SPATIAL PREDICTIONS OF CATEGORICAL ATTRIBUTES CONSTRAINED TO UNCERTAINTY ASSESSMENTS

Felgueiras, C. A.[1], Ortiz[1], J. O. and Camargo[1] E. C. G.

[1]Image Processing Division (DPI) of the Brazilian National Institute for Space Researches (INPE)

carlos@dpi.inpe.br

Abstract

This article explores the use of nonlinear geostatistical procedures, known as kriging and simulation indicator approaches, for spatial modeling of categorical attributes. The categorical information is initially represented by a set of sample points observed within a spatial region of interest. The original sample set is used to generate indicator fields take into account the classes of the categorical data. The indicator fields, or indicator samples, contain 0 and 1 attribute values according to the class they are representing. Empirical and theoretical semivariograms are built from the indicator samples to represent the spatial variation of each class in relation to the others. The geostatistical procedures, making use of the samples and the theoretical semivariograms, allow obtaining an approximation of the stochastic model, the conditioned probability distribution function (*cpdf*) of the categorical attribute at any desired spatial location. From any *cpdf* it is possible to assess optimal prediction, or estimate, and uncertainty values associated to the stochastic model. Optimal prediction as mean, median or any quantile values can be assessed. Uncertainty values are obtained by means of the maximum *cpdf* probability, Shannon entropy, or another criterion. The uncertainty values can be used to qualify the predictions and can also be considered to generate constrained spatial predictions, or constrained classifications, that are important in decision makings related to environmental planning activities, for example. The concepts here presented are applied and tested in a case study developed for a sample set of soil texture observed in an experimental farm in the region of São Carlos city in São Paulo State, Brazil. Four classes of soil texture are considered, sandy, medium clay, clay and too clay, in order to get the cpdf values. Some maps derived by constraints are presented and analyzed considering different probability values from the attribute stochastic models.

KEYWORDS: Spatial Analyzes, Indicator Geostatistics, Modeling of Categorical Attributes, Uncertainty Assessments, Constrained Classifications, Decisions Making in Environmental Planning.

INTRODUCTION

Computational spatial modeling for environmental phenomena has nowadays been developed using Geographical Information Systems (GISs). This task demands the integration of different environmental variables in a mathematical model and has the main objective of obtaining reports or maps that will be useful for decision makings applied to problems related to the earth surface. The spatial modeling issue involves the use of continuous and categorical variables which have to be well structured in order to facilitate their integration. In general, these variables are represented by sample sets organized as vectors, points or lines, or matrices. When a set of points are available, as samples of a variable, it is common to spatialize the samples in order to have information in any location inside a spatial region of interest. The variable spatialization is accomplished using interpolation procedures which can be deterministic or stochastic. Interpolations are very common when the values of a variable are numeric, i. e., when continuous numeric variables are handled. Nevertheless, it is difficult to find appropriated procedures to interpolate variables representing qualitative information, also called categorical or thematic variables.

Geostatistical approaches can be used to model numerical and thematic variables (Goovaerts, 1995; Delbari et al., 2011; Isaaks and Srivastava, 1989; Wasiullah and Bhatti, 2005). The geostatistics use stochastic methods, based on the inference of local or global probability distribution functions, *cdfs* or *pdfs,* of the variable. Geostatistical tools allow to perform exploratory analyses of the sample data and to infer prediction and simulation values of numerical and thematic information at no sampled spatial locations. These tools are conditioned to the sample sets and they allow to qualify the predictions and the simulations with uncertainty information associated to them. When the variable is continuous the predictions can be, for example, mean, median, or quantile values and the uncertainties are represented by confidence interval values based on variances, standard deviations or quantiles. When the variable is thematic the predictions and the uncertainties are assessed from the probability values of the *pdfs*. In this case its common to use the maximum probability value of the pdf to determine the prediction and the uncertainty values. Nonlinear approaches known as indicator geostatistics allow to spatialize environmental variables, numeric and thematic, from a set of sample points in a nonparametric method (Deutsch and Journel, 1998; Deutsch, 2006; Goovaerts, 1998; Felgueiras, 2000; Zaeri et al., 2013).

