-testi suoritetaan siten, että merkitään u_1:llä tunnusluvun U havaittua arvoa. Olkoon $u_i, i = 2, ..., s$ vastaavat simuloidut arvot, kun h_0 on voimassa. Oletetaan edelleen, että u_i:t ovat erisuuria, jolloin u_1:n sija on yksiselitteinen. Kuvatkoon $u_{(j)}$ j:nneksi suurinta arvoa u_i:ssä. Silloin h_0:n voimassa ollessa

$$P\{u_1 = u_{(j)}\} = s^{-1}, \quad j = 1, ..., s.$$

Nollahypoteesi hylätään u_1:n sijainnin perusteella suhteessa u_i:hin. Jos u_1 on k:nneksi suurin tai suurempi, saadaan tarkka yksisuuntainen testi p-arvonaan k/s.

4.3 Populaation klusteroitumisen testaus

Riskipopulaation klusteroitumisen tutkimiseksi on useita eri vaihtoehtoja, joita käytetään yhdessä. Ensimmäiseksi deskriptiivisenä menetelmänä piirretään tapahtumat kartalle, jotta saadaan kuva aineiston jakautumisesta alueella. Toiseksi voidaan tutkia tapahtumaparien välisiä etäisyyksiä (ks. luku 4.4.1) mikä johtaa K- tai L-funktion käyttöön. Kolmas vaihtoehto on ottaa käyttöön lähimmän naapurin menetelmät, jotka tarjoavat tehokkaan keinon pienen mittakaavan tapahtumien välisten vuorovaikutusten löytämiseksi. Lähimmän naapurin tunnusluvut, jotka vertaavat tutkittavaa aineistoa täysin satunnaiseen aineistoon, esitellään luvuissa 4.4.2-4.4.3.

4.4 Etäisyyksien jakaumista

4.4.1 Tapahtumaparien väliset etäisyydet

Eräs mahdollisuus kuvata tilajärjestystä A on laskea (satunnaisesti) valittujen pisteparien jakauma. Tällaisia pistepareja on $\frac{1}{2}n(n-1)$, kun A:ssa on n objektia. Jakauma riippuu paitsi tilajärjestyksestä myös alueen muodosta ja koosta. On huomattava, että K-funktio perustuu tapahtumaparien väliseen jakaumaan, sillä $\lambda^2 K(r)$ on niiden pisteparien odotettu lukumäärä, joissa ensimmäinen piste on on yksikköneliössä ja toinen enintään etäisyydellä r siitä.

Esitetään seuraavaksi pisteparietäisyyden kertymäfunktio, kun A on yksikköneliö ja yksikköympyrä (Diggle 1983). Lineaarisella muunnoksella näihin voidaan tehdä skaalan muunnos:

$$H(r) = \begin{cases} \pi r^2 - \frac{8r^3}{3} + \frac{r^4}{2} & ,0 \leq r \leq 1 \\ \\ \frac{1}{3} - 2r^2 - \frac{r^4}{2} + 4\frac{2r^2+1}{3}\sqrt{r^2-1} + 2\, r^2 \sin^{-1}(2\, r^{-2} - 1) & ,1 < r \leq \sqrt{2}. \end{cases}$$

Jos kyseessä on yksikköympyrä, niin esitys on muotoa

$$\begin{aligned} H(r) &= 1 + \pi^{-1}\left\{2\,(r^2-1)\cos^{-1}\frac{r}{2}\right\} \\ &\quad - \pi^{-1}\left\{r\,(1+\frac{r^2}{2})\sqrt{\frac{1-r^2}{4}}\right\},\ 0 \leq \mathrm{r} \leq 2. \end{aligned}$$

Tehdään testi CSR:lle. Jos $H(r)$ tunnetaan, saadaan tapahtumaparien väliselle etäisyydelle jakauma. Silloin

$$\hat{H}_1(r) = \{\frac{1}{2}\,n\,(n-1)\}^{-1} \#(r_{ij} \leq r),$$

missä $\hat{H}_1(r)$ kuvaa havaittua osaa tapahtumaparien välisistä etäisyyksistä, jotka ovat korkeintaan etäisyydellä r ja # lukumäärää. Jos CSR-hypoteesi on voimassa, se kuvautuu lineaariseksi, kun $\hat{H}_1(r)$ piirretään $H(r)$:ää vastaan. Kun lasketaan $\hat{H}_i(r), i = 1, ..., s$ (s = simulointien lukumäärä) ja oletetaan, että simuloinnit ovat riippumattomia ja samoin jakautuneita A:ssa, saadaan vaihteluväli $\hat{H}_1(r)$:lle helpottamaan kuvioiden tulkintaa. Vaihteluvälit ovat:

$$U(r) = \max_{i=1,..,s}\{\hat{H}_i(r)\}$$

$$L(r) = \min_{i=1,..,s}\{\hat{H}_i(r)\}$$

Jos vaihteluvälit piirretään $H(r)$:ää vastaan, saadaan CSR:n voimassa ollessa

$$P\{\hat{H}_1(r) > U(r)\} = P\{\hat{H}_1(r) < L(r)\} = s^{-1}.$$

Luottamusvälien on tarkoitus kontrolloida satunnaisvaihtelua ja siten helpottaa kuvaajan tulkintaa, kun $\hat{H}(r)$ piirretään $H(r)$:ää vastaan. Monte Carlo-testi voidaan konstruoida kahdella tavalla:

- Valitaan mittakaava r_0 ja määritellään $U_i = \hat{H}(r_0)$. Silloin havaitun u_1:n sija simuloitujen u_i:den joukossa antaa p-arvon (ks. luku 4.2).
- Määritellään r_1 ja tunnusluku u_i siten, että se on funktioiden $\hat{H}(r)$ ja $H(r)$ estimaattien kuvaajien ero eri mittakaavoilla $r \in (0, r_1)$. Etäisyysmittana voidaan käyttää esimerkiksi normia

$$U_i = \int_0^{r_1} \{\hat{H}(r) - H(r)\}^2 \, dr. \tag{29}$$

Erotuksen tilastollista merkitsevyyttä voidaan testata Monte Carlo-testillä.

Ensimmäinen testillä on tulkinta vain, jos r_0 voidaan valita järkevästi. Toinen testi on huomattavasti objektiivisempi, koska se toimii simultaanisesti mittakaavoilla $(0, r_1)$.

Jos $H(r)$:n jakaumaa ei tunneta, se voidaan korvata funktiolla

$$\bar{H}_i(r) = (s-1)^{-1} \sum_{j \neq i} \hat{H}_j(r).$$

Tunnusluvut u_i eivät enää ole riippumattomia. Monte Carlo-testin oletus, että nollahypoteesin voimassa ollessa kaikki u_i:den saamat arvot ovat yhtä todennäköisiä, on kuitenkin voimassa. Huomaa, että $\bar{H}_1(r)$ on $H(r)$:n harhaton estimaattori.

4.4.2 Lähimmän naapurin etäisyydet

Oletetaan, että joukossa A on n tapahtumaa ja y_i kuvaa etäisyyttä tapahtumasta i sen lähimpään naapuriin. Silloin y_i:tä sanotaan *lähimmän naapurin etäisyydeksi.* Se sisältää pisteparien väliset, molemmin puoliset, etäisyydet. Määritellään empiirinen tiheysfunktio:

$$\hat{G}_1(y) = n^{-1} \#(y_i \leq y).$$

Kun CSR on voimassa, riippuu Y:n teoreettinen jakauma n:stä ja A:sta. Sitä ei voida esittää suljetussa muodossa monimutkaisten reunavaikutusten takia. Jos reunavaikutuksista ei välitetä ja A:n pinta-ala on $|A|$, silloin $\pi y^2 |A|^{-1}$

on h_0:n voimassa ollessa todennäköisyys, että satunnaisesti valittu piste on etäisyydellä y valitusta pisteestä. Silloin Y:n tiheysfunktio on likimain muotoa

$$G(y) = 1 - (1 - \pi y^2 |A|^{-1})^{n-1}.$$

Suurilla n voidaan merkitä $\lambda = \frac{n}{|A|}$, jolloin

$$G(y) = 1 - \exp\{-\lambda \pi y^2\}, \tag{30}$$

kun $y \geq 0$.

Empiiriselle kertymäfunktiolle $\hat{G}_1(y)$ saadaan Monte Carlo-testillä luottamusvälit simuloiduista funkioista $\hat{G}_i(y)$, i=2,3,...,s. Testi voidaan muodostaa esimerkiksi seuraavasti. Perustetaan se aiemmin esitetyllä tavalla normiin

$$U_i = \int \{\hat{G}_i(y) - \bar{G}_i(y)\}^2 \, dy, \text{ missä}$$

$$\bar{G}_i(y) = (s-1)^{-1} \sum_{j \neq i} \hat{G}_i(y)$$

on määritelty samoin kuin $\bar{H}_1(r)$ luvussa 4.4.1.

Toinen vaihtoehto on perustaa testi lähinaapurietäisyyksien otoskeskiarvoon $\bar{Y}$. Asetetaan $\bar{Y}_i$ kuvaamaan i:nnestä simuloinnista laskettua keskiarvoa ($i = 1$ vastaa aineistosta laskettua). Merkitään $u_i = \bar{y}_i$ ja suoritetaan testi. Jos tapahtumaparien väliset etäisydet ovat pieniä, ei simulointia tarvita. Hyvä approksimaatio $\bar{Y}$:n jakaumalle on normaalisuusoletus parametrein

$$E(\bar{Y}) = \frac{1}{2}(n^{-1}|A|)^{1/2} + (0.051 + 0.042\, n^{-1/2})\, n^{-1}\, \ell(A) \tag{31}$$

$$Var(\bar{Y}) = 0.070\, n^{-2}|A| + 0.037(n^{-5}|A|)^{1/2}\, \ell(A), \tag{32}$$

missä $\ell(A)$ kuvaa A:n kehän pituutta. Huomattavan suuret keskiarvon $\bar{y}$ arvot tarkoittavat säännöllisyyttä eli satunnaisuutta ja pienet arvot klusteroitumista.

4.4.3 Pisteen ja lähimmän tapahtuman väliset etäisyydet

Merkitään x_i:llä lyhintä etäisyyttä satunnaisesti valituista otantapisteistä $\xi_1, ..., \xi_m$ lähimpiin tapahtumiin, joita on m kappaletta joukossa A. Silloin tyhjäntilan tunnusluku on

$$\hat{F}_1(x) = m^{-1} \# (x_i \leq x).$$

Joukon B_x, johon kuuluu kaikki A:n pisteet, jotka ovat vähintään etäisyydellä x A:n tapahtumista, pinta-alan estimaatti $|B_x|$ on $[1 - \hat{F}_1(x)]|A|$. Jos CSR on voimassa, niin

$$F(x) = 1 - \exp\{-\pi \lambda x^2\}, \tag{33}$$

missä $x \geq 0$ ja $\hat{\lambda} = n\,|A|^{-1}$.

Monte Carlo-testiä varten voidaan määritellä integraalinormiin perustuva tunnusluku lukujen 4.4.1-2 perusteella:

$$U_i = \int \{\hat{F}_i(x) - \bar{F}_i(x)\}^2\, dx, \tag{34}$$

missä $\bar{F}_i(x)$ on määritelty samoin kuin $\bar{H}(r)$ luvussa 4.4.1 ja $\bar{G}_i(r)$ luvussa 4.4.2..

