Revista Brasileira de Cartografia (2016), N⁰ 68/4, Edição Especial Geoinformação e Análise Espacial: 655-664
Sociedade Brasileira de Cartografia, Geodésia, Fotogrametria e Sensoriamento Remoto
ISSN: 1808-0936

SPATIAL MODELING OF CATEGORICAL ATTRIBUTES USING INDICATOR SIMULATION AND SOFT INFORMATION WITH UNCERTAINTY ANALYSES

Modelagem Espacial de Atributos Categóricos Usando Simulação por Indicação e Informação Secundária com Análises de Incertezas

Carlos Alberto Felgueiras[1], Suzana Druck[2], Antonio Miguel Vieira Monteiro[1], Jussara de Oliveira Ortiz[1] & Eduardo Celso Gerbi Camargo[1]

[1]Brazilian National Institute for Space Researches – INPE
DPI - Divisão de Processamento de Imagens
Av. dos Astronautas, 1758, Jardim da Granja, São José dos Campos, SP, Brazil, CEP: 12227-010, CP: 515
{carlos, miguel, jussara, eduardo}@dpi.inpe.br

[2]Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA
CPAC - Centro de Pesquisa Agropecuária dos Cerrados, Rodovia BR 020 Km18, Brasilia, Distrito Federal, Brazil,
CEP: 73310-970, CP: 08223
suzana@cpac.embrapa.br

Recebido em 18 de Outubro, 2015/ Aceito em 7 de Março, 2016
Received on October 18, 2015/ Accepted on March 7, 2016

ABSTRACT

This work explores a methodology to apply indicator geostatistical simulation approaches to geospatial modeling of categorical attributes using hard and soft information. Uncertainty analyses of the predictions are performed to evaluate the quality of classifications. 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 determining the soft information. These data are gathered from an experimental farm of agriculture researches. Uncertainty analyses of the results show that the use of soft information, along with the hard data, allows one to find out new specific regions of higher and lower uncertainties. The highest uncertainties regions should be considered as candidates for future resampling.

Keywords: Geostatistics, Geospatial Modeling of Categorical Attributes, Indicator Sequential Simulations, Hard and Soft data, Uncertainty Analyses.

RESUMO

Este trabalho explora uma metodologia de uso de procedimentos geoestatísticos de simulação por indicação na modelagem geoespacial de atributos categóricos usando informações primárias e secundárias. Análises de incertezas das predições são realizadas para avaliação da qualidade das classificações. Consideram-se amostras pontuais de um atributo categórico como informações principais, ou primárias, enquanto que dados de um mapa categórico são usados como informações auxiliares, ou secundárias. A informação auxiliar, correlacionada com a principal, é incorporada ao procedimento de simulação por indicação como valores médios a priori, tomados a partir de uma função de distribuição de probabilidade. Os valores médios a priori são, então, atualizados via simulação por indicação considerando-se os dados principais disponíveis em suas vizinhanças. Para ilustrar a metodologia, apresenta-se um estudo de caso com amostras de classes de textura do solo, dados primários, e com classes de um mapa de solos, dados secundários. Estas informações foram obtidas de uma fazenda experimental usada para pesquisas agrícolas. As análises dos resultados mostram que a utilização de informações secundárias, em conjunto com os dados primários, determinam novas regiões específicas de baixas e altas incertezas. As regiões de mais altas incertezas devem ser consideradas como candidatas para futuras reamostragens.

Palavras-Chave: Geoestatística, Modelagem Geoespacial de Atributos Categóricos, Simulações Sequenciais por Indicação, Dados Primários e Secundários, Análises de Incertezas.

1. INTRODUTION

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 correlated set of sample points of secondary information.

From the uncertainty models it is possible to derive attribute predictions and realizations along with uncertainty metrics as, for example, confidence intervals of the probability 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 modeling 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. *Sequential Indicator Simulation* (*SIS*) is a widely used geostatistical technique for modeling 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, conditioned to the hard data. The prior mean values are then updated via indicator simulation to account for the hard data available in their neighbourhoods.

In this context, the main objective of this work is to present a methodology for spatial modeling of categorical data applying indicator geostatistical simulation approaches on hard and soft information. In addition, important uncertainty analyses of the predictions are performed to evaluate the quality of the classifications. This article is an extension with significant improvements of the Felgueiras et al., 2015, article presented in the GeoComputation 2015 conference (1).

