Applied spatial data analysis

Jana Faganeli Pucer

University of Ljubljana, Faculty of Computer and Information Science

June, 2019

Spatial data analysis

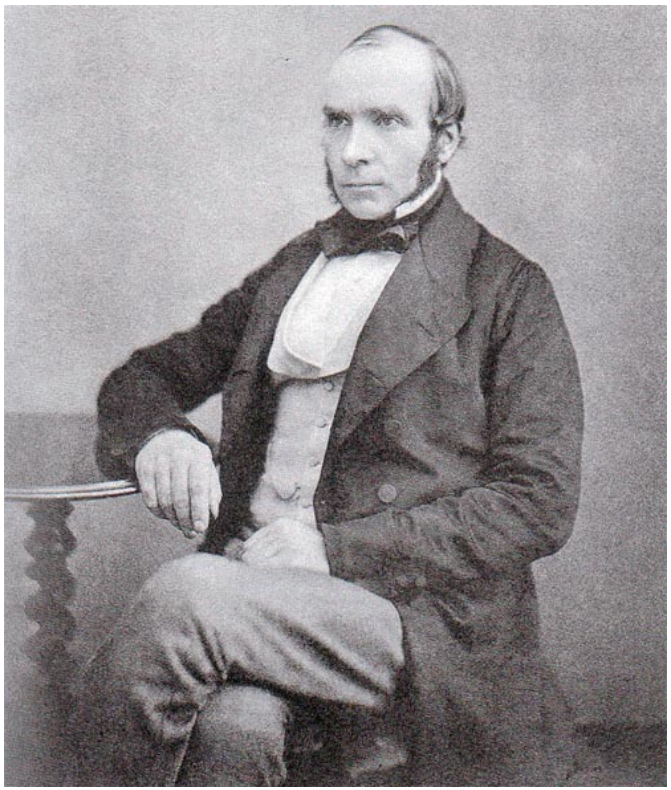
We are dealing with data $Z(s_1), \ldots, Z(s_n)$ observed at different locations $s_1, s_2, \ldots, s_n$
Usually the data are assumed random. Different types of spatial data:

- geostatistics
- spatial point patterns
- spatial process indexed over lattices

Spatial data are usually spatially dependent -> **spatial autocorrelation** (violation of standard statistics) Spatial regression models exploite this feature
Spatial autocorrelation: Values of a random variable, at paired points, are more or less similar as a function of the distance between them.

Beginnings of spatial analysis

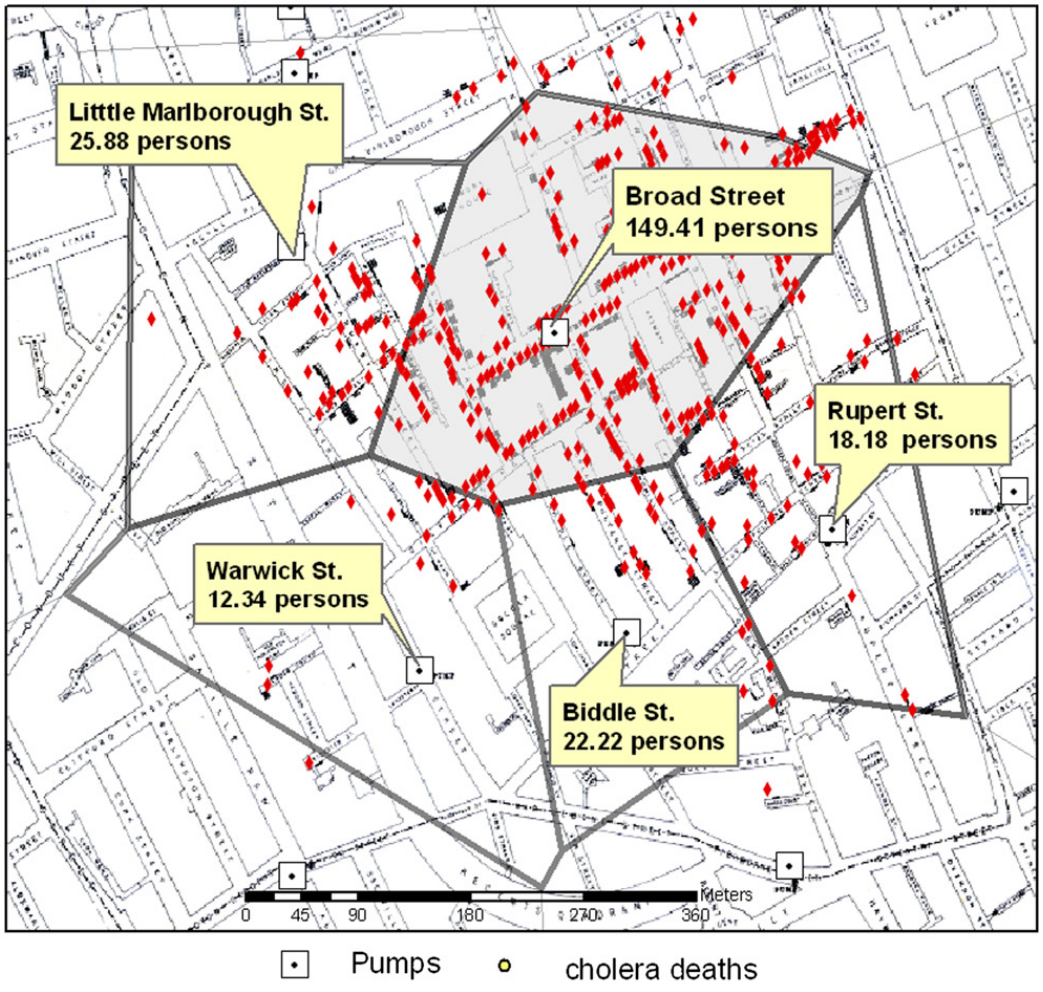
John Snow cholera analysis...John Snow not Jon Snow!



He analysed the deaths related to cholera and plotted them on a map. He discover that the cases were clustred around a particular pump in Soho.

