Raster Representations Of Spatial Attributes With Uncertainty Assessment Using Nonlinear Stochastic Simulation

Carlos Alberto Felgueiras [(1)]; Suzana Druck Fuks[(2)]; Antonio Miguel Vieira Monteiro[(3)]
[(1)]*DPI/INPE Brazilian Institute for Space Research*
carlos@dpi.inpe.br
[(2)]*CPAC/EMBRAPA Brazilian Agricultural Research Corporation*
suzana@cpac.embrapa.br
[(3)]*DPI/INPE Brazilian Institute for Space Research*
miguel@dpi.inpe.br

Abstract. Raster representations of thematic and numerical spatial attributes are very common in a GIS environment for computational simulation and analysis of spatial processes. This paper addresses the problem of predictions with uncertainty assessment for GIS raster representations created from a set of sample points of spatial attributes. The realizations of a stochastic simulation process, over numerical attribute samples, are used for inferencing the attribute values and the related uncertainties at non-sampled spatial locations. A case study, using elevation sample data, is presented in order to illustrate the used methodology with real data.

1. INTRODUCTION

GIS environment allows one to simulate and analyze different scenarios that can be used to support decisions made about a specific real spatial process. The main idea is to integrate representations of spatial attributes in order to analyze simulated spatial processes in a computational environment. The resulting scenarios will depend on the data representations and manipulations and, also, on the mathematical models used to integrate them. Raster representations of thematic and numerical spatial attributes are frequently used in a GIS environment for computational simulation and analysis of spatial processes. Raster representations of spatial attributes are derived from a set of attribute samples, commonly observed as sample points over a spatial region of interest. Nonlinear stochastic simulation procedures, based on the indicator kriging approach, can be used to create raster representations of spatial attributes. The realizations of the stochastic simulation inference process, over numerical attribute samples, are used in order to infer attribute values along with inference uncertainties at non-sampled spatial locations. The uncertainty of each representation can be propagated to the resulting scenarios of the computational spatial modeling (Heuvelink, 1998). The resulting uncertainties will qualify the scenarios, or the objects presented in the scenarios, yielding a quantitative information of the risk assumed for an adopted scenario. In this context, this work explores a methodology to create raster representations of numerical attributes, from a set of sample points, using a nonlinear stochastic approach called indicator sequential simulation. Furthermore, this work shows how to assess uncertainty values related to the attribute inferences obtained by this methodology. Different uncertainty metrics, based on confidence intervals, is addressed. A case study for an elevation sample set is presented in order to illustrate the use of the methodology applied to real data. Also, uncertainty metrics will be applied to the data realizations in order to qualify the elevation inferences

2. THE GEOSTATISTICAL PARADIGM FOR ATTRIBUTE INFERENCES WITH UNCERTAINTY ASSESSMENT

From a geostatistical point of view, the distribution of a spatial attribute in a region $A \subset \Re^2$ of the earth surface is represented as a random function $Z(\mathbf{u})$. For each position $\mathbf{u} \in A$ the attribute is considered as a random variable (RV) that can assume different values depending on the model of the spatial distribution of $z(\mathbf{u})$, i. e., depending on its probability distribution function (pdf). The conditional cumulative distribution function (ccdf) of a continuous RV $Z(\mathbf{u})$, conditioned to (n) sample points $z(\mathbf{u}_\alpha)$, $\alpha = 1,2,...,n$, can be denoted as

A random function (RF) is a set of RVs defined over some field of interest. A RF $Z(\mathbf{u})$ is characterized by a set

$$F(\mathbf{u}; z \mid (n)) = Prob\{Z(\mathbf{u}) \leq z \mid (n)\}$$

of all its K-variate ccdfs and its multivariate ccdf is defined as:

From the ccdf one can derive different optimal estimates for any unsampled value $z(\mathbf{u})$ in addition to the ccdf

$$F(\mathbf{u}_1,...,\mathbf{u}_K; z_1,...,z_K) = Prob\{Z(\mathbf{u}_1) \leq z_1,...,Z(\mathbf{u}_K) \leq z_K\}$$

mean, which is the least-squares error estimate (Deutsch, 1998). Also, the univariate ccdf of a RV is used to model uncertainty about the value $z(\mathbf{u})$ while the multivariate ccdf is used to model joint uncertainty about K values $z(\mathbf{u}_1),...,z(\mathbf{u}_k)$. Therefore, it is possible to derive various probability intervals that can be used as uncertainty metrics. These derivation processes will be addressed in the next sections.