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 clayey, clayey and too clayey. The classes of the soil map of the region of interest were taken

[1] This article is an extension with significant improvements of the Felgueiras et al., 2015, article presented in the GeoComputation 2015 Conference, Richardson, Texas, USA.

in order to get *cpdf* prior mean values of texture classes for each soil class. The soil textures were modelled using the hard data only and using the hard and the soft information. The resulting maps were presented, compared, and analyzed, mainly considering the information presented in the uncertainty maps obtained. The results show that the use of soft information, along with the hard data, can improve the quality of the final classifications showing regions with specific regions of higher and lower uncertainties. The highest uncertainties regions should be considered as candidates for future resampling.

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 for agriculture researches; section 5 presents results and discussions and; section 6 addresses final conclusions and new ideas for future researches related to the improvement of spatial data modeling.

2. CONCEPTS

The 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 *Simulation* process consists of 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:

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

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.

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

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 *Sequential Simulation* process works with the cpf estimates 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 gridded 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.

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}, while the uncertainty map may be assigned with the 1-P_{max} values. Other metrics of uncertainty can be used, as the Shannon Entropy, for example, that takes into account all the probability values of a cpdf (Shannon, 1949, Felgueiras, 2000).

3. METHODOLOGY

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

1. For a set of sample points of a categorical attribute, the hard data, evaluate the semi-variograms of the indicator sample sets related the attribute classes;
2. Determine the local prior *pdf* values for each spatial location of the output grid using a secondary information, the soft data;
3. Fill the parameter file of the *SIS* GSLIB functions known as *sisim* and *sisim_lm*;
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°54’46” to s 21°59’31” and the east-west coordinates from w 47°51’46” to w 47°48’18”

The hard 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 (left map) illustrates the borders of the Canchim farm along with location and the classification of the soil texture sample set. This categorical map was obtained with a nearest neighbor interpolation procedure showing the regions of influence of each class.

It was also considered the 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 shown in Table 1.

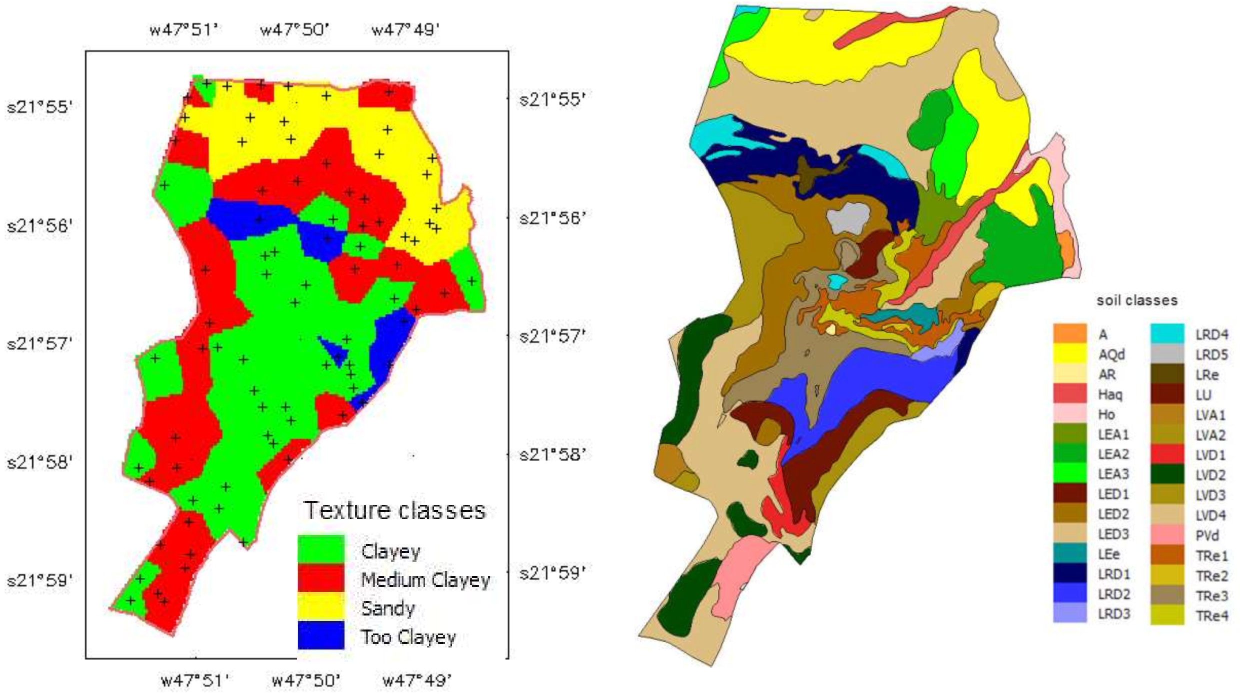


Fig. 1 - Distribution of the soil texture sample points (left map) and map of soil classes of the Canchim region (right map).

