USO DE SIMULAÇÃO ESTOCÁSTICA NÃO LINEAR PARA INFERÊNCIAS DE ATRIBUTOS ESPACIAIS NUMÉRICOS

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

[1]INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE/DPI
AV. DOS ASTRONAUTAS 1758 JARDIM DA GRANJA
CEP 12201-970 SÃO JOSÉ DOS CAMPOS SP BRASIL
FONE: (0XX12) 345 6519 FAX: (0XX12) 345 6468
E-MAILS: carlos@dpi.inpe.br, miguel@dpi.inpe.br, eduardo@dpi.inpe.br

[2]EMPRESA BRASILEIRA DE AGROPECUÁRIA – EMBRAPA/CPAC
BR 020 KM 18 RODOVIA BRASÍLIA FORTALEZA
PLANALTINA DISTRITO FEDERAL BRASIL
FONE: (0XX61) 389 1121 FAX: (0XX61) 389 2953
E-MAIL: drucks@ensam.inra.fr

RESUMO

O presente artigo explora o uso de uma ferramenta geoestatística conhecida por *Simulação Sequencial por Indicação* para inferir atributos numéricos a partir de um conjunto pontual de amostras. No trabalho utilizou-se um conjunto amostral de elevações, obtidos em uma fazenda experimental do Brasil para geração de modelos numéricos, representados por grades regulares, em ambiente de Sistemas de Informação Geográfica. Os métodos geoestatísticos assumem que o atributo numérico se comporta como uma variável aleatória em cada localização da superfície terrestre. Assim, os dados de altimetria, são inferidos a partir de um conjunto de realizações das variáveis aleatórias definidas para cada nó da grade. Além disso, essa técnica possibilita a obtenção de mapas de incertezas relacionados com as inferências de altimetria. Por isso, o trabalho explora, também, a definição de métricas de incerteza a partir dos conjuntos de realizações de mapas de altimetria, representados como grades regulares.

ABSTRACT

This work explores the use of a geostatistical tool named *Indicator Sequential Simulation* to infer numerical attributes from a sample point set. Elevation samples from a Brazilian experimental farm are used as input for creation of regular grids to be used as numerical models in Geographical Information Systems environment. The geostatistical procedures consider the numerical attribute as a random variable for each location of the earth surface. In this way, the elevation values are estimated from a set of random variable realizations simulated for each grid location. Furthermore, the presented simulation technique allows the generation of uncertainty maps related to the elevation inferences. On that account, this work also explores metrics for uncertainty definitions from a set of realizations of elevation grid maps.

INTRODUÇÃO

O desenvolvimento de aplicações relacionadas ao estudo, à análise e à simulação de processos ambientais reais, em Sistemas de Informação Geográfica – SIG –, requerem a modelagem de atributos espaciais da natureza. Essa modelagem envolve, em muitos casos, a obtenção de informação espacial derivada de inferências sobre amostras pontuais desses atributos. Conjuntos amostrais de atributos de natureza numérica, tais como, dados de altimetria, de geologia e de temperatura, e de natureza temática, tais como, classes de vegetação e de textura do solo, são exemplos típicos desses atributos espaciais. O procedimento geoestatístico não linear, para simulação e inferência de atributos espaciais, conhecido *como simulação estocástica sequencial* por indicação, pode ser usado, com vantagens, para modelagem de atributos espaciais em ambiente de SIG. Neste contexto, o presente trabalho tem os seguintes objetivos: desenvolver os aspectos conceituais, apresentar estudos de caso com dados reais e relatar vantagens relacionadas com o uso do procedimento de simulação estocástica sequencial por indicação para inferência de atributos ambientais de natureza numérica. O trabalho apresenta, ainda, métricas para obtenção de informação de incertezas associadas aos dados inferidos por esse procedimento. Os resultados do trabalho são apresentados em mapas que representam a espacialização, segundo uma estrutura matricial, do atributo espacial dentro de uma região de interesse. Além do mapa de atributos, apresentam-se mapas de incertezas, relacionadas ao processo de espacialização do atributo, obtidos por diferentes métricas de cálculo de incertezas. Exploram-se métricas de incertezas por intervalos de confiança e por intervalos de probabilidade. Neste trabalho utilizam-se dados de altimetria, amostrados pontualmente em uma fazenda experimental da Empresa Brasileira de Pesquisas Agropecuárias - EMBRAPA – do Brasil. As metodologias básicas, empregadas no trabalho, envolvem a técnica de simulação estocástica sequencial por indicação e os procedimentos de estimativa de incertezas para inferência de valores de altimetria em posições espaciais, geralmente não amostradas, de uma grade regular.