Remaking Snow's 1855 Map

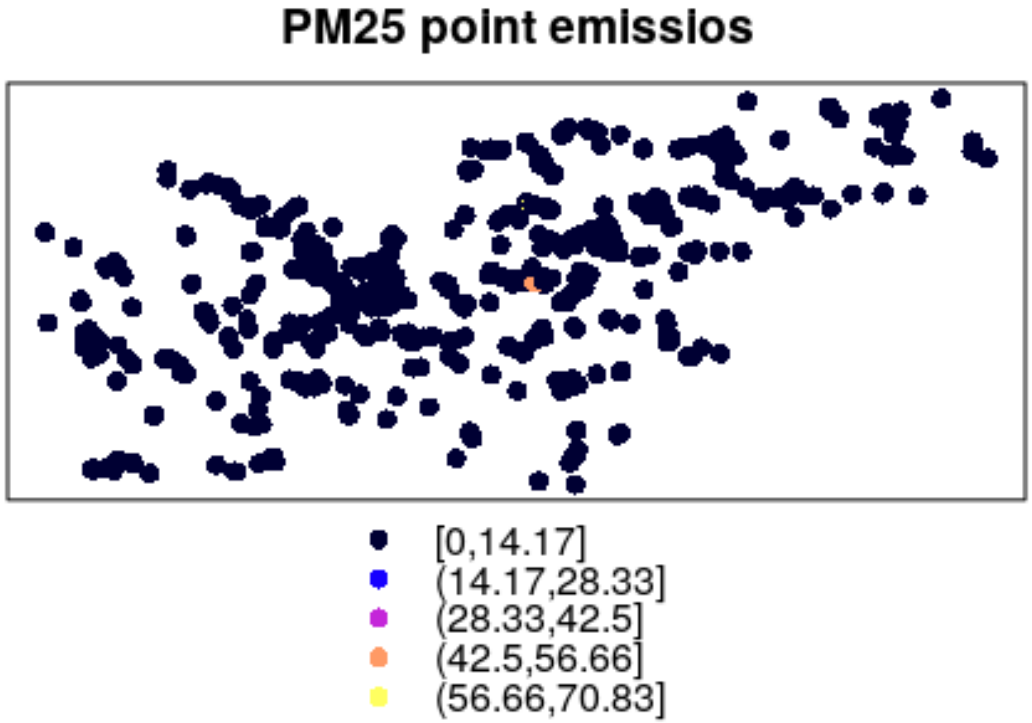
Cholera Mortality per 1,000 persons



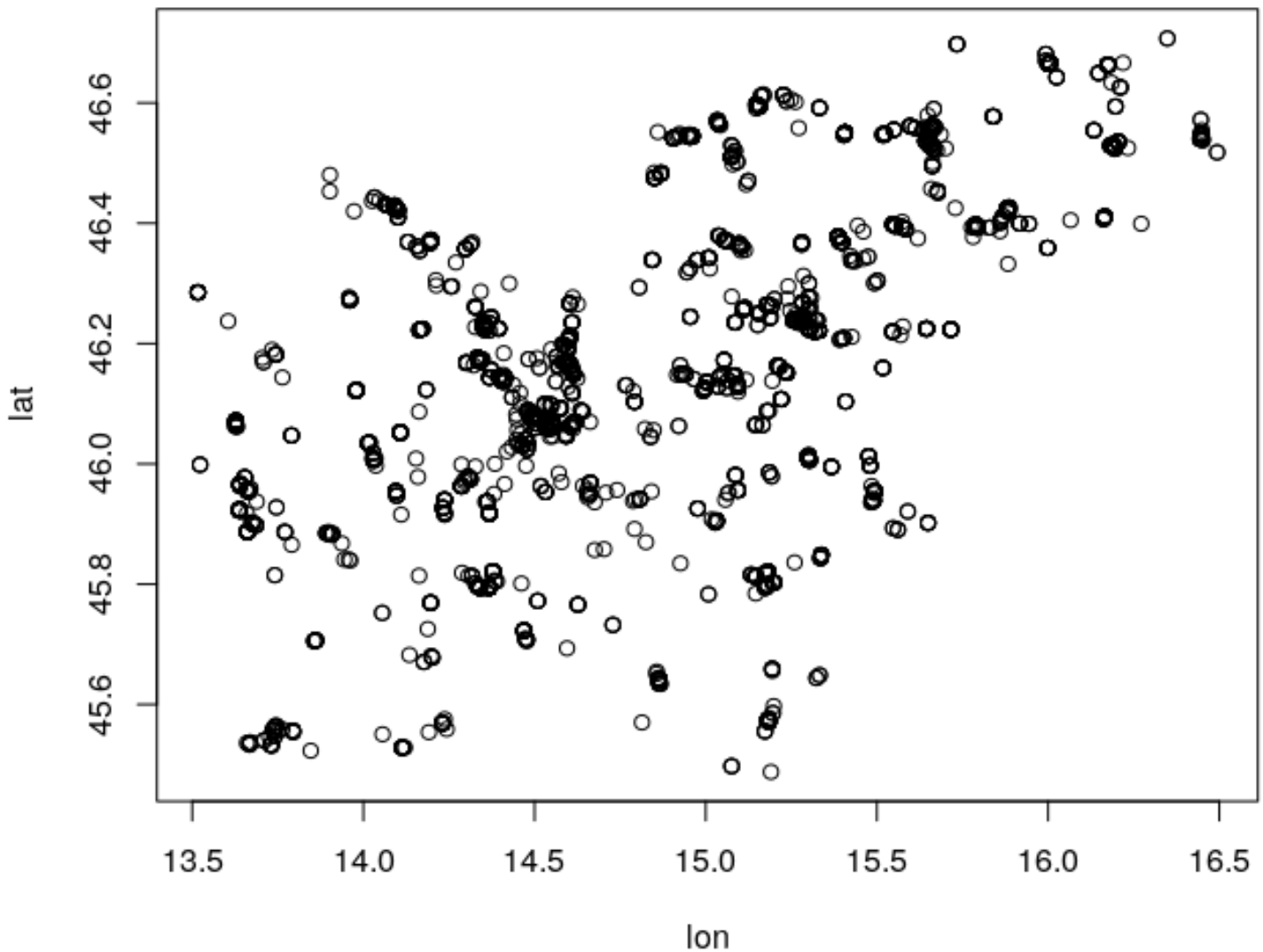
Cholera Mortaliy per 1,00 persons for central pump catchments.