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

Soil Class	Sandy	Medium Clayey	Clayey	Too Clayey
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

Figure 2 shows the soil map (left) and the result of a soil map reclassification (right) according to the maximum texture probability a priori of each soil class presented in Table 1. The reclassified map allows one to have a general idea of the a priori spatial distribution of the texture classes given the information of the soil classes.

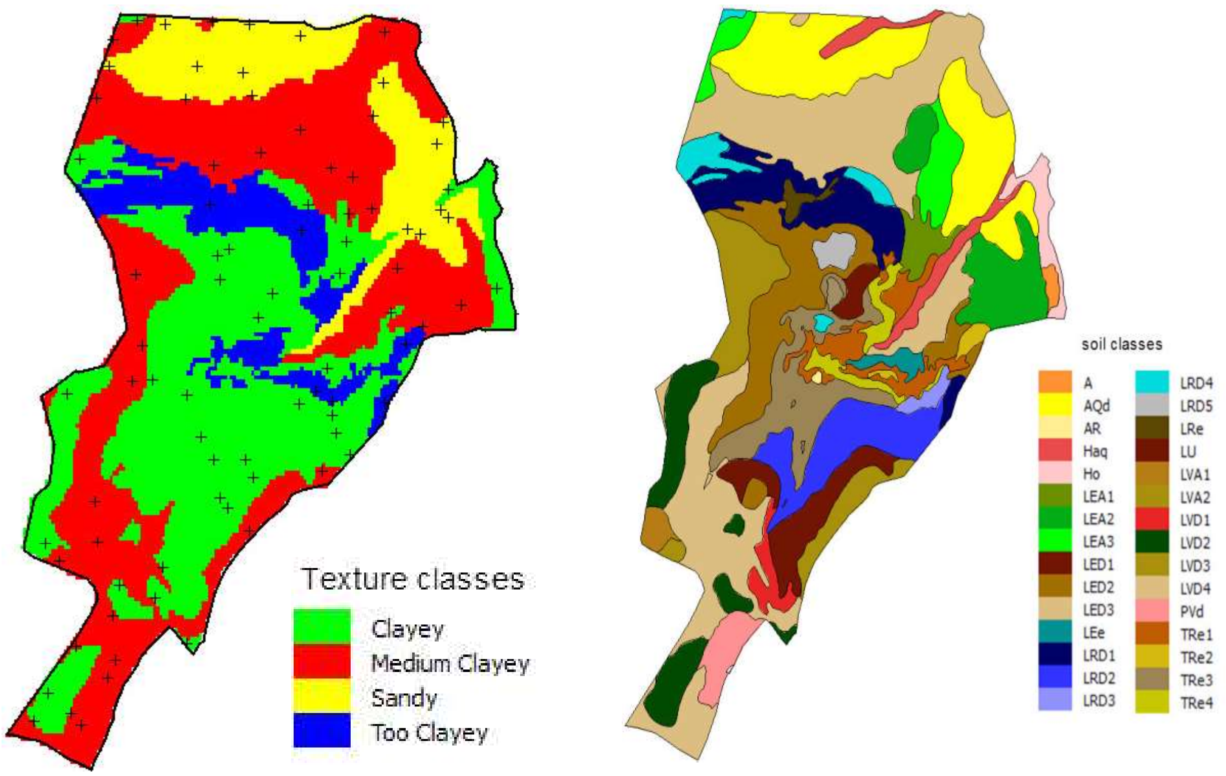
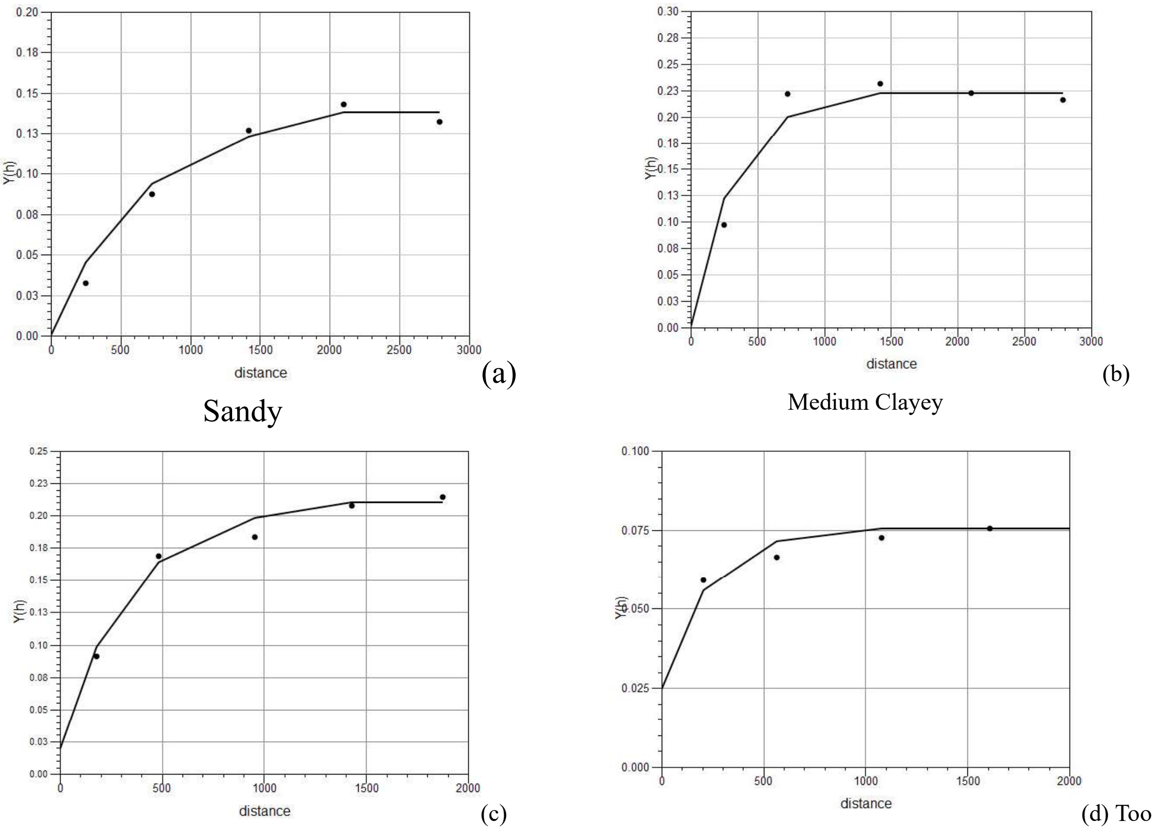


