4

ANÁLISE ESPACIAL DE SUPERFÍCIES: O ENFOQUE DA GEOESTATÍSTICA POR INDICAÇÃO

Carlos Alberto Felgueiras
Suzana Druck
Antônio Miguel Vieira Monteiro

4.1 Introdução

Os procedimentos de krigagem ordinária apresentados no capítulo anterior (vide Seção 3.4) buscavam predições ótimas da variável em estudo, em locais não observados, minimizando a variância do erro associado a essa estimativa. Neste capítulo, o foco será na análise de modelos de incerteza, ou seja, na inferência das distribuições de probabilidade para cada posição do espaço considerado, representadas pelos vetores $\boldsymbol{x}$. Os novos procedimentos vão permitir a definição de estimadores obtidos segundo a minimização de outras funções de erro inferencial, e não, como efetuado pela krigagem linear (vide Seção 3.5), um estimador baseado apenas na minimização da variância do erro. Situações em que a análise da incerteza é relevante podem ser ilustradas na aplicação da krigagem nos estudos de reposição de nutrientes nos solos. Neste caso, o que se deseja é determinar a quantidade de nutrientes que deve ser reposta nos solos de uma região de maneira a *maximizar* a produção e tornar *mínimo* os custos. O processo inferencial tem como objetivo evidenciar os locais em que um determinado fator dos solos, ***Z(x)***, é deficiente, ou seja, os locais em que o valor estimado, $\hat{Z}(\boldsymbol{x})$, seja igual ou abaixo de um valor crítico, z_{lim}, isto é, quando $\hat{Z}(\boldsymbol{x}) \leq z_{lim}$. Assim, o que interessa não é inferir exatamente um determinado valor, mas definir áreas com maior probabilidade que o evento ocorra, ou seja, áreas onde a probabilidade do valor estimado $\hat{Z}(\boldsymbol{x})$ ser menor ou igual a um limite z_{lim}, definida por $Prob\{\hat{Z}(\boldsymbol{x}) \leq z_{lim}\}$, tem um valor determinado.

Por outro lado, os erros inferenciais, que são a subestimação (estimar um valor menor do que seria o valor real) ou, a sobre-estimação (estimar um valor maior do que seria o valor real) vão produzir efeitos diferentes no processo produtivo. Enquanto a subestimação pode levar a repor nutrientes onde não é necessário, e contaminar os solos, a sobre-estimação pode conduzir a não repor nutrientes onde é necessário e prejudicar a produtividade. Dessa forma, esses erros inferenciais não podem ser tratados como se tivessem o mesmo impacto, e a minimização de um, ou de outro, ou de ambos, vai depender dos objetivos impostos pelo trabalho a ser executado. Neste contexto, o estimador de krigagem linear obtido pela

minimização da variância (vide Seção 3.5), que considera equivalentes e simétricos os impactos de subestimar ou sobre-estimar, seria insuficiente para apoiar as decisões necessárias a melhor solução do problema.

Este capítulo apresenta um conjunto de técnicas que procura construir o modelo de incerteza associado a uma determinada posição do espaço, representada pelo vetor $\boldsymbol{x}$. O modelo a ser produzido é condicionado a um conjunto de dados geográficos, coletados previamente a partir de suportes amostrais pontuais. Os exemplos, utilizados para ilustrar os conceitos deste capítulo, referem-se a conjuntos amostrais obtidos no levantamento de solos executado na região de Canchim (vide Seção 3.4, Figura 4-1 e Tabela 4-1). No que segue, admite-se que o leitor esteja familiarizado com os conceitos de krigagem apresentados no capítulo 3 (Seção 3.4 a Seção 3.7).

4.2 Incertezas locais

A geoestatística considera os valores de um atributo para cada posição $\boldsymbol{x} \in A$ (uma região da superfície terrestre) como uma realização de uma variável aleatória (VA), descrita como $Z(\boldsymbol{x})$. Isto significa que, na posição $\boldsymbol{x}$, $Z(\boldsymbol{x})$ pode assumir diferentes valores para o atributo considerado, cada valor com uma probabilidade de ocorrência associada a ele. Uma VA $Z(\boldsymbol{x})$ ordenada, contínua ou discreta, é caracterizada pela sua *função de distribuição de probabilidade acumulada, fdpa, univariada,* $F(\boldsymbol{x},z)$, definida como:

$$F(\boldsymbol{x};z) = Prob\{Z(\boldsymbol{x}) \leq z\} \tag{4.1}$$

Os procedimentos por indicação (também conhecidos por funções indicatriz) estão interessados na modelagem da *função de distribuição univariada acumulada condicionada* (*fdpac*), isto é, a *função de distribuição* que pode ser construída condicionada aos *n* dados amostrados, $F(\boldsymbol{x};z\,|\,(n))$, que é dada por:

$$F(\boldsymbol{x};z\,|\,(\mathrm{n})) = Prob\{Z(\boldsymbol{x}) \leq z\,|\,(n)\} \tag{4.2}$$

A $F(\boldsymbol{x};z\,|\,(n))$ modela a incerteza da V.A. $\boldsymbol{Z}$ no local $\boldsymbol{x}$, e uma vez estimada essa função de distribuição de probabilidade ela pode ser utilizada para:

- produzir uma estimativa de valores do atributo em posições não conhecidas;
- modelar a incerteza dos valores para o atributo nas posições estimadas;

O enfoque tradicional, oferecido pela krigagem linear, para modelar a incerteza em locais não amostrados , consiste em computar estimativas do valor desconhecido $\hat{z}(\boldsymbol{x})$ e de sua respectiva variância $\hat{\sigma}^2(\boldsymbol{x})$, e construir um intervalo de confiança do tipo gaussiano, centrado em $\hat{z}(\boldsymbol{x})$,

$$Prob\{Z(\boldsymbol{x}) \in [\hat{z}(\boldsymbol{x}) - 2\hat{\sigma}(\boldsymbol{x}),\ \hat{z}(\boldsymbol{x}) + 2\hat{\sigma}(\boldsymbol{x})]\} \quad (4.3)$$

A construção deste tipo de intervalo de confiança fundamenta-se nas hipóteses:

- os erros locais de estimação têm distribuição gaussiana;
- o intervalo de confiança pode ser construído através da variância destes erros.

Essas hipóteses são fortemente restritivas, uma vez que a distribuição local dos erros pode apresentar severas assimetrias, principalmente quando o histograma das amostras apresenta-se assimétrico, não se adequando a hipótese gaussiana sendo implicitamente considerada. Por outro lado, a variância obtida através da krigagem linear depende unicamente da configuração geométrica dos dados, e não do valor de seu atributo naquela posição, e uma variância com essas características pode não ser adequada para representar as incertezas na estimativa de valor para o atributo, principalmente em regiões onde amostras próximas apresentam valores para seu atributo, medido ou observado, muito discrepantes.

Um outro enfoque possível é considerar que primeiro é necessário modelar a incerteza, ou seja inferir as distribuições de probabilidades locais, as distribuições para cada ponto do espaço a ser estudado, representado pelo vetor $\boldsymbol{x}$. Uma vez estabelecidas as funções, $F(\boldsymbol{x}; z \mid (n))$, e só então deduzir as estimativas ótimas para cada ponto. Observe que o procedimento tradicional primeiro calcula a estimativa, os valores estimados para os pontos não observados, e depois acrescenta o intervalo de confiança, com base na variância dos erros produzidos pelo estimador. A modelagem da incerteza, sendo construída diretamente através da *fdpac*, $F(\boldsymbol{x}; z \mid (n))$, condiciona, por construção, essa *fdpac* aos dados amostrais, e produz então um modelo que é independente de uma particular estimativa $\hat{z}(\boldsymbol{x})$, obtida com base em um particular estimador, no nosso caso o estimador por krigeagem linear. Ficamos agora com o problema da inferência desta função de distribuição de probabilidade acumulada condicionada para cada ponto do espaço, da $F(\boldsymbol{x}; z \mid (n))$.

Vamos abordar dois enfoques, mais presentes na literatura :

- O *multigaussiano*, que estabelece o modelo de distribuição a ser considerado à priori;
- O *enfoque por indicação*, que não estabelece nenhum modelo de distribuição para os dados. A *fdpac* é modelada de forma aproximada pela sua discretização numa série de K cortes $z_k, k = 1,....\text{k}$.

O primeiro enfoque, o *multigaussiano*, é o mais fácil de ser utilizado, mas apresenta algumas restrições importantes:

1. estabelece a hipótese multigaussiana para a distribuição multivariada que não pode ser inteiramente verificada;
2. é inadequada para fenômenos que apresentam uma expressiva correlação em valores extremos da distribuição.

O *enfoque por indicação* pode ser considerado mais geral. Não restringe o fenômeno em estudo a ser representado por uma distribuição específica. Deve ser utilizado quando os dados não se ajustam a uma distribuição multigaussiana, ou quando os valores extremos da distribuição das amostras apresentam significante conectividade. Esse capítulo, por essas razões, focaliza esse procedimento.

