The local confidence uncertainty plume of SAKWeb$^{©}$

J. Negreiros[1], M. Painho[1], A. Cristina Costa[1], P. Cabral[1]
& F. Aguilar[2]
[1]Instituto Superior de Estatística e Gestão de Informação,
Universidade Nova de Lisboa, Lisboa, Portugal
[2]Escuela Politécnica Superior, Universidad de Almería, Almería, Spain

Abstract

The main goal of this research paper is to introduce a new uncertainty tool based on the Moran I correlogram, rescaled OK variance and local variance in a Web environment. It is hoped that this implementation will be used by users with problems to layout risk analysis environmental maps and plumes assessment. Spatial analysis, Moran I and other uncertainty measures are also reviewed.

Keywords: GIS, spatial analysis, Kriging, variogram, Moran I, local variance, local confidence, plumes assessment, SAKWeb$^{©}$.

1 Spatial analysis: overview

The geographical view of spatial analysis is essentially cartographical driven regarding the recognition and description of spatial patterns involving simple statistics and the direct use of visualization. Anselin [1] wrote that spatial data analysis could be defined as the study of statistical phenomena that manifest themselves in space. As a result, location, area, topology, spatial arrangement, distance and interaction become the focus of attention. As confirmed by Longley *et al.* [9], spatial analysis is a set of methods whose results change when the object location being analyzed changes too. Hence, spatial analysis is more widespread than statistical analysis of non-spatial information because it requires access not only to attributes but also to location and topological knowledge. It could be described as part of the process of transforming spatial data into geographical knowledge. Quoting Longley *et al.* [9], it can make what is implicit explicit.

doi:10.2495/RISK080091

Clark and Hosking [6] see spatial analysis as spatial modelling of a decision support such as GADS for solid waste spatial planning. In conjunction with the network and spatial analysis of GIS modules, the DSS Location Planner© analyzes market saturation, retail facilities accessibility, population mobility and demand-supply prediction based on demographic and socio-economic attributes, warehouse locations, distance or travel time between sites and expenditure flows between demand and supply chains (Arentze *et al.* [3]). However, according to Openshaw [18], an emphasis of DSS is a convenient distraction to hide a lack of the relevant GIS technology.

Ignoring technical matters and human technology resistance, spatial DSS, education and W3 services create political pressure on governments regarding public policy decisions in such fields as environmental protection, natural resource management, hazardous waste location and regional development ('government of the people, by the people and for the people'). This millennium will be about slowing the rate of deforestation, improving water quality, restoring wildlife habitats and understanding the earth's limitations (ArcNews [2]). Thus, the handling of spatial data should generate definitive answers.

McMaster [10] confirms this belief in his hazardous materials modelling for Santa Monica, CA. By identifying the explosives, flammable gases, solids and liquids, radioactive materials, corrosives and poisonous materials present within the community of 100,000 people, the evacuation plan by the Santa Monica Police and Fire Department became very clear. Yet, Yu [25] substantiate the difficulty of reaching a good balance between acceptable landscape planning in the Red Stone National Park, South China, and conflicts with ecologically protective boundaries for endangered medium-sized mammals and amphibian species, tourist preferences and newly reclaimed agriculture land. It is not easy to find the perfect solution. For Vale [20], resolution of this conflict for the same space involves three solutions: hierarchical dominance (certain matters are more important than others), multiple use (the same space may have several uses) and trade-off (certain issues are chosen to the detriment of others).

Another classical time example is shoreline limits, which are particularly useful for the study of global warming effects on coastal cities. Which shoreline should be adopted? Should the geographical database be dynamic and capable of tracking fluctuations? One possibility is to consider the shore slope classification, the sine angle and the time sea level above the height of the lowest tide. Since tides follow a deterministic formula, it is possible to calculate the exact shoreline location for a given *t* time. It is implicit that spatial analysis is a GIS component to support decision-making for solving problems with a spatial component.

Rossiter [19] also includes spatial flow modelling and deterministic processes like groundwater movement and environmental quality management based on economic criteria such as land use and transportation. It seems that geographical analysis comprises GIS (an applied computer-science view), spatial statistics including uncertainty issues (spatial autocorrelation, spatial autoregression, Kriging, stochastic simulation, morphologic geostatistics and space-time processes), classical aspatial statistics, remote detection and deterministic spatial analysis such as optimization routing, B-Splines, overlay, buffering and DEM operations (cartographic modelling).

