Improving Accuracy of Categorical Attribute Modeling with Indicator Simulation and Soft Information

C. A. Felgueiras[1], A. M. V. Monteiro[2], E. C. G. Camargo[3], J. O. Ortiz[4]

[1,2,3,4]INPE - Brazilian National Institute for Space Researches, P.O. Box 515, São José dos Campos, SP, Brazil
Telephone: +55 12 3208-6444
Email: {[1]carlos, [2]miguel, [3]eduardo,[4]jussara}@dpi.inpe.br

Abstract

The objective of this work is to apply an indicator geostatistical simulation approach to improve the accuracy of spatial modeling of categorical attributes using hard and soft information. Sample points of a categorical attribute are considered as the hard, or primary, information while a categorical map is used for determine the soft, or the secondary, information. The soft information is incorporated in the indicator simulation procedure as prior mean values, taken from a probability distribution function, related to the hard data. The prior mean values are then updated via indicator simulation to account for the hard data available in their neighborhoods. To illustrate the methodology a case study is presented with samples of soil texture classes, as the hard data, and with classes of a soil map defining the soft information. These data are gathered from an experimental farm of agriculture researches. The results show that the use of soft information, along with the hard data, improve the accuracy of the final products and show regions with higher uncertainties that are candidates to be sampled or resampled in the future.

Keywords: Geostatistics, Spatial Modelling of Categorical Attributes, Indicator Simulations, Uncertainty Assessments, Hard and Soft data.

1. Introduction

Categorical attributes can be modeled, as grid representations, from a set of their samples, distributed in a spatial region of interest, using geostatistical approaches (Delbari et al., 2011, Isaaks and Srivastava, 1989, Wasiullah and A.U. Bhatti, 2005,). Geostatistical indicator procedures, as the indicator kriging and the indicator simulation, are widely used mainly because they are able to estimate local or spatial uncertainty models, i. e., the joint conditional distribution functions, of continuous (ccdf) or categorical attributes (cpdf), at any unknown spatial location **u** (Juanga at al., 2004, Jaeri et al. 2013). The uncertainty models are conditioned to a set of sample points of an attribute of interest and optionally to a set of sample points of secondary information correlated with the attribute.

From the uncertainty models it is possible to derive attribute predictions and realizations along with uncertainty metrics as, for example, confidence intervals of the distributions. The final quality of the uncertainty models is greatly influenced by the number and the spatial distribution of the sample set. When the distribution of the samples is sparse, i. e., the number of samples is too small for the spatial region considered, the quality of the predictions and of the simulations tends to be low.

The geostatistical indicator approaches allow, also, to improve the uncertainty modelling of a spatial attribute when a secondary information, correlated with the primary one, is incorporated in the uncertainty estimation process. The secondary data is generally easier to obtain, sometimes at no cost on the internet, and densely distributed.

The objective of this work is to apply an indicator geostatistical simulation approach to improve the accuracy of the spatial modeling of categorical attributes using hard and soft information. Sequential Indicator Simulation (SIS) is a widely used technique for modelling uncertainties of continuous and categorical variables. (Goovaerts, 1997, Felgueiras, 2000, Deutsch, 2006). The SIS and the SIS with prior means, GSLIB (Deutsch and Journel, 1998) functions known as sisim and sisim_lm respectively, were used in this work. Sample points of a categorical attribute were taken as the hard, or primary, information while a categorical map is considered as the soft, or the secondary, information. The soft information is incorporated in the indicator simulation procedure as prior mean values, taken from a probability distribution function, related to the hard data. The prior mean values are then updated via indicator simulation to account for the hard data available in their neighborhoods.

To illustrate the applied methodology, a case study is presented with samples of soil texture classes, as hard data, and a soil map is used to determine the soft data. Four classes of soil texture were considered: sandy, medium clay, clay an too clayed. The classes of the soil map of the region of interest were taken in order to get pdf prior mean values of texture classes for each soil class. The soil texture were modeled using the hard data only and using the hard and the soft information. The resulting maps were presented, compared, and analyzed, mainly considering the improvement of the precision of the soil texture modeling. The results show that the use of soft information, along with the hard data, improve the quality of the final products and show regions with higher uncertainties that are candidates to be sampled or resampled in the future.

This article is organized as follows: Section 1 presents an introduction; section 2 refers to the main concepts of this work; section 3 describes the applied methodology; section 4 reports a case study in an experimental farm in the region of São Carlos city, in São Paulo, Brazil; section 5 presents results and discussions; and section 6 addresses the final conclusions and new ideas for futere researches related to the accuracy improvement of spatial data modeling.

2. Concepts

The indicator approaches allow for modeling the joint conditional distribution functions, of continuous (*ccdf*) or categorical attributes (*cpdf*), at any unknown spatial location **u** considering an available punctual sample set. The *Simulation* process consists of drawing realizations from the joint distribution functions.