3. THE CCDF DETERMINATION

The ccdf of a numerical RV, or of a numerical RF, can be obtained *parametrically* or *non-parametrically*. In the parametrical approach, the ccdf is determined by a limited set of statistical parameters. For example, the Gaussian ccdf is fully determined by two parameters, the mean and the variance of the distribution. Unfortunately it is a hard work to find out whether the distribution of a continuous attribute can be modeled by parametric ccdf or not. Non-parametrical distributions are more common for spatial attributes and can be estimated using the indicator kriging approach that will be explained in the next section.

4. THE CCDF APPROXIMATION USING THE INDICATOR KRIGING APPROACH

Instead of the variable $Z(\mathbf{u})$, consider its binary indicator transformation $I(\mathbf{u};z_k)$ defined as:

The expectation $E\{\ I(\mathbf{u};z_k)|(n)\ \}$ yields an estimation F^* for the ccdf of $Z(\mathbf{u})$ at the cutoff value z_k and conditioned to the n sample data (Deutsch, 1998), i. e.:

Using a linear kriging approach, as simple or ordinary

$$I(\mathbf{u};z_k)=\begin{cases}1, & \text{for } Z(\mathbf{u})\le z_k\\ 0, & \text{for } Z(\mathbf{u})> z_k\end{cases}$$

kriging (Camargo, 1997), to evaluate the expectation E defined in the above equation, the indicator kriging of a continuous variable aims to provide a least-squares

$$E\{I(\mathbf{u};z_k)\mid(n))\}=$$
$$1\cdot Prob\{I(\mathbf{u};z_k)=1\mid(n)\}+0\cdot Prob\{I(\mathbf{u};z_k)=0\mid(n)\}=$$
$$1\cdot Prob\{I(\mathbf{u};z_k)=1\mid(n)\}=F^*(\mathbf{u};z_k\mid(n))$$

estimate of the ccdf at cutoff z_k. A set of ccdf estimates in various cutoffs can lead to an approximation of the full ccdf of $Z(\mathbf{u})$. Some corrections for the follow order relation deviations:

and

must be performed to guarantee that the ccdf estimations are between 0 and 1 and increase

$$0\le F^*(\mathbf{u};z_k\mid(n))\le 1\quad\forall z_k,\ k=1,\ldots K$$

monotonically. Figure 1 illustrates the fitting process of

$$F^*(\mathbf{u};z_j\mid(n))\le F^*(\mathbf{u};z_k\mid(n))\ \text{ se } z_j\le z_k$$

the ccdf estimation using 5 cutoff values.

5. THE INDICATOR SIMULATION APPROACH

Stochastic simulation, hereafter called simulation for simplicity, is the process of drawing l alternative, l=1,...,L,, equally probable, joint realizations of the component RVs from an RF model (Deutsch, 1998). Each realization of $Z(\mathbf{u})$ is denoted by $z^{(l)}(\mathbf{u})$. A conditional simulation is the simulation conditioned to a set of n sample data. In this case the resulting realizations honor the sample data values at their $\mathbf{u}_a$ spatial locations, i. e., $z^{(l)}(\mathbf{u}_a)=z(\mathbf{u}_a),\ \forall\ l$.

