Journal of Geoscience and Environment Protection, 2019, 7, 92-113
http://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Spatial Analysis with myGeoffice.org: The Pb Jura Lake, Switzerland, Contamination Case

Joao Negreiros, Ana Neves

University of Saint Joseph, Macao, China
Email: joao.garrot@usj.edu.mo, ana.neves@usj.edu.mo

How to cite this paper: Negreiros, J., & Neves, A. (2019). Spatial Analysis with myGeoffice.org: The Pb Jura Lake, Switzerland, Contamination Case. *Journal of Geoscience and Environment Protection, 7,* 92-113.
https://doi.org/10.4236/gep.2019.72007

Received: January 3, 2019
Accepted: February 22, 2019
Published: February 25, 2019

Abstract

Protecting groundwater from lead contamination is an important public-health concern and a major national environmental issue worldwide. This article addresses myGeoffice Web Internet service for geographers, in general, and geo-statistics researchers, in particular, with the famous water contamination case at Jura lake, Switzerland (a typical rural-urban region). Based on 189 samples of lead (Pb), five key investigation steps for a scientific perspective of any pollution incident are presented: Descriptive analysis (including nearest neighborhood, G(h) and Kernel techniques), spatial autocorrelation (Moran location scatterplot and Moran I) and Ordinary Kriging (OK) interpolation. The uncertainty and cost assessments issues are exceptionally tackled with Indicator Kriging (including the conditional cumulative distribution function, Shannon local entropy, probabilistic intervals and E-type estimation) and Gaussian geo-simulation. The total lead pollution exhibited patterns of high and low levels of concentrations along the all lake, leading to the conclusion that water is unsuitable for human consumption, in general, and unsuitable for any living organism, in particular sub-areas. It is also hoped that future GIS readers will follow this approach for their spatial cases with myGeoffice.

Keywords

Geographical Information Systems, myGeoffice©, Spatial Autocorrelation, Kriging, Geo-Simulation, Costs Analysis

1. Introduction

The problem of statistical spatial analysis covers an escalating range of methods that address different spatial problems, from pattern recognition to spatial interpolation and economic trend modeling. In spite of each particular aspect, they

share a common factor: spatial statistics—geographically correlated raw data analyzed by statistical methods. Unfortunately, major software statistical applications are spatial since they ignore the special characteristics of spatial data such as spatial dependence (autocorrelation) and spatial heterogeneity (association). It is analysis of spatial data, not spatial analysis of data (Negreiros, 2015).

Therefore, spatial analysis should be centered on the search for patterns within spatial phenomena involving understanding, prediction and simulation, where location, time and geometry are significant. The differences between Geography and other sciences are locations, spatial structures and spatial processes, because no other study field concerns itself with the distribution of spatial phenomena (Negreiros, Aguilar, & Aguilar, 2012).

Preservation of lakes for their beauty and recreational benefits, as well as for their value as habitats for fish and wildlife, is of huge importance. The aim of lake monitoring is to identify small changes in conditions so that remedial work can start before a lake has degraded extensively and expensive restoration measures are necessary. This happens because clean water is important for our health, clean water supports our unique ecosystems, clean water helps us make a living and clean water keeps our power running. The effects of pollution of water ranges from diseases and food chain effects to destruction of ecosystems and eutrophication (decrease of the amount of oxygen in the water body, severely affecting the aquatic life there). The Jura Lake is no exception to all these concerns.

This writing tries to review a set of methods and procedures that are truly spatial (and special) such as Moran scatterplot, Indicator Kriging, stochastic simulation for spatial processes, error and uncertainty measurement. In this particular case, the spatial pollution case of lead (Pb) at Jura Lake, Switzerland, will take place. As expected, these fields hold a common process for any space study: Collection of data points (Section 2); Descriptive analysis (Section 3); Modeling of spatial variability for description of spatial patterns (Section 4); Spatial prediction at non-sampled locations (Section 5); Modeling of uncertainty such as what is the probability to exceed a critical concentration at any non-sampled location or which sub-areas should be considered for cleaning (Section 6); Geo-simulation (Section 7); Conclusions (Section 8).

2. The Lead Dataset of Jura Lake, Switzerland

The Jura Mountains are a sub-alpine mountain range located north of the Western Alps, mainly following the course of the France–Switzerland border. The Jura separates the Rhine and Rhône basins, forming part of the watershed of each.

The Jura contamination case set is a worldwide famous heavy metals dataset provided by J. P. Dubois of IATE-Pédologie, EcolePolytechniqueFèdérale de Lausanne and printed as appendix C in Pierre Goovaerts book, Geostatistics for Natural Resources Evaluation, Oxford University Press, of 1997. It comprises the spatial coordinates and values of categorical and continuous attributes at the 359 sampled sites:

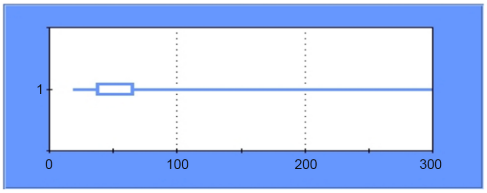
- Rock Types: Argovian, Kimmeridgian, Sequanian, Portlandian and Quaternary.
- Land uses: Forest, Pasture, Meadow and Tillage.
- Trace metals in the soil: Cd (Cadmium), Cu (copper), Pb (mg lead per Kg in topsoil), Co (cobalt), Cr (chromium), Ni (nickel) and Zn (zinc).
- Xloc, Yloc: Local grid in Km.

For the present analysis and myGeoffice© illustration capabilities purposes, only the lead contamination variable (subset of 189 samples) will take place. The proposed methodology was designed in order to suit various requirements of investigating water quality and based on the levels of Pb heavy metal pollutant. For that, myGeoffice Web software will be used intensively to help interpolation mapping, descriptive analysis, geo-simulation of different scenarios and cost study. Again, it is hoped that the present reader is able to fully understand this Web spatial analysis software for his/her future geographical studies.

3. Descriptive Analysis

Classical statistics constitute one method in the whole field of spatial analysis. Basically, it is just another analysis extension of spatial data whose role should be centered on the description phase, on the score analysis variability over space and on the check in the phenomena over time (Negreiros, Aguilar, & Aguilar, 2012).

According to Figure 1, the 189 available samples show a highly left skewness



DESCRIPTIVE STATISTICS
Sample Size: 189
Conventional Average or Mean: 59,7576719576719
95% Confidence Interval of the Mean: [54,3015360871015 ; 65,2138078282423]
Conventional Standard Error of the Average: 2,78374279110735
Conventional Variance: 1464,60332221096
Conventional Standard Deviation: 38,2701361666112
Maximum: 300
Median: 48
Minimum: 18,68
Mode: 48,8
Coeficient of Variation: 0,640422140168363
Absolute Average of Deviation: 24,448363707623
First Quartil: 37,88
Second Quartil: 48
Third Quartil: 65,2
Interquartil Range: 27,32
Skewness Index: 3,02727675002521
Pearson Index: 0,286324352491642
Kurtosis Index: 12,1723585911671

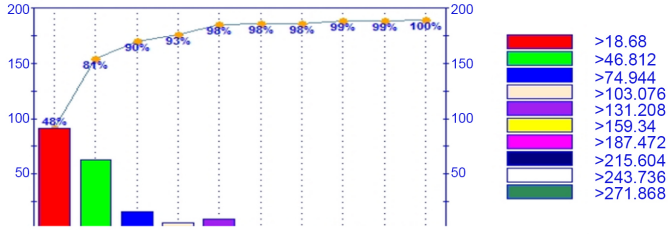


Figure 1. Descriptive analysis of myGeoffice© of the Pb pollution dataset.

(3.02) and a high leptokurtic kurtosis (12.17) distribution whose range values vary between 18.68 mg and 300 mg. Positively, it is difficult for this dataset to achieve interpolation. Yet, the spatial distribution of the sampling is a plus on this study whose R dispersion index equals 1.27, a fairly spatial dispersion of the samples (Figure 2 and Figure 3). As a result, the mean nearest distance among samples is 0.13 Km, not far apart from 0.10 Km of a completely random dispersed situation. Therefore, it can be concluded that any declustering methods, such as polygonal, nearest neighbor or cell, will not help our sampling representation in space.

