Praktikum z ekonometrie

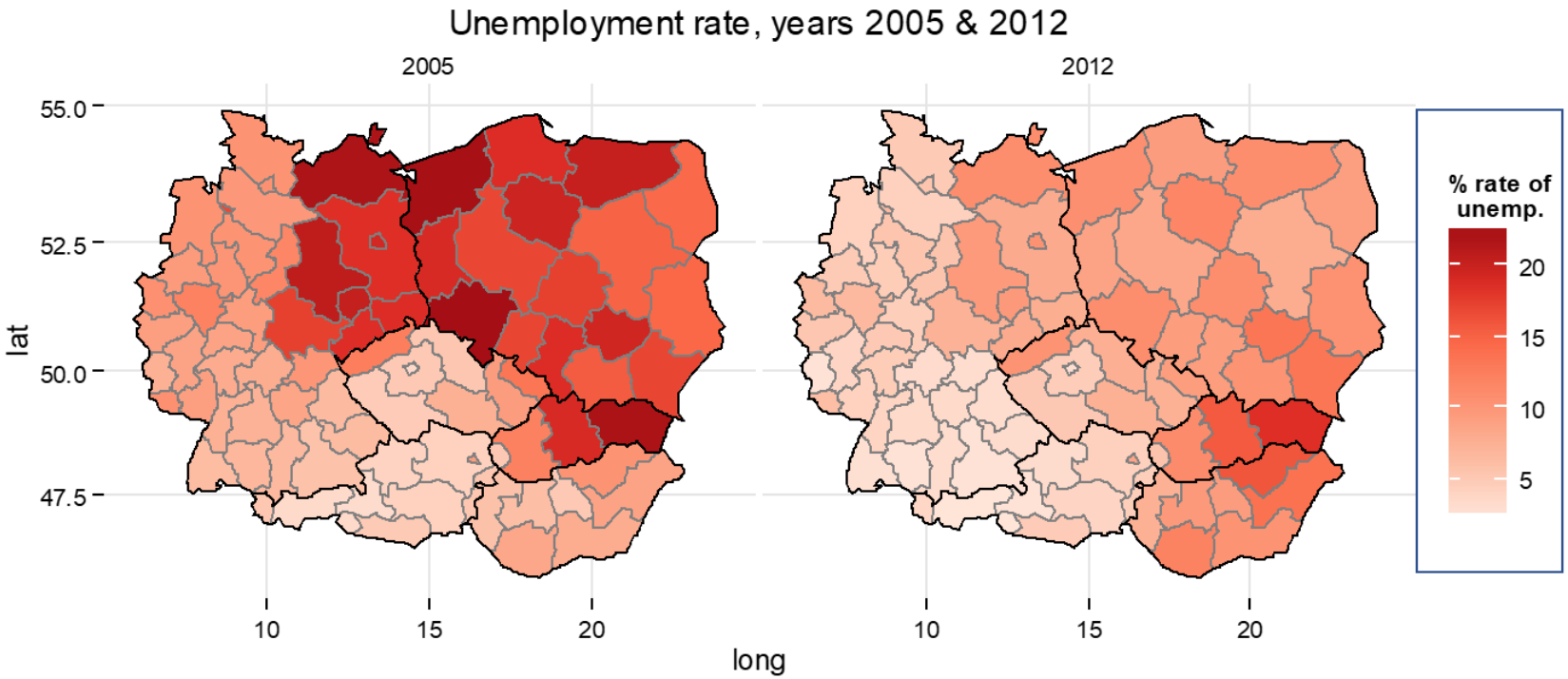
VŠE Praha

Tomáš Formánek

Outline

Introduction

Introduction - spatial analysis



Introduction - spatial analysis

Methods of quantitative spatial analysis:

- Visualization
  Maps, graphical display
- Data exploration & descriptive methods
  Tools to broadly look at spatial patterns
- Econometric modeling
  Fitting models, testing hypotheses, formalizing spatial dependence, discerning spatial effects from other factor (e.g. macroeconomic)

Introduction - spatial analysis

What are spatial data?

- Data that are location speciffic and that vary in space.
- Referenced by a spatial location $\boldsymbol{s}$ (usually 2D),
  $\boldsymbol{s} = (x; y)$; x is longitude (easting) and y is latitude (northing).

  May also be referenced by a zip code, county or state ID.
- Data that are close together in space (time) are often more alike than those that are far apart.
- Tobler's first law of geography:
  "Everything is related to everything else, but near things are more related than distant things."

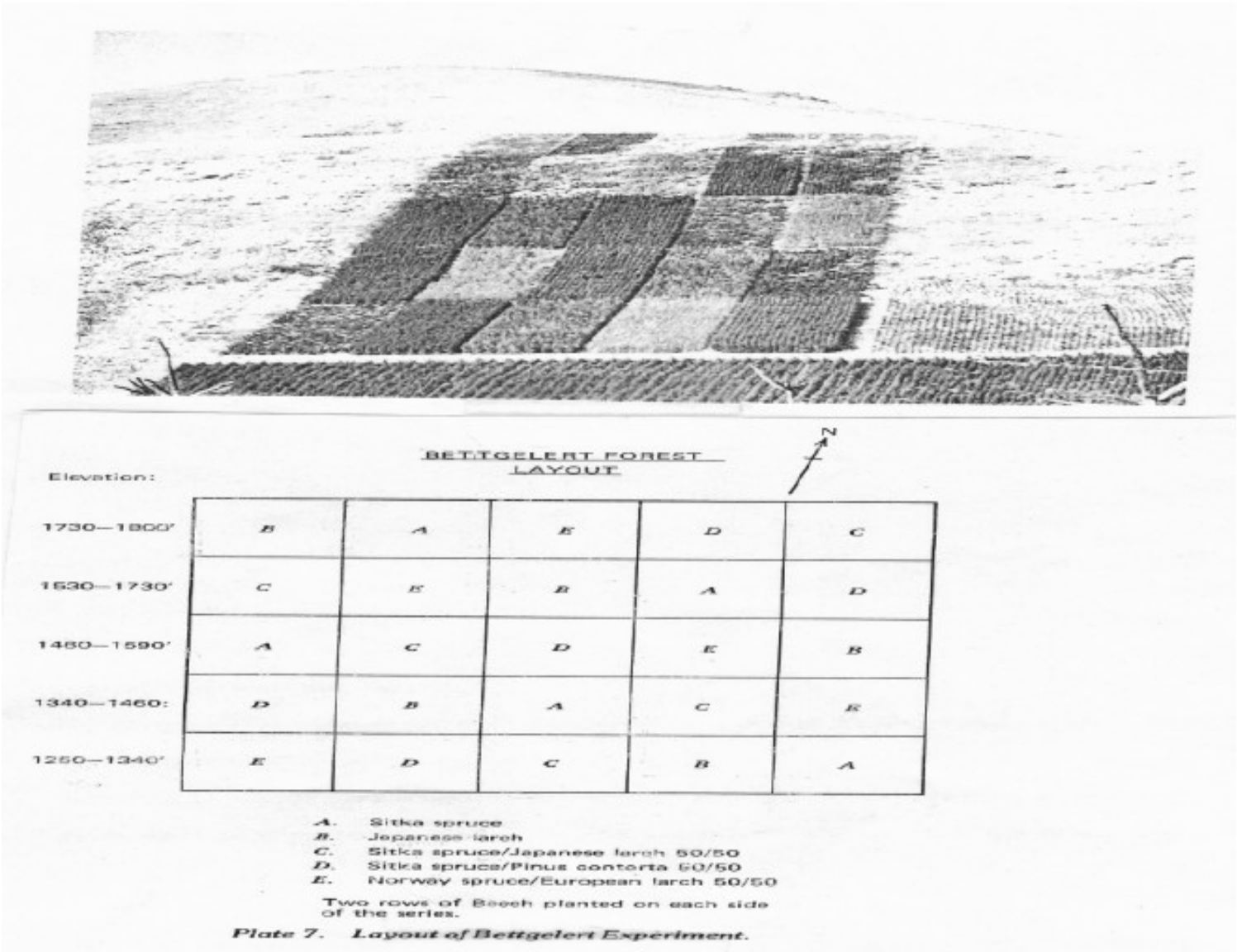
History of spatial analysis: 1854 – London – cholera

History of spatial analysis: 1854 – London – cholera

Dr. John Snow: Early spatial analysis

- In August 1854, there was a major Cholera outbreak in the Soho neighbourhood of London, UK. There were 127 cholera related deaths around the area.
- At the time, germ theory (microorganisms causing disease) was not generally accepted. Dr. J. Snow was a MD, pioneer of germ theory and a statistician.
- Dr. John Snow spoke to local residents and mapped where cholera cases occurred. As a result of his map, he was able to pinpoint the public water pump on Broad Street as the source of contaminated water causing the cholera outbreak.
- Dr. Snow used statistics to find a relationship between water sources and cholera cases and subsequently found out that the waterworks company supplying water to Broad Street pump was taking water from a sewage polluted area of the Thames river.

