

CAPÍTULO 2

CONCEITOS FUNDAMENTAIS

2.1 MODELAGEM POR TÉCNICAS DA GEOESTATÍSTICA

A geoestatística se constitui de um conjunto de ferramentas, determinísticas e estatísticas, que foram desenvolvidas para o entendimento e modelagem da variabilidade espacial de atributos ambientais.

Dentro desse conjunto de ferramentas, existem técnicas de inferência conhecidas por *krigeagem*, que partem do paradigma de que qualquer valor z , não amostrado, é caracterizado como uma variável aleatória Z (Deutsch e Journel, 1998). Desta forma, dentro de uma região A da superfície terrestre, para cada posição $\mathbf{u} \in A$, o valor do atributo ambiental $z(\mathbf{u})$ é modelado como uma variável aleatória $Z(\mathbf{u})$ (Felgueiras et al., 1999).

Esta variável aleatória $Z(\mathbf{u})$ pode assumir diferentes valores para o atributo e, cada um deles, com uma probabilidade de ocorrência associada. Nas posições amostradas \mathbf{u}_α , $\alpha=1,2,\dots,n$, os valores $z(\mathbf{u}_\alpha)$ são considerados determinísticos ou, ainda, podem ser considerados variáveis aleatórias cujo valor medido tem uma probabilidade de 100% de ocorrer. Já nas posições não amostradas, a incerteza associada aos valores $z(\mathbf{u})$ é modelada pela função de distribuição de probabilidade da variável aleatória $Z(\mathbf{u})$ (Felgueiras, 1999).

Segundo Deutsch e Journel (1998), a função de distribuição de probabilidade acumulada, fdpa, é definida como:

$$F(\mathbf{u}; z) = \text{Prob} \{Z(\mathbf{u}) \leq z\}.$$

Quando a fdpa é específica para um particular conjunto de informações, consistindo de n valores vizinhos, a fdpa é dita condicionada a n e, portanto, a função de distribuição de probabilidade acumulada condicionada, fdpac, é definida como:

$$F(\mathbf{u}, z | (n)) = \text{Prob} \{Z(\mathbf{u}) \leq z | (n)\}.$$

Se a variável aleatória $Z(\mathbf{u})$ for uma variável categórica que pode assumir qualquer um dos K valores, $k = 1, \dots, K$, a notação da fdpc é:

$$F(\mathbf{u}; k | (n)) = \text{Prob}\{Z(\mathbf{u}) = k | (n)\}.$$

Uma distribuição de probabilidade pode ser determinada por dois caminhos, um paramétrico e outro não paramétrico. Enquanto naquele se estabelece um modelo de distribuição a priori, que é completamente determinado por um conjunto limitado de parâmetros, neste estima-se um conjunto de valores que representam uma aproximação discretizada do modelo de distribuição. É a partir destes modelos, estabelecidos a priori ou estimados a posteriori, que são inferidos os valores dos atributos em posições não conhecidas e a incerteza sobre os valores inferidos (Felgueiras, 1999).

Segundo o mesmo autor, os estimadores de *krigeagem* linear, ou simplesmente *krigeagem*, podem ser usados para a inferência dos parâmetros média e variância de um modelo de distribuição de probabilidade gaussiano. Estes parâmetros correspondem, respectivamente, aos valores estimados nas posições não amostradas e à variância do erro de estimação.

Entretanto, alerta o autor, a modelagem por *krigeagem* linear possui algumas restrições, correspondentes a) ao fato de apenas modelar atributos de natureza numérica, b) à grande dificuldade para verificação da hipótese de distribuição multigaussiana dos atributos, e c) ao fato de que a estimativa da variância de *krigeagem*, que é totalmente condicionada ao tipo de estimador utilizado, ser independente dos valores dos atributos.

Diante deste cenário, uma alternativa é a modelagem por procedimentos não lineares da geoestatística, como exemplo a *krigeagem* por indicação, uma técnica não paramétrica que permite a construção de uma aproximação discretizada da distribuição de probabilidade do atributo, da qual são estimados os valores não amostrados e as incertezas associadas.

