GEOSTATISTICAL ANALYSIS OF SPATIAL DATA

Goovaerts, P.
Biomedware, Inc. and PGeostat, LLC, Ann Arbor, Michigan, USA

Keywords: Semivariogram, kriging, spatial patterns, simulation, risk assessment

Contents

Summary

This article presents an overview of the geostatistical tools available for processing spatial data and illustrates the different steps of a typical geostatistical analysis using an environmental data set.

First, geostatistics provides descriptive tools such as semivariograms to characterize the spatial pattern of continuous and categorical soil attributes. Various interpolation (kriging) techniques capitalize on the spatial correlation between observations to predict attribute values at unsampled locations using information related to one or several attributes. An important contribution of geostatistics is the assessment of the uncertainty about unsampled values, which usually takes the form of a map of the probability of exceeding critical values, such as regulatory thresholds in pollution or criteria for soil quality. This uncertainty assessment can be combined with expert knowledge for decision making such as delineation of contaminated areas where remedial measures should be taken or fertile areas where specific management plans can be developed. Last, stochastic simulation allows one to generate several models (images) of the spatial distribution of attribute values, all of which are consistent with the information available. A given scenario (remediation process, land use policy) can be applied to the set of realizations, allowing the uncertainty of the response (remediation efficiency, soil productivity) to be assessed.

1. Introduction

During the last decade, the development of computational resources and geoinformatics has fostered the use of numerical methods to process the large bodies of data that are

measured in the geosciences. A key feature of geoscience information is that each observation relates to a particular location in space. For example, Figure 1 shows the spatial distribution of five stratigraphic classes and of concentrations of two heavy metals recorded, respectively, at 359 and 259 locations in a 14.5km area in the Swiss Jura. Knowledge of an attribute value, say a pollutant concentration, is of little interest unless the location of the measurement is known and accounted for in the analysis. Another feature is that the information available is usually sparse which, in combination with the imperfect knowledge of underlying processes, leads to a large uncertainty about the actual spatial distribution of values. Such an uncertainty needs to be quantified and accounted for in decision-making, hence probabilistic (statistical) tools are increasingly preferred to a deterministic approach where a single (error-free) representation is sought.

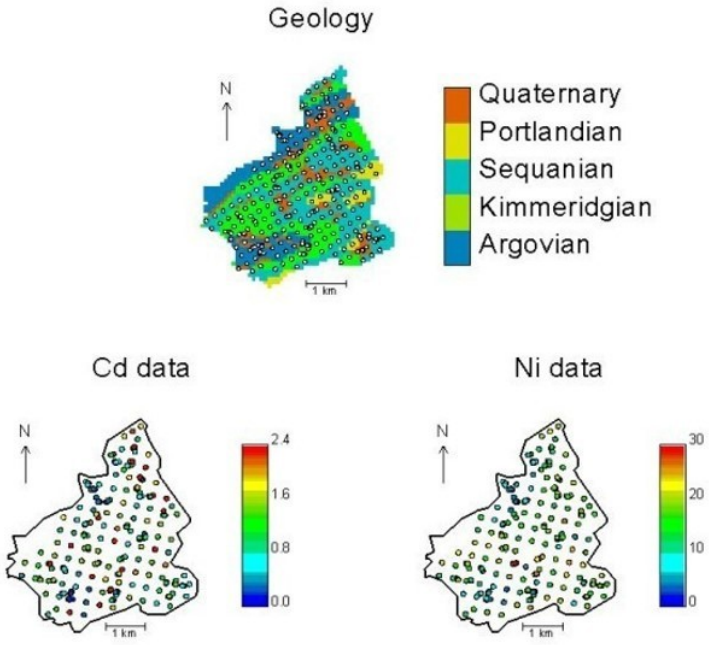

Figure 1. Locations of sampling sites superimposed on the geologic map, and concentrations in cadmium (Cd) and nickel (Ni) at 259 of these sites (units = mg kg^{-1}).

Different types of spatial data can be distinguished: lattice observations whose spatial locations are regularly spaced (e.g., gridded data, such as satellite sensor imagery or systematic soil survey), point patterns where the important variable to be analyzed is the location of "events", and geostatistical data which can be measured continuously in space (e.g., soil properties). This article deals with the latter type of data, which are the most common in the geosciences. It is noteworthy that geostatistics is also widely used for the analysis of remotely sensed data.

