Integrating Geostatistical Tools in Geographical Information Systems

CARLOS ALBERTO FELGUEIRAS[1]
ANTÔNIO MIGUEL VIEIRA MONTEIRO[1]
EDUARDO CELSO GERBI CAMARGO [1]
GILBERTO CÂMARA NETO[1]
SUZANA DRUCK FUKS[2]

[1]INPE—Instituto Nacional de Pesquisas Espaciais, Caixa Postal 515, 12201 São José dos Campos, SP, Brasil
{carlos,miguel,eduardo,gilberto}@dpi.inpe.br
[2]EMBRAPA – Empresa Brasileira de Pesquisa Agropecuária – CNPS – Rua Jardim Botanico, 1024, Jardim Botânico. CEP: 22460-000 , Rio de Janeiro, RJ, Brasil
suzana@cnps.embrapa.br

Abstract. Increasingly, geostatistical procedures have been used for analysis and for attribute modeling of spatial data. These procedures are applied to an attribute sample set, represented as points, and allow the estimation of numerical and categorical attribute values in locations not sampled. Furthermore, the geostatistical estimation processes permit the inference of uncertainties related to the estimated values. In Geographical Information Systems – GIS – environment, the attribute models and their uncertainties can be used as input data for computational modeling of environmental processes. This paper addresses the main aspects involved in the incorporation of geostatistical procedures in a GIS environment. The requirements for a geostatistical module, implemented as part of the GIS, are described and analyzed. Also, the extension of a model description language to consider the uncertainties of the data models and to propagate them to the computational modeling results is presented and discussed.

1 Introduction

Geostatistical analysis can be used for spatial modeling in a diversity of geographic applications. In particular, geostatistical methods and tools are of great utility when searching for surface models for fitting a data set sampled as points. By exploring the spatial structure of the data being modeled, the geostatistics yields procedures for surface estimation that contemplates the spatial variability inherent to, for instance, a large set of environmental data. At the same time, these developments can generate a spatial estimate of the uncertainty of the information conveyed by those surfaces representing the attribute information. When developing a geographical application on a GIS platform that has no geostatistical tools, these analysis are performed by exchanging the data to be modeled between the GIS and the geostatistical package where the analysis will take place. These back and forth procedure can be very cumbersome, and in fact, can take a lot of time out that should be used in the task of being thinking on the modeling for that data. These are some reasons for the current trend to incorporate geostatistical facilities in the most popular GIS packages. Initial efforts have been done in the direction to integrate a geostatistical module in the GIS software. Unfortunately, this is not enough to guarantee that the geostatistical power is fully incorporated in geographic analysis using GIS packages. Geostatistical procedures must be implemented as operators to be used in spatial modeling languages. Also, the uncertainty related to the modeled data should be considered and propagated to the results of a spatial model that contains non-deterministic parameters and data. This paper addresses the central aspects involved in the incorporation of geostatistical facilities in a GIS modeling environment. The requirements for a geostatistical module to be aggregated into a GIS package and the issues related to geostatistical and error propagation operators in GIS modeling languages are presented and discussed.

2 Geostatistics and GIS

Definitions

Geostatistics is concerned with “the study of phenomena that fluctuate in the space and/or time”, Olea (1991). Geostatistics offers a way of describing the spatial continuity that is an essential feature of many natural phenomena and provides adaptations of classical regression techniques to take advantage of this continuity, Isaaks and Srivastava (1989). Geostatistics offers a collection of deterministic and statistical tools aimed at understanding and modeling spatial variability, Deutsch and Journel (1998).

The above definitions show that the main purpose of the geostatistics is to represent the spatial variations of some attribute occurring within a region **A** of the earth surface, $\mathbf{A} \subset \Re^2$. Furthermore, this representation allows inferences for attribute values in any location **u** within **A**, $\mathbf{u} \in \mathbf{A}$. For the geostatistics, the attribute is characterized as a random variable (RV) *Z* whose probability distribution function models the uncertainty about realizations *z* of *Z*. It is assumed the hypothesis of second order stationarity for the random function (RF) *F* defined by the RV's representing the attribute within **A**. The second order stationarity hypothesis settles that the expected (mean) z value is constant, i.e., $E\{Z(\mathbf{u})\} = m$, and the autocovariance depends only on the vectorial distance **h** between pair of z-data values within **A**, i. e., $C(\mathbf{h}) = E\{Z(\mathbf{u}+\mathbf{h})Z(\mathbf{u})\} - [E\{Z(\mathbf{u})\}]^2$, $\forall\ \mathbf{u}, \mathbf{u}+\mathbf{h} \in \mathbf{A}$.