Fig. 2 - Soil map (left) and map of texture classes (right) according to the maximum probability a priori of the distributions presented in Table 1.

5. RESULTS AND DISCUSSIONS

The spatial dependences analyses are represented by the indicator semivariograms generated from the indicator sample set defined by each texture class. Figure 3 illustrates the four indicator semivariograms representing the four soil texture classes consi dered. This is necessary in order to run the geostatistical *SIS* GSLIB functions. The spatial dependence analyses are based on the sample set of the soil texture classes.



(a)

Sandy

(b)

Medium Clayey

(c)

Clayey

(d) Too

Clayey

Fig. 3 - Indicator semivariograms of the four soil texture classes.

The indicator semivariogram parameters, along with the global probabilities, of each texture class are reported in the Table 2.

Table 2: Parameters of the indicator semivariograms

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

All the semivariograms were fitted with exponential functions. The global probabilities are assessed by the ratio between the number of samples of each class and the total number The information presented in Table 2, together with a text file with the sample data set, are used as input parameters for the GSLIB *SIS* functions.

Figure 4 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. A qualitative, visual, comparison between this map of predictions and the map of nearest neighbours interpolation, left map of Figure 1, shows that the both maps globally agree with the spatial distribution of the texture sample set. The differences appear in the smoother class transitions presented in the map predicted from the geostatistical simulated values.