ASPECTOS CONCEITUAIS

Representação de atributos numéricos pela geoestatística

A geoestatística considera a distribuição espacial de um atributo, em uma região $A \subset \Re^2$ da superfície terrestre, é representada por uma função aleatória $FZ(\mathbf{u})$, $\mathbf{u} \in A$. Para cada posição $\mathbf{u}$ o atributo é representado como uma variável aleatória (VA) $Z(\mathbf{u})$,. A função de distribuição acumulada condicional (fdac), condicionada a (n) amostras, de uma VA contínua é denotada por:

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

A fdac univariada descreve o comportamento de uma VA $Z(\mathbf{u})$ é usada para modelar a incerteza sobre valores $z(\mathbf{u})$ assumidos para a VA.

Determinação da fdac de uma VA pela metodologia de krigeagem por indicação

A fdac univariada de uma VA numérica pode ser aproximada utilizando-se de uma metodologia geoestatística não paramétrica chamada *krigeagem por indicação*. Essa

metodologia requer a transformação das VA $Z(\mathbf{u})$ em VA por indicação $I(u; z_k)$ pela seguinte relação:

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

onde z_k é um valor de corte pertencente ao domínio do atributo.

O valor esperado da VA por indicação, $E\{ I(\mathbf{u}; z_k)|(n)$, fornece uma estimativa F^* da fdac de $Z(\mathbf{u})$ no valor de corte z_k e condicionado aos n dados amostrais do atributo $z(\mathbf{u}_\alpha)$.

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

Essa estimativa, quando realizada por uma krigeagem ordinária sobre o conjunto de valores por indicação, fornece uma inferência por regressão de mínimos quadrados para a fdac de $Z(\mathbf{u})$ no valor de corte z_k (Deutsch, 1998). Um conjunto de estimativas F^* em diferentes valores de corte leva a uma aproximação da fdac completa de $Z(\mathbf{u})$.

O procedimento de simulação por indicação

Uma *simulação estocástica* é um processo de obtenção de realizações, igualmente prováveis, a partir dos modelos de distribuição de probabilidades de VA. Dada uma VA $Z(\mathbf{u})$, cada realização l dessa VA é denotada por $z^{(l)}(\mathbf{u})$. Uma *simulação condicional* é uma simulação condicionada ao conjunto de n dados amostrais. Neste caso as realizações honram os valores do atributo nas posições dos dados amostrais, ou seja, $z^{(l)}(\mathbf{u}_\alpha) = z(\mathbf{u}_\alpha)$, $\forall\ l$.

Deutsch, 1998, apresenta uma metodologia chamada simulação sequencial por indicação que utiliza aproximações locais de fdac´s para obtenção de realizações de uma VA $Z(\mathbf{u})$. Essa simulação pode ser usada para criar representações matriciais de atributos contínuos e categóricos. Para isso, uma fdac univariada é definida para cada nó grade que é visitado em sequência aleatória. Para assegurar reprodução do modelo de covariância, cada fdac univariada é condicionada as amostras e também aos nós da grade previamente simulados (Goovaerts, 1997).

As realizações são obtidas a partir de valores de probabilidades, de um modelo de distribuição uniforme, que são mapeados para valores z segundo a fdac da VA que representa o atributo considerado. A Figura 1 ilustra esse processo.

