Aalto University
School of Engineering

Introduction to spatio-statistical methods

Kirsi Virrantaus
Department of Built Environment
School of Engineering

8.2.2019

ENY-C2005

Aalto University
School of Engineering

Johdanto spatiotilastollisiin menetelmiin

Kirsi Virrantaus
Department of Built Environment
School of Engineering

8.2.2019

ENY-C2005

Contents of the lecture

- What is **scientific research** ?
- What is **spatio-statistical** research ?
- What is **spatial autocorrelation** and how to identify it ?
- Concepts and elementary tasks in **spatio-statistical analysis**
- Examples on **descriptive spatio-statistical** methods and **eksplorative** analysis
- **Spatio-statistical inference**
- Problems and solutions in spatial analysis – an example on how standard **linear regression** can be extended into a **spatial method**

Luennon sisältö

- Mitä on **tieteellinen tutkimustapa** ?
- Mitä on **spatiotilastollinen** tutkimus?
- **Spatiaalinen autokorrelaatio** ja sen tunnistaminen datasta
- **Spatiotilastolliset** perustehtävät ja –käsitteet
- Esimerkkejä **spatiaalisista kuvailevista menetelmistä** ja **eksploratiivisesta analyysistä**
- Mitä on **spatiotilastollinen päättely**
- Spatiaalisen datan mallinnuksen problematiikkaa – esimerkkinä kuinka **regressiomallissa** voidaan huomioida spatiaalisuus

1.Scientific approach to research - using statistical analysis

- when researchers approach **phenomena and problems** of reality they use **scientific approach**
- the starting point of scientific approach is to define the **context and concepts**
- the available data is first **described**, in order to understand the **dependencies** between phenomena
- **hypotheses** are created and **models** are developed in order to test the hypotheses
- if the the hypotheses can be confirmed, **laws(rules)** and even **theories** can be developed

1. Tieteellinen tutkimustapa – tilastollinen analyysi

- kun tutkijat lähestyvät todellisuuden **ilmiöitä ja ongelmia**, he käyttävät tieteellistä lähestymistapaa
- tieteellinen lähestymistapa lähtee **kontekstin** (=asiayhteys) ja **käsitteiden** määrittelystä
- käytettävissä olevaa dataa **kuvaillaan**, jotta saataisiin käsitys vallitsevista **riippuvuuksista** asioiden välillä
- luodaan **hypoteeseja** ja kehitetään **malleja** hypoteesien testaamiseksi
- jos hypoteesit saavat vahvistusta niistä voidaan kehittää **lakeja(sääntöjä)** ja jopa **teorioita**

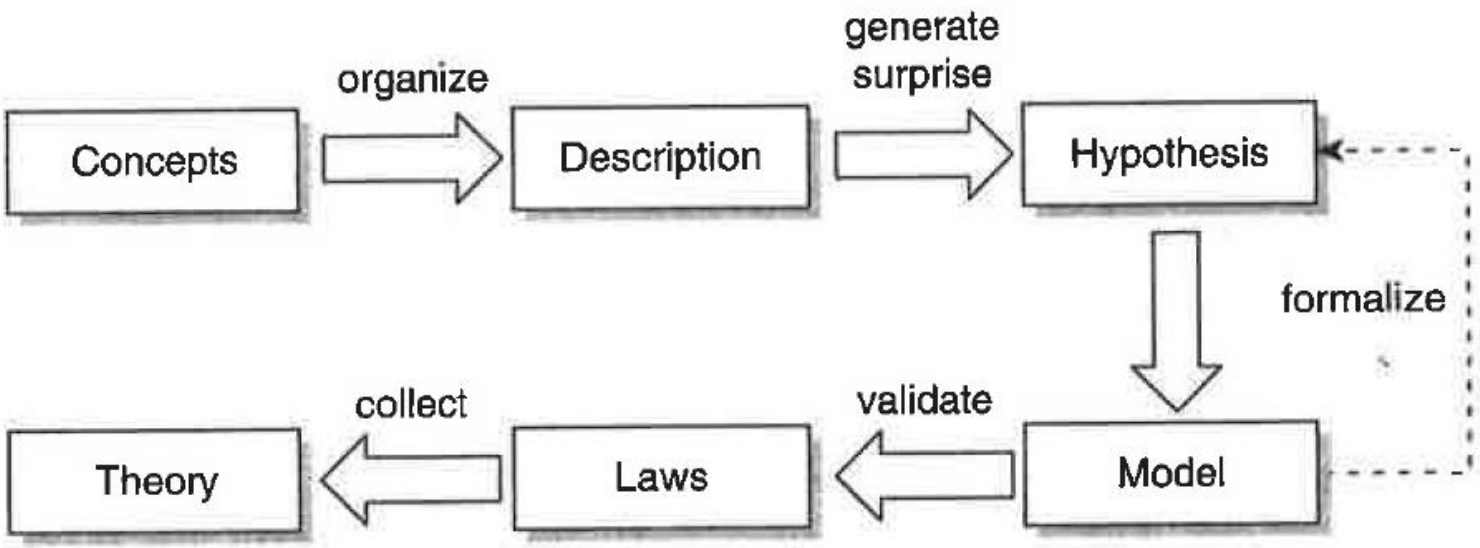


Figure 1.1 The scientific method

The core of scientific method is statistical analysis.
Tieteellisen menetelmän ytimen muodostavat tilastolliset menetelmät.

(Rogerson, 2015)

2. Statistical methods

- in statistical science iboth descriptive and inferential methods are used
- **descriptive statistics**
  - describing the **characteristics of the data**: mean, variance, standard deviation, median
  - processing a sample data set
  - explorative approach
- **inferential statistics**
  - based on a **hypothesis** of the behaviour of the phenomenon
  - a **model** is assumed to represent the whole population
  - attempt to be able to predict the behaviour of the phenomenon in the future
  - confirmatory methods

2. Tilastollinen tutkimus

- tilastotieteessä hyödynnetään kuvailevia menetelmiä sekä tilastollista päättelyä
- **kuvaileva spatiotilastotiede**
  - menetelmillä **kuvataan datan piirteitä**: keskiarvo, varianssi, keskihajonta, mediaani
  - otosaineiston käsittely (otos = näytejoukko koko populaatiosta)
  - eksploratiivinen, tutkiva lähestymistapa
- **tilastollinen päättely**
  - perustuu hypoteesiin ilmiön käyttäytymisestä
  - käytetään jotain **mallia** jonka oletetaan kuvaavan koko populaatiota
  - pyrkimys voida ennustaa ilmiön käyttäytymistä tulevaisuudessa
  - konfirmatoriset, vahvistavat menetelmät

Spatio-statistical methods

- spatial data – points, lines, polygons – are defined in 2d space (or in 3d space)
- when analyzing them with basic statistical measures - mean,variance,standard deviation – the dimensionality has to be taken into account
- Examples:
  - **mean center** (compare to mean)
    - the average x and y coordinate of all the points in the study area
  - **standard distance** (compare to standard deviation)
    - summary measure of feature distribution around their center

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} + \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}}$$

Spatiotilastolliset menetelmät

- spatiaalinen data – pisteet, viivat, alueet – on määritelty 2d-avaruudessa (tai 3d)
- kun pisteitä, viivoja ja alueita kuvataan tilastollisin mittarein (keskiarvo, varianssi, keskihajonta), tulee dimensioisuus ottaa huomioon
- esimerkkejä
  - Keskiarvopiste (vert keskiarvo)
    - x:n ja y:n keskiarvopiste kaikista tutkimusalueen pisteistä laskettuna
  - Keskietäisyys
    - mittari joka ilmaisee kohteiden keskiarvopisteeseen laskettujen etäisyyksien hajonnan

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n} + \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}}$$

Specialities of spatial statistics

- spatio-statistical analysis is based on regular statistical methods
- but, they **can not be applied as such**

- Special features of spatial phenomena and data:
  - **spatial autocorrelation (SAC)**
  - other structural dependiecies ja interrelationships, physical obstacles and structures that cause **heterogeneity**

- if we can identify autocorrelation or heterogeneity it can be managed and taken into account in the method

Spatiotilaston lähtökohtia

- tilastolliset menetelmät ovat spatiaalisen analyysin perusta
- mutta, tilastolliset menetelmät **eivät ole käyttökelpoisia sellaisenaan**

- spatiaalisten ilmiöiden ja datan erityispiirteet:
  - spatiaalinen autokorrelaatio; **spatial autocorrelation (SAC)**
  - muut spatiaaliset riippuvuudet ja keskinäiset suhteet, fyysiset esteet ja rakenteet joista seuraa **spatiaalinen heterogeenisyys**

- kun autokorrelaatio tai heterogeenisyys **tunnistetaan** se voidaan hallita ja ottaa huomioon menetelmässä

2. Spatial autocorrelation

- **autocorrelation** is mathematical method for describing dependencies in 1d- or 2d-data; for example in **time series** or **spatial data layer**

- as a phenomenon **spatial autocorrelation** expresses the Tobler´s first law of Geography:
  - **”Everything is related to everything else, but near things are more related to each other.”**