The main goal of point pattern analysis is to determine if there are points that may reveal a clustered (concentration of points) and uniform (every point is as

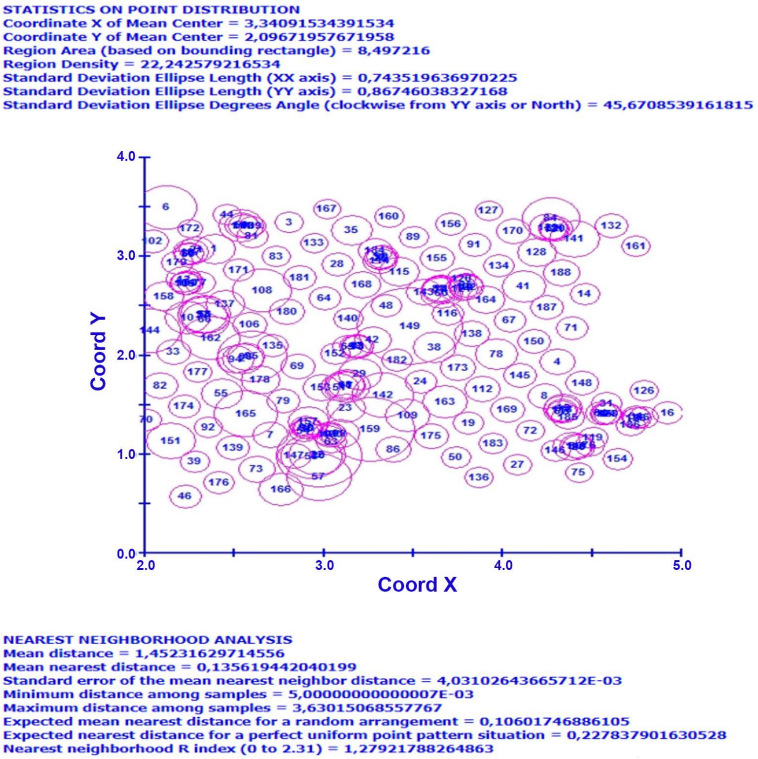


Figure 2. Nearest neighborhood analysis of myGeoffice©.

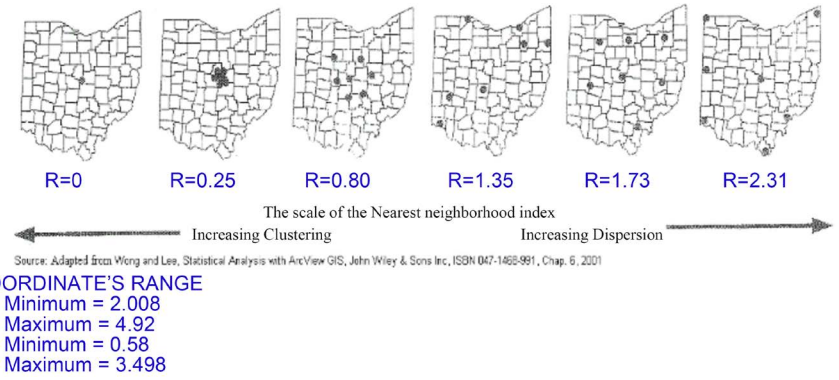


Figure 3. Coordinates range of the lead pollution dataset.

far from all of its neighbors as possible) pattern over an area, as opposed to being randomly (the location of any point is not affected by the position of any other point) distributed (Negreiros, 2011).

The G(h) cumulative distribution, based on nearest neighborhood analysis, was used to check this evidence of events (samples) interaction. Mathematically, the G(h) function equals the ratio between the number of samples for a particular h distance and the total samples number. As expected, there is a close relationship between the considered h distance (for myGeoffice©, the one-hundredth successive h increment equals 10% of the mean nearest distance) and the ratio numerator. By default, G(h) curves with faster growth for closer distances suggest interaction between samples (clustered scale presence). On the other hand, small values for closer distances and faster growth for greater ones imply a more regular or even spatial distribution (our case and according to **Figure 4** and **Figure 5**).

4. Spatial Autocorrelation

The role of location (the absolute coordinates and the relative typology) holds two major implications for the way statistical analysis should be carried out. Location leads to spatial dependence (correlation or variation that each neighbor holds in relation to a particular point) and spatial heterogeneity (clustering, concentration or proportion of neighborhood average in relation to a specific point) established by Tobler's First Law of Geography ("all things are related, but nearby things are more related than distant things") although the expression

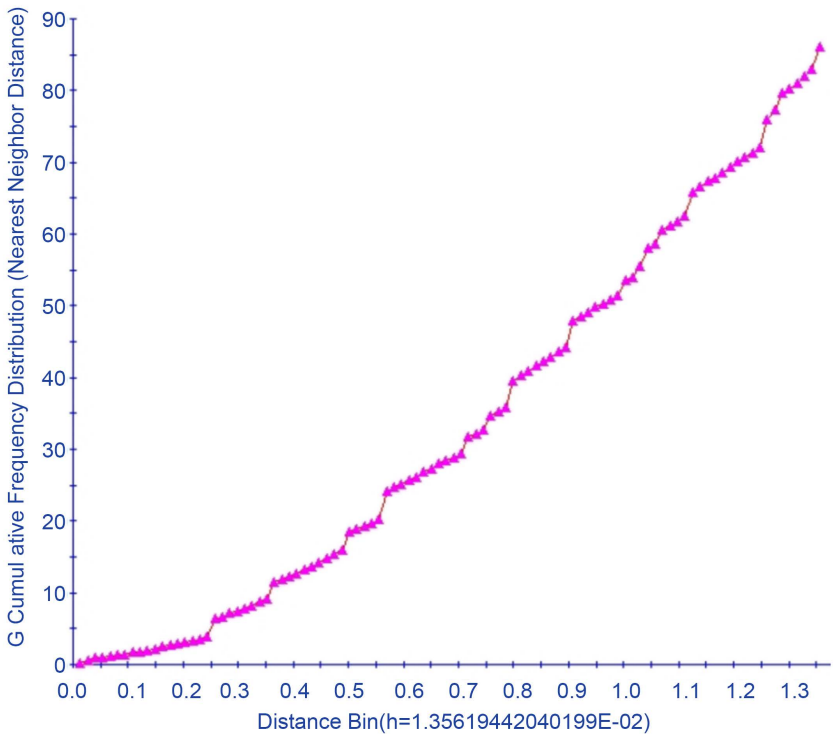


Figure 4. G cummulative function of myGeoffice©.

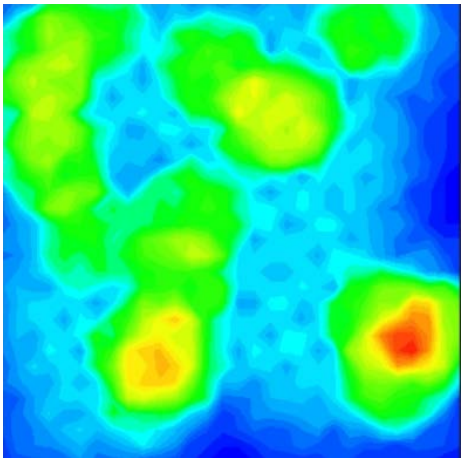


Figure 5. As a reference, the Kernel function generates a raster density map from point data using a moving Kernel function (under myGeoffice©, the Gaussian is used) to produce a continuous tapered surface and to identify hot spots. In this particular case, the h estimation radius (the locally-adaptive window approach) was 2.5 of the nearest average distance among samples (0.13 Km), although other options could be setup. The aim is to visually determine where lack and high concentration of samples can be found.

of Tobler's Law is not completely accurate (probably it should be called *Tobler's empirical postulation*) since a law, by definition, is an invariable relationship among variables, a geographical situation that may break down by Mother Nature. As regional differentiation respects the intrinsic uniqueness of each location, spatial autocorrelation can be viewed, hence, as a map pattern descriptor (Griffith, Morris, & Thakar, 2016).

One of the major steps in geo-statistics is to infer the real variogram in order to develop the Kriging equations. Mostly, the variogram is the inverse function of the spatial covariance and quantifies spatial autocorrelation along dissimilar distances. Typically, the most common models are the Spherical, Exponential and Gaussian models (**Figures 6-8**). These are sometimes called bounded models because the variogram increase with distance until they reach a maximum, named sill, at an approximate distance known as the range. The sill is the maximum variance, and represents variability in the absence of spatial dependence (Margaret & Webster, 2015). The range is the distance h at which the spatial correlation vanishes, i.e., observations separated by a distance larger than the range are spatially independent observations.

In theory, the variogram value at the origin (zero lag) should be zero. If it is significantly different from zero for lags very close to zero, then this variogram value is referred to as the nugget. Accordingly, the nugget is the variance as the separation approaches zero. It represents variability at a point that can't be explained by any spatial structure.