Esse processo é utilizado para a geração campos aleatórios, representados como grades regulares, que representam realizações da superfície do atributo numérico na região de interesse. ?? melhorar

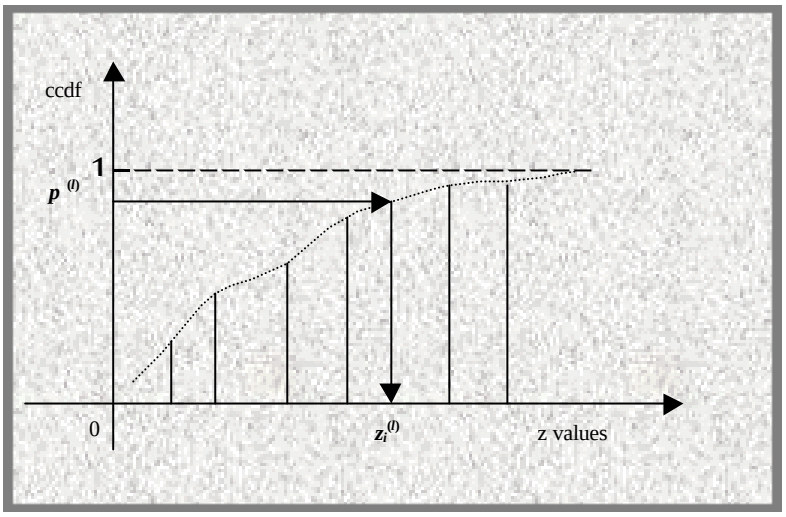

Figura 1: Ilustração do processo de obtenção de realizações a partir da fdac de uma VA

Estimativa de parâmetros da fdac a partir de realizações da VA

O conjunto de realizações, obtidas para um determinado nó dos campos aleatórios, pode ser usado para se determinar parâmetros estatísticos da fdac local na posição do nó.

A *média* $\boldsymbol{m}$ da fdac é calculada a partir da média simples de todas as realizações obtidas para o nó. A *variância* e o *desvio padrão* $\boldsymbol{s}$ da fdac são facilmente obtidos a partir de sua média $\boldsymbol{m}$ e dos valores realizados.

A *mediana* $\boldsymbol{q}_{.5}$ é determinada a partir da ordenação dos valores realizados $z(\mathbf{u})$ e posterior definição de um valor do atributo que divide o conjunto ordenado em dois subconjuntos de igual cardinalidade. Valores aproximadamente iguais da média e da mediana indicam similaridade da função de distribuição de probabilidade da VA. A mediana é um estimador mais robusto para distribuições de probabilidade não simétricas (Isaaks, 1989). O conjunto de realizações ordenado serve, ainda, para a determinação de outros quantis associados a da distribuição da VA.

A média e a mediana são tipicamente utilizados como valores estimados de atributos numéricos representados como variáveis aleatórias.

Incertezas das estimativas dos valores do atributo

Dado a fdac de uma VA é possível derivar-se vários *intervalos de probabilidade* que podem ser usados definir métricas de incertezas. Para atributos numéricos, as incertezas são comumente expressas por *intervalos de confiança*. Quando a fdac de uma VA $Z(\mathbf{u})$ apresenta alto grau de similaridade e sua distribuição tem um comportamento normal, é comum se representar as incertezas por intervalos de confiança Gaussiano $2\boldsymbol{s}$ e $4\boldsymbol{s}$, centrados na média, com intervalos de probabilidade do tipo:

$$Prob\{Z(\mathbf{u}) \in [m_Z(\mathbf{u}) \pm \sigma(\mathbf{u})]\} \approx 0.68$$

or

$$Prob\{Z(\mathbf{u}) \in [m_Z(\mathbf{u}) \pm 2\sigma(\mathbf{u})]\} \approx 0.95$$

onde $\boldsymbol{s}^2(\mathbf{u})=E\{(Z(\mathbf{u})-E\{Z(\mathbf{u})\})^2\}$ é a variância do atributo representado pela VA Z em $\mathbf{u}$.

Para distribuições não simétricas pode-se derivar intervalos de probabilidade baseado nos quantis da fdac da VA. Por exemplo, o intervalo de confiança $[q_{0.025};q_{0.975}]$ de 95% de probabilidade é dado por:

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

onde $q_{0.025}$ and $q_{0.975}$ sao os quantis 0.025 e 0.975 da fdac, ou seja, $F^*(\mathbf{u}; q_{0.025}|(n)) = 0.025$ e $F^*(\mathbf{u}; q_{0.975}|(n)) = 0.975$.