Com esta estimativa direta da distribuição, independente de um estimador, as incertezas condicionam-se apenas ao comportamento de variabilidade do atributo. É importante observar que, além da possibilidade de modelar atributos de natureza numérica, pode-se modelar atributos de natureza categórica.

Em seguida, serão apresentados e formalizados os procedimentos de *krigeagem* por indicação e algumas formas de propagação das incertezas estimadas nestes procedimentos. Para uma maior organização, a apresentação foi dividida em duas: *krigeagem* por indicação para atributos numéricos e para atributos categóricos.

2.1.1 KRIGEAGEM POR INDICAÇÃO APLICADA A DADOS DE NATUREZA NUMÉRICA

Segundo Felgueiras (1999), a *krigeagem* por indicação é uma técnica de estimação não linear, que corresponde a uma *krigeagem* linear aplicada sobre um conjunto amostral, cujos valores foram modificados por uma codificação por indicação.

Esta codificação por indicação é uma transformação não linear, que para um conjunto amostral $Z(\mathbf{u}=\mathbf{u}_\alpha)$, dado um valor de corte z_k , gera um conjunto amostral por indicação $I(\mathbf{u}=\mathbf{u}_\alpha; z_k)$, do tipo:

$$I(\mathbf{u}; z_k) = 1, \text{ se } Z(\mathbf{u}) \leq z_k$$

$$I(\mathbf{u}; z_k) = 0, \text{ se } z(\mathbf{u}) > z_k.$$

O conjunto amostral por indicação é usado para a inferência das variáveis aleatórias por indicação $I(\mathbf{u}; z_k)$ nas posições não amostradas. Assim, a esperança condicional da variável aleatória numérica por indicação é determinada por:

$$E\{I(\mathbf{u}; z_k) | (n)\} = 1 * \text{Prob}\{I(\mathbf{u}; z_k) = 1 | (n)\} + 0 * \text{Prob}\{I(\mathbf{u}; z_k) = 0 | (n)\} =$$

$$1 * \text{Prob}\{I(\mathbf{u}; z_k) = 1 | (n)\} = F^*(\mathbf{u}; z_k) | (n).$$

Da equação acima, verifica-se que a esperança condicional de $I(\mathbf{u}; z_k)$ fornece, para o valor de corte $z = z_k$, uma estimativa da função de distribuição de probabilidade

acumulada condicionada $F^*(\mathbf{u}; z_k|n)$ para atributos numéricos. Este resultado é extremamente importante, porque a esperança condicional $E\{I(\mathbf{u}; z_k|n)\}$ pode ser estimada por algoritmos de *krigeagem*, como a *krigeagem* por indicação ordinária, e, desta forma, podem ser estimados valores da fdpac de $Z(\mathbf{u})$ para vários valores de corte, possibilitando a construção de uma aproximação discretizada da fdpac de $Z(\mathbf{u})$. Quanto maior o conjunto de valores de corte, melhor será esta aproximação.

Alguns valores característicos da distribuição de probabilidade, como valor médio e variância, podem ser estimados diretamente dos valores discretizados, enquanto outros, como a mediana e quantis, dependem de uma função de ajuste para a distribuição. Serão mostrados aqui, apenas os estimadores do valor médio e da variância, pois eles são aplicados na metodologia do estudo de caso.