The Moran I is evaluated by measuring the covariance between attributes at each place and near sites towards the overall mean. If both neighboring values are above or below the average, the outcome becomes positive reflecting the presence of a similar pattern. Otherwise, the negative outcome of the two mean

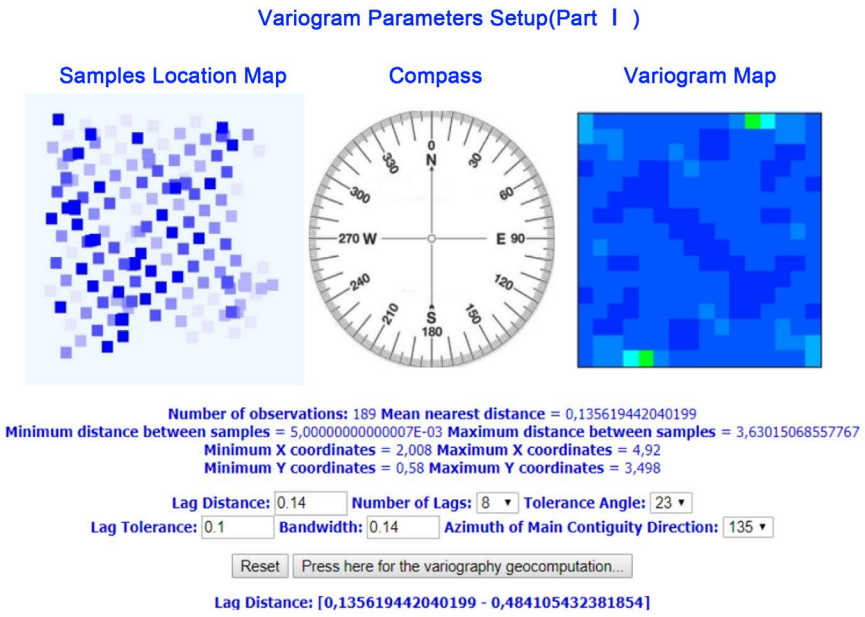


Figure 6. myGeoffice© step one regarding the variogram setup.

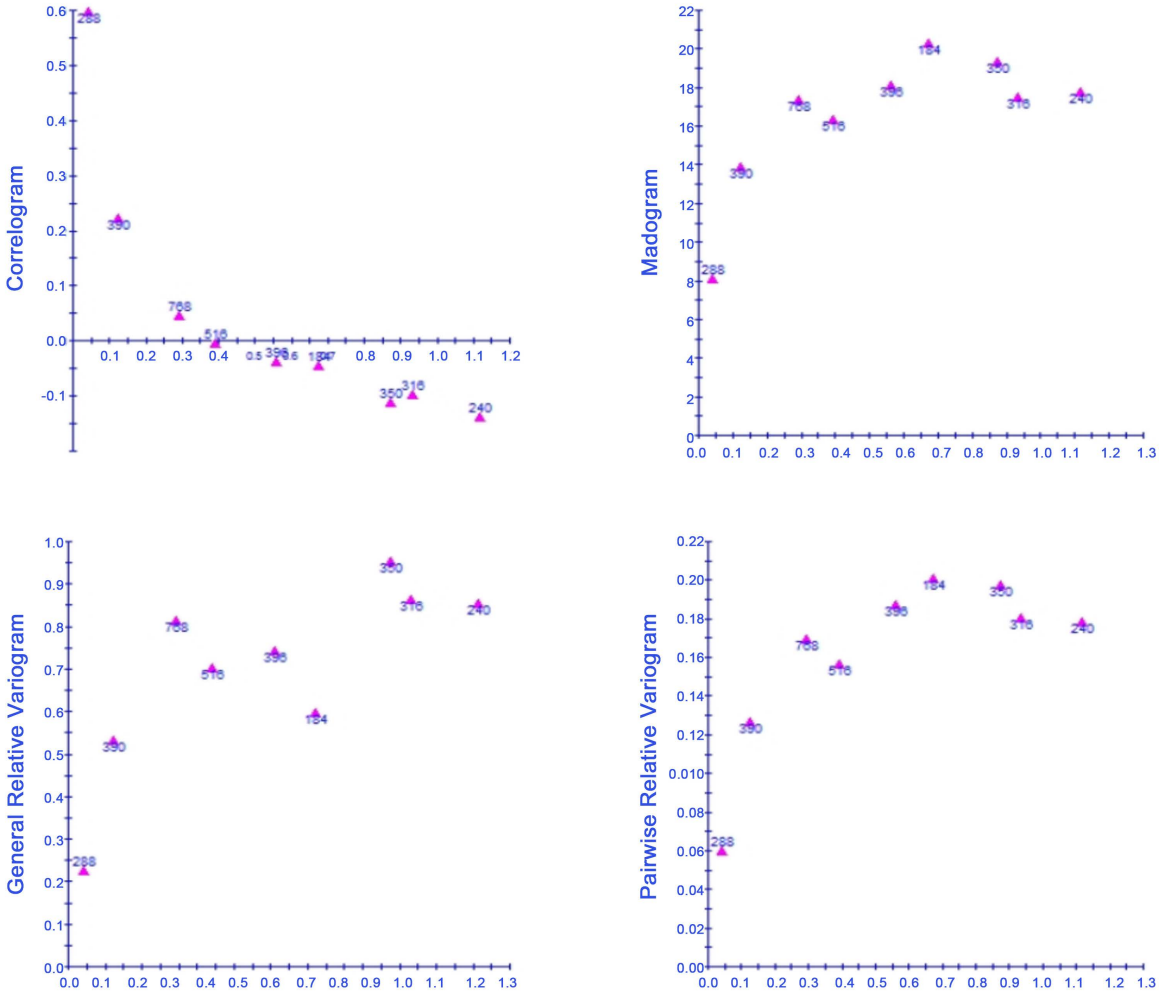


Figure 7. myGeoffice© step two regarding the variogram setup.

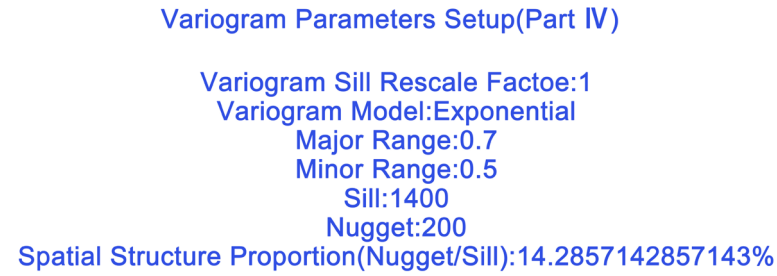


Figure 8. myGeoffice© step three regarding the variogram setup.

deviations reveals a neighborhood presence of high-low or low-high values. Theoretically, the expected random value of the Moran I is $-1/(n-1)$, where n equals the number of observations (Mera, Condal, Rios, & Silva, 2016). This means that the zero cutoff cannot be used as a reference point to distinguish positive and negative spatial autocorrelation (particular in the presence of a small dataset).

Under myGeoffice©, the vicinity weights of the Moran I index isassessed by Variogram Fitness option. The Moran I is calculated for five equal incremental distances (a scale issue) whose largest distance equals the chosen variogram maximum range. Therefore, it is expected that the Moran I for this range distance should be close to zero (a confirmation issue) as **Figure 9** illustrates.

The Moran location scatterplot allows to visualize and identify the degree of spatial instability based on the bivariate regression coefficient of the spatial lagged variable, a weighted average of the neighboring values, against the original variable. The four quadrants, centered on the global mean, are composed of the x-axis, deviations from the original variable mean, and the y-axis, neighborhood weight average mean deviations. As expected, this scatterplot is divided into four association types (**Figure 10**): upper right quadrant (high values above the mean surrounded by high values), lower left (high values surrounded by low values), upper left (low values surrounded by high values) and lower right (high values surrounded by low ones). Yet, this option holds a small issue concerning its layout. When an over crowed sampling happens, it is not possible to clearly see the samples ID because of overlap reasons as **Figure 10** testifies.

5. Ordinary Kriging (OK): Interpolation

Kriging is an exact interpolator in the sense that the sample and estimation are equal. This happens because the variogram value between sample k and estimation x_0, $\gamma(x_k, x_0)$, equals zero since the estimated site is the sample itself. By resolving the Kriging system (row k of matrix B changes to zero if the nugget-effect is zero), the weight w_k is one while all other weights and the LaGrangean dummy parameter become zero. Its Kriging variance also equals zero.

myGeoffice© presents three models to handle the nugget-effect component, the most mysterious factor of the variogram itself (Cardillo, Bromham, & Simon, 2016):

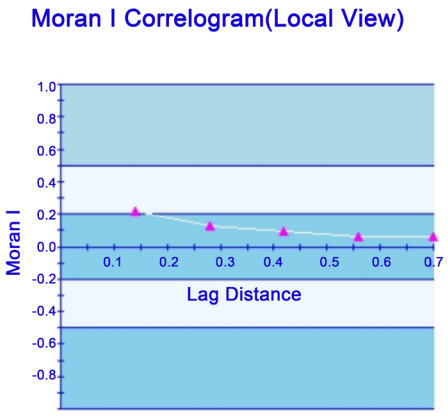


Figure 9. As expected, the major range of the variogram (0.7 Km) leads to a Moran I closer to zero for the same distance.