The geostatistical spatial data modeling begins with the study of the variability of a sample set, observed as points, that is considered representative of the attribute variation. A theoretical semivariogram is fitted for the sample set and is used to determine weights for the sample neighborhoods considered in the inference process. Therefore, the geostatistical inference procedures, the kriging and the stochastic simulation, use the sample set and a correlation model to estimate attribute values in spatial locations different from the samples locations.

The *kriging procedure* aims to estimate z values based on a weighted mean approach of the z-sampled values of a local neighborhood. The kriging weights are determined from the basic hypothesis of minimum variance of the error estimation and make use of the theoretical semivariogram in order to calculate the covariance between two locations within **A**.

Stochastic simulation is the process of drawing alternative, equally probable, joint realizations of the component RVs from a RF model, Isaaks and Srivastava (1989). The realizations represent L possible fields of the spatial distribution for the attribute values over **A**. When conditioned to a set of attribute samples this process is named *conditional stochastic simulation*.

There are many definitions for GIS, depending on the type of user and application domain, Maguirre at all (1991). Burrough and McDonnell (1998) point out three classes of definitions, toolbox-based definitions, database-definitions and organization-based definitions. The first one defines the GIS as "a powerful set of tools for collecting, storing, retrieving at will, transforming and displaying spatial data from the real world for a particular set of purposes ". This is most applicable definition for the discussions presented in this text. The database approach defines GIS as a non-conventional geographic database that "supports management of spatial data", De Oliveira at all (1997). In an organization based point of view, Cowen (1988) defines GIS as "a decision support system involving the integration of spatially referenced data in a problem solving environment".

Geostatistics and GIS, what do they have in common?

Geostatistics and GIS have at least two common characteristics:

- *They handle spatial data*. While geostatistics procedures allow spatial data modeling, with uncertainty, GIS yields tools for storing, manipulating and displaying spatial data models and products.
- *They are multidisciplinary* with interest in common disciplines related to the environment modeling. Soil sciences, agriculture, geology, hydrology and environmental sciences are typical examples of disciplines that make use of both systems.

3 Geostatistical procedures in a GIS environment

Geostatistical procedures are important in a GIS environment for the following main reasons:

- They add, to the SIG tools, new methodologies for interpolating attributes of spatial data. The interpolation procedures, in a GIS environment, are mostly used to create attributes fields represented as a regular rectangular grid. The geostatistical interpolation methodologies are different from the deterministic ones because they are based in the study and modeling of the attribute variability of a spatial data;

- When using geostatistical procedures based on indicator kriging, see Deutsch and Journel (1998) and Felgueiras at all (1999), it is possible to apply an unified methodology for estimating numerical and categorical attributes of spatial data;

- They allow assessing the uncertainty related to the attribute values estimated by these procedures. These uncertainties are used to qualify the spatial information in each location of the space considered.

- The equally probable numerical and categorical fields supplied by the conditional stochastic simulation can be used as inputs for computational modeling and error propagation modeling. Heuvelink (1998) describes the Monte Carlo method that can be used as an error propagation methodology in computational modeling for environment processes.

4 Geostatistical modules for GIS

Requirements Specification

A complete geostatistical module integrated in a GIS environment should have, at least, the following procedures:

- Procedures for *exploratory analysis* of the sampled data. These procedures have to be able to analyze the input data and to inform the user about their uni-variate and multi-variate general statistics. Typical functions of the exploratory analysis are: evaluation and presentation of simple descriptive statistics as, maximum, minimum, mean, variance, standard deviation, kurtosis coefficient, correlation coefficients, quantis and others values and; display graphically formatted information such as, histogram, probability and cumulative distribution functions, normal probability curves and scatterplot diagrams;
- Procedures for *data transformation*. These are procedures for editing and transforming the input data in order to improve the quality or the usability of the data. In this group the geostatistical module should supply: functions to filter erroneous and undesirable data, as outliers for example; functions to eliminate clusters or redundant data and; functions to transform and back-transform the original distribution of the data, such as, normal, logarithmic and indicator transformations
- Procedures for *modeling the spatial variability* of the attribute. These procedures use the input sampled and transformed, when it is the case, data in order to define its spatial autocorrelation model or its crosscorrelation model with another spatial attribute. Typical examples in this group of procedures are: stationarity analysis by moving windows; surface variography analysis for detecting anisotropy axes; fractal dimension analysis and; exploratory variography analysis to define an experimental and theoretical variogram model.
- Procedures for *attribute values inference*. These procedures use the input sampled, or transformed when it is the case, data and its spatial correlation model to infer attribute values in spatial positions not sampled. The geostatistical module should provide: linear interpolation procedures, such as, simple, ordinary and universal kriging and cokriging; nonlinear interpolation procedures as indicator kriging and cokriging and; linear and nonlinear stochastic simulation procedures.
- Procedures for *validations*. These procedures allow one to verify the quality of the variogram model used and the quality of the attribute value inferences. Cross validation procedures and functions for error and residues analysis are commonly implemented in this part of the geostatistical module.