Considering the above context, the objective of this article is to explore the use of the indicator kriging and simulation approaches for spatial modeling of categorical attributes. This article is an extension of the Felgueiras et al., 2003. The categorical information is initially represented by a set of sample points observed within a spatial region of interest. The original sample set is used to generate indicator fields take into account the classes of the categorical data. The indicator fields, or indicator samples, contain 0 and 1 attribute values according to the class they are representing. Empirical and theoretical semivariograms are built from the indicator samples to represent the spatial variation of each class in relation to the others. The geostatistical procedures, making use of the samples and the theoretical semivariograms, allow obtaining an approximation of the stochastic model, i. e., the conditioned probability distribution function (*cpdf*) of the categorical attribute at any chosen spatial location. From any *cpdf* it is possible to assess optimal prediction, or estimation, and uncertainty values associated to the stochastic model. Optimal prediction as mean, median or any quantile values can be assessed. Uncertainty values are obtained by means of the maximum *cpdf* probability, Shannon entropy (Shannon and Weaver, 1949), or another criterion. The uncertainty values can be used to qualify the predictions and can also be considered to generate constraint spatial predictions, or classifications, that are important in decision makings evolved in environmental planning activities for example. The concepts here presented are applied and tested in a case study developed for a sample set of soil texture observed in an experimental farm in the region of São Carlos city in São Paulo State, Brazil. Four classes of soil texture are considered, sandy, medium clay, clay and too clay, in order to get the *cpdf* values.

Constrained prediction maps are obtained when the uncertainty values are taken into account as levels of quality accepted for a specific application. Many derived constrained maps are presented and analyzed considering different probability values of uncertainties.

The results of this work show that environmental planners can get different possible classified scenarios, based on different uncertainty levels, which can be evaluated to support decisions on applications where the spatial uncertainty is considered important.

MATERIAL AND METHODS

Concepts

The geostatistical indicator approaches allow for modeling the joint conditional distribution functions, of continuous (*ccdf*) or categorical (*cpdf*) random variables, at any unknown spatial location **u,** considering an available set of sample points. The *Kriging* and *Simulation* processes work on, respectively, assessing predictions and drawing realizations from the joint distribution functions.

For categorical variables the *cpdfs* are built from estimations on indicator fields obtained by indicator transformations applied to the original sample set $S(\mathbf{u})$ considering K classes. Instead of the variable $S(\mathbf{u})$, consider its binary indicator transform $I(\mathbf{u};z_k)$ as defined by the relation of equation 1:

$$I(\mathbf{u};s_k) = \begin{cases} 1, & \text{if} \quad S(\mathbf{u}) = s_k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Kriging of the indicator random variable $I(\mathbf{u};z)$ provides an estimate that is also the best least square estimate of the conditional expectation of $I(\mathbf{u};z)$. Now the conditional expectation of $I(\mathbf{u};z)$ is equal to the local *pdf* of $Z(\mathbf{u})$ as presented in equation 2:

$$\begin{aligned} \mathrm{E}\{I(\mathbf{u};s_k) \mid (n)\} &= 1 \cdot \mathrm{Prob}\{I(\mathbf{u};s_k) = 1 \mid (n)\} + 0 \cdot \mathrm{Prob}\{I(\mathbf{u};s_k) = 0 \mid (n)\} \\ &= 1 \cdot \mathrm{Prob}\{I(\mathbf{u};s_k) = 1 \mid (n)\} = \mathrm{p}(\mathbf{u};s_k \mid (n)) \end{aligned} \tag{2}$$

In order to perform the above estimations using kriging procedures it is necessary to model indicator semivariograms that represent the spatial variability, or spatial dependence, of the indicator random variables.

The *cpdfs*, obtained from the indicator fields, their semivariograms, and using the property of the equation 2 are used directly by the *Indicator Kriging* procedure as it will be explained below.

Moreover, any *ccdf* can be built from the estimated *cpdf* of a categorical variable considering an order among its classes. The *Sequential Indicator Simulation* process works with these *ccdfs* and a random number generator. N realizations of each, continuous or categorical, random variable *Z* can be drawn from any *ccdf*. This process comprises draw randomly N decimal numbers, each between 0 and 1, and mapping these numbers to N classes according the *ccdf* function. The realization values allow the reproduction of the spatial *cpdfs* of a categorical random variable at any spatial location **u**.