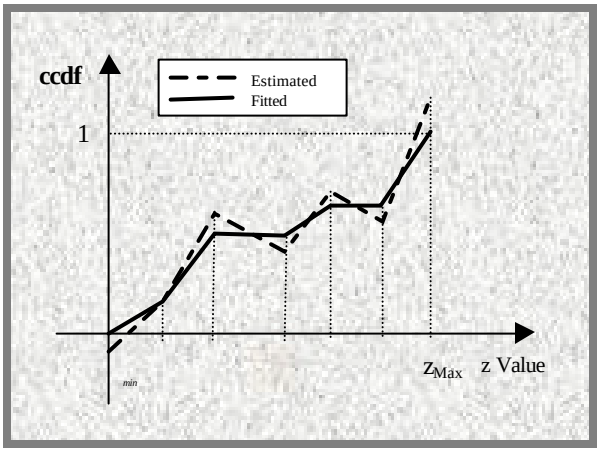
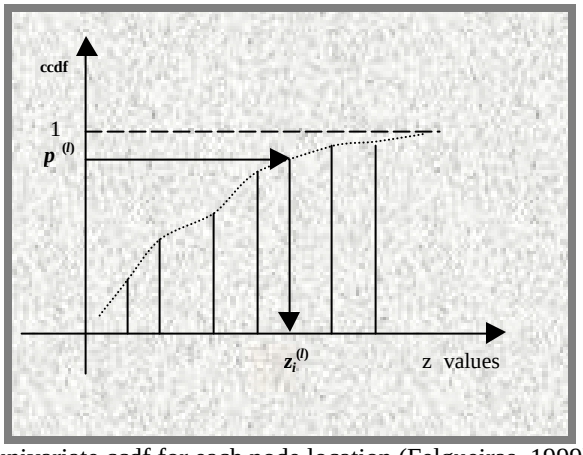


Figure 1: The ccdf estimation using indicator kriging approach with order relation corrections

Deutsch, 1998, presents a sequential indicator simulation approach that uses local ccdf approximation, determined by the indicator kriging approach, in order to obtain realizations of RVs $Z(\mathbf{u})$. For creating a raster representation, a univariate ccdf is modeled at each node of the all grid nodes visited along a random sequence. To ensure reproduction of the z-covariance model, each univariate ccdf is made conditional not only to the sample data but also to all simulated values at previously visited locations (Goovaerts, 1997).

The realizations are drawn using probability values, obtained from an uniform random model, that are mapped to z values taking into account the estimate



univariate ccdf for each node location (Felgueiras, 1999). Figure 2 illustrates this process.

Figure 2: Process of obtaining a realization from a estimated univariate ccdf

6. EVALUATION OF STATISTICAL PARAMETERS FROM THE REALIZATIONS

The set of realizations at a node location u can be used to determine a ccdf, along with its parameters, of a RV $Z(\mathbf{u})$.

The most popular predictive ccdf parameter is the mean value m. From a set of L realizations the mean value of a ccdf is evaluated as the average of all the realizations. The variance s^2 and the standard deviation s are easily evaluated using the realization values and the mean value.

The median value, $q_{.5,}$ can be determined splitting the set of realization into 2 subsets, each with equal number of elements. Also, the set of realizations can be split in more equal subsets to derive different quantile values. When the median and the mean values are closer the distribution can be considered symmetric The median is a more robust estimator for non-symmetrical distributions (Isaaks, 1989).

7. UNCERTAINTY ASSESSMENT FOR LOCAL ESTIMATES

As already emphasized, in section 2, the univariate ccdf of a RV is used to model uncertainty about the value $z(\mathbf{u})$ while the multivariate ccdf is used to model joint uncertainty about K values $z(\mathbf{u}_1),...,z(\mathbf{u}_k)$. Therefore, given a ccdf model it is possible to derive various probability intervals that can be used as uncertainty metrics.

For numerical attributes usually the uncertainties are expressed as confidence intervals. When the ccdf of a RV $Z(\mathbf{u})$ presents a high degree of symmetry and the normality of the distribution can be assumed, the estimated value $z^*(\mathbf{u})$, typically the mean value m, and the standard deviation s are combined to derive Gaussian-type confidence intervals, centered on $m_Z(\mathbf{u})$, such as:

$$Prob\{Z(\mathbf{u}) \hat{I} [\mu_Z(\mathbf{u}) \pm s(\mathbf{u})]\} @ 0.68$$

or

$$Prob\{Z(\mathbf{u}) \hat{I} [\mu_Z(\mathbf{u}) \pm 2s(\mathbf{u})]\} @ 0.95$$

where $s^2(\mathbf{u})=E\{(Z(\mathbf{u})-E\{Z(\mathbf{u})\})^2\}$.