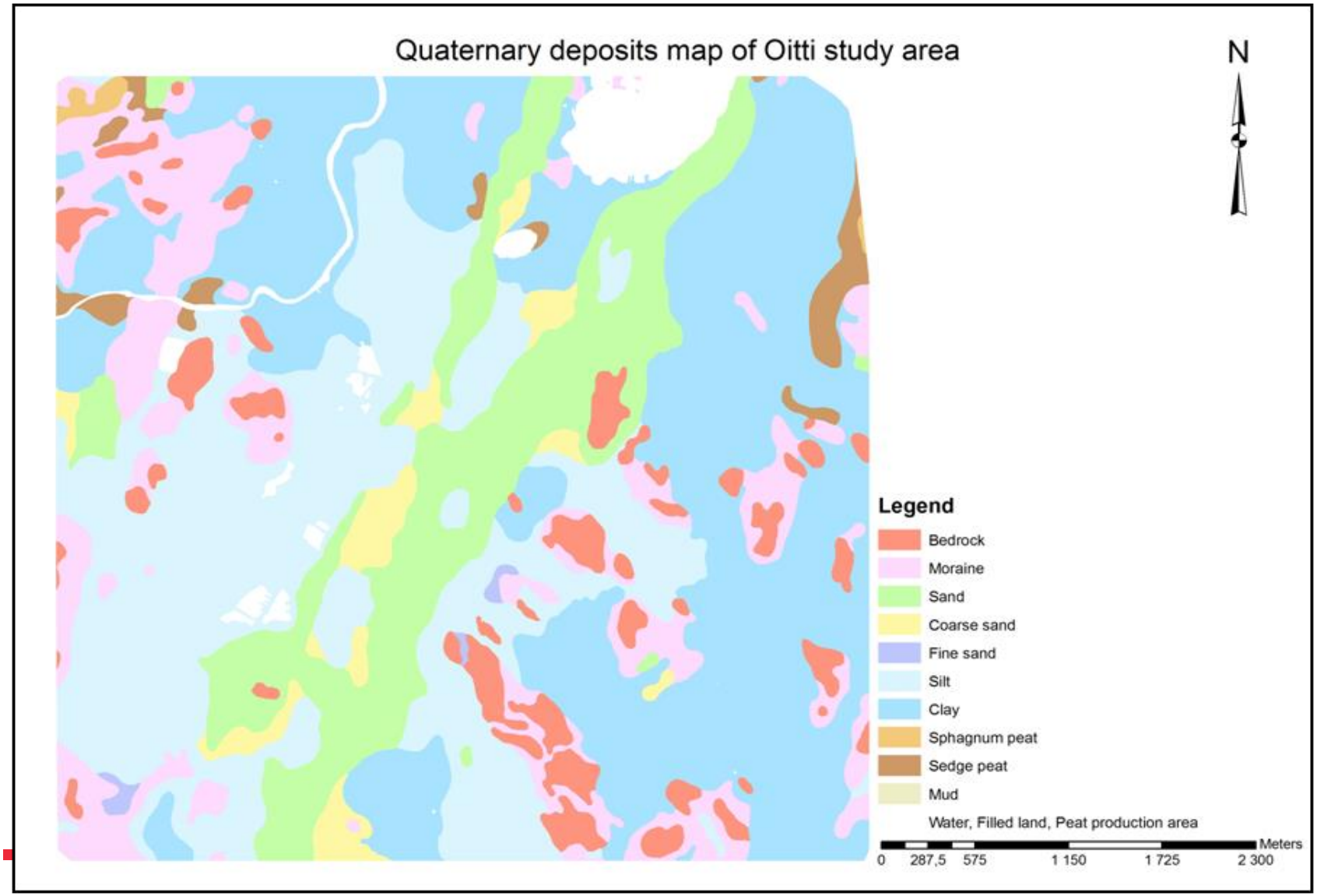
- autocorrelation can be tested by global or local methods
  - global methods: Moran´s I
  - local methods: Local Moran´s I

2. Spatiaalinen autokorrelaatio

- **autokorrelaatio** on matemaattinen menetelmä, 1d- tai 2d-datan keskinäisten riippuvuuksien kuvaamiseen; esimerkiksi **aikasarjat** tai **spatiaalinen tietotaso**

- ilmiönä **spatiaalinen autokorrelaatio** ilmaisee Tobler´in maantieteen ensimmäisen lain
  - **”Everything is related to everything else, but near things are more related to each other.”**

- autokorrelaatiota voidaan mitata globaaleilla mittareilla mutta myös spatiaalisilla lokaaleilla mittareilla
  - globaalit menetelmät: Moran´s I
  - lokaalit menetelmät: Local Moran´s I



Sunila, R., 2009

Autocorrelation as a phenomenon

- spatial dependency in 2d can be
  - equal to all directions or different to the directions
  - homogenous or varying

- if variation is homogeneous, the process in question is **stationary**, if not, then it is **non-stationary**

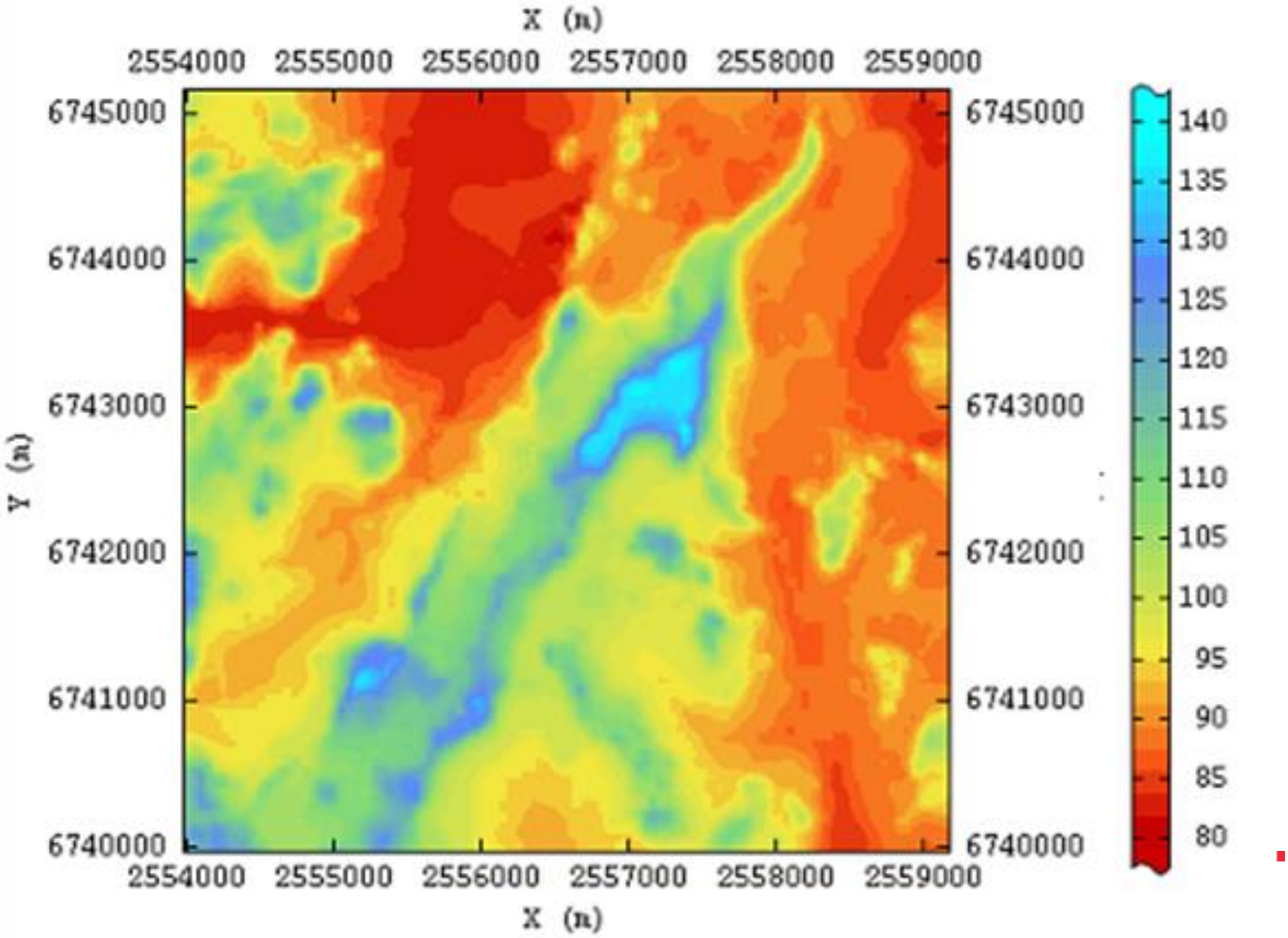
- if the variation in x- and y-directions is equal, then the process is **isotropic**, otherwise it is **anisotropic**

Autokorrelaatio ilmiönä

- riippuvuus 2d tilassa voi olla
  - joka suuntaan samaa tai sitten eri suuntiin erilaista
  - homogeenista tai vaihtelevaa

- jos vaihtelu on homogeenista, kyseessä oleva spatiaalinen prosessi on **stationäärinen**, jos ei, silloin prosessi on **ei-stationäärinen**

- jos vaihtelu x- ja y-suuntiin on samanlaista silloin prosessi on **isotrooppinen**, muutoin se on **aniostrooppinen**



Sunila,R., 2009

Special methods required

- **identifying autocorrelation**
  - answers to question whether there is autocorrelation in the data
    - Moran´s I
    - variogram method
- **modelling autocorrelation**
    - spatial autoregressive models, SAR
      - Idea idea of SAR models is that the when calculating the value oif an area or pixel in grid**, the values of neighbours are taken into account**

Tarvitaan erityisiä menetelmiä

- **autokorrelaation tunnistamisen menetelmät**
  - vastaa kysymykseen onko autokorrelaatiota vai ei
    - Moranin indeksi
    - variogrammi-menetelmä
- **autokorrelaation mallinnus**
    - spatiaalinen autoregressiivinen prosessi/Spatial autoregressive models, SAR
    - perustuu siihen, että alueen tai gridin solun arvoa mallilla laskettaessa esimerkiksi luokittelussa tai erilaisissa simulaatioissa **huomioidaan naapurisolujen arvot**

Identifying spatial autocorrelation in the data

- the most popular test is **Moran´s I**
  - calculates one index value for the whole data set
  - gives the first estimate about the behaviour of the data
- **Local Moran** is an index that calculates a spatially varying value for the spatial autocorrelation
  - more accurate information about the variation

- **variogram**-method reveals even more about the variation of the data
  - strength and spatial extent
  - can be used in interpolation (Kriging)

Spatiaalisen autokorrelaation tunnistaminen datasta

- tunnetuin testi on **Moranin indeksi**
  - laskee yhden tunnusluvun koko aineistolle
  - menetelmällä saadaan ensimmäinen arvio datan jakautumisesta
- **Local Moran** on indeksi, joka laskee autokorrelaation spatiaalisesti vaihtelevana tunnuslukuna
  - saadaan tarkempaa tietoa spatiaalisesta vaihtelusta

- **variogrammipilvi**-menetelmä paljastaa enemmän spatiaalisesta autokorrelaatiosta
  - saadaan tietoa autokorrelaation vahvuudesta ja vaikutusalueesta
  - voidaan käyttää hyödyksi interpoloinnissa (Kriging)

Moran´s I

- spatial form of a non-spatial correlation measure
- fits for objects having numerical attribute data in ratio or interval scales; can be modified for nominal scale as well