4.3 O Enfoque por Indicação

O enfoque por indicação está fundamentado na interpretação da probabilidade condicional $Prob\{Z(\boldsymbol{x}) \in [\hat{z}(\boldsymbol{x}) - 2\hat{\sigma}(\boldsymbol{x}), \hat{z}(\boldsymbol{x}) + 2\hat{\sigma}(\boldsymbol{x})]\}$ como uma esperança condicional de uma variável aleatória por indicação, $I(\boldsymbol{x}, z_k|(n))$, considerada a informação disponível nas *(n)* amostras, isto é:

$$F(\boldsymbol{x}; z_k \,|\,(n)) = E\{I(\boldsymbol{x}, z_k)\,|\,(n)\} \quad k = 1,...,K \tag{4.4}$$

onde $I(\boldsymbol{x}, z_k(\mathrm{n})) = 1 \; se \; Z(\boldsymbol{x}) \leq z_k$ **e** $I(\boldsymbol{x}, z_k(\mathrm{n})) = 0 \; se \; Z(\boldsymbol{x}) > z_k$

A estimativa de krigagem de uma variável por indicação, $I(\boldsymbol{x}, z_k|(n))$, é também uma estimativa de sua esperança condicional. Portanto, as estimativas de $\hat{F}(\boldsymbol{x}, z_k|(n)), \; para \; k = 1,...,\mathrm{K}$, podem ser calculadas estimando-se o valor $\hat{i}(\boldsymbol{x}, z_k|(n))$, que utiliza para sua estimativa os dados transformados para dados por indicação, com valores *1* e *0*.

Dessa forma, os procedimentos por indicação iniciam-se por uma transformação não linear, chamada de *codificação por indicação*, que transforma cada valor do conjunto amostral, $z(\boldsymbol{x})$, em valores por indicação, $i(\boldsymbol{x}, z_k)$.

A codificação por indicação dos dados amostrais

Na distribuição de um conjunto de dados amostrais, um determinado número de cortes K e seus respectivos valores de cortes $z_k, k = 1,....\mathrm{k}$, são definidos. A codificação por indicação, se processa para cada valor de corte z_k, e gera um *conjunto amostral por indicação* $i(\boldsymbol{x}, z_k)$ do tipo:

$$i(\boldsymbol{x}; z_k) = \begin{cases} 1, & \text{se } z(\boldsymbol{x}) \le z_k \\ 0, & \text{se } z(\boldsymbol{x}) > z_k \end{cases} \tag{4.5}$$

A codificação por indicação é aplicada sobre todo conjunto amostral criando, para cada valor de corte, um conjunto cujos valores são 0 ou 1. Os K valores de corte, são definidos em função do número de amostras e devem ser escolhidos de tal forma que os $K+1$ cortes contenham aproximadamente as mesmas frequências. Entretanto, existem algumas critérios para a escolha de *K:*

1. Os valores de k, devem ser representativos de toda a gama de valores apresentados pelos dados.
2. Os valores de k devem destacar os pontos importantes da distribuição.
3. O número de cortes *K* não deve ser muito grande, o que demandaria grande esforço computacional, mas principalmente não deve ser muito pequeno, pois pode resumir aspectos relevantes da distribuição. Uma regra razoável é considerar que o valor de *K* não deve ser menor que cinco (5), nem maior que quinze (15).

Se para um determinado conjunto de dados cujos valores variam no intervalo $[5, 43]$ podemos definir $z_k = 20, 30, 39$ correspondentes respectivamente a três quantis de sua distribuição ($p = 0.25, 0.50, 0.75$). A codificação associará a cada valor amostral um vetor com 3 dados por indicação com valores 0 ou 1. Por exemplo, se o valor amostral $z(u_j) = 25.2$, então o valor por indicação $i(u_j, 20) = 0$ e representa a probabilidade de $Z(u_j)$ assumir valores menores ou iguais a *20*, dado que o valor de $z(u_j) = 25{,}2$, $Prob\left[Z(u_j) \le 20 \middle| z(u_j) = 25.2\right]$. Considerando os três valores de z_k, o vetor por indicação seria representado como abaixo descrito:

$$\begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} i(\boldsymbol{x}_j, 20) \\ i(\boldsymbol{x}_j, 30) \\ i(\boldsymbol{x}_j, 39) \end{bmatrix} \rightarrow \begin{bmatrix} Prob\{Z(\boldsymbol{x}_j) \le 20 | z(\boldsymbol{x}_j) = 25.2\} \\ Prob\{Z(\boldsymbol{x}_j) \le 30 | z(\boldsymbol{x}_j) = 25.2\} \\ Prob\{Z(\boldsymbol{x}_j) \le 39 | z(\boldsymbol{x}_j) = 25.2\} \end{bmatrix}$$

4.3.2 A variografia por indicação

A análise de variografia se processa de forma semelhante a realizada na krigeagem linear (vide Seção 3.5), considerando-se separadamente o conjunto de valores por indicação para cada valor de corte, z_k. Dessa forma, para cada valor de corte z_k um modelo de variograma deve ser estabelecido, o que corresponderia, no exemplo anterior, ao ajuste de 3 modelos de semivariogramas a partir de 3 variogramas experimentais computados como:

$$\gamma_I(\boldsymbol{h}, z_k) = \frac{1}{2N(\boldsymbol{h})} \sum_{\alpha=1}^{N(h)} [i(\boldsymbol{h}_\alpha; z_k) - i(\boldsymbol{h}_\alpha + \boldsymbol{h}; z_k)]^2 \quad (4.6)$$

Como os valores das variáveis por indicação são *0* e *1*, o variograma por indicação é, usualmente, bem comportado e resistente a valores extremos ("outliers"). Também as amostras de $i(u, z_k)$ para cada z_k são considerados como amostras de uma distribuição Bernouilli cuja variância máxima é 0.25. Dessa forma o efeito pepita somado ao patamar, que são aproximadamente iguais ao valor da variância, terá como valor máximo 0.25. Calcular os variogramas é relativamente simples, sendo a única dificuldade prática o número de variogramas a ser modelados.

4.3.3 A estimação dos valores por indicação

Como mencionado anteriormente para cada valor de corte $z_k, k = 1,....\text{k}$, a $\hat{F}(\boldsymbol{x}, z_k | (n))$ pode ser estimada através da combinação linear dos dados por indicação $i(\boldsymbol{x}, z_k)$. O estimador linear é expresso em termos de VAs por indicação.

$$\hat{F}(\boldsymbol{x}; z_k | (n)) = \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha(\boldsymbol{x}; z_k) I(\boldsymbol{x}_\alpha; z_k) + \left[1 - \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha(\boldsymbol{x}; z_k)\right] I(\boldsymbol{x}; z_k) \quad (4.7)$$

onde $\lambda_\alpha(\boldsymbol{x}; z_k)$ é o peso assinalado a cada dado convertido interpretado como uma realização de uma variável aleatória por indicação. Se a média por indicação, $E[I(\boldsymbol{x}; z_k)]$, é considerada constante dentro da área em estudo dois procedimentos podem ser considerados, descritos a seguir.

Krigeagem por Indicação Simples

Neste caso a média por indicação é conhecida e constante, isto é:

$$E\{I(\boldsymbol{x}; z_k)\} = F(z_k) \quad (4.8)$$

e o preditor linear (4.6) é então rescrito,

$$\hat{F}_{KS}(\boldsymbol{x}; z_k | (n)) = \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha^{KS}(\boldsymbol{x}; z_k) I(\boldsymbol{x}_\alpha; z_k) + \left[1 - \sum_{\alpha=1}^{n(\boldsymbol{x})} \lambda_\alpha^{KS}(\boldsymbol{x}; z_k)\right] F(z_k) \quad (4.9)$$

onde os pesos $\lambda_\alpha^{KS}(\boldsymbol{x}, z_k)$ são determinados através do sistema de krigeagem simples.

$$\sum_{\beta=1}^{n(\boldsymbol{u})} \lambda_\beta^{KS}(\boldsymbol{x}; z_k) C_I(\boldsymbol{h}_{\alpha\beta}; z_k) = C_I(\boldsymbol{h}_\alpha; z_k) \quad \forall \alpha = 1,2,...,n(\boldsymbol{x}) \quad (4.10)$$

onde $\boldsymbol{h}_{\alpha\beta}$ é o vetor de separação definido pelas posições $\boldsymbol{x}_\alpha$ e $\boldsymbol{x}_\beta$, $\boldsymbol{h}_\alpha$ é o vetor definido entre as posições $\boldsymbol{x}_\alpha$, e a posição a ser estimada $\boldsymbol{x}_0$, $C_I(\boldsymbol{h}_{\alpha\beta}; z_k)$ é a

autocovariância definida por $\boldsymbol{h}_{\alpha\beta}$ e $C_I(\boldsymbol{h}_\alpha; z_k)$ é a autocovariância definida por $\boldsymbol{h}_\alpha$ em $z = z_k$. As autocovariâncias são determinadas pelo modelo de variografia teórico definido pelo conjunto I para $z = z_k$.