[0]Koch, Tom, and Kenneth Denike. "Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854." Social science medicine 69.8 (2009): 1246-1251

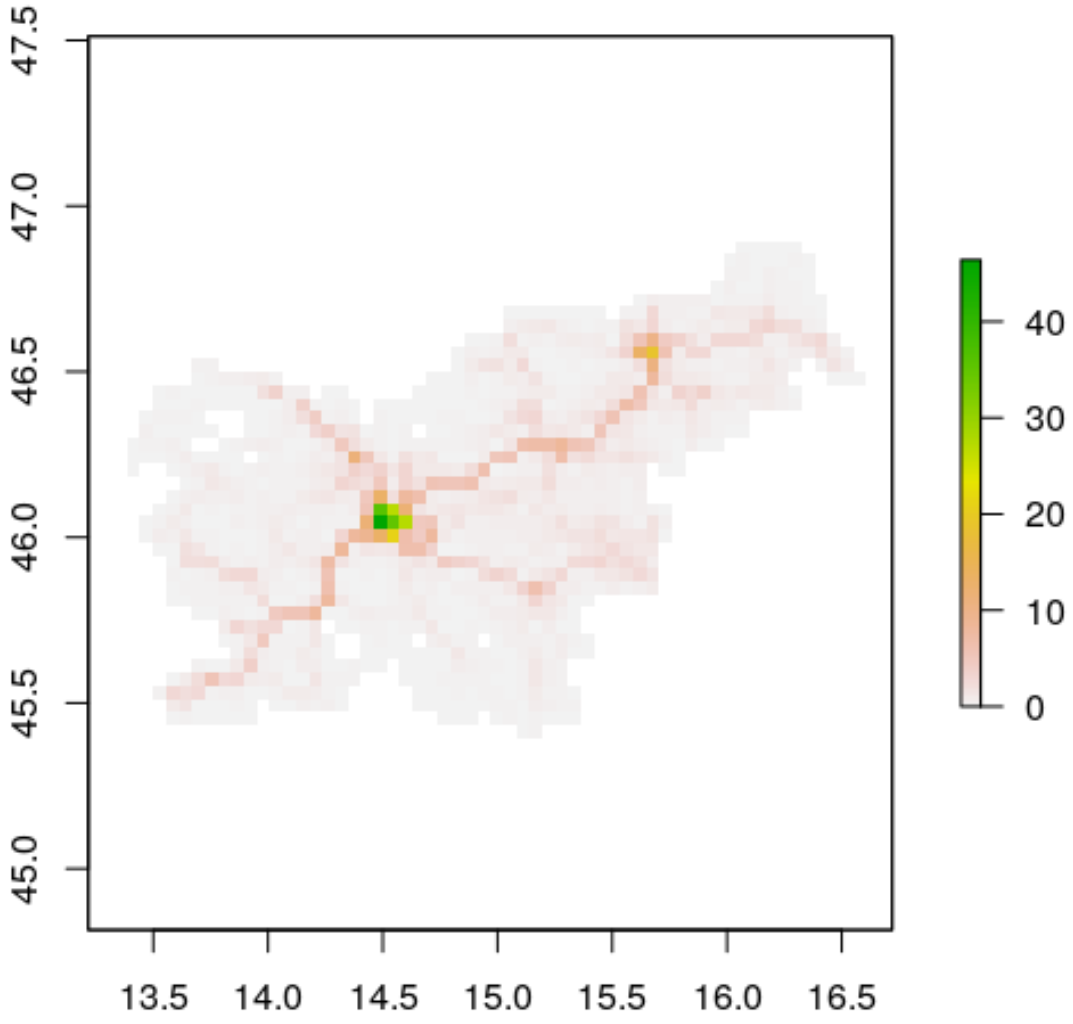
Geostatistics



Spatial point patterns



Lattice



Geostatistics

Includes a set of statistical methods that concern random variables with spatial and/or temporal variability (random fields). The methods are based on the assumption that the spatiotemporal variability includes a random component that has space–time correlation.

Geostatistics deals with the analysis of random fields $Z(s)$ -> Z random and s the non-random spatial index:

$$\{Z(s) : s \in D\} \tag{1}$$

Its applications are in ore mining, petroleum geology, hydrology, oceanography, meteorology, geochemistry, soil science, agriculture.

We can represent a stationary random process by the model value of $Z(s)$ can be decomposed into the mean of the process plus a random component drawn from a distribution.

$$Z(s) = u + e(s) \tag{2}$$

Assumption of intrinsic stationary: The variance of the difference between is the same between any two points that are at the same distance apart no matter which two points are chosen.

$$2\gamma(h) = var[Z(s) - Z(s+h)] = E[Z(s) - s(s+h)^2] \tag{3}$$

The function $\gamma(h)$ is called the variogram The variance of Z is constant, the spatial correlation of Z does not depend on location s, but only on the separation distance h. With further assumption of *isotropy* (direction independence) of semivariance, h can be replace with $||$h$||$.

The variogram

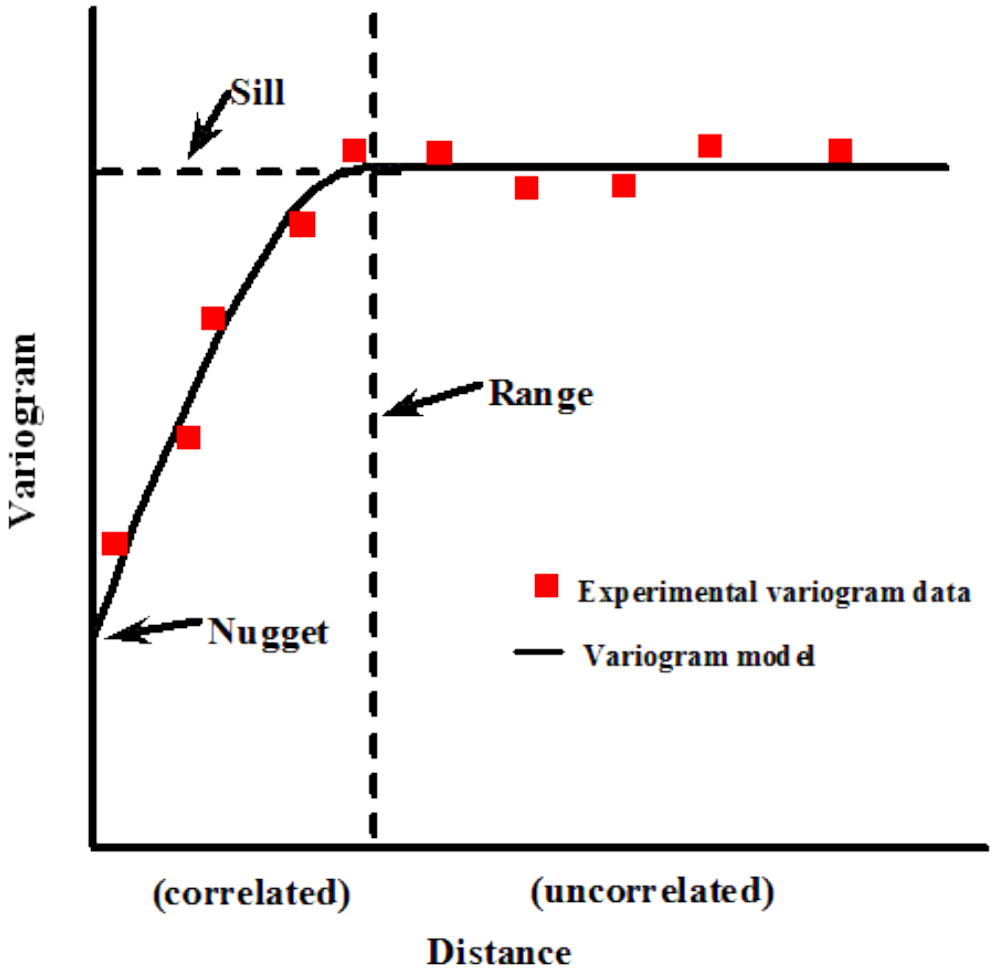
Characteristics of the variogram:

- It is non-negative ($(h) \geq 0$)
- The variogram is symmetric in space $\gamma(h) = \gamma(-h)$
- A function is a semivariogram if it is a conditionally negative definite function
- The variance increases with increasing la
- It might be non-continuous only at the origin- $>$ Nugget variance; for microscale variations $\gamma(0) \neq 0$
- If there is no spatial dependence, the variogram is the constant $var(Z(s))$ everywhere except ta the origin

Different variograms for different data

- Unbounded variogram
- Decrease of the variogram after its local maximum
- Anysotropy -> spatial variation is the the same in every direction
- Trend $Z(s) = u(s) + e(s)$ where $u(s)$ is the trend function. A variogram that appears parabolic at the origin suggests local trend.

The variogram



- **Sill** Limit of the variogram tending to infinity
- **Nugget** The jump of the variogram at the beginning, du to miscroscale variation or measurement error
- **Range** The distance at which the data are no longer autocorrelated, the difference o the variogram from the sill becomes negligible

Estimating semivariances

The variogram cloud, compute the variance for every pair of points x_i and x_j:

$$\gamma(x_i, x_j) = \frac{1}{2}(z(x_i) - z(x_j))^2 \tag{4}$$

Plot values against lag distance as a scatter diagram -> the VARIOGRAM CLOUD

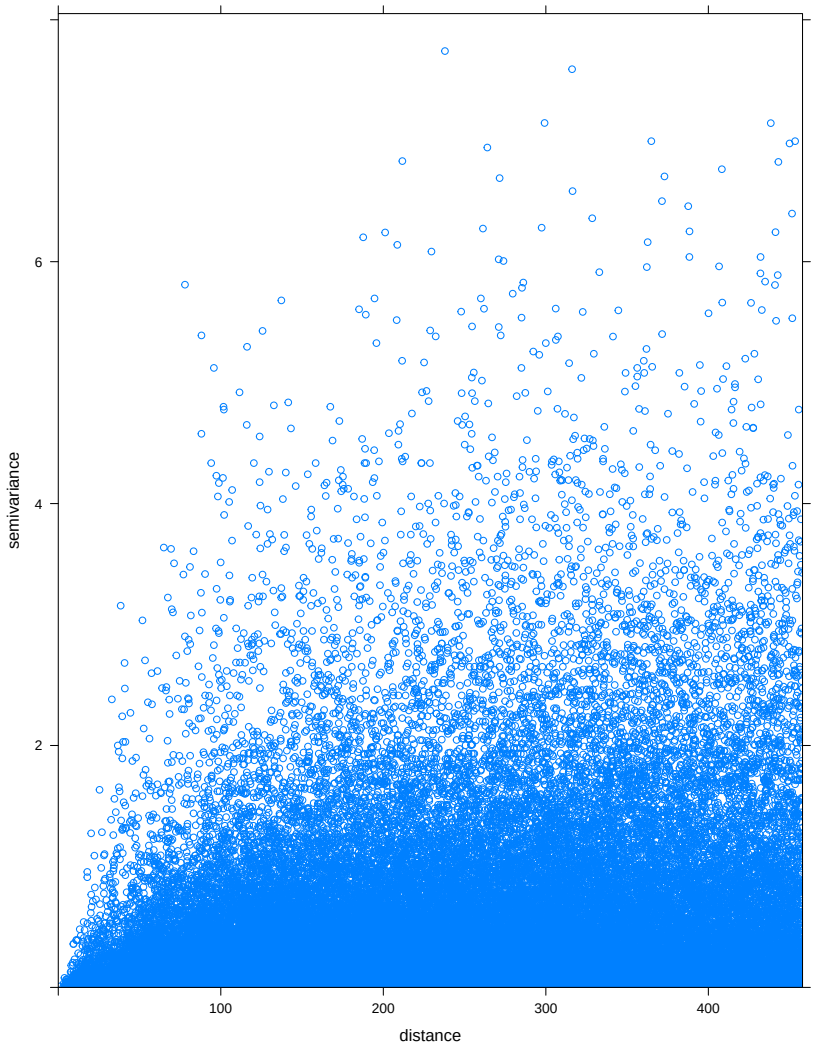
h-Scattergram The closer the points lie to the diagonal -> correlation is stronger and semivariance is smaller

Estimator of semivariance

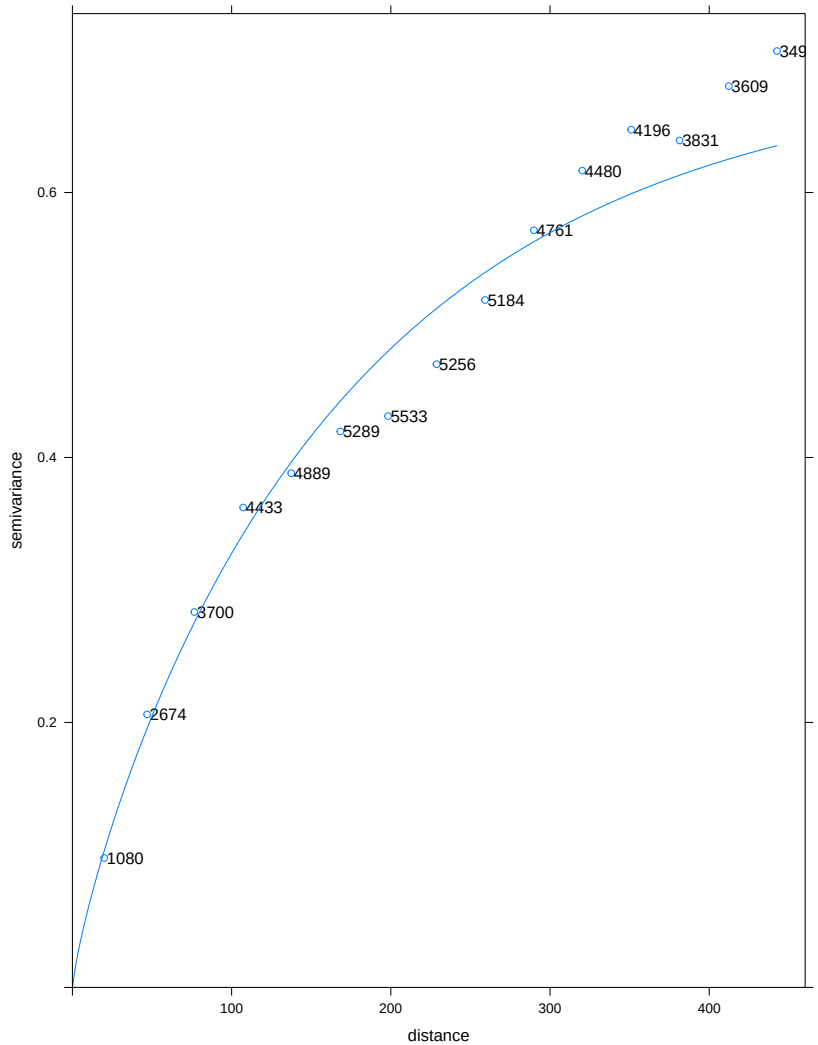
$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} (z(x_i) - z(x_i + h))^2 \tag{5}$$