Current Status

Because of the reasons presented in the beginning of section 3, recently it can be detected a trend to incorporate geostatistical modules in the GIS. Follows a brief description of the geostatistical module capabilities supplied for some GIS found in the market:

The geostatistical module implemented in the SPRING GIS, Camargo (1997) and DPI/INPE (1999), version 3.3, allows one to perform the most common exploratory analysis in the input data, to model the spatial variability of a spatial attribute via an interactive function, to create fields in regular rectangular representation using simple and ordinary kriging and performing cross validations for the input data in order to verify the quality of the modeling parameters assumed. The module functions are based in the GSLIB package, Deutsch and Journel (1998).

The IDRISI32, a version of the IDRISI GIS (http://www.idrisi.com/03prod/Idrisi.htm) for 32 bits, offers facilities for exploratory analysis of the sampled data, for spatial variation analysis that include two interfaces, the Spatial Dependence Modeller and the Curve Fitting, and for attribute inferences by way of kriging and simulation procedures. The implementation of these facilities was based in the gstat (http://www.geog.uu.nl/gstat/) package, a program for modeling, interpolation and simulation of random variables in 1, 2 and 3 dimensions.

The PCRaster geostatistical module (http://curie.ei.jrc.it/software/pcraster.htm) contains functions for performing exploratory analysis, including variogram modeling, for interpolations based on kriging methodology and simulation. This system has a language for dynamic modeling that allows the interactive construction of space-temporal environment models. The geostatistical modeling uses the functions of the Gstat 2.0.g (http://curie.ei.jrc.it/software/Gstat.htm) that is strongly coupled to the system.

The version 8 of the ArcInfo GIS is including an extension called Geostatistical Analyst Designed for Advanced Spatial Modeling" (http://www.esri.com/news/arcnews/spring99articles/07-geostatistical.html). As advertised in this URL address "The Geostatistical Analyst--an extension to ArcInfo Version 8--is aimed at an emerging advanced spatial modeling audience. These tools were developed

specifically for surface generation using geostatistical tools and analyzing the error of the resulting estimation (surface). Numerous kriging and predictive tools are the center of the extension, but it is the close integration with GIS functionality and the easy-to-use and rich user interface that distinguishes it from any other geostatistical offering". Also, from ESRI's GIS tecnology efforts, there is an extension named "Splus for Arcview & S+Spatial Stats 1.0" (http://www.mathsoft.com/splus/splsprod/arcview.htm) based in the geostatistical packages SPLUS and S+Spatial Stats (http://www.mathsoft.com/splus/) in the ArcView system (http://www.esri.com/software/arcview/index.html).

5 Spatial Modeling with uncertainty treatment

The efforts spent for integrating geostatistical modules in a GIS environment are important and imperative. Unfortunately, this is not enough to guarantee that the geostatistical power is fully incorporated in geographic analysis using GIS. Spatial analysis for computational modeling can be carried out, in a GIS environment, using spatial modeling languages, also named model description language (MDL).

In a MDL the user must be able to *define and initialize* different *type of variables* that will represent parameters and spatial attributes involved in the computational modeling. Furthermore, the language must supply a *set of operators* that will be used for transforming, analyzing and integrating spatial data according to a chosen spatial model. Nowadays the spatial modeling languages supplied by the GIS are, in general, deterministic, i. e., they do not allow the user define stochastic types and operators.

In order to work with uncertainty in spatial modeling the current MDLs have be extended to incorporate:

- stochastic types to define stochastic variables;
- geostatistical operators to implement geostatistical procedures and;
- error propagation operators to calculate uncertainties of the modeling results.