The *cpdfs* are then used to assess to the most frequent class, mode or higher probability, in order to produce prediction and uncertainty maps. In this case the prediction map may be assigned with the classes with higher probabilities, P_{max}, among of them. The uncertainty map may be assigned with the 1-P_{max} values. Other metrics of uncertainty can be used, as the Shannon Entropy that takes into account all the probability values of a *cpdf* (Shannon, 1949, Felgueiras, 2000).

Using the typical geostatistical procedures, the prediction maps are constructed without probability restrictions. This means that even the maximum probability of a *cpdf* is lower than a minimum value, the spatial location is assigned to the class of this value. A constrained prediction map for a categorical variable can be produced considering a minimum probability, or a maximum uncertainty, value defined by application demand. In this way many, different scenarios for decision making purposes can be built according the maximum uncertainty value considered by the application planner. In the constrained predictions maps the classes with probabilities higher than a threshold are maintained, otherwise the classes are changed to a user defined background class.

Material

For this work it was used as primary information a punctual sample set of a categorical attribute. The samples were imported to a geographic database organized in the GIS software known as Sistema de Processamento de Informações Georeferenciadas, SPRING, (Câmara et al., 1996). The SPRING was the main tool to perform some exploratory analyses in the spatial data, as the assessment of the indicator semivariograms, and to display contents of the derived database infolayers. From the geostatistical library, named GSLIB (Deutsch and Journel, 1998), it was used the *sisim* procedure for performing the sequential indicator simulations.

Methodology

The methodology of this work follows the sequence:

1. Exploratory analyzes of the categorical sample set;
2. Generation of indicator sample sets using the considered set of categorical classes;
3. Assessment of experimental and theoretical semivariograms for the indicator sample sets;
4. Running of indicator simulation program, using the original samples and the semivariograms, to get the prediction and uncertainty maps;
5. Creating constrained prediction maps using the uncertainty models.

Case Study

In order to illustrate the methodology of this work, it was used a set of points of soil texture data sampled in the region of an experimental farm known as Canchim. The studied 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^{o}54’46’’ to s 21^{o}59’31’’ and the east-west coordinates from w 47^{o}51’46’’ to w 47^{o}48’18’’.

The sample data set consists of 84 samples of soil texture information each classified as one of the following four classes: sandy, medium clayey, clayey or too clayey. Figure 1 illustrates the borders of the Canchim farm along with location and the classification of the soil texture sample set. This categorical map of figure 1 was obtained with a nearest neighbor interpolation procedure showing the regions of influence of each spatial sample along with its class.

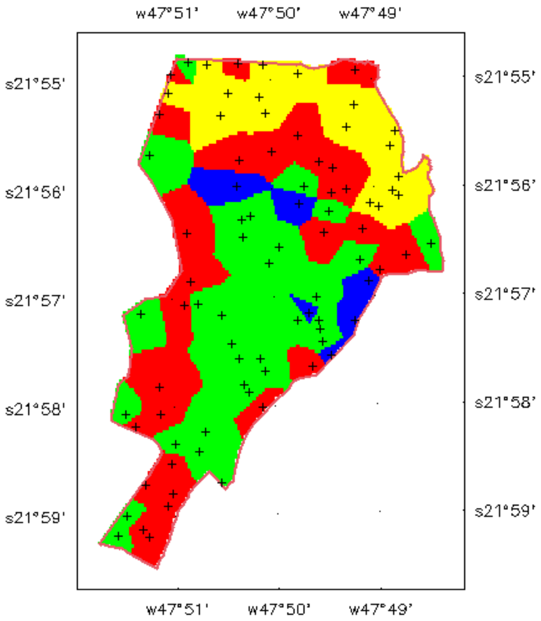


Figure 1. Spatial localization of the Canchim farm, the categorical sample set distribution and a classified map showing the region of influence of each location and its class.