Where $m(h)$ it the number of pairs of data separated by lag h.
By changing h -> EXPERIMENTAL SEMIVARIOGRAM

The variogram



Variogram cloud



Experimental variogram

Modelling the variogram

Valid models

- **Unbounded random variation** -> power functions
- **Bounded models**:
  - Bounded linear model
  - Circular model
  - Spherical model (most frequently used in geostatistics)
  - Pentaspherical model
  - Exponential model
  - Gaussian model
  - Matern funcion
  - Pure nugget

Variogram types

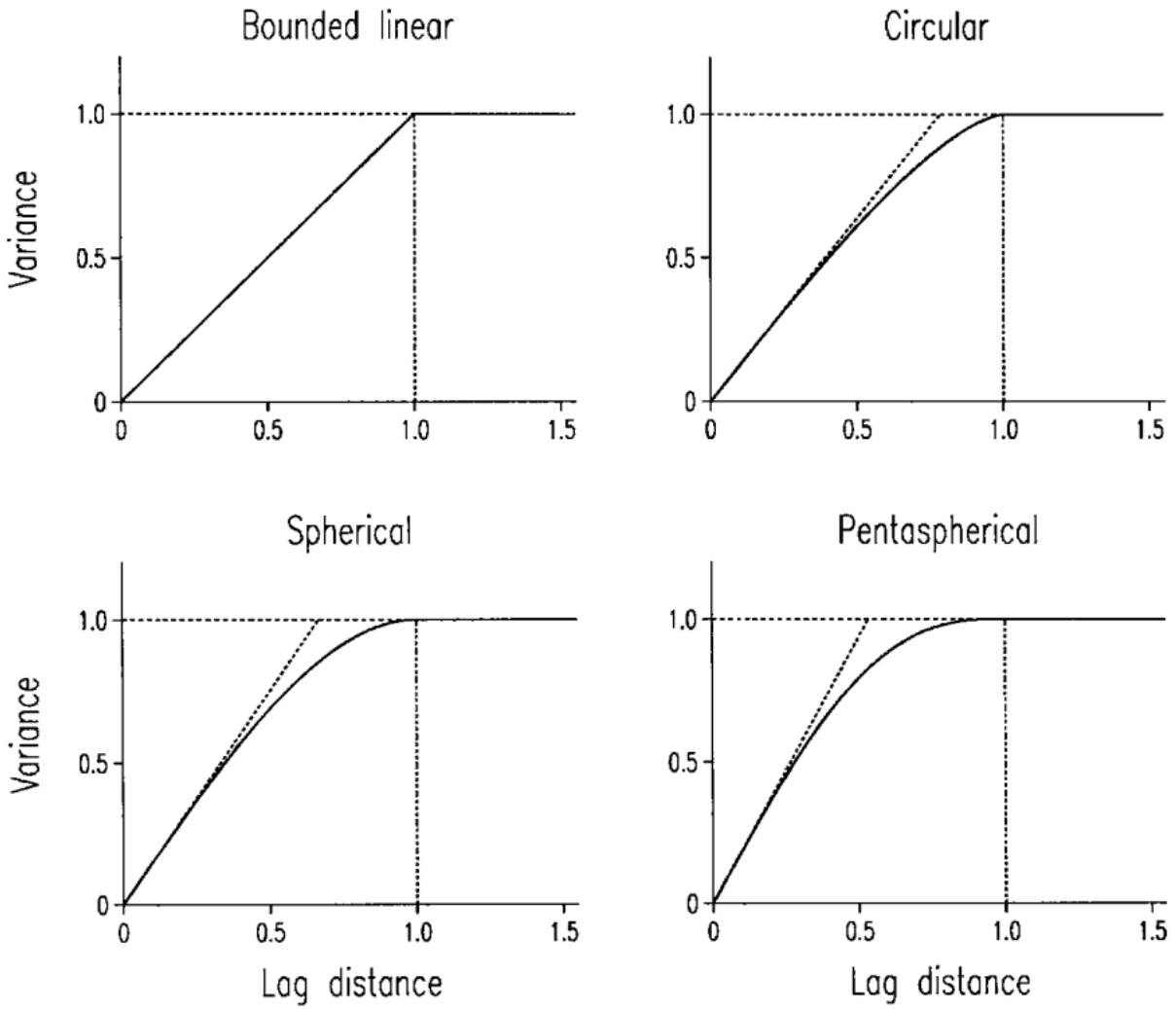


Figure 5.4 Bounded models with fixed ranges: (a) bounded linear; (b) circular; (c) spherical; (d) pentaspherical.

[0]Image from see References 3

Fitting the variogram

Fitting models is difficult for several reasons:

- The accuracy of the observed semivariances is not constant.
- The variation may be anisotropic.
- The experimental variogram may contain much point-to-point fluctuation.

Local estimation or prediction - KRIGING

Interpolation across space according to spatial lag relationship (variogram).
Exploits the aurocorrelation of sampled data in spatial data.
Gaussian process regression -> is a method of interpolation where the interpolated values are modelled by a Gauss process ruled by prior covariances.
The kriging predictior is an "optimal linear predictior " and an exact interpolator (each interpolated value is calculated to minimize the prediction error for that point).
The kriging is the best linear unbiased predictor

It is a two step process:

- The spatial covariance structure is determined by fitting a variogram
- Weights are derived from the covariance structure and used to interpolate values at unsampled locations

Kriging assumptions

- Normal distribution of data
- Stationarity
- Anisotropy

Ordinary kriging

$Z(s) = m + e(s)$, stochastic Z composed by a mean and a residual with a constant mean $E(Z(s)) = m$ and a variogram. The mean is not known. Weighed averaged of measurements where the covariance structure among the observed locations is incorporated:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_1 z(x_i) \quad (6) \qquad \sum_{i=1}^{N} \lambda_i = 1 \quad (7)$$

The kriging estimator incorporates the covariance structure among $z(x_i)$ into the weights doe prediction or $\hat{Z}(x_0)$.