- core part: **covariance term**
  - describes the difference of the object values from the mean, by summing up the multiplicated difference of two object values from the mean over the whole area
- **adjacency matrix** is used for eliminating the effect of objects in too long distances (no touch)
- all other terms in the equation just normalize the result (amount of areas and amount of adjacencies, y values)
- **positive value/positive correlation; neg. value/neg.correlation**

Moranin indeksi

- ei-spatiaalisen korrelaatiomittarin spatiaalinen muoto
- sopii numeerista suhde- tai intervalliasteikolla skaalattua ominaisuustietoa omaaville kohteille, myös versio nominaaliasteikolliselle datalle

- keskeinen osa: **kovarianssitermi**
  - kuvaa tutkittavan alueen alkioiden eroa keskiarvosta laskemalla kahden alkion eron tulo yli koko alueen
- **viereisyysmatriisia** käytetään eliminoimaan toisistaan kaukana (ei kosketusta) olevien alkioiden ominaisuuksien vaikutus
- kaikki muut termit kaavassa vain normalisoivat saatua tulosta (alueiden ja viereisyyksien määrä, y:n arvo)
- **positiivinen arvo/positiivinen korrelaatio; neg. arvo/neg.korrelaatio**

Moranin indeksin laskenta

- y= alueen arvo
- w=viereisyysmatriisi (adjacency matrix)
- kovarianssitermi kertoo kuinka muuttujat vaihtelevat yhdessä
- nominaaliarvoisella datalla arvojen erotus korvataan kontingenssimatriisilla

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

Calculating Moran´s I

- y= attribute value of the area
- w=adjacency matrix
- covariance term
- if attribute values are nominal, the difference of values is replaced by contingency matrix

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

Example of the adjacency of areas

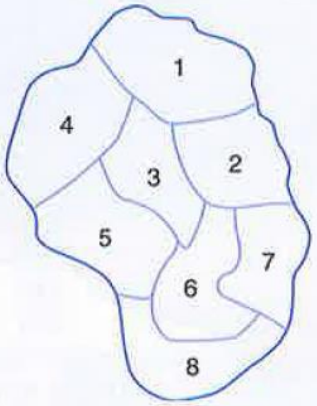


Figure 4.12 A simple mosaic of zones

Table 4.1 The weights matrix **W** derived from the zoning system shown in Figure 4.12

	1	2	3	4	5	6	7	8
1	0	1	1	1	0	0	0	0
2	1	0	1	0	0	1	1	0
3	1	1	0	1	1	1	0	0
4	1	0	1	0	1	0	0	0
5	0	0	1	1	0	1	0	1
6	0	1	1	0	1	0	1	1
7	0	1	0	0	0	1	0	1
8	0	0	0	0	1	1	1	0

adjacency matrix describes the adjacency relations of areas
if adjacent, then 1
if not adjacent , then 0

adjacency here is defined to be
direct touch

(Longley et al., 2005)

Esimerkki alueiden viereisyydestä

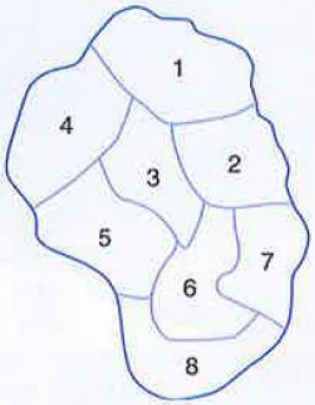


Figure 4.12 A simple mosaic of zones

Table 4.1 The weights matrix **W** derived from the zoning system shown in Figure 4.12

	1	2	3	4	5	6	7	8
1	0	1	1	1	0	0	0	0
2	1	0	1	0	0	1	1	0
3	1	1	0	1	1	1	0	0
4	1	0	1	0	1	0	0	0
5	0	0	1	1	0	1	0	1
6	0	1	1	0	1	0	1	1
7	0	1	0	0	0	1	0	1
8	0	0	0	0	1	1	1	0

viereisyysmatriisi=
adjacency matrix
kuvaa alueiden
viereisyys-
relaatiot
jos viereiset, arvo 1
jos ei viereiset, arvo 0

viereisyys tässä
määritelty **välittömäksi**
kosketukseksi

(Longley et al., 2005)

Interpretation of Moran´s I value

- Moran´s I value
  - positive, when there is positive autocorrelation
  - negative, when there is negative autocorrelation
  - 0 when the data is randomly distributed
- one value for the whole data set

Moranin indeksin arvon tulkinta

- Moranin indeksin arvo
  - positiivinen, kun datassa positiivinen autokorrelaatio
  - negatiivinen, kun datassa negatiivinen autokorrelaatio
  - 0 kun data on satunnaista ja riippumatonta
- yksi arvo koko datalle

Autocorrelated or not ? Positively, negatively? Autokorreloitunutta vai ei? Positiivisesti, negatiivisesti ?

Rook's case

$J_{BB} = 27$

$J_{WW} = 27$

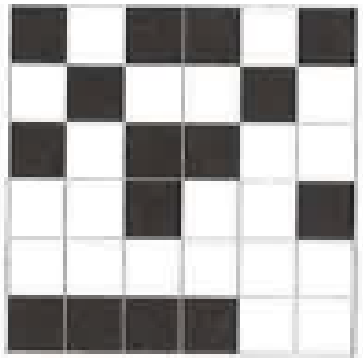
$J_{BW} = 6$

Queen's case

$J_{BB} = 47$

$J_{WW} = 47$

$J_{BW} = 16$



$J_{BB} = 6$

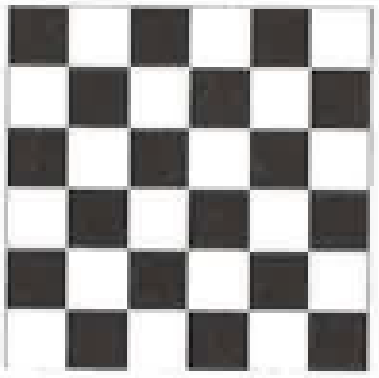
$J_{WW} = 19$

$J_{BW} = 35$

$J_{BB} = 14$

$J_{WW} = 40$

$J_{BW} = 56$



$J_{BB} = 0$

$J_{WW} = 0$

$J_{BW} = 60$

$J_{BB} = 25$

$J_{WW} = 25$

$J_{BW} = 60$

O´Sullivan&Unwin, 2003

Variogram cloud

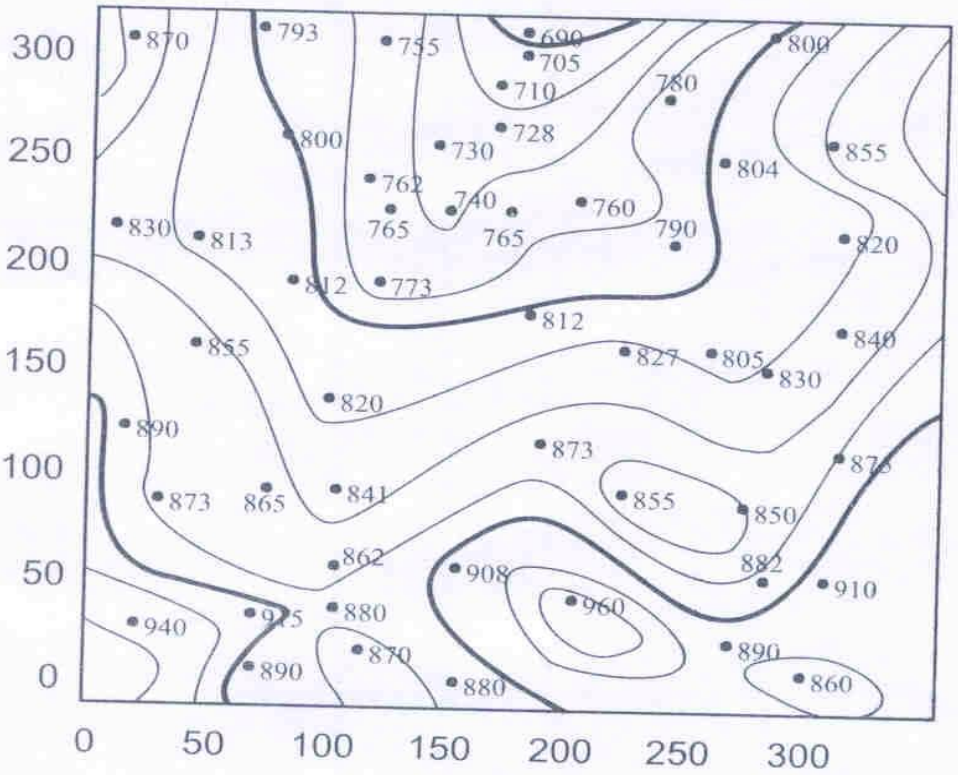
- describes the relation between the location and attribute values of the points in the data set
- example is from the book: O'Sullivan&Unwin: Geographic Information Analysis

  - **contour pesentation** of the topography of the terrain
  - same information contents as a **variogram cloud**:
    - the differences of **all possible point pairs** in the data set is processed and presented as a graphic;
    - x-axis is **the distance between the points** and in
    - y-axis **the square root of the difference of attribute values of the points** is presented
  - interpretation of the variogram

Variogrammipilvi

- kuvaa sijainnin ja ominaisuustiedon välistä suhdetta
- esimerkki on kirjasta: O'Sullivan&Unwin: Geographic Information Analysis

  - **korkeuskäyräesitys** maaston pinnanmuodoista
  - sama tieto **variogrammipilvenä**: kaikkien mahdollisten pisteparien etäisyys (x-akseli) ja pisteiden ominaisuustietoerotuksen neliöjuuri (y-akseli)
  - variogrammin tulkinta



.6 Spot heights and their contour pattern. Note that this contour pattern is ation only and was done by hand.