Stochastic types are necessary to allow the definition of stochastic variables that will be initialized with their deterministic and statistic members. Geostatistical operators are required in order to apply geostatistical estimation procedures, various kriging and simulation options, over an input attribute sampled as points. These operators will require the definition of some geostatistical parameters, such as, number of sample neighbors, variogram model, etc., to perform the estimations. Error propagation operators must be supplied for different data types and computational models.

Heuvelink (1998) describes an error propagation software tool, named **ADAM,** which allows the definition of a quantitative computational model via an MDL that supports the items described above. The ADAM tool functions as a compiler, translating a user's propagation problem into standard GIS operations ready be executed in a GIS environment. Unfortunately, the ADAM tool is limited and must be extended to consider categorical data and categorical error propagation functions.

6 Conclusions

The inclusion of a geostatistical module in a GIS environment is a reality. As presented above many commercial GIS are announcing efforts in this direction. In general, all the geostatistical modules were implemented using procedures already developed for geostatistical packages that runs independently of a GIS. The innovation is their integration in the GIS that allows the user work in a unique environment to develop GIS applications requiring geostatistical analysis.

Although all these modules have been implemented with almost all the requirements described in the section 4, it can not be found explicit implementations for non-linear procedures based in the indicator kriging. It seems that the current attentions are mostly concentrated in supply geostatistical analysis rather than work on uncertainties and quality of spatial data. The indicator kriging approach can be used to obtain uncertainties that really represent the attribute variation independently of the pattern of the accumulated probability distribution function of the random variable or field used. Furthermore, the indicator approach can be used with categorical attributes what is not possible with procedures based in the linear kriging approach.

Finally, with the integration of geostatistical capabilities in a GIS environment it can be expected that GIS users should be able:

- to analyze and interpolate attributes of spatial data using geostatistical procedures;
- to obtain uncertainties associated with the estimation process (the GIS database management must yield ways to store and retrieve the uncertainty information along with the data representation) and;
- to work with spatial modeling that accounts for the uncertainties of the input data model and that propagates these uncertainties to the GIS products.

References

[1] P. A. Burrough and R. A. McDonnell, *Principles of Geographical Information Systems*, Oxford University Press, 1998

[2] E. C. G. Camargo, *Desenvolvimento, implementação e teste de procedimentos geoestatísticos (krigeagem) no Sistema de Processamento de Informações Georeferenciadas (SPRING).* Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1997.

[3] D. J. Cowen, "GIS versus CAD versus DBMS: what are the differences?", *Photogrammetric Engineering and Remote Sensing*, 54, (1988), 1551-1554.

[4] J. L. De Oliveira, F. Pires and C. B. Medeiros, "An environment for modeling and design of geographic applications", *GeoInformatica*, 1, (1997), 29-58.

[5] C. V. Deutsch and A. G. Journel, *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press, 1998.

[6] C. A. Felgueiras, *Desenvolvimento de um Sistema de Modelagem Digital de Terreno para microcomputadores.* Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 1987.

[7] C. A. Felgueiras, A. M. V. Monteiro, S. D. Fuks and E. C. G. Camargo, "Inferências e Estimativas de Incertezas Utilizando Técnicas de Krigeagem Não Linear" [CD-ROM]. In: V *Congresso e Feira para Usuários de Geoprocessamento da América Latina*, 7, Salvador, 1999. Anais. Bahia, gisbrasil'99. Seção de Palestras Técnico-Científicas.

[8] G. B. M. Heuvelink, *Error Propagation in Environmental Modeling with GIS*, Bristol, Taylor and Francis Inc, 1998.

[9] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*, Oxford University Press, 1989.

[10] DPI/INPE.- (SPRING) Sistema de Processamento de Informações Georeferenciadas – Divisão de Processamento de Imagens (DPI) do Instituto Nacional de Pesquisas Espaciais (INPE).. http://www.dpi.inpe.br/spring/, 1999.

[11] D. J. Maguire, M. F. Goodchild and D. W. Rhind, *Geographical Information Systems: Principles and Applications* – Volume I. John Wiley and Sons, 1991.

[12] R. A. Olea, *Geostatistical Glossary and Multilingual Dictionary*. International Association for Mathematical Geology Studies in Mathematical Geology. Oxford University Press, New York. No. 3, 1991. 177p.