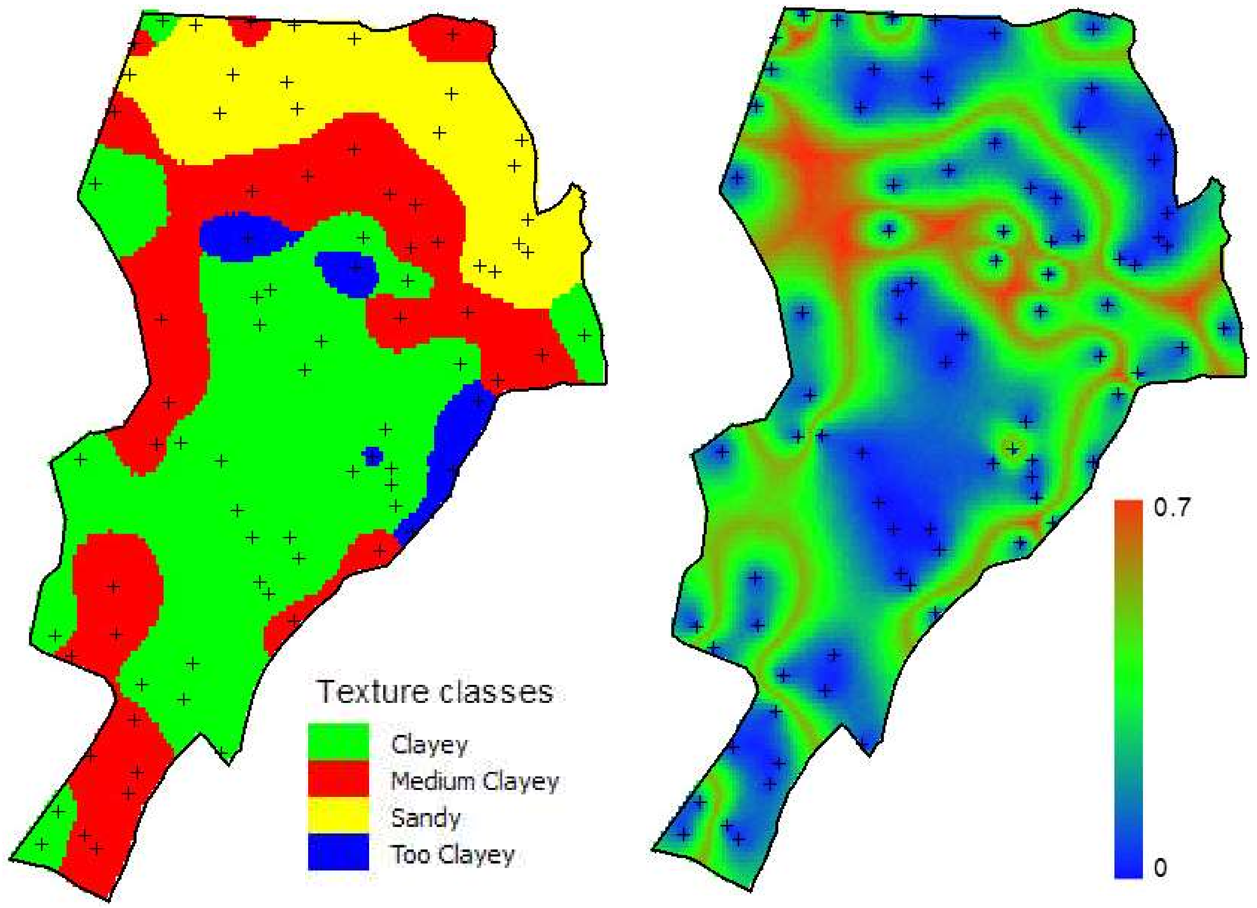


Fig. 4 - Map of predictions of texture classes (left) and map of uncertainties (right) estimated using the output of the sisim function.

The uncertainties depicted in Figure 4 were defined as 1-P_{max} (the higher probability of the *cpdf*). 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

Figure 5 shows the map of predicted texture classes (left) and respectively uncertainties (right) obtained from the realizations of the *sisim_lm* GSLIB function. The estimations were also assessed from the higher probabilities of the *cpdfs* estimated at each spatial location.

As for the texture class map of Figure 4 the class transitions of this predicted texture map are smoother than that of the Figure 1. The general class distributions presented in this map also agree with the spatial distributions of the hard data, but show important differences at some regions. The differences are caused by the inclusion of the probabilities a priori, from the secondary information, in the simulation process. For example, the region of too clayey class, highlighted in the map by the black polygon, had a significant increase caused by the soil class and the lack of primary samples in this area.

The maps of Figure 5 show significant differences from those presented in Felgueiras et al., 2015, where the search radius parameters of the GSLIB functions were defined smaller than that used in this work. This fact implied in predictions considering too local input information that privileges the denser secondary information.

In order to facilitate visual comparisons, between the texture class maps of figures 4 and 5, Figure 6 depicts these both maps with the reclassified map of Figure 2.

As for the Figure 4, the uncertainty map of Figure 5 shows higher values in the border of the soil texture classes of the predicted map of this figure. Consequently the probability uncertainty values are lower in the middle of those regions.

With the aim of make easier visual comparisons Figure 7 depicts the two uncertainty maps resulting of the simulation processes.