(O´Sullivan &Unwin, 2003)

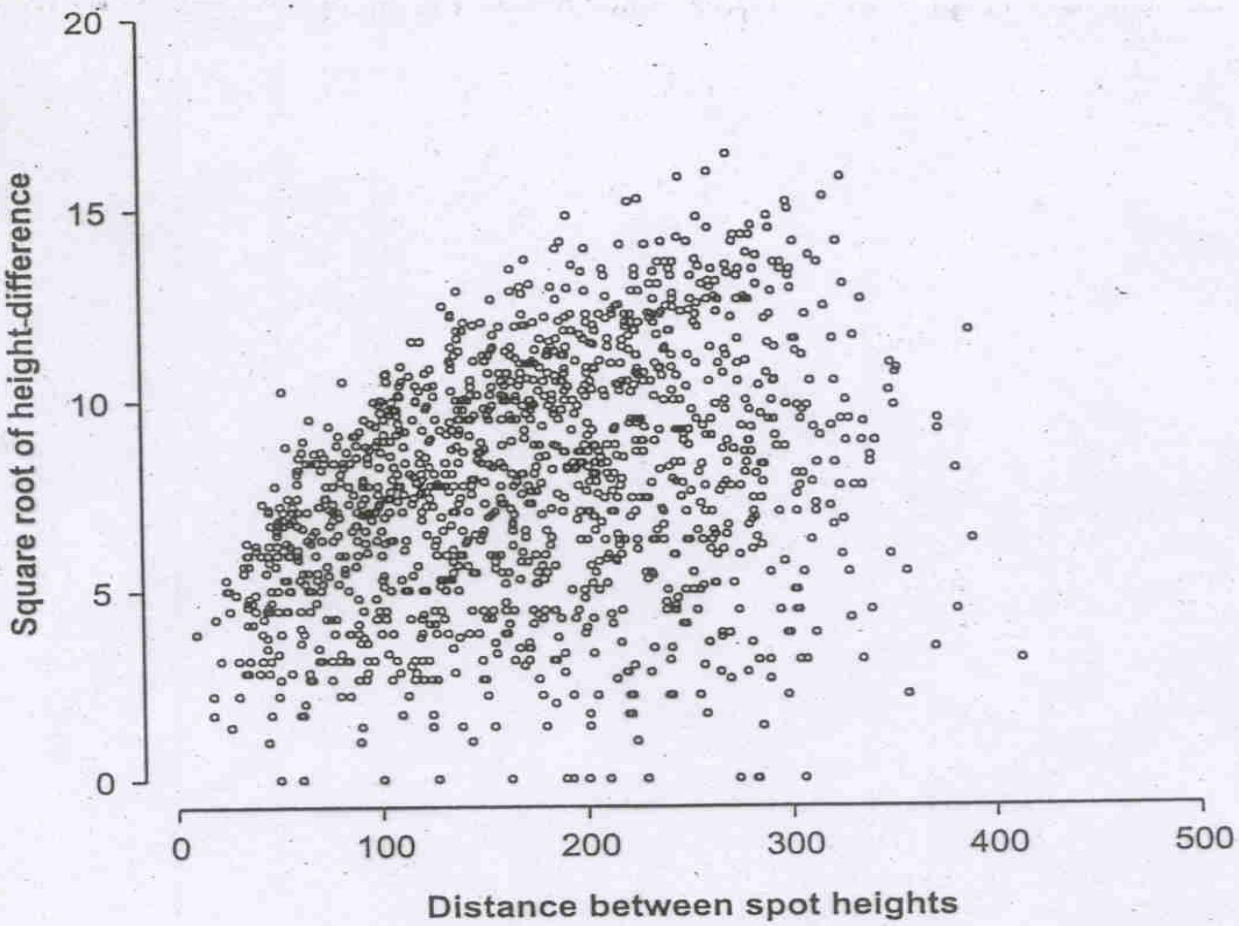


Figure 2.7 Variogram cloud for the spot height data in Figure 2.6.

(O´Sullivan &Unwin, 2003)

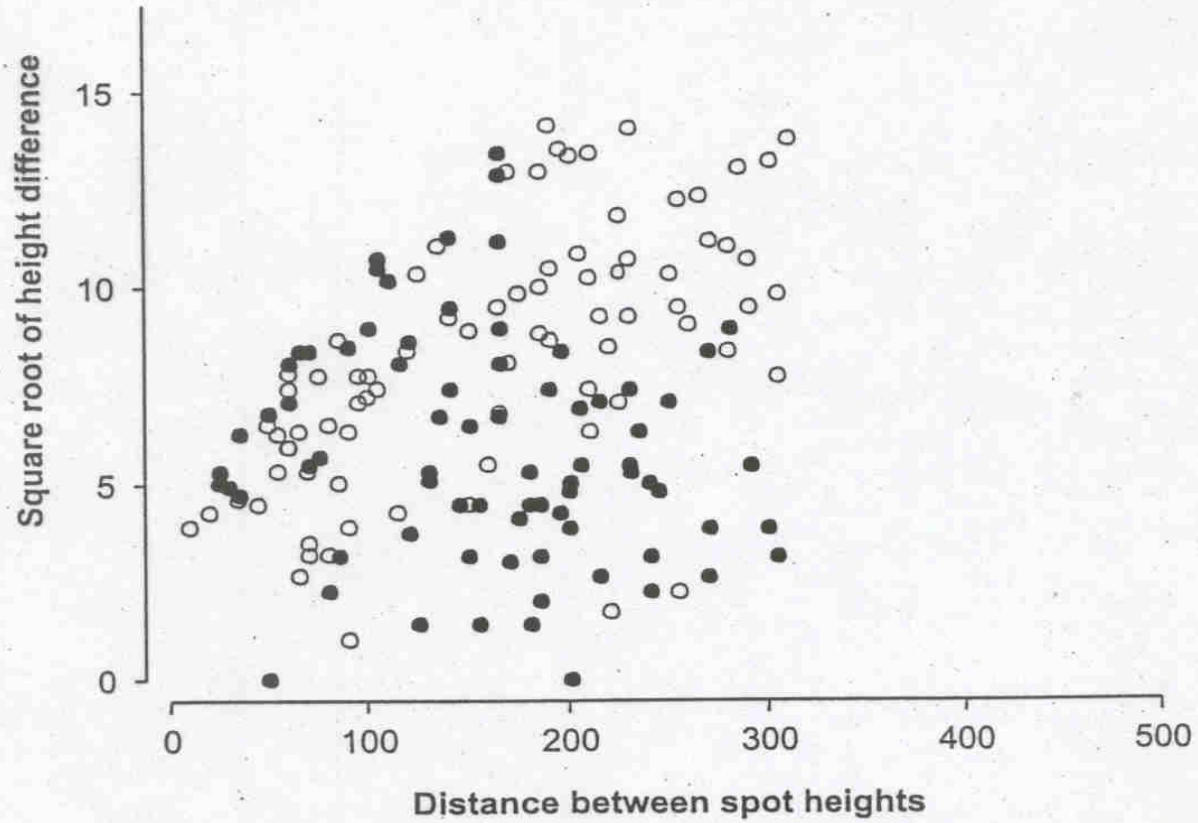


Figure 2.8 Variogram clouds for N–S oriented pairs in Figure 2.6 (open circles), and for E–W oriented pairs (filled circles).

(O´Sullivan &Unwin, 2003)

Interpretation of the variogram

- East-West and North-South directions are analysed
  - in North-South direction some behaviour can be recognized
  - also the height differences between North-South point pairs are bigger than the corresponding values in East-West direction; fits with the contour presentation
  - this phenomenon is called as anisotrophy; there is a direction-related behaviour in the data set values

Variogrammin tulkintaa

- analysoidaan itä-länsi -suuntaa ja pohjois-etelä -suuntaa
  - havaittavissa pohjois-etelä -suunnassa tietty käyttäytyminen
  - nähdään myös, että pohjois-etelä -suunnassa olevien pisteparien korkeuserot ovat suurempia kuin itä-länsi -suunnassa olevien pisteparien; vastaa korkeuskäyräesitystä
  - tätä ilmiötä nimitetään anisotropiaksi; datan spatiaalisessa vaihtelussa on suuntaan liittyvä piirre

4. Basic tasks in spatial statistics

- we are interested in **the spatio-temporal distribution of the data set**
- we are interested **in possible correlations and other dependencies** (correlation = dependency between two variables shown by using statistical methods)

- **autocorrelation, clustering, hotspots and outliers** are analysed
- data sets are analysed together, dependencies

- we try to discover typical behaviour of the data sets, so that **models** can be built
- models are used for **predicting the future** and identifying **anomalies** (=situations when the data is not behaving as the model)

4. Spatiaalisen analyysin perustehtävät

- ollaan kiinnostuneita ilmiön **spatio-temporaalisesta jakautumisesta**
- ollaan kiinnostuneita mahdollisista **korrelaatioista ja muista riippuvuuksista**

- tutkitaan aineiston **autokorrelaatiota, klusteroitumistaipumusta, hotspotteja, outliereitä**
- tutkitaan **aineistoja yhdessä, keskinäisiä suhteita**
- pyritään löytämään **lainalaisuuksia**, jotta voitaisiin laatia **malleja**
- malleilla halutaan **ennustaa tulevaa** ja tunnistaa **anomalioita**

Autocorrelation, hotspot, cluster ?

- **autocorrelation**: observations that have similar attribute values are also close to each othes in spatial location; autocorrelation is identified by analysing how observation values change together
- **hotspot**: the density of similar observations is high (vs. cold spot)
- **cluster**: observations /objects are spatially close to each others in attribute space; identification is based on calculation of the distances between objects in the attribute space
- **heterogeneity vs. homogeneity**: spatial data if often distributed in a heterogeneous way, the question is not always on autocorrelation

Autokorrelaatio, hotspot, klusteri ?