For non-symmetrical distributions one can derive probability intervals based on quantiles of the ccdf. For example, the 95% interval $[q_{0.025};q_{0.975}]$ is taken as:

$$Prob\{Z(\mathbf{u}) \hat{I} [q_{0.025};q_{0.975}]|(n)\} = 0.95$$

with $q_{0.025}$ and $q_{0.975}$ being the 0.025 and 0.975 quantiles of the ccdf, i. e., $F^*(\mathbf{u}; q_{0.025}|(n)) = 0.025$ and $F^*(\mathbf{u}; q_{0.975}|(n)) = 0.975$

8. A CASE STUDY FOR ELEVATION DATA

In order to illustrate the concepts presented above, the following case study uses a set of elevation data sampled in the region of an experimental farm called 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 data set consists of 662 elevation samples distributed in the Canchim region. Some statistic values of this sample set is shown in the Table 1.

Table 1: Univariate statistics of the elevation sample set of the Canchim region

Statistic	*Value*
Number of Samples	662
Mean Value	800.596
Variance	4481.662
Standard Deviation	66.945
Coefficient of Variation	0.084
Coefficient of Skewness	-0.296
Coefficient of kurtosis	1.562
Minimum Value	687.000
Lower Quartile	732.500
Median	827.000
Upper Quartile	859.500
Maximum Value	911.000

The histogram graph, presented in the Figure 3, shows the distribution of the elevation sample set compared with a normal curve distribution. It can be seen that the sample data distribution approximates a bimodal behavior and differs considerably from the Gaussian (normal) or symmetrical distribution.

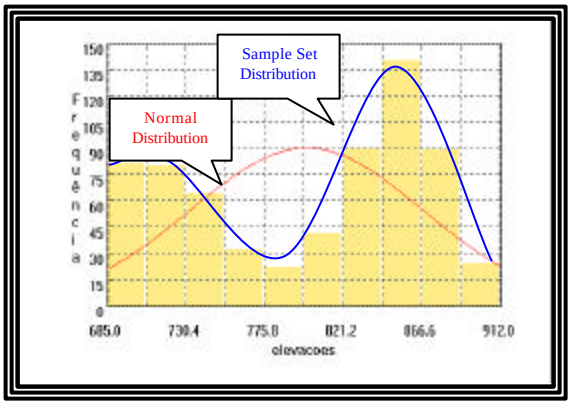


Figure 3: Histogram of the elevation sample set emphasizing the non-normal and non-symmetrical behavior of the distribution

The spatial distribution of the elevation sample set in the Canchim region is illustrated in the Figure 4

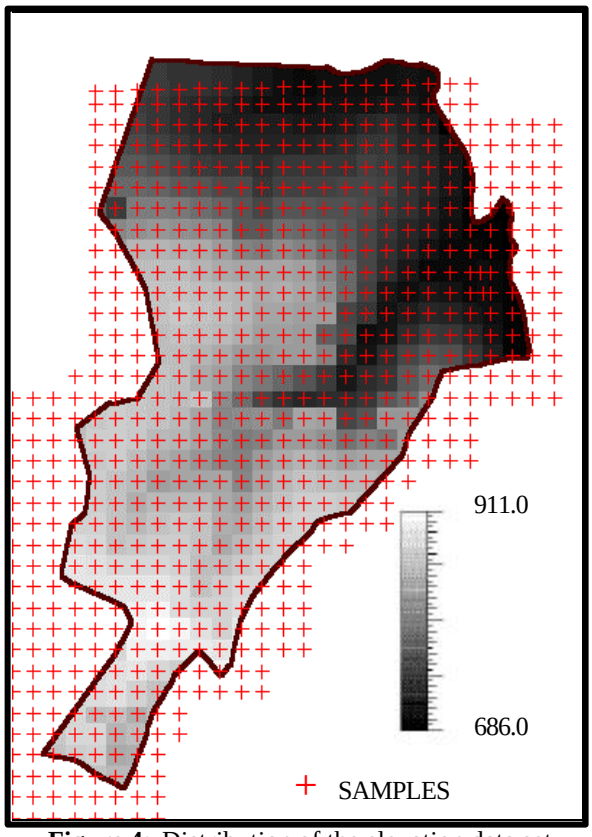


Figure 4: Distribution of the elevation data set observed in the Canchim region.

