On the Use of Indicator Simulation for Numerical Predictions with Uncertainty Assessment

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

[1]INPE/DPI—Instituto Nacional de Pesquisas Espaciais, Caixa Postal 515, 12201 São José dos Campos, SP, Brasil
{carlos,miguel,eduardo}@dpi.inpe.br
[2]EMBRAPA/CPAC – Empresa Brasileira de Agropecuária , Rodovia Brasília Fortaleza Br 020 Km 18, Planaltina
Distrito Federal, Brasil
drucks@ensam.inra.fr

Abstract. Raster representations of thematic and numerical spatial attributes are used very commonly 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 inference process, over numerical attribute samples, are used in order to infer the attribute values and related uncertainties at non-sampled spatial locations. A case study, with elevation sample data, is presented to illustrate the concepts used in this work.

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 spatial data attribute representations in order to study the spatial process in a computational environment. The final scenarios depend on the data representations and also on the mathematical model used to integrate them. The attribute representations are derived from a set of attribute samples, commonly sample points, obtained in a spatial region of interest. Nonlinear stochastic procedures, based on the indicator kriging approach, can be used to create attribute representations along with uncertainty information related to a set of estimated attribute values. The uncertainty of each representation can be propagated to the resulting scenarios of a spatial process modeling. The resulting uncertainties will qualify the scenarios, or the objects presented in the scenarios, yielding a quantitative information of the risk assumed when a determined scenario is chosen. In this context, this work presents a methodology to create attribute representations, from a set of sample points, using a nonlinear stochastic approach called *indicator simulation*. Furthermore, this work shows how to obtain uncertainty values related to the attribute value inferences created by this methodology. Different uncertainty metrics, based on confidence intervals, will be addressed. A case study for an elevation sample set will be presented to show how the methodology can be applied to real data and how to use the uncertainty metrics in order to qualify inferences of numerical attribute representations.

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 $z(\mathbf{u}_\alpha)$, α =1,2,...,n, can be denoted as:

$$F(\mathbf{u};\mathrm{z}\,|\,(n)) = Prob\{Z(\mathbf{u}) \le \mathrm{z}\,|\,(n)\}$$

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

$$F(\mathbf{u}_1,...,\mathbf{u}_\mathrm{K};\mathrm{z}_1,...,\mathrm{z}_\mathrm{K}) = Prob\{Z(\mathbf{u}_1) \le \mathrm{z}_1,...,Z(\mathbf{u}_K) \le \mathrm{z}_K\}$$

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

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

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, i. e.:

$$\begin{aligned}&E\{I(\mathbf{u};z_k)\,|\,(n)\}=\\&1\cdot Prob\{I(\mathbf{u};z_k)=1\,|\,(n)\}+0\cdot Prob\{I(\mathbf{u};z_k)=0\,|\,(n)\}=\\&1\cdot Prob\{I(\mathbf{u};z_k)=1\,|\,(n)\}=F^*(\mathbf{u};z_k\,|\,(n))\end{aligned}$$

Using a linear kriging approach, as simple or ordinary kriging, to evaluate the expectation E defined in the above equation, the indicator kriging of a continuous variable aims to provide a least-squares 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:

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

and

$$F^*(\mathbf{u};z_{\mathbf{j}}\,|\,(n))\le F^*(\mathbf{u};z_{\mathbf{k}}\,|\,(n))\ \text{ se } z_{\mathbf{j}}\le z_{\mathbf{k}}$$

must be performed to guarantee that the ccdf estimations are between 0 and 1 and increase monotonically. Figure 1 illustrates the fitting process of the ccdf estimation using 5 cutoff values.

