Applied spatial data analysis

Jana Faganeli Pucer

University of Ljubljana, Faculty of Computer and Information Science

June , 2019

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 scince, agriculture.

Assumption of intrinsic stationarity: $Z(s) = m + e(s)$, stochastic Z composed by a mean and a residual with a constant mean $E(Z(s)) = m$ and a variogram :

$$\mathbb{E}[Z(s) - Z(s+h)] = 0 \tag{2}$$

and

$$\gamma(h) = \frac{1}{2}E(Z(s) - s(s+h))^2 \tag{3}$$

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 indepencence) of semivariance, h can be replace with $||$h$||$.

Characteristics of the variogram

- The variogramis symmetric in space $\gamma(h) = \gamma(-h)$
- It is semidefinite
- Nugget variance; for microscale variations $\gamma(0) \neq 0$
- The variance increases with increasing lag
- Sill and range: Sill is the upper bound of a variogram, the maximum is the sill variance. The varigram reaches the sill at a finite lag distance -> range
- Unbounded variogram
- Decrease of the variogram after its local maximum
- Anysotropy -> spatial variation is the the same in every direction
- Trend $Z(x) = u(s) + e(s)$ where $u(s)$ is the trend function. A variogram that appers parabolic at the origin suggests local trend.

The variogram - the cornerstone of geostatistics

The variogram cloud, compute the variance for every pair od 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

Estimaor of semivariance

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

By changing h -> EXPERIMENTAL SEMIVARIOGRAM

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

Fitting the variogram

Fitting models is difficult for several resons:

- the accuracy of the observed semivariances is not constant.
- the variation may be anisotropic.
- the experimental variogram may contain much point-to-point fluctuation.
- most models are non-linear in one or more parameters.

Local estimation or prediction - KRIGING

Gaussian process regression is a method of interpolation.
The kriging predictor is an “optimal linear predictor” and an exact interpolator, meaning that each interpolated value is calculated to minimize the prediction error for that point.
Exploits the aurocorrelation of sampled data in spatial data.
It is a two step process:

- The spatial covariance structure is determied 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

Weighed averaged of measurements where the covariance struction among the observed locations is incorporated:

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

The weights are based on the covariance among points in the sample and the covariances between points and the point to be predicted.

The prediction variance is: $Var(\hat{Z}(x_0)) = Var[(Z(\hat{x}_0) - Z(x_0))^2]$ which is minimized to obtain the kriging weights λ

Kinds of kriging

- Ordinary kriging-> most robust in mostly used
- Lognormal krigning -> for skewed data
- Universal kriging -> data with trend or drift
- Facotrial kriging -> nested variation
- Ordinary cokriging ->more input variables

Others:

Dealing with trend

If there is trend in the data $Z(x) = \mu + \epsilon(x)$ where the μ is constant. When the spatial process includes trend or drift:

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

Trend is usally 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)

Cokriging

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)

Geographycally weighted regression

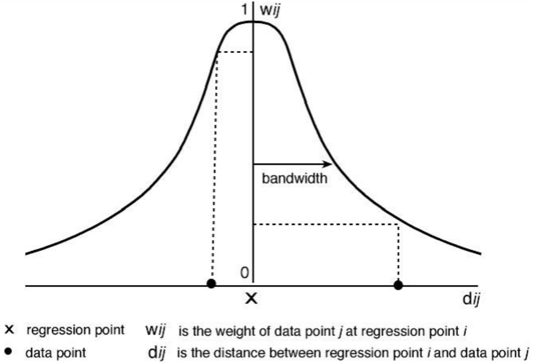
A form of ordinary least squares. In a stationaty process: $y_i = \beta_0 + \beta_1 x_{1i} + \beta x_{2i} + \cdots + \beta_n n_{ni} + \epsilon_i$ Where $\hat{\beta} = (X_T X)^{-}1X^T Y$ If spatial non-stationarity-$>$ relationships can vary over space: $y_i = \beta_0(i) + \beta_1(i)x_{1i} + \beta_2(i)x_{2i} + \cdots + \beta_n(i)n_{ni} + \epsilon_i$, where $\hat{\beta}(i) = (X_T W(i)X)^{-1}X^T W(i)Y$

where 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.

A useful as an explanatory technique. It is able to indicate where non-stationarity is taking place on the map.

