Risk assessment: the global confidence uncertainty plume of SAKWeb©

J. Negreiros[1], M. Painho[1], M. Aguilar[2] & F. Aguilar[2]
[1]*Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa, Portugal*
[2]*High Polytechnic School, Agriculture Department, Almería University, Spain*

Abstract

SAKWeb© (Spatial Autocorrelation and Kriging Web) is the first Internet geosoftware that provides access to a wider audience to boost geostatistical knowledge in conjunction with new tools. The enhancement of a global confidence uncertainty measure, based on the rescaled Ordinary Kriging (OK) variance of Isaaks and Srivastava (An Introduction to Applied Geostatistics. Oxford University Press, New York, 1989.), is the core of this paper. Hence, OK variance, variography, declustering methods and nearest neighbourhood analysis are reviewed in this paper. In addition, a detailed analysis regarding Geographical Information Systems (GISs) software implementation is presented, as well as SAKWeb© overall functionalities. It is expected that this accomplishment might be used by geostaticians with risk assessment troubles to layout a first raw global confidence plume after any Kriging surface interpolation.

Keywords: geographical information systems, spatial interpolation, Kriging, uncertainty, risk analysis, clean cost assessment.

1 Preamble

1.1 Implementation strategies for spatial analysis

At present, one software implementation solution is to relate the GIS to existing statistical packages, such as SAS-GIS®, SPSS-X®, Glim®, Systat® or Minitab®. The difficulty involved in this question is that statistical packages do not provide

WIT Transactions on Information and Communication Technologies, Vol 43, ©2010 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/RISK100031

the GIS functionality needed because they were not developed to handle spatial data or, particularly, the structure of spatial data switching back and forth. Furthermore, the large volume of spatial data, the complexity of the topological structure and the lack of interface transparency are a reality.

A second implementation strategy is to add statistical functionality into the GIS by modular integration, such as the Geostatistical Analyst® of ESRI. A major advantage of full integration of spatial analysis into commercial a GIS is good software documentation and ready availability of spatial analysis functionality for all GIS users. This solution can be more secure for users and simultaneously saves time. Moreover, users prefer to buy complete systems. Its deficiencies are closely related with its high costs in terms of software and maintenance contracts.

Created by universities and researcher groups, independent geosoftware becomes the third option. This procedure was applied by Goovaerts [7], for instance, with GSLib® in his multivariate geostatistical prediction by incorporating the raster Digital Elevation Model into rainfall predictions for Algarve, Portugal. Similarly, Nalder and Wein [10] used the GStat® Kriging package with Visual Basic® to generate the required data and command files for GStat®, while Chainey and Stuart [2] report the use of Turbo Pascal® to develop the Voronoi interpolator. Stein *et al.* [16] ran a groundwater flow simulation package, TRIWaco®, to model the dual aquifer of Goeree, Netherlands. Boykova [1] used GeoEAS® to estimate the depth discontinuity at the Moho Balkan Peninsula, a seismically unstable area. Geolith® is also referenced by the Chevron Petroleum Corporation (Frank *et al.* [4]). Contrary to common users, it seems that advanced researchers do not use proprietary GIS technology and choose to write their own programs or prefer specialized software in order to carry out their own research.

1.2 The Internet platform

Regardless of the chosen option, what the users appreciate and interact with is the Graphical User Interface (GUI), because they do not care about the technical structure, as long as the results are trustworthy, prompt and compatible with their operating system and hardware. Users want intuitive and easy-to-use software in order to give immediate results without having to read pages of documentation. The standard Web browser, whatever the background computer code adopted, fulfils this strategy quite well. It is cost free and already provided in many operating systems. In other words, does the user need to see a Word® document? It can be viewed in a browser. Does the user need to work on an Excel® spreadsheet? It can be opened in the browser. Does the user need to find local, network and Internet files? The browser can search for them. Ultimately, the user will work with all available software in the same way, regardless of the location of the data or its purpose (Negreiros and Painho [12]).

In the beginning, the trend was to find the data with the browser. Today, we can download the software and install it. In the near future we will just run the programs from the Internet. This represents a significant advance of the user

interface concept, because users will no longer have to worry about the software location and the technical knowledge necessary to connect data. The standard Web browser is the present-future interface. This solution relies on an enhanced standard front-end acting as a protected wall against the computer code in an embeddable, extendable and reusable development. The capability to undercover technical implementation to the final user via WWW is essential.