Clearly, two global methods emerge regarding spatial analysis: the deterministic and the stochastic view. If the former has no capability for assessing uncertainty, the latter must be viewed as a process instead of a single function with an increase in computer power demand, particularly with variography and Kriging spatial interpolation.

Briefly, Kriging is a geostatistical estimation technique. It uses a linear combination of surrounding sampled values to make such predictions. To make such predictions, the Kriging system needs to know the weights applied to each surrounding sampled data. In fact, it allows deriving weights that result in optimal and unbiased estimates (within a probabilistic framework, Kriging attempts to minimize the error variance and systematically set the mean of the prediction errors to zero, so that there are no over or under estimates). However, it is the variogram that underpins Kriging. The variogram allows one to quantify the correlation between any two values separated by a lag distance h and uses this information to make predictions at unsampled locations by assigning different weights within the Kriging system.

2 The enhanced local confidence interval of SAKWeb©

The main issue discussed here regards risk analysis and uncertainty because spatial data lives with it. Uncertainty is a dimensionless parameter for which high values are bad and lower ones are optimal. Thus, uncertainty must be space geometry dependent because areas away from sample locations hold higher uncertainty. It must also take into account the variability of sample values.

Since different interpolation procedures may give dissimilar results and ground truth can never be known, it may be useful to know what the predicted chance of exceeding a given upper limit is, so decisions about expensive cleanup operations can be well founded, for instance. With agricultural applications, administrators might be interested to know how much of the whole population would give a higher return than the value of a certain crop while supervisors might be looking at toxicity levels. The same question arises when the fire department calculates tree density in tropical forests, when the fishery service computes water salinity and the density of shellfish or when engineers evaluate slope stability conditions to decide on the best route for a new road. Irrespective of the circumstances, the key question is to determine how much of the population is likely to lie above or below a cutoff value.

Two topics emerge from this perspective. Firstly, the choice of the probability threshold can be subjective. Secondly, the estimation error may be ignored because the contaminated location can be declared safe on the basis of an estimate of pollutant concentration, which is incorrect but slightly less than the regulatory threshold. If the true and the estimated values belong to quadrant II on an Estimated versus True grade plot, a good opportunity to invest can be missed. If it falls within quadrant IV, expensive consequences can be expected.

2.1 Overview of uncertainty measures

How can we quantify these uncertainties? The misclassification risk associated with a particular physical cutoff definitely increases at threshold location boundaries. If the goal of a manager's decision is to minimize unnecessary cleansing and ill health costs (in conjunction with a pre-setup deterministic cost) then it is possible to layout the total spatial health and remediation costs based on the resulting expected false negative error and false positive error models.

With regard to economic land evaluation, linear programming including sensibility analysis is a new possibility. Burrough [4] presents a gradual deterministic response of PH crop illustrating soil acidity impact on crop growth: No crops, if PH>=7; Normal growth, if PH<=5; (7-PH)/2 of crops, if 5<PH<7.

By assuming an unknown constant mean, non-linear Probability Kriging (PK) and Indicator Kriging (IK) present an alternative uncertainty method whose final estimates show the probability mapping for exceeding a given cutoff. Because of the Kriging smoothing effect, the local distribution of Kriging estimates is conditionally biased leading to a false and biased probability distribution, especially when the cutoff value becomes very high or very low. Juang and Lee [8] point out that PK accuracy is much higher than that of IK in their probability estimation of heavy–metal concentration in Taiwan. According to both authors, it yields more space variability and it behaves better with a screen-effect situation by reducing the risk of getting inconsistent probabilities. In addition, with both approaches, the local uncertainty cumulative distribution is taken into consideration, which avoids the strong assumption about the spatial distribution.