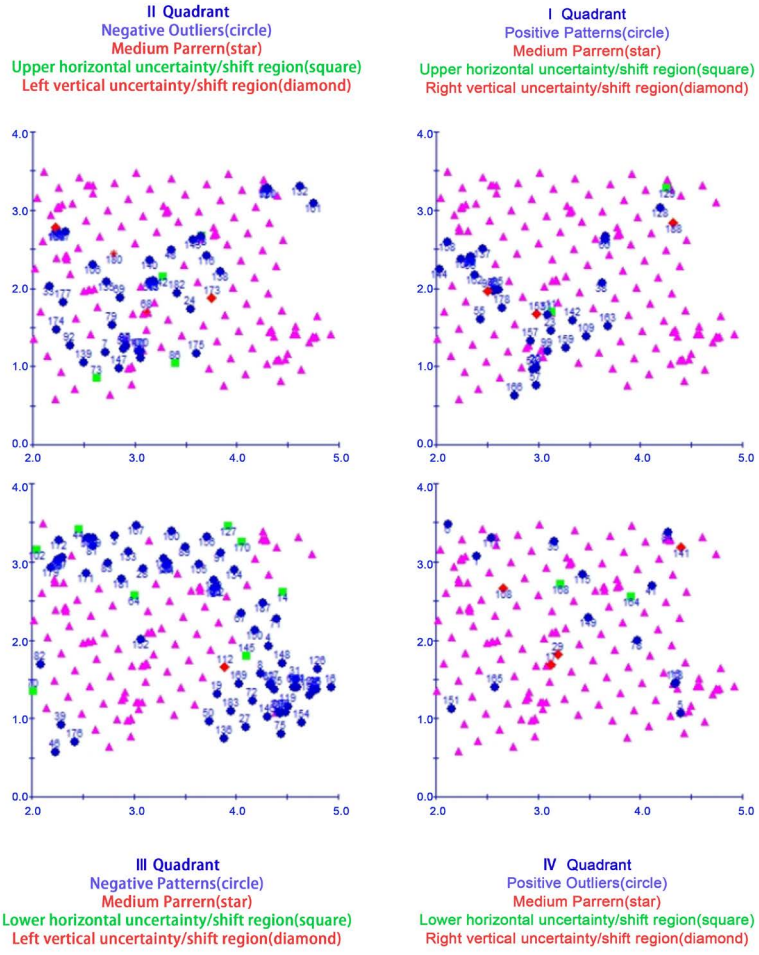


Figure 10. Globally, it seems that the bottom left region is characterized by high concentrations of Pb (I quadrant) while the bottom right and upper regions (III quadrant) are portrayed by low concentration ones. Negative outliers (II quadrant) are located in a strong pattern, a layout not followed by positive outliers.

- With Model 1 (**Figure 11**-left), the observations are considered precise and accurate, although sudden jumps at the variogram origin may emerge. This means forcing the nugget-effect to zero at zero lag distance: Variogram (0) = 0; Variogram (h) = Spherical/Exponential/Gaussian model, if 0 < h ≤ range; Variogram (h) = sill = C_0(nugget-effect) + C_1(partial sill), if h > range.
- A second possibility is to consider Variogram (0) = 0 in a continuous mode (Model 2).
- With the micro-scale components of Model 3, the nugget-effect is divided into two nested factors: Variogram (0) = 0 and Variogram1 (0 ≤ h ≤ shortest sampling interval) = Spherical/Exponential/Gaussian first model. For lag distances between the shortest sampling interval and the actual range, Variogram2 (h) = Spherical/Exponential/Gaussian second model. The first micro-scale range equals the shortest sampling interval (SSI) lag, its nugget-effect is zero while its total sill matches the extrapolated value of the second variogram structure (in other words, the given value of the second variogram structure at SSI lag distance becomes the first variogram sill parameter). For (Negreiros, 2017), this micro-vision reflects different processes of variability at different scales: Z(s) = NIU(s) + W(s) + NANO(s) + ERROR(s), where NIU(s) denotes the large-scale deterministic variation, W(s) is the smooth small-scale variation (the intrinsically stationary process whose variogram range exists and is larger than min{||si-sj||}), NANO(s) equals the micro-scale variation whose variogram range exists and is smaller than min{||si-sj||}) whilst ERROR(s) represents a zero-mean white-noise process, independent of W and NIU.

The OK (**Figure 11**-right) uncertainty is closely related to its variance in the following way:

$$\sigma_e^2 = C_{00} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C_{ij} - 2\sum_{i=1}^{n} w_i C_{i0} = -\gamma_{00} - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma_{ij} + 2\sum_{i=1}^{n} w_i \gamma_{i0},$$

where C_{00} is the variance of the estimated point value, C_{ij} is the covariance between the i^{th} and j^{th} sample, w_i and w_j are the OK weights, C_{i0} represents the covariance between the i^{th} sample and the unknown value being estimated. If errors respect the Gaussian curve then real values will fall within the Kriging_predictor $\pm 2\sigma_{\text{OK}}^2$ interval for a 95% confidence level (as expected, this implies symmetry of the local distribution of errors).

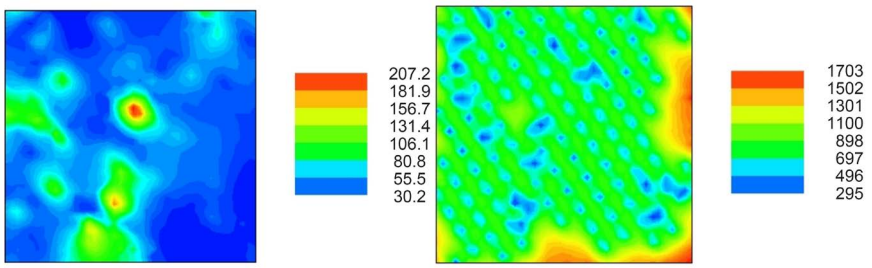


Figure 11. OK with nugget-effect; OK variance.

However, uncertainty is not included with variogram estimation and, quite often, prediction variance is underestimated. But even more critically, OK variance is not sensitive to local error for three major reasons: 1) It is based on the same global variogram; 2) Distances among locations are the only relevant factor, that is, two identical data configuration own? The same Kriging variance, whatever the data values; 3) It decreases as spatial continuity increases. OK variance is mainly a geometry-dependent measure, heading the assumption that an OK true error map is a better substitute. OK variance is too much of a spatial operation.

As expected, OK with no nugget-effect (**Figure 12**-left) produces the least smoothing effect of the three linear spatial estimations, an outcome not followed by Model 3 (**Figure 12**-right). Yet, it is noteworthy that both these interpolation versions are the ones that achieve the best results when a cross-validation procedure takes place (**Figure 13**-left). At last, if measurement error of the sampling is added to each sample (**Figure 13**-right) and, unsurprisingly, OK becomes a non-exact interpolator in the sense that the interpolation and the sample value are not equal (Negreiros, 2017).

6. Indicator Kriging

IK is a well-developed geo-statistical model for the probabilistic mapping of local conditional probability distribution function (cpdf) since it allows different types of information to be processed together, regardless of their origins. The objective is to evaluate the Z variable, at any location x, the conditional cumulative distribution function (ccdf) value or posterior probability, that is, $F(x; Z^*|(n)) = \text{Prob}\{Z(x) \leq Z^*|(n)\}$, where the conditioning information consist of n data measurements (Cressie, 1993).

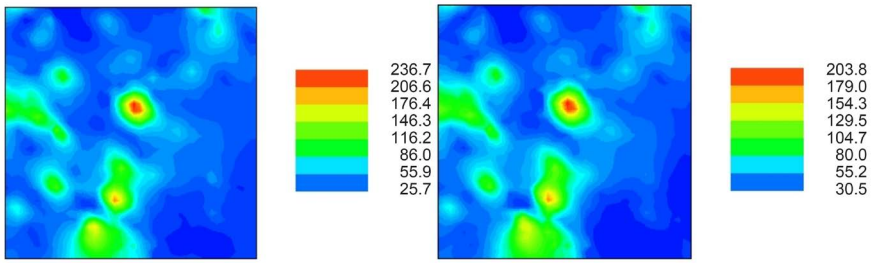
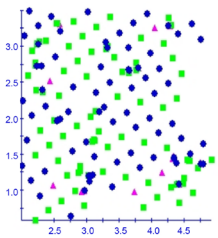


Figure 12. OK with no nugget-effect; OK with nested-structures for the nugget-effect.



Sample ID	X Coordinate	Y Coordinate	Sample Value	Estimation with 9% of measurement error
172	2.251	3.288	18.68	19.412362
133	2.942	3.137	25.88	26.081336
189	2.593	3.312	26.2	26.801349
167	3.017	3.482	26.32	26.33839
75	4.426	0.821	27.2	27.140148
127	3.919	3.466	28.6	28.708401
44	2.461	3.423	29.36	29.65826
123	4.59	1.414	30.76	31.011458
3	2.807	3.347	30.8	30.698619
136	3.87	0.762	30.84	30.670599

Figure 13. OK cross-validation; OK with 9% of measurement error.