ESTUDO DE CASO

A Figura 2 apresenta a distribuição de um conjunto amostral de elevações utilizado para estudo de caso deste trabalho. A região escolhida é uma fazenda experimental da EMBRAPA (Empresa Brasileira de Pesquisas Agropecuárias do Brasil) chamada Canchim. A fazenda Canchim situa-se no município de São Carlos, no estado de Sao Paulo, Brasil, envolvida pelas coordenadas s 21^o 55' 00'' to s 21^o 59' 00'' e o 47^o 48' 00'' to o 41^o 52' 00''.

Na Figura 1, as amostras de altimetria estão superpostas ao mapa de altimetria (grade regular) obtido por inferência pelo valor da amostra vizinha mais próxima. Esse mapa serve de referência primária para se ter uma idéia da distribuição espacial do atributo em estudo.

Fig. 01

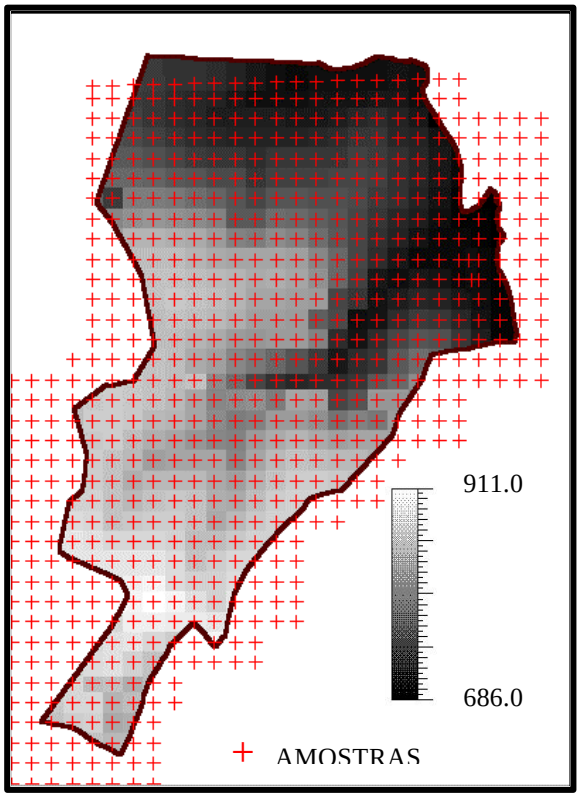

Figura 2: Distribuição das amostras de elevação superposta ao mapa de elevações obtido por interpolação por vizinho mais próximo.

Para se aplicar o procedimento de simulação sequencial, o conjunto amostral de altimetria foi transformado por indicação, definindo-se 10 novos conjuntos de amostras por indicação

segundo os seguintes valores de corte: 703.1 719.1 744.1 779.1 826.1 841.1 854.1 863.1 876.1, correspondentes aos decis (quantis de 10%) do conjunto amostral original de altimetria.

Modelos de variografia foram gerados para cada um dos conjuntos por indicação utilizando-se os procedimentos de análise exploratória, principalmente os procedimentos de geração de variograma de superfície e por indicação, do módulo de geoestatística do Sistema de Informação Geográfica SPRING (SPRING, 2000).

Os modelos de variografia foram inseridos no arquivo de parâmetros (sisim.par) do programa Sisim.exe da GSLIB (Deutsch, 1998), juntamente com os demais parâmetros necessários para se rodar o programa sisim.exe. A Figura 3 contém uma cópia desse arquivo de parâmetros.

Fig. 03

RESULTADOS E ANÁLISES

O programa sisim.exe gerou como saída um conjunto de 400 realizações de grades regulares, cada grade regular formada por 200 linhas x 200 colunas de valores de altimetria.

Esse conjunto de realizações foi, então, utilizado para se definir os mapas de inferências locais, de média e de mediana, do atributo altimetria utilizando-se as metodologias anteriormente descritas. Os mapas de valores médios e medianas de altimetria são apresentados na parte superior da Figura 4,.