The *Sequential Simulation* process works with the *ccdfs* and a random number generator. For categorical variables, the *ccdfs* can be built from the *cpdfs* considering one order among the classes. N realizations of each, continuous or categorical, Random Variable *Z* can be drawn from a *ccdf* repeating n times the following steps: generating a random *cp* number between 0 and 1 (*cp* - cumulative probability value) and mapping the *cp* value to the z_{cp} attribute value using the given *ccdf*.

The *Sequential Indicator Simulation* takes the following steps (Govaerts, 1997):

- Draw a value $z_1^{(l)}$ from the univariated ccdf of Z_1, Prob$\{Z_1 \leq z_1|(n)\}$, conditioned to the (*n*) original samples.
- Update the original sample data set (*n*) to a new information set (*n*+1) :

$(n+1)=(n) \cup \{Z_1 = z_1^{(l)}\};$

- Draw a new value $z_2^{(l)}$ from the univariated ccdf of Z_2, Prob$\{Z_2 \leq z_2 | (n+1)\}$, conditioned to the information set $(n+1)$:
- Update the information set $(n+1)$ to a new information set $(n+2)$:

$(n+2)=(n+1) \cup \{Z_2 = z_2^{(l)}\};$

- Sequentially consider all the J Random Variables Z_j*'s.*
- Repeat the above sequence for a new l realization (up till L Random Fields)

The *Sequential Indicator Simulation* with *Prior Mean* allows incorporating prior *pdf/cdf* information obtained from a secondary (soft) data. The prior *cdfs/pdfs* are updated via indicator kriging (Bayesian framework), i.e., each prior local are updated to account for the hard data available in its neighborhood (Deutsch and Journel, 1997).

The realizations at each location **u** are used to create prediction maps and uncertainty maps. From the realization values of continuous variables one can assess to the mean, the standard deviation or any quantile value to build a prediction, or estimated, map. Confidence intervals, based on the standard deviation or quantile values, are used to create the uncertainty maps. From the realization values of categorical variables one can assess to the most frequent class, higher probability, to built prediction and uncertainty maps. In this case the prediction map contains the classes with higher probabilities, P_{max}, while the uncertainty map contains the 1-P_{max} values. Other metrics of uncertainty can be used, as the Shannon Entropy, that take into account all the probability values of a cpdf (Felgueiras, 2000).

3. Methodology

Given a spatial region of interest, the methodology applied has the following steps:

1. For a sample set of points of a categorical attribute, the hard data, evaluate the variograms for residuals of the indicator sample sets related the attribute classes;
2. Determine the local prior *pdf* values for each output grid spatial location using a secondary information, the soft data;
3. Fill the parameter file of the *SIS*, *sisim* and *sisim_lm*, GSLIB functions;
4. Run the *SIS* functions to obtain grids with realizations of the hard information;
5. Creating maps of predicted, or estimated, classes and uncertainties, 1-P_{max}, values from the output file of the *SIS* functions;
6. The final resulting maps of predictions and uncertainties are analyzed and compared.

4. A Case Study

In order to illustrate the methodology of this work, it was used as hard information a set of points of soil texture data sampled in the region of an experimental farm known as Canchim. The study region is located in the city of São Carlos, SP, Brazil, and cover an area of 2660 ha between the north-south coordinates from s 21°55'00" to s 21°59'00" and the east-west coordinates from w 47°48'00" to w 41°52'00".

The hard data set consists of 86 samples of soil texture information each classified as one of the following four classes: sandy, medium clay, clay or too clayed. Figure 1 (left map) illustrates the borders of the Canchim farm along with location and the classification of the soil texture sample set. This map was obtained with a nearest neighbor estimation procedure showing the regions of influence of each class.