Krigeagem por Indicação Ordinária

A krigeagem por indicação ordinária permite considerar flutuações locais da média limitando seu domínio de estacionariedade a vizinhança local $W(\boldsymbol{x})$

$$E\{I(\boldsymbol{x}; z_k)\} = \text{ constante mas desconhecida para } \forall \boldsymbol{x} \in W(\boldsymbol{x})$$

$$E\{I(\boldsymbol{x}; z_k)\} = \hat{F}(\boldsymbol{x}; z_k) \text{ estimado no domínio } W(\boldsymbol{x})$$

O estimador de *krigeagem por indicação ordinária* tem a seguinte expressão:

$$\hat{F}_{KS}(\boldsymbol{x}; z_k \mid (n)) = \sum_{\alpha=1}^{n(\boldsymbol{x})} \lambda_\alpha^{KS}(\boldsymbol{x}; z_k) I(\boldsymbol{x}_\alpha; z_k) + \left[1 - \sum_{\alpha=1}^{n(\boldsymbol{x})} \lambda_\alpha^{KS}(\boldsymbol{x}; z_k)\right] \hat{F}(\boldsymbol{x}; z_k) \qquad (4.11)$$

sendo que os pesos $\lambda_\alpha^{KS}(\boldsymbol{x}, z_k)$ são obtidos pela solução do seguinte sistema de equações de krigagem por indicação ordinária:

$$\begin{cases} \sum_{\beta=1}^{n(\boldsymbol{x})} \lambda_\beta^{KO}(\boldsymbol{x}; z_k) C_I(\boldsymbol{h}_{\alpha\beta}; z_k) + \phi(\boldsymbol{x}; z_k) = C_I(\boldsymbol{h}_\alpha; z_k) \quad \forall \alpha = 1,2,\ldots,n(\xi) \\ \sum_{\beta=1}^{n(\boldsymbol{x})} \lambda_\beta^{KO}(\boldsymbol{x}; z_k) = 1 \end{cases} \qquad (4.12)$$

onde $\phi(\boldsymbol{x}; z_k)$ é o multiplicador de Lagrange.

A krigagem por indicação, simples ou ordinária, fornece, para cada valor de corte z_k , a melhor estimativa da esperança condicional da VA $I(\boldsymbol{x}, z_k)$, $\hat{I}(\boldsymbol{x}, z_k)$. Utilizando esta propriedade, e o teorema que estabelece que $\hat{I}(\boldsymbol{x}, z_k) = \hat{F}(\boldsymbol{x}, z_k)$ pode-se calcular estimativas dos valores da *fdpac* de $Z(\boldsymbol{x})$ para vários valores de $z = z_k$, pertencentes ao domínio de $Z(\boldsymbol{x})$. O conjunto dos valores das *fdpac's*, estimados nos valores de corte, é considerado uma *aproximação discretizada* da *fdpac* real de $Z(\boldsymbol{x})$. Quanto maior a quantidade de valores de corte, melhor é a aproximação. A aproximação é complementada pela definição de *uma função de ajuste para a distribuição*, que deve ser utilizada para se inferir a *fdpac* para valores diferentes dos valores de corte. Um ajuste linear é o mais simples de se definir, porém funções de maior grau podem ser usadas.

4.3.4 Correção dos Desvios de Ordem

A aproximação da função de distribuição apresenta alguns problemas, conhecidos como *desvios de relação de ordem*, que devem ser corrigidos automaticamente pelo procedimento. Os valores de probabilidades acumuladas condicionadas, para cada valor de corte, são inferidos independentemente. Para que esses valores de probabilidade constituam uma distribuição legítima, eles devem verificar as seguintes relações de ordem:

1. Os valores inferidos de $\hat{F}(\boldsymbol{x}, z_k|(n))$ devem satisfazer a seguinte relação $0 \le F^*(\boldsymbol{x}; z_k \mid (n)) \le 1 \quad \forall z_k,\ k = 1,...,K$
2. O valor estimado de $\hat{F}(\boldsymbol{x}, z_k|(n))$ não deve ser maior do que a $\hat{F}(\boldsymbol{x}; z_{k+1}|(n))$ quando $z_k \le z_{k+1}$, ou seja $\hat{F}(\boldsymbol{x}; z_k \mid (n)) \le \hat{F}(\boldsymbol{x}; z_{k+1} \mid (n))$ *se* $z_k \le z_{k+1}$

A primeira condição pode ser garantida quando todos os pesos do estimador são positivos e somam 1. A krigeagem não garante que os pesos sejam todos positivos. Por isso é possível a inferência de valores da *fdpac* fora do intervalo [0,1]. A solução para este problema é ajustar os valores estimados para as bordas, ou seja, valores negativos são mapeados para 0 e valores maiores que 1, para 1. A segunda condição é garantida com o uso de ponderadores positivos que somam 1, e com a utilização dos mesmos pesos de estimação para todos os valores de corte, o que não pode ser garantido pela krigeagem por indicação. Portanto, estas inconsistências podem ocorrer e devem ser corrigidas. Um procedimento simples de correção é verificar pares de *fdac's* estimadas, em valores sucessivos de cortes, e ajustá-los para o valor médio das duas, sempre que a relação de ordem, $\hat{F}(\boldsymbol{x}; z_k \mid (n)) \le \hat{F}(\boldsymbol{x}; z_{k+1} \mid (n))$ *se* $z_k \le z_{k+1}$**,** não for satisfeita. A Figura 4-2 ilustra os problemas e as soluções das 2 condições acima descritas.

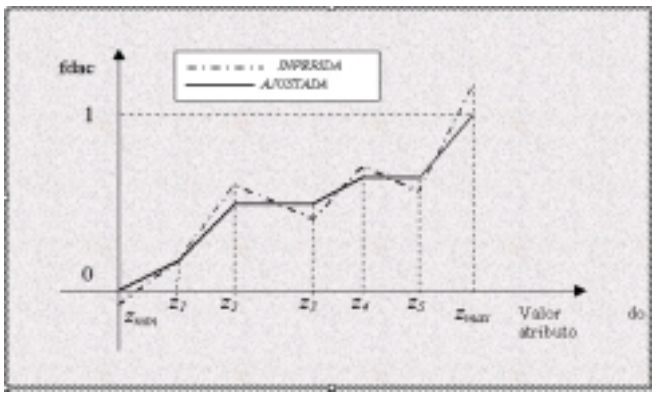

Figura 4-2 Correção dos desvios de relação de ordem

A Figura 4-3 e a Figura 4-4 que seguem buscam ilustrar as etapas descritas para a obtenção do modelo de incerteza para um conjunto amostral tomado conceitualmente como variáveis aleatórias.