RESULTS AND DISCUSSIONS

The spatial dependences analyses are represented by the indicator semivariograms generated from the indicator sample set defined by each texture class. It was fitted four semivariograms representing the four soil texture classes considered. The spatial dependence analyses are based on the indicator sample set of the soil texture classes.

The indicator semivariogram parameters and the global probabilities of each texture class are reported in the table 1.

Table 1: Parameters of the indicator semivariograms and global probability of each class

Texture Class	Nugget Effect	Contribution	Range	Global Probability
Sandy	0.00	0.14	1915	0.20
Medium Clayey	0.00	0.22	902	0.35
Clayey	0.01	0.20	1059	0.38
Too Clayey	0.03	0.05	695	0.07

The spatialization of the soil texture classes in the Canchim region was accomplished by using the Sequential Indicator Simulation, *sisim*, function available in the geostatistical library package GSLIB. 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 probabilitie P_{max}, the mode, of the *cpdfs* estimated at each spatial location.

A qualitative, visual, comparison between this map of predictions and the map of nearest neighbours' interpolation, map of figure 1, shows that the both maps globally agree with the spatial distribution of the texture sample set. The differences appear mainly in regions of class transitions of the predicted map obtained from the geostatistical simulated values.

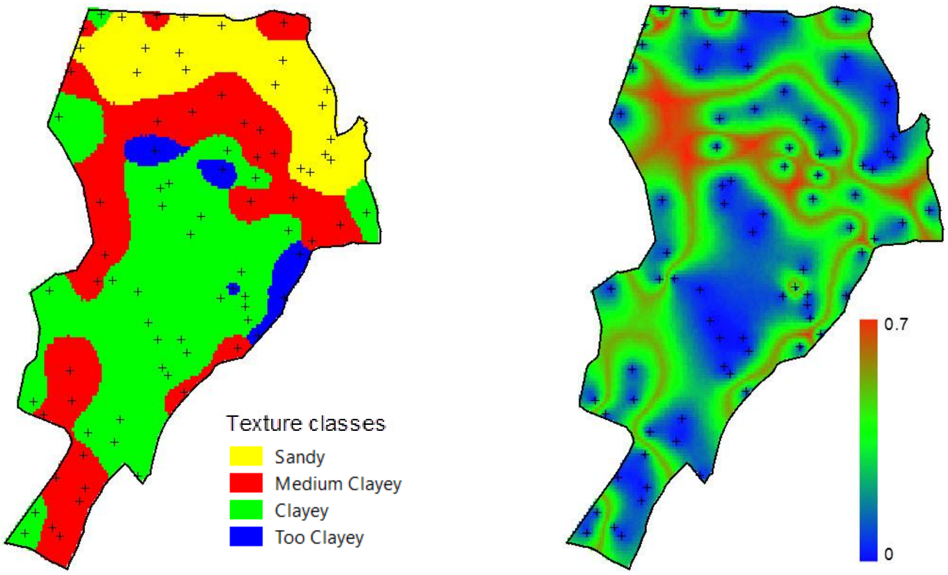
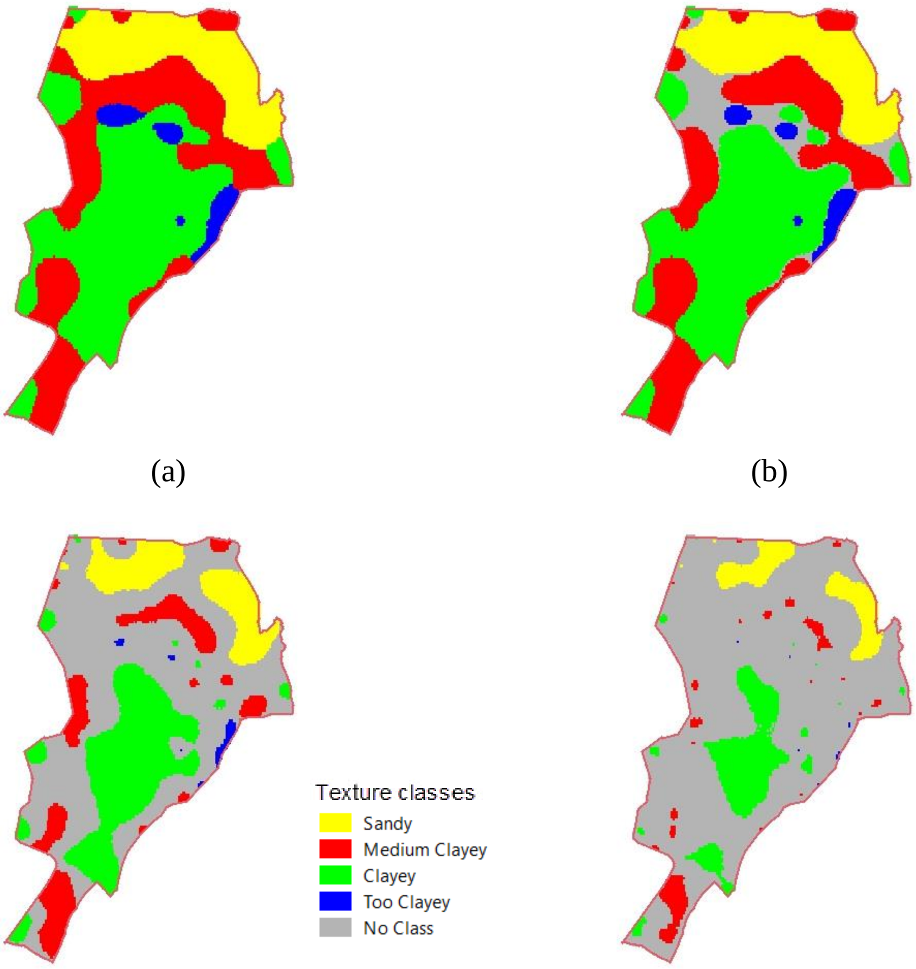