History of spatial analysis: 1935 – field experiments



BETTGELERT FOREST
LAYOUT

N

Elevation:					
1730—1800′	B	A	E	D	C
1530—1730′	C	E	B	A	D
1460—1590′	A	C	D	E	B
1340—1460′	D	B	A	C	E
1250—1340′	E	D	C	B	A

A. Sitka spruce
B. Japanese larch
C. Sitka spruce/Japanese larch 50/50
D. Sitka spruce/Pinus contorta 50/50
E. Norway spruce/European larch 50/50

Two rows of Beech planted on each side of the series.

Plate 7. Layout of Bettgelert Experiment.

History of spatial analysis: 1935 – field experiments

R.A. Fisher: Early spatial analysis

- R.A. Fisher was probably the first to recognize implications of spatial dependency for statistical analysis.
- In his work on design of experiments in agricultural science, he wrote (Fisher, 1935, p. 66):
  "After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart."
- Observed spatial variability, i.e. field-to-field variability, was largely due to physical properties of the soil and environmental properties of the field. He avoided the confounding of treatment effects with plot effect with the introduction of randomization.
- Fisher's solution was to eliminate spatial dependency bias by localizing the crops under scrutiny into randomly assigned blocks.

Spatial stochastic processes

Measuring spatial variables

Spatial data: measurements and measurement scales

- **Generally, data vary continuously over space, but are measured only at discrete locations.**
  - To characterize spatial variables, spatial aggregation is necessary.
  - Aggregation of spatial variables may be just another source of bias and potential data mis-manipulation
    Summary values are influenced by shape and scale of spatial units.
    Shape (administrative boundaries) may change over time.
- Scale, consistency and relevance should be carefully considered when collecting and analyzing spatial data

Spatio-temporal data:

- Data that are location specific and repeated in time.
- Each variable-observation has a location, time and value.
- Similar methods for analysis, with an added time dimension (choice of sampling frequency, spatial panel data analysis)

Measuring spatial distances

Distances (d) can be defined in a variety of ways, yet the following technical conditions should always apply (invariant to spatial translation, i.e. "shift"):

1. $d(\boldsymbol{s}_i, \boldsymbol{s}_j) = d(\boldsymbol{s}_j, \boldsymbol{s}_i)$
   (symmetry)
2. $d(\boldsymbol{s}_i, \boldsymbol{s}_i) = 0$
   (dist. between a point and itself is zero)
3. $d(\boldsymbol{s}_i, \boldsymbol{s}_j) \leq d(\boldsymbol{s}_i) + d(\boldsymbol{s}_j)$
   (triangle inequality; $d(\boldsymbol{s}_i)$ is the distance from origin)

Measuring spatial distances

Euclidean distances, measured between two point in the "ordinary" Euclidean space. In 2D, the Euclidean distance (L_2 norm) is defined as

$$d(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sqrt{(s_{ix} - s_{jx})^2 + (s_{iy} - s_{jy})^2}\,,$$

where the x and y subscripts handle planar coordinates. For smaller distances, the computational simplicity is attractive.

Great circle distances - for larger distances, planar projection accumulates non-negligible errors. The shortest path between two points on a sphere (given their longitudes and latitudes):

$$d(\boldsymbol{s}_i, \boldsymbol{s}_j) = 2r \arcsin \sqrt{\sin^2\left(\frac{\phi_j - \phi_i}{2}\right) + \cos(\phi_i)\cos(\phi_j)\sin^2\left(\frac{l_j - l_i}{2}\right)}\,,$$

where r is the radius of the sphere, ϕ_1 and ϕ_2 are the latitudes of $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ in radians, l_i and l_j are the longitudes (in radians). Its only an approximation when applied to the Earth, which is not a perfect sphere (correct within a 0.5%; alternative: Vicenty's formulae).

Manhatan distances - L_1 norm, a function on a fixed grid.

Spatial stochastic processes

For a generic location $\boldsymbol{s}$ given by a vector of d coordinates in a d-dimensional Euclidean space, spatial stochastic process ("random field") is often denoted as

$$Z(\boldsymbol{s}) : \boldsymbol{s} \in D \subseteq \mathbb{R}^d \,.$$

- Typically, $d = 2$ for most economic and econometric applications, $d = 3$ is often used in fields such as geology or astronomy.
- D is a fixed finite set of N spatial locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N$.
- Individual $\boldsymbol{s}_i$ units are points in space (say, with GPS-based latitude and longitude coordinates). Sometimes, such points can be associated with non-zero surface area elements.
- Much like in time-series analysis, the individual realizations of a spatial stochastic process – random field – are often denoted $z(\boldsymbol{s}_i)$ or, simply, z_i.

Spatial stochastic processes

Stationarity is a common assumption: a spatial process under scrutiny repeats itself over the domain D. If we translate the entire set of coordinates by $\boldsymbol{h}$ – a specific distance in a specified direction, the stochastic process and its features remain unchanged.

Strong stationarity of a random filed: We start with a finite-dimensional distribution:

$$F_{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_m}(z_1,\ldots,z_m) = P[Z(\boldsymbol{s}_1) \leq z_1, Z(\boldsymbol{s}_2) \leq z_2, \ldots, Z(\boldsymbol{s}_m) \leq z_m] \,.$$

Strong stationarity $\leftrightarrow$ F is invariant under spatial translation $\boldsymbol{h}$.
Unlike d_{ij} (Euclidean distance between two spatial units $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$), $\boldsymbol{h}$ is an orientated distance "shift" (spatial translation) vector.
For strong stationarity:

$$\begin{aligned} &P\left[Z(\boldsymbol{s}_1) \leq z_1, Z(\boldsymbol{s}_2) \leq z_2, \ldots, Z(\boldsymbol{s}_m) \leq z_m\right] \\ &= P\left[Z(\boldsymbol{s}_1 + \boldsymbol{h}) \leq z_1, Z(\boldsymbol{s}_2 + \boldsymbol{h}) \leq z_2, \ldots, Z(\boldsymbol{s}_m + \boldsymbol{h}) \leq z_m\right] . \end{aligned}$$

Spatial stochastic processes

Weak stationarity (also called second order stationarity) assumes that the first two moments exist, are invariant (and finite) and covariance only depends on spatial translation (orientated distance) $\boldsymbol{h}$:

$$E[Z(\boldsymbol{s})] = \mu\,,$$
$$\text{var}[Z(\boldsymbol{s})] = \sigma^2\,,$$
$$\text{cov}[Z(\boldsymbol{s}+\boldsymbol{h}), Z(\boldsymbol{s})] = C(\boldsymbol{s}+\boldsymbol{h}, \boldsymbol{s}) = C(\boldsymbol{h})\,.$$

As autocovariance is a function of $\boldsymbol{h}$ only (under weak st.), for any spatial points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ such that $\boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}$, we can write:

$$\text{cov}\left[Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right] = C(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{h})\,.$$

Covariogram $C(\boldsymbol{h})$ is the covariance between two spatial units, separated by $\boldsymbol{h}$. For $\boldsymbol{h} = \boldsymbol{0}$, it simply describes variance:

$$\text{cov}\left[Z(\boldsymbol{s}+\boldsymbol{0}), Z(\boldsymbol{s})\right] = C(\boldsymbol{0}) = \text{var}\left[Z(\boldsymbol{s})\right].$$

Under weak dependency, covariance disappears with growing distance:

$$C(\boldsymbol{h}) \to 0 \text{ as } ||\boldsymbol{h}|| \to \boldsymbol{\infty}$$

Spatial stochastic processes

Intrinsic stationarity is less restrictive than weak (second order) stationarity and it is defined in terms of first differences.

A spatial process is intrinsically stationary if the difference between two observed spatial points is weakly stationary:

$$E[Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})] = 0\,,$$
$$\mathrm{var}[Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})] = 2\gamma(\boldsymbol{h})\,,$$

where $2\gamma(\boldsymbol{h}) \geq 0$ is the variogram. Generally, $2\gamma(\boldsymbol{h})$ increases with growing oriented distance $\boldsymbol{h}$.

The two types of relaxed stationarity are related: weak stationarity implies intrinsic stationarity but not vice versa. For weakly stationary spatial processes (where $E(Z(\boldsymbol{s}+\boldsymbol{h})) = E(Z(\boldsymbol{s})) = \mu$) the variogram simplifies to:

$$2\gamma(\boldsymbol{h}) = E\left[(Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s}))^2\right],$$

i.e. to the expected squared difference between two observed realizations of a spatial stochastic process.

Spatial stochastic processes

Semivariogram is denoted as $\gamma(\boldsymbol{h})$ and it equals to half the variogram.

Since $2\gamma(\boldsymbol{h})$ is calculated as expectation of a square, $\gamma(\boldsymbol{h}) \geq 0$ for both weakly and intrinsically stationary random fields.