- **autokorrelaatio, itsekorrelaatio**: samaa ominaisuuden arvoa omaavat havainnot ovat maantieteellisesti lähekkäin; tunnistaminen havaintojen yhteisvaihteluun perustuen

- **hotspot/kuuma piste**: samanlaisten havaintojen/objektien tiheys on suuri (vs. cold spot)

- **klusteri**: havainnot/objektit ovat lähellä toisiaan ominaisuustietoavaruudessa, tunnistaminen perustuu samanlaisuuden/etäisyyden laskemiseen; menetelmä riippuu ominaisuuksien määrästä (vain tyyppi, yksi ominaisuus, useita ominaisuuksia)
- **heterogeenisyys vs. homogeenisuus**

Distance and density as basic measures in the methods

- in the previous methods the calculations were based on the distances between observations/objects

- In spatial analytics methods spatial character can be present in two measures
  - **distance** and
  - **density**

- methods measure the same phenomenon, how observations/objects are located in relation to each others, **spatial distribution** either
  - uniformly, randomly or clustering

Etäisyys ja tiheys menetelmien perustana

- edellä esitetyissä menetelmissä laskennan perustana on **kohteiden välinen etäisyys**

- Spatiaalisuus voidaan saada menetelmiin kahdella eri tavalla:
  - mitataan etäisyyksiä
  - mitataan tiheyttä

- menetelmät mittaavat samaa asiaa, kohteiden sijoittumista toisiinsa nähden
  - tunnistavat kohteiden jakautumisen, joko tasaisesti, satunnaisesti tai klusteroituneesti

Examples on density and distance based analysis methods

- density based methods for analysis of univariate data
  - **Kernel –density estimate**
    - visual method for identifying **hotspots/clustering** in the data set
  - **quadrat method**; comparing the observed empirical data on quadrats and the mathematical model (assumed distribution)
    - simple quadrat method: VMR (variance /mean ratio) calculation; if VMR > 1, tendency for **clustering**
- distance based methods for analysis of univariate/bivariate data sets
  - **G-function** for identifying spatial **clustering**

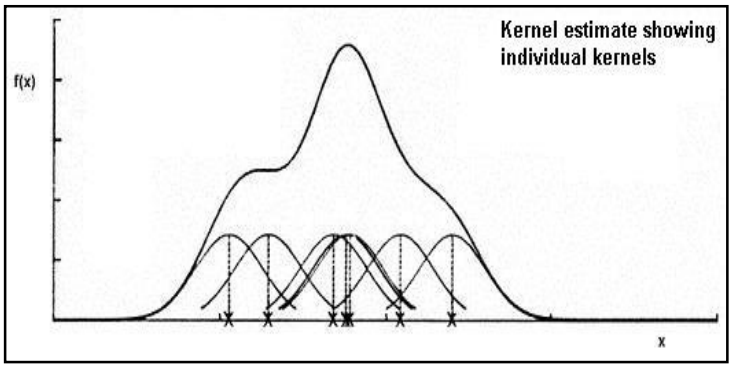
Esimerkkejä tiheys- ja etäisyys-perusteisista menetelmistä

- yhden muuttujan analyysi tiheysmenetelmällä
  - **Kernel –tiheysestimaatti**
    - visuaalinen menetelmä **hotspottien** tunnistamiseen
  - **koealamenetelmä**; verrataan koealoista kerättyä empiiristä dataa matemaattiseen malliin
    - yksinertainen koealamenetelmä: VMR (variance/mean ratio) laskenta; jos VMR > 1, klusteroitumistaipumus
- yhden muuttujan/kahden muuttujan analyysi etäisyysperusteisella menetelmällä
  - **G-funktio** spatiaalisen klusteroitumisen tunnistamiseen

Kernel-menetelmän käyttö/ Use of Kernel density method

- Kernel menetelmä
- Kernel method

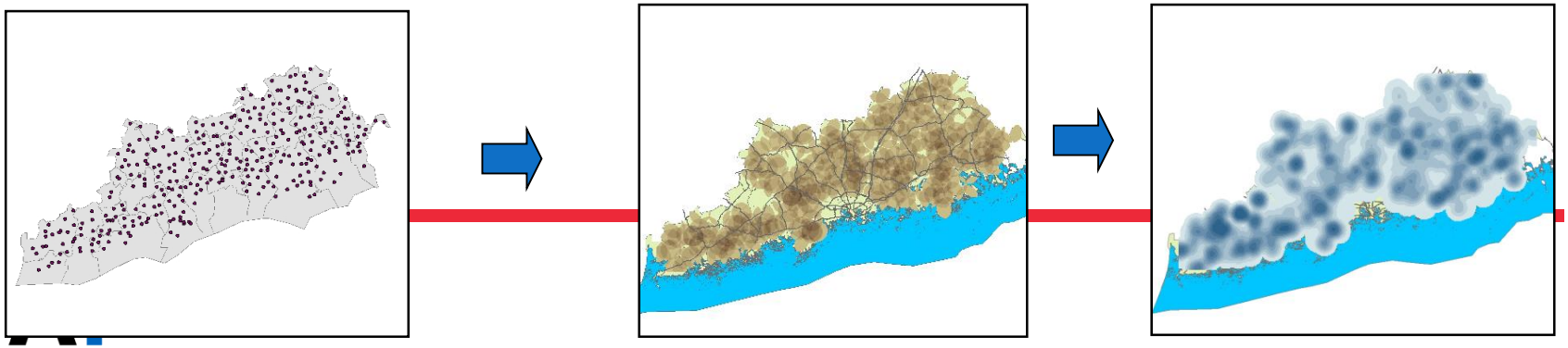
(Krisp,2006)



- yksittäisistä havainnoista tiheyspinnaksi/ from discrete observation data to a continuous density surface

yksittäiset havainnot

tiheyspinta

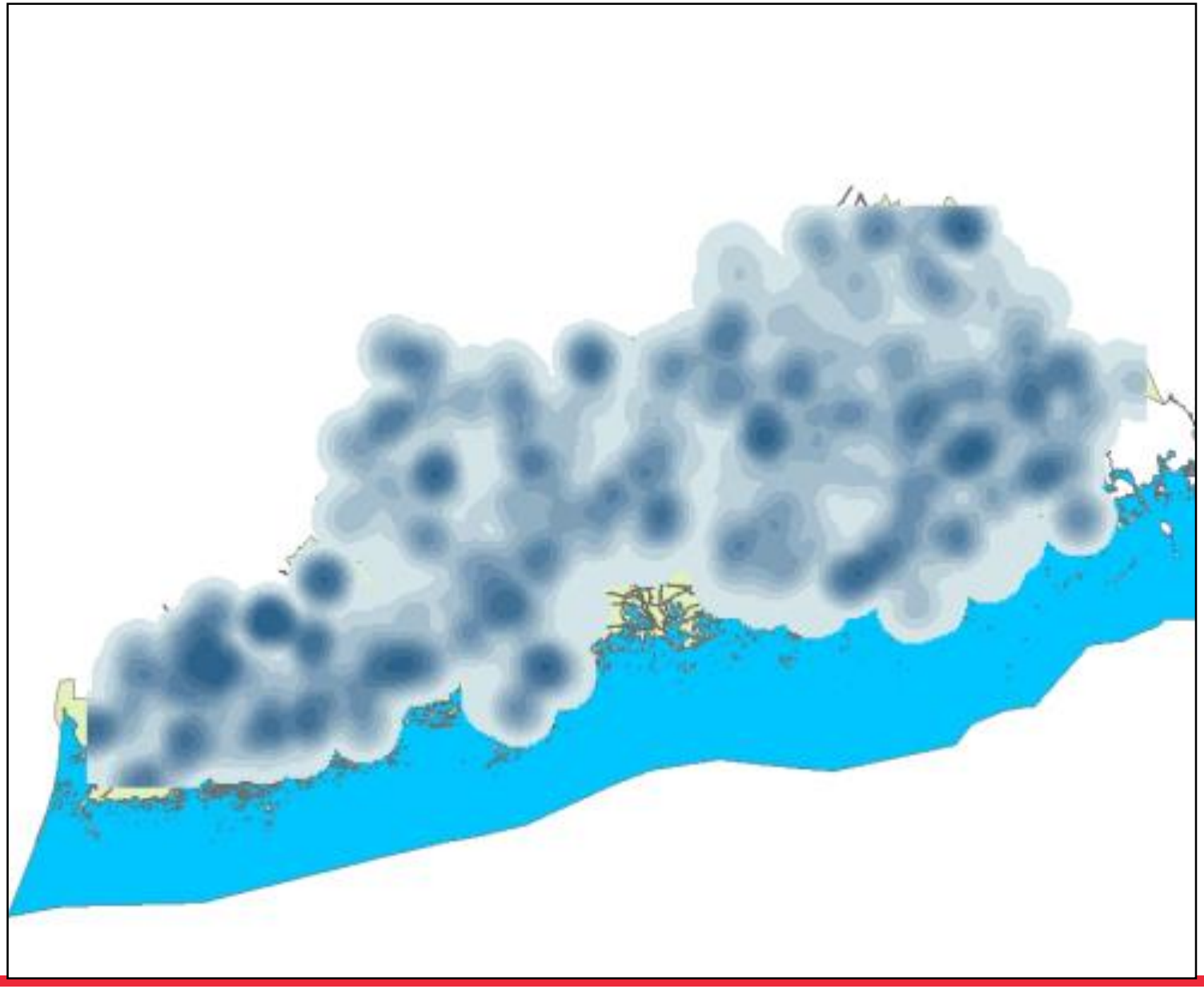


Kernel -density estimation

- weighting the observations by a mathematical model
  - k, **2-dimensional density function** (kernel) is selected, for weighting the observations (points); in the center point weight is 1, on the edge weight is 0
  - passing through each point and summing up in each location the values of overlapping kernel functions the density durface is created
  - **bandwidth** defines the area in which the function reaches; the suitable range is decided by the user; if it is too small a lot of details are created, if it is too big the results is ”flat”
  - in correct use the result is a surface with dark areas among the light background