Figure 2. Map of predictions of texture classes (left) and map of uncertainties (right) estimated using the output realizations of the GSLIB sisim function.

The uncertainties depicted in figure 2 were assessed by 1-P_{max} value, the complement to 1 of the *cpdf* higher probability. As expected for environmental attributes, the uncertainties are higher in the borders, the transitions areas, of the soil texture class regions. Consequently, the probability uncertainty values are lower in the middle of the classes.

Figure 3 presents maps of predictions of texture classes constrained to different uncertainty threshold values. These maps were built by integrating the classified information with their respective uncertainties. From these maps it can be observed that the smaller the value of the uncertainty threshold, the more restricted areas of the classes will be accepted to be effectively used. These means that the planner can control the size of the regions to be handled according the uncertainty or the maximum region extent accepted by the user application.



(a) (b)

(c) (d)

Figure 3. Prediction maps of texture classes constrained by different uncertainty threshold values: (a) 1; (b) .5; (c) .25 and; (d) .12.

The following program developed in LEGAL language, that is a language for spatial analyses of the SPRING GIS, was applied to the predicted and uncertainty information of figure 2 in order to obtain the maps of figure 3. Uncertainty threshold values of 1, 0.5, 0.25 and 0.12 were assigned to the threshold variable of the program in order to create different output scenarios.

```
{
// Variable Declarations
  Thematic predictionsin, predictionsout ("Texture-Classes");
  Numeric  uncertainty ("Texture-DTM");

// Variable Instantiations
  predictionsin = Retrieve (Name = "Mode_Sisim-1000rfat_rec" );
  uncertainty = Retrieve (Name = "Uncert_Sisim-1000r_rec");

  predictionsout = New (Name = "Mode-12%", ResX = 35 , ResY = 50 , Scale = 10000 );

// Operations
  threshold = 0.12;
  predictionsout = (uncertainty <= threshold) ? predictionsin : uncertainty < 1 ? Class("No Class"):Class(0);

}
```

