Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 6 (3): 237-255, 2011
ISSN 1819-3579 / DOI: 10.3923/tasr.2011.237.255

Evaluation of Stochastic Geographical Matters: Morphologic Geostatistics, Conditional Sequential Simulation and Geographical Weighted Regression

J. Negreiros, A.C. Costa and M. Painho
ISEGI B Universidade Nova de Lisboa, Campus de Campolide, Lisboa, Portugal

Corresponding Author: J. Negreiros, ISEGI B Universidade Nova de Lisboa, Campus de Campolide, Lisboa, Portugal

ABSTRACT

The aim of this study is to highlight four main stochastic modeling procedures for spatial data within Geographical Information Systems (GIS) which are still unknown by most GIS users: Morphologic Geostatistics (MG), Geographical Weighted Regression (GWR), Conditional Sequential Simulation (CSS) for continuous and categorical variables. Sequential simulation, for instance, is a widely used geostatistical tool for obtaining a set of equiprobable simulated realizations of variables from natural phenomena, conditional to observed data, honoring their spatial distribution and uncertainty. While Gaussian simulation involves the generation of many independent realizations of a Gaussian random field but requiring the transformation of original variables, direct sequential simulation (DSS) has been proposed for simulating directly in the original data space and does not rely on multi-Gaussian assumptions. A generic Pb contamination dataset is used to illustrate the MG and CSS procedures. Major relationships among Kriging estimation, spatial autocorrelation, geographical regression and the missing data issue are also reviewed in the last section.

Key words: Geographic information systems, spatial analysis, morphologic geostatistics, conditional sequential simulation, geographical weighted regression

INTRODUCTION

The problem of statistical spatial analysis encompasses an expanding range of methods which address different spatial problems, from image enhancement and pattern recognition to spatial interpolation and socio-economic trend modeling. Each of these methods focuses on a particular aspect but what emerges is something that is clearly identifiable as spatial statistics, statistical methods like MG, CSS and GWR techniques which address geographical raw data that are spatially correlated.

According to Ferreira and Simões (1993, 1994), the kernel of geography is to think geographically, that is, to study the spatial distribution of the phenomena and their correlations. Traditional statistics must be reformulated to properly account for spatial correlation and spatial heterogeneity within georeferenced data (Anselin, 1996, 1998). Spatial autocorrelation is a reality but also a requirement to carry out spatial interpolation and spatial regression modeling. For instance, if regression residuals reveal a medium-strong spatial autocorrelation then any missing variable within the initial regression model can be significant. Certainly, MG, CSS and GWR try to include this spatial autocorrelation issue within their spatial interpolation and simulation computation.

To re-examine these procedures becomes, hence, the goal of this research. Section two highlights Morphologic Geostatistics (MG) while sections three and four underline Conditional Gaussian Sequential Simulation (CGSS) and Conditional Categorical Sequential Simulation (CCSS), respectively. Direct Sequential Simulation (DSS) is also introduced while Geographical Weighted Regression (GWR) is presented in section five. Close relationships among these spatial issues are presented in present study.

It is vital to stress that the comprehension of this essay underlines the knowledge of variography and Kriging topics (Aguilar *et al*., 2008; Negreiros *et al*., 2010). Additionally, the input Pb spatial data, exploratory, variography and Kriging results that follow section two and three are presented in the Appendix (contamination data of Aljustrel, Portugal, within a global area of 4500 m×2950 m). The software used here was GeoMS[8] from Instituto Superior Técnico, Lisbon.

MORPHOLOGIC GEOSTATISTICS

This section is dedicated to MG for categorical variables (body X and body X^C) such as the presence/absence of a particular type of pine tree in an open field, for example. The main idea is

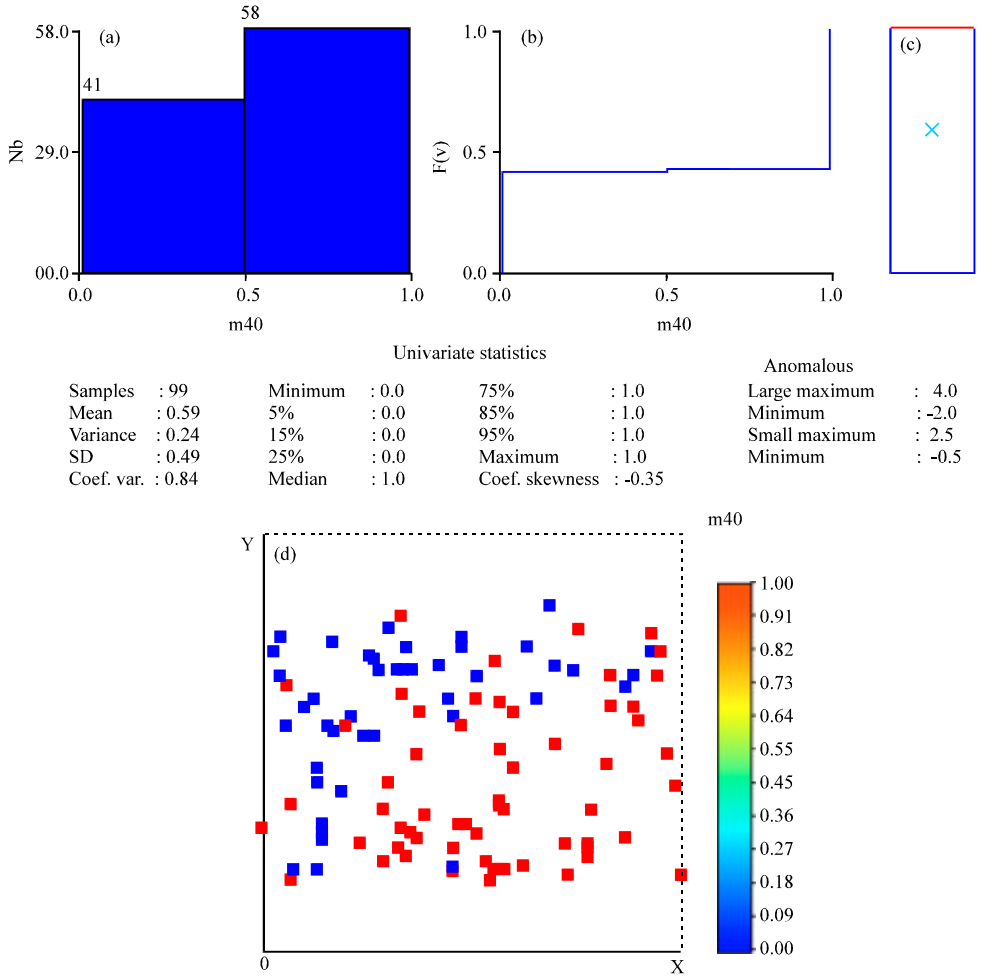


Fig. 1: (a-d) Histogram, cumulative distribution functions, box-plot, descriptive statistics (top) and spatial location map (bottom) of the categorical variable (blue dots versus red ones) based on Eq. 1

to analyze the spatial structure for problems closely related with qualitative variables which is useful to understand the degree of presence of a certain type of objects in space. Quite often, farmers do not want to measure the number of grasshoppers (quantitative variable) but wish to classify sub-regions according to a pre-define classification (categorical variable) such as lower, medium and high infestation.

For illustration purposes lets consider the Pb contamination dataset (Appendix). By considering an indicator variable whose cutoff value equals the median of the Pb values, a categorical variable can be created according to the next rule:

$$I_{Morf}(x) = \begin{cases} 1 \text{ if } x > 40 \in X \\ 0 \text{ if } x <= 40 \in X^c \end{cases} \quad (1)$$

That is, a particular x location equals 1 if that particular site is classified as an X body (all x>40 ppm). Otherwise, it equals 0, which means that the location is classify as X^C (all x No. 40 ppm). It is central to highlight that this I_{morf} (x>40 ppm) condition (body X) used in this example works as a classification criteria for the present contamination dataset creating, thus, a two-phase structure: population X and population X^C (Fig. 1a-d).