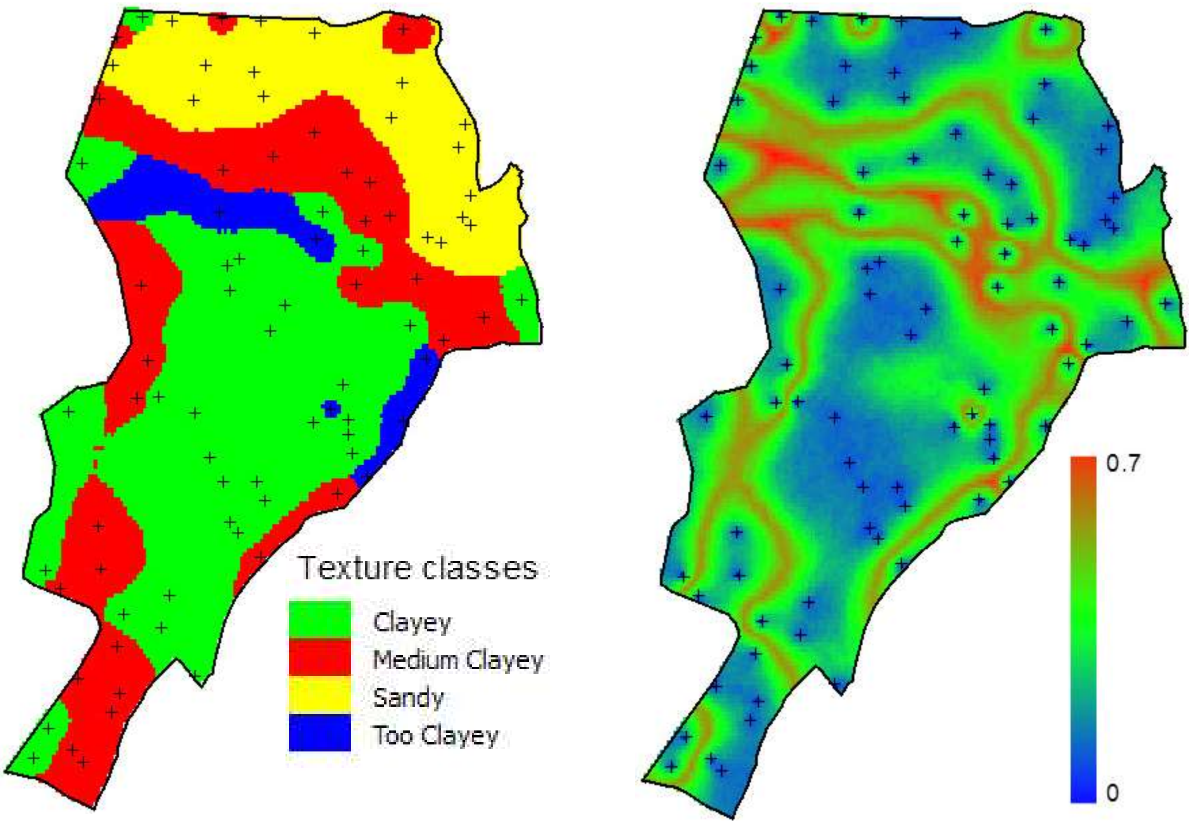


Fig. 5 - Map of predictions of texture classes (left) and map of uncertainties (right) estimated using the output of the sisim_lm function.