4.4.4 Neliölukumäärät

Vaihtoehto etäisyysperusteiselle lähestymistavalle on jakaa joukko A m pinta-alaltaan yhtäsuuriin alijoukkoihin, esim. neliöihin. Neliöiden tapahtumien lukumääriä käytetään CSR:n testaamiseen. Oletetaan, että A on yksikköala, joka on jaettu $k \times k$:hon neliöön, jolloin $m = k^2$. Olkoon vielä n_i, $i = 1, ..., m$, neliöissä olevien tapahtumien lukumäärät ja merkitään $\bar{n} = n/m$. Silloin tunnusluku

$$X^2 = \sum_{i=1}^{m} \frac{(n_i - \bar{n})^2}{\bar{n}}, \tag{35}$$

on likimain χ^2_{m-1} jakautunut, kun $n \to \infty$. Nollahypoteesin hylkäys voi tässä yhteydessä johtua joko siitä, että tapahtumat eivät ole jakautuneet tasaisesti joukossa A tai siitä, että tapahtumat riippuvat toisistaan. Edelleen pienet arvot kuvaavat säännöllisyyttä ja suuret klusteroitumista.

5 Spatiaalinen tapaus-kontrolli -asetelma

5.1 Johdanto

Verrokkitutkimuksessa pyritään vakioimaan sekoittavia tekijöitä. Esimerkiksi ikä-ja sukupuolivakiointi tarkoittaa sitä, että tapaukselle valitaan populaatiosta verrokki, joka on samaa sukupuolta ja samaa ikäryhmää.

Sairaustapausten muodostamassa kartassa tärkein sekoittava tekijä on populaation heterogeenisuus: havaittava tapausklusteri ei välttämättä ilmaise kohonnutta alueellista riskiä vaan voi yksinkertaisesti johtua populaation klusteroitumisesta.

Tapahtumien sijainnit kuvaavat alueella A tiettynä ajanjaksona asuneiden ja sairastuneiden henkilöiden asuinpaikan koordinaatteja. Vastaavasti kontrollihenkilöt eli verrokki valitaan siten, että he ovat satunnaisotos samalta alueelta samana ajanjaksona asuneesta riskipopulaatiosta. Ikä- ja sukupuoli vakiointi tehdään kontrolleja valittaessa siten, että kontrollit ovat jakutuneet samoin iän ja sukupuolen mukaan, kuin tapaukset.

Tapauksia ja kontrolleja kuvataan kahden lajin pisteprosessilla. Ensimmäisen lajin pisteet ovat tapauksia ja toisen kontrolleja.

Menetelmällä on paljon etuja: Se on teoreettisesti perusteltu ja kontrollit on saatavissa nopeasti ja edullisesti käyttöön rekisteristä. Myöskään aineistot eivät ole aggregoituja, jolloin kaikki mittakaavat ovat simultaanisesti käytössä tai ainakin aggregointitaso on tutkijan päätettävissä.

5.2 Tapaus-kontrolli -lähestymistavan teoria

Oletetaan, että data, joka koostuu tyyppiä 1 olevista (keuhkosyöpä-) tapauksista $\{x_i \in A,\ i = 1,\ldots,n_1\}$ joukossa A. Edellisen luvun perusteella voidaan data ajatella realisaatioksi spatiaalisesta prosessista. Ympäristön heterogeenisuudesta ja/tai tapahtumien klusteroitumisesta johtuvaa tapahtumien vaihtelua ei voida sinällään erottaa ilman lisäinformaatiota. Tarvitaan tyyppiä 2 olevia (kontrolli-) tapauksia, joilla pyritään mallintamaan ympäristön heterogeenisuutta.

Merkitään riskipopulaatiosta satunnaisesti valittuja kontrollitapauksia (-henkilöitä) $\{x_i \in A,\ i = n_1 + 1,\ldots,n\}$. Diggle (1993) suosittelee, että kontrolleja valitaan 3-4-kertaa tapausten lukumäärä.

Silloin kun tapaukset eivät ole alueellisesti klusteroituneet, tapahtumat ovat riippumaton harvennus populaatiosta. Silloin $\lambda_i K_{ij}(r)$-funktio on tyyppiä j olevien tapahtumien odotettu lukumäärä etäisyydellä r satunnaisesti

valitusta tyyppiä i olevasta tapahtumasta. Näillä funktioilla K_{11}, K_{22} ja K_{12} on seuraavat tulkinnat:

1. K_{11} sisältää tietoa tapausten klusteroitumisesta, jossa on mukana populaation heterogeenisuus ja mahdollinen tapausten klusteroituminen populaatiossa.
2. K_{12} sisältää tapausten ja populaation heterogeenisuuden
3. K_{22} kuvaa ympäristön heterogeenisuudesta johtuvaa klusteroitumista, populaation heterogeenisuutta.
4. Jos tyypittely on satunnaista, niin $K_{11} = K_{22} = K_{12}$

On huomattava, että $K_{12} = K_{21}$ aina, mutta tämä ei välttämättä päde K_{12}:n estimaattorille johtuen reunakorjauksista. Mikäli tapaukset ovat klusteroituneet ympäristöefektiä enemmän, niin $K_{11}(r) > K_{22}(r)$, jollain $r > 0$.

5.3 Klusteroitumisen testaus

Klusteroitumisen testausta on käsitelty esimerkiksi artikkeleissa Diggle ja Chetwynd, 1991 "Second-order analysis of spatial clustering for inhomogeneus populations" sekä Besag ja Newell, 1991 "The detection of rare diseases".

Edellisen perusteella määritetään tunnusluku klusteroitumiselle:

$$D(r) = K_{11}(r) - K_{22}(r) \tag{36}$$

ja sille estimaattori

$$\hat{D}(r) = \hat{K}_{11}(r) - \hat{K}_{22}(r).$$

Jos siis $D(r) > 0$, on sairaustapaukset klusteroituneet ympäristöefektiä enemmän. Tunnusluvulle $\hat{D}$ voidaan laskea varianssi ehdolla, että klusteroitumista ei ole. Varianssin avulla voidaan määritellä tunnusluvulle vaihteluväli, joka helpottaa $\hat{D}$ ja r:n välisen kuvaajan tulkitsemista.

Klusteroidulle pisteprosessille on tyypillistä, että $D(r)$ on positiivinen ja että $\mathrm{Var}\{\hat{D}(r)\}$ kasvaa voimakkaasti, kun $r \to \infty$. Sen takia tilastollinen informaatio on $\hat{D}(r)$:ssä suurinta pienillä r:n arvoilla. Tämä tekee m:n valinnan kaavassa (37) vähemmän merkitseväksi, kuin mitä se yleensä on. Kuitenkin m:n arvo vaikuttaa testin tehokkuuteen suhteessa muihin vaihtoehtoihin.

Tunnusluku, jolle voidaan laskea p-arvo *Monte-Carlo*-menetelmällä, on:

$$\hat{D} = \int_0^{r_0} \left\{ \frac{\hat{D}(r)}{\sqrt{\mathrm{Var}\hat{D}(r)}} \right\} dr \approx \sum_{k=1}^{m} \left\{ \frac{\hat{D}(r_k)}{\sqrt{\mathrm{Var}\hat{D}(r_k)}} \right\}, \tag{37}$$

missä r_k $(k = 1, \ldots, m)$ on epidemiologisesti relevanttien tapausten väliset etäisyydet. Kun h_0 on voimassa, on kaavan (37) otosjakauma approksimatiivisesti normaali parametrein $(0,\ m + 2\sum_{j=2}^{m}\sum_{k=1}^{j-1} \mathrm{corr}\{\hat{D}(r_j), \hat{D}(r_k)\})$, kun $j,\ k = 1, \ldots, m$.

Kaavasta (37) saadaan kaksi testiä merkitsevyylle:

1. standardoitu normaalitesti $T = \frac{\hat{D}}{\sqrt{\hat{V}}}$, missä $\hat{V} = \widehat{\mathrm{Var}(\hat{\mathrm{D}})}$
2. Monte-Carlo-testi

Monte-Carlo-testi ei edellytä asymptotiikkaa ja soveltuu siten pienille aineistoille mutta se saattaa, tosin algoritmista ja kovarianssin määrittämisestä riippuen, olla raskas. Etuna Monte Carlo-testillä on, että se säilyttää luonnollisen tulkinnan: $\hat{\lambda}_1 \hat{D}(r)$ on seuraavien tapahtumien lukumäärän odotusarvon estimaatti etäisyydellä r tutkittavasta tapahtumasta keskiarvoistettuna yli kaikkien tapahtumien. Vastaavasti, $\hat{D}(r)/\hat{K}_{22}(r)$ on suhteellisen riskin kohoamisen estimaatti ympyrällä, jonka säde on r ja keskipisteenä satunnaisesti valittu tapahtuma.

Toinen vaihtoehto D:lle on korvata (37) varianssimatriisilla painotetulla neliömuodolla. Kun h_0 on voimassa, se on approksimatiivisesti χ^2_m-jakautunut. Neliöön korottaminen hävittää etumerkin. Tällöin menetetään yksisuuntaisen testin mahdollisuus, koska vain suuret testisuureen arvot ovat tilastollisesti merkitseviä.

Kolmas vaihtoehto on määritellä

$$D = \max_{k=1,\ldots,m} \left\{ \frac{\hat{D}(r_k)}{\sqrt{\mathrm{Var}\{\hat{D}(r_k)\}}} \right\}. \tag{38}$$

Kirjoittamalla $a_m = \{2 \log m\}^{\frac{1}{2}}$ ja

$$b_m = \{2 \log m\}^{\frac{1}{2}} - \frac{1}{2}\{2 \log m\}^{-\frac{1}{2}}[\log\{\log m\} + \log(4\pi)],$$

kun h_0 on voimassa ja $m \to \infty$, saadaan

$$Pr(D \leq d) \approx \exp[-\exp\{-a_m(d - b_m)\}]. \tag{39}$$

5.4 Menetelmän arviointia

Tapaus-verrokkitutkimukseen liittyy paljon ongelmia (Andersson & Titterington 1997). Suurin ongelma on varmasti kontrollihenkilöiden ja heidän lukumäärän valinta. Lapsia tutkittaessa ei eri muuttujilla ole suurta vaikutusta. Sen sijaan aikuisilla, jotka ovat eläneet erilaisissa olosuhteissa, on muuttujien valinta vaikeaa. Juuri heterogeenisuuden vuoksi kontrollien valinta tulee suunnitella huolella ja ottaa huomioon tulkintaa tehtäessä. Tärkeitä valintaan vaikuttavia muuttujia voivat olla esimerkiksi asuinpaikka, ikä, sukupuoli. Varsinaiset ongelmat alkavat, jos halutaan ottaa moniin asioihin vaikuttava, erittäin hankalasti määriteltävä, sosiaaliluokka (tai vast.) huomioon. Näitä tietoja ei ole saatavissa väestörekisteristä.

Klusterointi on jo sinänsä epämääräinen käsite, minkä vuoksi spatiaalisen skaalan eli mittakaavan valinta saattaa myös vaikuttaa niiden esiintymiseen. Jos valitaan pieni mittakaava, esiintyy klustereita helpommin. Vastaavasti, jos valitaan suuri mittakaava, ei klustereita esiinny. Tämä johtuu siitä, että tapahtumat, joita on yleensä vähän suhteessa alueen kokoon, esiintyvät usein varsinkin suurella mittakaavalla hajallaan alueessa A, jolloin ne "hukkuvat" kontrollien joukkoon. Vastaavasti, jos mittakaava on pieni saattaa pienikin klusteri "hypätä" esille.