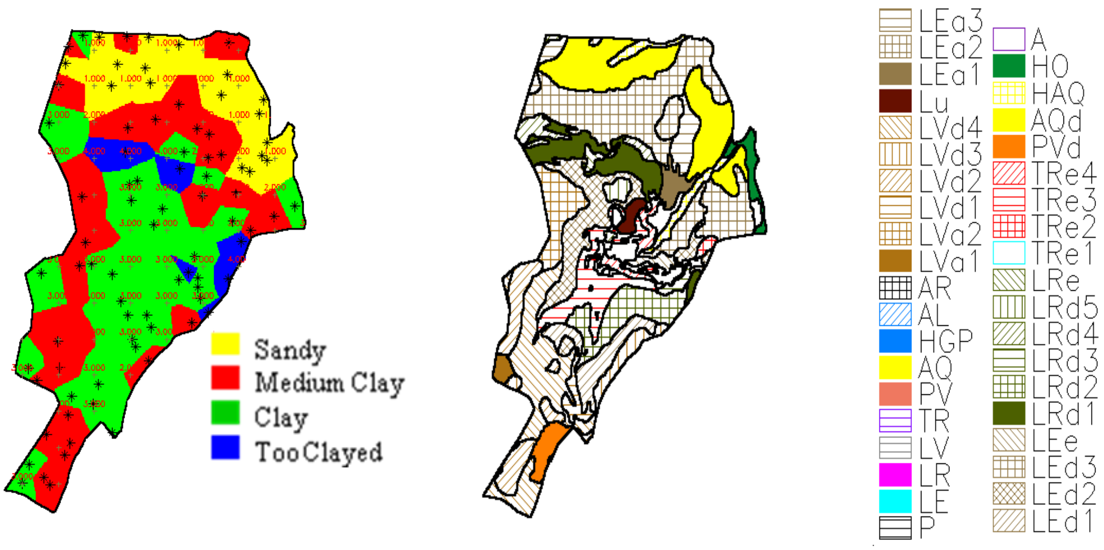

Figure 1. Distribution of the soil texture sample points (left map) and map of soil classes of the Canchim region (right map)

It is also considered a soil map of the figure 1 (right map) in order to assess the secondary (soft) information, the probabilities a priori of the texture class for each soil class. These probabilities a priori are presented in table 1.

Soil Class	Sandy	Mediu Clay	Clay	Too Clayed
LVA1	0	0	1	0
LVA2	0	1	0	0
LVD1	0	0	1	0
LVD2	0	0	1	0
LVD3	0	1	0	0
LVD4	0	1	0	0
LU	0	0	1	0
LEA1	0	0.4	0.6	0
LEA2	0	1	0	0
LEA3	0	1	0	0
LED1	0	0	1	0
LED2	0	0	1	0
LED3	0	1	0	0
LEe	0	0	1	0
LRD1	0	0	0	1
LRD2	0	0	0.8	0.2
LRD3	0	0	0.7	0.3
LRD4	0	0	1	0
LRD5	0	0	1	0
LRe	0	0	0.4	0.6
TRe1	0	0	0.4	0.6

TRe2	0	0	0	1
TRe3	0	0	1	0
TRe4	0	0	0.7	0.3
PVd	0	1	0	0
AQd	1	0	0	0
Haq	0.8	0	0.2	0
Ho	0	0	1	0
A	0	0	1	0

Table 1. Probabilities a priori of the texture classes for each soil class

5. Results and Discussions

Figure 2 shows the map of predicted soil texture classes (left) and respectively uncertainty map (right) obtained from the realizations of the *sisim* approach. The estimations were assessed from the higher probabilities of the *cpdfs* estimated at each spatial location. The uncertainties were defined as 1- the higher probability of the *cpdf* and, as expected for environmental attributes, are higher in the borders, the transitions areas, of soil texture class regions. Consequently the probability uncertainty values are lower in the middle of those regions.

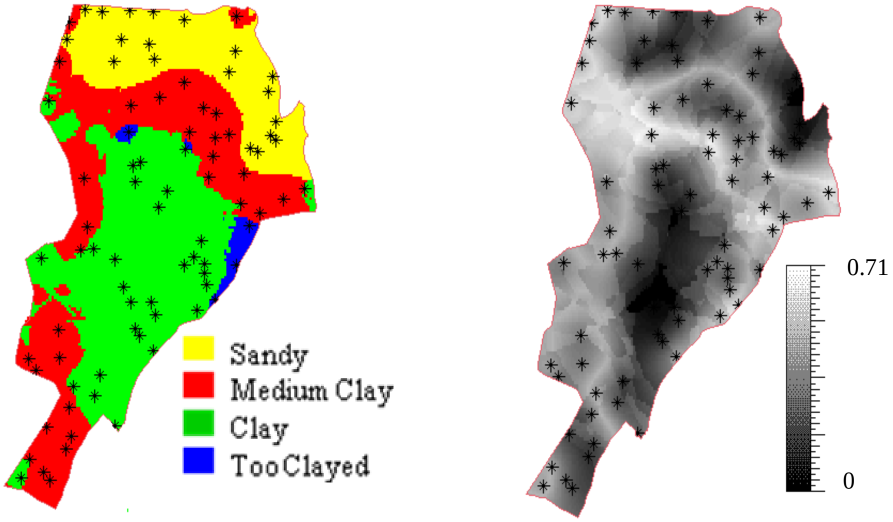

Figure 2. Map of predictions of texture classes (left) and map of uncertainties (right) estimated using the output of the sisim function

Figure 3 shows the map of predicted texture classes (left) and respectively uncertainties (right) obtained from the realizations of the *sisim_lm* approach. The borders of the soil classes is overlapped in these maps to aid the analyzes of the results. This final soil texture map presents a configuration more similar to that of the soil map. The uncertainty map presents lower global uncertainty probability indicating that the use of soft information can produce higher precision distributions. As to the previous case higher uncertainties appear in the borders of the classes. In addition it was found a region of high uncertainties, highlighted in blue, in the middle eastern of the map. This region contains samples of texture classes conflicting with the soil map information. Regions with these characteristics should be considered as candidates to be sampled or resampled in order to get more reliable results.