Geostatistics provides a set of statistical tools for incorporating the spatial and temporal coordinates of observations in data processing (see *Stochastic Modeling of Spatio-temporal Phenomena in Earth Sciences)*. It is a relatively new discipline, which was

developed in the 1960s, primarily by mining engineers who were facing the problem of evaluating recoverable reserves in mining deposits.

Priority was given to practicality, a current trademark of geostatistics that explains its success and application in such diverse fields as mining, petroleum, soil science, oceanography, hydrogeology, remote sensing, agriculture, and environmental sciences. Geostatistics can be used for three main purposes: 1) description of spatial patterns, 2) spatial interpolation, and 3) modeling of local and spatial uncertainty. In other words, looking at the example of Figure 1, here are some of the key issues that geostatistics allows one to address. What are the main features of the spatial patterns of heavy metals and how do they relate to the distribution of potential sources, such as rock types and human activities? What is the metal concentration that could be expected at an unsampled location? What is the probability that the regulatory threshold is exceeded at an unsampled location? Which areas should be remediated and what is the risk of making a wrong decision that is classifying as safe contaminated locations or classifying as contaminated safe locations? Where should additional observations be measured to increase the accuracy of the predictions and reduce the risk of misclassification? These few examples illustrate the potential of geostatistics for improving the understanding and characterization of the environment, leading to more accurate models of the spatial distribution of attribute values, and subsequently more informed decision-making.

2. Description of Spatial Patterns

Analysis of spatial data typically starts with a spatial "posting" of data values such as in Figure 1. For both continuous (metal concentrations) and categorical (rock type) attributes, the spatial distribution of values is not random, in that observations close to each other on the ground tend to be more alike than those further apart. The presence of such a spatial structure is a prerequisite to the application of geostatistics, and its description is a preliminary step towards spatial prediction or modeling of uncertainty.

Consider the problem of describing the spatial pattern of a continuous attribute z, say a pollutant concentration such as cadmium or nickel. The information available consists of values of the variable z at n locations:

$$\mathbf{u}_\alpha,\ \{z(\mathbf{u}_\alpha), \alpha = 1, 2, \ldots, n\},$$

where $\mathbf{u}_\alpha$ is a vector of spatial coordinates in up to three dimensions.

Spatial patterns are usually described using the experimental semivariogram $\hat{\gamma}(\mathbf{h})$, which measures the average dissimilarity between data separated by a vector $\mathbf{h}$.

It is computed as half the average squared difference between the components of data pairs:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + h)]^2 \qquad (1)$$

where $N(\mathbf{h})$ is the number of data pairs within a given class of distance and direction.

Figure 2 (the top graphs) shows the semivariograms computed from the data of Figure 1 using distance classes of 100~m. Data pairs in all directions were pooled, and such semivariograms are called omnidirectional. For both metals, semivariogram values increase with the separation distance, reflecting the intuitive feeling that two concentrations close to each other on the ground are more alike and, thus, their squared difference is smaller than those further apart. The two semivariograms stop increasing at a given distance, called the range, which can be interpreted as the distance of dependence, or zone of influence, of metal concentrations. The discontinuity at the origin of the semivariogram (i.e., at very small separation distances) is called the nugget effect and arises from measurement errors or sources of spatial variation at distances smaller than the shortest sampling interval or both. These graphs point out distinct spatial behaviors of the two metals: Ni concentrations appear to vary more continuously than Cd concentrations, as illustrated by the smaller nugget effect and larger range of its semivariogram. In combination with knowledge about the phenomenon and the study area, such a spatial description can enhance our understanding of the physical underlying mechanisms controlling spatial patterns. In the present example, the long-range structure of the semivariogram of Ni concentrations is probably related to the control asserted by rock type, while the short-range structure for cadmium suggests the impact of local human-induced contamination.