As expected, the spatial continuity of both bodies can be measured by the indicator variogram based on the sill, anisotropy (range) and nugget-effect parameters. In this morphologic context, the range factor measures the average dimension of X and X^C bodies while the nugget-effect denotes the transition frequency between both bodies (1 and 0 states) at small scale (Fig. 2). The variogram that lies behind this spatial autocorrelation structure is a spherical one with a major and minor range of 2500 and 1000 m, respectively. The nugget-effect equals zero. The main direction of the geometric anisotropy is 901.

The probability map of Fig. 2 can be transformed into a morphologic binary map to reproduce the X and X^C bodies as described next (Soares, 1990):

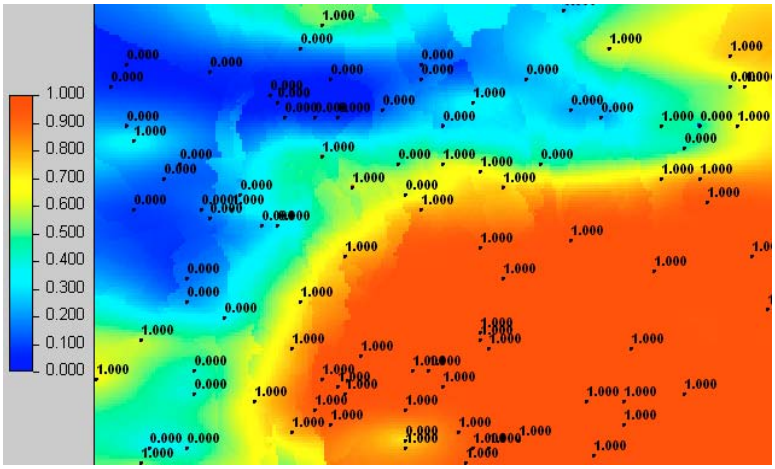


Fig. 2: Probability estimation map for the X body obtained by Indicator Kriging of the I_{Morf} variable (the highest probabilities of Pb to be greater than 40 ppm are located at the lower right corner)

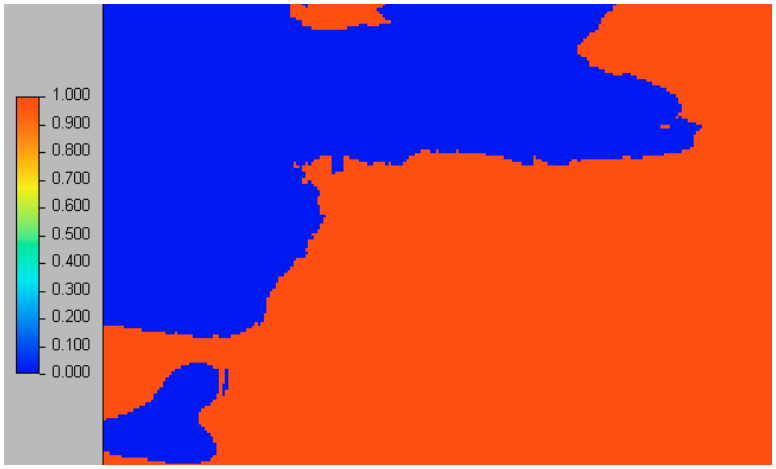


Fig. 3: The two-strata structure morphologic map (body X in red and body X^C in blue)

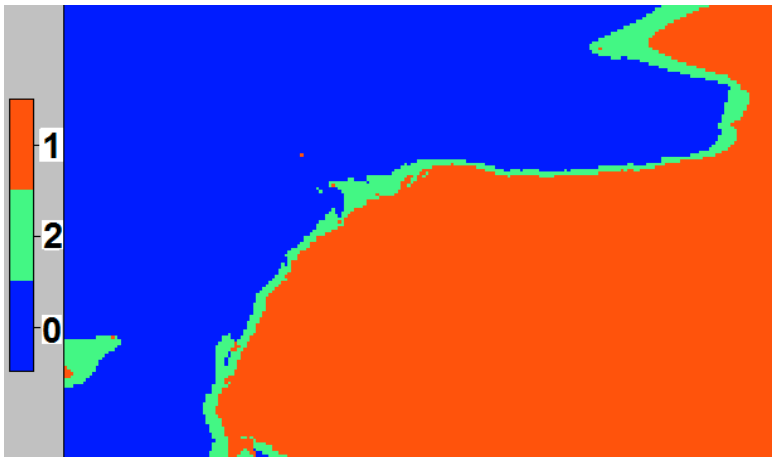


Fig. 4: The previous two-phase structure morphologic map including uncertainty regions

- Determine the global average, $(m_X)^*$, of the Indicator Kriging probabilities estimates. Averaging all grid locations = probabilities allows obtaining an estimate of the proportion of locations of the region that should be classified as X
- Sort (descending) the probabilities estimates of all grid locations
- Create a new binary variable according to the following rules: assign the value 1 to all locations with probability estimate greater than $(m_X)^*$; assign the value 0 otherwise
- Create a map of this new binary variable

In the Pb example, $(m_X)^*$ equals 0.605786 which means that the body X covers 60.6% of the region. Among the 34126 kriged blocks of Fig. 2, the 20673 locations (34126H0.605786) with the highest probabilities were assigned the value 1 and the remaining ones were set to 0. Mapping this new binary variable allows to produce the estimated morphologic map of the body X (Fig. 3).

Concerning the uncertainty assessment of the morphologic map, which is higher on the border between different bodies, the following four steps were computed (Fig. 4):

- Determine the global average, $(m_X)^*$, of the Indicator Kriging probabilities estimates
- Sort (descending) the probabilities estimates of all grid locations
- Create a new variable according to the following rules: assign the value 1 to all locations with probability greater than $105\%H(m_X)^*$; assign the value 0 to those smaller than $95\%H(m_X)^*$; otherwise, the new variable will assume the value 2
- Generate a map based on this new variable with three phases

In the example:

- $105\% \times (m_X)^* = 0.636075$
- $0{:}95\% \times (m_X)^* = 0.575497$
- Total number of locations with values equal to 2: 34126H10%=3412

CONDITIONAL GAUSSIAN AND DIRECT SEQUENTIAL SIMULATION

Interpolation methods, such as Ordinary Kriging (OK), typically overestimate small values and underestimate large ones. As with the traditional regression, the final estimates are less variable than the true values. Although a kriged map shows the best estimates of a variable, it does not represent the variability in a proper way, that is, this loss of variance could lead to wrong decisions. To estimate a surface that retains the original samples variability, we need other techniques such as geostatistical stochastic simulation (Webster and Olivier, 2007).

Within spatial analysis, the stochastic simulation is a tool to evaluate the spatial uncertainty based on the generation of N sets of equiprobable values (Soares, 2000). Each simulated image should hold the statistical properties of the original observations, such as the mean, variance, histogram and covariance. Additionally, simulated values honor, at their locations, the measured data values. Based on the N realizations, it is possible to compute uncertainty measures and to assess the extreme spatial behavior of any particular phenomenon. For instance, from a set of maps, it is possible to assess all sub-regions with values above a certain threshold and the probability of that to happen. Even more important for pollution studies, it is possible to generate scenarios yielding the smallest (best scenario) and the largest (worst scenario) contaminations costs.