Also, at $\boldsymbol{h} = \mathbf{0}$, $\gamma(\mathbf{0}) = 0$ because

$$E\left[(Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_i))^2\right] = 0 \text{ for } \forall\, i\,.$$

Variogram (semivariogram) is a generalization of the covariogram $C(\boldsymbol{h})$ and under weak stationarity, the two functions are related by:

$$\gamma(\boldsymbol{h}) = C(\mathbf{0}) - C(\boldsymbol{h})\,.$$

If a stationary stochastic process has no spatial dependency at all (i.e. $C(\boldsymbol{h}) = 0$ for $\boldsymbol{h} \neq \mathbf{0}$), the semivariogram is constant: $\gamma(\boldsymbol{h}) = \text{var}[Z(\boldsymbol{s})]$ everywhere, except for $\boldsymbol{h} = \mathbf{0}$, where $\gamma(\mathbf{0}) = 0$.

Isotropic spatial process may be defined through a semivariogram:

$$\gamma(\boldsymbol{h}) = \gamma(||\boldsymbol{h}||) = \gamma(d).$$

Isotropy means that the semivariogram depends only on the distance d between two points and not on direction.

The lack of isotropy – anisotropy – means the semivariogram depends on direction as well as distance.

To assess and test anisotropy, we can estimate and plot directional semivariograms (shown next).

Empirical semivariogram

To perform empirical analysis of distance-based data correlations, we construct the so called empirical semivariogram. First, we divide the distances observed over the domain D into K conveniently chosen intervals:

$$I_1 = (0, d_1],\ I_2 = (d_1, d_2],\ \ldots\ ,\ I_K = (d_{K-1}, d_K]\,.$$

Here, d_1 is the maximum distance within the I_1 interval and d_K is the maximum distance observed over the field of data.

The intervals can be proportional in terms of distance or in terms of sets of observation pairs allocated to each interval (to adjust for unevenly spaced observations).

Note that distances are determined by d (distance magnitudes) only – here, we do not use the orientated distances $\boldsymbol{h}$.

Empirical semivariogram is calculated using the following formula:

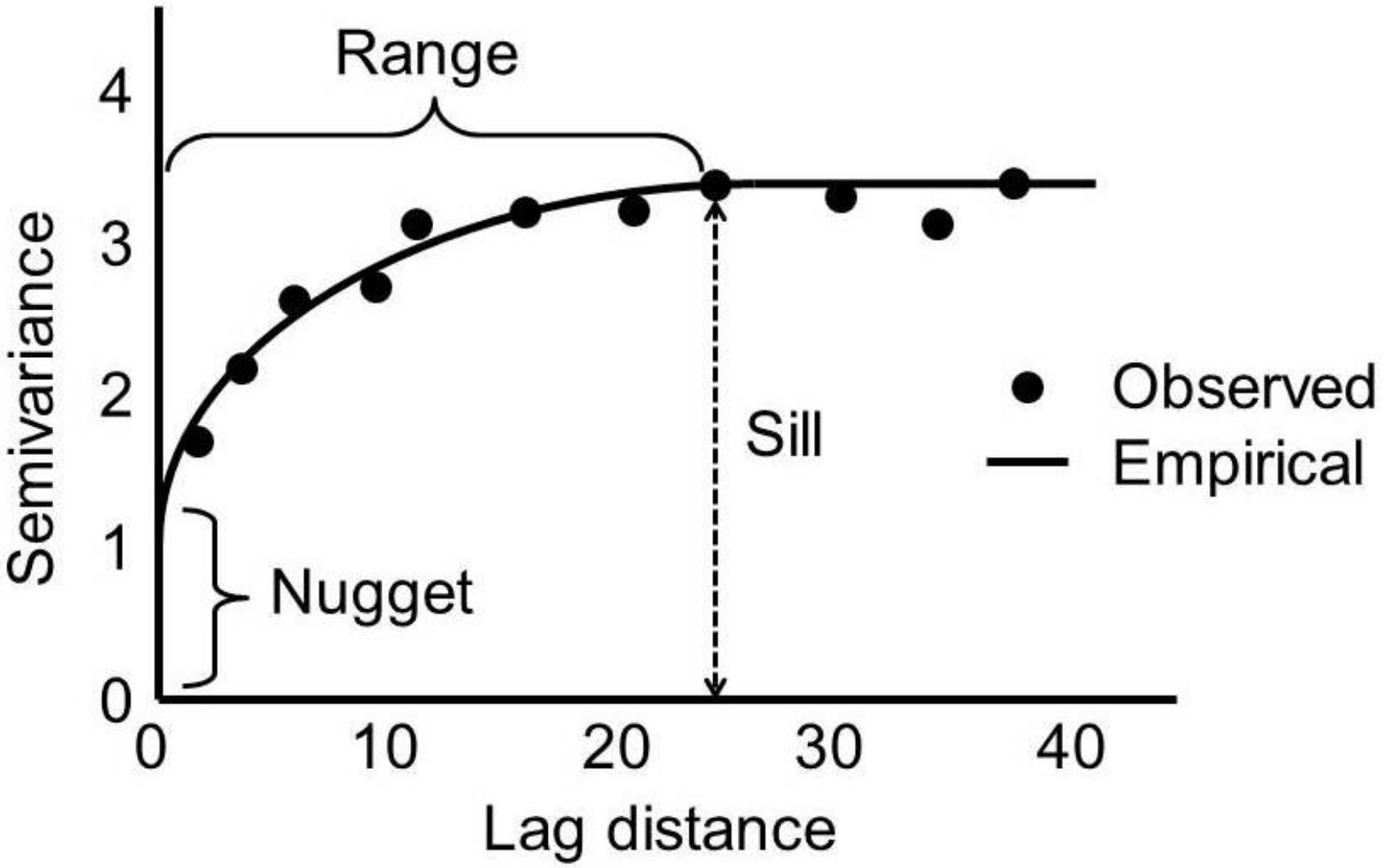
$$\hat{\gamma}(d_k) = \frac{1}{2N(d_k)} \sum_{N(d_k)} [Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)]^2 \,,$$

where $N(d_k)$ is the number of distinct observation pairs in the interval I_k and $\hat{\gamma}(d_k)$ is the semivariogram estimate for its corresponding group (interval) of distances.

Usually, we fit a convenient parametric function (exponential, spherical, Gaussian, etc.) to the estimated $\hat{\gamma}(d_k)$ values (shown next).

The main goal of empirical semivariogram construction is to estimate and visualize the spatial autocorrelation structure of the observed stochastic process.

Empirical semivariogram



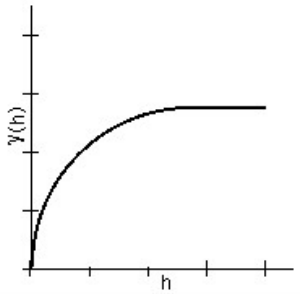
Empirical semivariogram

Three main features of an estimated empirical semivariogram:

- *Nugget* (nugget effect) describes the micro-scale variations or measurement errors in data. Theoretically, at zero distance, $\gamma(0) = 0$. However, two factors play a role here: First, $\gamma(d_1)$ is estimated over the $N(d_1)$ set of pairs, i.e. for the first interval where $d_{ij} \in (0, d_1]$. Second, fitting the empirical semivariogram curve to observed values often causes the non-zero nugget.
- *Sill* amounts to $\lim_{d\to\infty} \gamma(d)$. The sill corresponds to variance of the stochastic field at distances where spatial dependency (which reduces $\gamma(d)$) no longer applies. $\lim_{d\to\infty} \gamma(d) = C(\mathbf{0}) = var[Z(\boldsymbol{s})]$.
- *Range* is the spatial distance (if any) beyond which the data are not autocorrelated. In a way, range describes the strength of spatial structure – based on where the semivariogram "reaches" its asymptote (sill).