Aineiston ryhmittelyssä alijoukkoihin on omat ongelmansa. Kuinka valita joukot niin, että ne kuvaavat yleisjakaumaa? Varsinkin lukumääriin perustuvat menetelmät ovat herkkiä erilaisille ryhmittelyille.

Muita ongelmia ovat esimerkiksi: tilastollisen merkitsevyyden määrittely ja johtopäätösten teko tulosten perusteella on usein hankalaa. Kausaalitulkintaan ei päästä. Lisäksi datan laatu saattaa myös aiheuttaa ongelmia, koska paikkareferenssin hyvyys voi olla ongelmallinen. Ongelmia saattaa aiheuttaa myös datan saantia ja käyttöä rajoittavat tietosuojakysymykset.

Niin kuin luvussa 5.1 sanottiin on menetelmällä paljon etuja. Menetelmä soveltuu käytettäväksi, kun epäillään alueellista sairastuvuusylimäärää ja mietitään mahdollisia jatkotoimenpiteitä tarkasteltavalla pienalueella. On kuitenkin syytä käyttää muitakin menetelmiä ja epidemiologista tietoa esimerkiksi sairauden etiologiasta yhdessä tilastollisten päätelmien kanssa ennenkuin tehdään kovin pitkälle meneviä johtopäätöksiä.

6 Sairauksien klusteroituminen kiinteän kohteen ympärille

Tämä luku esitellään Digglen artikkelin "Point prosess modelling in environmental epidemiology" (Diggle, 1993) pohjalta.

Oletetaan, että tietty tunnettu kohde (esim. teollisuusalue) on yhdistetty tietynlaiseen sairauteen joukossa A. Tavoitteena on tutkia sairastumisriskin vaihtelua altistumislähteen ympärillä sekä riskipopulaation spatiaalista rakennetta.

Koska tapausten sijainnit tunnetaan, voidaan käyttää epähomogeenista Poisson-prosessia mallintamaan taudin esiintymistä. Merkitään kiinteää kohdetta x_0:lla ja tapahtumia $x_i \in A$, $i = 1, ..., n_1$. Oletetaan, että

$$\lambda(x) = \rho\, \lambda_0(x)\, f(x - x_0; \theta). \tag{40}$$

Parametri $\lambda(x)$ on siis tapausten intensiteetti, joka muodostuu kolmesta komponentista: $\lambda_0(x)$ on riskipopulaation intensiteetti ilman altistusta, ρ on riskisuhde populaatiossa ilman altistusta ja $f(u; \theta)$ kuvaa tapahtumien riskin muutosta x_0:n ympärillä. Jos $f(x - x_0; \theta) = 1$, niin sairastumisriskissä ei havaita muutosta

Jos $\lambda_0(x)$ tunnetaan, niin ρ:n ja θ:n log-uskottavuus on muotoa

$$\begin{aligned} \ell(\rho, \theta) &= n_1 \log \rho + \sum_{i=1}^{n_1} \log\, \lambda_0(x_i) + \sum_{i=1}^{n_1} \log\, f(x_i - x_0; \theta) \\ &- \rho \int_A \lambda_0(x)\, f(x - x_0; \theta)\, dx. \end{aligned} \tag{41}$$

Jos θ tunnetaan, su-estimaattori ρ:lle on muotoa

$$\hat{\rho}(\theta) = n_1 \Big/ \int_A \lambda_0(x)\, f(x - x_0; \theta)\, dx.$$

Jos $\hat{\rho}(\theta)$:n arvo sijoitetaan kaavaan (41), saadaan

$$\ell^*(\theta) = n_1 \log \left\{ \int \lambda_0(x)\, f(x - x_0; \theta)\, dx \right\} + \sum_{i=1}^{n_1} \log\, f(x - x_0; \theta).$$

Eräs vaihtoehto on korvata $\lambda_0(x)$ kontrollipopulaatiosta $\{x_{n+1}, ..., x_n\}$ lasketulla kernel-estimaattorilla $\hat{\lambda}_0(x)$, joka on muotoa

$$\hat{\lambda}_0(x) = (n - n_1)^{-1} \sum_{i=n_1+1}^{n} G\{(x - x_i)/h\}, \tag{42}$$

missä

$$G(u) = (2\pi)^{-1} exp(-u'u/2)$$

ja $h > 0$ valitaan minimoimaan $\hat{\lambda}_0(x)$:n keskineliövirhettä.

Vaikka kernel-estimaattori (42) itsessään on kiinnostava, vaikuttaa $\lambda_0(x)$:n estimointimenetelmä θ:n arvoon. Yksi mahdollisuus on ehdollistaa tapahtumien ja kontrollien sijainnit $x_i, i = 1, \ldots, n_1 + n_2$, ja sitten arpoa tapahtumat x_i uudelleen kiinnitettyihin sijainteihin riippumattomilla Bernoulli-kokeilla. Jos $p(x_i)$ on todennäköisyys, että piste x_i on tapahtuma, niin kaavan (7 perusteella saadaan

$$p(x) = \frac{\rho f(x - x_0; \theta)}{1 + \rho f(x - x_0; \theta)}. \tag{43}$$

Binääristä regressiomallia (43) ei voida yleistää lineaariseksi, koska on perusteltua olettaa, että $f(u; \theta) \to 1$, kun $\|u\| \to \infty$ eli $\{p(x) - p/(1-p) \in [0, 1]\}$, kun $\|x - x_0\| \to \infty$. Siitä huolimatta p:n ja θ:n uskottavuus on johdettavissa suhteellisen suoraan. Ehdollistavan mallintamisen johtopäätökset ovat selkeämmin tulkittavia, kuin ehdollistamattoman mallintamisen.

Mallit yleistyvät suoraan usealle kiinteälle kohteelle ns. monipistemallina $x_{0j} : j = 1 \ldots, p$, kun funktio $f(u; \theta)$ korvataan muodolla

$$\prod_{j=1}^{p} f(\|x - x_{oj}\|; \theta).$$

Mallintamisvaihtoehtoja on useita. Ainoastaan data asettaa rajoitteita mallintamiselle, ei teoria, ja johtaa laskennallisiin ongelmiin.

Poisson-pisteprosessi- ja Poisson-regressiomallintamisen välillä on vahva yhteys taudin riskien vaihteluiden mallintamisessa. Oletetaan, että $N(A_j)$ on tapausten lukumäärä erillisissä joukoissa A_j. Niitä voidaan käsitellä riippumattomina Poisson-muuttujina parametreilla μ_j, jotka määrittelevät yhden tai useamman A_j:hin liittyvän kovariaatin. Regressiomallilla on selvästi spatiaalinen tulkinta. Jos kaikki kovariaatit, jotka eroavat spatiaalisesti suhteellisen vähän toisistaan, ovat paloittain jatkuvia vakiofunktioita, joiden epäjatkuvuus kohdat ovat alijoukkojen rajat, on käytännössä paras ratkaisu tunnistaa yleishajonta ja sallia hajontaa aliryhmissä, jolloin $\mu_j \to \mu$. Pienin mahdollinen joukko on tietysti piste.

7 Ohjelmistotyökalujen esittely

Tutkimuksen laskennallinen osuus suoritetaan S-Plus-ohjelmistolla. Ohjelmistoon kuuluu suuri joukko standardifunktioita, joita voidaan käyttää joko yksin tai sopivasti yhdistelemällä (ohjelmoimalla) suorittamaan joustavasti erilaisia toimintoja. Käyttäjä voi vielä lisäksi liittää myös C-, tai Fortran-kielisiä funktioita. Liitteissä ja kappaleessa 7.5 on esitelty tätä tutkimusta varten ohjelmoidut funktiot.

Tässä luvussa esitellään SpatialStats-modulin sekä Splancs-kirjaston tärkeimmät ja yleisimmät funktiot. Eri kirjastojen käyttämät funktiot erotellaan toisistaaan siten, että SpatialStats:iin kuuluvat funktiot on **tummennettu** ja Splancs:iin kuuluvat on *kallistettu*. Loput funktiot ovat S-Plus:n perusfunktioita.

7.1 S ja S-Plus

S on interaktiivinen objektiorientointunut kieli data-analyysille ja grafiikalle. Sen ominaisuuksiin kuuluu esimerkiksi:

1. datan käsittely ja tallennus
2. numeeriset operaatiot skalaareille, vektoreille, matriiseille ja taulukoille
3. erinomainen grafiikka
4. pystyy korkeatasoiseen ja yksityiskohtaiseen analyysiin
5. sallii käyttäjälle mahdollisuuden ohjelmoida funktioita
6. toimii UNIX- ja WINDOWS-ympäristössä

7.2 S+SpatialStats

Tämä luku esitellään Kaluzny et. al. (1996) kirjoittaman: "S+SpatialStats users manual" pohjalta, jossa on esitetty selkeästi S-Plus:n käyttö spatiaalisen aineiston analysoinnissa.

S+SpatialStats vaatii alkuasetuksina spatial-modulin, **>module(spatial)**, käyttöönoton, jolloin data, joka on koordinaattimatriisi, pitää määritellä spp-objektiksi (kohta 2). Datan määrittelyn matriisiksi voi tehdä monella tavalla. Kohdassa 1 muodostetaan kahdesta vektorista ($v1$ ja $v2$) list-komennolla matriisi. Edellytyksenä on, että vektoreissa $v1$ on x- ja $v2$ on y-koordinaatit ja että niiden dimensiot ovat samat.

1. $>$ *pts* $<-$ list(x = v1, y = v2)

2. $>$ *pts* $<-$ **spp** (*pts*)

missä funktio **spp** tekee objektista *pts* spp-objektin *pts*. On huomattavaa, että viimeinen sijoitus jää voimaan eli Splus ei sekoita objektejapts keskenään.

7.2.1 Intensiteetti

Pisteprosessin tiheys kuvaa pisteiden lukumäärää alueessa A. Sen laskemiseksi käytetään funktiota **intensity**:

$$> \textit{tiheys} <- \textbf{intensity}(\textit{pts}, \textit{method} = "", \textit{span} = 0.1),$$

missä method voi olla joku seuraavista optioista: basic, binning, kernel tai gauss2d. Basic laskee pisteiden lukumäärän jaettuna alalla A, kun taas binning-, kernel-, gauss2d-menetelmät estimoivat paikallisia tiheyksiä A:ssa. Span-optiolla voidaan valita alueiden koot, joissa paikalliset tiheydet lasketaan. Tiheydet voidaan piirtää image-funktiolla (ks. kuva 7):

$$> \text{image}(\text{tiheys}).$$

Kolmiulotteisen tiheyskuvaajan voi piirtää **wireframe**-funktiolla (ks. kuva 8). **Wireframe**-funktion käyttö on esitelty mm. S+Spatialstats users manual s.178.