Under the GIS perspective and after struggling for years for digital information, spatial analysis needs to concentrate on what the information means, sharing it through the W3 and other distributed architectures. Extended Markup Language (XML) and Wireless Application Protocol (WAP) technologies are important players in this context. In effect, Web-enabled wireless devices are enabling millions of people to access the Internet while on the go. Therefore and under the GIS umbrella, improving data exploration by applying the available tools of spatial analysis and back-office technology integration with the Internet becomes critical.

1.3 SAKWeb© overview

SAKWeb© is not a comprehensive statistical package in the traditional of solving everyone's problems. Written for an Internet Information Service® (main Web Server technology of Microsoft®) environment, it was developed with the philosophy that spatial autocorrelation and Kriging software is needed as a learning tool by individuals with limited geostatistical knowledge. SAKWeb© deals with Kriging interpolation in conjunction with spatial association measures in a Web continuum process, instead of a loose local spatial function. From this view point, an element of its originality and innovation can, thus, be appreciated.

Basically, SAKWeb© presents four critical: 1) Data Input and Exploring View, which focuses on MS-Excel® input, control management of the user session and descriptive analysis. 2) Exploratory Spatial Data Analysis (ESDA) and Variography, which covers variogram setup, the Moran I correlogram, the Moran location scatterplot and the Moran variance scatterplot. 3) SAKWeb© Ordinary Kriging (OK), that concentrates on OK calculus and surface mapping in accordance with four nugget-effect strategies. Validation with an extra dataset, 3D-2D surface profiles, cross-validation and region plumes based on threshold values and confidence levels are also included. 4) SAKWeb© Help, which presents ten options regarding help with the software and e-Learning tools.

Still, its architecture, interfaces, comparative feedback and technical details are not presented here. This paper mainly focuses on the assessment and mapping of the highest and lowest Kriging interpolation surface (the plume concept) based on a certain confidence level of uncertainty and the Gaussian error interval assumption. Hence, it will be divided into six more sections. The impact of preferential sampling in common statistics and variography, in particular, is stressed in section 2 while the following one presents the main strategies currently available for samples weight declustering. Section 4 handles SAKWeb© capability to handle both previous issues in order to produce a global confidence interval of Kriging predictions. An illustrative theoretical example of sill rescale assessment is presented in section 5. SAKWeb© geocomputation

concerning this global confidence interval is analyzed in section 6 using a contaminant spatial dataset. As expected, the conclusion section follows next.

2 The impact of preferential sampling

Ordinary Kriging (OK) is a geostatistical estimation technique. It uses a linear combination of surrounding sampled values to make such predictions. Within a probabilistic framework, Kriging attempts to minimize the error variance and systematically set the mean of the prediction errors to zero, in order to avoid over or under estimates. Hence, this stochastic methodology describes the best linear unbiased estimator in the sense of least variance. Kriging is B.L.U.E. (best linear unbiased estimator) and B.U.E. (best unbiased estimator, if the data histogram respects the Normal curve).

Yet, it is the variogram that underlies Kriging. This spatial autocorrelation tool allows one to quantify the correlation between any two values separated by a lag distance of h and uses this information to make predictions at unsampled locations by assigning different weights within Kriging equation system. Generally, the variogram quantifies Tobler's Law at all scales by summarizing the degree of similarity between data values for all possible data pairs as a distance function.

The upper limit of the variogram is the sill which implies no spatial dependence between data points because all variances are invariant with the sample separation distance. The separation distance at which samples are spatially autocorrelated is the range. The nugget-effect, C_0, represents the measurement error variance and the spatial variation at distances much shorter than sample spacing, which cannot be resolved (GSLib [8]). The percentage ratio between the nugget-effect and total sill is called the relative nugget-effect.

The variogram sill is not a good global variance estimator because it is often higher than the global variance (Isaaks and Srivastava [9]). Due to preferential sampling, particularly in rich areas, the arithmetic mean becomes a poor global mean estimate. As well, preferential sampling affects variability measures. Regrettably, it is almost inevitable in mining that a geologist will schedule more samples in rich regions than in poor ones (Goovaerts [6]).