The essence of the indicator approach relies on the binomial data coding into either 1s or 0s, depending upon its relationship to a particular cutoff value, z_k. For a given observation of value $z(x)$, the indicator variable is decoded as Equation (1) states.

$$I(x) = \begin{cases} 1 \text{ if } z(x) > z_K \\ 0 \text{ if } z(x) \le z_K \end{cases} \quad (1)$$

This is a non-linear transformation of the input data into either a binary form. Values, which are much greater than a given cutoff, z_k, will receive the same indicator value as those values which are only slightly greater than that cutoff. Thus, indicator transformation of data is an effective way of limiting the very high values effect (positive outliers). Based on a set of indicator-transformed values, IK will provide a resultant value between 0 and 1 for each point estimate. Basically, this is an estimate of the proportion (probabilities) of the values in the neighborhood which are greater than z_k threshold.

6.1. IK Uncertainty Interpolation

For a cutoff of 48 mg/kg (the median value), a Spherical model with a major range of 1 Km, a minor range of 0.7 Km, a sill of 0.26 and a nugget-effect of 0.12 was setup. For a cutoff of 70 mg/kg (the safety value for non-play areas), an Exponential model with a major range of 0.5 Km, a minor range of 0.3 Km, a sill of 0.18 and a nugget-effect of 0.06 was setup. For reference and under myGeoffice©, if sample values are below the chosen threshold, these are decoded as one (pink triangle). Otherwise, they become zero (green square).

The resulting interpolation map of **Figure 14** shows the probability of falling below (2D and 3D layout mapping) or exceeding (**Figure 15**) the chosen 48 mg/kg threshold by transforming the continuous variable into a (binary) indicator one (**Figure 16** and **Figure 17** shows this same logic but based on a threshold of 70 mg/kg). For computation purposes, it assumes an unknown mean while the phenomenon is considered continuous in space. Statistically, IK provides a least square estimate of the conditional cumulative distribution function (CCDF) at any cutoff. Thus, the ccdf for any location can be built by assembling the IK system for several cutoffs (Cressie, 1993).

Regarding the drawbacks of this non-parametric method, five issues should be stressed: 1) Loss of information because it does not distinguish among observations if they are both below or above the threshold; 2) Setup of as many variograms as the levels to be considered; 3) Possibility of obtaining estimates greater than 1 and below 0; 4) Regarding extreme values, the variogram may correspond to a pure nugget-effect; 5) Concerning a series of several cutoffs, order relation correction should be taken into account with a posteriori correction procedure of the conditional cumulative distribution function (a non-decreasing one).

6.2. CCDF Model

IK provides the conditional cumulative distribution function (ccdf) at different

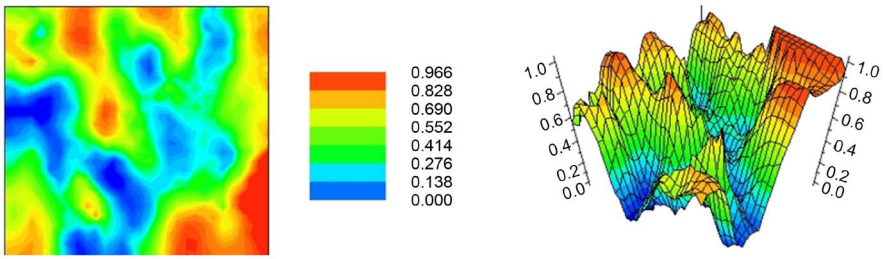


Figure 14. IK of myGeoffice©.

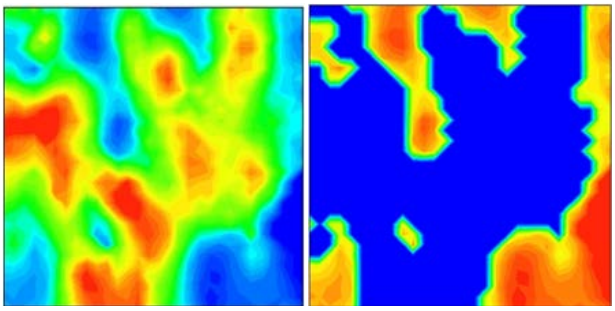


Figure 15. The probability of any spatial location between 67% and 100% of being upper (left) and lower (right) than 48 mg/kg; that is, the safest areas of the Jura lake to take a bath, for instance, are highlighted by the non-blue areas.

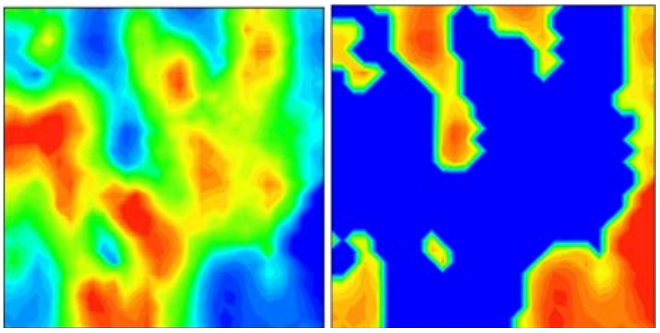


Figure 16. Analogous to **Figure 14**, both maps show the probabilities of being below (left) and above (right) 70 mg/kg.

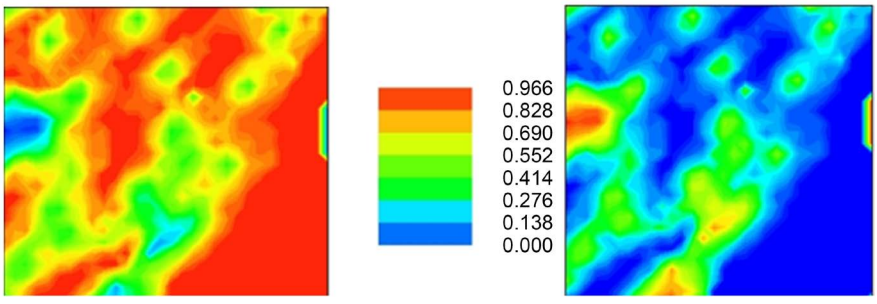


Figure 17. Analogous to **Figure 15**, both maps show the probability of any spatial location between 67% and 100% of being higher (left) and lower (right) than 70 mg/kg.

known location to assess uncertainty about an unknown X location. The aim of this option is to plot the probability model for different observations, using the

median IK (MIK). Henceforth, sample 108 (2.656, 2.656, 138.56) was simply randomly chosen and the below graph was generated for illustration purposes.

For reference, the cumulative distribution function F(x) is defined as the probability that all values in a random vector X are less than or equal to the corresponding values of vector x, that is, Prob($X \leq x$). Hence, this function (**Figure 18**) is monotone increasing, right-continuous while its range varies between zero and one (Negreiros, Aguilar, & Aguilar, 2012). One of the interesting features of IK is that ccdf can take very different shapes from place to place (working as a local uncertainty measure).

6.3. Shannon Local Entropy

IK allows the possibility of estimating the ccdf for any particular spatial X location. By defining L cutoffs (nine, under myGeoffice©, which means ten non-overlapping classes), it is then possible to infer about the probability uncertainty for each class (differences between the ccdf of any adjacent cutoffs).

Using this information, the Shannon entropy allows the inference about the interpolation local uncertainty of unknown sites. Given $L = 9$ sets of IK probabilities for any X location ($P_i(X)$, where $i = 1 \ldots 9$), the Shannon entropy (a disorder measure connected to the spatial organization of an attribute) equals -sum($P_i(X) \times \ln(P_i(X))$), where ln() denotes the Neperian logarithm, $P_i()$ represents the IK probability estimation for each class while -sum() is the negative sum of all probability classes.

Red color signifies high lack of estimation confidence (low trust) which means that the researcher should consider these sub-areas as the future candidates for further sampling measurements. Dark blue means low instability (high certainty) regarding spatial estimation (the smaller the entropy at (X, Y) location, the greater certainty associated with that location).

Take notice that under myGeoffice©, this Shannon entropy is not standardized (**Figure 19**). At last, median IK (the indicator variogram of the median cutoff value) is used for the nine thresholds because of geo-computation simplicity (Negreiros, 2017).