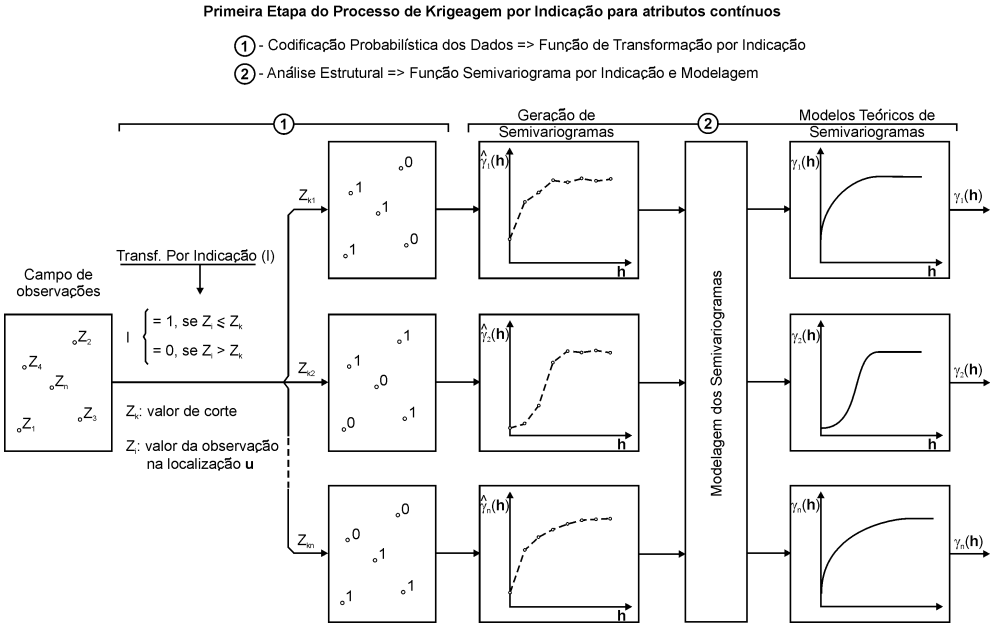

Figura 4-3 Primeira etapa do processo de krigagem por Indicação

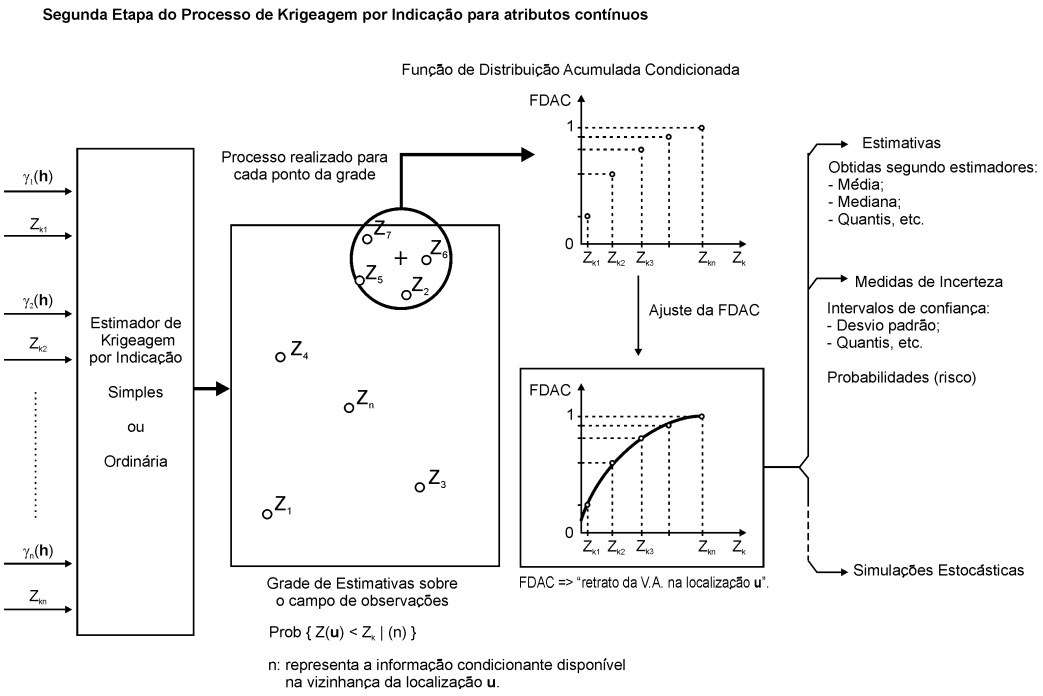

Figura 4-4 Segunda etapa do processo de krigagem por Indicação

4.4 Estimativa de incertezas locais

O conhecimento da *fdpac*, $F(\boldsymbol{x}; z_k|(n))$, em uma localização $\boldsymbol{x}$, possibilita a estimativa direta da incerteza, sobre o valor não conhecido $z_k(\boldsymbol{x})$, independente da escolha de um estimador para $z_k(\boldsymbol{x})$. Vamos ver agora como a incerteza pode ser estimada quando adotamos o enfoque por indicação aqui apresentado.

Intervalos de probabilidade

A incerteza pode ser estimada através de intervalos de valores do atributo. A probabilidade de um valor $z_k(\boldsymbol{x})$ estar dentro de um intervalo $(a,b]$ qualquer, chamado *intervalo de probabilidade*, é computado como a diferença entre os valores da *fdpac* para os limiares *b* e *a*, ou seja:

$$Prob\{Z(\boldsymbol{x}) \in (a,b]|(n)\} = F(\boldsymbol{x}; b|(n)) - F(\boldsymbol{x}; a|(n)) \quad (4.13)$$

Um intervalo de probabilidade dado por $Prob\{Z(\boldsymbol{x}) \in (a,b]|(n)\} = 0.7$, significa que $z(\boldsymbol{x})$ tem 70% de chance de estar dentro e, portanto, 30% de chance de estar fora do intervalo (a, b]. Quando $b = \infty$ obtêm-se a probabilidade de se exceder um limiar *a*, ou seja:

$$Prob\{Z(\boldsymbol{x}) \in (a,+\infty]|(n)\} = Prob\{Z(\boldsymbol{x}) > a|(\boldsymbol{n})\} = 1 - F(\boldsymbol{x}; a|(n)) \quad (4.14)$$

Esta probabilidade é particularmente importante em aplicações ambientais focadas em medir os riscos de se exceder limites regulatórios. Para exemplificar a utilização dessas medidas de incerteza, numa situação real, considere o conjunto amostral de altimetria de Canchim, apresentado na Figura 4-5. Esse conjunto amostral foi utilizado como entrada para produção do mapa temático de altimetria e do mapa de incertezas apresentados na Figura 4-6 (a) e (b), respectivamente.

A classificação apresentada no mapa da Figura 4-6(a) foi obtida a partir dos modelos de distribuição probabilística inferidos pelo procedimento de krigeagem por indicação condicionado às amostras de altimetria. Neste caso, foram definidas 3 faixas distintas de valores de altimetria, 3 classes, e para cada ponto desse mapa, as probabilidades de pertinência a cada um dos intervalos de valores, definidos para as classes, foram calculadas pela formulação apresentada na equação 4.13. Para classificação de cada ponto do mapa temático de altimetria, utilizou-se o critério de máxima probabilidade, ou seja, atribuiu-se, a cada ponto do mapa, a classe de maior probabilidade de ocorrência nesse local. Os valores de incerteza apresentados na Figura 4-6(b), mapa da direita, foram calculados a partir do valor da probabilidade da classe que foi associada a cada ponto do mapa temático de altimetria gerado. Assim, calculou-se a incerteza como:

$$Inc(\boldsymbol{x}) = 1 - Prob\{z(\boldsymbol{x}) \in s_k(\boldsymbol{x}), k = 1, 2 \text{ ou } 3\} \quad (4.15)$$

onde $s_k(\boldsymbol{x})$ é a classe atribuída a localização $(\boldsymbol{x})$.

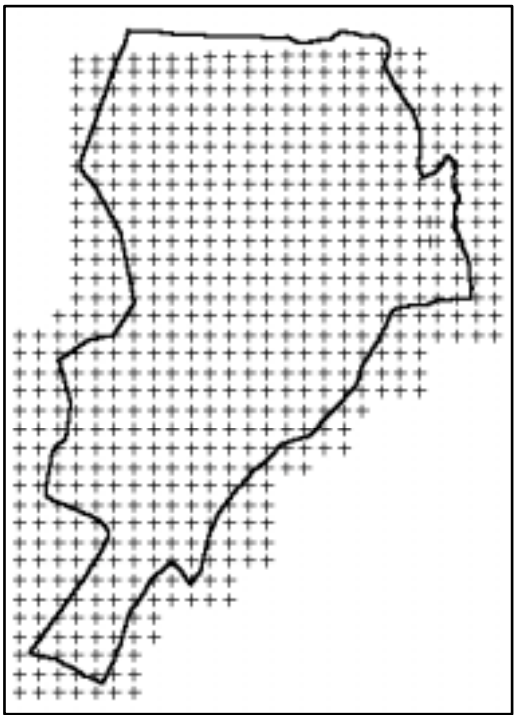

Figura 4-5 Distribuição espacial das amostras de altimetria na região de Canchim

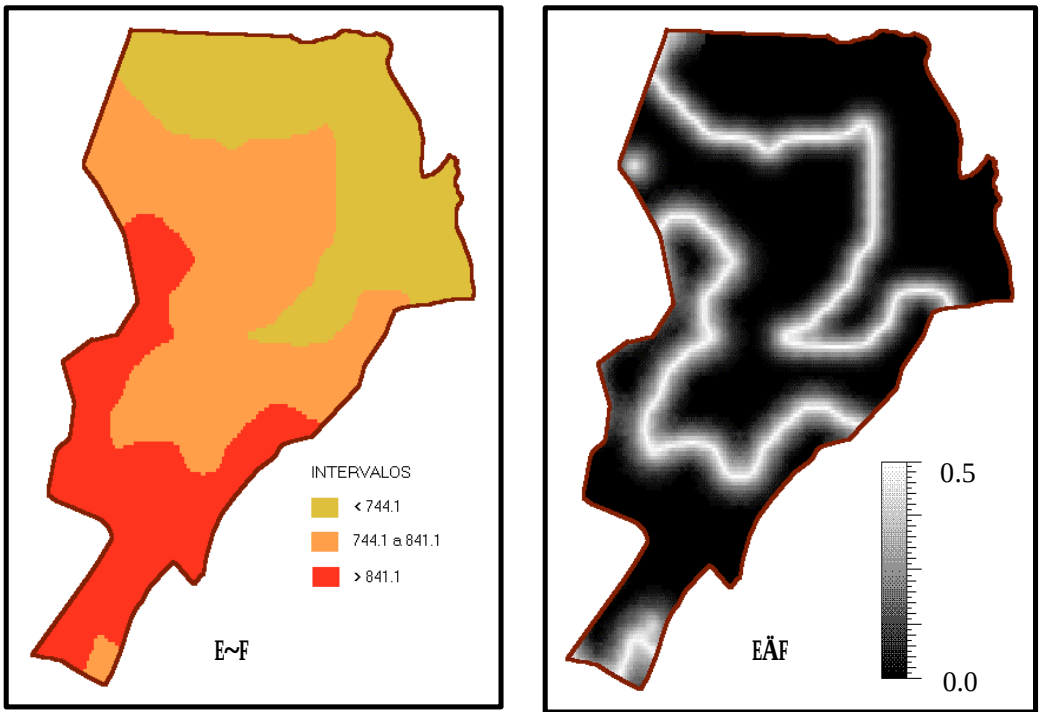

Figura 4-6 Mapa temático de altimetria *(a)* e respectivas medidas de incerteza *(b)*

Distância interquantil

Uma medida mais robusta de espalhamento é um intervalo interquantil. Por exemplo, o intervalo interquartil, $q_R(\boldsymbol{x})$ é definido por:

$$q_R(\boldsymbol{x}) = q_{0.75}(\boldsymbol{x}) - q_{0.25}(\boldsymbol{x}) = F^{-1}(\boldsymbol{x}; 0.75 \mid (n)) - F^{-1}(\boldsymbol{x}; 0.25 \mid (n)) \quad (4.16)$$