Empirical semivariogram (fitting)

CIRCULAR



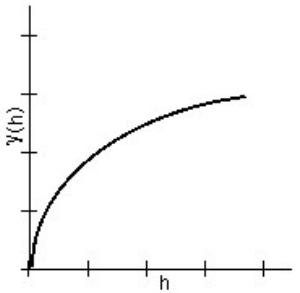
$$\gamma(\mathbf{h}) = c_0 + c\left(1 - \frac{2}{\pi}\cos^{-1}\left(\frac{h}{\alpha}\right) + \sqrt{1 - \frac{h^2}{\alpha^2}}\right)$$
$$0 < h \leq \alpha$$

$$\gamma(\mathbf{h}) = c_0 + c \qquad h > \alpha$$

$$\gamma(0) = 0$$

Circular semivariance model illustration

EXPONENTIAL

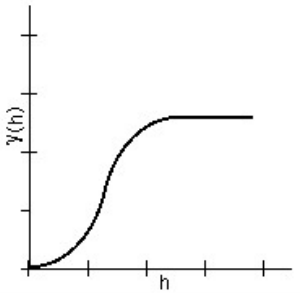


$$\gamma(\mathbf{h}) = c_0 + c\left(1 - exp\left(\frac{-h}{r}\right)\right) \qquad h > 0$$

$$\gamma(0) = 0$$

Exponential semivariance model illustration

GAUSSIAN

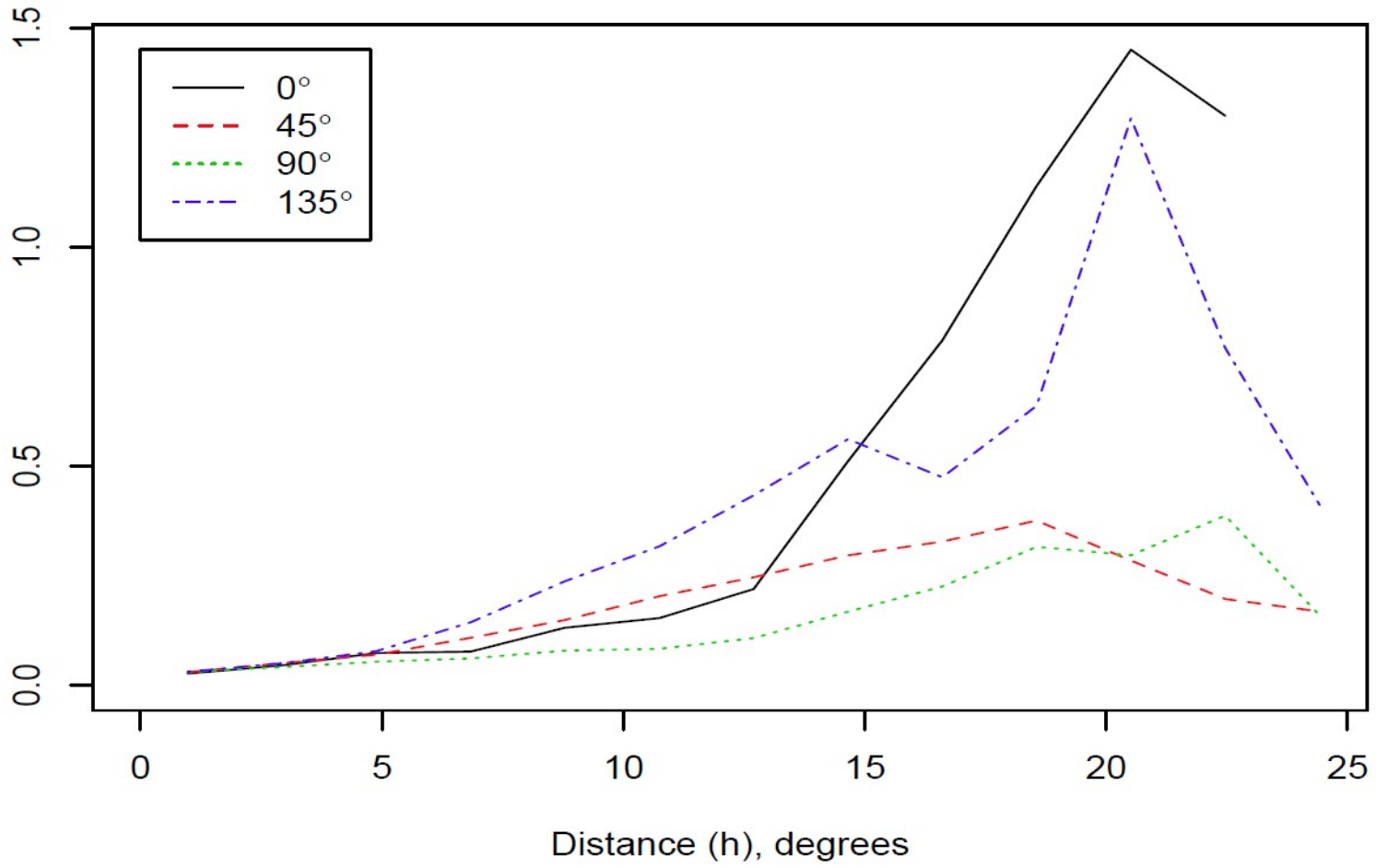


$$\gamma(\mathbf{h}) = c_0 + c\left(1 - exp\left(\frac{h^2}{r^2}\right)\right) \qquad h > 0$$

$$\gamma(0) = 0$$

Gaussian semivariance model illustration

Empirical semivariogram (directional semivariogram)



Spatio-temporal stochastic processes

The above discussion can be generalized to accommodate processes that are observed repeatedly over time.

Such observations usually exhibit both spatial and temporal dependency and variability.

Given the frequency and density limitations of empirical measurements in continuous space and time, we model our observations as realizations of a spatio-temporal random function (random field)

$$Z(\boldsymbol{s}, t), \qquad \text{where } (\boldsymbol{s}, t) \in \mathbb{R}^d \times \mathbb{R}\,,$$

the spatio-temporal domain is indexed in space by $\boldsymbol{s} \in \mathbb{R}^d$ and in time by $t \in \mathbb{R}$.

The separation between spatial and time dimensions is substantial, which is reflected in the notation.

Spatio-temporal stochastic processes

Weak and intrinsic stationarity concepts can be easily expanded from spatial to spatio-temporal data.

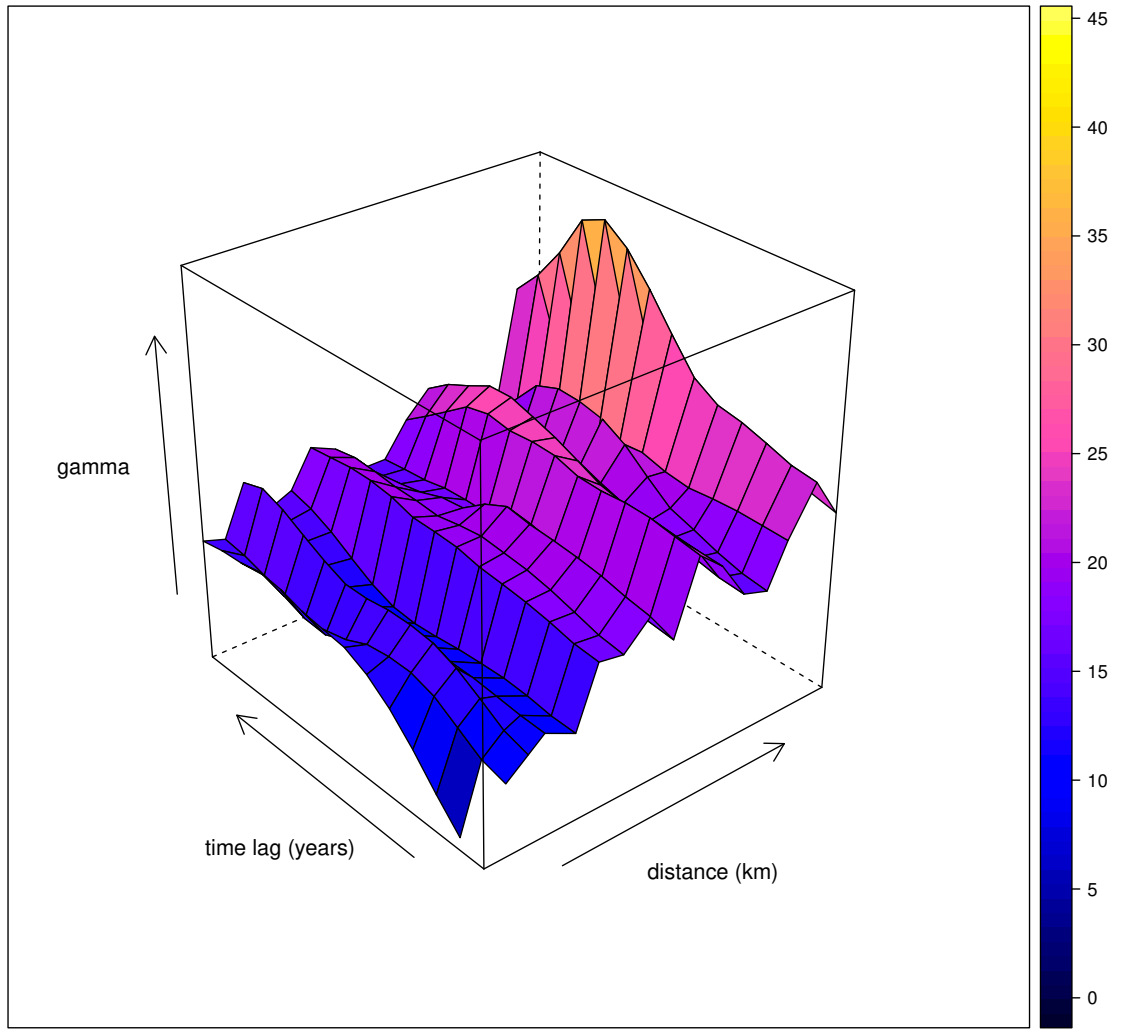
For an intrinsically stationary process $Z(\boldsymbol{s}, t)$, spatio-temporal semivariograms (STSV) is:

$$\gamma(\boldsymbol{h}; t) = \frac{1}{2}\text{var}\left[Z(\boldsymbol{s}_0 + \boldsymbol{h}\,;\, t_0 + t) - Z(\boldsymbol{s}_0; t_0)\right], \qquad (\boldsymbol{h}, t) \in \mathbb{R}^d \times \mathbb{R}.$$

STSV does not depend on the selection of origin $(\boldsymbol{s}_0, t_0) \in \mathbb{R}^d \times \mathbb{R}$ (under intrinsic stationarity).