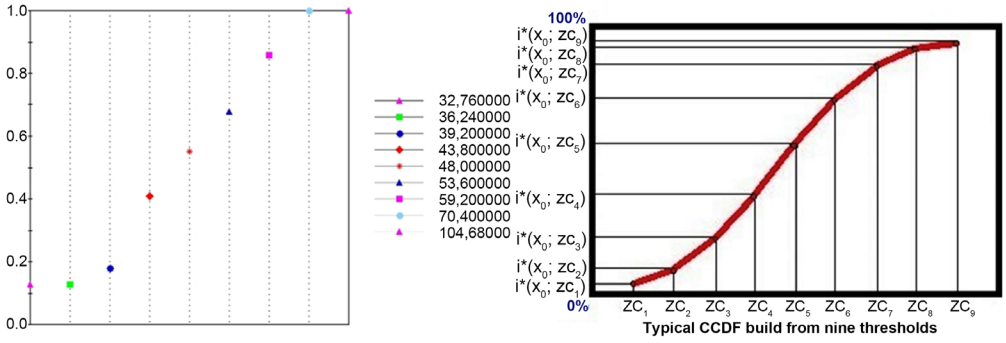


Figure 18. For instance, the estimation probability of this variation sample at this location varies between 39.2 and 70.4 mg/kg and equals the probability difference of both cutoffs, that is, 82.15%. As well, the probability of being higher than 104.68 mg/kg is 0%.

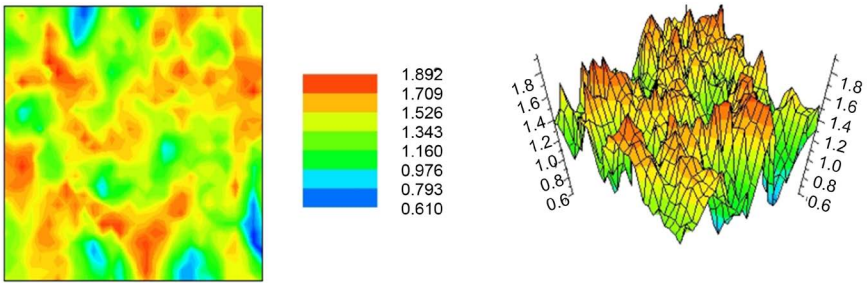


Figure 19. Recall that OK variance is data-independent while this local uncertainty measure is data dependent.

6.4. Probabilistic Intervals

Similarly to the Shannon local entropy, Figure 20 maps can be interpreted as a local uncertainty measure. Concretely, if the difference between two generic IK thresholds, t1 and t2, is small, for instance, then the spatial distribution tails become significant. Otherwise, if IK(t2)-IK(t1) is considerable larger, then all major spatial interpolations will fall between those two cutoffs. That being said, a red pattern signifies that there is a high probability for that particular area to fall between the chosen [t1-t2] interpolation range. On the contrary, blue color denotes a negligible probability of that region falling between the selected interval (Cressie, 1993).

6.5. E-Type Estimation

E-type is considered to be an optimal estimator under the least square criterion whose estimates depend on data values. According to (Goovaerts, 1997), the advantage of the E-type estimate lies in the availability of a model of uncertainty much richer than the mere OK variance. Analogous to ccdfmodel and entropy of Shannon, the MIK parameters (the indicator variogram of the median cutoff) will be used by myGeoffice© to calculate these maps (Figure 21).

6.6. Uncertainty and Cost Analysis

Many investigations lead to important decisions being made such as cleaning hazardous areas or correcting soil deficiencies. There is a risk of declaring contaminated a safe location (false positive or Type I Alpha risk) for a specific cutoff. Conversely, one might declare safe a contaminated location (false negative or Type II Beta risk). These two misclassification risks can be assessed from the ccdf model, according to the following math guidelines (Goovaerts, 1997):

- Risk Alpha = Prob$[z(x_0) \leq \text{Cutoff} | z^*(x_0) > \text{Cutoff}]$
- Risk Beta = Prob$[z(x_0) > \text{Cutoff} | z^*(x_0) \leq \text{Cutoff}]$.

Henceforth, for any particular location x_0, the magnitude of this misclassification risk depends on the ccdf model, not on the particular estimate retained. Bear in mind that this type of misclassification risk depends on the E-type estimation assessed previously.

According to the above Alpha and Beta statistical risk definitions, some inferences can be drawn from Figure 22:

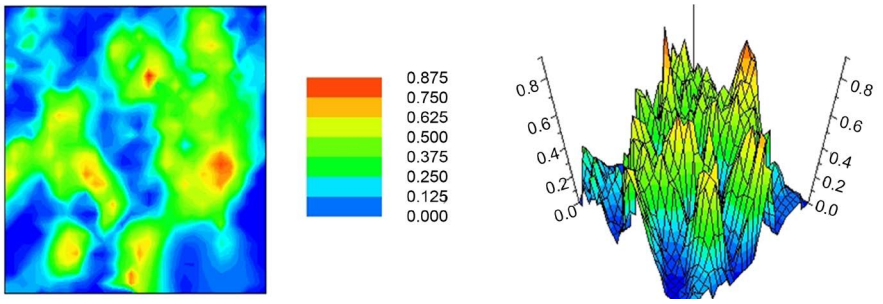


Figure 20. In this particular case, the t1 and t2 thresholds range equal to 48 mg/kg and 70 mg/kg, respectively (basically, this mapdisplays the difference between **Figure 16**-left and **Figure 14**-left).

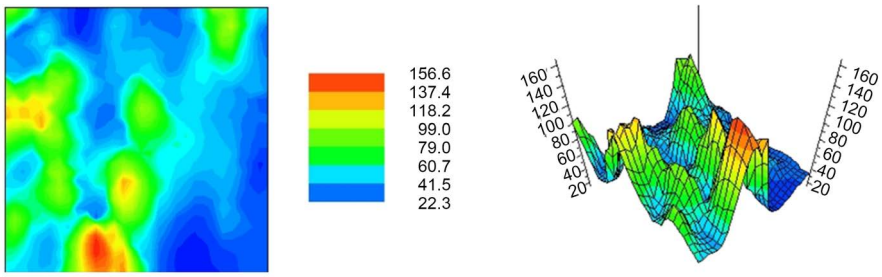


Figure 21. Unfortunately, this spatial interpolation suffers from highly smooth effect, an outcome not desirable for any researcher, particularly when the investigator is searching for the high and low concentrations of contamination.

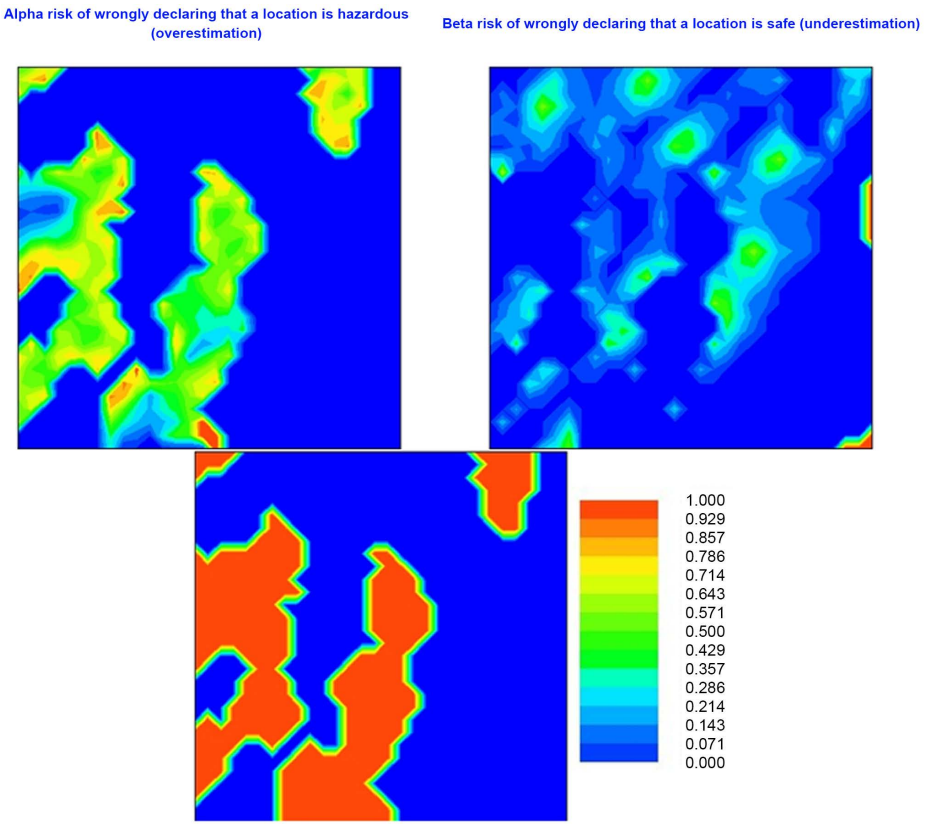


Figure 22. Alpha and Beta risks of myGeoffice©.

- Red areas of the Alpha risk map show contaminated regions (the E-type estimation is greater than the given 70 mg/kg cutoff) that hold a major probability of being incorrect (Alpha risk). Yellow and green spatial patterns present some doubts regarding our hazardous classification (30% - 70%). As expected, these locations are larger between high and low transition sub-areas (close to the borders of the cutoff regions). On the contrary, the dark blue color may hold two notations: 1) The E-type estimation is smaller or equal than the chosen cutoff; 2) The associated probability of being higher than 70 mg/kg is close to zero.
- Regarding the Beta risk map, this same dark blue color indicates the sub-regions whose E-type estimation is greater than the chosen cutoff or the associated probability of being lower than 70 mg/kg is close to zero. Yet, a red pattern indicates that those sites were classified as clean (the E-type estimation is smaller the given cutoff) but with a major certainty (80% - 100%) of being considered hazardous instead (Beta risk). Yellow and green colors represent clean sections with some probability of wrongly declaring that they are safe (curiously, the Beta risk is not defined where the Alpha risk is and vice-versa).