Fig. 04

O conjunto de realizações foi, também, utilizado para se definir mapas de incertezas baseadas nos intervalos de confiança, $2s$ (68% de probabilidade) e $q_{.9}$ e $q_{.1}$ (80% de probabilidade), locais da distribuição da altimetria, utilizando-se as metodologias anteriormente descritas. Os mapas de incertezas são apresentados na parte inferior da Figura 4.

A partir da análise qualitativa, visual, dos mapas da Figura 4 podem-se destacar as seguintes considerações:

1. Os mapas de valores de altimetria, inferidos por média e por mediana, têm coerência com os valores das amostras utilizadas, uma vez que esses mapas tem aspectos gerais similares ao mapa de vizinhança mais próxima da Figura 2.

   Fig. 02

2. Apesar do aspecto geral ser similar, os mapas de média e de mediana são ligeiramente diferentes sugerindo uma assimetria nas funções de distribuição de probabilidades locais inferidas para os pontos da grade. Neste caso o mapa de medianas é um estimador mais robusto do que o mapa de médias.

3. Os mapas de incertezas estão de acordo com a variabilidade do atributo pois a maiores incertezas aparecem em regiões de maior variabilidade do atributo e também em faixas de transição. Por outro lado, os mapas de incertezas apresentam valores menores em áreas mais homogêneas de altimetria.. Para o dado de altimetria considerado, sugere-se utilizar um mapa de intervalos de confiança baseado em quantis, para representar as incertezas relacionadas com os valores de medianas inferidos.

```
                  Parameters for SISIM
                  **********************

START OF PARAMETERS:
1                           \1=continuous(cdf), 0=categorical(pdf)
9                           \number thresholds/categories
703.1 719.1 744.1 779.1 826.1 841.1 854.1 863.1 876.1 \   thresholds / categories
.1 .2 .3 .4 .5 .6 .7 .8 .9    \   global cdf / pdf
canchim.pts                  \file with data
1   2   0   3                \   columns for X,Y,Z, and variable
direct.ik                    \file with soft indicator input
1   2   0   3                \   columns for X,Y,Z, and indicators
0                           \   Markov-Bayes simulation (0=no,1=yes)
0.61  0.54  0.56  0.53  0.29 \      calibration B(z) values
-1.0e21    1.0e21            \trimming limits
687.5    911.0               \minimum and maximum data value
1      1.0                   \   lower tail option and parameter
1      1.0                   \   middle     option and parameter
1      1.0                   \   upper tail option and parameter
nenhum.dat                   \   file with tabulated values
3   0                        \      columns for variable, weight
2                           \debugging level: 0,1,2,3
sisimiso.dbg                 \file for debugging output
sisimiso.out                 \file for simulation output
20                           \number of realizations
100   204035.0    70.0       \nx,xmn,xsiz
100   7565050.0   100.0      \ny,ymn,ysiz
1    1.0   10.0              \nz,zmn,zsiz
69069                        \random number seed
4                           \maximum original data  for each kriging
12                           \maximum previous nodes for each kriging
1                           \maximum soft indicator nodes for kriging
1                           \assign data to nodes? (0=no,1=yes)
0     3                      \multiple grid search? (0=no,1=yes),num
0                           \maximum per octant    (0=not used)
2000.0  2000.0  0.0          \maximum search radii
 0.0   0.0   0.0             \angles for search ellipsoid
0    2.5                     \0=full IK, 1=median approx. (cutoff)
1                           \0=SK, 1=OK
1    0.02                    \One   nst, nugget effect
1    0.06 0.0   0.0   0.0       \      it,cc,ang1,ang2,ang3
     3172.0 3172.0 3172.0       \      a_hmax, a_hmin, a_vert
1    0.014                   \Two   nst, nugget effect
1    0.15  0.0   0.0   0.0      \      it,cc,ang1,ang2,ang3
     4874.0 4874.0 4874.0       \      a_hmax, a_hmin, a_vert
1    0.015                   \Three nst, nugget effect
1    0.228  0.0   0.0   0.0     \      it,cc,ang1,ang2,ang3
     5955.0 5955.0 5955.0       \      a_hmax, a_hmin, a_vert
1    0.011                   \Four  nst, nugget effect
1    0.202  0.0   0.0   0.0     \      it,cc,ang1,ang2,ang3
     4855.0 4855.0 4855.0       \      a_hmax, a_hmin, a_vert
1    0.01                    \Five  nst, nugget effect
1    0.218  0.0   0.0   0.0     \      it,cc,ang1,ang2,ang3
     4950.0 4950.0 4950.0       \      a_hmax, a_hmin, a_vert
1    0.026                   \Six   nst, nugget effect
1    0.2  0.0   0.0   0.0       \      it,cc,ang1,ang2,ang3
     5049.0 5049.0 5049.0       \      a_hmax, a_hmin, a_vert
1    0.03                    \Seven nst, nugget effect
1    0.162  0.0   0.0   0.0     \      it,cc,ang1,ang2,ang3
     4016.0 4016.0 4016.0       \      a_hmax, a_hmin, a_vert
1    0.024                   \Eight nst, nugget effect
1    0.123  0.0   0.0   0.0     \      it,cc,ang1,ang2,ang3
     3607.0 3607.0 3607.0       \      a_hmax, a_hmin, a_vert
1    0.014                   \Nine  nst, nugget effect
1    0.065  0.0   0.0   0.0     \      it,cc,ang1,ang2,ang3
     2061.0 2061.0 2061.0       \      a_hmax, a_hmin, a_vert
```