Geostatistical simulation tries to reflect the mean, the histogram, the covariance structure and the spatial data variance characteristics of the original dataset while, at the same time, it makes the simulation value close to the real one. According to Wang and Zhang [23] and Chainey and Stuart [5], by generating multiple and unique interpretations that respect the spatial dataset, simulation generates multiple configurations of possible realities, a realism issue, based on the search window Gaussian distribution but not on the optimum estimation. Certainly, the greatest potential of geostatistical simulation lies in the production of uncertainty estimations for a given cutoff value and, therefore, the assessment of impact costs.

The uncertainty layout of the conventional Kriging software is closely related to OK variance, too. According to Soares [22], if the sum of the weights is one and the average estimation error equals zero then the Kriging error variance becomes $\sigma^2_{OK} = \sum_{i=1}^{n} w_i \gamma(x_i, x_0) + \Psi$, where Ψ equals the LaGrange multiplier of the OK system, w_i is the OK weights and $\gamma(x_i, x_0)$ is the variance between the i^{th} and the estimated point. Hence, if errors respect the 'bell' curve then real values will fall within the Kriging_predictor $\pm 2\sigma^2_{OK}$ interval for a 95% confidence level. However, uncertainty is not included with variogram estimation and, thus, prediction variance is underestimated. But even more critically, OK variance is not sensitive to local error for two major reasons: 1) It is based on the same global variogram; 2) Distances among locations are the only relevant factor. 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.

2.2 The Moran I

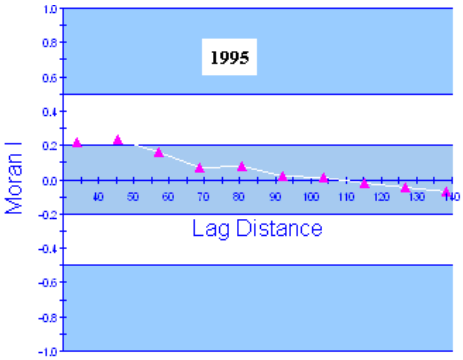
In order to understand the Local Confidence Interval of SAKWeb©, the first free Internet application for spatial interpolation and spatial autocorrelation indices (Negreiros [11], Negreiros and Painho [15]), the Moran I is reviewed here since this measure is an inherent element of it. Under the spatial independency null hypothesis, the global Moran I index (see equation 1 where $\bar{x}$ symbolizes the overall mean, W_{ij} is the weight of matrix W between ith and jth, n represents the total number of observations, x_i and x_j are ith and jth sample value while S_0 equals the sum of the spatial weights) is evaluated by measuring the covariance between attributes at each place and near sites towards the overall mean. If both neighbouring values are above or below the mean (similar high-high or low-low values), the product is positive, reflecting the presence of a similar spatial autocorrelation. Otherwise, the product of the two mean deviations will be negative (unrelated high-low and low-high values), indicating a non-positive one.

$$I=\frac{n\sum_{i=1}^{n}\sum_{j=1(i<>j)}^{n} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{S_0\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{1}$$

Its output domain varies between +1.0 (similar patterns were found) and –1.0 (nearby areas tend to be dissimilar) although a zero outcome for all neighbourhood distances denotes a pure nugget-effect situation for Kriging. For Wong and Lee [24], the mean of Moran I equals $E(I)=-(n-1)^{-1}$ although this value will tend to zero as the sample size increases. In addition, the Moran I cumulative peak should be analogous with the range for the sample variogram and it can be expected to have a similar value. If these ranges do not match then larger-scale patterns were not modelled for larger lags by the variogram (Skinner [21]).

2.3 SAKWeb© strategy

One new extension of SAKWeb© to enhance the local confidence interval is based on the local error variance as a true representative of the local pattern of spatial continuity. Since the shape of the variogram is the same for the whole study region, a hard assumption, to rescale the local variogram should be computed to reflect local spatial variability. According to Isaaks and Srivastava [7], the error variance of a relative variogram with a sill of one multiplied by the conventional local variance achieves this aim in a proper way. In practical terms, to produce a variogram sill of one, thus, each coefficient of the original model