The prediction variance is: $Var(\hat{Z}(x_0)) = Var[(\hat{Z(x_0)} - Z(x_0))^2]$ which is minimized to obtain the kriging weights λ

Kriging weights

When calculating the weights the two goals are no bias and minimal variance of estimation. In general the only large weights are are those of the points near the point to be kriged. Factors affecting the weights:

- Near points carry more weight than more distant. If nugget is large closer points are not as large as without nugget
- Clustered points carry less weights individually than isolated ones. In ordinary kriging difference over large distances has little influence-> we can accept the enotion of quasi-stationarity

Kinds of kriging

- Ordinary kriging-> most robust in mostly used
- Lognormal krigning -> for skewed data (problems with the inverse of log)
- Block kriging -> estimation over a block
- Universal kriging -> data with trend or drift
- Facotrial kriging -> nested variation
- Ordinary cokriging ->more input variables

Others: indicator kriging, disjunctive kriging, probability kriging, bayesian kriging

Dealing with trend

If there is trend in the data $Z(s) = m + e(s)$ where the m is not constant. When the spatial process includes trend or drift:

$$Z(s) = u(s) + e(s) = \sum_{k=0}^{K} \beta_k f_k(s) + e(s) \tag{8}$$

Trend is usually modelled with low order polynomial. What we can do?

- Use universal kriging
- Model the trend first and subtract it from data (use trend surface analysis)
- Variogram estimation by residual maximum likelihood (REML)

Evaluate kriging performance

Use of the cross validation method:

- Compute experimental variogram for the whole sample data and fit variogram models.
- Divide spatial data in n random subsets, use one subset for testing all others for fitting kriging.
- Evaluate the results of as the mean of some statistical measure (SE, MSE,...) from all n subsets.

Cokriging

Use variables that are also or more available than the response variable (the variable we want to interpolate). More variables interest us simultaneously. Each variable individually is treated as random. Multiple variables are cross correlated-> the spatial variability of A is cross correlated with B and can be used in its prediction Modelling:

- Select variogram model
- Fit models to the direct and cross-variograms for each variable simultaneously-> all models have the same shape and range but different silly and nuggets a (Linear Model of Coregionalization)

Geographically weighted regression

A form of ordinary least squares. In a stationary process:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta x_{2i} + \cdots + \beta_n n_{ni} + \epsilon_i \tag{9}$$

, where $\hat{\beta} = (X_T X)^{-}1 X^T Y$
If spatial non-stationary-$>$ relationships can vary over space:

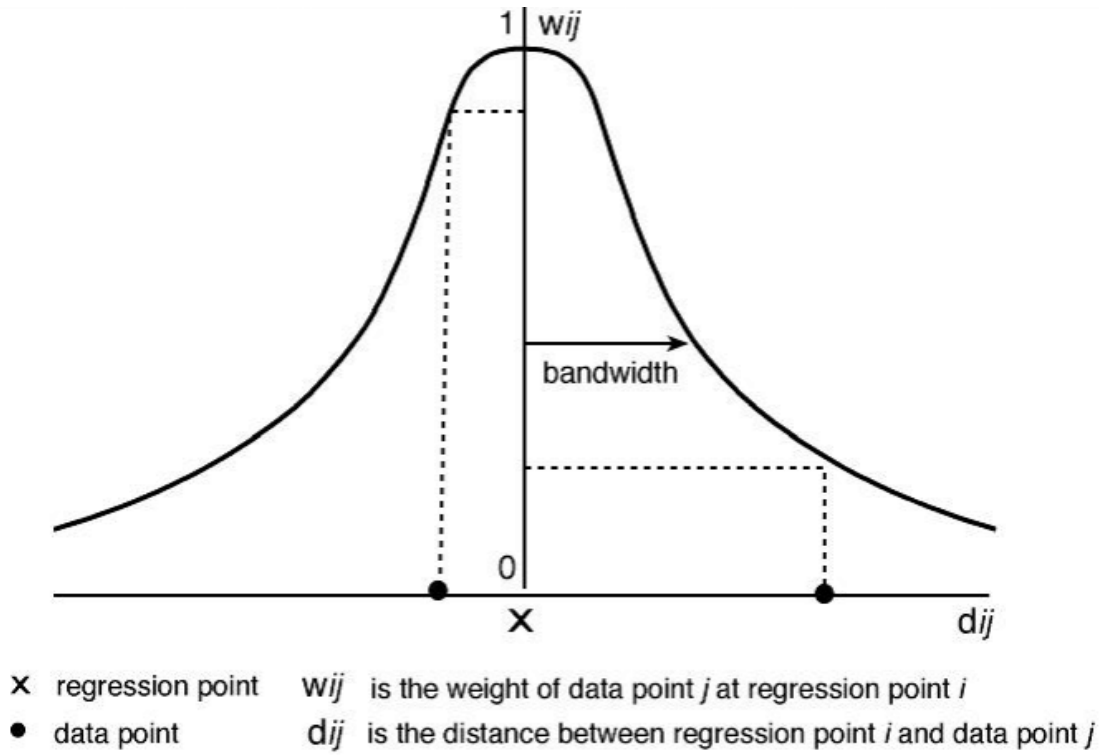
$$y_i = \beta_0(i) + \beta_1(i) x_{1i} + \beta_2(i) x_{2i} + \cdots + \beta_n(i) n_{ni} + \epsilon_i, \tag{10}$$

where $\hat{\beta}(i) = (X^T W(i) X)^{-1} X^T W(i) Y$

W(i) is a matrix of weights specific to location i such that observation nearer to i are given greater weight than observations further away.

GWR weighting function

The weighting function:



moving a weighted window over the data, estimating one set of coefficient values at every chosen 'fit' point. Functions usually comprise: Gaussian, Exponential, Box-car, Bi-square, Tri-cube functions.