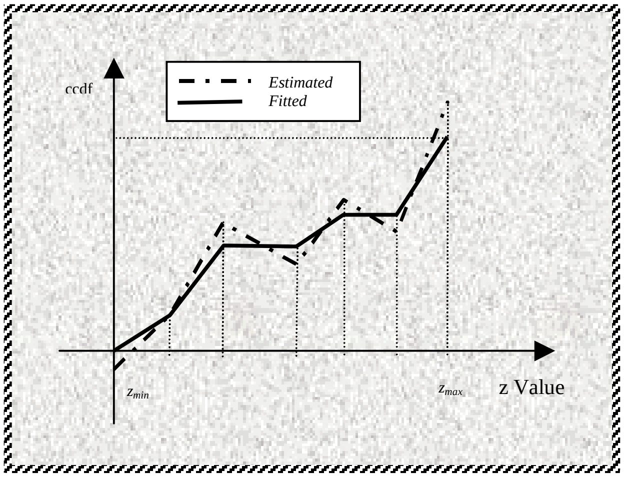

Figure 1 The ccdf estimation using indicator kriging approach with order relation corrections

5 The indicator simulation approach

Stochastic simulation, hereafter called simulation for simplicity, is the process of drawing alternative, equally probable, joint realizations of the component RVs from an RF model (Deutsch, 1998). Each realization of $Z(\mathbf{u})$ is denoted by $z^{(l)}(\mathbf{u})$, $\mathbf{u}\in A$. 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 location, i. e., $z^{(l)}(\mathbf{u}_\alpha)=z(\mathbf{u}_\alpha)$, $\forall\ l$.

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 $Z(\mathbf{u})$, $\mathbf{u}\in A$. For creating a raster representation, one 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 values simulated 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.

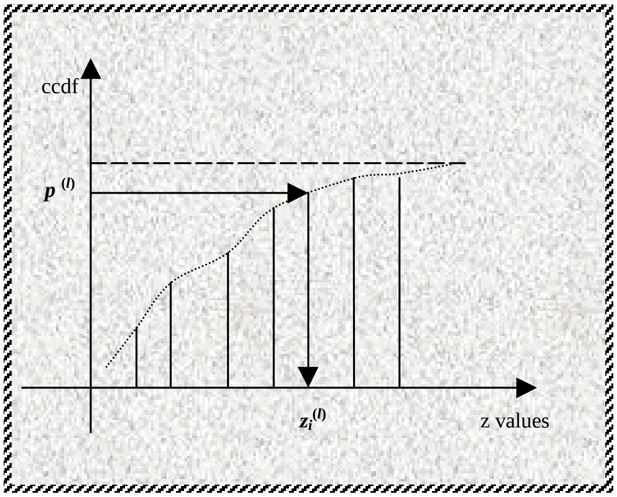

Figure 2 Process of obtaining a realization from a estimated univariate ccdf

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 μ. From a set of realizations the mean value of a ccdf is evaluated as the average of all the realizations. The variance and the standard deviation, σ, is 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 μ, and the standard deviation σ are combined to derive a Gaussian-type confidence intervals, centered on $z^*(\mathbf{u})$, such as:

$$Prob\{Z(\mathbf{u})\in [\text{ì}_Z(\mathbf{u}) \pm \text{ó}(\mathbf{u})]\} \cong 0.68$$

or

$$Prob\{Z(\mathbf{u})\in [\text{ì}_Z(\mathbf{u}) \pm 2\text{ó}(\mathbf{u})]\} \cong 0.95$$

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

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:

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

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, this case study uses a set of elevation data sampled in the region of a 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 as illustrated in the Figure 3. Some statistic values of this sample set is shown in the Table 1.

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

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

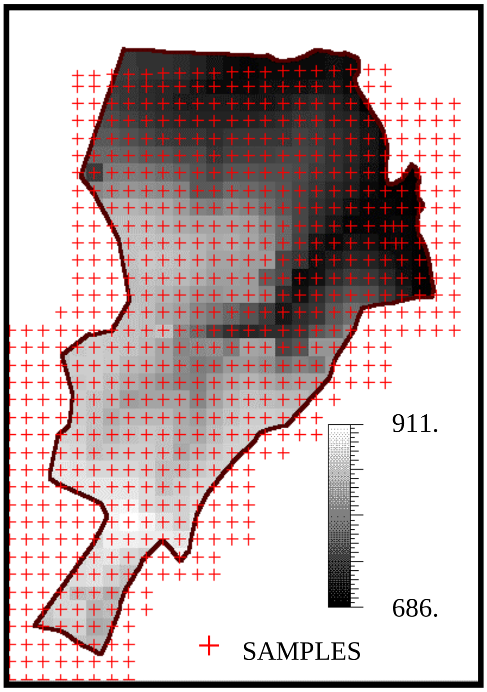

Figure 3 Distribution of the elevation data set observed in the Canchim region.