Broadly, there are three classes of simulation algorithms B sequential, p-field and annealing-although only the first one will be covered in this study. The Conditional Gaussian Sequential Simulation (CGSS) technique is based on the assumption that the spatial distribution of a continuous random variable can be modeled by a multivariate Gaussian model. Through the sequential simulation algorithms, instead of modeling the N-points conditional cumulative distribution function (ccdf), a one-point ccdf is modeled and sampled at each of the N nodes visited along a random sequence. To ensure the reproduction of the variable covariance model, each one-point ccdf is made conditional not only to the original data but also to all values simulated at previously visited locations (Goovaerts, 1997).

With CGSS, each variable is simulated sequentially according to its Normal ccdf, which is fully characterized through a Simple Kriging (SK) system. The conditioning data consist of all original data and all previously simulated values found within a neighborhood of the location being simulated. The conditional simulation of a continuous variable $Z(u)$ modeled by CGSS proceeds, then, as follows:

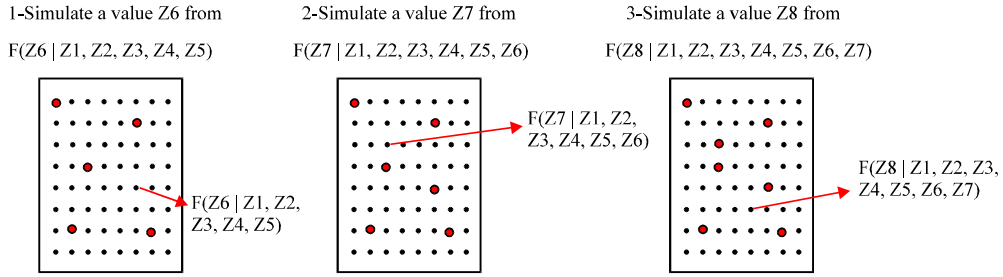


Fig. 5: Random path example of the sequential simulation technique (Soares *et al.*, 2010)

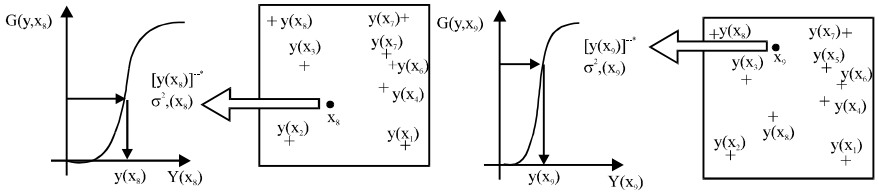


Fig. 6: Diagram of step 5 (Soares *et al.*, 2010)

- Based on all available samples, compute the univariate cumulative distribution function (cdf) representative of the entire study area
- Using this cdf, it is necessary to perform a normal-score transformation of the z(u) original data into y-data (the Gaussian restriction)
- Define a random path that visits each node of the grid once (Fig. 5). At each node u, keep hold of a specified number of neighboring conditioning data including both initially transformed y-data and previously simulated grid node values (the conditional restriction)
- Apply Simple Kriging (SK) with the Normal score variogram model (based on the y-values) to determine the estimate $z^*(u)$ and the variance $\sigma^*(u)$, at the u location
- Randomly draw a value between 0 and 1 from the cumulative Gaussian distribution based on step 4 results, that is, $N(z^*(u), \sigma^*(u))$. The simulated value is then equal to $G^{-1}(z^*(u), \sigma^*(u))$. Afterwards, this simulated value is added to the current dataset (Fig. 6)
- Proceed to the next node and loop until all nodes are simulated (the sequential restriction)
- Back-transform the simulated normal values $y^*(u)$ into the original variable for each location u (Fig. 7a-f)

Journel (1994) showed that for this conditional sequential simulation algorithm to reproduce a specific covariance model it suffices that the simulated values are drawn from the local distributions centered at the Simple Kriging estimates with a variance corresponding to the Simple Kriging estimation variance. This result guarantees that the spatial covariance and the global sample mean and variance, of the original variable are reproduced but not the histogram. To overcome this limitation, Soares (2001) proposed a Direct Sequential Simulation (DSS) algorithm that uses the local Simple Kriging estimates of the mean and variance, not to define the local cumulative distribution function (cdf) but to sample from the global cdf.

Recently, Costa *et al.* (2008a) and Costa and Soares (2009) proposed a new method for the homogenization of climate data using the DSS algorithm. The DSS procedure is used to calculate the local probability density function (pdf) at a candidate station's location. The algorithm generates

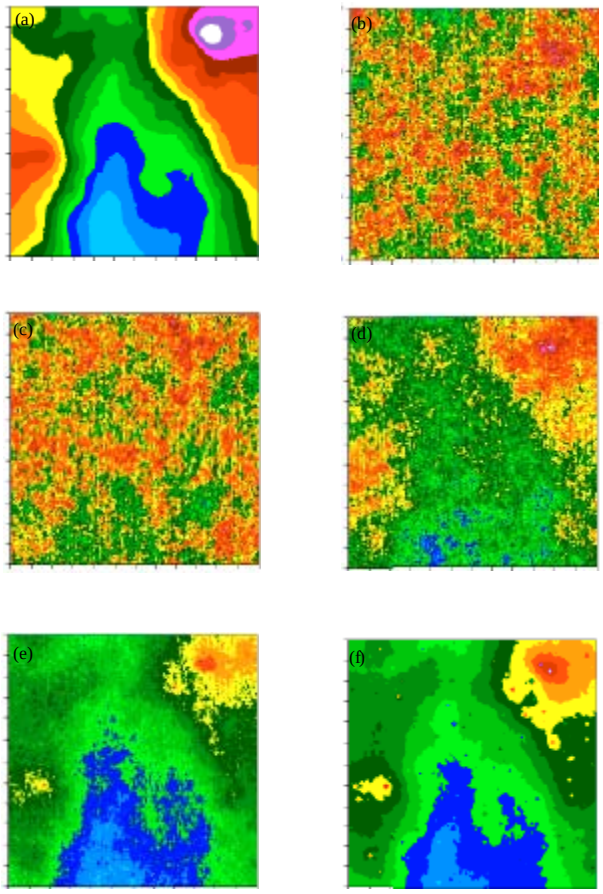


Fig. 7: With conditional simulation, the number of realizations used to produce the final estimates can strongly affect the outcome of the interpolation. The upper six images correspond to 80 m×80 m grid interpolated maps (density of 0.5 m) with different numbers of simulations: (a) block Kriged data; (b) the same data interpolated for n = 1 conditional simulation; (c) n = 10 simulations; (d) n = 100 simulations; (e) n = 1000 simulations; (f) n = 10000 simulations (Robertson, 2008)

realizations of the climate variable through the resampling of the global pdf using the local mean and variance of the candidate station, which are estimated through a spatiotemporal model. The local pdf from each instant in time is used to verify the existence of irregularities: a breakpoint is identified whenever the interval of a specified probability p, centered in the local pdf, does not contain the observed (real) value of the candidate station. If irregularities are detected in a candidate series, the time series can be adjusted by replacing the inhomogeneous records with the mean of the pdf(s) calculated at the candidate station's location for the inhomogeneous period(s).

Recently, Costa *et al.* (2008b) used the direct sequential cosimulation (coDSS) algorithm to map a flood indicator and extreme precipitation frequency in Southern Portugal using elevation as auxiliary information. The methods incorporate space-time models that account for long-term trends of extreme precipitation and local changes in the relationship between elevation and extreme precipitation through time.