GWR weighting function

The weighting function:



x regression point

● data point

w*ij* is the weight of data point *j* at regression point *i*

d*ij* is the distance between regression point *i* and data point *j*

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 patteren x is the realisation of a stochastic process X in two dimensional space D:

$$\{X(x) : x \in D\} \tag{9}$$

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

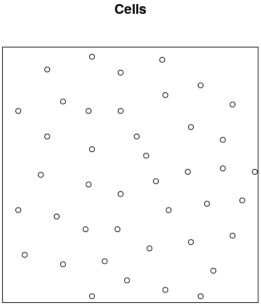
$$x = \{x_1, x_2, ..., x_n\}, x_i \in W_i, n \geqslant 0 \tag{10}$$

Application of point pattern analysis

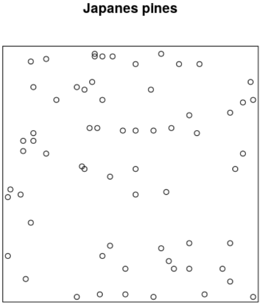
Analysis of point patters appear if different areas of research; ecology, biology, epidemiology, seismology, meteorology Examples of point patterns:

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

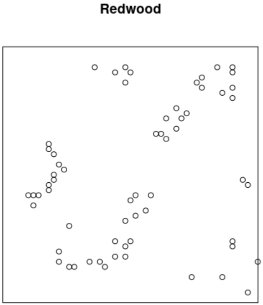
Types of spatial point patterns



regular pattern



random pattern



clusterd pattern

Complete spatial randomness

Point patterns are usually compared to a completely spatially random process. Complete spatial randomness is synonimous to "homogenous 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 indentically distributed and uniformly distributed inside A.

homogenous Poisson process

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

1. The number of $N(\mathbf{X} \cup B)$ falling in any region B is a Poisson random variable;
2. the expected number of points falling in B is $\mathbb{E}[N(\mathbf{X} \cup B)]$

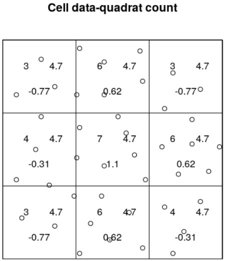
Test of CSR

Find evidence ageinst 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 homogenous Poisson process*



Cell data-quadrat count

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

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 :

$$d(u, x) = min||u - x_i|| : x_i \in x \tag{11}$$

from a location in $\mathbb{R}^2$ to the nearest point in a point pattern. For this distance we estimate the cumulative distribution function:

$$F(r) = \frac{1}{m} \sum_j 1\{d(u_j, x) < r\}; \tag{12}$$

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

$$F_{pois} = 1 - exp(-\lambda \pi r^2) \tag{13}$$

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

Distance measures

G function based on the nearest neighbour distance:

$$t_i = min||x_i - x_j|| : x_i \in x \tag{14}$$

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

$$G(r) = \frac{1}{n(x)} \sum_i 1\{t_i < r\}; \tag{15}$$

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

$$G_{pois} = 1 - exp(-\lambda \pi r^2) \tag{16}$$

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

Distance measures

K function based on the pairwise distances function:

$$s_{ij} = ||x_i - x_j| \tag{17}$$

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

$$K(r) = \frac{1}{\lambda}\mathbb{E}[numbersofextraeventswithindistancerofanarbitraryevent] \tag{18}$$

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

$$K_{pois} = \pi r^2 \tag{19}$$

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

Envelopes

The Monte Carlo testing principle

- take the theoretical F, G or K funcion 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 esimated fuctions get the upper and lower limit of the envelope

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

Departure from csr

Departure from csr:

- Clustering:
  - Inhomogenious Poisson process
  - Cox process
  - Poisson cluster process
- Regularity:
  - Simiple inhibition processes

Inhomogenious Poisson Processes (IPP)

The intensity is a deterministic function of spatial location (λ is not constant but location dependent)
Estimations can be done:

- Non-parametric estimation-$>$ kernel smoothing:

$$\hat{\lambda} = \frac{1}{h^2} \sum_{i=1}^{h} \kappa(\frac{||x - x_i||}{h})/q(||x||) \tag{20}$$

  where $\kappa(u)$ is a bivariate and symmetric kernel function, $q(||x||)$ is a border corection and h is the bandwidth (level of smoothness)

- Parametric estimation -$>$ proposing a function for the intensity - parameters are estimated by maximasing the likelihood of the point process:

$$L(\lambda) = \sum_{i=1}^{n} log \lambda(x_i) - \int_A \lambda(x) dx \tag{21}$$