Also, for intrinsically stationary random fields $Z(\boldsymbol{s}, t)$, the STSV $\gamma(\boldsymbol{h}; t)$ is non-negative and $\gamma(\boldsymbol{0}; 0) = 0$.

Empirical STSV (EU's Unemp., NUTS0, 2002—2016)



Data exploration & descriptive methods

Interpolation

Inverse distance weighting (IDW): a deterministic method of interpolation with a known scattered set of points.

Assigned values to unknown points are calculated with a weighted average of the values available at the known points.

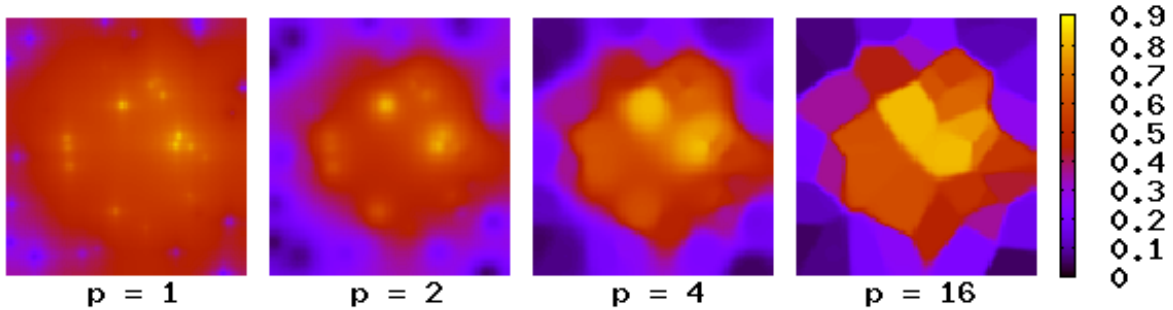
To find an interpolated value x_0 at a given point $\boldsymbol{s}_0$ based on samples $x_i = x(\boldsymbol{s}_i)$ with $i = 1, 2, \ldots, N$ we use a IDW interpolating function:

$$x(\boldsymbol{s}_0) = \begin{cases} \dfrac{\sum_{i=1}^{N} w_i(\boldsymbol{s_0}) x_i}{\sum_{i=1}^{N} w_i(\boldsymbol{s_0})}, & \text{if } d(\boldsymbol{s}_0, \boldsymbol{s}_i) \neq 0 \text{ for all } i \\ x_i & \text{if } d(\boldsymbol{s}_0, \boldsymbol{s}_i) = 0 \text{ for some } i \end{cases}$$

where $w_i(\boldsymbol{s}_0) = \frac{1}{d(\boldsymbol{s}_0, \boldsymbol{s}_i)^p}$, $\boldsymbol{s}_0$ is an interpolating (known) point, d is a distance, and p is a positive *real* number (power parameter).

Weight decreases as distance increases from the interpolated points. High p assigns greater influence to closest values – result may turn into a mosaic of tiles (a Voronoi diagram).

IDW interpolation – illustration



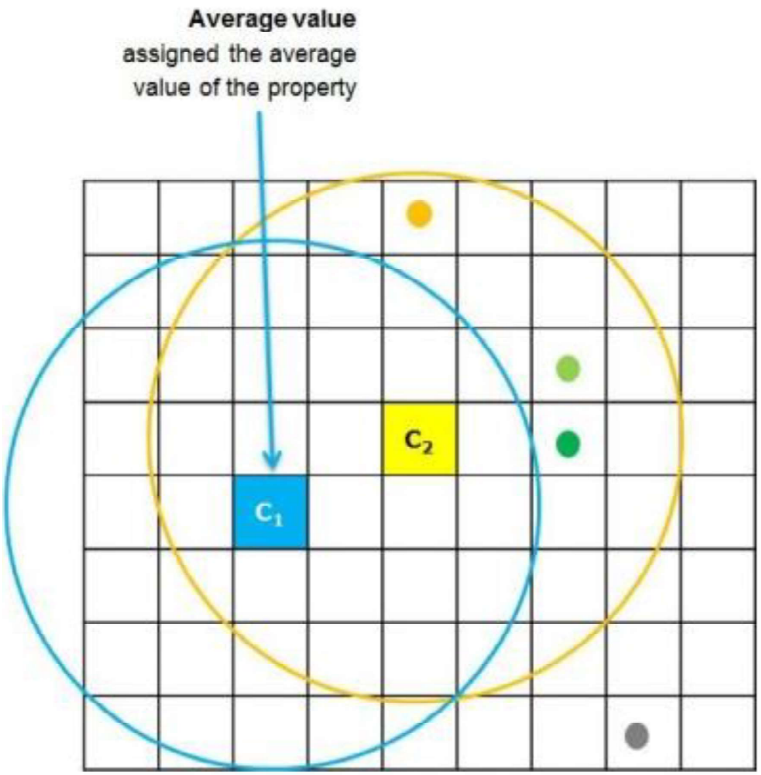
Krigging

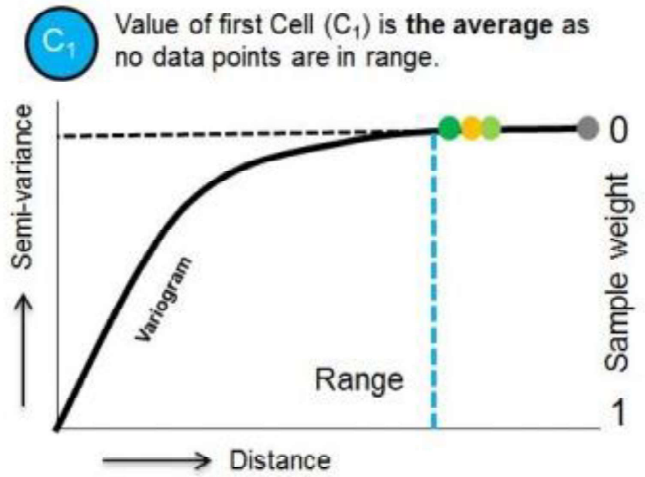
- developed by Daniel G. Krige (1919-2013)
  (originally called "weighted moving averages" method)
- Also known as BLUP (best linear unbiased prediction) and closely related to OLS estimation.
- Returns observed values at sampling locations.
- Interpolates values using the intensity and shape of the empirical (semi)variogram. Results depend on the choice of fitting model (Gaussian, spherical, exponential, etc.).
- Uses neighborhood and/or distance search radius.
- Provides standard errors of interpolated values.
- Multiple approaches and generalizations exist (Block Krigging).
  `http://desktop.arcgis.com/en/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm`

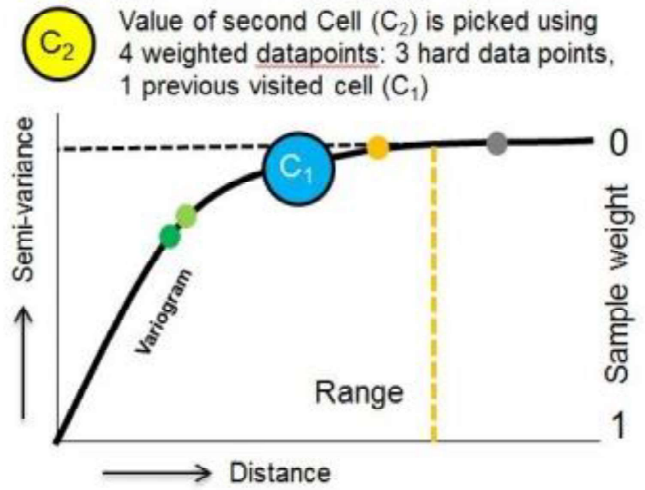
Kriging

- $x(\boldsymbol{s}_0) = \sum_{i=1}^{N} \lambda_i \, x(\boldsymbol{s}_i) \quad \text{s.t.} \quad \sum_{i=1}^{N} \lambda_i = 1.$
- In IDW, λ_i depend only on distance to the prediction location.
- With krigging, λ_i depend on a fitted model to the measured points, the distance to the prediction location, and the spatial relationships among the measured values around the prediction location.

Krigging – illustration







Krigging

Ordinary kriging assumes constant unknown mean only over the search neighborhood of $\boldsymbol{s}_0$.