SEQUENTIAL SIMULATION FOR CATEGORICAL VARIABLES

Variables, such as the concentration of a metal in the soil, may appear to change abruptly in space. In this case, the phenomenon should be modeled as a mixture of two populations each of them may have different patterns of spatial continuity (Goovaerts, 1997). To deal with this type of variables, Sequential Indicator Simulation (SIS) is presented in this section as it allows modeling the relative geometry of each population (strata) in order to create an exhaustive categorical map.

Analogous to the CGSS, seven major steps are involved in the SIS algorithm:

- Transform each categorical data (e.g., tillage, meadow, pasture and forest) into a vector of 1s and 0s (e.g., (0,0,0,1) if a particular site is classified as forest only)
- Assess the occurrence probability for each category using Indicator Kriging (IK) at all locations
- Correct these category probabilities in terms of relation order
- Build the cumulative distribution function (cdf) at each spatial node. As expected, the sum of the probabilities of all categories at each location equals one
- Draw a random number p between 0 and 1 from that cdf. The simulated category at that location is the one that corresponds to the probability interval that includes p
- Add the simulated value to the conditioned dataset
- Proceed to the next location and repeat steps 3 to 6

Sequential simulation will be exemplified for non-continuum variables using the input dataset detailed in Appendix. First, a new categorical variable with four classes (Fig. 8a-d), named CI (u), is computed using the Pb contamination data and Eq. 2:

$$CI(u) = \begin{cases} 1 \text{ if } z(u) < Q_1 \\ 2 \text{ if } Q_1 \le z(u) < Q2 \\ 3 \text{ if } Q2 \le z(u) < Q_3 \\ 4 \text{ if } z(u) \ge Q_3 \end{cases} \quad (2)$$

Where:
Q_1 = first quartile = 36.2 ppm
Q2 = median = 43.6 ppm
Q_3 = third quartile = 59.2 ppm
u stands for a particular spatial location

Several experimental variograms for the four phase variable were computed according to two main directions (geometric anisotropy): 0° (N/S direction) and 901 (E/W direction). Notice that this multiphase variogram equals:

$$\gamma(h) = 0.5HE\left(\sum_{b=1}^{4}[I_b(u) - I_b(u+h)]\right)$$

that is, each variogram lag corresponds to a global average of four individual covariances (Fig. 9a, b).

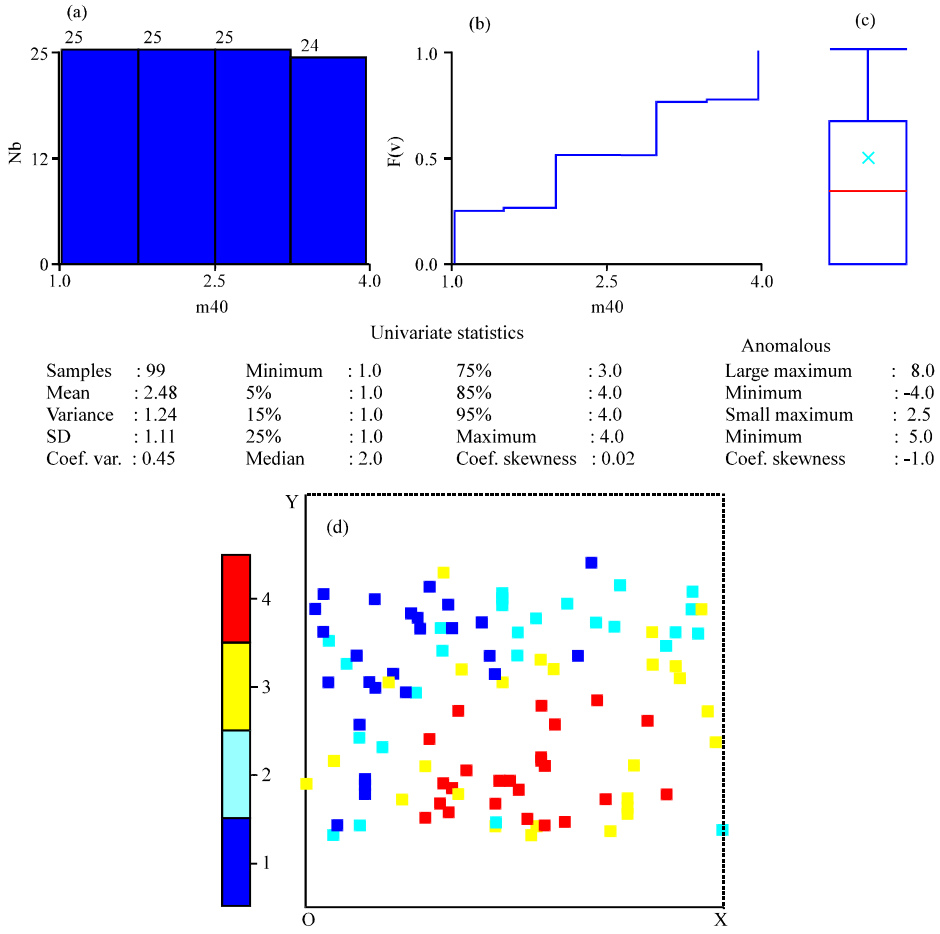


Fig. 8: (a-d) Histogram, cumulative distribution function, box-plot and univariate statistics (top) spatial layout (bottom) of the categorical variable CI(u)

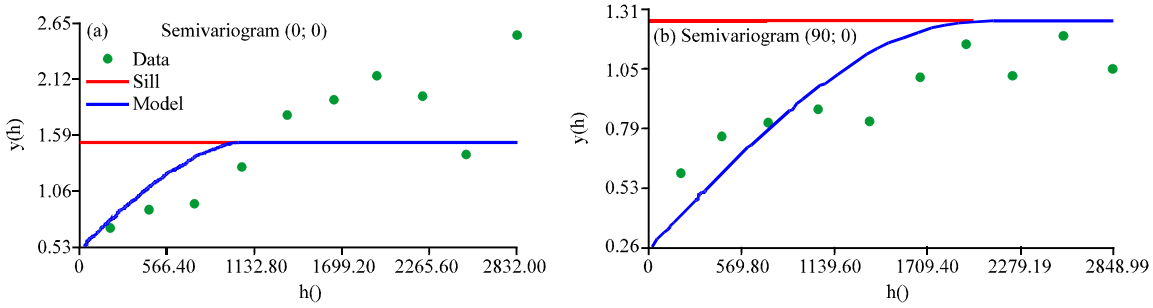


Fig. 9: (a, b) Average variogram for the multiphase CI(u) variable adjusted by a spherical model with the sill equal to 1.24, the major range equal to 2000 m (E/W direction, bottom) and the minor one equal to 1200 m (N/S direction, top)

Regarding the simulated data of Fig. 10, it is curious to confirm that the histogram of these 34126 values estimated by the Sequential Indicator Simulation (SIS) is similar to the original