The program initiates with an opened curly bracket, {, and terminates with a closed curly bracket, }, symbol. All the line commands preceded by // are comments and are not considered as part of the program interpretation. All language commands must be terminated with semicolon character. Two *Thematic* variables, named predictionsin and predictionsout, are declared as belonging to the Texture-Classes database category. One *Numeric* variable, named uncertainty, is declared as belonging to the Texture-DTM category. Two input infolayers are assigned to the predictionsin and uncertainty variables through the *Retrieve* language command. An empty output infolayer, called Mode-12%, is created, through the *New* language command, with spatial resolution x and y, ResX and ResY, equal to 35m and 50m, respectively, and with a Scale equal to 1:10000. The uncertainty threshold value is user defined and has 0.12 value in this example. The last command is used to fill each point of a rectangular grid of the output infolayer, represented by the predictionsout variable, with classes determined by a macro command that means: if the uncertainty value is lower than a threshold, then (? symbol) maintain the information of the predictionsin data, else (: symbol), if the uncertainty is lower than 1 set the grid value equal to the class named No Class and, otherwise, set the grid value to the background class 0. As a result of the program above, the classes are preserved when the uncertainty value is lower than the user defined threshold and is assigned to the No Class value otherwise.

The areas of the classes in the maps of figure 3 are reported in table 2. Those areas were calculated in raster format, considering the pixel resolutions 35 m and 50 m in the x and y directions respectively.

Table 2: Table with areas of the remaining classes in each constrained classification

Uncertainty level (%)	**Area of the Texture Classes (m^2)**						
	Sandy	**Medium Clayey**	**Clayey**	**Too Clayey**	**No Class**	**Total Classified**	**Total Classified**
100	5439000	8452500	11620000	1109500	0	26621000	100%
50	5141500	7323750	10711750	824250	2619750	24001250	90.16%
25	3020500	3010000	5978000	199500	14413000	12208000	45.86%
12	1701000	878500	2941750	22750	21077000	5544000	20.83%

Based on the information of figure 3 and table 2, and considering different uncertainty values, environmental planners can build different scenarios with different area values for the classified regions. The area values can be used for making decisions on evaluating risks, or profits, to be assumed for agricultural project implementations, for example.

In the maps of figure 3 the uncertainty restriction was applied to all the classes. The planner can also work on create scenarios considering only a subset of the available classes. This idea is illustrated in figure 4 that shows two classified maps containing only the Medium Clayey and Clayey classes and the uncertainty threshold values were set to 0.5 and 0.25. The following command:

```
predictionsout = (predictionsout == Class("Sandy") || predictionsout == Class("Too Clayey")) ? Class("No Class") : predictionsout;
```

was inserted as the final command of the above LEGAL program in order to take the classes Sandy and Too Clayey off the final maps.

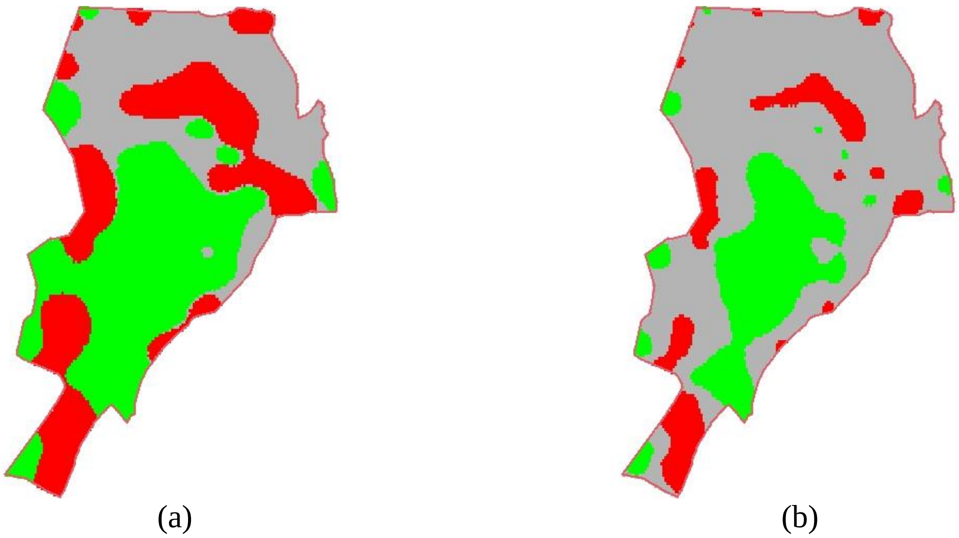


Figure 4. Prediction maps of Medium Clayey and Clayey texture classes constrained by uncertainty threshold values (a) 0.5 and (b) 0.25.