Isaaks and Srivastava [9], once again, demonstrated that additional sampling is most necessary in anomalous areas since it improves estimates where the proportional effect makes them least accurate. Even so, variograms with a proportional effect associated with clustered samples might not work well either, due to the apparent hole-effect, a dip variance at distances greater than the range. Certainly, a major variogram hard assumption is that only one single dominant correlation scale is assumed. In addition, heterogeneity on a larger scale is not apparent while heterogeneity on small scales is reflected on the nugget-effect. The variogram is a middle-aged man with difficulty seeing close objects and far away ones.

Hence, the global mean and global variance estimation will be inflated, leading to a higher variogram sill. Block Kriging (BK), for instance, should also include this discrimination concern since its global estimate is computed as an

even linear combination of block estimates. BK prediction should take into account the area of samples considered as a proportion of the total block area. For Simple Kriging (SK), the input sampling mean needed to achieve this Kriging variation is a requirement for this interpolation procedure. To include some type of declustering to avoid preferential sampling in order to attain a more accurate mean value should be a pre-requirement, as well. For normal score transformation (NST), once again, it is key that samples histogram reflects population histogram. If preferential sampling exists then its histogram will become wrong while NST results will become bias. Whatever is the situation, preferential sampling would lead to wrong estimates since our input dataset is not representative of the unknown reality.

3 The conventional declustering strategies

A plausible solution to overcome this preferential sampling issue for certain regions (quite often, high ones) is to assign different weights to samples (see figure 1). The clarification for preferential sampling is, then, to weight samples where representative regions of dense observations receive less weight while sparse samples receive additional one. Conventional cell declustering, for instance, assigns weights according to the cell size. Because each sample receives an inversely proportional weight to the number of points that fall within the same cell, consequently, several outcomes may emerge. With Geostatistical Analyst® of ESRI®, the cell size is determined from the maximum value of Morisita´s index (ESRI [3]).

With the polygonal method, it may also generate several layouts and, therefore, different results can be achieved (see figure 2). Confirmed by Frank *et al.* [4], when sampling size does not suggest a natural cell dimension, several cell sizes and origins must be tried. In addition, both methods have a difficulty in defining the limits across the research area edges, quite often leading too excessive weighting. With Geostatistical Analyst®, the outer rectangle boundary is formed by taking the largest (x,y) coordinates of all available samples plus $\sqrt{area/2}/2$ factor.

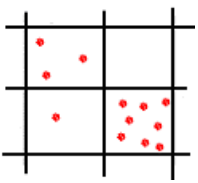

Figure 1: Method of cell declustering weight where each sample weight is inversely proportional to the total number of available samples of its respective square cell.

Figure 2: Two different weight assignments (left and centre) for the same sampling layout (right).

4 SAKWeb© declustering approach

Based on Kriging predictions, one possibility to create a raw Gaussian confidence interval (sometimes a hard spatial assumption) could be the use of the error Kriging variance (σ^2_{OK}) for a particular confidence level index (CLI): Kriging_Prediction $\pm CLI \times \sigma^2_{OK}$, where CLI factor follows the values of the Normal distribution with zero mean and variance of one. This means that for each Kriging prediction and for a certain confidence level, the OK variance should be added and subtracted to each interpolated value by a CLI factor. As expected, the CLI parameter equals 1.645, for instance, if the confidence level is 90%. Although OK variance is not a good uncertainty measure in terms of risk assessment since it does not depend on samples values, it could be used if the variogram sill would reflect somehow the local variance. Due to preferential sampling, some type of transformation should, then, be carried out to take into account the discrepancy between the sample data variance and the variogram sill. This happens because, most of the times, the latter is quite higher than the former one because sampling is not representative of the true population distribution.

SAKWeb© declustering choice is based on the nearest neighbourhood analysis where each sample weight relies on the nearest neighbour distance among all samples and the estimated one. According to Fig. 2, sample 1 should hold the highest weight influence, quantified by the nearest distance between samples 1 and 4. As expected, the nearest neighbour distances among samples 2, 3 and 4 are smaller and their weights for representativeness become also less. Notice that the sum of these weights equals one. Further, political boundaries or coastlines delimitation areas do not affect this methodology because nearest neighbour distance only depends on samples coordinates.