Para distribuições altamente assimétricas, uma medida mais robusta é o intervalo interquantil, que é definido como a diferença entre dois quantis, simétricos em relação a mediana. A partir da função de distribuição acumulada condicionada inferida, $\hat{F}(\boldsymbol{x}; z|(n))$, pode-se derivar vários intervalos de probabilidade tais como o intervalo 95%, $[q_{0.025}; q_{0.975}]$, tal que:

$$Prob\{Z(\boldsymbol{x}) \in [q_{0.025}; q_{0.975}] | (n)\} = 0.95 \quad (4.17)$$

com $q_{0.025}$ e $q_{0.975}$ sendo os quantis relativos aos valores de probabilidade da *fdpac* 2,5% e 97.5%, ou seja, $F^*(\boldsymbol{x}; q_{0.025}|(n)) = 0.025$, e $F^*(\boldsymbol{x}; q_{0.975}|(n)) = 0.975$. Os valores do atributo, referentes aos quantis, são estimados a partir da função de ajuste e dos valores de corte usados na krigeagem por indicação. Um mapa de incertezas obtido pelos valores de uma grade de intervalos interquartis, diferença entre o primeiro e o terceiro quartil de altimetria, e estimados segundo a equação 4.16, está apresentado na Figura 4-7.

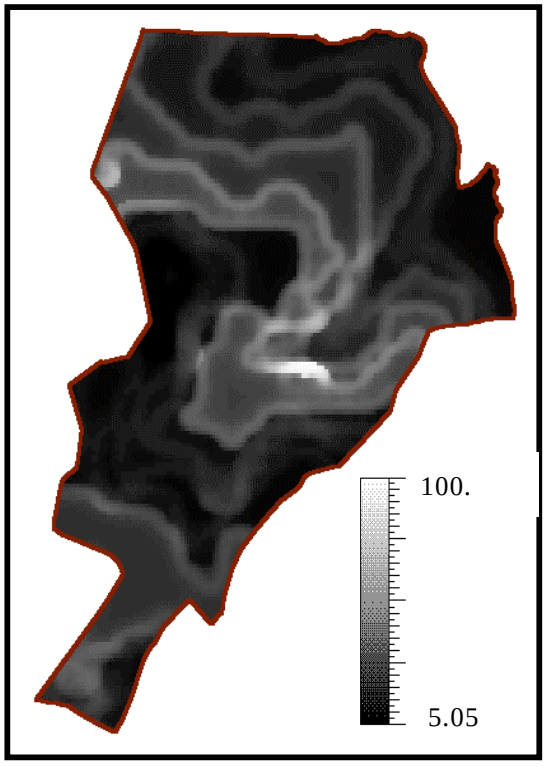

Figura 4-7 Mapa de incertezas locais obtido a partir dos quartis, primeiro e terceiro, dos modelos de distribuição probabilística locais inferidos pela krigagem por indicação

Variância condicional

Uma medida importante de espalhamento de uma distribuição é a variância condicional que mede os desvios da *fdpac* em torno da média da distribuição, $\bar{z}_{Z_k}(u)$. Diferente das medidas de incerteza anteriormente descritas, esta necessita da estimação da média da distribuição, isto é, da definição desse estimador. É possível obter-se uma estimativa da variância da distribuição condicionada, $\hat{\sigma}^2(\boldsymbol{x})$, pela seguinte formulação:

$$\begin{aligned}\left(\hat{\sigma}^2\right)(\boldsymbol{x}) &= \int_{-\infty}^{\infty}\left[z-\bar{z}_{Z_k}(\boldsymbol{x})\right]^2 dF(\boldsymbol{x};z\mid(n)) \\ &\approx \sum_{k=1}^{K+1}\left[z_k^{'}-\bar{z}_{Z_k}(\boldsymbol{x})\right]^2\left[\hat{F}(\boldsymbol{x};z_k\mid(n))-\hat{F}(\boldsymbol{u};z_{k-1}\mid(n))\right]\end{aligned} \tag{4.18}$$

onde $\bar{z}_{z_k}$ é o valor da média da classe $(z_{k-1},z_k]$.

A Figura *4-8* apresenta um mapa de variâncias para os valores de altimetria, da região de Canchim, obtidas pela equação 4.18.

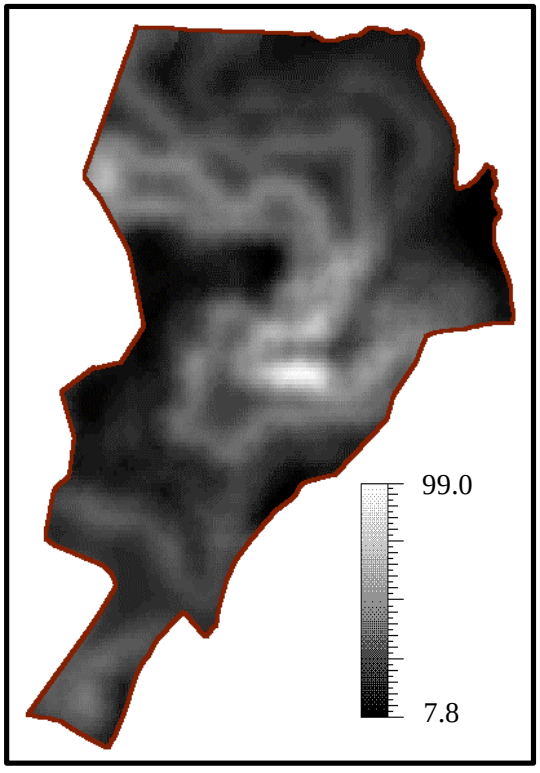

Figura 4-8 Mapa de incertezas locais obtido a partir das variâncias inferidas dos modelos de distribuição probabilística construídos pela krigeagem por indicação.

Entropia de Shannon

Uma medida de incerteza local, não relacionada a qualquer intervalo $(a, b]$, é dada pela medida de entropia da função de densidade de probabilidade local. Essa medida é definida como:

$$H(\boldsymbol{x}) = \int_{-\infty}^{\infty} -\left[ln\ f(\boldsymbol{x}; z|(n))\right] \cdot f(\boldsymbol{x}; z|(n))dz \tag{4.19}$$

onde $f(\boldsymbol{x}; z|(n)) = \partial F(\boldsymbol{x}; z|(n))/\partial z$ é a função de distribuição de probabilidade. Na prática a amplitude de variação de z é discretizada em K classes, que não se interceptam, $(z_{k-1}, z_k]$, computando-se a probabilidade desses K intervalos como:

$$p_k(\boldsymbol{x}) = \left[\hat{F}(\boldsymbol{x}; z_k \mid (n)) - \hat{F}(\boldsymbol{x}; z_{k-1} \mid (n))\right] \tag{4.20}$$

A entropia para a distribuição condicional em $\boldsymbol{x}$ é computada como:

$$H(\boldsymbol{x}) \cong -\sum_{k=1}^{K} \left[ln(p_k(\boldsymbol{x}))\right] \cdot p_k(\boldsymbol{x}) \geq 0, \quad \forall p_k \neq 0 \tag{4.21}$$

4.5 Estimadores Ótimos para as Superfícies Interpoladas

O processo inferencial visa calcular uma estimativa do valor de $z(\boldsymbol{x})$ através de um estimador que é caracterizado por uma determinada função dos dados. Esse estimador, no que concerne aos objetivos do processo inferencial, deve minimizar algum tipo de erro que se deseja evitar, maximizando os acertos de interesse. Por essa razão, um estimador é dito ótimo quando minimiza perdas, isto é, uma particular função dos erros inferenciais, $L(\varepsilon)$, onde $\varepsilon = z(\boldsymbol{x}) - \hat{z}(\boldsymbol{x})$. Entretanto, minimizar $L(\varepsilon)$ significa conhecer $z(\boldsymbol{x})$, que é desconhecido. Portanto, a idéia é utilizar o modelo de incerteza definido para determinar a perda esperada, $E[L(\varepsilon)]$.

$$\begin{aligned} E[L(\varepsilon)] &= E\{L(\varepsilon(\boldsymbol{x}))|(n)\} \\ &= \int_{-\infty}^{+\infty} L(\varepsilon(\boldsymbol{x}))dF(\boldsymbol{x}, z|(n)) \end{aligned} \tag{4.22}$$

Na prática, a seguinte aproximação é utilizada

$$E[L(\varepsilon)] \cong \sum_{k=1}^{K+1} L(\hat{z}(\boldsymbol{x}) - \bar{z}_k)\left[\hat{F}(\boldsymbol{x}, z_k|(n)) - \hat{F}(\boldsymbol{x}, z_{k-1}|(n))\right] \tag{4.23}$$

Assim sendo a determinação de estimativas ótimas se processa em duas etapas:

1. A incerteza sobre o valor desconhecido $z(\boldsymbol{x})$ é inicialmente modelada pela sua *fdpac* $\hat{F}(\boldsymbol{x}, z_k|(n))$;