- Matrix form: $x(\boldsymbol{s}_0) = \boldsymbol{\lambda}' \boldsymbol{x} \quad \text{s.t.} \quad \boldsymbol{\lambda}'\mathbf{1} = 1.$
  where $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_N)$ and $\mathbf{1}$ is a vector of ones.
- Krigging – solve: $\min E(x(\boldsymbol{s}_0 - \boldsymbol{\lambda}' \boldsymbol{x})^2 \text{ s.t. } \boldsymbol{\lambda}'\mathbf{1} = 1.$
- Minimum variance of $x(\boldsymbol{s}_0)$: $\sigma^2 = \sum_{i=1}^{N} \lambda_i \gamma(\boldsymbol{s}_i, \boldsymbol{s}_0) \; + \; m$
- is obtained when $\sum_{i=1}^{N} \lambda_i \gamma(\boldsymbol{s}_i, \boldsymbol{s}_j) \; + \; m = \gamma(\boldsymbol{s}_j, \boldsymbol{s}_0)$ for all j.

  Here, γ is the semivarogram, m is an additional LM parameter (mean estimate) that ensures unbiasedness of the estimate.
  For additional discussion and alternative estimation methods (simple, universal krigging, etc.), see
  `https://en.wikipedia.org/wiki/Kriging`

Definition of neighbors

- Fotheringham et al (2002): "Spatial dependency is the extent to which the value of an attribute in one location depends on the values of the attribute in nearby locations."
- Different definitions of spatial dependency are possible.
- To discuss spatial dependency, spatial autocorrelation, corresponding tests and spatial econometric models, we need to formalize the concept of **nearby locations – neighbors**

Definition of neighbors

- **Distance-based approach** defines two units as neighbors if their distance does not exceed some ad-hoc predefined threshold: τ.
  - Can generate "islands" (units with zero neighbors), if τ is low compared to minimum distances among unit pairs.
  - Less suited for analysis of areas with uneven geographic density (of measurements).
- **Centroids** are used for measuring distances between units with non-zero areas (e.g. regions)
  - Centroids can be purely geographical, "main" city locations, population-weighted, transportation-weighted (highway/railway), etc.

Definition of neighbors

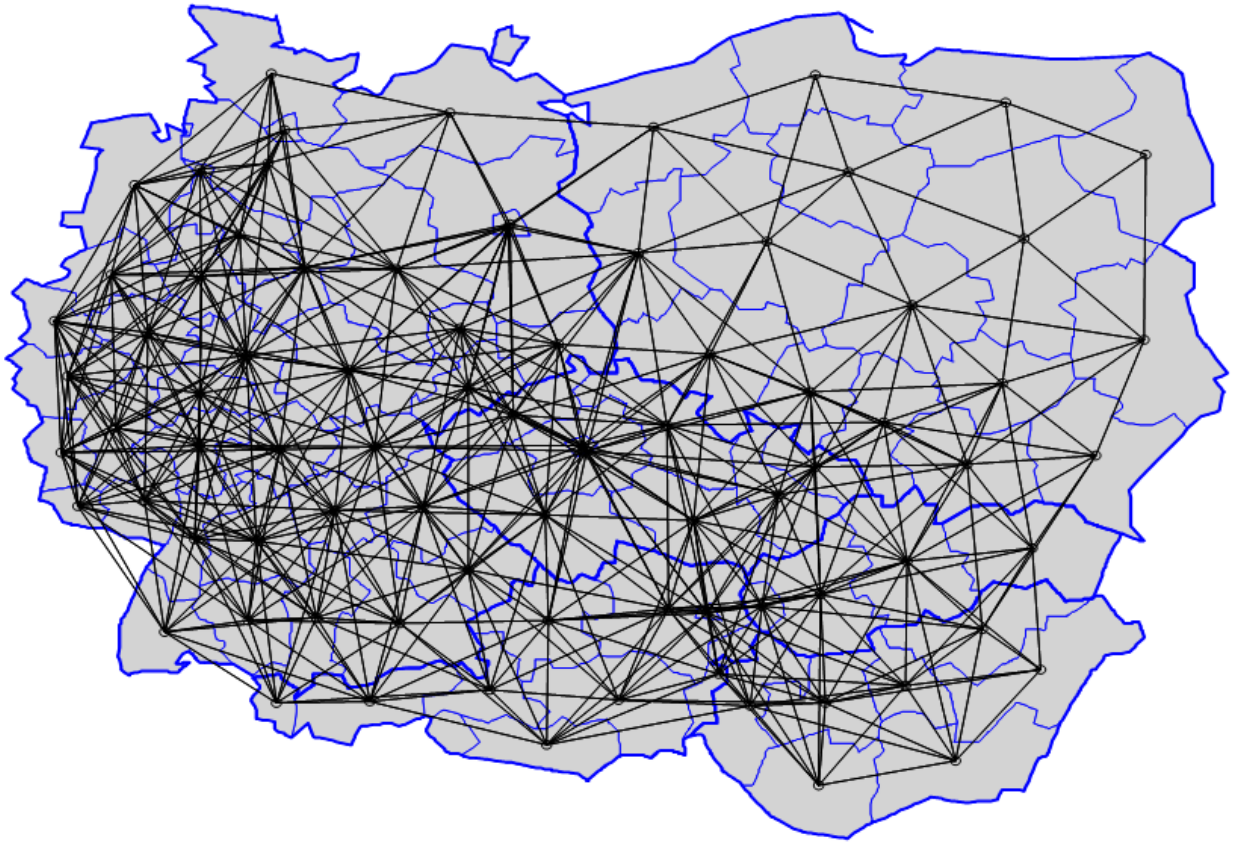


Figure: Plot for distance-based neighbors (NUTS2), maximum neighbor distance threshold at 250 km

Definition of neighbors

- **Contiguity-based approach** spatial units (regions) are neighbors if they share a common border (at least one point).

- **Generalized contiguity approach** is based on the premise that a "second order" neighbor is the neighbor of a first order neighbor (the actual contiguous neighbor).

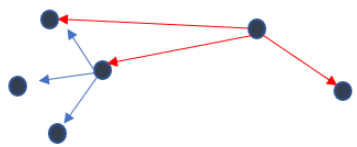
  With this type of approach, we can define a maximum neighborhood lag (order) to control for the highest accepted number of neighbors traversed (not permitting cycles).

Definition of neighbors

k Nearest neighbors (kNN)

For each spatial unit, we search for a preset number of k nearest units that we define as its neighbors.

- Solves for differences in areal densities (k neighbors are ensured for each unit).
- Usually leads to asymmetric spatial connectivity matrices with potentially flawed neighborhood interpretation.
- Illustration for $k = 3$ (neighbors only shown for 2 units):



Spatial connectivity matrix ($\boldsymbol{C}$)

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \qquad \text{a 4-unit example}$$

$$c_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are neighbors,} \\ 0 & \text{if } i \text{ and } j \text{ are not neighbors.} \end{cases}$$

- Zeros on the diagonal – units are not neighbors to themselves.
- Spatial connectivity matrix interpretation:
  - row/column 1: unit 1 is neigbor to units 2,3,4
  - row/column 2: unit 2 is neigbor to units 1,3 (not 4)
- Matrix $\boldsymbol{C}$ is symmetric (for kNN, transformations are available).

Spatial weights matrix ($\boldsymbol{W}$)

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \rightarrow \boldsymbol{W} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

- Usually, $\boldsymbol{W}$ is row-standadrized (to unity): $w_{ij} = \frac{c_{ij}}{\sum_{j=1}^{N} c_{ij}}$.
- Binary (connectivity) indicators c_{ij} may be generalized prior to standardization. Hence, instead of binary c_{ij}, we may use inverse distances (linear, squared, ...) – given assumed decay in spatial dependency over distance.

  Validity of any such prior information (decay pattern) will influence subsequent analysis (spatial dependency tests, spatial regression models).

Spatial weights matrix ($\boldsymbol{W}$)

$$\boldsymbol{W} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix}$$

- Each row of $\boldsymbol{W}$ "provides" weights for an expected value of an observed spatial variable y_i – weighted averages (fitted values) can be calculated as $\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{y}$. For example:

$$\hat{y}_1 = \frac{1}{3}y_2 + \frac{1}{3}y_3 + \frac{1}{3}y_4$$

$$\ldots$$

$$\hat{y}_4 = \frac{1}{2}y_1 + \frac{1}{3}y_3$$

- Observed values of variables can be used to predict corresponding values for neighboring spatial units.
- In this context, $\hat{y}_i$ is often called *spatial lag* of y_i.