Another main cause of this enhanced declustering approach is closely related with its geocomputation because it can be difficult to implement the polygonal approach over the Web platform since it would be necessary to upload all GIS topological structures. As well, cell declustering depends heavily on the size and the layout considered of the cell grid, thus, creating a critical decision for the common user. Suppose that if samples layout follows a U shaped. With polygonal approach, samples that are located on the inner border are given too much weight because they represent a vast area without any observations. Nearest neighbourhood analysis does not reflect this concern.

A last topic relates to accuracy. By considering Fig. 2, samples weights fluctuate quite different: 0.49, 0.17, 0.17 and 0.17, if nearest analysis is considered, versus 0.27, 0.33, 0.16 and 0.24, for polygonal approach, versus 0.35, 0.15, 0.15 and 0.35, for cell declustering. It is imperative to stress than none of these approaches is considered 'the best'. However, it can be guaranteed that all these declustering methods for sampling weight will always lead to a tremendous improvement in local and global distribution estimates when compared with the traditional statistical approach where all observations have the same weight (Isaaks and Srivastava [9]).

5 Rhetorical case study of rescale assessment

The purpose of this example is to demonstrate nearest neighbourhood declustering approach and variogram sill impact, regarding their computation. The city of San Diego, CA, is not a uniform area concerning housing costs where reasonable small portion of the urban area follow a distinct spatial autocorrelation affinity similar to main cities in this world (Getis and Ord [5]). The housing costs mean of each ward, in 1989, equals $192.81 while skewness and kurtosis are 0.83 and 0.20, respectively.

As expected, the non-convergence trend regarding sill variogram overestimation against conventional global variance was a reality: 6497 (sill variogram) versus 5523 (standard variance). When nearest neighbourhood distance was applied to samples weights, the estimated global mean (EGM) and the estimated global variance (EGV) became $EGM=\sum_{i=1}^{n} w_i x_i = 199$ and $EGV=\sum_{i=1}^{n} w_i (x_i - EGM) = 5379$, where x_i represents the value of each ward housing cost while w_i equals each weight of each ward among the San Diego county region. As mention before, all these weights are based on nearest neighbourhood analysis whose weights sum equals one.

The computation ratio between EGV and variogram sill can lead to a reliable improvement of OK variance although this sill rescaling operation does not affect Kriging estimation. The initial variogram was as follows: $\gamma(\mathbf{h})=6497\times(1-e^{(-h/5.31)})$. Since the index ratio between EGV and variogram sill equals 5379/6497=82%, therefore, the rescaled variogram becomes $\gamma_1(\mathbf{h})=\gamma(\mathbf{h})\times 0.82=6382\times(1-e^{(-h/5.31)})\times 0.82 =5379\times(1-e^{(-h/5.31)})$, where e=2.71 while h symbolizes the lag distance between two generic spatial locations. This means an OK variance decrease of 7.86% (5272), on average.

6 The global confidence interval option of SAKWeb©

It is than possible to generate a 90% (Kriging_Prediction $\pm$ $1.645\times\sigma_{OK}$) or 80% (Kriging_ Prediction $\pm$ $1.282\times\sigma_{OK}$) global confidence interval, for instance, with a major result improvement due to a neighbourhood declustering and a sill

rescaled operation. On the basis of the Normal error distribution, SAKWeb© simulates the minimum and maximum global plume for three confidence intervals (80%, 90% and 95%) versus three Ordinary Kriging models (OK with two structures, OK with and without nugget-effect). Hence, nine configurations can be achieved. Afterwards, the computation and mapping of the largest and smallest area above a specific threshold can, then, be achieved. To setup the cutoff value is a user responsibility.

In order to demonstrate this procedure within SAKWeb©, the Pb contamination default dataset (128 samples) of GS+® geosoftware (Gamma Design) was used. The conventional mean, standard deviation, skewness, kurtosis and standard error deviation of the mean are 0.382, 0.206, 1.31, 2.74 and 0.019, respectively. The variogram adopted was a spherical isotropic one with the following parameters: C_0=0.0148, C_1=0.0319, range=73.9 (the fitness R^2 equals 94.5%).