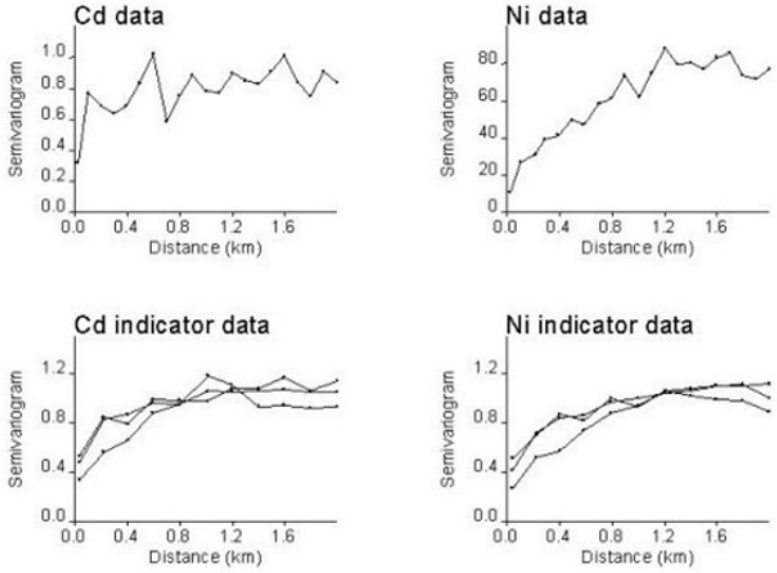

Figure 2. Experimental omnidirectional semivariograms for Cd and Ni: original concentrations and indicator transforms using thresholds corresponding to the second- (---), fifth- (--~--) and eighth-decile (-~-~-) of the sample histogram.

Spatial patterns may differ depending on whether the attribute value is small, medium, or large. For example, in many environmental applications, a few random “hot spots” of large concentrations coexist with a background of small values that vary more continuously in space. Depending on whether large concentrations are clustered or

scattered in space, the interpretation of the physical processes controlling contamination and the decision for remediation may change.

The characterization of the spatial distribution of z-values above or below a given threshold value z_k requires a prior coding of each observation $z(\mathbf{u}_\alpha)$ as an indicator datum $i(\mathbf{u}_\alpha; z_k)$, defined as:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) \le z_k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Indicator semivariograms can then be computed by substituting indicator data $i(\mathbf{u}_\alpha; z_k)$ for z-data $z(\mathbf{u}_\alpha)$ in the equation (1):

$$\hat{\gamma}_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [i(\mathbf{u}_\alpha; z_k) - i(\mathbf{u}_\alpha + h; z_k)]^2 \quad (3)$$

The indicator variogram value $2\hat{\gamma}_I(\mathbf{h}; z_k)$ measures how often two z-values separated by a vector $\mathbf{h}$ are on opposite sides of the threshold value z_k. In other words, $2\hat{\gamma}_I(\mathbf{h}; z_k)$ measures the transition frequency between two classes of z-values as a function of $\mathbf{h}$. The greater is $\hat{\gamma}_I(\mathbf{h}; z_k)$, the less connected in space are the small or large values.

Figure 2 (the bottom graphs) shows the omnidirectional indicator semivariograms computed for the second-, fifth- and eighth-decile of the distributions of cadmium and nickel concentrations. For both metals, indicator semivariograms for small concentrations have smaller nugget effect than those for larger concentrations, which suggests that homogeneous areas of small concentrations coexist within larger zones where large and medium concentrations are intermingled. Two clusters of small concentrations are indeed apparent on the location maps of Figure 1 and correspond to the Argovian rocks.

Many variables in geosciences, such as texture or land use classes, take only a limited number of states, which might be ordered or not. The spatial patterns of such categorical variables can also be described using geostatistics. Let S be a categorical attribute with K possible states $s_k, k = 1, 2, \ldots, K$. The K states are exhaustive and mutually exclusive in the sense that one and only one state s_k occurs at each location $\mathbf{u}_\alpha$. The pattern of spatial variation of a category s_k can be characterized by semivariograms of type (3) defined on an indicator coding of the presence or absence of that category:

$$i(\mathbf{u}_\alpha; s_k) = \begin{cases} 1 & \text{if } s(\mathbf{u}_\alpha) = s_k \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The indicator variogram value $2\hat{\gamma}_I(\mathbf{h}; s_k)$ measures how often two locations a vector $\mathbf{h}$ apart belong to different categories $s_{k'} \neq s_k$. The smaller is $2\hat{\gamma}_I(\mathbf{h}; s_k)$, the more connected is category s_k. The ranges and shapes of the directional indicator semivariograms reflect the geometric patterns of s_k.

Figure 3 shows the indicator semivariograms of two stratigraphic classes of Figure 1 computed in four directions with an angular tolerance of 22.5°. For both classes the indicator semivariogram value equals zero at the first lag, which means that any two data locations less than 100~m apart belong to the same formation. The longer SW-NE range (larger dashed line) reflects the corresponding preferential orientation of these two lithologic formations.