Another uncertainty approach concerns the economic impact evaluation of two possible decisions by using the concept of an economical loss function. Consider, for example, the spatial estimation of a toxic concentration. Underestimation of that concentration (negative estimation error) may cause ill health, leading to insurance claims and lawsuits. Conversely, overestimation of toxic concentration (positive estimation error) may cause costly and unnecessary cleaning (Goovaerts, 1997).

Given any damage (financial) cost function, any x_0 site estimate should be chosen so as to minimize this resulting loss. Since the real $z(x_0)$ is quite often unknown, the idea is to use the conditional cumulative distribution function model (E-Type estimate) to determine this expected loss based on two types of misclassification (**Figure 23**): 1) Cleaning places that are not in need: Alpha risk due to overestimation; 2) Leaving contaminated territories untouchable: Beta risk due to underestimation.

- The financial loss of each site declared contaminated due to overestimation equals 0, if $z(x_0) > z_c$ (the site is indeed polluted), or w_1, otherwise (cost of cleaning a site which, in fact, does not need it).
- The financial loss of each site declared safe due to underestimation equals 0, if $z(x_0) \leq z_c$ (the site is indeed clean), or $w_2 \times [z(x_0) - z_c]$, otherwise (cost of not cleaning a site which, in fact, is in need), where w_2 is a relative cost of underestimating the same toxic concentration. According to (Goovaerts, 1997), this health cost variable is in units of money/concentration such as dollar/ppm.

After each health and remediation cost has been computed for each x_0 location, the site is categorized. The location is declared safe or contaminated so as to minimize the resulting expected loss based on the following two principles:

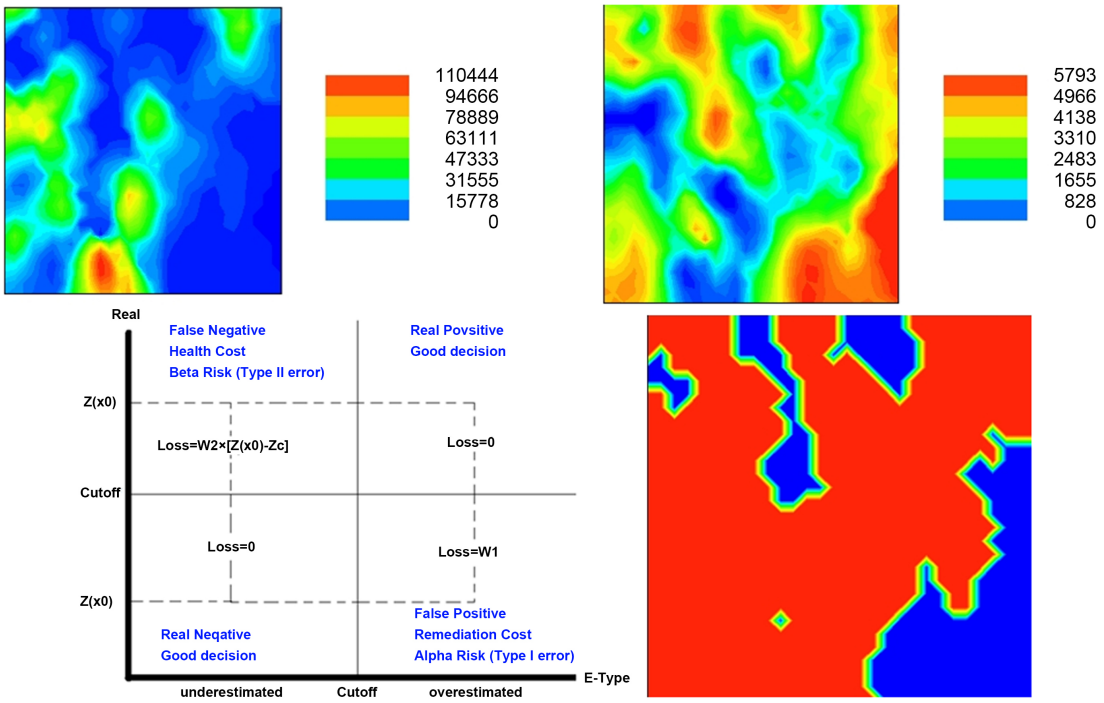


Figure 23. Estimation costs for a hypothetical health cost (w_1) of 1000 USD and remediation cost (w_2) of 6000 USD.

- If the health loss cost is greater or equal to the remediation loss cost, then x_0 site is classified as contaminated (red regions).
- If the health loss costs are lower than the remediation loss cost, then x_0 site is classified as safe (blue regions).

Considering the present case and cost charges, the range of health expenditures varies between 0 and 114,388 USD. Similarly, the remediation fee goes up to 6000 USD. Finally, the total health and remediation cost of the whole region equals 19,470,704 USD and 2,765,446 USD, respectively.

7. Conditional Gaussian Geo-Simulation

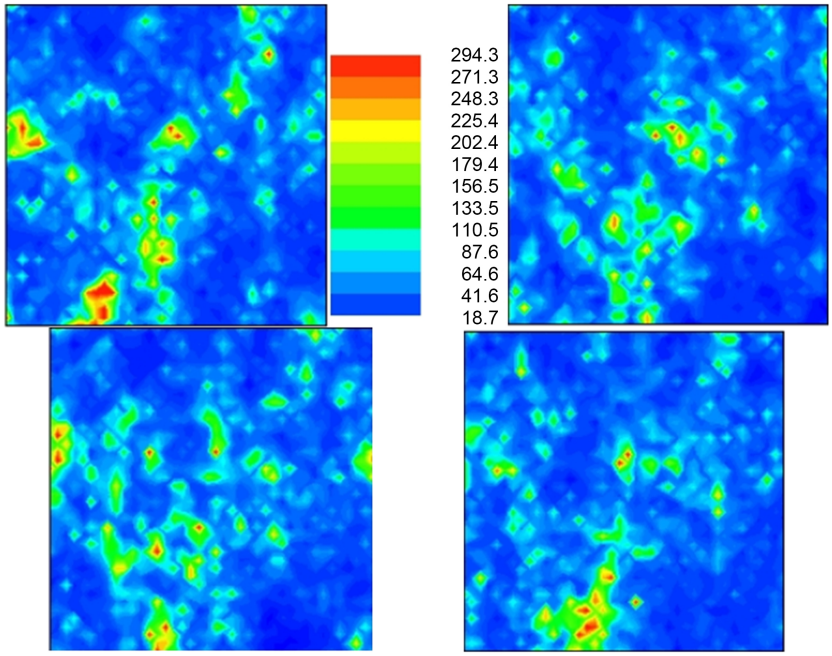
Ordinary Kriging (OK) tends to underestimate values that are larger than the average and to overestimate those that are smaller. It behaves in the same way as the conventional regression: its final estimates are less variable than the true values (the smooth effect). Although a Kriged map shows our best estimates of a variable, it does not represent the variation in a proper way, that is, this loss variation detail could be misleading. To obtain a statistical surface that retains the original samples variation, we need some other technique. Simulation is such a technique (Goovaerts, 1997).

It consists in the generation of N data sets of simulated values. Each realization holds the same statistical features such as the mean, variance, histogram and covariance of the original observations. As well, simulated values honor the measured data values at their locations. Based on N reproductions of any spatial phenomenon, it is feasible to compute uncertainty and the extreme spatial beha-

vior of that particular phenomenon. For instance, from a set of maps, it is possible to assess all sub-regions that overcome a certain threshold and their probability. But, even more important, it is possible to generate a set of realizations (Figure 24), yielding the smallest (best scenario) and the largest (worst scenario) of contamination costs (if that is the study case).

Based on ten realizations generated by myGeoffice©, Figure 25 (top-left, top-right and bottom) shows the E-Type (average) estimate and the minimum and maximum extreme behavior of this spatial Pb contamination case, respectively.

The map on the left in Figure 26 highlights the probability of the ten realizations being greater than 400 mg/kg (a highly dangerous area for any type of living organism). Computationally, if two or six out of the ten realizations (for a particular location) hold a value greater than the chosen threshold, for instance, then the probability assigned becomes 0.2 and 0.6, respectively.