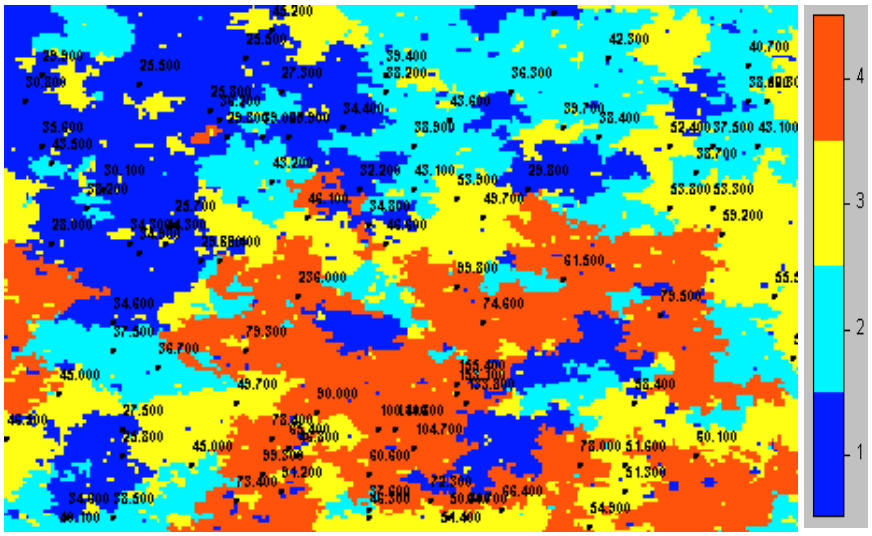
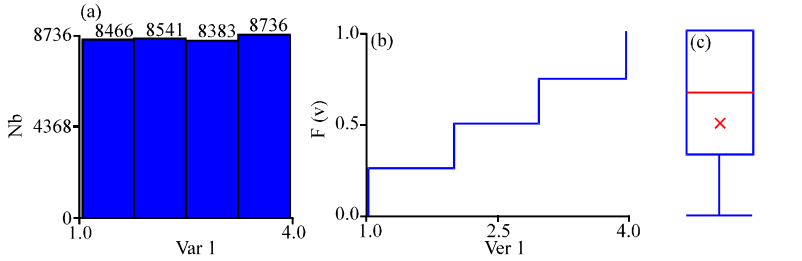


Fig. 10: Example of an indicator realization map of CI(u) generated by SIS



Univariate statistics

Samples	: 34126	Minimum	: 1.0	75%	: 4.0	Anomalous	
Mean	: 2.51	5%	: 1.0	85%	: 4.0	Large maximum	: 9.0
Variance	: 0.24	15%	: 1.0	95%	: 4.0	Minimum	: -3.0
SD	: 1.26	25%	: 2.0	Maximum	: 4.0	Small maximum	: 6.0
Coef. var.	: 0.054	Median	: 3.0	Coef. skewness	: -0.01	Minimum	: 0.0

Fig. 11: (a-c) Histogram, cumulative distribution function, box-plot and univariate statistics of the simulated values

distribution histogram (Fig. 11a-c). It is also important to highlight that all simulated data honors the experimental data.

In average, the simulated spatial images hold the same statistical features of the original observations. In fact, the variograms of the simulated map hold a similar shape (Fig. 12a, b). In this case, a spherical model with a major range of 2000 m (E-W direction) and a minor one of 1200 m (N-S direction) was fitted. The fluctuations between both models are known as ergodic fluctuations (Goovaerts, 1997). In fact, Gaussian and Indicator simulation algorithms only reproduce the original observations variogram in the presence of an average simulated map that results over many realizations. The ergodic fluctuations of the realizations' variograms are generally important when the range of the variogram model is larger with respect to the size of the simulated area (particularly if the relative nugget-effect is small).

As already stated before, Indicator Kriging (IK) allows to estimate the conditional cumulative distribution function (ccdf) for a particular spatial location. By defining L cutoff values, it is then

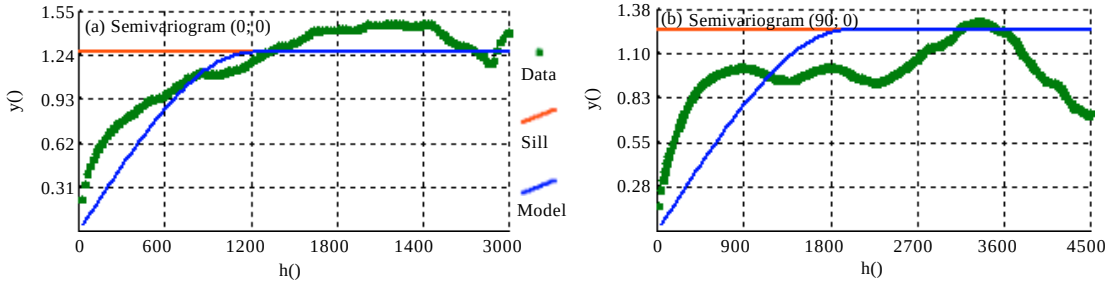


Fig. 12: (a, b) Comparison of the SSI multiphase experimental variogram with the one fitted to the original data in Fig. 9.

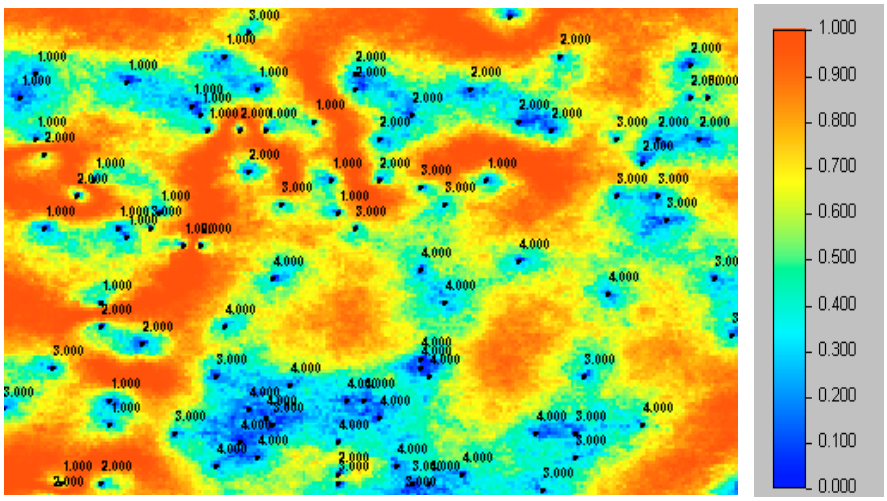


Fig. 13: Local entropy map generated by the 100 indicator simulation maps

possible to infer about the uncertainty probability of each class (ccdf difference of any adjacent cutoffs). Using these partial probabilities, the Shannon entropy procedure allows inferring the local uncertainty of the interpolation at unknown sites. Given L sets of IK probabilities at any u location $P_i(u)$, with i = 1...L, the Shannon entropy (a disorder measure closely connected to the spatial organization of an attribute) equals BSUM($P_i(u)$H LN($P_i(u)$)), where LN() denotes the Neperian logarithm, $P_i()$ is the IK estimation probability for each class while BSUM() denotes the negative sum of all classes' probabilities at that u location. As expected, red color signifies high lack of estimation confidence while dark blue denotes low uncertainty (Fig. 13). Unsurprisingly, uncertainty is smaller near the samples locations.

GEOGRAPHICALLY WEIGHTED REGRESSION (GWR)

Statistically, Ordinary Least Squares (OLS) regression is a technique that allows to relate k-1 independent variables (X_1YX_{k-1}) to a dependent one (Y) in the following form: $Y = X_0+X_1\beta_1+X_2\beta_2+Y+X_{k-1}\beta_{k-1}+\epsilon$. Denoting by n the number of sub-regions considered in the spatial problem, Y is a (nH1) vector, X is a (nHk) matrix, β is a (kH1) regression coefficients vector concerning each independent variable and X is a (n×1) residuals vector (Fig. 14). The regression

errors should follow a Gaussian distribution with zero mean and constant variance σ^2 (Druck *et al.*, 2004).