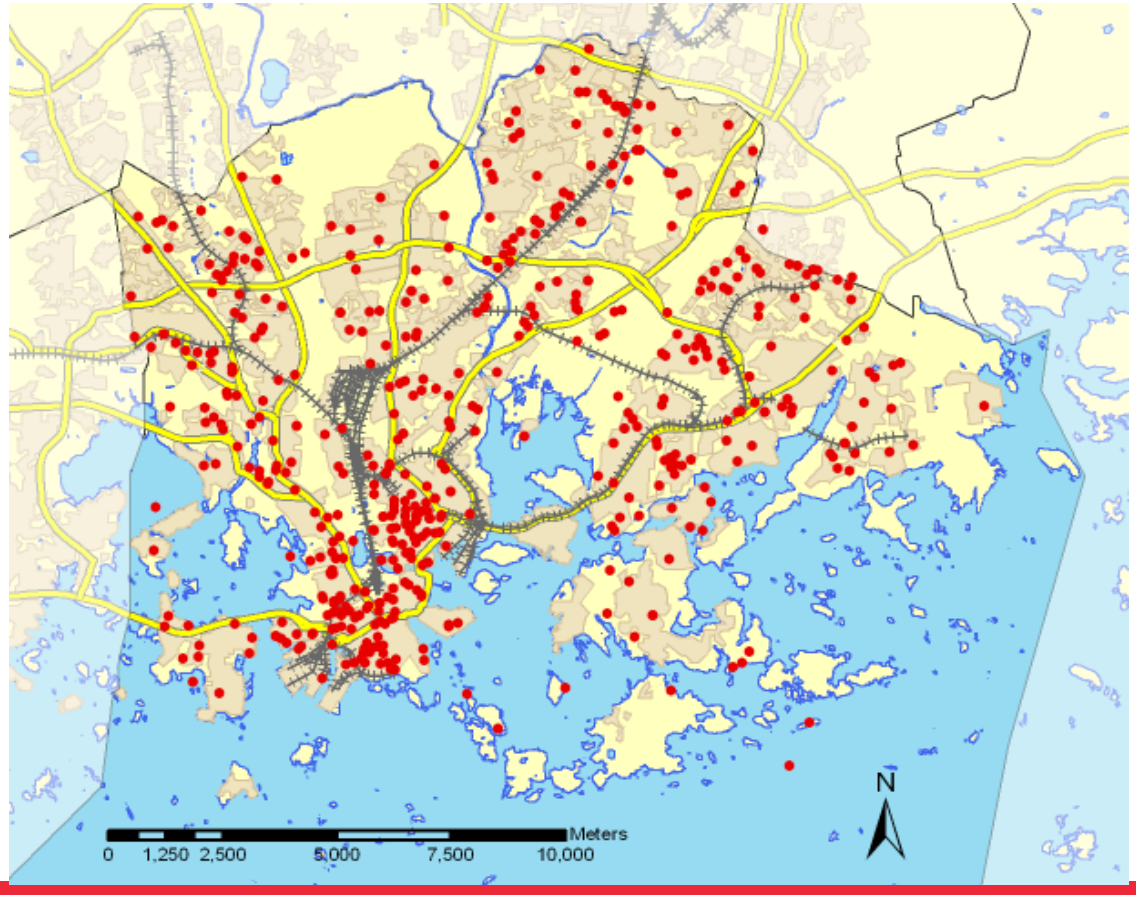
Kernel-tiheysestimointi

- käytetään matemaattista funktiota painotukseen
  - valitaan k, 2-ulotteinen **tiheysfunktio** (kernel, ydin), jonka avulla **pisteet painotetaan**, keskipisteessä paino max, reunalla 0;
  - kuljetaan jokaisen pisteen kautta ja **summataan pisteen arvoksi**
  - **bandwidth** (ytimen leveys) määrittää alueen, jolle funktio ulottuu; haettava sopiva leveys, kun b kasvaa tulos on ”litteä”, kun b on pieni paljon detaljeja
  - syntyy ”tummia” alueita kun pisteet klusteroituu



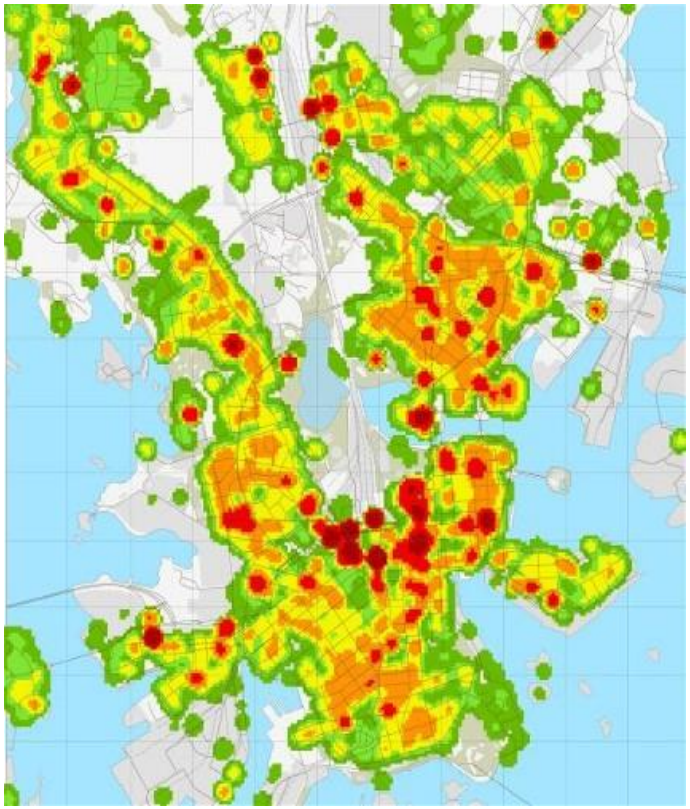
Example of analysis of fire and rescue data: building fires in Helsinki area
The goal is to find causes of building fires

Building fires data set



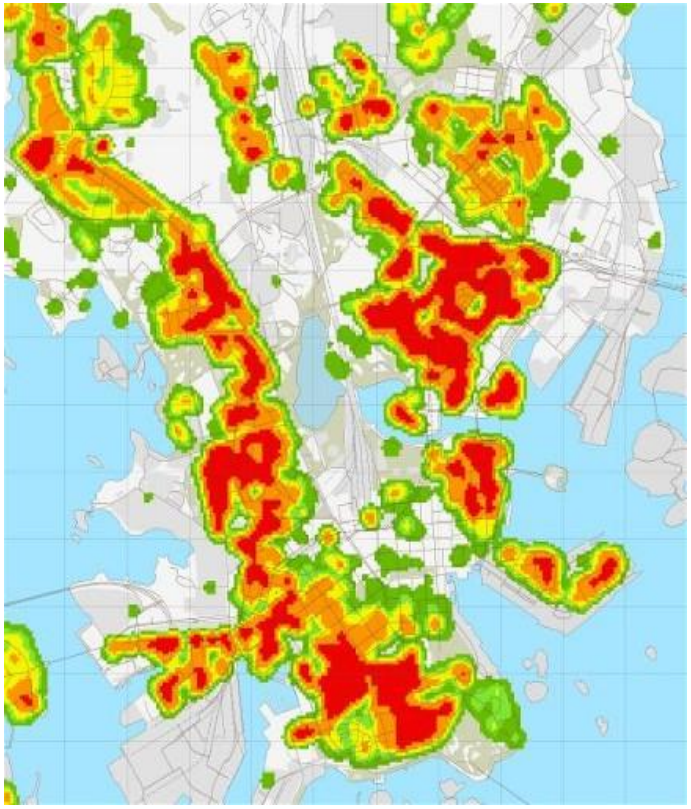
First the Kernel densities are calculated, during the day and night separately

Päivä/Daytime



Lasketut Kernel tiheydet rakennus-paloista päivä- ja yö-aikaan

Yö/Nighttime



200 - 3 000
3 000 - 6 000
6 000 - 12 000
12 000 - 24 000
24 000 - 65 000
65 000 - 300 000

Distance based cumulative frequency graphs for cluster identification

Question: how does the curve look like when points are very clustered?

$$G(d) = \frac{\text{no. } [d_{\min}(\mathbf{s}_i) < d]}{n} \qquad (4.7$$

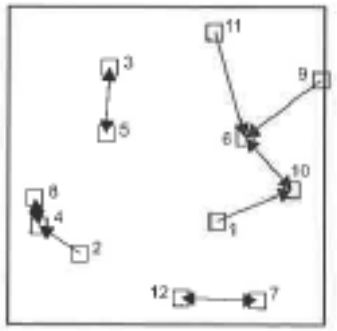


Figure 4.9 Distances to nearest neighbor for a small point pattern. The nearest neighbor to each event lies in the direction of the arrow pointing *away* from it.

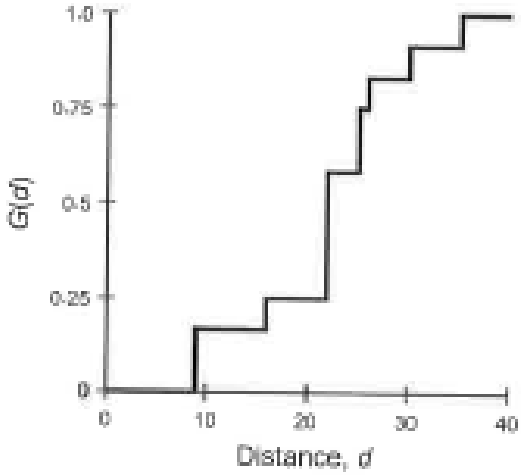


Figure 4.10 *G* function for the point pattern of Figure 4.9 and Table 4.2.

O´Sullivan&Unwin, 2003

Kumulatiivinen frekvenssikäyrä etäisyyksistä lähimpään naapuriin

Kysymys: millainen käyrä syntyy, kun pisteet ovat klusteroituneet ?

$$G(d) = \frac{\text{no. } [d_{\min}(\mathbf{s}_i) < d]}{n} \qquad (4.7$$

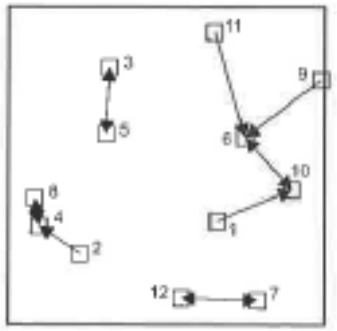


Figure 4.9 Distances to nearest neighbor for a small point pattern. The nearest neighbor to each event lies in the direction of the arrow pointing *away* from it.

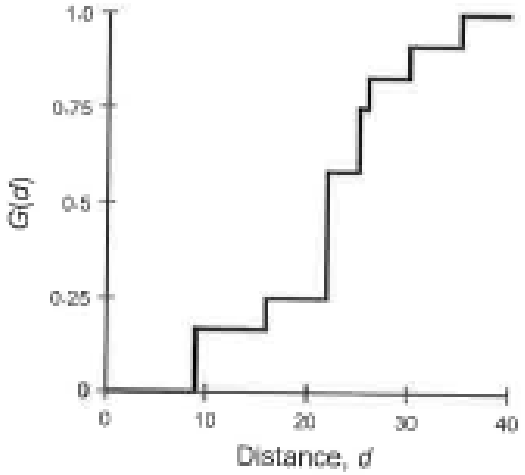


Figure 4.10 *G* function for the point pattern of Figure 4.9 and Table 4.2.