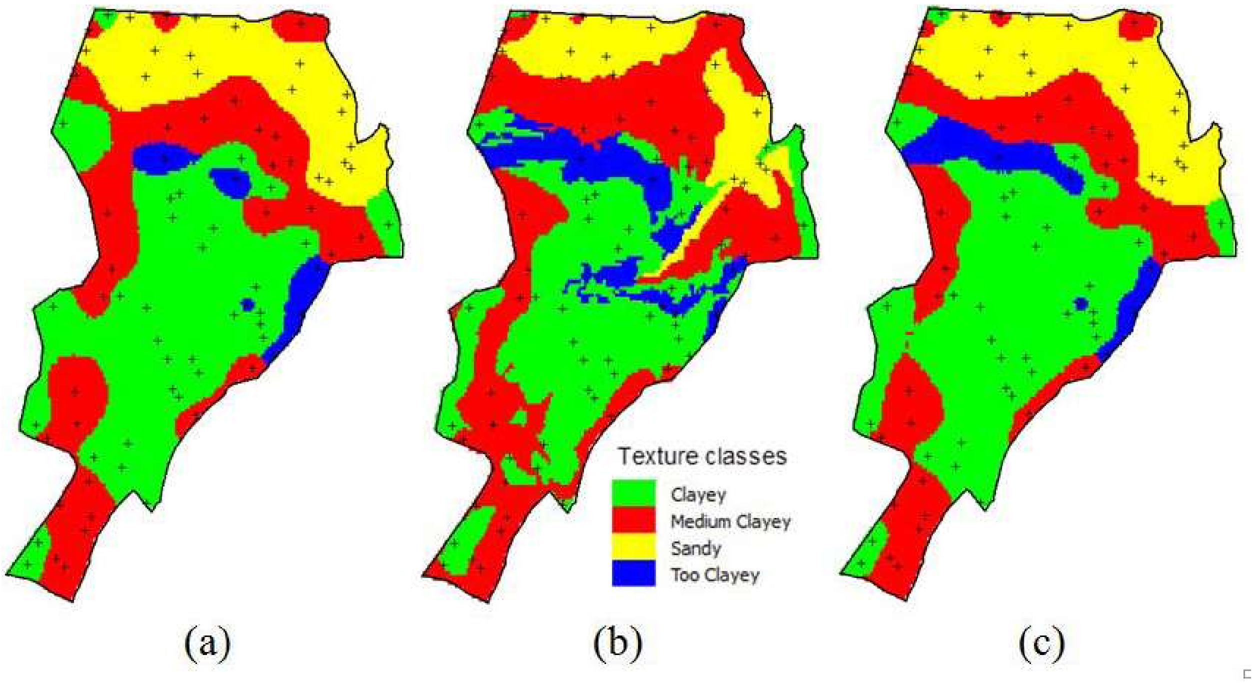


Fig. 6 - Soil texture class maps obtained by: (a) simulation with hard data, (b) reclassification according to soil classes and (c) simulation with hard and soft data.

Both maps of Figure 7 show similarities and differences in the spatial distributions of their uncertainty values. Differences appear in spatial regions where the estimated classes from hard and soft information do agree or not. Where there are agreements, as shown in highlighted regions A, B and C of Figure 7(b), the uncertainties decreased. Where there are no agreements, as shown in highlighted regions D, E and F of Figure 7(b), the uncertainties increased. These areas are natural candidates for future sampling or resampling field works. The agreements and disagreements occur mainly because the number of the hard samples is too small for the whole considered map area.

Table 3 reports some global statistics of the uncertainty maps of the soil texture modeling depicted in Figure 7.

The global maximum and the mean values of the Table 3 are higher for the uncertainty map of Figure 7(b). It means that the disagreements between hard and soft data are higher than the agreements in this experiment.

Table 3: Statistics information of the uncertainty maps of Figure 7

Input Information	Mi*ni*-mum	Maxi-mum	Mean	Variance	Standard Deviation
Sample Set	0.0	0.67	0.28	0.03	0.16
Sample Set plus Soil Information	0.0	0.70	0.31	0.02	0.14

The variance and the standard deviation are smaller for this map. It means that the global uncertainty variation around the mean value is lower, i. e., the uncertainty distribution is more homogeneous for the map of Figure 7b.