7.2.2 Simulointi

Ohjelmistoon kuuluuvalla **make.pattern**-funktiolla voidaan simuloida viiden eri jakauman mukaan jakautuneita pisteprosessien realisaatioita. Pisteprosessit ovat: binomial- (oletus), Poisson-, SSI- Strauss- ja cluster-prosessi. Simuloidaan esimerkkinä klusteroitunut prosessi:

$$> \textbf{make.pattern}(\text{n}, \textbf{process} = "\textbf{cluster}", \textbf{radius} = 0.1, \textbf{cpar} = 15),$$

missä n on simuloitavien pisteiden lukumäärä, radius on klustereiden säde sekä cpar on äitipisteiden eli klusterikeskusten lukumäärä.

Toinen vaihtoehto on ohjelmoida itse funktio simuloimaan halutun kaltainen pisteprosessi (ks. liite1).

7.3 Splancs-kirjasto

Splancs eli spatial point pattern analysis code in S-Plus esitellään Rowlingsonin & Digglen (1991) tekemän monisteen sekä Diggle P.J.(1993) artikkelin pohjalta.

7.3.1 Datan määrittely

Tallennuksen standardimuoto on siis kaksisarakkeinen taulukko, jota merkitään *pts*:llä. Taulukkoon tallennetaan n kappaletta tapahtumia x- ja y-koordinaatteineen. Seuraavassa näytetään esimerkki, jossa tehdään kahdesta vektorista taulukko.

1. > *v1* <- rnorm(20)
2. > *v2* <- runif(20)
3. > *pts* <- *as.points(v1,v2)*,

missä 1-kohdassa generoidaan 20 normaalisti jakautunutta pistettä, 2-kohdassa generoidaan 20 välillä [0,1] jakautunutta pistettä ja 3-kohdassa tehdään taulukko. Jos data on tallennettu tiedostona (S-Plus:n ulkopuolelle), joissa on tapahtumien koordinaatit, niistä saadaan tehdyksi taulukko, joka voidaan lukea scan-funktiolla ja konvertoida taulukoksi *spoints*-funktiolla. Esimerkiksi komennoilla

$$> pts < - spoints\,(\text{scan}\,('\text{points.dat}')),$$

missä 'points.dat' on tallennustiedoston nimi. Tulokseksi saadaan taulukko, *pts*, johon on tallennettu tapahtumien koordinaatit.

Kaksi tai useampia pistejoukkoa saadaan yhdistetyksi funkiolla rbind. Funktio *npts* laskee taulukossa olevien tapahtumien lukumäärän. Esimerkiksi

1. > *pts*1 $< -$ *spoints* (scan ('file1'))
2. > *pts*2 $< -$ *spoints* (runif (20))
3. > *pts*3 $< -$ rbind (pts1, pts2)
4. > *pts*4 $< -$ *pts*3 [1 : 5,]
5. > *npts* (*pts*4)

eli ensimmäiseksi luetaan tiedosto ja sitten siihen simuloidaan kymmenen tapahtumaa. Kolmanneksi yhdistetään kaksi joukkoa, jonka jälkeen valitaan niiden viisi ensimmäistä arvoa ja lopuksi tulostetaan tapahtumien lukumäärä.

Komennolla *pointmap* (*pts*) saadaan tulostettua ruudulle edellisestä pistekuvio. Jos taas halutaan lisätä tai poistaa tapahtumia pistekuviosta, käytetään komentoja > *pts* $< -$ *addpoints* (*pts*) tai > *pts* $< -$ *delpoints* (*pts*), joissa hiirellä näytetään uuden tapahtuman sijainti tai poistettava vanha tapahtuma.

7.3.2 Visualisointi

Prosessi saattaa olla jaettu mielivaltaisiin joukkoihin A_j. Ko. joukot voivat olla esimerkiksi monikulmioita. Alijoukot tallennetaan samoin kuin tapahtumat eli kaksisarakkeiseen taulukkoon. Data määrittee niiden tapahtumien järjestyksen, jotka määrittelevät alijoukkojen rajat.

Alijoukot luetaan samalla tavalla, kuin tapahtumat. Ne voidaan tulostaa näytölle funktiolla *getpoly*. Jos alijoukkojen tapahtumat on tallennettu järjestyksessä, voidaan ne piirtää pistekartaksi komennolla

$$> \; polymap\,(poly, add = T)$$

Muut alijoukkoihin liittyvät funktiot ovat:

1. *area* tulostaa alijoukkojen pinta-alat
2. *bbox* tulostaa pistekuviolle rajat
3. *pip* tulostaa ne pistejoukot, jotka kuuluvat alijoukkoon
4. *sbox* tulostaa pistekuviolle kehyksen

7.3.3 Kernel-tasoitus

Kernel-tasoitus suoritetaan komennolla

$$> \; kernel2d(pts, poly, h_0, nx, ny),$$

joka antaa estimaatin $\hat{\lambda}(x)$ arvon laskettuna tapahtumille *pts*. Parametri h_0 skaalaa estimaattorin keskihajonnan. Neliömuotoinen kernel-estimaattori on muotoa:

$$\hat{\lambda}(x) = \sum_{i=1}^{n} \left\{ 1 - \frac{d_i^2}{2h_0^2} \right\}^2, \tag{44}$$

missä d_i kuvaa etäisyyttä tapahtumasta i pisteeseen x. Funktion *kernel2d* tulostukset saadaan tulostettua (näytölle) standardi S-Plus:n image -funktiolla.

Berman ja Diggle (1989) määrittelemän optimaalisen h_0-parametrin saa lasketuksi funktiolla *mse2d*.

7.3.4 Data-analyysi

Havaitun pistekuvion nollahypoteesin testaamiseksi on siis kolme funktiota: $\hat{K}(\cdot)$, $\hat{G}(\cdot)$ ja $\hat{F}(\cdot)$. Vastaavat Splancs-funktiot ovat *khat(pts,poly,r)*, *Ghat(pts,r)* ja *Fhat(pts,poly,k,r)*.

Täydellistä satunnaista vaihtelua voidaan tutkia graafisesti piirtämällä hajontakuvio $\hat{G}(\cdot)$:sta ja $\hat{F}(\cdot)$:sta tai piirtämällä ne teoreettista muotoa (eli oletetaan, että $\hat{G}$ ja $\hat{F}$ ovat samat (ks. kaava (20)) vastaan, kun tuntematon λ korvataan aineistosta lasketulla estimaatilla $\hat{\lambda}$. K-funktion lasketaan komennolla:

$$> khat\,(pts, poly, r),$$

joka antaa K-funktion estimaatin arvon laskettua tyyppiä pts olevista pisteistä ja annetulla etäisyysvektorilla r. Vastaavasti Ripleyn (SpatialStats-) K-funktio lasketaan funktiolla **Khat**(pts).

Kahden lajin pistekuvioille saadaan K-funktio laskemalla

$$> k12hat\,(pts1, pts2, poly, r).$$

7.3.5 Simulointi

Ohjelmistoon kuuluu funktioita, joiden avulla voidaan simuloida standardijakaumien mukaan jakautuneita stokastisten pisteprosessien realisaatioita. Perusrakenne on yleensä täysin satunnainen pistekuvio eli n tapahtumaa arvotaan riippumattomasti ja samoin jakautuneesti tason mielivaltaiseen joukkoon. Splancs'ssa se toteutetaan komennolla:

$$> csr\,(poly, n),$$

joka tuottaa n tapahtumaa alueeseen, jonka polygoniobjekti *poly* määrittelee.

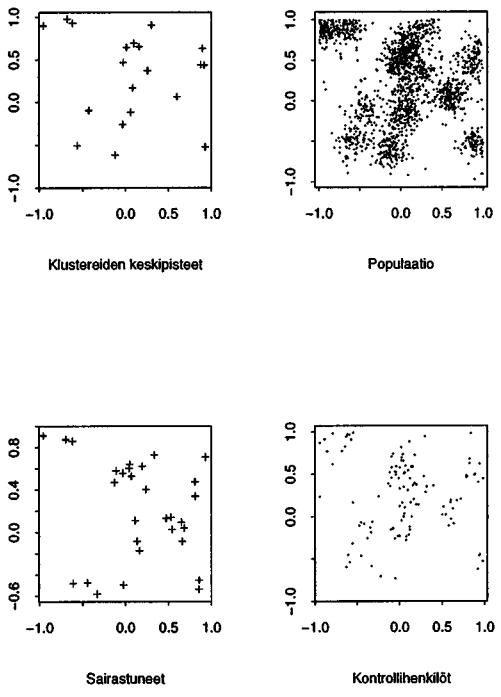
Splancs'ssa on myös funktioita, jotka laskevat simuloimalla vaihteluvälit data-analyyttisille funktioille helpottamaan niiden tulkintaa. Funktio *Kenv.csr* arpoo annetun määrän tapahtumia mielivaltaiseen joukkoon ja laskee siitä K-funktion estimaatin. Kun se lasketaan p kertaa, saadaan estimaatille vaihteluväli, jota voidaan verrata aineistosta laskettuun K-funktioon.

Funktio *Kenv.label* laskee kahden pistekuvion K-funktioiden vaihteluvälien erotuksen, kun CSR-oletus on voimassa. Sitä voidaan käyttää vertaamaan eroa simuloitujen- ja datasta lasketuiden K-funktioiden välillä eli ovatko pistekuviot satunnaisia vai ei.

Funktio *Kenv.tor* tuottaa yhdelle joukolle määrätyn määrän satunnaisia toroidi-vaihtoja ja laskee kahden lajin K-funktion jokaiselle vaihdolle ja kiinnitetylle joukolle, jolloin saadaan kahden alkuperäisen pistekuvion riippumattomuustestin keskiarvot. Toroidiksi sanotaan kappaletta, joka jatkuu samanlaisena, kun mennään ulos yhdestä reunasta ja tullaan samaan aikaan sisään vastakkaiselta reunalta.

7.4 Aineiston simulointi

Luvun 8 simulointiaineisto laadittiin Splus-ohjelmistoon itseohjelmoidulla funktiolla, jonka koodi on esitetty liitteessä 1. Funktio simuloi ensin satunnaisesti 20 klusterin keskipistettä (kuvan 4 vasen yläkulma). Seuraavaksi simuloidaan 100 pistettä taustakohinaksi alueeseen A varmistamaan, että kovin suuria valkoisia alueita ei ole. Sitten simuloidaan jokaiseen klusteriin odotusarvoltaan 200 normaalistijakautunutta pistettä (kuva 3 ja kuva 4 oikea yläkulma), joista poistetaan pisteet, jotka eivät kuulu alueeseen A eli $x \in [-1, 1]$, $y \in [-1, 1]$. Kolmanneksi "riskipopulaatiosta" arvotaan satunnaisesti 30 pistettä sairaustapauksiksi (kuvan 4 vasen alakulma). Lopuksi arvotaan vielä 120 pistettä kontrollihenkilöiksi (kuvan 4 oikea alakulma). Yksi piste kuvaa siis yhden henkilön asuinpaikan koordinaatteja riskipopulaatiossa A tietyllä ajanhetkellä.



Kuva 4: Simulointiaineiston konstruoinnin vaiheet

7.5 Ohjelmoidut S-Plus-funktiot

Liitteissä 1-4 on esitetty itseohjelmoitujen S-Plus-funktioiden koodit. Ensimmäisenä on siis funktio, jolla simuloidaan klusteriprosessi luvun 3.6.3. algoritmin mukaan. Liitteen 2 funktio piirtää kuvat 5 ja 6. Liitteen 3 funktiot piirtävät vastaavasti kuvat 6 ja 7. Liitteen 4 funktio suorittaa *Monte-Carlo*-testin aineiston ryvästyneisyyden testaamiseksi. Liitteen 5 funktio piirtää kuvan 9.