By using the option Global Region Confidence Interval of SAKWeb©, a confidence level of 80% was chosen against the OK without nugget-effect model. The second step concerns the definition of the cutoff limit (in this case, 0.5 ppm was chosen). At this stage (middle image of figure 3), the left map depicts the highest plume map for the OK estimation according to a confidence level of 80%. As the reader may confirm, the left map always presents higher estimation values than the right one. The last step of this process regards the display of the lowest and highest plume for a given threshold value (bottom image of figure 3). In both cases, all interpolated values lower than the cutoff limit were setup to zero.

How can this risk measure be useful? 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, within environmental issues, supervisors might be looking at toxicity levels. The same question arises when the fishery service computes water salinity and the density of shellfish (Goovaerts [6]). Irrespective of the circumstances, the question is to determine how much of the population is likely to lie above or below a cutoff limit. To define those risk assessment areas is the aim of this plume based on a probabilistic thinking. In conjunction with a cost analysis technique (not available in SAKWeb©, yet), it would be, then, possible to assess a first comprehensible cleanness cost for this particular region by using a pre-defined price per meter square, for instance.

7 Final thoughts

In this particular article, a true geographical issue takes place in terms of risk assessment: spatial analysis computation of a global uncertainty measure via a rescaling operation (estimated global variance against variogram sill). Although OK interpolation procedure is not affected, the OK variance surface and global confidence interval can be improved, particularly when compared with traditional geo-software.