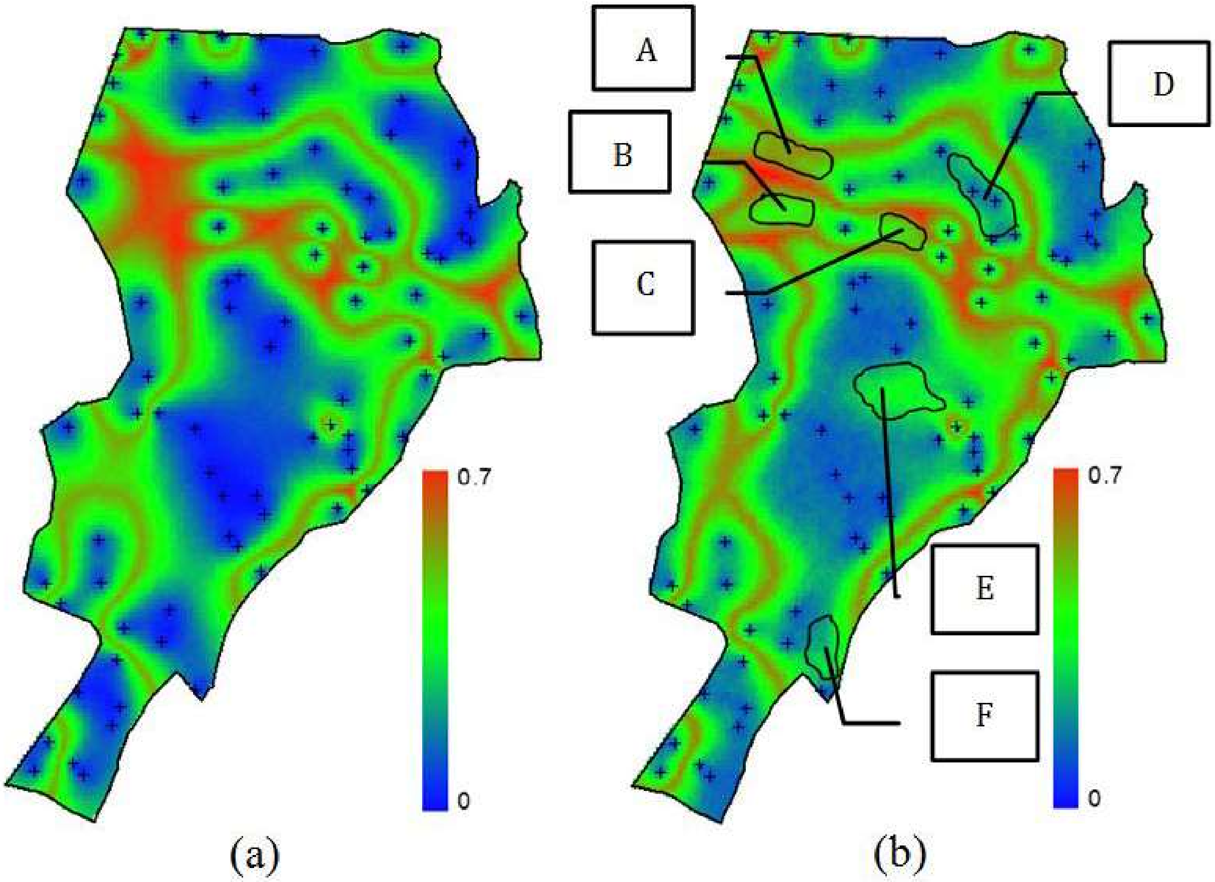


Fig. 7 - Uncertainty maps of the soil texture modeling using (a) only the sample set and (b) the sample set along with the soil classes information.

6. CONCLUSIONS

Spatial modeling of categorical attributes can be accomplished by 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 quality of the prediction and of the uncertainty representations.

Although the global uncertainty increased, as occurred in this experiment, the quality of classification can be considered superior after inclusion of the correlated secondary information in the simulation. The availability of secondary information is used to fill gaps for the lack of information in regions without primary samples. Moreover, in class conflicting regions, high uncertainties are indications that such areas should be sampled or resampled. Thus, even with increased global uncertainty, the classified map using secondary information is more reliable and should be applied in decision making procedures for planning activities on environmental applications, for example.

Although this work considers only predictions and uncertainties information, it is important know that the set of realizations of the indicator simulations can be used also 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 hard and secondary variables.

REFERENCES

DELBARI, M.; AFRASIAB, P.; LOISKANDL, W. Geostatistical Analysis of Soil Texture Fractions on the Field Scale. **Soil & Water Resources**, 6(4): 173–189. 2011.

DEUTSCH, C. V. & JOURNEL, A. G. **GSLIB: geostatistical software library and user's guide**. Oxford University Press, New York, USA. 1998. 384p.

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

FELGUEIRAS, C. A. **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. 2000.

FELGUEIRAS, C. A.; MONTEIRO, A. M. V.; CAMARGO, E. C. G.; ORTIZ, J. O. **Improving Accuracy of Categorical Attribute Modeling with Indicator Simulation and Soft Information**. Proceedings of the 13th International Conference on GeoComputation, Richardson, Texas, USA. Available online at: http://www.geocomputation.org/2015/index.html 2015. 25-31pp.

GOOVAERTS, P. **Geostatistics for natural resources evaluation**. Oxford University Press, New York, USA, 1997. 496p.

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

ISAAKS, E. H.; SRIVASTAVA R. M. **An Introduction to Applied Geostatistics**. Oxford University Press, New York, USA. 1989. 561p.

JUANGA, K.; CHENB, Y.; LEEB, D. Using sequential indicator simulation to assess the uncertainty of delineating heavy-metal contaminated soils. **Environmental Pollution**, 127: 229–238. 2004.

SHANNON, C. E.; WEAVER, W. **The mathematical theory of communication**. Urbana: The University of Illinois Press, 117p. 1949.

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

ZAERI, K.; HAZBAVI, S.; TOOMANIAN, N.; ZADEH, J. T. 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, 2013.