Liitteiden funktiot on ohjelmoitu simuloitua aineistoa varten. Niitä on käytetty varsinaisen aineiston analysoinnissa joko sellaisenaan tai modifioituna versiona. Modifioinnit ovat lähinnä marginaalisia, joten niitä ei ole kuvattu aina erikseen.

8 Simuloidun aineiston analysointi

Aineiston analysointi aloitetaan piirtämällä tapaukset ja kontrollit kartalle (vrt. kuvan 4 alareuna). Tarkoituksena on saada kuva aineiston jakautumisesta alueessa A. Seuraavaksi lasketaan ja piirretään empiiriset funktiot kuvaamaan aineistoa (empiiriset funktiot on esitelty luvussa 3). Kuvassa 5 on esitetty sairaustapauksista lasketut funktiot. Kuvassa 6 on vastaavasti esitetty kontrolleista lasketut funktiot. Jos kuvissa 5 ja 6 on F- ja G-funktioiden estimaattien kuvaajat ovat erilaiset, on aineisto klusteroitunut (kuten simuloidussa aineistossa). Kuvia 5 ja 6 vertaamalla saadaan myös käsitys mahdollisista eroista sairaustapausten ja kontrollien jakautumisessa.

Kuvassa 5 huomataan, että $\hat{G}(0.20) = 0.8$ ja $\hat{F}(0.20) = 0.4$. Tämä on indikaatio tapahtumien ryvästymisestä (niinkuin simuloinnin mukaan pitikin). L-funktio kertoo, että etäisyyksillä $0.1 < r < 0.3$ on ryvästyneisyyttä mutta ei enää suuremmilla etäisyyksillä. K-funktio kertoo saman kuin L-funktio mutta on vaikeampi tulkita.

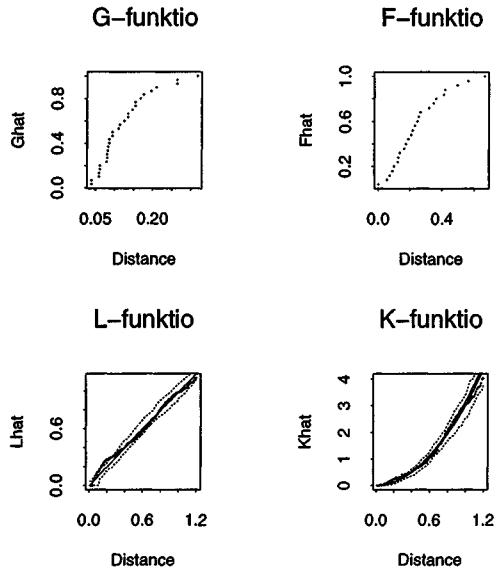
Kuvassa 6 huomataan, että $\hat{G}(0.10) = 0.8$ ja $\hat{F}(0.10) = 0.3$ eli kontrollit ovat selvästi ryvästyneet. L ja K-funktion mukaan tilastollisesti merkitsevää ryvästyneisyyttä on havaittavissa kaikilla etäisyyksillä, sillä funktioiden laskettujen (havaittujen) arvojen kuvaajat kulkevat kokoajan simuloitujen luottamusvälien yläpuolella tai ainakin ylärajan tuntumassa, joten voimme päätellä, että myös kontrollit ovat ryvästyneet.

Liitteessä 4 on esitetty ohjelma funktiosta, jolla on laskettu testisuure:

$$C = \max |\hat{F}(s_j) - \hat{G}(s_j)|, \quad j = 1, ..., m,$$

missä m pisteprosessin pisteiden lukumäärä. Testisuuretta voidaan käyttää oletuksen CSR testaamiseen *Monte-Carlo*-testillä. Ensin on laskettu aineistosta c_1. Sitten on simuloitu 99 kpl:tta täysin satunnaisia pisteprosesseja ja laskettu niistä $c_i, i \in [2, 100]$. Testille saadaan p-arvo vertaamalla c_1:n suuruutta simuloiduista aineistoista laskettuihin d_i-arvoihin. Tapauksista laskettu $c_1 = 0.75$ p-arvolla $p = 0.96$. Vastaavasti kontrolleista laskettu $c_1 = 0.7119$ ja $p = 0.01$. Eli molemmissa tapauksissa h_0 hylätään 5%:n merkitsevyystasolla ts. molemmat aineistot ovat vahvasti ryvästyneitä. Ryvästyminen voi johtua joko tapausten tai populaation ryvästymisestä. Tulokset vahvistavat edelleen käsitystä, että riskipopulaatio on heterogeeninen. Vielä ei kuitenkaan ole saatu vastausta sairauden ryvästyneisyydelle.

S-plus:n grafiikalla saa kätevästi kuvatuksi myös aineiston intensiteettiä. Kokonaistiheyksien estimaatit ovat $\lambda_k = 36.3$ (kontrollihenkilöiden) ja

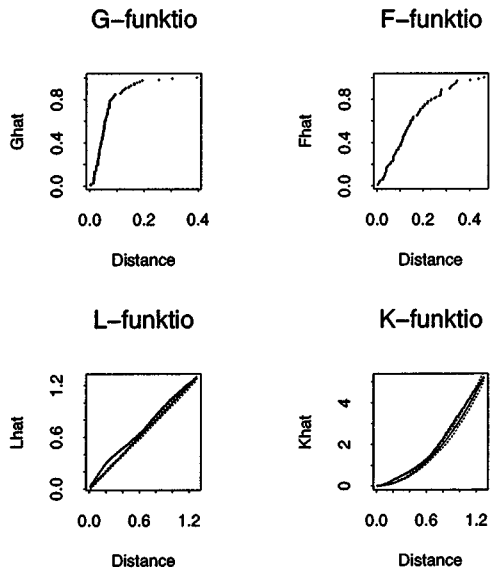


Kuva 5: Sairastapausten empiiriset funktiot

$\lambda_s = 10.6$ (sairastuneiden). Tapausten tiheydellä on tulkinta mutta kontrollien määrä alueessa on tutkijan päätettävissä, joten se ei ole mielenkiintoinen. Kuvassa 7 on kuvattu sairastuneiden- ja kontrollihenkilöiden suhteellista tiheyttä (tumma väri tarkoittaa tapahtumien tihentymää). Kuvasta huomataan, että vaikka kuvat muistuttavat toisiaan eivät ne silti ole aivan samanlaiset eli tapahtumien suhteelliset tiheydet eivät ole aivan samanlaiset. Kuvassa 8 on piirretty kolmiulotteinen kuva kontrollihenkilöiden jakautumisesta alueessa A. Tässä kuvaajan korkeus kertoo (suhteellisen) tiheyden; mitä korkeampi kuvaaja sitä suurempi tihentymä on kyseessä.

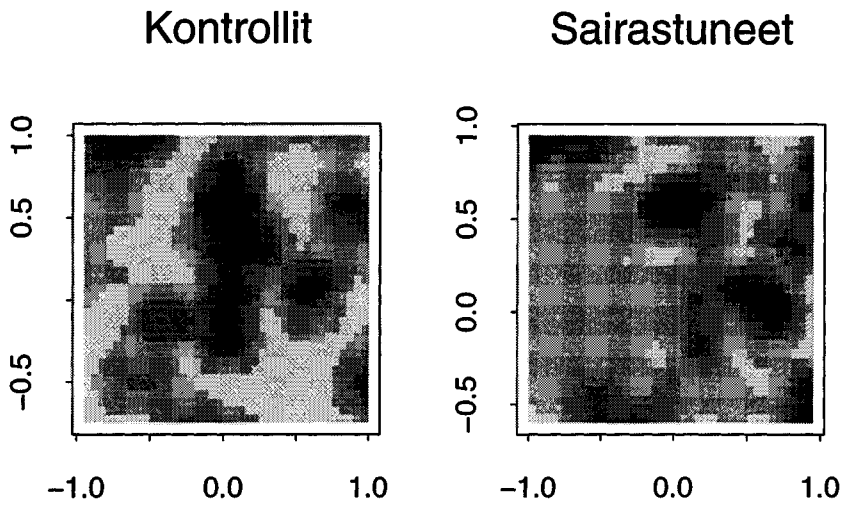
Luvussa 3 esitelty kahdenlajin K_{12}-funktion estimaatti on piirretty kuvassa 9. Funktion K_{12} tulkintahan on: $\lambda_2 K_{12}(r)$ on enintään etäisyydellä r olevien 2-pisteiden lukumäärän odotusarvo satunnaisesti valitusta 1-lajin pisteestä. Kuvan 9 tulkinta on, että $\hat{\lambda}_k \hat{K}_{12}(r)$ on enintään etäisyydellä r satunnaisesti valitusta tapauksesta olevien kontrollien lukumäärän odotusarvon estimaatti. Kuviosta 9 huomataan, että kontrollihenkilöiden lukumäärän odotusarvo ylittää satunnaisen aineiston odotusarvon, kun etäisyys > 0.3. Tulos ei edelleenkään ole yllättävä, koska simuloimme tarkoituksella klusteriprosessin. Tämä ei siis johda tulkintaan tapausten klusteroitumisesta.

Tähän asti on tutkittu tapausten ja kontrollien klusteroitumista. Seuraa-

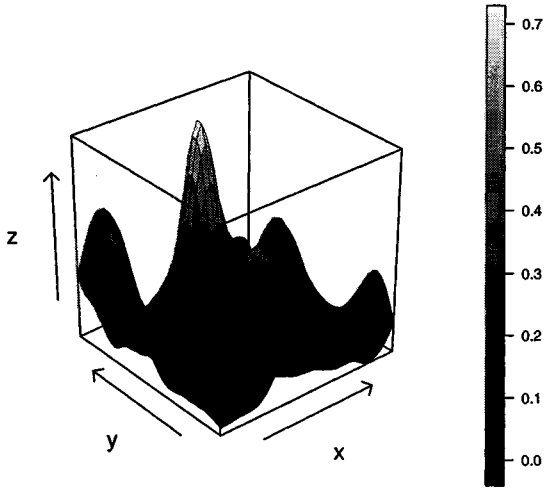


Kuva 6: Kontrollien empiiriset funktiot

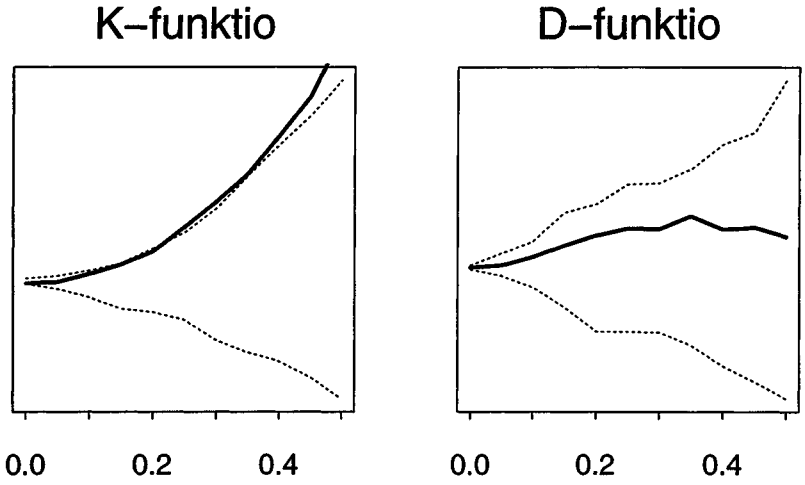
Tähän asti on tutkittu tapausten ja kontrollien klusteroitumista. Seuraavaksi tutkitaan, onko tapausten klusteroituminen pelkkää satunnaisvaihtelua. Luvussa 4 esitetyn kaavan (37) mukainen analyysi on tehty liitteen 5 funktiolla, jossa on siis laskettu tunnusluku $\hat{D}$ eri mittakaavoilla ja sen jälkeen on simuloitu sille luottamusväli. Kuvasta 9 huomataan, että $\hat{D}$ on koko ajan selvästi luottamusvälien sisäpuolella. Johtopäätöksenä on, että sairastapausten ryvästyneisyys johtuu ympäristön heterogeenisuudesta eli ne eivät ole ryvästyneet populaation heterogeenisuutta enempää. Tulkinta tuntuu luonnolliselta, koska ensin simuloitiin klusteriprosessi ja sen jälkeen arvottiin satunnaisesti niin sairastuneet kuin kontrollitkin riskipopulaatiosta. Todennäköisyys, että tapaukset olisivat ryvästyneet riskipopulaatiota enemmän on hyvin pieni.