O valor esperado $\mu_z(\mathbf{u})$ de uma variável aleatória, pode ser estimado por:

$$E[Z(\mathbf{u})] = \int_{-\infty}^{\infty} z \cdot f(\mathbf{u}; z | n) dz = \int_{-\infty}^{\infty} z \cdot dF(\mathbf{u}; z | n)$$

e a partir de K valores de corte, pode ser estimado pela aproximação:

$$\mu_z(\mathbf{u}) = \int_{-\infty}^{\infty} z \cdot dF(\mathbf{u}; z | n) \approx \sum_{k=1}^{K+1} z'_k [F^*(\mathbf{u}; z_k | n) - F^*(\mathbf{u}; z_{k-1} | n)]$$

onde os valores das $F^*(\mathbf{u}; z_k|n)$, $k = 1, 2, \dots, K$, são os valores estimados das fdpc's acumuladas para cada valor z_k do atributo, $z_0 = z_{min}$, $z_{K+1} = z_{max}$, $z'_k = (z_k + z_{k-1})/2$, $F(\mathbf{u}; z_0|n) = 0$ e $F(\mathbf{u}; z_{K+1}|n) = 1$.

Da mesma forma, a variância $(\sigma^2)^*(\mathbf{u})$ de uma variável aleatória, pode ser estimada por:

$$(\sigma^2)^*(\mathbf{u}) = \int_{-\infty}^{\infty} [z - \mu_z(\mathbf{u})]^2 dF(\mathbf{u}; z | n) \approx \sum_{k=1}^{K+1} [z'_k - \mu_z(\mathbf{u})]^2 [F^*(\mathbf{u}; z_k | n) - F^*(\mathbf{u}; z_{k-1} | n)]$$

A partir destas métricas podem ser estimados intervalos de confiança, que são a forma mais comum de expressar incertezas. Tem-se um intervalo de confiança quando há uma probabilidade de um valor desconhecido estar entre um valor mínimo e um valor máximo, por exemplo, um intervalo de 95% de probabilidade do valor estimado para a

média μ de uma distribuição gaussiana estar entre os valores $\mu \pm 2\sigma$, onde σ é o desvio padrão da distribuição.

Um cuidado que se deve ter com os resultados obtidos dos procedimentos de *krigeagem* por indicação é quanto aos desvios de relação de ordem. Eles ocorrem porque em cada valor de corte os pesos de estimação da *krigeagem* são únicos, ou seja, não são usados os mesmos pesos para todos os valores de corte e, ainda, porque a *krigeagem* não garante que todos os pesos sejam positivos.

Assim, para atributos numéricos, podem ocorrer dois problemas com os valores inferidos de probabilidade acumulada. O primeiro, é que podem ser inferidos valores menores que 0 e maiores que 1, e o segundo, é que o valor estimado para um valor de corte z_j , pode vir a ser maior que o valor estimado para um valor de corte z_k , quando $z_j \leq z_k$ (Felgueiras, 1999).

A solução para o primeiro problema é um ajuste dos valores estimados para as bordas, mapeando os valores negativos para 0 e os valores maiores que 1 para 1. Já para o segundo problema há várias soluções, e uma delas é a construção de uma fdpac máxima e uma mínima, em função da fdpac estimada, e obter uma média entre elas.

A construção da função máxima parte do valor estimado para o primeiro corte z_k , e segue a seguinte regra: o valor do corte seguinte z_{k+1} , deve ser maior ou igual ao valor do corte z_k . Se o valor do corte z_{k+1} for menor, ele é ignorado, e para este corte é atribuído o valor correspondente ao corte z_k , e assim, sucessivamente, até a construção da função máxima.

Para a construção da função mínima, a regra é inversa e parte do valor do último corte, sendo que o valor do corte anterior deve ser menor ou igual ao valor do corte que está sendo considerado (Deutsch e Journel, 1998). A Figura 2.1 ilustra este procedimento de correção.