2. Desse modelo uma estimativa de $\hat{z}(\boldsymbol{x})$ é obtida tal que minimiza $E[L(\varepsilon)]$.

Estimativa do valor esperado

A estimativa do valor esperado para cada valor espacial da distribuição é realizada a partir do de mínimos quadrados onde $L[\varepsilon(u)] = [\varepsilon(u)]^2$. Mostra-se que essa função é minimizada quando z é o valor esperado, $\hat{z}(\boldsymbol{x}) = z_E(\boldsymbol{x})$. A estimativa do valor esperado, $\hat{z}_E(\boldsymbol{x}) = E\{Z(\boldsymbol{x})\}$ onde:

$$E[Z(\boldsymbol{x})] = \int_{-\infty}^{\infty} z \cdot f(\boldsymbol{x}; z \mid (n)) dz = \int_{-\infty}^{\infty} z \cdot dF(\boldsymbol{x}; z \mid (n)) \tag{4.24}$$

é obtida pela função de densidade de probabilidade condicionada as *n* amostras, $f(\boldsymbol{x}, z_k \mid (n))$, e a partir dos K valores de corte, z_k, pela aproximação:

$$E[Z(\boldsymbol{x})] = \int_{-\infty}^{\infty} z \cdot dF(\boldsymbol{x}; z \mid (n)) \approx \sum_{k=1}^{K+1} \bar{z}_k \left[\hat{F}(\boldsymbol{x}; z_k \mid (n)) - \hat{F}(\boldsymbol{x}; z_{k-1} \mid (n))\right] \tag{4.25}$$

A estimativa do valor esperado como definida em (4.25) e aquela obtida por krigagem linear são ambas ótimas no sentido de minimizar variâncias inferenciais, entretanto produzem resultados diferentes. São diferentes porque, no caso do enfoque aqui adotado, derivam de uma *fdpac* que depende dos valores dos dados.

Estimativa da mediana

O estimador de mínimos quadrados não é a única função de otimização de erros possível. Uma outra função $L(\boldsymbol{\varepsilon}(\boldsymbol{x}))$ pode também ser considerada. Podemos tomá-la como sendo dada pelo valor absoluto dos erros estimados $L(\boldsymbol{\varepsilon}(\boldsymbol{x})) = |\boldsymbol{\varepsilon}(\boldsymbol{x})|$. Mostra-se que o valor de z que minimiza $E[\ L(\boldsymbol{\varepsilon}(\boldsymbol{x}))\ ]$, quando $L(\boldsymbol{\varepsilon}(\boldsymbol{x}))$ é o modulo de $\boldsymbol{\varepsilon}(\boldsymbol{x})$, é a mediana da distribuição $q_{0.5}(\boldsymbol{x})$, definida como:

$$q_{0.5}(\boldsymbol{x}) = F^{-1}(\boldsymbol{x}; 0.5 \mid (n)) \tag{4.26}$$

A mediana é inferida aplicando-se a função de ajuste da distribuição sobre os valores de corte com probabilidades acumuladas vizinhas ao valor 0.5. Para distribuições com alto grau de assimetria, a mediana é um estimador mais robusto do que a média. Os mapas de média e mediana, dos dados de altimetria de Canchim, estão mostrados na Figura 4-9.

Estimativa de quantis

A função de perda considerada nos dois estimadores anteriormente definidos não discriminava as diferenças de impacto dos erros de sub-estimação ou sobre-estimação. Entretanto, existem situações, como a descrita no início desse capítulo (vide Seção 4.1), em que cada um desses erros produz diferentes impactos, e essas diferenças devem ser também consideradas no processo inferencial. Assim, funções de perdas assimétricas devem ser utilizadas

$$L[\varepsilon(\boldsymbol{x})] = \begin{cases} w_1 \cdot \varepsilon(\boldsymbol{x}) & se\ \varepsilon(\boldsymbol{x}) \geq 0 \quad sobrestimado \\ w_2 \cdot \varepsilon(\boldsymbol{x}) & se\ \varepsilon(\boldsymbol{x}) < 0 \quad subestimado) \end{cases} \tag{4.27}$$

onde w_1 e w_2 são parâmetros não negativos, e medem o relativo impacto de sub ou sobre estimar. O estimador que minimiza essa função $L(\varepsilon(\boldsymbol{x}))$ é chamado de p-quantil, e definido como:

$$\hat{z}_q = F^{-1}(\boldsymbol{x}; p|(n)) = q_p(\boldsymbol{x}) \tag{4.28}$$

onde $p = \dfrac{w_2}{w_1 + w_2}$

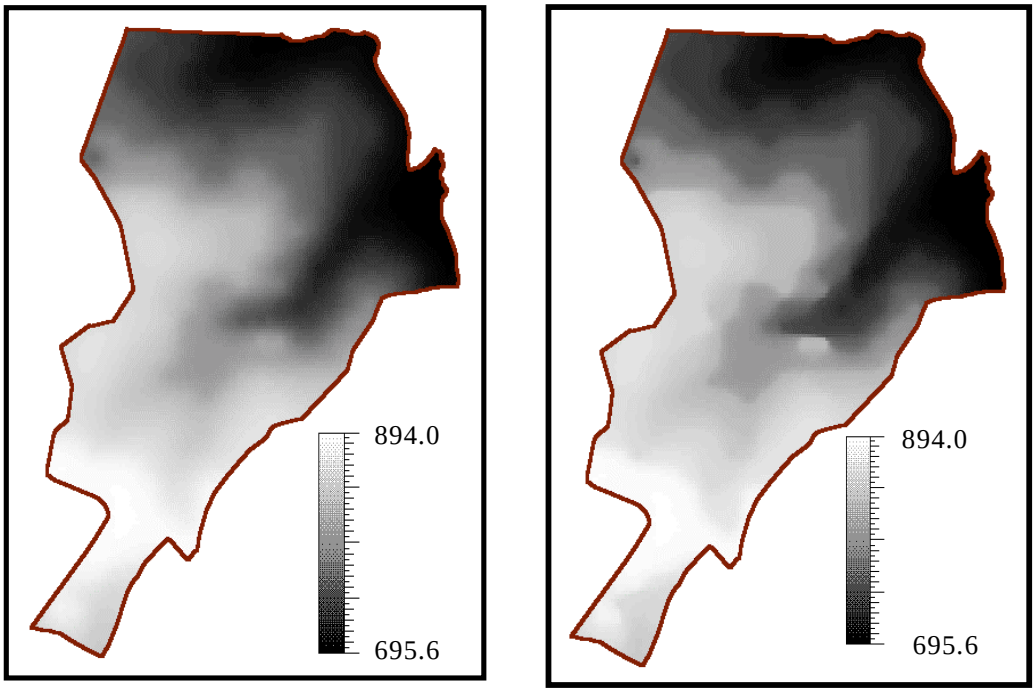

Figura 4-9 Mapas de média (a) e mediana (b) inferidos pelo procedimento por indicação, para os dados de altimetria da região de Canchim.

Considerando o exemplo de aplicação apresentado na introdução desse capítulo, seja w_1 o impacto de sobre-estimar um determinado nutriente no solo, e w_2 o impacto de subestimar este mesmo nutriente. Vamos supor que se deseja estimar $z(\boldsymbol{x})$ de forma a reduzir o risco de comprometimento da produção, que é motivado pelos erros de sobre-estimação. Dessa forma, $w_1 > w_2$ e $p < 0.5$, ou seja, um estimador ótimo seria um quantil menor do que a mediana, onde $p = 0.5$. Ou ainda, se $w_1 = 0.9$ e $w_2 = 0.1$, $p = 0.1$. A estimativa ótima seria considerando o quantil de 10%.

4.6 Incertezas locais para atributos Categóricos

O enfoque por indicação, semelhante àquele aplicado aos dados com atributos numéricos, pode ser também aplicado a dados com atributos categóricos, também chamados dados temáticos. O dado categórico é aqui considerado como o dado cujo atributo é discreto e sem ordenação, para o qual não é possível um cálculo de distribuições acumuladas, a menos que se defina uma ordenação para os mesmos. Um exemplo típico de dados categóricos é o atributo textura do solo, cujas classes são derivadas de atributos granulométricos do solo. Outros exemplos podem ser: tipos de rochas, classes de solo, etc. A metodologia geoestatística, aqui apresentada, utilizada para espacialização de dados categóricos, baseia-se na krigeagem por indicação e, equivale a um processo de classificação de dados categóricos a partir de amostras individuais. Os principais conceitos abordados aqui são exemplificados a partir do mesmo conjunto de dados coletados na região de Canchim (vide Seção 3.4, Figura 4-10 e Tabela 4-2).