The histogram graph, presented in the Figure 4, 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.

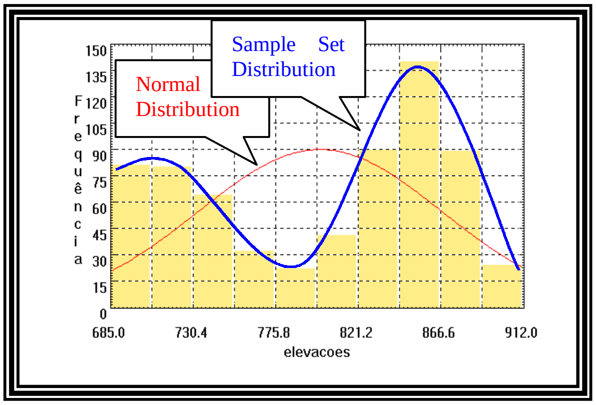

Figure 4 Histogram of the elevation sample set emphasizing the non-normal and non-symmetrical behavior of the distribution.

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 3.3 (SPRING (DPI/INPE), 1999).

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 and the median value 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.

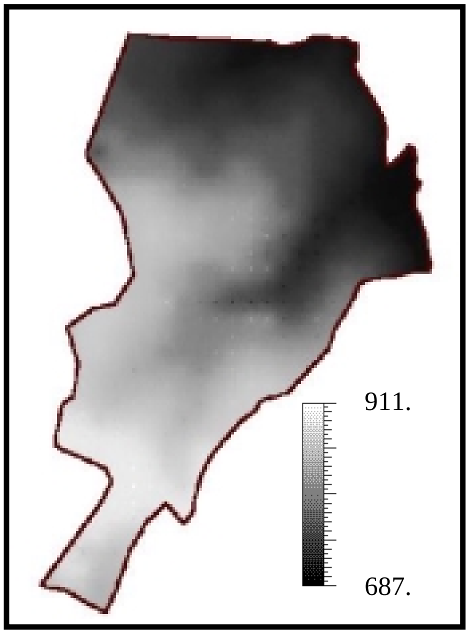

Figure 5 Elevation grid map of local mean values estimated from the 400 grid realizations

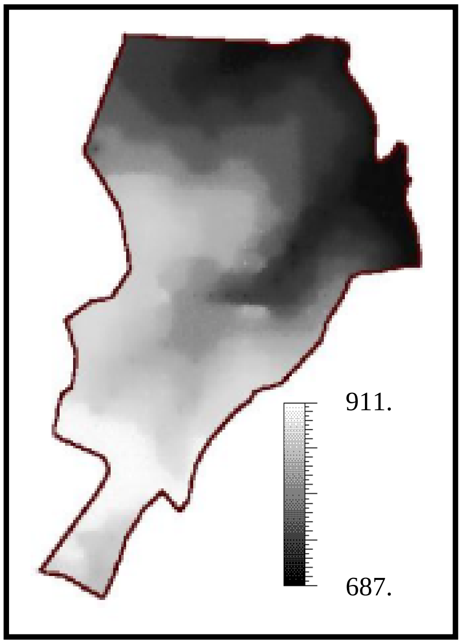

Figure 6 Elevation grid map of local median values estimated from the 400 grid realizations

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

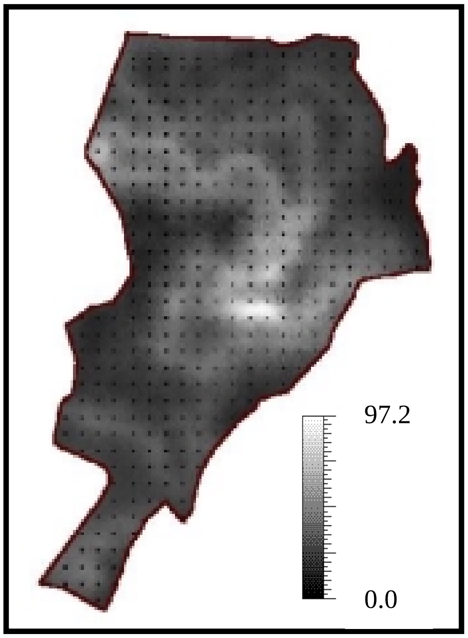

Figure 7 Map of local uncertainties based on Gaussian-type confidence intervals ($Prob\{Z(\mathbf{u}) \in (\boldsymbol{\mu} \pm \boldsymbol{\sigma})\} \cong 0.68$)