Highest Moran I Range: 45.6700 **Moran I:** 0.2310

Second Highest Moran I Range: 34.0778 **Moran I:** 0.2181

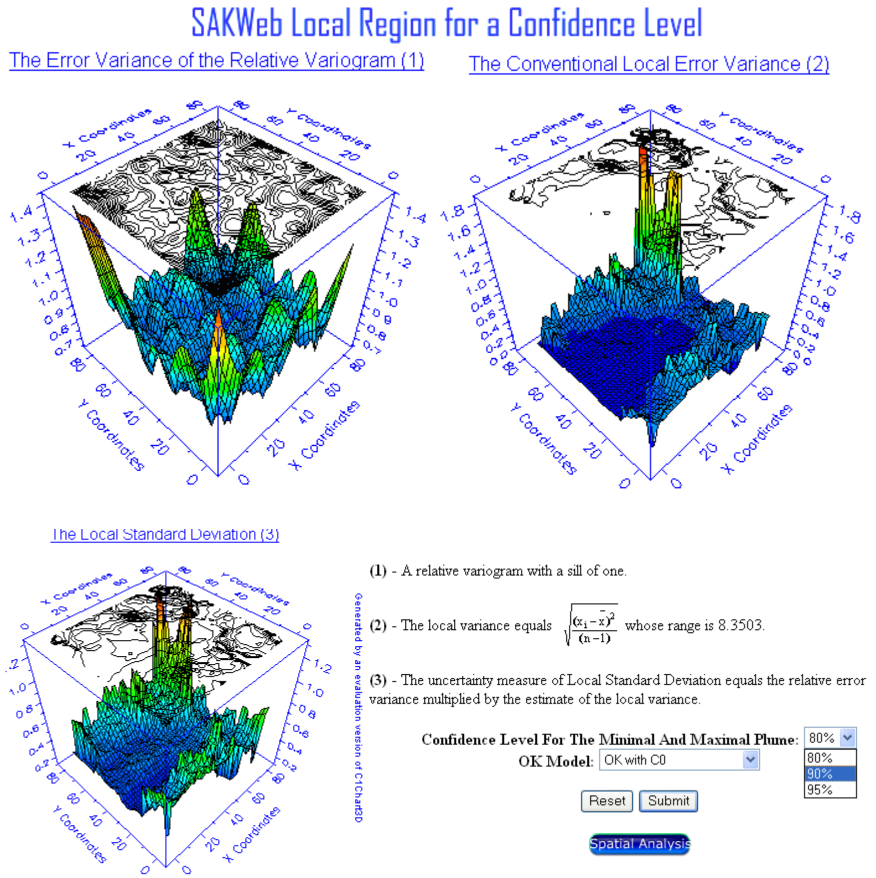
Figure 1: Moran I correlogram of SAKWeb©.

should be divided by its sill. With regard to the conventional local variance assessment, moving neighbourhood statistics can help this estimation, especially with the close relationship between local mean and local variance.

Under SAKWeb©, this local variability is computed based on the local error variance whose local range equals the greatest value between the lag with the highest and second highest found in the Moran I correlogram (Negreiros *et al.* [13], Negreiros and Painho [14]). That is, the global Moran I is computed for ten lag distances whose search range is between 3/2 of the average distance among samples, a nearest neighbourhood analysis parameter, and the variogram range. The goal is to identify the best scale of autocorrelation within spatial data. This happens because the proportional effect is not always true for all datasets and it is central that this factor becomes the most accurate as possible. According to the First Law of Geography, it is expected that this scale range varies between the first and third lag distance as Figure 1 (above) demonstrates.

In this particular case, the grasshopper 1995 infestation dataset of Colorado was used for layout purposes only. The highest spatial autocorrelation index achieved was at the second lag. Once the local error variance has been assessed based on the samples values that are within the range of the estimation point, then each local standard deviation uncertainty is multiplied by the rescaled OK error variance (see figure 2). The top left image shows the Ordinary Kriging variance based on a rescaled variogram, the top right presents the conventional local error variance while the bottom figure presents the local standard deviation estimated.

As expected, SAKWeb© offers the Normal 80%, 90% or 95% confidence intervals for three OK models (OK with nugget-effect, OK without nugget-effect and OK with a micro-scale component) but reflecting local conditions, as figure 3 shows. As expected, the user can setup a threshold limit (in this case, 0.5 grasshoppers per m^2) to assess the highest and lowest confidence plume (Negreiros [12], Negreiros *et al.* [16, 17]).