O Enfoque por Indicação para Atributos Categóricos

Considere-se um dado espacial cujo atributo é categórico, podendo assumir *K* classes, ou estados diferentes, $s_k, k = 1,..., K$. Para cada posição $(\boldsymbol{x})$ do espaço, o dado categórico pode ser representado por uma variável aleatória $S(\boldsymbol{x})$ que pode assumir s_k estados, cada um associado a uma probabilidade de ocorrência. Os procedimentos por indicação para atributos categóricos baseiam-se na modelagem da *função de distribuição de probabilidade condicionada*, (fdpc), isto é, a modelagem da distribuição condicionada aos *n* dados amostrados, $p(\boldsymbol{x}; s_k \mid (n))$, que é definida como:

$$p(\boldsymbol{x}; s_k \mid (n)) = Prob\{S(\boldsymbol{x}) = s_k \mid (n)\} \tag{4.29}$$

A $p(\boldsymbol{x}; s_k \mid (n))$ modela a incerteza da variável aleatória S no ponto $(\boldsymbol{x})$ e, uma vez estimada, essa função de distribuição de probabilidade pode ser utilizada para:

- classificar o atributo em posições não conhecidas;

- modelar a incerteza das classificações efetuadas.

Pela metodologia por indicação, a definição da *fdpc* depende, inicialmente, da definição de um conjunto de valores de cortes para a variável em questão. Para um conjunto de amostras de uma variável aleatória categórica qualquer, o número de cortes *K* é definido pela quantidade de classes que essa variável pode assumir no seu domínio. Neste caso, a codificação por indicação, se processa em valor de cortes s_k, e gera um *conjunto amostral por indicação* $i(\boldsymbol{x};s_k)$ do tipo:

$$i(\boldsymbol{x};s_k)=\begin{cases}1, & \text{se } s(\boldsymbol{x})=s_k\\ 0, & \text{se } s(\boldsymbol{x})\neq s_k\end{cases} \tag{4.30}$$

A codificação por indicação é aplicada sobre todo conjunto amostral criando, para cada corte s_k, um conjunto amostral por indicação, $I(\boldsymbol{x};s_k\,|(n))$, cujos valores são 0 ou 1. Cada probabilidade condicional $p(\boldsymbol{x};s_k\,|(n))$ é, também, a esperança condicional da variável aleatória por indicação $I(\boldsymbol{x};s_k\,|(n))$, a saber:

$$p(\boldsymbol{x};s_k\,|(n))=E\{I(\boldsymbol{x};s_k|(n))\} \tag{4.31}$$

onde $I(\boldsymbol{x};s_k)=1$ se $S(\boldsymbol{x})=s_k$, e 0 (zero) caso contrário .

Assim, a *fdpc* da variável categórica $S(\boldsymbol{x})$ pode ser modelada usando-se um enfoque por indicação, semelhante àquele aplicado às variáveis de natureza contínua. Para cada um dos *K* conjuntos $I(\boldsymbol{x};s_k\,|(n))$, define-se um variograma experimental, ajustado *a posteriori* por um modelo teórico, que busca representar a variabilidade espacial do conjunto de dados codificados por indicação sendo considerados. Cada modelo de variograma teórico, em conjunto com as amostras, codificadas por indicação, é usado para se estimar o valor da probabilidade condicional $[p(\boldsymbol{x};s_k\,|(n))]^*$. O conjunto dessas probabilidades estimadas, considerando-se os *K* valores de corte, determina uma aproximação discreta da *fdpc* de $S(\boldsymbol{x})$. Essa *fdpc* deve, ainda, sofrer uma correção dos desvios de relação de ordem para se garantir as relações:

$$[p(\boldsymbol{x};s_k\,|(n))]^*\in[0,1]\quad k=1,\ldots,K \tag{4.32}$$

$$\sum_{k=1}^{K}[p(\boldsymbol{u};s_k\,|(\mathrm{n}))]^*=1 \tag{4.33}$$

ou seja, cada valor deve estar no intervalo [0,1] e a soma total desses valores deve ser igual a1.

4.7 Classificadores para Atributos Categóricos

No enfoque por indicação, os classificadores locais para atributos categóricos são definidos a partir da distribuição de probabilidade inferida para cada uma das s_k classes de $S(\boldsymbol{x})$. Em geral, esse classificador é implementado segundo um *estimador de moda*, que determina o valor de $S(\boldsymbol{x})$ como sendo a classe com a maior probabilidade inferida em $(\boldsymbol{x})$, ou seja:

$$S^*(\boldsymbol{x}) = s_{k\,max}(\boldsymbol{x}) = s_k(\boldsymbol{x}) \; sse \; [p(\boldsymbol{x}; s_k \mid (n))]^* > [p(\boldsymbol{x}; s_i \mid (n))]^* \; \forall i = 1,...,K \text{ e } k \neq i \tag{4.34}$$

Uma variante do classificador de moda considera também a reprodução das proporções globais definidas a priori. O mapa da Figura 4-11 mostra o resultado de uma classificação, pelo estimador de moda, a partir de um conjunto de amostras do atributo textura do solo.

4.8 Medidas de incerteza para atributos Categóricos

Apresentam-se, a seguir, dois procedimentos de medida de incertezas para atributos categóricos, a incerteza do classificador de moda e a incerteza por entropia de Shannon.

A Incerteza do classificador de moda

A incerteza local $Inc(\boldsymbol{x})$ pode ser definida como 1(um) menos a maior probabilidade condicional, estimada em $\boldsymbol{x}$ para as diversas classes de corte s_k :

$$Inc(\boldsymbol{x}) = 1 - [p(\boldsymbol{x}; s_{k\,max}(\boldsymbol{x}) \mid (n))]^* \tag{4.35}$$

A Figura 4-12 mostra o mapa de incertezas locais do classificador de moda usado na geração do mapa da Figura 4-11. Analisando-se a classificação apresentada na Figura 4-11 e o mapa de incertezas da Figura 4-12, observa-se que este último mostra um campo com variação proporcional ao comportamento do atributo na região. Nas regiões de transição entre as classes, os valores de incerteza por moda aumentam, com os valores mais baixos longe das transições, como ocorre naturalmente com muitas propriedades naturais nas proximidades de zonas de fronteira.

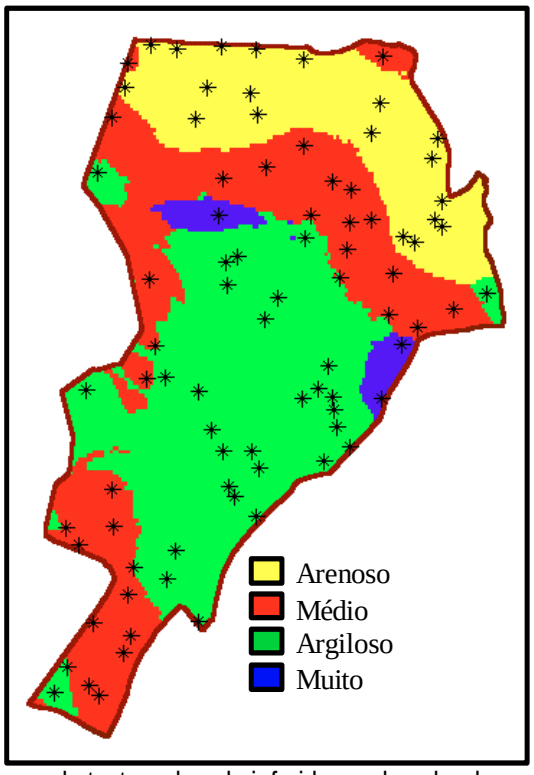

Figura 4-11 Mapa de valores de textura do solo inferidos, pelo valor de moda, a partir do procedimento de krigeagem por indicação

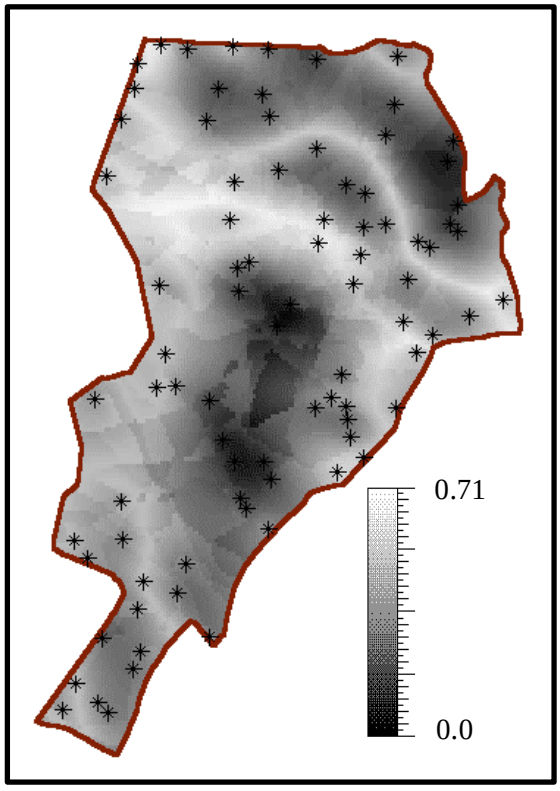

Figura 4-12 Mapa de incerteza por moda estimado a partir do procedimento de krigeagem por indicação usado para inferir o mapa da Figura 4-11

Incerteza por entropia de Shannon

Outra medida da incerteza local ***Inc(x)*** é a entropia de Shannon das probabilidades condicionais das diversas classes de corte s_k, definida como:

$$Inc(\boldsymbol{x}) = H(\boldsymbol{x}) \cong -\sum_{k=1}^{K} ln[p(\boldsymbol{x}; s_k \mid (n))]^* [p(\boldsymbol{x}; s_k \mid (n))]^* \geq 0 \qquad (4.36)$$

A entropia de Shannon é maximizada para distribuições uniformes, ou seja, quando as probabilidades de ocorrência das classes se igualam. Assim, os valores de incerteza por entropia de Shannon são maiores onde existe uma confusão maior entre as classes consideradas. Isto pode ser observado no mapa de incertezas da Figura 4-13.