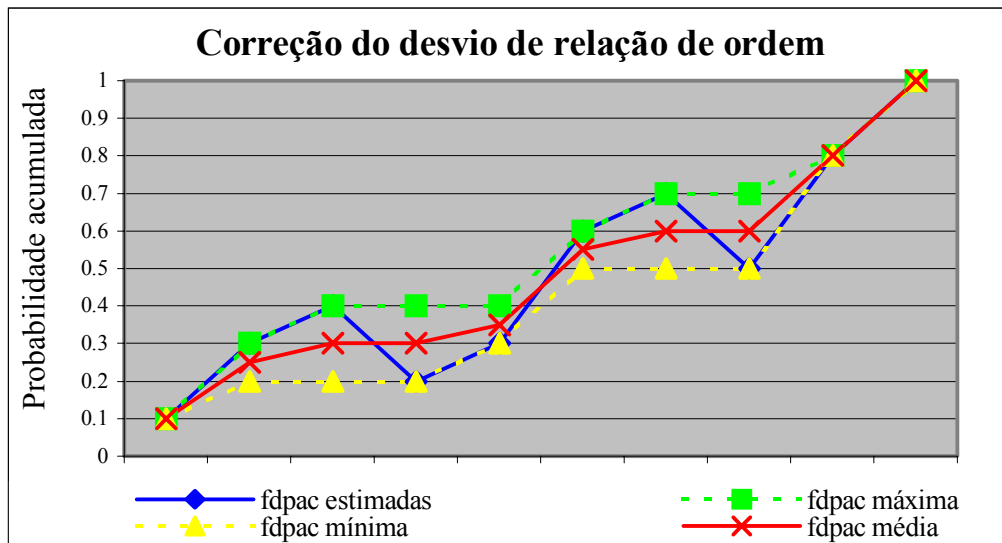


Fig. 2.1 – Ilustração da correção do desvio de relação de ordem, na modelagem de atributos numéricos, utilizando *krigeagem* por indicação

FONTE: Adaptada de Deutsch e Journel, 1998.

2.1.2 KRIGEAGEM POR INDICAÇÃO APLICADA A DADOS DE NATUREZA CATEGÓRICA

Todo o processo de *krigeagem* por indicação para atributos categóricos é semelhante ao processo aplicado para atributos numéricos. Basicamente, a diferença fica por conta da codificação por indicação, que gera conjuntos amostrais por indicação, pelos seguintes critérios (Felgueira, 1999):

$$I(\mathbf{u}; z_k) = 1, \text{ se } Z(\mathbf{u}) = z_k$$

$$I(\mathbf{u}; z_k) = 0, \text{ se } Z(\mathbf{u}) \neq z_k$$

onde os valores de corte z_k , $k = 1, \dots, K$, correspondem às K classes que pertencem ao domínio da função aleatória $Z(\mathbf{u})$.

A esperança condicional da variável aleatória por indicação categórica, inferida por procedimentos de *krigeagem*, permite a estimativa de valores da função de densidade de probabilidade condicionada, para todas as classes z_k . Desta forma, em cada localização \mathbf{u} , é inferido um valor de probabilidade $p_k(\mathbf{u})$ para cada uma das classes z_k .

Partindo das probabilidades $p_k(\mathbf{u})$, a estimativa do valor não amostrado da variável aleatória categórica $Z(\mathbf{u})$ pode ser obtida pelo estimador de moda, o qual atribui para cada posição \mathbf{u} a classe de maior probabilidade (Soares, 1992 citado por Felgueiras, 1999), ou seja:

$$Z(\mathbf{u}) = z_k \text{ se } p_k(\mathbf{u}) > p_i(\mathbf{u}) \forall i = 1, \dots, K \text{ e } k \neq i$$

A incerteza associada a esse valor estimado, pode ser inferida por uma medida chamada de incerteza pela moda, que corresponde à probabilidade de não ser atribuída, para a posição \mathbf{u} , a classe de maior probabilidade $p_k(\mathbf{u})$ (Felgueiras, 1999), e pode ser estimada por:

$$\text{Inc}(\mathbf{u}) = 1 - p_k(\mathbf{u})$$

Quanto aos desvios de relação de ordem, para o caso categórico, também existem dois problemas que devem ser verificados. O primeiro é a estimativa de valores de probabilidade fora do intervalo $[0, 1]$. A solução é o mesmo mapeamento para as bordas apresentado no caso numérico. O segundo problema é a somatória dos valores de probabilidade inferidos para as K classes ser diferente de 1. Neste caso, uma solução é aplicar uma função linear que transforme os valores de probabilidade inferidos em valores de probabilidade cuja a soma total seja 1 (Deutsch e Journel, 1998).