The major aim of regression analysis is to uncover which variables contribute in a significant way for the linear relationship of the dependent variable. Still, it is expected, among other factors, that regression errors are independent. According to Anselin (1992, 1994), this OLS model fails quite often due to the occurrence of spatial autocorrelation. The global spatial lag model (SAR) for stationary processes overcomes this drawback. It is defined by $Y = \rho WY+X\beta+\epsilon$ where, W represents the neighborhood matrix and ñ is the spatial autoregressive coefficient. Another possibility is to use different spatial regimes (i.e., an individual regression for each region) in SpaceStat[8]. Polynomial trends is another option to work with the region overall tendency and non-stationary phenomenon. Another possibility is the local continuous variation framework computed by the Geographically Weighted Regression (GWR) system. The idea is to adjust a regression model by weighting the neighborhood observations. In this way, the estimation computation will reflect automatic adjustments according to the distance of the available samples.

GWR is a specific model which allows representing non-stationary local phenomena by generating a separate regression equation for every feature analyzed as a means to address spatial variation (Fotheringham *et al.*, 2002). Thus, GWR allows the modeling of processes that vary over space. Since, it usually works with aggregation data, this inferential model considers that spatial data may change abruptly (or not) at region boundaries only (Fig. 15).

Another example of this method is presented by Legg and Bowe (2009) regarding the analysis of the listed sales price for single family houses in Marquette, Michigan and it is based on location and three other variables: number of bedrooms and bathrooms, house square footage and lot size (Fig. 16). According to these authors, the OLS model was found to be significant and had a high $R^2 = 0.782$. Yet, the GWR model improved on this statistic and increased the model's goodness of fit to an $R^2 = 0.865$. In addition, the range of the residual error decreased by $160,000 when using the GWR model instead of the Ordinary Least Square (OLS) model. The coefficients surface was

$$\begin{bmatrix}\gamma_1\\ \gamma_2\\ \vdots\\ \vdots\\ \gamma_n\end{bmatrix}=\begin{bmatrix}1 & X_{11} & \cdots & X_{1k-1}\\ 1 & X_{21} & \cdots & X_{2k-1}\\ \vdots & \vdots & \cdots & \vdots\\ \vdots & \vdots & \cdots & \vdots\\ 1 & X_{n1} & \cdots & X_{nk-1}\end{bmatrix}\begin{bmatrix}\beta_0\\ \beta_1\\ \vdots\\ \vdots\\ \beta_{k-1}\end{bmatrix}+\begin{bmatrix}\epsilon_1\\ \epsilon_2\\ \vdots\\ \vdots\\ \epsilon_n\end{bmatrix}$$

Fig. 14: The Ordinary Least Squares (OLS) regression system in matrix notation

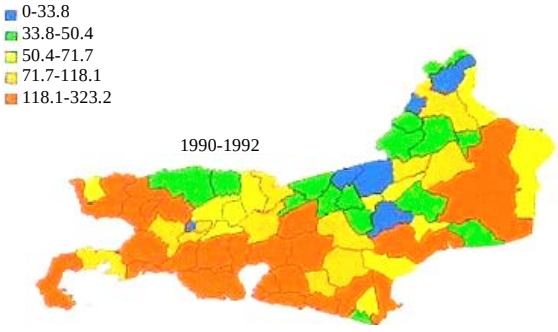


Fig. 15: The number of murders per ward in Rio de Janeiro between 1908 and 1992 (Druck *et al.*, 2004). Values do not vary within each ward because of the polygon spatial structure

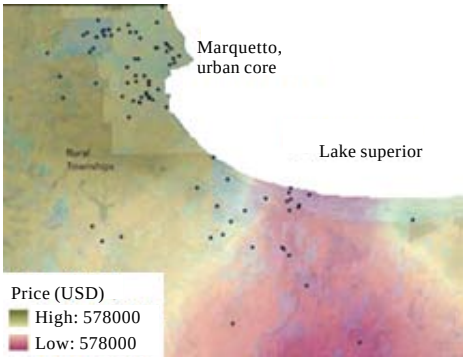


Fig. 16: Listing price map of a typical house modeled using spatially varying regression coefficients generated using GWR tools of ArcGIS© (Legg and Bowe, 2009)

$$W(u_i)=\begin{bmatrix} W_{i1} & 0 & 0 & \cdots & 0 \\ 0 & W_{i2} & 0 & \cdots & 0 \\ 0 & 0 & W_{i3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & W_{in} \end{bmatrix}$$

Fig. 17: $W(u_i)$ is an n by n matrix representing the geographical weights around location u_i. This weighting system is known as kernel

also helpful for identifying the spatial patterns apparent in the study area. For example, the lot coefficient indicates that located groups nearer the urban core and farther from the rural townships increases its price. In contrast, coefficients suggest that the larger the house, the less it contributes to the listing price.

For Lloyd (2007), one key decision emerges from this inferential spatial approach: the choice of a weighting function (kernel shape and kernel bandwidth). Given two independent variables, y_1 and y_2, the local estimation of GWR parameters equals $z(u_i) = \beta_0(u_i)+\beta_1(u_i)Hy_1+\beta_2(u_i)Hy_2$, where z is the dependent variable while (u_i) is the (x,y) location at which the parameters are estimated. By solving the system, the parameters can be estimated by $\beta(u_i) = (Y^THW(u_i)HY)^{-1}HY^THW(u_i)Hz$, where, W (u) is a weight square matrix closely related to the position of u towards the available samples (Fig. 17), $Y^THW(u_i)$ HY represents the geographically weighted variance-covariance matrix (the estimation requires its inverse) while *z* corresponds to the vector of values of the dependent variable (original observations).

Typically, these weights follow the Gaussian weighting function: $W(u_i) = \exp(B0.5(d/b)^2)$ if $d_i \leq b$, where d is the Euclidean distance between the location of the neighborhood observations and the location u_i to be estimated, while b represents the bandwidth of the kernel. As the bandwidth gets larger, the weights approach one and the local GWR model approaches the global OLS model. As expected, if $d_i > b$, a zero value should be produced.

Overall, GWR extends the traditional regression framework by allowing local rather than global parameters to be estimated. Regarding the choice of the weighting system $W(u_i)$, the majority of software follows an adaptative kernel decoded in Eq. 3 (Bocci *et al*., 2007). A fixed kernel may be another option where b is a fixed vicinity distance.

$$W_{ij} = \left[1 - \left(\frac{d_j}{b}\right)^2\right]^2, \text{if j is one of the Nth nearest neighbors of i. Otherwise, } W_{ij} = 0 \quad (3)$$

where, b is the distance between i and the Nth nearest neighbor (between 8 and 16 neighbors is a good start). Hence, this kernel function distance varies in space and presents an adaptive bandwidth depending on the data points' density: b is relatively small in areas where the data points are densely distributed and the bandwidth is relatively large where the data points are sparsely distributed.

Charlton and Fotheringham (2009) use the corrected Akaike Information Criterion (Eq. 4) in the GWR as a measure of goodness of fit. Two separate models being compared are held to be equivalent if the difference between the two AICc values is less than 3. The AICc value can also be used to determine the optimal value of the kernel bandwidth (the lowest, the better).

$$AIC_c = 2n\log_e(\hat{\sigma}) + n\log(2\pi) + n\left[\frac{n + tr(s)}{n - 2 - tr(s)}\right] \quad (4)$$

where, n is the number of observations in the dataset; $\hat{\sigma}$ is the estimate of the standard deviation of the residuals; and tr(S) is the trace of the hat matrix.