Kuva 7: Kontrollien ja sairastuneiden tiheydet



Kuva 8: Kontrollihenkilöiden 3-d-tiheys

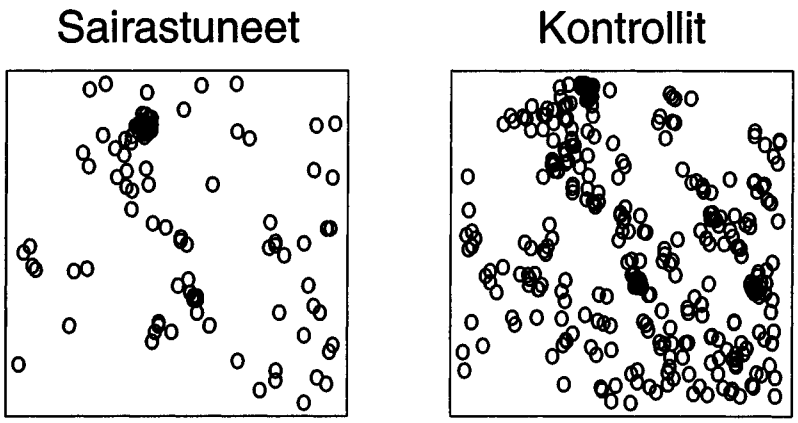


Kuva 9: $\hat{K}_{12}-$ ja $\hat{D}$-funktio simuloidulle aineistolle luottamusväleineen. $\hat{K}_{12}$-funktiossa on mukana ympäristön heterogeenisuus, joka on vähennetty $\hat{D}$-funktiosta.

9 Klusteroitumisen tutkiminen keuhkosyöpäaineistolle

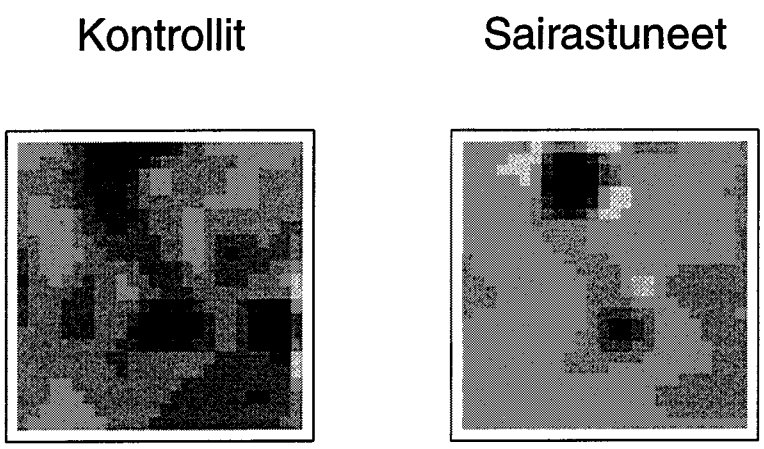
9.1 Aineiston analysointi

Keuhkosyöpäaineisto on siis kerätty 50×50 olevalta alueelta Pohjois-Savosta. Kuvaan 10 on piirretty jo luvussa 2 esiintynyt keuhkosyöpäaineiston hajontakuvio. Kuten jo luvussa 2 todettiin, ei tapausten ryvästymisestä voi kuvan perusteella tehdä johtopäätöksiä. Tarvitaan aineiston tarkempaa analysointia.



Kuva 10: Keuhkosyöpäaineisto Pohjois-Savosta. Vasemmassa kuvassa on sairastuneiden henkilöiden ja oikeassa kontrollihenkilöiden sijainnit. Alueen sivujen pituudet ovat 50 km.

Kuvissa 11, 12 ja 13 on kuvattu tapahtumien ja kontrollien tiheyttä. Kuvista huomataan, että alueella on yksi selkeä tapaustihentymä. Kontrolleilla on yksi suuri ja pari pienempää tihentymää alueella. Kontrollihenkilöiden suurin tihentymä osuu samaan kohtaan, kuin tapausten selkein tihentymä. Edellisestä voidaan otaksua, että tapaustihentymä voi olla seurausta riskipopulaatiossa olevasta klusterista. Muuten niin tapaukset kuin kontrollitkin ovat melko säännölisesti jakautuneet alueella.



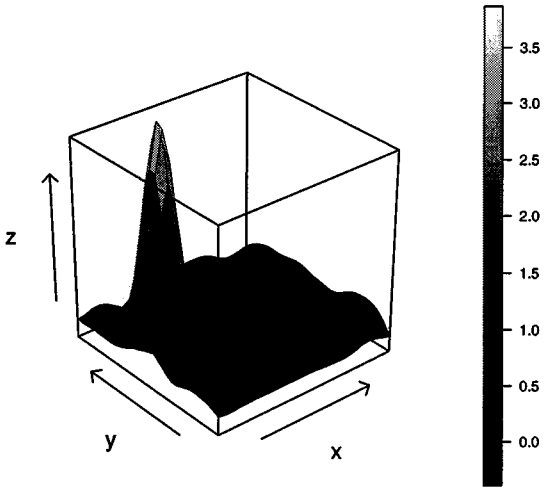
Kuva 11: Kontrollien vs. sairastuneiden tiheys

Kuten luvussa 2 jo todettiin, ei silmämääräisesti tulkittuna voida tehdä johtopäätöksiä aineiston klusteroitumisesta. Kuvassa 10 on esitetty sama kuva kuin kuva 1 luvussa 2. Sen sijaan kuvista 11, 12 ja 13 on havaittavissa selvää heterogeenisuutta aineistossa. Aineiston varsinainen analysointi aloitetaan laskemalla ja piirtämällä G-, F-, L- ja K-funktiot niin sairastuneiden kuin kontrollien prosesseille. Kuvasta 14 huomataan, että $\hat{G}(2000) = 0.4$ ja $\hat{F}(4000) = 0.4$, joten ainakin näiden perusteella sairastapaukset ovat ryvästyneet (huom. 2000 = 2 km). Vastaavasti kontrollihenkilöistä (kuva 15) havaitaan, että $\hat{G}(2000) = 0.8$ ja $\hat{F}(4000) = 0.8$, joten myös ne ovat ryvästyneet.

L- ja K-funktiot, jotka pitävät sisällään saman informaation, kertovat, että tapaukset ovat ryvästyneet, koska molemmat funktiot ovat simuloitujen luottamusvälien yläpuolella 0-20 km:n välillä ja alapuolella välillä 25-30 km:iin.

Kontrollihenkilöiden L- ja K-funktioiden estimaatit kulkevat samalla lailla kuin tapaustenkin kuvassa 16. Kuvan 17 funktiossa aineistosta laskettu kuvaaja kulkee luottamusvälin yläpuolella 0-10 km ja alapuolella 25-30 km etäisyyksillä.

K-funktion tulkintahan on, että $\lambda\, K(r)$ tapahtumien lukumäärän odotusarvon estimaatti etäisyydellä r olevasta satunnaisesti valitusta tapahtumasta. Silloin kuvista 16 ja 17 päätellään, että molemmat aineistot ovat



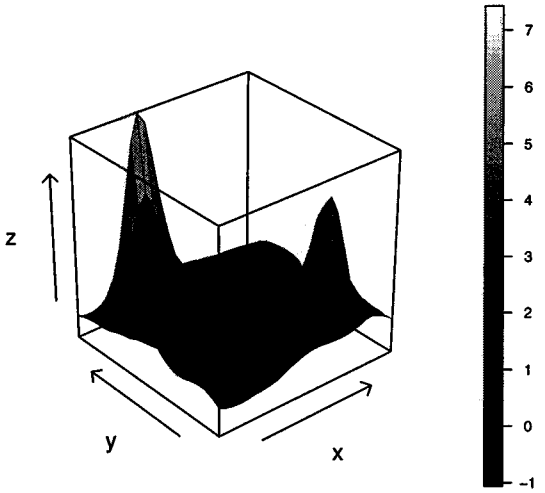
Kuva 12: Sairastuneiden tiheys

ryvästyneitä eli että riskipopulaatio on heterogeeninen.

Edellä on päätelty, että molemmat aineistot ovat ryvästyneitä. Ryvästyminen voi siis johtua joko tapausten tai kontrollien ryvästymisestä, joten vielä ei ole saatu vastausta sairauden ryvästymisestä.

Sairauden ryvästymisen tutkiminen suoritetaan piirtämällä kuvat 18 ja 19. Kuvissa on piirretty K- ja D-funktiot luottamusväleineen, kun säde $r_{\max} = 1.0 \cdot 10^4$ m (18) ja kun $r_{\max} = 1.5 \cdot 10^4$ m (19). Funktioiden (piste-) estimaatit on piirretty yhtenäisellä viivalla ja simuloidut luottamusvälit pisteinä. K_{12}-funktiossa on mukana tapausten ja ympäristön heterogeenisuus (1=sairastuneet, 2=kontrollit). D-funktiosta on poistettu ympäristön heterogeenisuus, joten jäljelle on jäänyt tapausten heterogeenisuus eli sairauden "ryvästyneisyys".

Kuvassa 18 on K_{12}-funktio ryvästynyt käytännössä koko välillä 0-10 km. D-funktio taas kertoo syöpätapausten säännöllisyydesstä välillä 0-5 km. Kun sädettä kasvatetaan 15:sta kilometriin, niin molemmat funktiot kertovat aineiston olevan satunnainen. Johtopäätöksenä on siis, että tästä aineistosta ei löytynyt todistetta epäillylle keuhkosyövän alueelliselle sairastuvuusylimäärälle.