SAKWeb Local Region for a Confidence Level

The Error Variance of the Relative Variogram (1)

The Conventional Local Error Variance (2)

The Local Standard Deviation (3)

(1) - A relative variogram with a sill of one.

(2) - The local variance equals $\sqrt{\frac{(x_i-\bar{x})^2}{(n-1)}}$ whose range is 8.3503.

(3) - The uncertainty measure of Local Standard Deviation equals the relative error variance multiplied by the estimate of the local variance.

Confidence Level For The Minimal And Maximal Plume: 80%

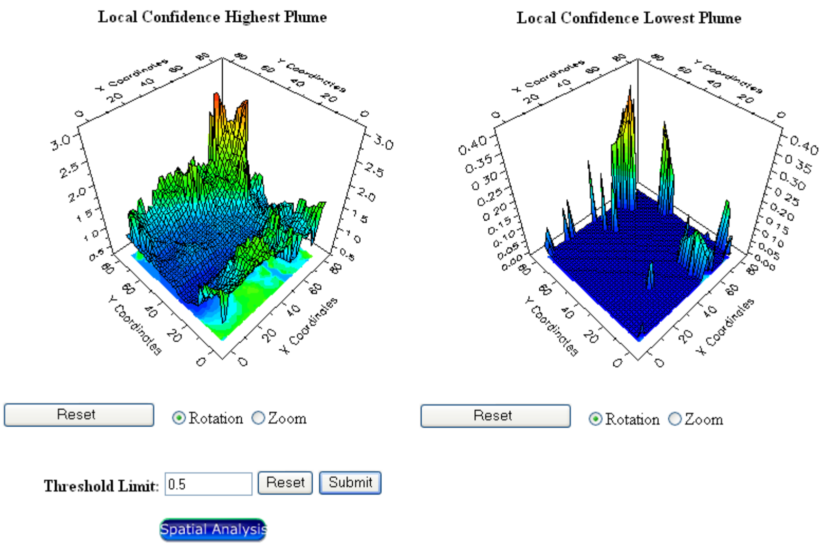
OK Model: OK with C0

Figure 2: Estimation of the local standard deviation for the same spatial dataset.

3 Conclusions

Although mathematics will always have limitations when describing space events because Earth has single and complex terrain processes (Negreiros and Painho [15]), new applied space formulations are becoming crucial. This happens because the kernel of geography is to think geographically, that is, to study spatial distribution phenomena and their correlations.

Thus, traditional statistics must be reformulated to properly account for spatial correlation and spatial heterogeneity within georeferenced data. In addition, users now have the ability to collect and to explore large amount of georeferenced data. With the advent of Web technology and modern wireless computing, it has become necessary to develop a W3 software for interpolation (a major inspiration for SAKWeb©) to understand the often complex spatial autocorrelation that exist among the samples collected in space.



● Rotation ○ Zoom

● Rotation ○ Zoom

Threshold Limit: 0.5

Figure 3: The third step process of the Local Region Confidence Interval option of SAKWeb©.

Another relevant issue lies in the implementation philosophy of theoretical research papers produced by SAKWeb©, e.g., the local confidence interval presented here. Quite often, the research papers end up on a library shelf without any application for the common GIS user. It is essential for theoretical research to be reflected in practical outcomes. OK variance is an index that does not depend on data although most users demand the use of this feature within geosoftware. Therefore, OK variance uncertainty was taken into consideration via a variogram rescale procedure with a sill equal to one. The Moran I correlogram must be produced in order to uncover the optimal highest scale of spatial autocorrelation among samples. Based on that lag distance, it is possible to compute the local conventional variance for each estimated point. By multiplying both factors and for a particular threshold value and confidence level, it is possible to draw two final maps for the study region that layout the highest and lowest plume of contamination, for instance. Although the OK interpolation procedure is not affected, certainly this local confidence interval can improve results when compared with traditional software.

References

[1] Anselin, L., Exploratory Spatial Data Analysis in Geocomputational Environment in Geocomputation A Primer, John Wiley & Sons, 1998.

[2] ArcNews, Newspaper, Fall 2000.