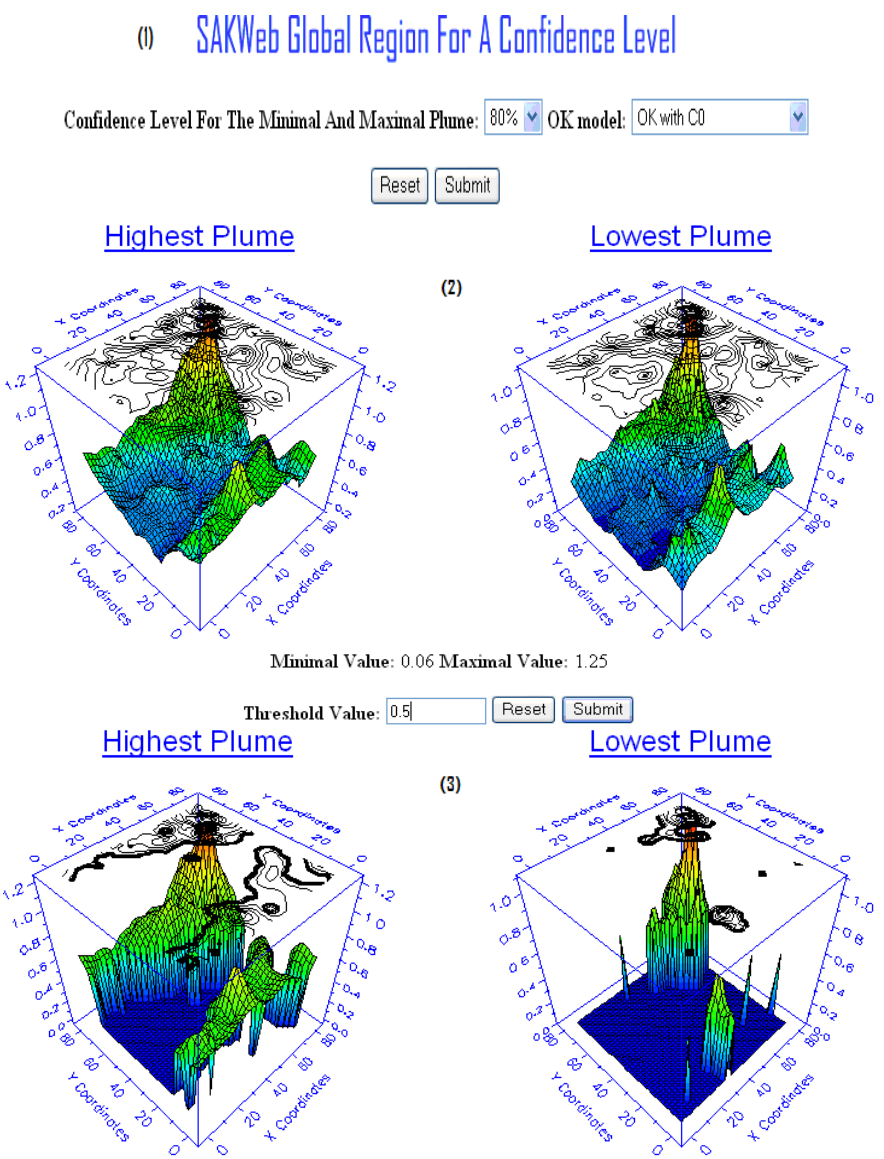

Figure 3: The Global Region Confidence Interval option of SAKWeb©, a three step process: (1) OK model and confidence level setup; (2) plume mapping plus definition of threshold limit; (3) regions that are above and below the threshold limit, that is, the 'worst' scenario (highest plume) and 'best' scenario (lowest plume) if a cleaning operation is considered.

SAKWeb© also follows the belief of bringing spatial analysis tools for the W3 environment since Internet is an ingredient of our future (Negreiros *et al.* [14]).

A final relevant question of SAKWeb© lays in the implementation philosophy of theoretical research papers produced by others researchers. The variogram rescaling operation of Isaaks and Srivastava [9] is an example of this conviction. 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 (Negreiros *et al.* [15]).

References

[1] Boykova, A., Moho Discontinuity in Central Balkan Peninsula in the Light of the Geostatistical Structural Analysis, Physics of the Earth and Planetary Interiors, 114, 1999.
[2] Chainey, S., Stuart, N., Stochastic Simulation: An Alternative Interpolation Technique for Digital Geographical Information in Innovations in GIS 3, Taylor & Francis, 1997.
[3] ESRI, Using ArcGIS Geostatistical Analyst. USA, 2001.
[4] Frank, J., Reet, E., Jackson, W., Combining Data Helps Pinpoint Infill Drilling Targets in Texas Field, Oil & Gas Journal, 1999.
[5] Getis, A., Ord, J., The Analysis of Spatial Association by Use of Distance Statistics in Geographical Analysis, 1992.
[6] Goovaerts, P., Geostatistics for Natural Resources Evaluation. Oxford University Press, 1997.
[7] Goovaerts, P., Geostatistical Approaches for Incorporating Elevation into The Spatial Interpolation of Rainfall, Journal Of Hydrology, Elsevier Science, 2000.
[8] GSLib, http://www.gslib.com, 2002.
[9] Isaaks, E., Srivastava, R., An Introduction to Applied Geostatistics. Oxford University Press, New York, 1989.
[10] Nalder, I., Wein, R., Spatial Interpolation of Climatic Normals: Test of a New Method in the Canadian Boreal Forest, Agriculture and Forest Meteorology, 92, 1998.
[11] Negreiros, J., SAKWeb (Spatial Autocorrelation and Kriging Web) – A W3 Computation Perspective. Unpublished Ph.D. Thesis, 449p, 2004.
[12] Negreiros, J. & Painho, M., The Web Platform for Spatial Statistical Analysis. The Portuguese Conference of Information Systems 05 (#92), 2005.
[13] Negreiros, J. & Painho, M., SAKWeb© – Spatial Autocorrelation and Kriging Web Service. Geo-Environment and Landscape Evolution II, 79-89, WIT Press, 2006.
[14] Negreiros, J., Painho, M., Costa, A., Santos, J., Lopes, I., Geostatistical Analysis: Software Flashpoint, Geocomputation Conference, Dublin, Ireland, 2007.
[15] Negreiros, J., Painho, M., Oliveira, T., Aguilar, M., Aguilar, F., Synopsis of an Internet Spatial (Decision Support) Prototype, WorldComp 08 -

International Conference on E-Learning, e-Business and e-Government, Hamid Arabnia and Azita Bahrami (Eds), ISBN 160-132-063-9, USA, 2008.
[16] Stein, A., Zaadnoordijk, W., Improved Parameter Estimation for Hydrological Models using Weighted Object Functions, Hydrological Processes, 13, 1999.