The original sample set was split in 10 equal subsets (deciles) using 9 cutoff elevation values. Each cutoff value was considered in order to create indicator subsets using the indicator transformation explained in section 4. The variability of the indicator subsets are analyzed allowing the definition of an experimental and a theoretical variogram model for each subset. These tasks were performed using the geostatistical module of the SPRING GIS version 3.5 (SPRING V.3.5, 2001).

The variogram models, along with the original sample set, were used to set the parameter values of the gslib (Deutsch, 1998) sequential simulation program named sisim. This program was modified and used for estimating 400 realizations of 200 rows by 200 columns elevation grids (rectangular regular grids). Considering the 400 elevation realizations at any grid location **u** it was possible to render the mean m and the median value $q_{.5}$ maps using the methodologies defined in section 6. These maps are shown in the Figures 5 and 6. A qualitative (visual) comparative analysis of the two maps shows that they differs. This is explained by the non-symmetrical distribution of the elevation distribution model. Because of these, the median map can be considered more representative as central measure for this attribute in the region considered.

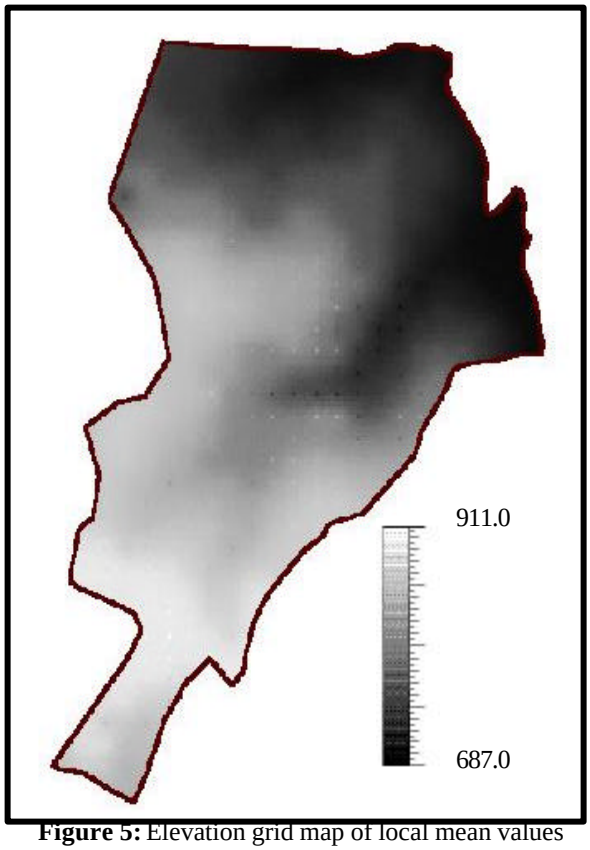


Figure 5: Elevation grid map of local mean values estimated from the 400 grid realizations

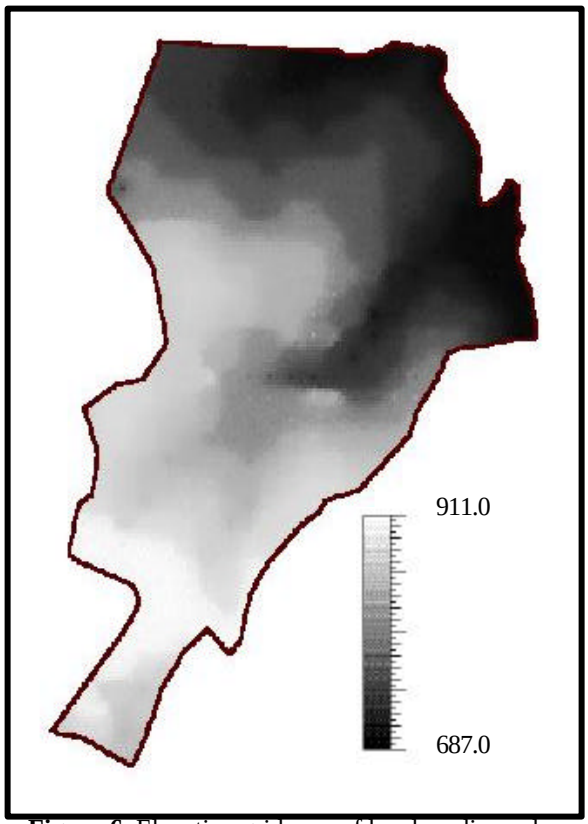


Figure 6: Elevation grid map of local median values estimated from the 400 grid realizations

The Figures 7, 8 and 9 show uncertainty maps rendered using also the 400 elevation realizations and the confidence interval methodologies as explained in the section 7.