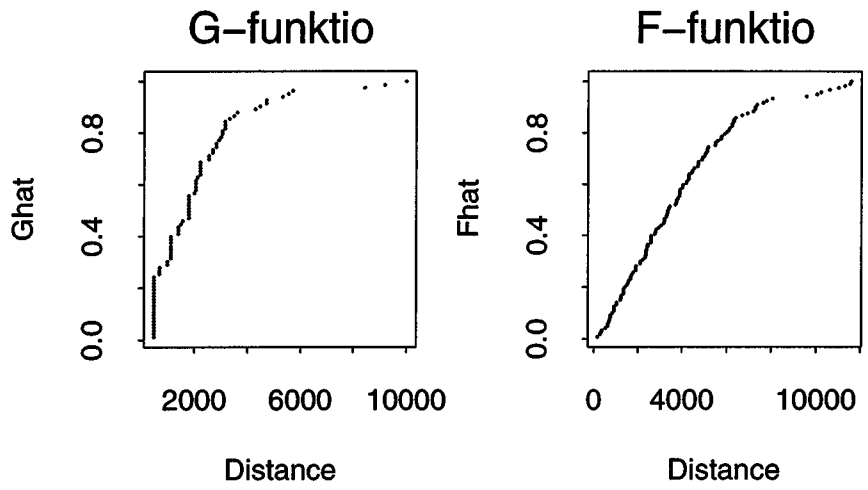
Kuva 13: Kontrollien tiheys

9.2 Yhteenveto ja pohdintaa

Keuhkosyöpäaineiston analysoinnin johtopäätöksenä on siis, että käytetty menetelmä ei löytynyt todisteita epäillylle keuhkosyövän alueelliselle sairastuvuusylimäärälle. Itse asiassa tapahtumien jakautuminen tuntuu olevan varsinkin pienillä etäisyyksillä (0-5 km) pikemminkin säännöllistä kuin satunnaista puhumattakaan ryvästyneisyydestä (ks. kuva 18).

Se, että epäilyistä huolimatta, ei alueelta löytynyt sairastuvuusylimäärää voi johtua kahdesta syystä. Ensinnäkin siitä, että sairastuvuusylimäärää ei yksinkertaisesti esiinny. Toinen syy voi olla ongelmat kontrollien valinnassa. Kuinka valita spatiaalisesti edustava otos kontrolleja? Tämän tutkimuksen kontrollit valittiin satunnaisotoksena alueella samaan aikaan asuneista miehistä. Ainoa sekoittava tekijä, joka otetiin valinnassa huomioon oli ikä (asuinpaikan lisäksi). Muita mahdollisia muuttujia olisi ainakin, tupakointitottumukset, yleiskunto ja sosiaaliluokka, jotka varmasti vaikuttavat henkilön terveydentilaan ja sairastumisriskiin. Lisäksi kontrollien lukumäärän valinta saattaa aiheuttaa vaikeuksia. Yleisesti (mm. Diggle, 1993) pidetään riittävänä, jos kontrolleja on 3-4 kertaa tapausten lukumäärä. Tässä tutkimuksessa kontrolleja oli käytettävissä lähes 4-kertainen määrä tapauksiin nähden.

Kolmas, vaikkakin teoreettinen, syy voisi olla menetelmän riittämättömyys mikä ei tunnu uskottavalta, koska menetelmä tuntui olevan herkkä ai-



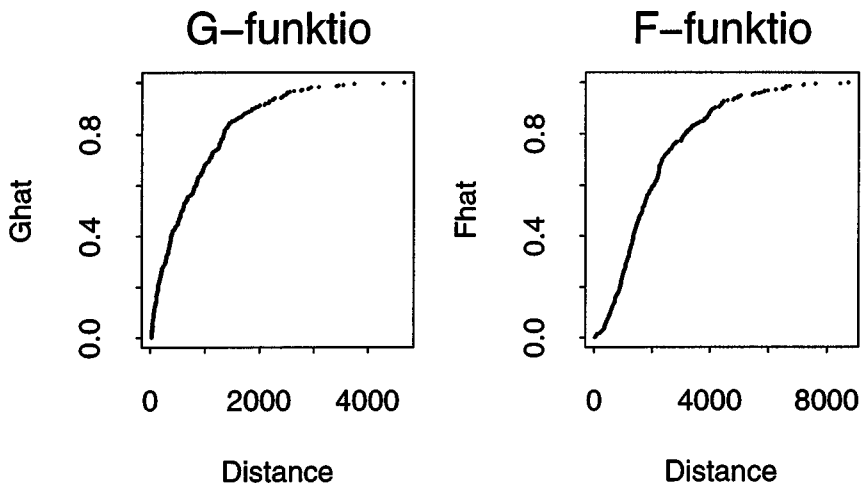
Kuva 14: Sairastuneiden $\hat{G}$- ja $\hat{F}$-funktiot

neistojen vaihtelulle eri simulointien suhteen. Lisäksi menetelmällä on päästy hyviin tuloksiin muissa yhteyksissä (ks. esim. Gardner, 1991). Tämän menetelmän etu aggregoitujen aineistojen analysointimenetelmiin nähden on se, että taustalla olevan stokastisen prosessin spatiaalinen rakenne ei häviä.

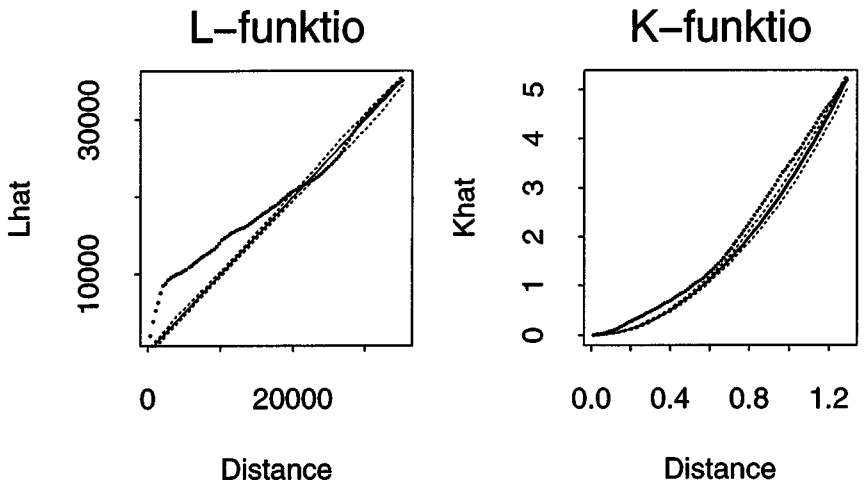
Tutkimuksen teoriaosassa esiteltiin ensin pisteprosessien teoriaa ja keskityttiin varsinkin kahden lajin prosessien toisen kertaluvun ominaisuuteen. Seuraavaksi käytiin läpi spatiaalisen klusteroitumisen hypoteesit ja klusteroitumisen testaus sekä tapaus-kontrolliasetelma.

Tutkimuksen empiirisessä osassa (luvut 8-9) tehtiin empiiriset tarkastelut ensin simuloidulle ja sitten varsinaiselle aineistolle.

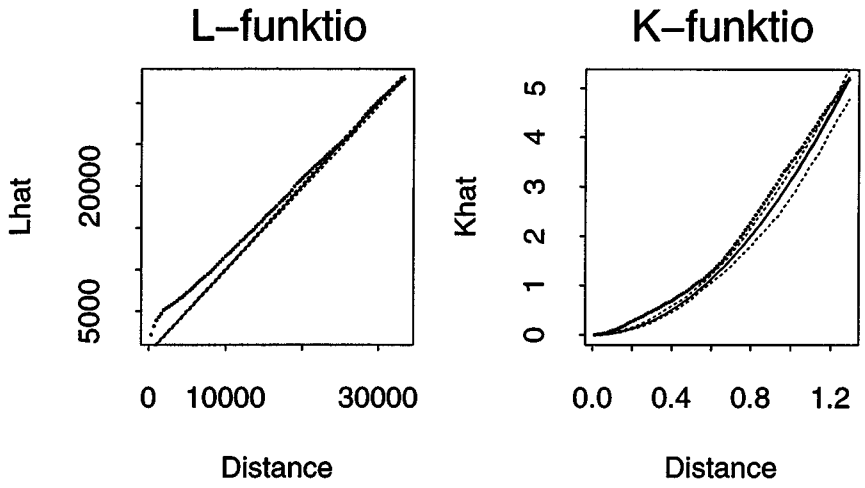
Tässä tutkimuksessa käsiteltiin klusteroitumista vain tilassa. Monissa yhteyksissä yhtä tärkeää on ottaa huomioon ajanvaikutukset ja tutkia klusteroitumista tilan lisäksi ajassa (Diggle, 1993).



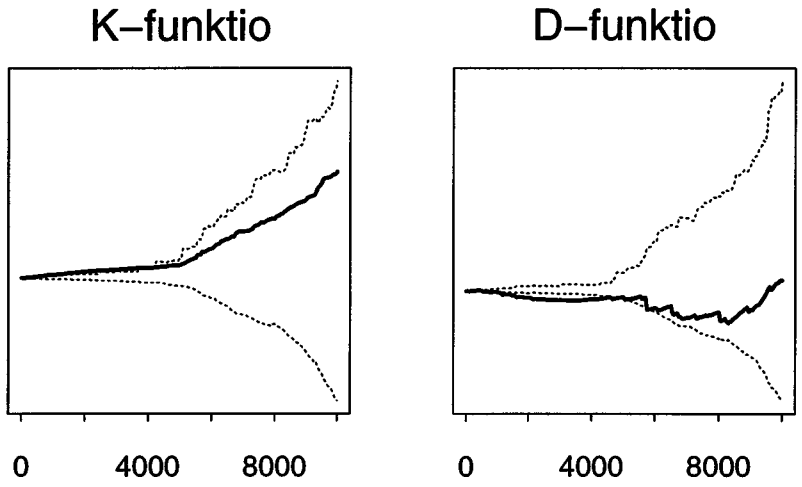
Kuva 15: Kontrollien $\hat{G}$- ja $\hat{F}$-funktiot



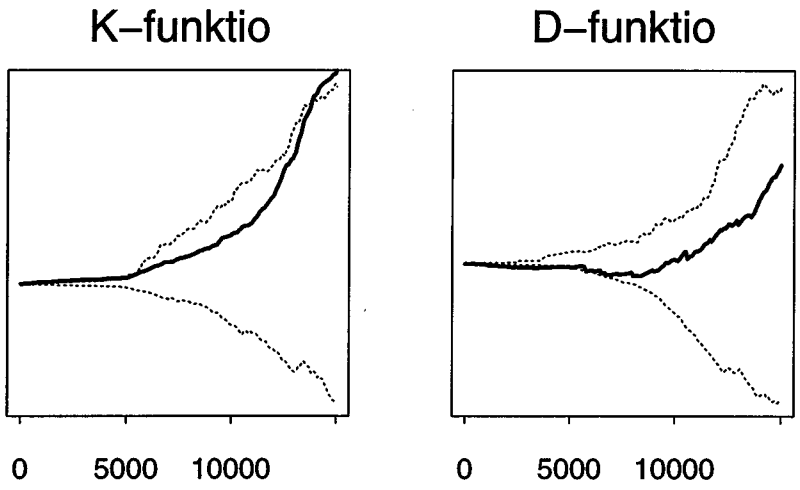
Kuva 16: Sairastuneiden $\hat{L}$- ja $\hat{K}$-funktiot. $\hat{K}$-funktiossa etäisyys 1.2 vastaa $\hat{L}$-funktion etäisyyttä 20 km.