O´Sullivan&Unwin, 2003

Empirical and simulated G-function

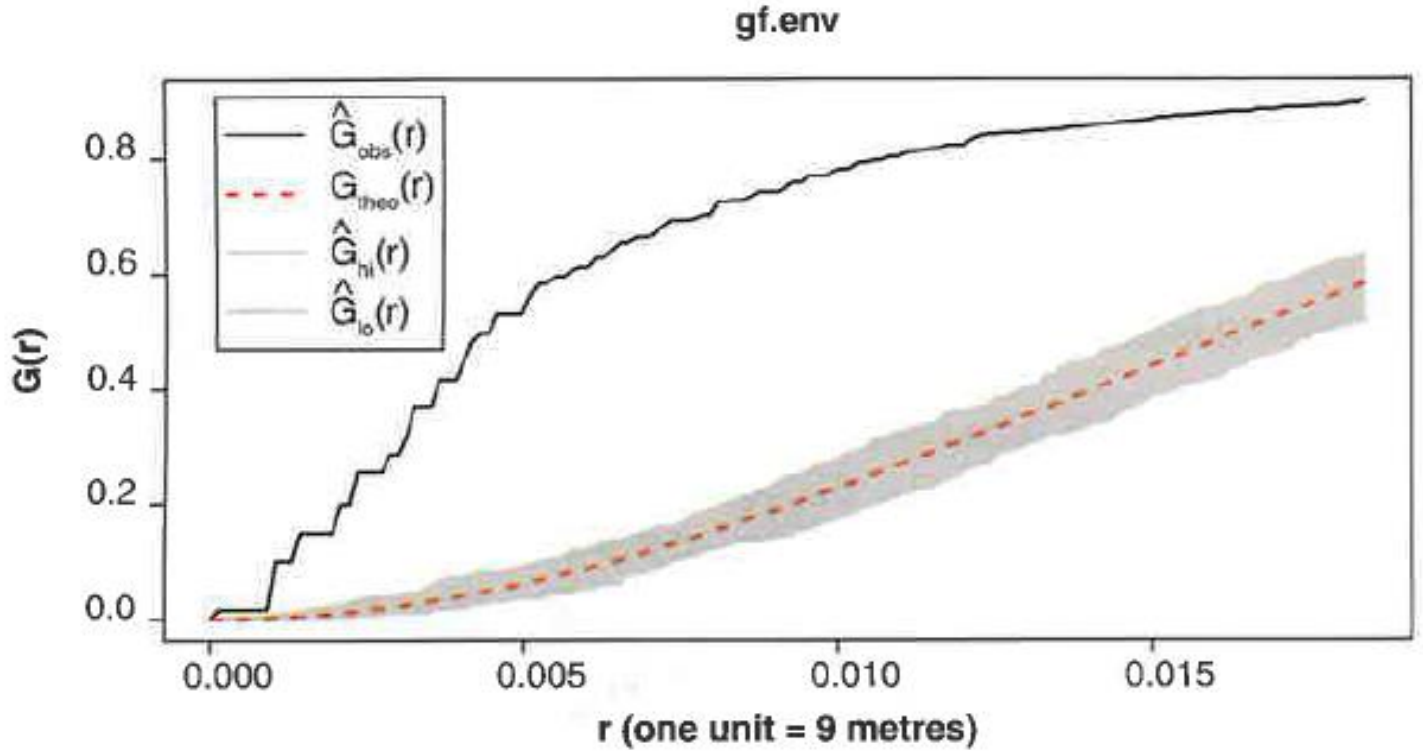


Figure 6.14 *G*-function with envelope

Brundson&Comber,2015

Empiirinen ja simuloitu G-funktio

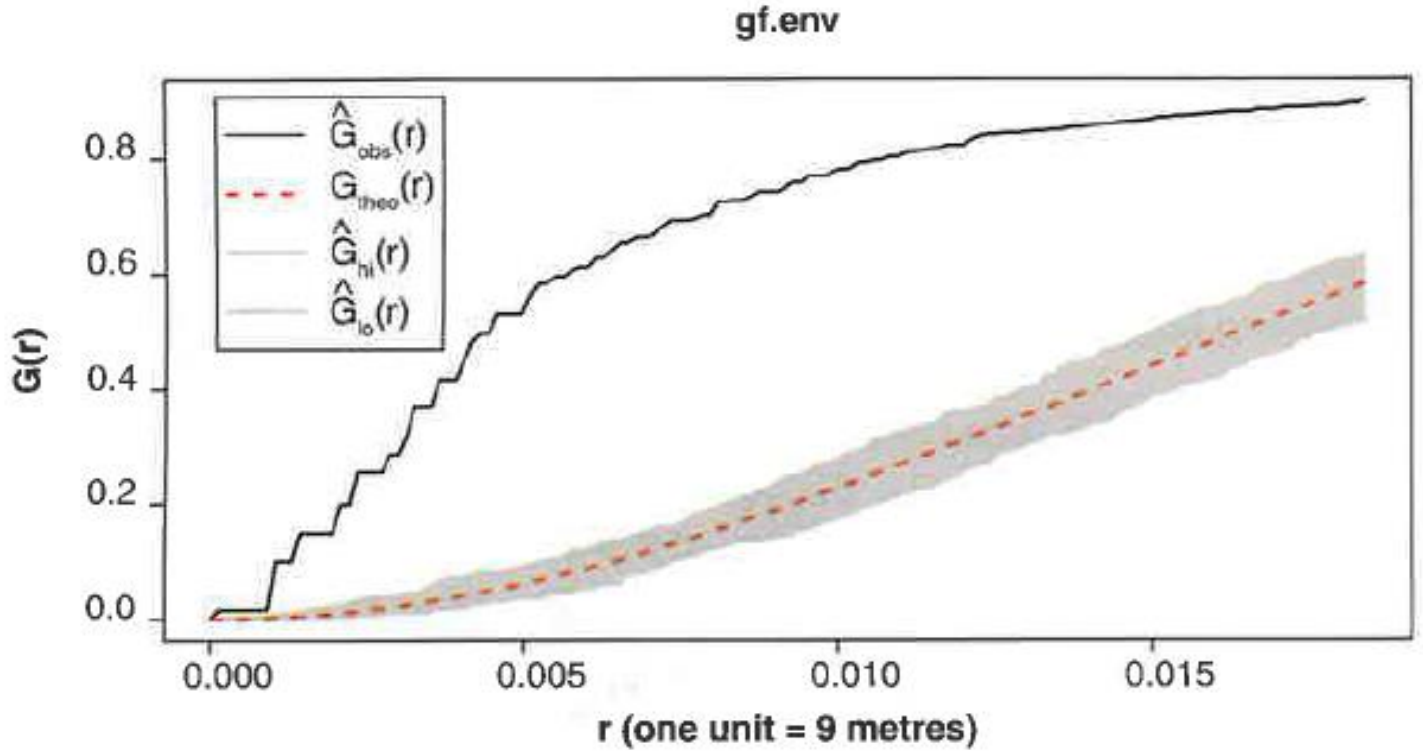


Figure 6.14 *G*-function with envelope

Brundson&Comber,2015

Complete spatial randomness – Independent random process

- because of autocorrelation we can not assume
  - **complete spatial randomness (CSR) of things**

- spatial processes are not always
  - **independent random processes (IRP)**

Täydellinen spatiaalinen satunnaisuus - Riippumaton satunnaisprosessi

- autokorrelaatio aiheuttaa, että emme voi suoraan olettaa
  - **täydellistä spatiaalista satunnaisuutta (Complete spatial randomness), CSR**

- spatiaaliset prosessit eivät ole aina
  - **riippumattomia satunnaisprosesseja (Independent random processes), IRP**

Independent random process

- independent random process (IRP), complete spatial randomness (CSR), mean

  - **condition of equal probability**
    - each point has equal probability to have any location
  - **condition of indenpendency**
    - location of a point is independent of location of any other point

  - independent random process is the **hypothesis** that is used when the spatial distribution of a data set is analysed
  - as we did in the previous example

Riippumaton satunnaisprosessi

- riippumaton satunnaisprosessi, independent random process (IRP), täydellinen spatiaalinen satunnaisuus, complete spatial randomness (CSR), tarkoittaa
  - **yhtäsuuren todennäköisyyden ehto**
    - jokaisella pisteellä on yhtä suuri todennäköisyys sijaita missä tahansa paikassa
  - **riippumattomuusehto**
    - pisteen sijainti on riippumaton toisten pisteiden sijainnista
  - riippumaton satunnaisprosessi on se **hypoteesi** jota vasten empiiristä dataa testataan **spatiaalisen jakautumisen** analysoimiseksi
  - näin tehtiin mm. edellisessä esimerkissä

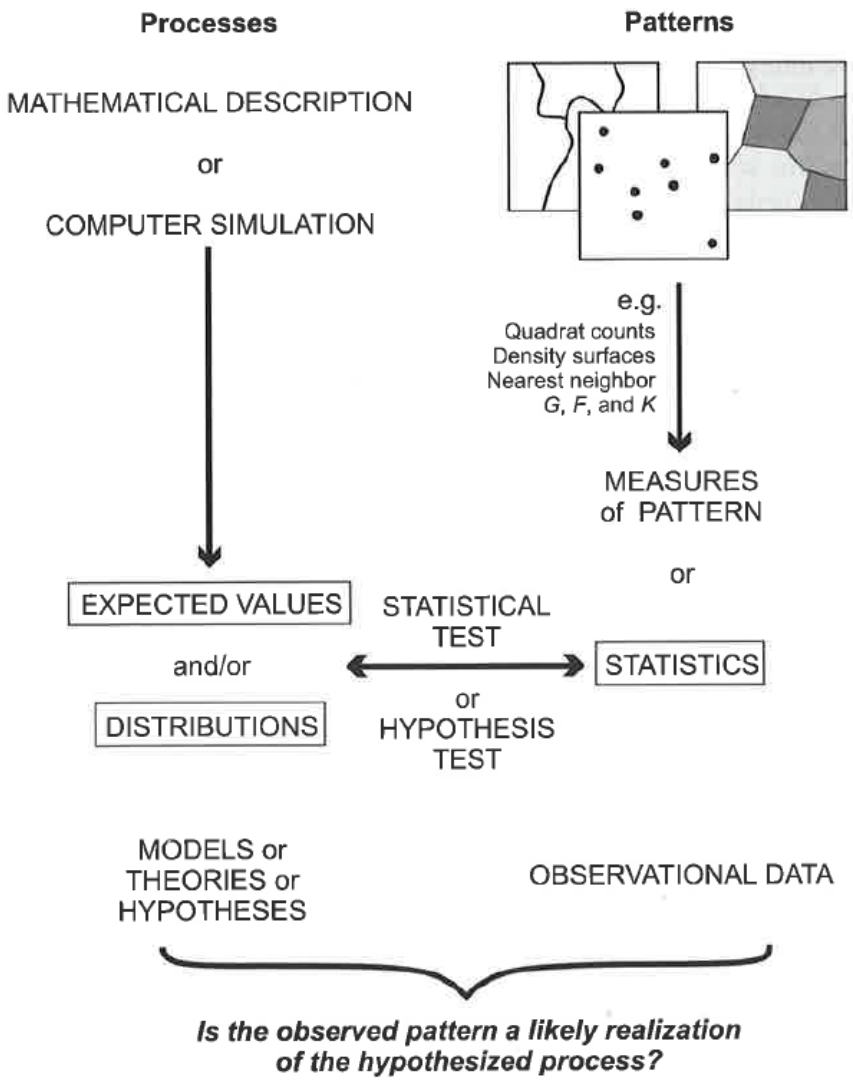


Figure 4.16 Conceptual framework for the statistical approach to spatial analysis.