2.2 PROPAGAÇÃO DAS INCERTEZAS ASSOCIADAS AOS ATRIBUTOS NUMÉRICOS

A propagação de incertezas em SIG corresponde ao tratamento das incertezas associadas aos dados de entrada, quando estes são submetidos a operações de análise espacial. Em outras palavras, pode-se dizer que uma representação de saída $U(\cdot)$ originada de um processo de análise espacial, corresponde às representações de entrada $b_i(\cdot)$ e suas incertezas $V_i(\cdot)$, manipuladas pelas operações $g(\cdot)$, ou seja (Heuvelink, 1998):

$$U(\cdot) = g([b_1(\cdot) + V_1(\cdot)], \dots, [b_m(\cdot) + V_m(\cdot)]).$$

Quando as representações $b_i(\cdot)$ são representações de atributos numéricos, as operações $g(\cdot)$ correspondem a operações locais e as incertezas $V_i(\cdot)$ são expressas em função da variância dos atributos numéricos, o mesmo autor apresenta quatro métodos de propagação de incertezas, que são: método de Taylor de primeira ordem, método de Taylor de segunda ordem, método de Rosenblueth e método de simulação de Monte Carlo.

O método de Taylor de primeira ordem, que foi o método aplicado no estudo de caso, consiste de uma expansão das séries de Taylor de primeira ordem centrada no vetor \mathbf{b} , que é um vetor de valores médios de n variáveis aleatórias Z_i , $i = 1, \dots, n$. A expansão pode ser escrita como (Heuvelink, 1998 e Heuvelink et al., 1989):

$$U = g(Z) = g(\mathbf{b}) + \sum_{i=1}^n \left\{ (Z_i - b_i) \left(\frac{\delta g}{\delta z_i}(\mathbf{b}) \right) \right\} + \text{residuo}$$

supondo que o termo de resíduo possa ser desconsiderado, o valor médio μ e a variância σ^2 de U são obtidos por:

$$\begin{aligned} \mu = E[Y] &= E \left[g(\mathbf{b}) + \sum_{i=1}^n \left\{ (Z_i - b_i) \left(\frac{\delta g}{\delta z_i}(\mathbf{b}) \right) \right\} \right] = g(\mathbf{b}) \\ \sigma^2 = E[(Y - E[Y])^2] &= E \left[\left(g(\mathbf{b}) + \sum_{i=1}^n \left\{ (Z_i - b_i) \left(\frac{\delta g}{\delta z_i}(\mathbf{b}) \right) \right\} - g(\mathbf{b}) \right)^2 \right] \\ &= E \left[\left(\sum_{i=1}^n \left\{ (Z_i - b_i) \left(\frac{\delta g}{\delta z_i}(\mathbf{b}) \right) \right\} \right) \left(\sum_{j=1}^n \left\{ (Z_j - b_j) \left(\frac{\delta g}{\delta z_j}(\mathbf{b}) \right) \right\} \right) \right] \\ &= \sum_{i=1}^n \left\{ \sum_{j=1}^n \left\{ \tau_{ij} \sigma_i \sigma_j \frac{\delta g}{\delta z_i}(\mathbf{b}) \cdot \frac{\delta g}{\delta z_j}(\mathbf{b}) \right\} \right\} \end{aligned}$$

onde σ_i e σ_j são respectivamente os valores de desvio padrão das variáveis aleatórias Z_i e Z_j e, τ_{ij} é o coeficiente de correlação destas mesmas variáveis.

Partindo das equações acima, verifica-se que U depende apenas das representações b_i e que a variância de U , além de depender das correlações e desvios padrões das variáveis aleatórias Z_i , depende também, das primeiras derivadas de $g(\cdot)$, o que impõe, para que esse método de propagação possa ser aplicado, o uso de operações diferenciáveis de análise espacial.