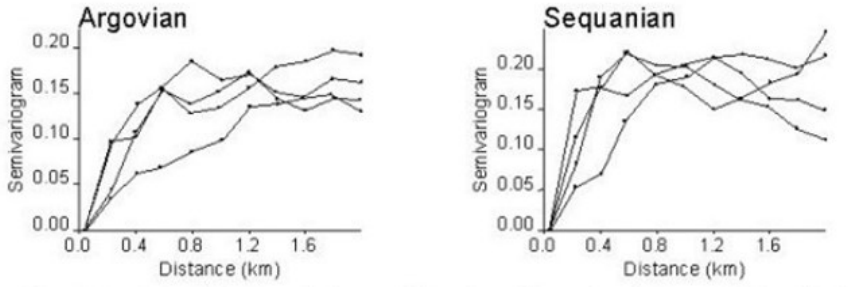

Figure 3. Experimental indicator semivariograms of Argovian and Sequanian rocks computed in four directions measured in degrees clockwise from North (---: 22.5, --~--: 67.5, -~-~- : 112.5, ~~: 157.5; angular tolerance = 22.5).

Information in the geosciences is often multivariate, and the semivariogram can be generalized to the bivariate case, allowing one to investigate how the correlation between two attributes (e.g., Ni and Cd concentrations) varies in space. Three-dimensional data can be analyzed using the same tools although the description and visualization of variability along the different dimensions becomes more complicated.

-
-
-

Bibliography

Cressie N. (1993). *Statistics for Spatial Data*, 900pp. New-York: John Wiley & Sons. [This book provides statisticians with a comprehensive presentation of tools for analysing the three main types of spatial data].

Chiles, J.P., Delfiner, P. (1999). Geostatistics. Modelling Spatial Uncertainty, 695pp. New York, NY, USA: John Wiley & Sons. [This recent book provides a thorough exploration of a wide spectrum of theoretical and practical aspects of geostatistics].

Deutsch C.V., Journel A.G. (1998). GSLIB: *Geostatistical Software Library and User's Guide*. Second Edition, 369pp. New-York: Oxford University Press. [This represents the most widely used library of public-domain geostatistical codes].

Dungan J. (1998). Spatial prediction of vegetation quantities using ground and image data. *International Journal of Remote Sensing* **19**, 267-285. [This work presents an application of geostatistics to the integration of remotely sensed and field data].

Goovaerts P. (1997). *Geostatistics for Natural Resources Evaluation*. 483pp. New York: Oxford University Press. [This book provides practitioners with a comprehensive presentation of the state-of-the-art in geostatistics].

Isaaks E.H., Srivastava R.M. (1989). *An Introduction to Applied Geostatistics*. 561pp. New York: Oxford University Press. [This book is a primer for a self-taught introduction to geostatistics].

Journel A.G., Huijbregts C.J. (1978). *Mining Geostatistics*. 600pp. New York: Academic Press. [This is a seminal work in geostatistics, reflecting its mining origin].

Rossi R.E., Mulla D.J., Journel A.G., Franz E.H. (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecology Monographs* **62**, 277-314. [This is one of the few review papers dealing with the application of geostatistics to ecological data].

Yarus, J.M., Chambers, R.L. (editors) (1994). *Stochastic Modeling and Geostatistics. Principles, Methods, and Case Studies*, Vol. 3, 379pp. Tulsa, OK, USA: The American Association of Petroleum Geologists (AAPG). [This book presents an overview of geostatistics as applied to the petroleum industry].

Biographical Sketch

Pierre Goovaerts is currently Chief Scientist for the Research and Development company Biomedware, Inc and is the president of the consulting company PGeostat, LLC. He obtained his Ph.D. in agriculture engineering from the Catholic University of Louvain-la-Neuve, Belgium. His general expertise is in the sampling and geostatistical treatment (semivariogram analysis, spatial prediction and stochastic simulation) of data, with an emphasis on the assessment of uncertainty attached to prediction and its use in decision-making process, such as identification of locations for additional sampling or delineation of contaminated areas.

He is the author of "Geostatistics for Natural Resources Evaluation", which has become a reference textbook in the field. Pierre Goovaerts is the author or co-author of more than 50 publications in refereed journals, and is a reviewer for more than 40 journals. He was the recipient of the 1999 Andrei Borisovich Vistelius Research Award, given by the International Association of Mathematical Geology for an original and outstanding contribution by a young scientist to the application of mathematics and informatics to the earth sciences.