O´Sullivan&Unwin, 2003

5. Spatial data in models

- as a result of analysis similar distribution (dependency) can be found in two data sets
- a **hypothesis on spatial correlation** can be derived and it can be tested by statistical methods
- the goal can be a model that can be used for **prediction** of unknown events by using the known events

- for example building fires could be predicted by the known data on population density or other variables like attributes of the buildings

- **regression models** is one way of predicting

5. Spatiaalisen datan käyttö mallinnuksessa

- analyysin tuloksena voidaan tunnistaa useissa datatyypeissä samanlaista spatiaalista jakautumista
- tästä voidaan johtaa hypoteesi spatiaalisesta korrelaatiosta ja sitä voidaan analysoida edellä mainituilla menetelmillä
- tavoitteena voi olla malli, jolla voidaan ennustaa tuntemattomia asioita tunnettujen asioiden perusteella

- esimerkiksi rakennuspaloja voitaisiin ennustaa tiettyjen väestön ominaisuuksien tai rakennusten ominaisuuksien perusteella

- eräs tapa on muodostaa **regressiomalli**

Linear regression model

- regression model is based on the correlation relationship between the known and unknown variables (dependent and independent)
  - for example among a group of schoolchildren the age of the child explains quite well the length of the child of his/her knowledge on multiplication table
  - mathematical form of regression model – simple linear regression model is as below; Y=unknown, the variable we want to predict; X=the known variable

$$Y = a + bX$$

Regressiomalli

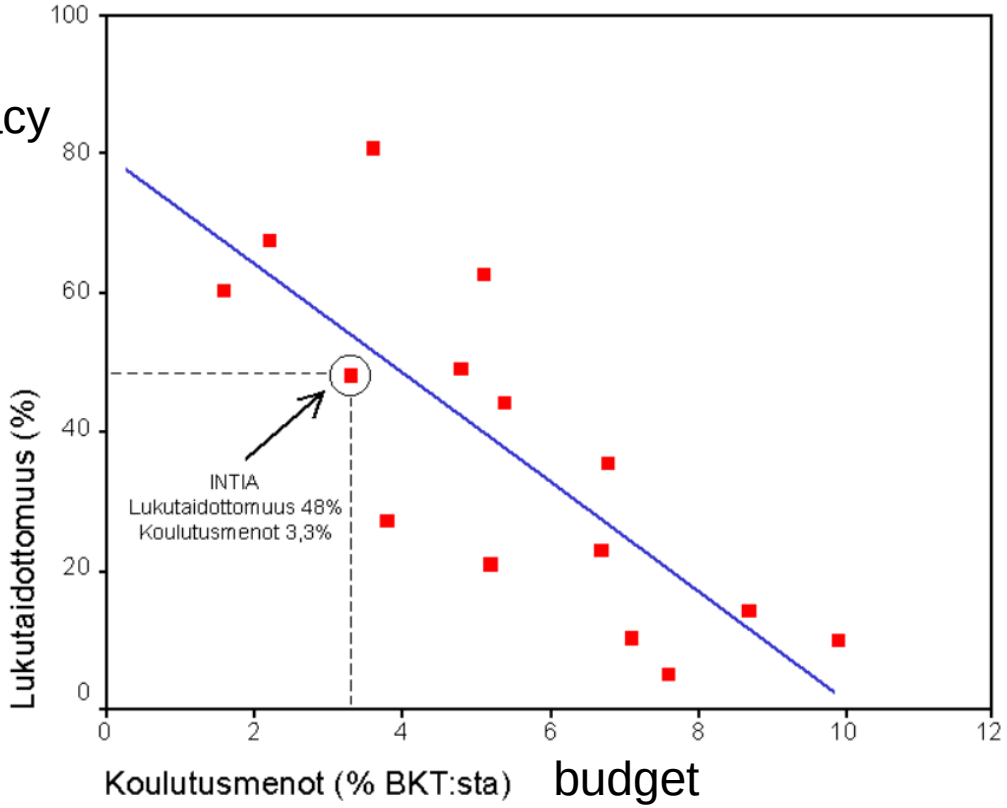
- regressiomalli perustuu tunnettujen ja tuntemattoman muuttujan (selittävät ja selitettävät) väliseen korrelaatiosuhteeseen
  - esimerkiksi peruskoululaisten joukossa oppilaan ikä selittää melko hyvin oppilaan pituutta tai sitä kuinka hyvin hän osaa kertotaulun

- regressiomallin matemaattinen esitys – yksinkertainen lineaarinen regressiomalli on alla olevaa muotoa; Y=tuntematon, jota halutaan ennustaa; X=tunnettu muuttuja

$$Y = a + bX$$

Esimerkki regressiosuorasta – lineaarinen regressiomalli; Example

illiteracy



budget

Kuinka eri valtioiden koulutusbudjetin koko selittää lukutaidottomuusprosenttia

Suhde tässä melko selvä ja lineaarinen

Kaava

$$Y = a + bX$$

How the size of budget on education explains the illiteracy
Linear relationship

Spatial regression models

- spatial phenomena can also be derived into regression models
- the problem is thet the relationship between variables is not always as straightforward as in non-spatial data

- because of spatial autocorrelation the relationship between variables is also spatially varying
  - in clusters the relationship is stronger
  - in empty areas the relationship is weaker

- spatial autocorrelation/spatial heterogeneity must be added to the spatial regression model

Spatiaaliset regressiomallit

- myös spatiaalisista ilmiöistä voidaan muodostaa vastaavia regressiomalleja
- ongelmana on se, että aivan yhtä suoraviivaista selityssuhdetta ei aina löydy
- spatiaalinen autokorrelaatio aiheuttaa sen, että selityssuhde vaihtelee alueittain
  - klustereissa selitys on vahvempaa
  - tyhjillä alueilla selitys on heikompaa

- spatiaaliseen regressiomalliin tulee ottaa mukaan spatiaalinen autokorrelaatio/spatiaalinen heterogeenisyys

Spatial autocorrelation and heterogeneity in regression model

- there are many ways if including spatial autocorrelation/heterogeneity in models
- adjacency matrix W can be used
  - same matrix that was used in Moran´s I
- there are also methods like GWR (Geographically weighted regression), that takes heterogeneity into account by using kernel;

Spatiaalinen autokorrelaatio regressiomallissa

- spatiaalinen autokorrelaatio/heterogeenisyys voidaan huomioida regressiomallissa (ja muissakin malleissa) useilla eri tavoilla

- klusteroitumistaipumusta voidaan kuvata esimerkiksi viereisyysmatriisilla W
  - tämä on sama matriisi, jota käytettiin autokorrelaation tunnistamisen menetelmisssä (Moranin indeksi)

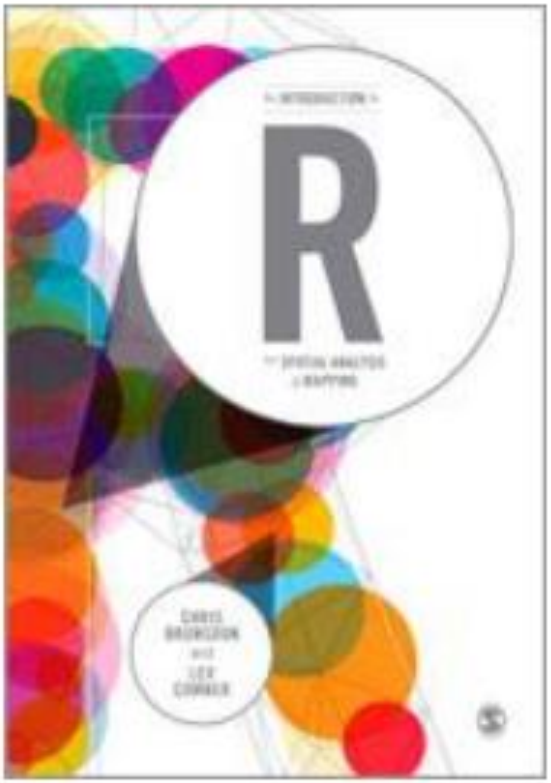
- on myös menetelmiä, kuten GWR (Geographically weighted regression), joka huomioi heterogeenisyyden käyttämällä kerneliä (vrt Kernel-estimointi); syntyy useita malleja yhden sijaan

Literature, references

- Rogerson, P.A., Statistical methods for geography, A Student´s Guide, 2015
- Longley,P., Goodchild,M., Maguire,D., Rhind,D., Geographic information systems and science.
- Doctoral dissertations:
  - Sunila,R., Spatial data modelling using fuzzy and geostatistical applications, 2009.
  - Krisp,J., Geovisualization and knowledge discovery for decision-making in ecological network planning, 2006.
  - Spatenkova,O., Discovering spatio-temporal relationships: A case study of risk modelling of domestic fires, 2009.
- O´Sullivan,D., Unwin,D., Geographic information analysis, 2003.
- Brundson,C., Comber,L:, An introduction to R for spatial analysis and mapping, 2015.

- Reading for the exam:
  - In MyCourses you will find few copied pages from the books above.

If you want to program yourself



An Introduction to R for Spatial Analysis and Mapping

Chris Brunsdon - National University of Ireland, Maynooth
Lex Comber - University of Leeds, UK