REALIZATION 1
Sample Size: 1600
Conventional Average or Mean: 63,6660573399998
Maximum: 300
Minimum: 18,68
Conventional Variance: 2089,11834139983
Conventional Standard Deviation: 45,7068741153869
Skewness Index: 2,87010300151402

REALIZATION 2
Sample Size: 1600
Conventional Average or Mean: 59,4010070799999
Maximum: 300
Minimum: 18,68
Conventional Variance: 1359,44460115894
Conventional Standard Deviation: 36,8706468774138
Skewness Index: 2,53700954894525

REALIZATION 3
Sample Size: 1600
Conventional Average or Mean: 59,87656195625
Maximum: 300
Minimum: 18,68
Conventional Variance: 1388,11951483973
Conventional Standard Deviation: 37,2574759590573
Skewness Index: 2,78406018110022

REALIZATION 4
Sample Size: 1600
Conventional Average or Mean: 59,2080473937499
Maximum: 300
Minimum: 18,68
Conventional Variance: 1490,3566254157
Conventional Standard Deviation: 38,6051372930559
Skewness Index: 3,01340653055019

Figure 24. Four realizations of the spatial distribution of lead (Pb) contamination values over the Swiss Jura region. Curiously, the descriptive statistics of each realization (bottom) is quite similar. (Although not shown here, the variogram structure would have similar characteristics, too).

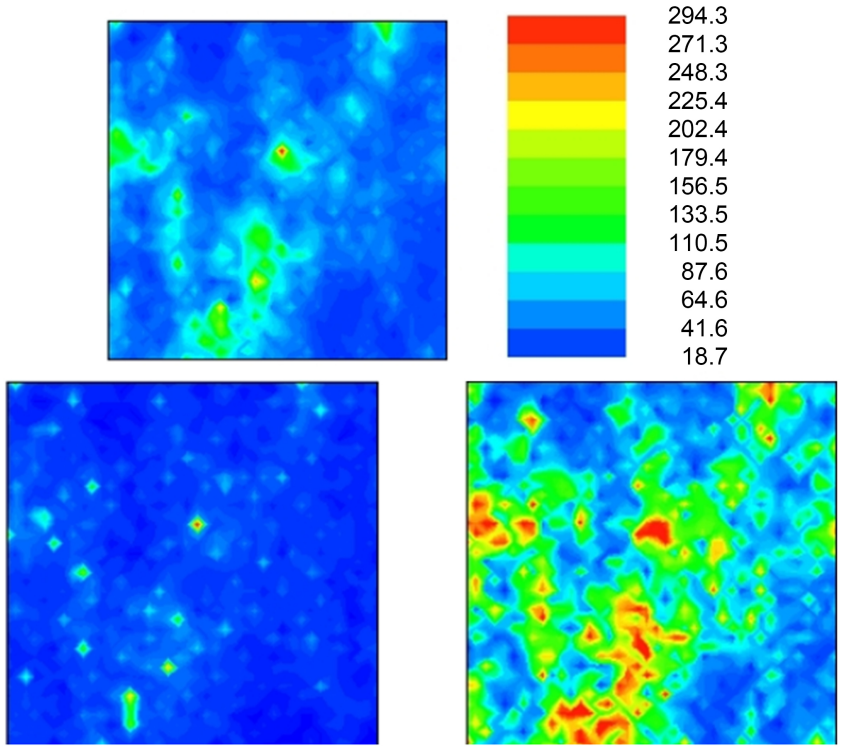


Figure 25. It is important to highlight that this technique is based on the assumption that any spatial distribution of a continuous random variable can be modeled by a multivariate Gaussian model.

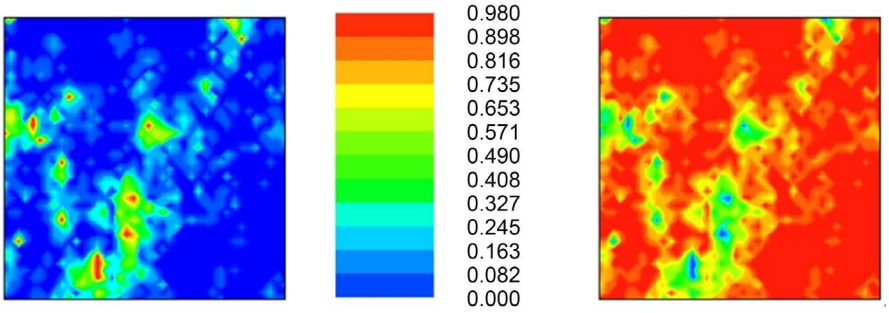


Figure 26. The right-layout represents the probability of the ten realizations being less than or equal to 100 mg/kg.

8. Conclusion

In a world driven by pollution issues, water is the major one due to its scarcity in certain regions of the globe. Known as one of the highest qualities for standard of living, Switzerland should be an inspiring example to the world (Volken & Rüesch, 2012), starting with its beautiful Jura lake. Regardless of this marketing perception, a lake is always a key asset for agriculture and tourism purposes (Wei et al., 2014).

Four goals should be highlighted under these conditions: 1) to describe the variability and spatial distribution of lead, 2) to compare the results of OK (Ordinary Kriging) and SGS (Sequential Gaussian Simulation), 3) to determine the

uncertainty in predicting the spatial variability of Pb and 4) to estimate a theoretical cost analysis.

Ten realizations were computed, the minimum, medium (E-type), maximum and standard deviation maps were produced. It seems that there is a prevalence of high Pb values in the lower-left and center of the lake whereas low values are distributed throughout the whole region. As well, a high degree of similarity between the E-type estimates and those generated by Ordinary Kriging (OK) can be found. Essentially, the average map represents the dominant patterns in the region while the minimum and maximum estimations represent the best and worst possible scenario for this Pb variable.

The Jura Lake is an optimal organizing unit for dealing with the management of water and closely related indirect resources, particularly fishes and other mammals. The responsible range of stakeholders varies with scale and must be clearly defined so that the costs and benefits associated with any plan are fully taken into account. Yet, it is easiest to implement them at the local ward level where the political, institutional and funding decision making grows especially complex.

As well, scientists and managers should strive to educate the public and private companies to avoid the concentration of heavy metals in the Jura Lake. Section 6.1 (IK uncertainty interpolation), Section 6.4 (probabilistic intervals) and Section 7 (Conditional Gaussian geo-simulation) really confirms that certain areas requires a future cleaning if the environment is of concern by the local authorities.

The costs presented here highlights a first approach associated with this water resource and environmental issue. Positively, different thresholds will lead to different budgets. Yet, the indirect profits on tourism and agriculture, for instance (not assessed here in this writing), will probably overcome by far the direct cleaning costs.

Concerning the municipal authorities of the Jura district, decision-making may lead to different policies regarding prevention measures, like increasing overall hydraulic capacity of the sewage collection system or promote self-awareness publicity to their own citizens, such as not to pour fat oils under the sink or to not flush pills in the toilet. On the other hand, if no measures are taken, at least, some areas of this lake must be prohibiting to any living organism.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Cardillo, M., Bromham, L., & Simon, J. (2016). Greenhill Links between Language Diversity and Species Richness Can Be Confounded by Spatial Autocorrelation. *Proceedings of the Royal Society B: Biological Sciences,* 78-86.

Cressie, N. (1993). *Statistics for Spatial Data (Revised Edition).* USA: John Wiley & Sons,

Inc., 928 p.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. USA: Oxford University Press, 483 p.

Griffith, D., Morris, E., & Thakar, V. (2016). Spatial Autocorrelation and Qualitative Sampling: The Case of Snowball Type Sampling Designs. *Annals of the American Association of Geographers, 106,* 773-787. https://doi.org/10.1080/24694452.2016.1164580

Margaret, A., & Webster, O. R. (2015). *Basic Steps in Geostatistics*: *The Variogram and Kriging*. USA: Springer, 106 p.

Mera, E., Condal, A., Rios, C., & Silva, L. (2016). Characteristic Variogram for Land Use in Multispectral Images. *Journal of Physics: Conference Series, 720,* 012047. https://doi.org/10.1088/1742-6596/720/1/012047

Negreiros, J. (2011). *Spatial Statistics for Special Data*. Macau: Lulu Press, 165 p.

Negreiros, J. (2015). myGeoffice©: Overview of a Geographical Information System. *Proceedings of 14th ISERD Conference,* Malaysia, 12-16.

Negreiros, J. (2017). *Spatial Analysis Techniques with myGeoffice©*. USA: IGI Global, 409 p.

Negreiros, J., Aguilar, F., & Aguilar, M. (2012). *Lectures on Spatial Statistics for Geographical Information Systems*. Macau: Saint Joseph Academic Press, 207 p.

Volken, T., & Rüesch, P. (2012). Risk of Overweight and Obesity among Migrants in Switzerland. *Health, 4,* 514-521. https://doi.org/10.4236/health.2012.48082

Wei, X., Peng, J., & Cao, L. (2014). An Empirical Study of the Relationship between Agriculture Environmental Efficiency and Economic Growth. *Modern Economy, 5,* 598-608. https://doi.org/10.4236/me.2014.55056