Kuva 17: Kontrollien $\hat{L}$- ja $\hat{K}$-funktiot



Kuva 18: Aineistosta estimoidut $\hat{K}_{12}$- ja $\hat{D}$-funktiot (paksu yhtenäinen viiva) ja niiden pisteillä piirretyt luottamusvälit, kun $r_{\max} = 1.0 \cdot 10^4$ m



Kuva 19: Aineistosta estimoidut $\hat{K}_{12}$- ja $\hat{D}$-funktiot, kun $r_{\max} = 1.5 \cdot 10^4$ m

Kiitosmaininnat

Tutkimusta on ohjannut dos. Antti Penttinen. Työtä on osittain rahoittanut Suomen Akatemian projektit "Small-area Analysis of Canser Incidence around a Point Source" n:o 42559 ja "Analysis of Spatial Environmental Data" no: 39372. Lisäksi haluan kiittää dos. Juha Pekkasta ja FM Esa Kokkia Kansanterveylaitoksen ympäristöepidemiologian osastolta tutkimuksen aineiston hankintaan ja käsittelyyn liittyvästä tuesta.

Lähdeluettelo

Anderson, N.H. and Titterington, D.M. (1997) Some methods for investigating spatial clustering, with epidemiological applications. *J.R. Statist. Soc. A* **160**, 87-105.

Besag, J. and Newell, J. (1991) The detection of clusters in rare diseases. *J.R.Statist. Soc.* **A 154**, 143-155.

Cressie, N. (1991) *Statistics for Spatial Data.* New York: Wiley.

Diggle, P.J. (1983) *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.

Diggle, P.J. (1993) *Point Process Modelling in Environmental Epidemiology.* Statistics for the Environment (V. Barnett and F. Turkman, eds.) New Tork: Wiley.

Diggle, P.J. and Chetwynd, A.G. (1991) Second-order analysis of spatial clustering for inhomogeneus populations. *Biometrics.* **47**, 1153-63.

Gardner, M.J. (1991) Investicating childhood leaukemia rates around the Sellafield nuclear plant. *Bull. Int. Statist. Inst.* **54**

Kaluzny, S.P., Vega, S.C., Cardoso, T.P. and Shelly, A.A. (1996) *S+SpatialStats User's Manual,* Version 1.0, Seattle :MathSoft, Inc.

Ripley, B.D. (1981) *Spatial Statistics.* New York: Wiley.

Stoyan, D., Kendall, W.S. and Mecke, J. (1987) *Stochastic Geometry and its Applications.* Berlin: Wiley.

Liitteet

Liite 1: Klusteriprosessin generointiohjelma.

```
function(lkm = 20, n = 200, alue = 1, case = 30, otos = 120)
{
# Tämä funktio generoi klusteriprosessin. Kutsuparametrit ovat: lkm =
# klustereiden lkm n = pisteiden lkm:n odotusarvo klusterissa, case kertoo
# populaatiosta arvottujen sairastapausten lukumäärän ja otos, joka vastaa
# kontrollihenkilöiden lukumäärää.

# Alustellaan:
        tulos <- matrix(0, nrow = lkm * n, ncol = 2)
        xcase <- rep(0, length(case))
        xotos <- rep(0, length(otos))
        alku <- 1
        loppu <- 0
        luku <- 1
        k <- 1
        luku2 <- 1
        l <- 1
# Generoidaan klustereiden keskipisteiden x- ja y-koordinaatit:
        x0 <- runif(rpois(lkm, 1), - alue, alue)
        y0 <- runif(length(x0), - alue, alue)
        z0 <- list(x = x0, y = y0)
# "Taustakohina":
        x <- runif(100, - alue, alue)
        y <- runif(length(x), - alue, alue)
        x <- x[x * x + y * y < 2]
        y <- y[x * x + y * y < 2]
# Generoidaan pisteet klustereihin ja niiden x- ja y-koordinaatit:
        for(i in 1:lkm) {
                x1 <- rnorm(rpois(n, 1), x0[i], 0.1)
                y1 <- rnorm(length(x1), y0[i], 0.1)
# Pienennetään aluetta:
                xpiste <- x1[x1 * x1 < 1]
                ypiste <- y1[y1 * y1 < 1]
# Tallennetaan klusterin pisteiden sijainnit (piste kerrallaan) matriisiin:
                loppu <- loppu + length(xpiste[xpiste != 0])
                for(j in alku:loppu) {
                        tulos[j, 1] <- xpiste[j - alku + 1][xpiste[j - alku + 1 ] != 0]
```

```
                tulos[j, 2] <- ypiste[j - alku + 1][ypiste[j - alku + 1 ] != 0]
            }
            alku <- alku + length(x1[x1 != 0])
        }
        pituus <- length(tulos[, 1][tulos[, 1] != 0]) + length(x[x != 0])
        populaatio <- matrix(0, nrow = pituus, ncol = 2)
        populaatio[, 1] <- c(x, tulos[, 1][tulos[, 1] != 0])
        populaatio[, 2] <- c(y, tulos[, 2][tulos[, 2] != 0])
# Arvotaan sairastuneet populaatiosta:
        while(k <= case) {
            luku <- runif(1, 1, length(populaatio[, 1])) %/% 1
            xcase[k] <- luku
            k <- k + 1
        }
        sairaat <- populaatio[xcase, ]
# Arvotaan otos populaatiosta:
        while(l <= otos) {
            luku2 <- runif(1, 1, length(populaatio[, 1])) %/% 1
            xotos[l] <- luku2
            l <- l + 1
        }
        otos <- populaatio[xotos, ]
# Tulostukset:
        list(klu = z0, populaatio = populaatio, otos = otos, case = sairaat)


}
```

Liite 2: Pisteprosessitiivistelmien (G-, F-, K- ja L- funktioiden) lasku- ja piirto-ohjelma.

```
function(data)
{

# Tämä funktio laskee ja piirtää G-, F-, K- ja L-funktiot.
        par(mfrow = c(2, 2), pty = "s")
# Tehdään datasta spp-objekti:
        data.spp <- spp(data)
# G-funktio:
        ghat <- Ghat(data.spp)
        title(main = "G-funktio")
# F-funktio:
        fhat <- Fhat(data.spp)
        title(main = "F-funktio")
# L-funktio:
        lhat <- Lhat(data.spp)
        Lenv(data.spp, nsims = 50)
        title(main = "L-funktio")
# K-funktio(vertailu kohtana analyyttisesti lasketut arvot):
        khat <- Khat(data.spp)
        anal <- khat$values[, "dist"]
        lines(anal, pi * anal^2)
        Kenv(data.spp, nsims = 50)
        title(main = "K-funktio")

        par(mfrow = c(1, 1))

}
```

Liite 3: Kaksi tiheyden estimointi- ja piirto-ohjelmaa.

```
# Funktio laskee otosten ja tapahtumien tiheydet (vrt. luku 7.22) ja
# piirtää kuvan 7.

function(data)
{       par(mfrow = c(1, 2), pty = "s", mar=c(2,2,4,1))
        data1.spp <- spp(data$otos)
        data2.spp <- spp(data$case)

        int1 <- intensity(data1.spp, method = "binning", nx = 30, ny = 30, span = 0.1)
        int2 <- intensity(data2.spp, method = "binning", nx = 30, ny = 30, span = 0.1)

        image(int1, main = "Kontrollit")
        image(int2, main = "Sairastuneet")

        par(mfrow = c(1, 1))
}



# Funktio laskee tiheyden ja piirtää kuvan 8

function(data)
{
# Tehdään "3-ulotteinen"kuvaaja!
#
        data.spp <- spp(data)
        int <- intensity(data.spp, method = "binning", nx = 30, ny = 30, span = 0.1)

        trellis.device(postscript,print.it=F,color = F)
        mgrid <- expand.grid(x = int$x, y = int$y)
        mdf <- data.frame(x = mgrid$x, y = mgrid$y, z = c(int$z))
        print.trellis(wireframe(z   x * y, data = mdf, drape = T))
        dev.off() }
```

Liite 4: Täydellisen satunnaisuuden testaus F- ja G- funktioiden avulla.

```
function(data)
{
#
# CSR-testi simuloidulle aineistolle käyttäen F- ja G-funktioita.
#
        pts <- spoints(data)          # luetaan data
        lkm <- npts(pts)              # pisteiden lkm
        s <- seq(0, 1, 0.05)          # mittakaava
#
# Alustetaan alue simuloidulle aineistolle ja lasketaan d1 = |F^(s) - G^(s)|
#
        s2 <- seq(0, 1, 0.05)
        alue <- cbind(s, s2)
        g <- Ghat(pts, s, plot.it = F)
        f <- Fhat(pts, alue, s, plot.it = F)
        d <- max(abs(f - g))
#
# simuloidaan 99 CSR dataa ja lasketaan di = |F^(s) - G^(s)|
#
        for(i in( 2:100) {
                x <- runif(lkm, -1, 1)
                y <- runif(length(x), -1, 1)
                spts <- cbind(x, y)
                g <- Ghat(spts, s, plot.it = F)
                f <- Fhat(spts, alue, s, plot.it = F)
                d <- c(d, max(abs(f - g)))          # tallennetaan erotukset vektoriin d
        }
#
# tulostetaan p-arvo eli d1:n sijainti simuloituihin verrattuna
#
        list(p = p <- rank(d)[1])
}
```

Liite 5: K_{12}- ja D-funktioiden ohjelmat.

```
function(case1, case2)
{
#
# Tällä funktiolla lasketaan K12- ja D-funktiot (d=k2-k1) tapauksille
# ja kontrolleille.
#
        mitta <- seq(0, 0.5, 0.05)                    # Mittakaava
        par(mfrow = c(1, 2), pty = "s", mar = c(2, 2, 2, 1))
#
# Alustetaan alue (kehkosyöpäaineiston analysoinnissa käytetään taulukon ??
# mukaista määritystä):
#
        x <- c(-1, 1, 1, -1)
        y <- c(-1, -1, 1, 1)
        alue <- cbind(x, y)
#
# Lasketaan K12-funktio. Simuloidaan sille ylä- ja alarajat ja piirretään ne
# samaan kuvaan aineistosta lasketun kanssa.
#
        k12 <- k12hat(case1, case2, alue, mitta)
        alue2 <- spp(x = alue[, 1], y = alue[, 2])
#       alue2 <- list(x = alue[, 1], y = alue[, 2])     # Toinen tapa alustaa alue.
#
        k12case <- Kenv.label(case1, case2, alue2, 50, mitta, quiet = T)
        matplot(mitta, cbind(k12case$lower, k12case$upper), col = 1,
                type = "l", lty=2, ylim = c(-0.55, 0.55))
        par(lwd = 2)
        lines(mitta, k12)
        title(main = "K-funktio")
#
# Lasketaan d12-funktio ja simuloidaan sille teoreettiset ylä- ja alarajat.
# Piirretään ne yhdessä samaan kuvaan.
#
        k1 <- khat(case1, alue, mitta)
        k2 <- khat(case2, alue, mitta)
        d12 <- k2 - k1
#
        kenvcase <- Kenv.label(case1, case2, alue, 50, mitta, quiet = T)
        matplot(mitta, cbind(kenvcase$lower, kenvcase$upper), col = 1,
```

```
                type = "l",lty=2, ylim = c(-0.55, 0.55))
        par(lwd = 2)
        lines(mitta, d12)
        text(0.4, 0, c("d=k2-k1"))
        title(main = "D-funktio")
}
```