It can be seen that all the uncertainty map values are related to the attribute behavior. These uncertainty maps present maximum uncertainty values on regions (whiter regions) where the attribute values behave more erratically. Minimum uncertainty values (blacker regions) appear where attribute values vary smoothly.

The map of Figure 7 shows uncertainty values based on Gaussian-type confidence intervals. This map was generated using one standard deviation centered in the mean value ($Prob\{Z(\mathbf{u}) \in (m \pm s)\} \cong 0.68$). It is common to use this map as the uncertainty map related to the map estimated by mean values (Figure 5). A care has to be taken on using this type of uncertainty representation. It must be used only when the attribute variation can be modeled as RV with symetric-distributions (normal one, for example).

The maps of Figures 8 and 9 represent uncertainties as confidence intervals based on quantiles. The map of Figure 8 was obtained using interquartile confidence intervals ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.25};\boldsymbol{q}_{0.75}]\} = 0.5$) while the map of Figure 9 was generated with interdecile confidence intervals ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.10};\boldsymbol{q}_{0.90}]\} = 0.8$).

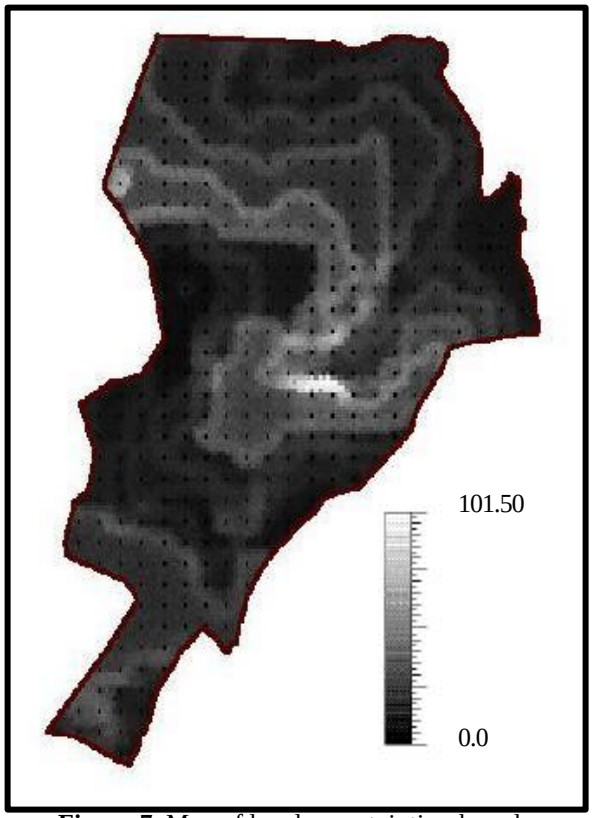


Figure 7: Map of local uncertainties based on Gaussian-type confidence intervals ($Prob\{Z(\mathbf{u}) \in (m \pm s)\} \cong 0.68$)

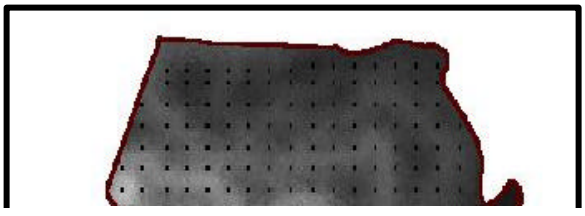


Figure 8: Map of local uncertainties based in the interquartile confidence intervals ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.25};\boldsymbol{q}_{0.75}]\} = 0.5$)