Figura 3: Parâmetros do programa Sisim.exe da GSLIB usados para simulação de altimetrias

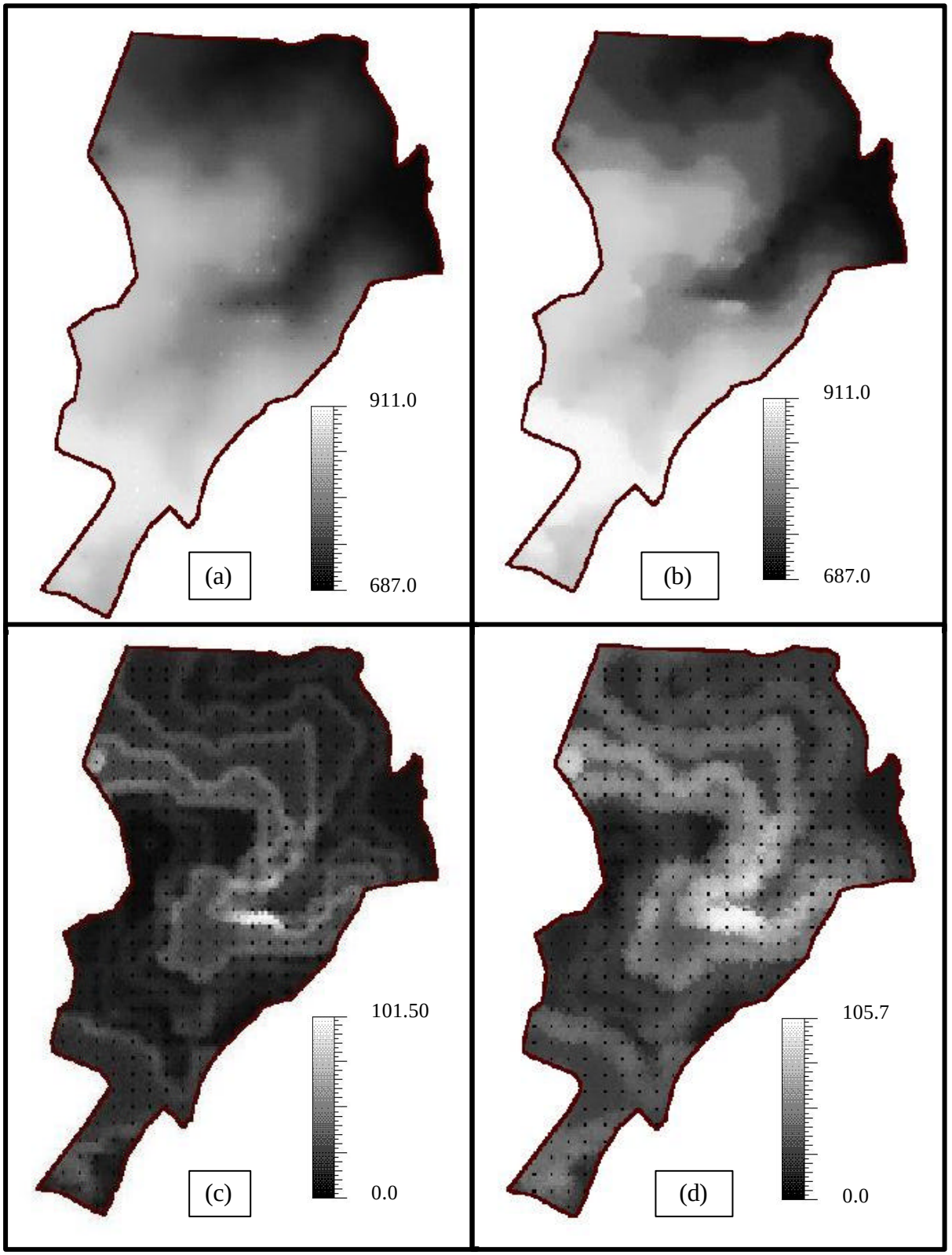

Figura 4: Ilustração dos mapas de médias (a), de medianas (b), de incertezas de $2s$ (c) e de incertezas por quantis .1 e .9 (d), relacionados com a espacialização do atributo altimetria na região de Canchim

CONCLUSÕES

A partir da aplicação das metodologias descritas e dos resultados obtidos neste trabalho conclui-se que o procedimento de simulação estocástica sequencial pode ser aplicado, com sucesso, para espacialização de atributos numéricos com as seguintes vantagens: reprodução do comportamento da variabilidade do atributo a partir da análise exploratória das amostras; inferência dos valores segundo critérios estatísticos baseados no comportamento da distribuição de probabilidade do atributo e; estimativa de incertezas associadas aos valores inferidos.

Ressalta-se, ainda, que o sucesso na utilização dessa metodologia de inferência depende da existência de um conjunto suficiente de amostras que possibilite a determinação do comportamento da variabilidade do atributo em estudo.

Resultados semelhantes aos aqui apresentados poderiam ser obtidos por krigeagem por indicação. Porém os resultados teriam menor variabilidade já que a krigeagem é um estimador de mínima variância. Outros estimadores determinísticos forneceriam resultados mais pobres pois não baseiam suas inferências no estudo prévio de variabilidade do atributo. Também as metodologias determinísticas não fornecem mapas locais de incertezas.

Por fim, destaca-se que esta metodologia, com pequenas adaptações, pode ser aplicada a atributos temáticos e que permite, ainda, a inclusão de dados indiretos, dados relacionados ao atributo em estudo, para melhorar a acurácia da inferência. Também, os dados simulados podem ser utilizados em modelagens de processos ambientais baseadas em simulações de Monte Carlo. Estes temas não foram aqui explorados mas devem fazer parte de futuros trabalhos.

BIBLIOGRAFIA

Burrough P. A. and McDonnell R. A. *Principles of Geographical Information Systems*, Oxford University Press, 1998.

Camargo E. C. G. *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.

Cowen D. J. "GIS versus CAD versus DBMS: what are the differences?", *Photogrammetric Engineering and Remote Sensing*, 54, (1988), 1551-1554.

De Oliveira J. L., Pires F. and Medeiros C. B., "An environment for modeling and design of geographic applications", *GeoInformatica*, 1, (1997), 29-58.

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

Felgueiras C. A. *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.

Felgueiras C. A., Monteiro A. M. V., Fuks S. D. 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.

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

Isaaks E. H. and Srivastava R. M. *An Introduction to Applied Geostatistics*, Oxford University Press, 1989.

SPRING V.3.4, (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/ , 2000.