[3] Arentze, T, Borgers, A., Timmermans, H., A Generic Spatial Decision Support System for Planning Retail Facilities in Geographical Information Research, Taylor & Francis, 1998.

[4] Burrough, P. (1991). Principles of Geographical Information Systems for Land Resources Assessment, Clarendon Press-Oxford, 193 p.

[5] Chainey, S., Stuart, N., Stochastic Simulation: an Alternative Interpolation Technique for Digital Geographical Information, in Innovations in GIS 3, Taylor & Francis, p. 3–24, 1997.

[6] Clark, W., Hosking, P., Statistical Methods for Geographers, John Wiley & Sons, 1986.

[7] Isaaks, E., Srivastava, R., An Introduction to Applied Geostatistics, Oxford University Press, New York, 551 p., 1989.

[8] Juang, K., Lee, D., Comparison of Three Nonparametric Kriging Methods for Delineating Heavy-Metal Contaminated Soils, Journal of Environmental and Quality, 29, 2000.

[9] Longley, P., Goodchild, M., Maguire, D., Rhind, D., Geographical Information Systems and Science, John Wiley & Sons, 2001.

[10] McMaster, R., Modelling Community Vulnerability To Hazardous Materials Using GIS in Introductory Readings in GIS, Taylor & Francis, 1993.

[11] Negreiros, J., SAKWeb© (Spatial Autocorrelation and Kriging Web) – A W3 Computation Perspective, Unpublished Ph.D. Thesis, 449 p., 2004.

[12] Negreiros, J., Painho, M., The Web Platform for Spatial Statistical Analysis (#92), The Portuguese Conference of Information Systems 05, ISBN 13:978-972-789-219-8, Bragança, Portugal, 2005.

[13] Negreiros, J., Costa, A., Painho, M., Lopes, I., Spatial Autocorrelation and Association Measures, Encyclopedia of Networked and Virtual Organizations, Idea Group Reference, 2006.

[14] Negreiros, J., Painho, M., SAKWeb© – Spatial Autocorrelation and Kriging Web Service, Geo-Environment & Landscape Evolution II, WIT Press, p 79–89, ISBN 1-84564-168-X, Rhodes, Greece, 2006.

[15] Negreiros, J., Painho, M., SAKWeb© – Spatial Autocorrelation and Kriging Web Service (Part II), Internet Research 7.0: Internet Convergences (http://paginas.ulusofona.pt/p2203), Brisbane, Australia, 2006.

[16] Negreiros, J., Costa, A., Painho, M., Santos, J., Autocorrelation, Autoregression and Kriging: The Spatial Interpolation Issue, International Statistic Institute 56th Conference, Lisboa, Portugal, 2007.

[17] Negreiros, J., Costa, A., Painho, M., Santos, J., Lopes, I., Geostatistical Analysis: Software Flashpoint, Geocomputation 2007, Dublin, Ireland, 2007.

[18] Openshaw, S., Building Automated Geographical Analysis and Explanation Machines in Geocomputation A Primer, John Wiley & Sons, 1998.

[19] Rossiter, D., Geographical Information Systems, Cornell University, 1999.

[20] Vale, M., Construção de um SIG para Avaliação e Implementação do Plano de Ordenamento da Albufeira do Castelo de Bode, Dissertação Mestrado, ISEGI-UNL, 1994.

[21] Skinner, K. Spatial and Multivariate Analysis of Colorado Rangeland Grasshopper Abundances: Pattern and Process, Ph.D. Thesis, Colorado State University, p. 225, 1999.

[22] Soares, A., Geoestatistica Para As Ciências Da Terra E Do Ambiente, IST Press, 2000.

[23] Wang, X., Zhang, Z., A Comparison of Conditional Simulation, Kriging and Trend Surface Analysis for Soil Heavy Metal Pollution Pattern Analysis, in Journal of Environmental Sciences and Health, A34(1), Mercel Dekker Inc., p. 73-89, 1999.

[24] Wong, D., Lee, J., Statistical Analysis with ArcView® GIS, John Wily & Sons, Inc., p. 146, 2001.

[25] Yu, K., Ecologists, Farmers, Tourists - GIS Support Planning of Red Stone Park, China in Geographical Information Research, Taylor & Francis, 1998.