2.3 PROPAGAÇÃO DAS INCERTEZAS ASSOCIADAS AOS ATRIBUTOS CATEGÓRICOS

A propagação de incertezas de representações de atributos categóricos parte da mesma idéia da propagação de incertezas de representações de atributos numéricos. A única diferença, que na verdade não diz respeito diretamente à propagação, é a limitação de funções de análise espacial aplicadas às representações categóricas. Isto ocorre principalmente com as funções de integração, que se restringem aos operadores lógicos da Álgebra *Booleana*, como os operadores “AND”, “OR” e “XOR”.

Na literatura foi encontrado apenas um meio de propagação de incertezas de atributos categóricos, que é pela aplicação da teoria de probabilidade. Esta forma de propagação foi proposta por Newcomer e Szagin (1984), citado por Walsh et al. (1987) e Lanter e Veregin (1992), para a propagação de medidas globais de incerteza, que normalmente correspondem à probabilidade de classificação correta das representações categóricas.

Os procedimentos da *krigeagem* por indicação para atributos categóricos permitem a estimativa de medidas pontuais de incertezas, que também são probabilidades, mas com o significado inverso das medidas globais, ou seja, correspondem à probabilidade de classificação incorreta. Neste caso, o meio de propagação proposto não pode ser usado na íntegra, necessitando de adaptações. Assim, toda a formulação apresentada em seguida, é a formulação proposta para a propagação das medidas globais, porém, adaptada neste estudo.

A aplicação da teoria de probabilidade depende do operador lógico usado na integração. Se for usado o operador “AND”, que corresponde a uma operação de interseção de dois

eventos, a incerteza propagada é estimada pela probabilidade de não ocorrer esta interseção, ou seja:

$$Inc = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - (P(A) \cdot P(B/A))$$

onde $P(A)$ é a probabilidade de ocorrer o evento A e $P(B/A)$ é a probabilidade condicional de ocorrer B, dado que está ocorrendo A.

Se os eventos A e B forem independentes, o fato de ocorrer A não indica nada a respeito da ocorrência de B, então $P(B/A) = P(B)$ (Meyer, 1983). A incerteza propagada passa a ser estimada por:

$$Inc = P(\overline{A \cap B}) = 1 - P(A \cap B) = 1 - (P(A) \cdot P(B))$$

Para verificar a independência entre os eventos existe um teste estatístico, cuja hipótese nula é a igualdade entre a probabilidade de interseção dos eventos e o produto das probabilidades individuais destes eventos. A estimativa da estatística do teste parte das probabilidades marginais dos eventos, as quais podem ser estimadas das próprias amostras. Assim, supondo que será verificada a independência entre A e B (Freitas, 1998):

	A				
B	P_{11}	P_{12}	...	P_{1j}	P_{1*}
	P_{21}				P_{2*}

	P_{i1}			P_{ij}	P_{i*}
	P_{*1}	P_{*2}	...	P_{*j}	

	Amostras		
	A	não A	
B	n_{11}	n_{12}	n_{1*}
não B	n_{21}	n_{22}	n_{2*}
	n_{*1}	n_{*2}	n

as probabilidades marginais, p_{i*} e p_{*j} , podem ser estimadas por:

$$\hat{p}_{i*} = \frac{n_{i*}}{n} \quad \text{e} \quad \hat{p}_{*j} = \frac{n_{*j}}{n}$$

Destas probabilidades marginais, podem ser estimadas as probabilidades p_{ij} , e delas, E_{ij} :

$$\hat{p}_{ij} = \hat{p}_{i*} \cdot \hat{p}_{*j} \quad \text{e} \quad \hat{E}_{ij} = \hat{p}_{ij} \cdot n$$

A estatística do teste corresponde a:

$$Q = \sum_{i=1}^L \sum_{j=1}^C \left(\frac{n_{ij} - \hat{E}_{ij}}{\hat{E}_{ij}} \right)^2$$

que segue uma distribuição χ^2 , com $((L - 1) * (C - 1))$ graus de liberdade.