As stated by Lloyd (2007), the goodness-of-fit of a GWR model can also be assessed using the geographically weighted coefficient of determination: $R_i^2 = (TSS^w\text{-}RSS^w)/TSS^w$ where TSS^w denotes the geographically weighted total sum of squares:

$$TSS^w = \sum_{j=1}^{n} w_{ij}(y_j - \bar{y})^2$$

and RSS represents the geographically weighted residual sum of squares:

$$RSS^w = \sum_{j=1}^{n} w_{ij}(y_i - \hat{y}_j)^2$$

FINAL THOUGHTS

Spatial autocorrelation and statistical heterogeneity hold the ability to compare two regions and to characterize texture differences. Quite often, distant pairs are less similar (competitive spatial processes) than closer ones (cooperative spatial processes). Probably, some landscapes can exhibit extremely irregular shapes. As a consequence, indices of spatial autocorrelation calculated globally and locally are valuable for descriptive purposes because they provide a measure of how similar objects are to their spatial neighbors. This spatial dependence impact is also crucial on spatial inference interpolation such as Kriging, spatial simulation and geographical regression.

The word Kriging is synonymous with the optimal prediction of unknown values from observed data at known locations (Journel and Huijbregts, 1978; Aunon and Hernandez, 2000). After the variogram has been defined, the algebraic relationship between values at different distances is used

to estimate Kriging weights. Mostly, four factors are taken into account in assigning weights to the spatial observations: closeness to the location being estimated, redundancy between data values (clustering), anisotropy (direction) and magnitude of continuity.

Generally, the estimation of missing spatial data when undertaking GWR for discrete data, Kriging and spatial autocorrelation lead to a close linkage among them: the missing data issue (Griffith and Layne, 1999). Spatial autocorrelation can also be used with spatial prediction. For instance, a high degree of spatial autocorrelation suggests an equally likely chance of predicting neighboring values. Also, a low value reveals a low level of spatial data redundancy.

Eight relationships emerge among these concepts (Griffith and Layne, 1999):

- Spatial autocorrelation is the progenitor of Kriging and spatial regression models
- Spatial autocorrelation itself seeks description and diagnosis while spatial regression and Kriging seeks prediction
- The variance-covariance matrix is included within spatial regression and Kriging
- Once a variogram is fitted to the sample data, Kriging can be used to estimate de variable at locations where data are not sampled
- With spatial regression models, the missing data can be regarded as an interactive re-estimation solution fashioned with updated variable imputations based on R^2 in Maximum Likelihood (ML), OLS and bootstrap procedures
- Kriging is primarily concerned with more or less continuous attributes while spatial regression involves aggregations of phenomena into discrete regions such as areal units
- While autoregressive and trend surface methods assume that samples follow an underlying trend plus the random residuals, in Universal Kriging the trend component is modeled as a linear combination of functions of the spatial coordinates and Ordinary Kriging accounts for local variations of the mean
- Since, Kriging honors data at sampled locations, spatial regression residuals are not precisely similar to Kriging estimation errors

APPENDIX

Table A1: The Pb dataset (99 samples): Contamination data of Aljustrel, Portugal, within a global area of 4500 m×2950 m

X	Y	Pb	X	Y	Pb	X	Y	Pb	X	Y	Pb
650	1050	37.5	350	800	45.0	3800	1850	53.8	2450	200	72.3
4050	2200	37.5	50	550	46.5	1400	1050	79.3	1300	2250	29.8
4400	1350	55.5	2350	500	104.7	3950	2050	38.7	2650	100	64.7
650	100	38.5	1650	500	65.4	2100	150	37.6	1100	400	45.0
2900	2500	36.3	300	2100	43.5	350	0	40.1	3950	450	60.1
1200	2400	25.8	1800	700	90.0	800	1600	34.9	2100	1750	34.8
1550	2850	45.2	1550	550	78.4	1250	1550	38.4	2600	1400	99.8
2200	2500	38.2	4350	2450	52.3	250	2200	35.6	3600	750	58.4
800	2550	25.5	250	2600	29.9	400	100	34.6	3400	2250	38.4
4300	2200	43.1	950	1650	44.3	1250	2350	36.2	1500	350	99.3
1700	1350	236.0	2850	150	66.4	4250	2650	40.7	1700	450	46.8
1000	1750	25.7	4550	50	41.9	2600	1900	53.9	600	1950	30.1

Table A1: Continued

X	Y	Pb	X	Y	Pb	X	Y	Pb	X	Y	Pb
2100	350	60.6	1350	200	73.4	500	1850	38.2	2200	1650	46.6
3000	1950	29.8	1950	2300	34.4	1500	2250	39.0	3550	250	51.3
2550	2350	43.6	700	450	25.8	1600	250	91.2	650	1200	34.6
2750	1800	49.7	3800	2200	52.4	2550	100	50.3	2350	1950	43.1
2150	600	100.4	2250	600	111.8	1650	2250	27.9	750	1650	34.3
700	600	27.5	3200	2300	39.7	4100	1700	59.2	900	950	36.7
2100	100	46.3	3550	400	51.6	3300	400	78.0			
3200	1450	61.5	3150	2950	26.0	1150	1550	29.6			
3450	2700	42.3	3350	50	54.9	4050	1850	53.3			
2650	750	133.8	2200	2600	39.4	2600	800	153.1			
1750	1800	46.1	3750	1250	79.5	4500	1000	57.0			
1600	2500	27.3	1550	2000	43.2	2600	850	155.4			
2500	0	51.4	2050	1950	32.2	2750	1200	74.6			
4250	2450	38.6	1350	750	49.7	1400	2700	25.5			
300	1650	28	2350	2200	38.9	150	2450	30.8			

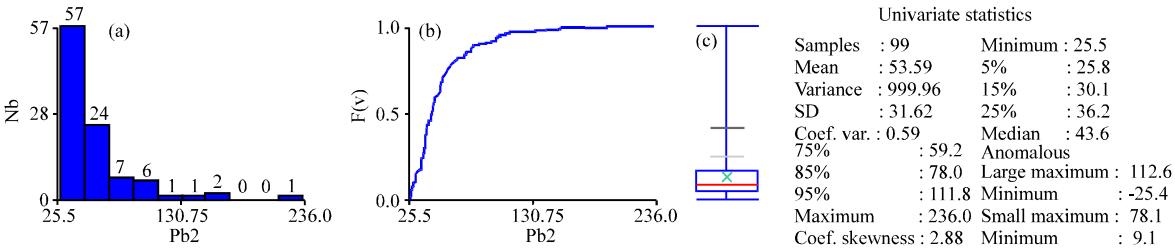


Fig. A1: Histogram, cumulative distribution functions, box-plot and descriptive statistics of the Pb dataset. Clearly, it follows a positive asymmetric distribution

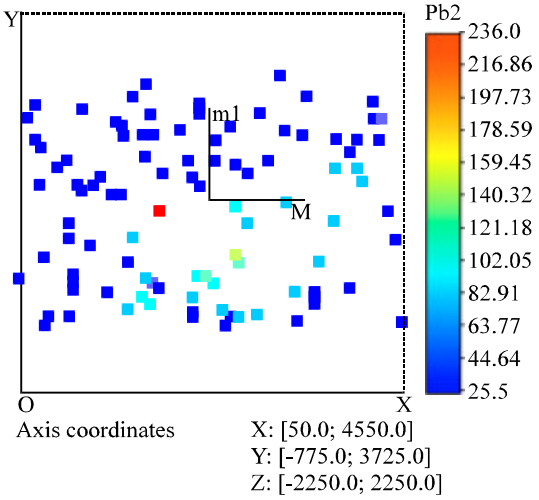


Fig. A2: Spatial distribution of the Pb contamination dataset. The attribute has a more continuous spatial pattern from East to West, thus it is anisotropic