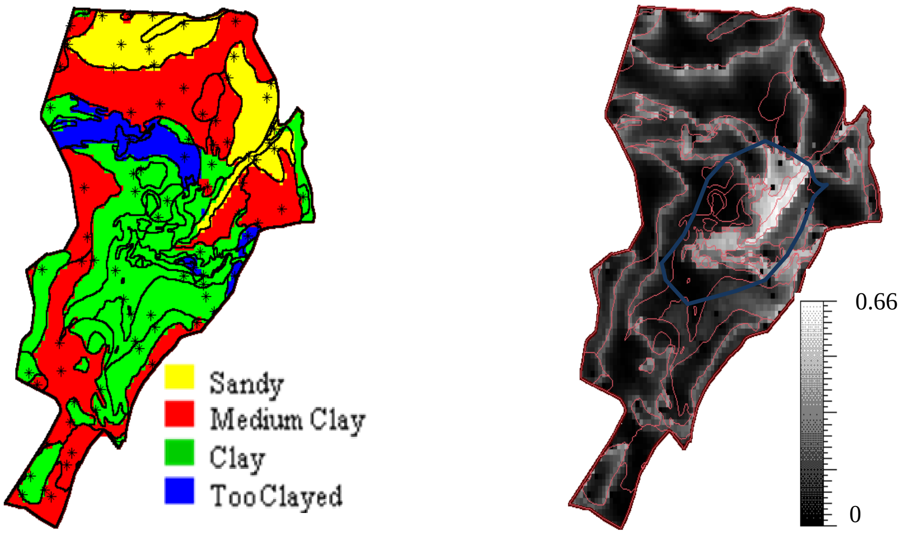

Figure 3. Map of predictions of texture classes (left) and map of uncertainties (right) estimated using the output of the sisim_lm function

6. Conclusions

Spatial modeling of categorical attributes can be accomplished from geostatistical indicator sequential simulation approaches using hard and also soft information when it is available. Secondary variables can be incorporated in the simulation to improve the accuracy of the predictions and of the uncertainty representation.

The uncertainties of the modeling are used to qualify the estimates and should be applied in decision making procedures for planning activities on environmental applications, for example. Moreover, regions where the uncertainties are higher must be considered candidates to be sampled or resampled in the future.

The set of realizations of the indicator simulations can be used as input for multivariable spatial modeling of categorical variables in Monte Carlo approaches, for example.

In the future we intend to explore similar methodology for spatial modeling of continuous attributes considering also secondary variables.

7. References

Delbari M, Afrasiab P, and Loiskandl W., 2011, Geostatistical Analysis of Soil Texture Fractions on the Field Scale. *Soil & Water Resources*, 6(4): 173–189.

Deutsch CV and Journel AG, 1998, *GSLIB: geostatistical software library and user's guide*. Oxford University Press, New York, USA.

Deutsch, C. V. (2006), "A sequential indicator simulation program for categorical variables with point and block data: BlockSIS", *Computer and Geoscience*, 32(10): 1669-1681

Felgueiras CA, 2000, *Modelagem ambiental com tratamento de incertezas em sistemas de informação geográfica: o paradigma geoestatístico por indicação*. 165p. PhD Thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, São Paulo, Brazil.

Goovaerts P, 1997, *Geostatistics for natural resources evaluation*. Oxford University Press, New York, USA

Goovaerts P, 2001,"Geostatistical modeling of uncertainty in soil science". *Geoderma* 103:3–26

Isaaks, EH and Srivastava RM, 1989, *An Introduction to Applied Geostatistics*. Oxford University Press, New York, USA.

Juanga K, Chenb Y and, Leeb D, 2004, " Using sequential indicator simulation to assess the uncertainty of delineating heavy-metal contaminated soils". *Environmental Pollution*, 127: 229–238

Wasiullah and A.U. Bhatti, 2005, Mapping of soil properties and nutrients using spatial variability and geostatistical techniques. *Soil and Environment*, 24(2): 88-97.

Zaeri K, Hazbavi S, Toomanian N, Zadeh JT, 2013, Creating surface soil texture map with indicator kriging technique: A case study of central Iran soils. *IJACS Journal International Journal of Agriculture and Crop Sciences, 6 (9), 518-521.*