CONCLUSIONS

This work illustrated how the spatial modelling of categorical attributes, initially represented by a sample set of points, can be accomplished by indicator geostatistical approaches known as indicator kriging and simulation. These approaches have allowed to create prediction maps of categorical attributes along with their uncertainty maps which can be used to qualify the predictions.

Moreover, the article showed that constraints could be applied to the predictions using the uncertainty information. It is important to note these constrained predictions aim to minimize risks of loss in economy, agricultural productivity or any other application where the uncertainties are considered important. Thus, the constrained prediction maps of categorical attributes should be considered by decision makers responsible to adapt the resulting classifications to user demands, as maximum area to be used or maximum budget (investment) to be spent in a project related to the earth surface supplies. An advantage of the presented methodology is to provide the user different planning possibilities, or scenarios, to be applied in the geographical application being developed.

Even though this work considers only predictions and uncertainties information, it is important to know that the set of realizations of the indicator simulations can be used also as input for multivariable spatial modelling of categorical variables in Monte Carlo approaches, for example.

In the future we intend to explore similar methodology for spatial modelling of continuous attributes and for different uncertainty metrics as, for example, the Shannon entropy.

AKNOWLEDGMENTS

The authors would like to thank FAPESP – the São Paulo Research Foundation – for the support, via process number 2015/24676-9, given to the researches that allow the realization of this article.

REFERENCES

Câmara, G.; Souza, R.C.M.; Freitas, U.M.; Garrido, J.. 1996. "SPRING: Integrating remote sensing and GIS by object-oriented data modelling". Computers & Graphics. Volume 20:(3). Pages 395-403.

Delbari, M.; Afrasiab, P.; Loiskandl, W.. 2011. “Geostatistical Analysis of Soil Texture Fractions on the Field Scale”. Soil & Water Resources. Volume 6(4). Pages173–189.

Deutsch, C.V. and Journel, A. G.. 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. Volume 32(10). Pages 1669-1681.

Felgueiras, C.A.. 2000. “Modelagem ambiental com tratamento de incertezas em sistemas de informação geográfica: o paradigma geoestatístico por indicação”. 165 pages. PhD Dissertation, INPE, São José dos Campos, São Paulo, Brazil.

Felgueiras, C.A.; Fuks, S.D..; Monteiro, A.M.V.. 2003. “Classificação de Atributos Espaciais baseada em Informações de Incertezas. Uma Metodologia de Apoio a Decisões.”. Anais do XI Simpósio Brasileiro de Sensoriamento Remoto. Belo Horizonte, Brasil, Pages 967-974

Felgueiras, C.A.; Monteiro, A.M.V.; Camargo, E.C.G.; Ortiz, J. O.. 2014. “Improving Accuracy of Categorical Attribute Modeling with Indicator Simulation and Soft Information”. Proceedings of the 13th International Conference on GeoComputation, Richardson, Texas, USA.

Goovaerts, P.. 1997. “Geostatistics for Natural Resources Evaluation”. Oxford University Press, New York, USA.

Goovaerts, P.. 2001. “Geostatistical modeling of uncertainty in soil science”. 2001. Geoderma. Volume103. Pages 3–26.

Isaaks, E.H.; Srivastava R.M.. 1989. “An Introduction to Applied Geostatistics”. Oxford Univ. Press, New York, USA.

Juanga, K.; Chenb, Y.; Leeb, D.. 2004. “Using sequential indicator simulation to assess the uncertainty of delineating heavy-metal contaminated soils”. Environmental Pollution. Volume 127. Pages 229–238.

Shannon, C. E.; Weaver, W.. 1949. “The mathematical theory of communication”. Urbana: The University of Illinois Press. 117 Pages.

Wasiullah; Bhatti, A.U.. 2005. “Mapping of soil properties and nutrients using spatial variability and geostatistical techniques”. Soil and Environment. Volume 24(2). Pages 88-97.

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