Definition of neighbors

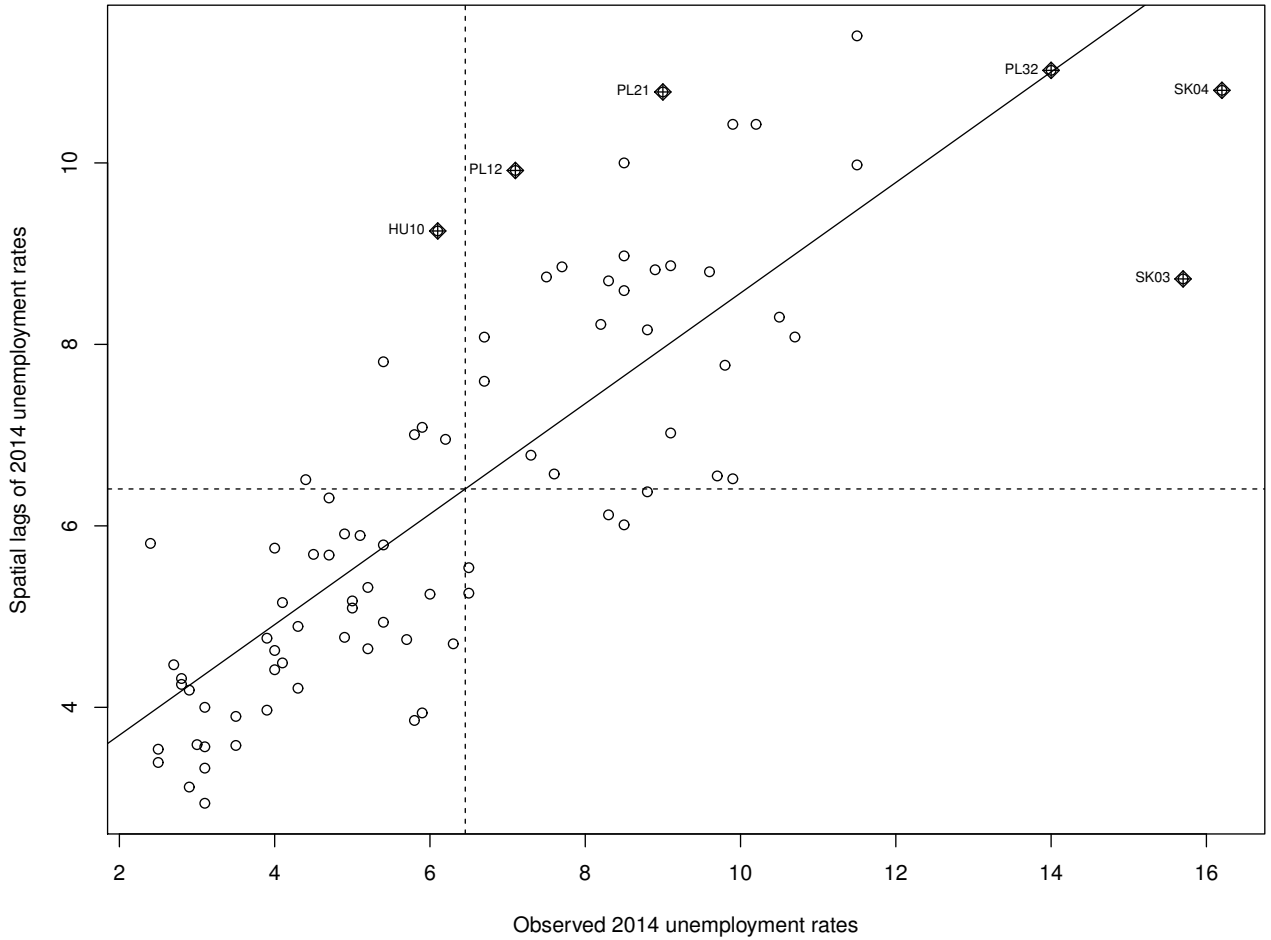


Figure: Moran plot for unemployment rate, 2014, NUTS2 (AT, CZ, DE, HU, PL, SK): observed values vs. spatial lags.

Sample selection in spatial data analysis

- Spatially autocorrelated processes are defined in terms of individual units and their interactions with neighbors.
- Clearly, we can only assess the impact of neighboring units if such units are part of our sample. Hence, in spatial econometrics, we usually do not draw limited samples from a particular area.
- Instead, we work with data from adjacent units located in unbroken ("complete") study areas.
- Otherwise, $\boldsymbol{C}$ and $\boldsymbol{W}$ matrices would be misleading and we could not consistently estimate spatial interactions and effects.
- Generally speaking, spatial analysis should include the whole geographically defined area/region instead of using random sampling (from a "population" of regions within the relevant area).

Spatial dependency tests

- Positive spatial autocorrelation occurs if high or low values of a variable cluster in space.
- For negative spatial autocorrelation, spatial units tend to be surrounded by neighbors with very dissimilar observations.
- Spatially random data lack any spatial pattern.
- Sometimes, spatial dependency patterns are easy to discern visually using choropleths.
- However, a formal approach towards evaluation of spatial dependency is often required.

Moran's I

Measure of global spatial autocorrelation, overall clustering of data:

$$I = \frac{N}{W} \boldsymbol{z}'\boldsymbol{W}\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1},$$

where

N is the number of spatial observations (units) of the variable under scrutiny (say, $\boldsymbol{y}$),

$\boldsymbol{z}$ is the centered form of $\boldsymbol{y}$: $z_i = y_i - \bar{y}$.
(Note a recast: $y_i = \beta_0 + z_i$. If SLRM is expanded by regressors, Moran's I can be applied to regression residuals).

The standardization factor $W = \sum_i \sum_j w_{ij}$ corresponds to the sum of all elements of the spatial weights matrix $\boldsymbol{W}$.
For row-standardized $\boldsymbol{W}$ matrices, $\frac{N}{W} = 1$. Moran's I can be used with $\boldsymbol{C}$ matrices (binary, generalized) as well.

Moran's I

$$I = \frac{N}{W}\, \boldsymbol{z}'\boldsymbol{W}\boldsymbol{z}(\boldsymbol{z}'\boldsymbol{z})^{-1},$$

In most empirical circumstances, $I \in [-1, 1]$. Under the null hypothesis of spatial randomness, Moran's I is asymptotically normally distributed with the following first two moments:

$$E(I) = -\frac{1}{N-1} \quad \text{and} \quad \text{var}(I) = \frac{N^2 W_1 - N W_2 + 3W^2}{(N^2-1)W^2},$$

where $W_1 = \sum_i \sum_j (w_{ij} + w_{ji})^2$ and $W_2 = \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2$.

Normality assumption $\rightarrow$ calculate a z-score

$$z = \frac{I - E(I)}{\sqrt{\text{var}(I)}}$$

and test for statistical significance of Moran's I: whether neighboring units are more similar ($I > E(I)$) or more dissimilar ($I < E(I)$) than they would be under the null hypothesis of spatial randomness.

Local Moran's I

- Moran's I yields only one statistic that summarizes the nature of spatial dependency in the observed variable – it assumes geographical homogeneity (stationarity) in the data.
- If such assumption does not hold (spatial dependency varies over space), then Moran's I test loses power and the "global" statistic is non-descriptive.
- To address this problem, we can use Local Moran's I statistic (for row-standardized $\boldsymbol{W}$):

$$I_i = \frac{z_i N}{\boldsymbol{z}'\boldsymbol{z}} \boldsymbol{w}_i \boldsymbol{z} \,.$$

- The expected value of Local Moran's I under the null hypothesis of no spatial autocorrelation is: $E(I_i) = -w_i/(N-1)$. Here, w_i is the sum of elements in the i-th row of $\boldsymbol{W}$.

Local Moran's I

$$I_i = \frac{z_i N}{\boldsymbol{z}'\boldsymbol{z}} \boldsymbol{w}_i \boldsymbol{z} \, .$$

- Values of $I_i > E(I_i)$ indicate positive spatial autocorrelation, i.e. that the i-th region is surrounded by regions that, on average, are similar to the i-th region with respect to the observed variable y.
- $I_i < E(I_i)$ would suggest negative spatial autocorrelation.
- Significance of spatial dependency is then evaluated using $var(I_i)$ and the corresponding z-score.

- We may see the global nature of Moran's I from

$$I = \frac{1}{N} \sum_{i=1}^{N} I_i \, .$$

Geary's C

Variance test similar (in principle) to the Durbin-Watson test statistic for residuals' autocorrelation in time-series regressions.

$$C = \frac{N-1}{2W} \frac{\sum_i \sum_j w_{ij}(y_i - y_j)^2}{\sum_i (y_i - \bar{y})^2},$$

where all elements follow from previous slides. Empirical Geary's C values range from 0 to 2. However, occurrences of $C > 2$ are possible.

- Under the null of no spatial autocorrelation, first two moments are:

$$E(C) = 1\,, \qquad \text{var}(C) = \frac{(N-1)(2W_1 + W_2) - 4W^2}{2(N+1)W^2},$$

- Positive spatial dependency: $C < 1$.
  Negative spatial autocorrelation is reflected in $C > 1$.
- z-transformation is asymptotically normally distributed.
  Therefore, $z(C)$ can be used for testing spatial randomness.