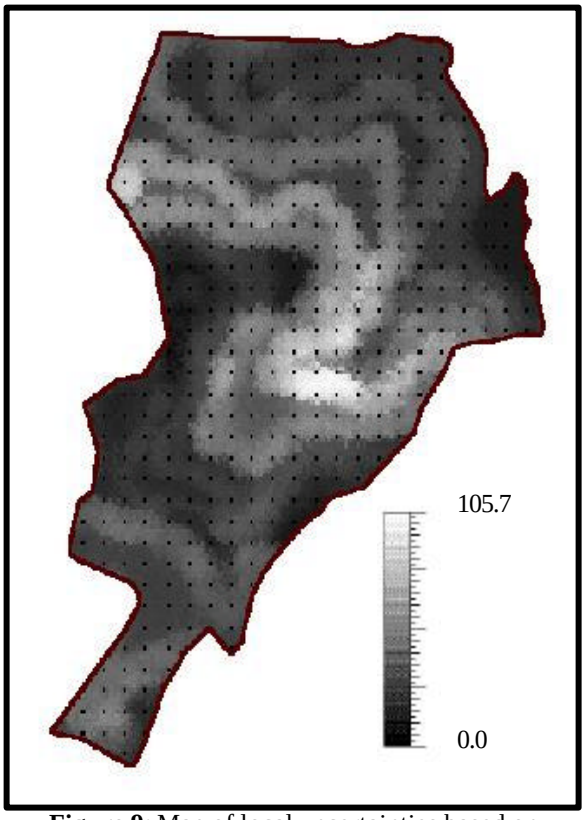


Figure 9: Map of local uncertainties based on interdecile confidence intervals ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.10};\boldsymbol{q}_{0.90}]\} = 0.8$)

As expected the map of Figure 9 contains larger uncertainty values than the one of Figure 8. The decision about which one to use depends on the accuracy demanded by an application. Finally the interquantile uncertainty maps are more appropriated to be used when the RV distributions can not be proven to

have symmetrical behavior, as for the elevation attribute considered in this work.

9. CONCLUSIONS

The concepts and results presented in this work show that the indicator simulation methodology is an interesting option to be considered when estimates with uncertainty assessments for numerical spatial attributes are required. The use of indicator simulation approach presents the following advantages:

- the indicator approach is non-parametric, so, it can be used independently of the attribute distribution model;
- the indicator approach allows assessment of uncertainties related to the attribute variability using the fitted local attribute distribution models;
- the sequential indicator algorithm determines the univariate ccdfs taking into account the attribute values of the sample data set and all the previously simulated values. This ensure reproduction of the z-covariance model, better representing the attribute variability;
- the various equally probable outcome realizations of the indicator simulation can be used as input for complex spatial modeling (with multi-layer analysis) performed by Monte Carlo simulation method, for example. Also, the outcomes of the spatial analysis results can be used to define their ccdf´s and, therefore, modeling their local uncertainties.
- Finally, it can be emphasized that the indicator simulation methodology can be applied, also, to thematic attributes with minor modifications. This has been the subject of researches that will be reported in the near future.

REFERENCES

Burrough, P. A. .and McDonnell, R. A. (1998) *Principles of Geographical Information Systems.* Oxford University Press: New York

Camargo, E. C. G. (1997) *Desenvolvimento, implementação e teste de procedimentos geoestatísticos (krigeagem) no Sistema de Processamento de Informações Georeferenciadas (SPRING).* Msc. Dissertation on Remote Sensing – INPE – The Brazilian Institute for Space Research, São José dos Campos, São Paulo.

Deutsch, C. V. and Journel, A. G. (1998) *GSLIB Geostatistical Software Library and User's Guide*. Oxford University Press: New York

Felgueiras, C. A. (1999) *Modelagem Ambiental com Tratamento de Incertezas em Sistemas de Informação Geográfica: O Paradigma Geoestatístico por Indicação.* Phd thesis on Applied Computation – INPE – The Brazilian Institute for Space Research, São José dos Campos, Publ. in: http://www.dpi.inpe.br/teses/carlos/.

Goovaertz, P. (1997) *Geostatistics for Natural Resources Evaluation*. Oxford University Press: New York.

Heuvelink, G. B. M. (1998) *Error Propagation in Environmental Modeling with GIS*. Bristol, Taylor and Francis Inc: London.

Isaaks, E. H. and Srivastava, R. M. (1989) *An Introduction to Applied Geostatistics*, Oxford University Press: New York.

SPRING V.3.5 (2001) Sistema de Processamento de Informações Georeferenciadas – DPI - Image Processing Division - INPE – The Brazilian Institute for Space Research, São José dos Campos, São Paulo. URRL: http://www.dpi.inpe.br/spring/ .