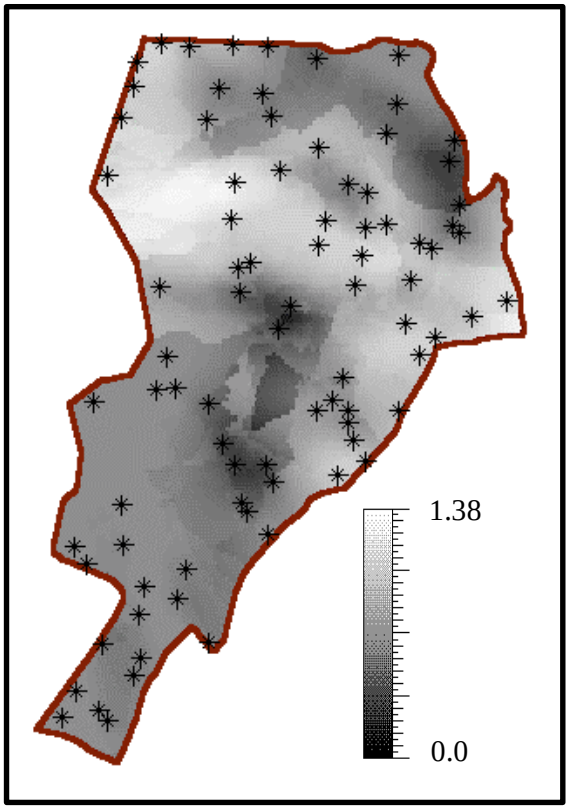

Figura 4-13 Mapa de incerteza por entropia de Shannon estimado a partir do procedimento de krigeagem por indicação usado para inferir o mapa da Figura 4-11

Comparando-se os mapas das Figura 4-12 e Figura 4-13, pode-se analisar as diferenças existentes entre o mapa de incertezas por moda e o mapa de incertezas por entropia. As diferenças são mais aparentes nas regiões onde várias classes se confundem. Este é um resultado esperado, uma vez que, nestas regiões a

distribuição de probabilidade das variáveis aleatórias está mais próxima de uma distribuição uniforme, quando então a incerteza medida pela entropia tem seus valores maximizados. A incerteza por moda mostra um crescimento a partir da parte central de uma classe em direção as zonas de transição. Os valores máximos de incerteza por moda aparecem nas bordas entre as classes e, não têm influência do número de classes próximos as bordas. Dependendo da aplicação, o especialista é responsável por decidir sobre qual medida de incerteza estará trabalhando. Quando a confusão entre classes é importante deve-se optar pela incerteza por entropia. Caso o interesse seja somente nas transições entre as classes, a incerteza por moda deve ser priorizada.

Conclusões

Apresentamos neste capítulo a formalização do procedimentos geoestatísticos da krigagem por indicação. Estes procedimentos servem não apenas para produzir uma predição de valores sobre uma superfície, mas essencialmente como uma poderosa ferramenta para produzir modelos de incertezas locais para dados geográficos que compartilham uma base de informações Estes dados são sempre usados em conjunto com outros para produzir novas informações, através de operações e transformações. Os procedimentos da geoestatística, em seu enfoque por indicação, nos permitem produzir informações espaciais qualificadas por uma métrica de "confiança" nas informações representadas naqueles suportes, os mapas. Temos a possibilidade concreta de produzir e operar com os mapas e suas "barras de erro". Podemos ainda ressaltar as seguintes características, específicas do procedimento de krigagem por indicação:

- a krigagem por indicação é não paramétrica. Não considera nenhum tipo de distribuição de probabilidade a priori para a variável aleatória. Ao invés disso, ela possibilita a construção de uma aproximação discretizada da *fdpac*. Os valores de probabilidades discretizados podem ser usados diretamente para se estimar valores característicos da distribuição, tais como: quantis, valor esperado e variância. Portanto, ela não se restringe a modelagem de atributos com distribuições simétricas como, por exemplo, a gaussiana;
- a krigagem por indicação fornece uma metodologia única para espacialização, com estimativa de incertezas, para atributos espaciais tanto de natureza temática quanto numérica;
- diferentemente da krigagem linear, que estima a variância do erro de estimação em função do estimador e da distribuição geométrica das amostras, a krigagem por indicação possibilita a estimativa de incertezas, utilizando a função de distribuição acumulada condicionada da VA que representa o atributo, independentemente do estimador;

- a krigagem por indicação pode ser usada para modelar atributos com alta variabilidade espacial sem a necessidade de se filtrar amostras cujos valores estão muito distantes de uma tendência (“outliers”);
- a krigagem por indicação permite melhorar a qualidade de estimação com o uso de amostras indiretas, retiradas de fontes auxiliares, que são acrescidas ao conjunto amostral do atributo, as amostras diretas.

No entanto, os procedimentos de krigagem por indicação apresentam também alguns problemas, além das probabilidades negativas e funções acumuladas inválidas já mencionados. Este procedimento requer, do especialista, um alto grau de interatividade para a definição da quantidade e dos valores de corte a serem utilizados. Também, exige que seja definido um variograma para cada valor de corte considerado.

A ferramenta geoestatística de krigagem é utilizada para inferir valores de atributos, em posições não observadas, e também incertezas associadas aos valores inferidos. Mostrou-se que a krigagem por indicação tem aplicação mais geral, principalmente porque não supõe nenhum tipo de distribuição de probabilidade a priori e pode ser usado com atributos numéricos e temáticos. Por exemplo, a krigagem por indicação permite a inferência de valores temáticos e, portanto, pode ser considerada um classificador estocástico, que fornece estimativas de incertezas associadas aos valores das classes atribuídos a cada ponto do espaço. Apresentou-se, ainda, alternativas para estimativas de incertezas que devem ser escolhidas de acordo com a natureza do atributo, que está sendo modelado, e também de acordo com os objetivos de uma aplicação.

Salienta-se que os procedimentos geoestatísticos por indicação incluem também os simuladores estocásticos, que não foram abordados neste capítulo. Também não foi abordado o uso de informação indireta para a melhora das inferências. Estes tópicos são de extrema relevância para o contexto do uso efetivo da geoestatística em análise de dados geográficos e deverão ser considerados em futuras edições. Mesmo no método por indicação algumas limitações da krigeagem permanecem – uso dos dados para estimar o variograma e predizer a incerteza, deficiência na extrapolação, ou seja, avaliar a incerteza fora dos dados. Novas generalizações começam a surgir, tomando como base a teoria dos campos aleatórios espaço-temporais.

REFERÊNCIAS BIBLIOGRÁFICAS

A estrutura teórica da geoestatística em seu enfoque por indicação está bem apresentada em Goovaerts (1997) e em Isaaks e Srivastava (1989). Algoritimos implementados e explicações didáticas sobre como operar a Krigeagem por indicação pode ser encontrada no livro de Deutsch e Journel (1992). Com relação à integração entre geoestatística e SIGs e modelagem e tratamento de incertezas em SIG, o leitor deve referir-se a Felgueiras C. A. (1999), Felgueiras et al (1999) e Heuvelink (1998). As questões sobre medidas de entropia podem ser apreciadas no clássico Shannon, and Weaver (1949). Para uma discussão sobre diferentes medidas de incerteza no enfoque por indicação veja Soares(1992). Referente a modelagem espaço-temporal, deve-se consultar o artigo de Kyriakidis e Journel (1999) e o livro do George Christakos (2000). Referências básicas sobre os dados da Fazenda Canchim podem ser encontrados em Calderano Filho et al. (1996). Estes dados também estão disponíveis no site do livro (www.dpi.inpe.br/gilberto/livro/analise).

Calderano Filho, B.; Fonseca, O. O. M.; Santos, H. G. e Lemos A. L.. *Levantamento Semidetalhado dos Solos da Fazenda Canchim São Carlos - SP*. Rio de Janeiro, EMBRAPA- CNPS, 1996. 261p.

Christakos, G. *Modern Spatiotemporal Geostatistics*; IAMG Studies no. 6, Oxford University Press, 2000

.Deutsch e Journel (1992). *GSLIB: Geostatistical Software Library and user's guide*. New York, Oxford University Press, 1992. 339p.

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.

Goovaerts, P. *Geostatistics for Natural Resources Evaluation*. New York, Oxford University Press, 1997. 481p.;

Isaaks E. H. and Srivastava R. M. **A***n Introduction to Applied Geostatistics*, Oxford University Press, 1989. 560p.

Kyriakidis, P. C. e Journel, A. G. Geostatistical Space-Time Models: A Review. Mathematical Geology, Vol. 31, No. 6, 1999

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

Shannon, C. E. e Weaver, W. *The Mathematical Theory of Communication*. Urbana, The University of Illinois Press, 1949. 117p.

Soares, A. *Geoestatistical Estimation of Multi-Phase Structures.* Mathematical Geology, 24(2):140-160, 1992.