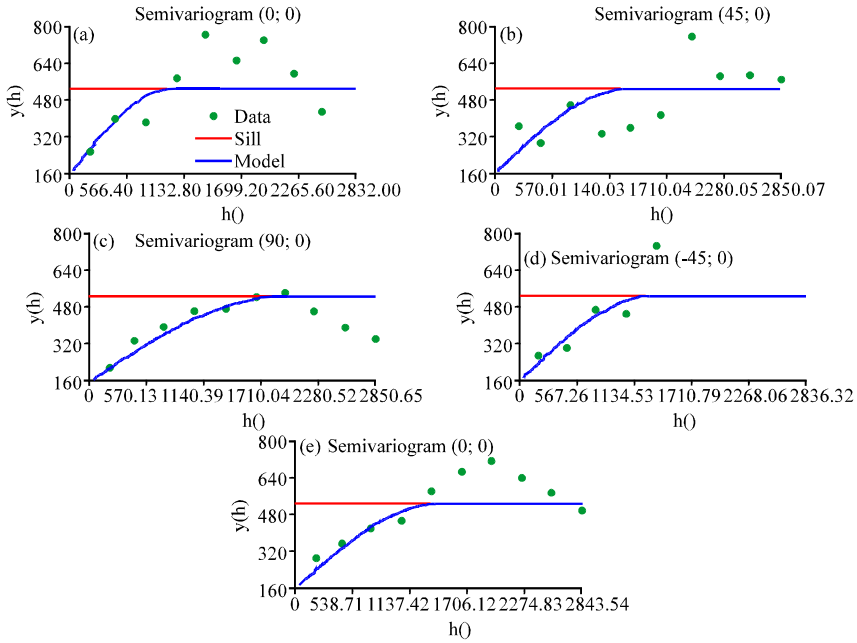


Fig. A3: Experimental variograms according to four directions plus the omnidirectional one with spherical models fitted

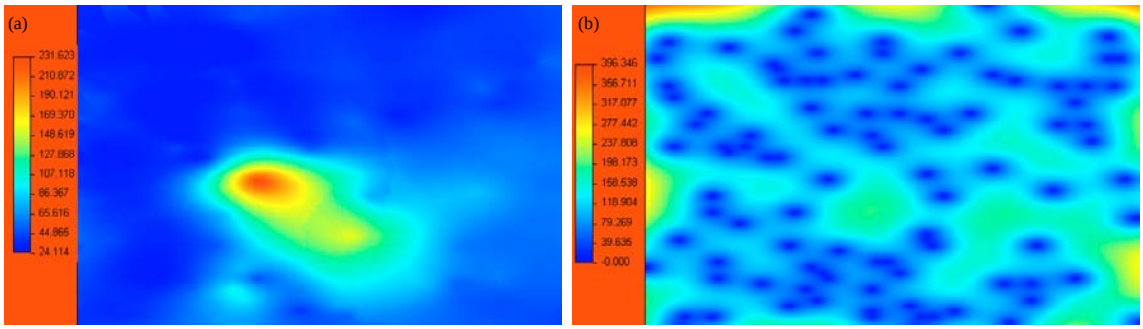


Fig. A4: Ordinary Kriging estimation (a) and Kriging variance (b) maps. As an uncertainty measure, OK variance reflects the geometry of the observations but not the variability among them

REFERENCES

Aguilar, F.J., J. Negreiros, M. Painho and M.A. Aguilar, 2008. Spatial error and interpolation uncertainty appraisal within geographic information systems. Res. J. Applied Sci., 3: 471-479.

Anselin, L., 1992. SpaceStat Tutorial-A Workbook for Using SpaceStat in the Analysis of Spatial Data. University of Illinois, Urbana, Champaign.

Anselin, L., 1994. Exploratory spatial data analysis and geographical information systems. Proceedings of the New Tools for Spatial Analysis, Nov. 18-20, Eurostat, Louxembourg, pp: 1-17.

Anselin, L., 1996. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association in Spatial Analytical Perspectives on GIS. Taylor and Francis, United Kingdom.

Anselin, L., 1998. Exploratory Spatial Data Analysis in Geocomputational Environment in Geocomputation A Primer. John Wiley and Sons, USA.

Aunon, J. and J. Hernandez, 2000. Dual kriging with local neighborhoods: Application to the representation of surfaces. Math. Geol., 32: 69-85.

Bocci, C., A. Petrucci and E. Rocco, 2007. An application of geographically weighted regression to agricultural data for small area estimates. http://www.unavarra.es/metma3/Papers/PDFS_ORAL/Petrucci.pdf.

Charlton, M. and A. Fotheringham, 2009. Geographically weighted regression. Proceedings of White Paper, National Centre for Geocomputation, March 3, Maynooth, Ireland, pp: 17-17.

Costa, A.C., J. Negreiros and A. Soares, 2008a. Identification of Inhomogeneities in Precipitation Time Series Using Stochastic Simulation. In: geoENV VI-Geostatistics for Environmental Applications, Soares, A., M.J. Pereira and R. Dimitrakopoulos (Eds.). Springer Verlag, Netherlands, pp: 275-282.

Costa, A.C., R. Durao, M.J. Pereira and A. Soares, 2008b. Using stochastic space-time models to map extreme precipitation in Southern Portugal. Nat. Hazards Earth Syst. Sci., 8: 763-773.

Costa, A.C. and A. Soares, 2009. Homogenization of climate data: Review and new perspectives using geostatistics. Mathe. Geosci., 41: 291-305.

Druck, S., M. Carvalho, G. Cβmar and A. Monteiro, 2004. Spatial Analysis of Geographical Data. Embrapa Publisher, Atibaia, Brazil, pp: 209.

Ferreira, C. and N. Simões, 1993. The Evolution of Geographical Thinking. 8th Edn., Gradiva Publications, New York.

Ferreira, C. and N. Simões, 1994. Graphics and Statistical Computation within Geography. 3rd Edn., Gradiva Publications, New York, pp: 130.

Fotheringham, A.S., C. Brunsdon and M. Charlton, 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester, UK.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. 1st Edn., Oxford University Press, London.

Griffith, D. and L. Layne, 1999. A Casebook for Spatial Statistical Data Analysis: A Compilation of Analyses of Different Thematic Data Sets. Oxford University Press, USA., pp: 506.

Journel, A.G. and C.J. Huijbregts, 1978. Mining Geostatistics. 1st Edn., Academic Press, London, pp: 600, ISBN-10: 0123910501.

Journel, A.G., 1994. Modelling Uncertainty: Some Conceptual Thoughts. In: Geostatistics for the Next Century, Dimitrakopoulos, R. (Eds.). Kluwer Academic Press, Dordrecht, The Netherlands, pp: 30-43.

Legg, R. and T. Bowe, 2009. Applying geographically weighted regression to a real estate problem. GEOconnexion International Magazine, December/January 2009.

Lloyd, C., 2007. Local Models for Spatial Analysis. CRC Press, USA., pp: 244.

Negreiros, J., M. Painho, F. Aguilar and M. Aguilar, 2010. Geographical information systems principles of ordinary kriging interpolator. J. Applied Sci., 10: 852-867.

Robertson, G.P., 2008. GS$^+$: Geostatistics for the Environmental Sciences. Addison-Wesley Professional, Plainwell, USA., pp: 179.

Soares, A., 1990. Geostatistical estimation of orebody geometry: Morphological kriging. Mathematical Geol., 22: 787-802.

Soares, A., 2000. Geoestatistica Para as Ciências da Terra e do Ambiente. IST Press, USA., pp: 206.

Soares, A., 2001. Direct sequential simulation and cosimulation. Mathe. Geol., 33: 911-926.

Soares, A., M. Joao and J. Sousa, 2010. Estimation of Extreme Values-Part II and Stochastic Simulation (Academic Powerpoints). CMRP-IST., Lisbon.

Webster, R. and M.A. Oliver, 2007. Geostatistics for Environmental Scientists. 2nd Edn., John Wiley and Sons, UK., ISBN-13: 9780470209394.