Spatial point patterns

The observed point pattern x is the realisation of a stochastic process X in two dimensional space D:

$$\{X(x) : x \in D\} \tag{11}$$

points are usually not observed in X but only in a limited space, a window W, the "sampling window“:

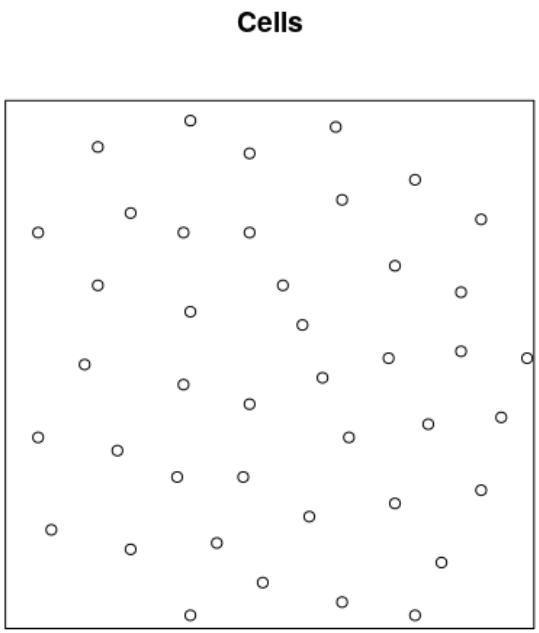
$$x = \{x_1, x_2, ..., x_n\}, x_i \in W_i, n \geqslant 0 \tag{12}$$

Application of point pattern analysis

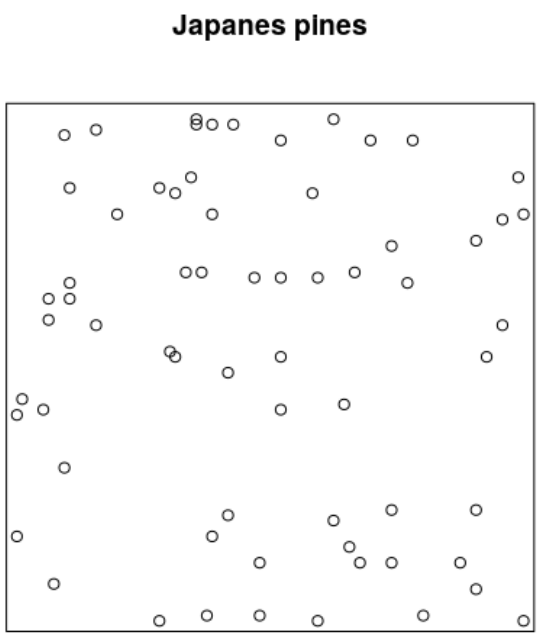
Analysis of point patters appear if different areas of research; ecology, biology, epidemiology, seismology, meteorology Examples of point patterns:

- The spatial distribution of plant species
- Spread of disease
- Locations of earthquakes
- Locations of tornados
- Moving of animals

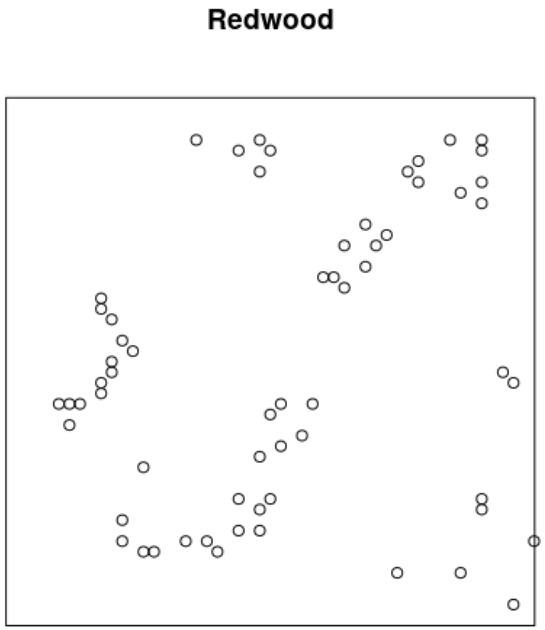
Types of spatial point patterns



regular pattern



random pattern



clustered pattern

Complete spatial randomness (crs)

Point patterns are usually compared to a completely spatially random process. Complete spatial randomness is synonymous to "homogeneous Process" in $\mathbb{R}$ (null model) defined by:

1. The number of events occurring within a finite region A is a random variable following a Poisson distribution with mean $\lambda|A|$ for some positive constant λ and $|A|$ denoting the area of A.
2. Given the total number of points N occurring within an area A, the locations of the N points are independent and identically distributed and uniformly distributed inside A.

homogeneous Poisson process

A homogenní *homogeneous* Poisson process of intensity $\lambda > 0$ has the properties:

1. The number of $N(\mathbf{X} \cap B)$ falling in any region B is a Poisson random variable;
2. the expected number of points falling in B is $\mathbb{E}[N(\mathbf{X} \cap B)] = \lambda area(B)$
3. if B_1 and B_2 are disjoint sets then $N(\mathbf{X} \cap B_1)$ and $N(\mathbf{X} \cap B_2)$ are independent random variables
4. given that $N(\mathbf{X} \cap B) = n$ the na points are independent and uniformly distributed ,

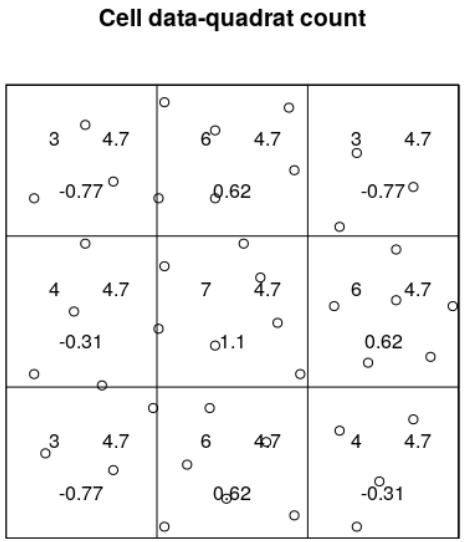
Test of CSR

Find evidence against CSR:

- Quadrat test
- Distance methods
- Second order properties

Quadrat count

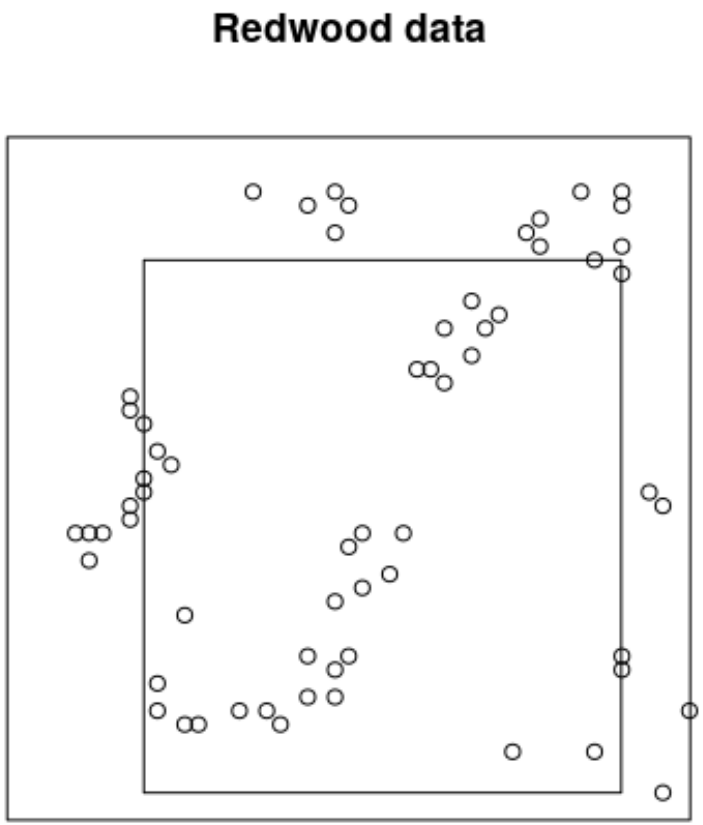
Collecting counts of the number of events in subsets of the study region A.
Test of null hypothesis that *the event is a homogeneous Poisson process*



- left -> observed number
- right -> expected number
- Pearson residuals $= \frac{observed - expected}{\sqrt{expected}}$

Edge effects

The window introduces bias in the distance estimation. The point process extends over X in $\mathbb{R}^2$, but it is only observed in a window W. The nearest point in W could be greater than the nearest point in X.



Edge effect has to be corrected. A lot of edge correction algorithms exist (see Ripley, 1988).

Distance measures

F function based on the empty space distance :

$$d(u, x) = min||u - x_i|| : x_i \in x \tag{13}$$

from a location in $\mathbb{R}^2$ to the nearest point in a point pattern. For this distance we estimate the cumulative distribution function on a grid of locations u_j, $j = 1, ..., ,$:

$$F(r) = \frac{1}{m} \sum_j 1\{d(u_j, x) < r\}; \tag{14}$$

which is biased and needs edge correction ($\hat{F}(r)$). We compare it with :

$$F_{pois} = 1 - exp(-\lambda \pi r^2) \tag{15}$$

- $\hat{F}(r)) > F_{pois}(r)$ suggest regularly spaced pattern
- $\hat{F}(r)) < F_{pois}(r)$ suggest clustered pattern

Distance measures

G function based on the nearest neighbour distance:

$$t_i = min||x_i - x_j|| : x_i \in x \tag{16}$$

distance from each point x_i to its nearest neighbour. For this distance we estimate the cumulative distribution function:

$$G(r) = \frac{1}{n(x)} \sum_i 1\{t_i < r\}; \tag{17}$$

which is biased and needs edge correction ($\hat{G}(r)$). We compare it with :

$$G_{pois} = 1 - exp(-\lambda \pi r^2) \tag{18}$$

- $\hat{G}(r)) < G_{pois}(r)$ suggest regularly spaced pattern
- $\hat{G}(r)) > G_{pois}(r)$ suggest clustered pattern

Distance measures

K function based on the pairwise distances function:

$$s_{ij} = ||x_i - x_j| \tag{19}$$

distance between all distinct pairs of points x_i an x_j in the pattern. For this distance we estimate the cumulative distribution function:

$$K(r) = \frac{1}{\lambda}\mathbb{E}[\text{num. of extra events within distance } r \text{ of an arbitrary event}] \tag{20}$$

which is biased and needs edge correction ($\hat{K}(r)$). We compare it with :

$$K_{pois} = \pi r^2 \tag{21}$$

- $\hat{K}(r)) > K_{pois}(r)$ suggest clustering
- $\hat{K}(r)) < K_{pois}(r)$ suggest regular pattern

Envelopes

The Monte Carlo testing principle

- take the theoretical F, G or K function as the reference for a completely random point process;
- generate M independent simulations of this process inside the study region W;
- compute the functions for each simulated realisation;
- from the estimated functions get the upper and lower limit of the envelope

The limits of the envelope are not "confidence intervals“ but critical values for a test of the hypothesis $K(r) = \pi r^2$

Departure from csr

Departure from csr:

- Clustering:
  - Inhomogenous Poisson process
  - Cox process
  - Poisson cluster process
- Regularity:
  - Simple inhibition processes

Inhomogenous Poisson Processes (IPP)

The intensity is a deterministic function of spatial location (λ is not constant but location dependent)

Estimates of the intensity: Non-parametric estimation-$>$ kernel smoothing:

$$\hat{\lambda} = \frac{1}{h^2} \sum_{i=1}^{h} \kappa(\frac{||x - x_i||}{h})/q(||x||) \tag{22}$$

where $\kappa(u)$ is a bivariate and symmetric kernel function, $q(||x||)$ is a border correction and h is the bandwidth (level of smoothness)
The problem of fitting such a function is the estimate of the bandwidth $||h||$.
Different kernels produce similar estimaes for equivalent bandwidths, while the same kernel with different bandwidths will produce very different results.

Parametric estimation -> proposing a function for the intensity - parameters are estimated by maximazing the likelihood of the point process. The expression of the likelihood is very difficult to define for abbitrary processes. For n-independent events of IPP it has the extression:

$$L(\lambda) = \sum_{i=1}^{n} log\lambda(x_i) - \int_A \lambda(x)dx \tag{23}$$

A log-linear model is usally used:

$$log(\lambda(x)) = \sum_{j=1}^{p} \beta_j z_j(x) \tag{24}$$

Models of non-Poisson patterns

- Poisson cluster process- > parents generated with a Poisson process, offspring according to some stochastic mechanism (Matern cluster process)
- Cox process -> A Poisson process with a random intensity function
- Thinned process
- Sequential models

Hands-on exercises

1. Open the script "do_it_yourself1.R" and estimate the ozone concentrations (OZHMX1HR) at the specifiec locations by *LAT* and *LON*
2. Open the script "do_it_yourself2.R" and estimate from the sightings of Bigfoot if they live in clusters or at random locations across the US. Estimate where it is most probable to find a bigfoot near the US Great Lakes.

Literature

1. Cressie, Noel. "Statistics for spatial data." Terra Nova 4.5 (1992): 613-617.
2. Bivand, Roger S., et al. Applied spatial data analysis with R. Vol. 747248717. New York: Springer, 2008.
3. Webster, Richard, and Margaret A. Oliver. Geostatistics for environmental scientists. John Wiley Sons, 2007.
4. Baddeley, Adrian J., and Rolf Turner. "Spatstat: An R package for analyzing spatial point pattens." (2004): 1-42.