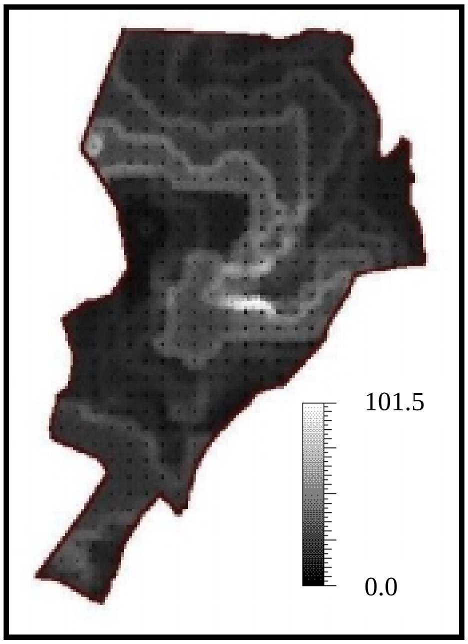

Figure 8 Map of local uncertainties based in the interquartile confidence intervals ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.25};\boldsymbol{q}_{0.75}]\} = 0.5$)

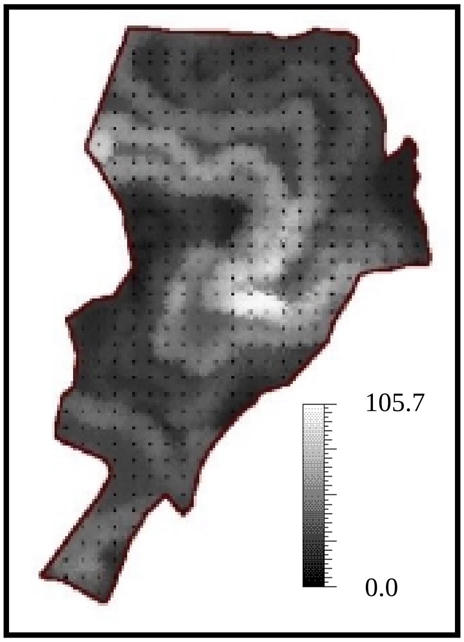

Figure 9 Map of local uncertainties based on interdecile confidence intervals ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.10};\boldsymbol{q}_{0.90}]\} = 0.8$)

It can be seen that all the above 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 the 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 ($Prob\{Z(\mathbf{u}) \in (\boldsymbol{\mu} \pm \boldsymbol{\sigma})\} \cong 0.68$). 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 quantile values are estimated directly from the 400 realizations as explained in the section 7.

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 ($Prob\{Z(\mathbf{u}) \in [\boldsymbol{q}_{0.10};\boldsymbol{q}_{0.90}]\} = 0.8$). 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.

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 an approximation of attribute distribution model;
- the sequential indicator algorithm determine the univariate ccdfs taking into account the sample data set and all values simulated at previously locations. This ensure reproduction of the z-covariance model, representing better the attribute variability;
- the various equally probable outcome realizations of the indicator simulation can be used as input for complex spatial modeling (with multilayer analysis) performed by Monte Carlo simulation method. Also, the outcomes of the spatial analysis results can be used to define their ccdf´s and, therefore, modeling their uncertainties.

Finally, it can be emphasized that the indicator simulation methodology can be applied, also, to thematic spatial attributes with minor modifications. This has been the subject of researchs that will be reported in the near future.

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, *Modelagem Ambiental com Tratamento de Incertezas em Sistemas de Informação Geográfica: O Paradigma Geoestatístico por Indicação.* Tese (Doutorado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Publicado em http://www.dpi.inpe.br/teses/carlos/, 1999.

[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] SPRING (DPI/INPE) 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.