Clusters – hotspots and coldspots

Getis' G^*: spatial clusters and hotspot analysis

- Clustering analysis by Getis can only be performed for positively autocorrelated spatial data.
- $G_i^*(\tau) = \frac{\sum_{j=1}^{N} c_{ij}^* y_j}{\sum_{j=1}^{N} y_j}$,

  where c_{ij}^* come from amended distance-based (arbitrary τ used) connectivity matrix $\boldsymbol{C}^* = \boldsymbol{C} + \boldsymbol{I}_N$; i.e. y_i observations enter $G_i^*(\tau)$ calculation. Observations of y are assumed to have a natural origin and positive support.
- $G_i^*(\tau)$ is a local stastic, a proportion of the aggregated y_j values that lie within τ of i to the total sum of y_j observations.
- Alternatively, Getis' statistic is calculated using $\boldsymbol{C}$ (not $\boldsymbol{C}^*$) and denoted $G_i(\tau)$

Clusters – hotspots and coldspots

- If we observe high values of y_j within distance τ of unit i, then $G_i^*(\tau)$ would be relatively high compared to its expected value under the null hypothesis of full spatial randomness:

$$E\left[G_i^*(\tau)\right] = \frac{c_i^*}{N},$$

  where c_i^* is the sum of elements of i-th row of $\boldsymbol{C}^*$.

- Also, under the H_0 of spatial randomness, we can write

$$\text{var}\left[G_i^*(\tau)\right] = \frac{c_i^*(N - c_i^*)}{N^2(N-1)}\left(\frac{Y_{i2}^*}{(Y_{i1}^*)^2}\right),$$

  where $Y_{i1}^* = \frac{\sum_j y_j}{N}$ and $Y_{i2}^* = \frac{\sum_j y_j^2}{N} - (Y_{i1}^*)^2$.

- High positive z-score indicates "hotspot" (cluster of high values) and vice versa. Critical values provided by Getis and Ord. Say, for $N = 100$ and $\alpha = 5\%$, the z-scores would have to exceed ± 3.289 for a statistically significant hot/cold spot.

Clusters – hotspots and coldspots

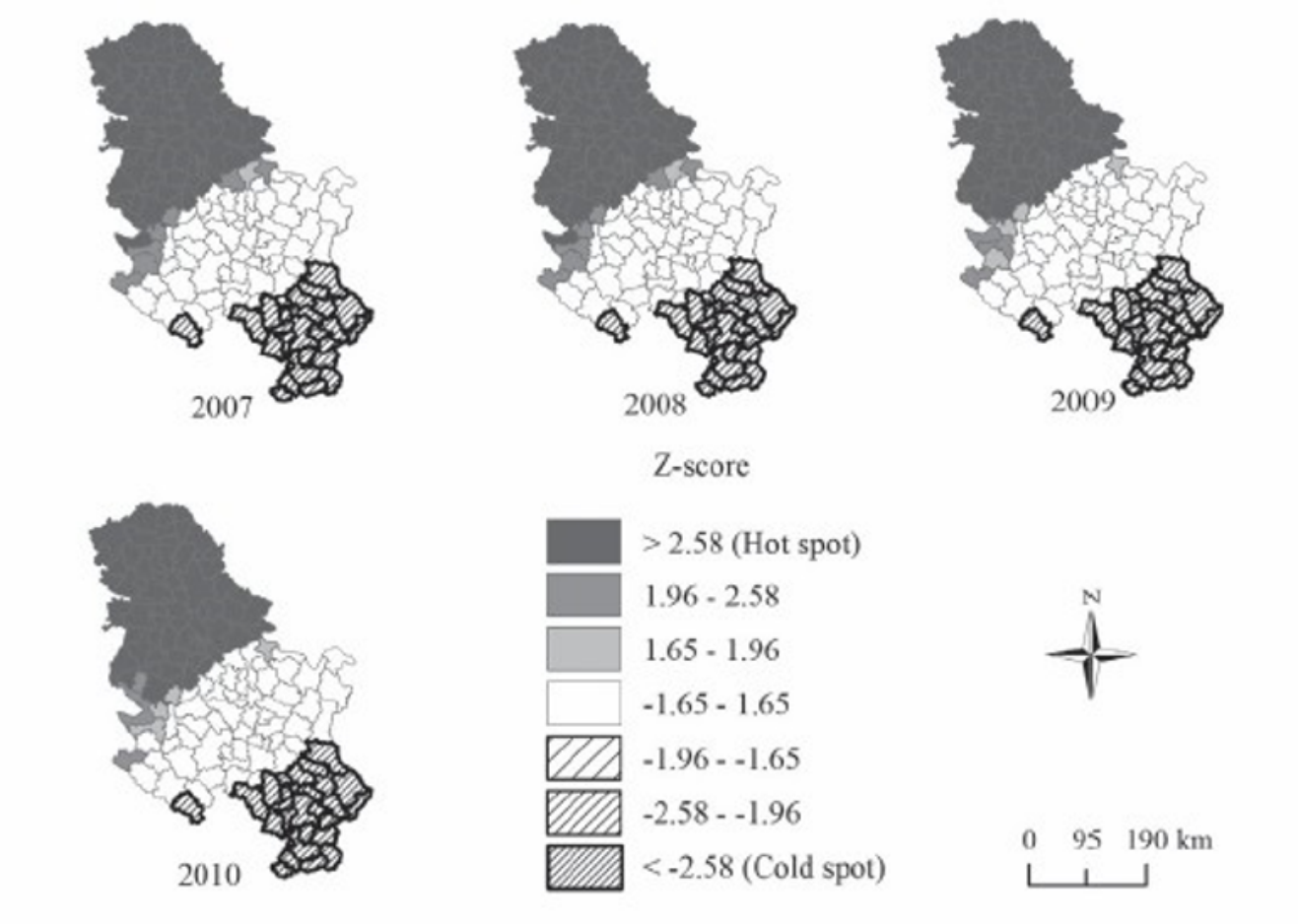


Figure: Spatial clusters (hot and cold spots) of the municipalities in Serbia by the level of average monthly net earning from 2001 to 2010

Spatial regression models

Spatial regression models

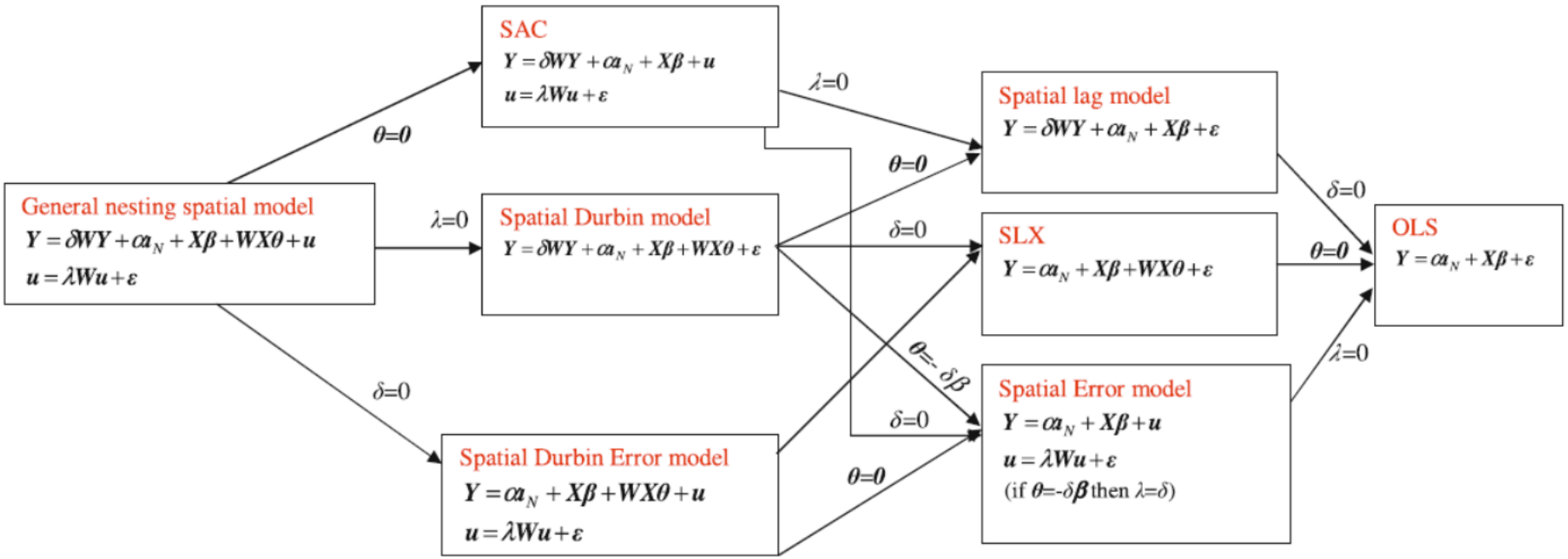


Figure: The relationship between different spatial dependence models for cross-sectional data (source Halleck Vega and Elhorst 2012)