Abordagem geoestatística por indicação com uso de copulas bivariadas empíricas para modelagem de incertezas associadas a imagens de sensoriamento remoto

Carlos Alberto Felgueiras[1]
Eduardo Celso Gerbi Camargo[2]
Jussara de Oliveira Ortiz[3]

[1,2,3]Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12227-010 - São José dos Campos - SP, Brasil
{carlos, eduardo, jussara}@dpi.inpe.br

Abstract. This paper presents an indicator geoestatistical methodology based on empirical bivariate copulas for spatial uncertainty modeling for remote sensing images. As the size of remote sensing images is usually very large it is used a random sample set sufficient to represent the spatial variability of the entire image. The sample set is considered as input to establish the structure of the spatial correlation via indicator semivariograms using empirical bivariate copulas. A set of cutoff values is considered to obtain the indicator semivariograms. The indicator semivariograms are fitted by mathematical models in order to be used as input, along with the samples, for indicator geostatistical approaches of kriging estimations. A case study is presented with China-Brazil Earth Remote Satellite (CBERS) images from the Amazon forest region considering deforested and no deforested areas. The results of the case study are reported along with spatial analyses considering aspects of the uncertainty related to the representations and estimations.

Palavras-chave: estrutura de correlação espacial, semivariograma, copula bivariada empírica, geostatística por indicação, sensoriamento remoto.

1. Introdução

Muitos estudos relacionados à área de sensoriamento remoto geralmente apresentam seus resultados e conclusões na forma de mapas e imagens, sem a preocupação de considerar ou avaliar a incerteza associada aos dados. É importante apontar nos resultados obtidos como essa incerteza se distribui espacialmente, ou seja no espaço geográfico considerado, pois tal informação pode, por exemplo, ser útil ou essencial na construção de planos de tomadas de decisões mais elaborados. Uma forma de modelar a incerteza associada aos dados é adotar uma abordagem probabilística, que tem como base o conceito de variável aleatória. Assim, todo dado espacial, observado e estimado, é considerado uma variável aleatória regionalizada. Por exemplo, numa imagem de sensoriamento remoto, cada pixel da imagem pode ser visto como uma das possíveis realizações de uma variável aleatória. Isto significa que o valor do pixel em uma determinada localização pode assumir diversos valores, dependendo da função de distribuição acumulada que caracteriza tal variável aleatória. Neste sentido, este trabalho emprega procedimentos geoestatísticos, que têm sido amplamente utilizados para se estimar modelos de incertezas, locais e globais, de variáveis e funções aleatórias representando atributos ambientais do espaço geográfico [Isaaks e Srivastava (1989), Deutsch e Journel (1998); Felgueiras (1999)]. Atributos espaciais, tais como, altimetria, temperatura, teores geofísicos e geoquímicos, respostas espectrais de sensoriamento remoto, entre outros, podem ser analisados e modelados pela geoestatística.

Os procedimentos geoestatísticos utilizam como informação de entrada um conjunto de pontos observados, amostras pontuais, que estão distribuídos em uma região da geografia terrestre. Cada ponto amostral tem associado a sua localização geográfica (**u**) e um valor agregado, denotado de $z(\mathbf{u})$, que é o atributo de interesse. Além disso, os procedimentos geoestatísticos requerem a modelagem da estrutura de correlação espacial do atributo investigado. Isto é realizado através da função semivariograma, que pode ser determinada

pelos valores e distâncias entre os pares de pontos amostrais. Com base nos semivariogramas, procedimentos geoestatísticos por indicação, krigeagem e simulação, são aplicados para obtenção de estimativas, valores simulados e incertezas dos atributos espaciais considerados.

Nesse contexto, o objetivo deste trabalho é construir modelos de incertezas locais para dados de imagens de sensoriamento remoto, usando procedimentos geoestatísticos por indicação. Como diferencial, os semivariogramas por indicação foram obtidos com uso de copulas empíricas bivariadas que representam a estrutura de correlação espacial, ou dependência espacial, necessária para modelagem da incerteza. O uso de cópulas bivariadas permite estabelecer a dependência espacial em toda a gama de quantis do atributo investigado, enquanto que o uso de semivariogramas tradicionais descrevem a dependência espacial a partir de valores médios, ou seja a dependência média. Além disso cópulas não são sensíveis a "outliers", Bardossy (2008), Kazianka (2010).

Para avaliar a metodologia imposta neste trabalho, um estudo de caso é apresentado com imagens do sensor CCD ("Charge Coupled Device") do satélite sino-brasileiro CBERS ("China-Brazil Earth Resources Satellite") da região da floresta amazônica considerando áreas com e sem desmatamentos.

A estrutura deste trabalho inicia-se com uma introdução na Seção 1. A Seção 2 apresenta os principais conceitos teóricos para a compreensão da metodologia imposta neste trabalho. A Seção 3 apresenta um estudo de caso, seguido dos resultados obtidos e discussões. Conclusões e sugestões para continuação deste trabalho são considerados na Seção 4.

2. Conceitos Básicos e Metodologia de Trabalho

2.1 O Semivariograma

A semivariância é definida como a metade da média das diferenças, ao quadrado, dos valores do atributo z entre pares de amostras pontuais. O semivariograma é uma função que relaciona as semivariâncias com os módulos dos vetores de distâncias espaciais **h**. Um semivariograma empírico, ou experimental, pode ser estimado diretamente de um conjunto amostral de um atributo espacial pela formulação apresentada na Equação 1.

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [z(\mathbf{u}_i) - z(\mathbf{u}_i + \mathbf{h})]^2 \quad (1)$$

onde N(**h**) é o número de pares a uma distância aproximada |**h**| entre as posições espaciais $\mathbf{u}_i$ e $\mathbf{u}_i$ +**h**. Semivariogramas direcionais consideram também a direção do vetor **h** definido por cada par de amostras. Semivariogramas omnidirecionais não consideram essa direção e são usados para representar fenômenos isotrópicos, Camargo et al. (2001).

Os semivariogramas empíricos são ajustados por modelos matemáticos conhecidos como semivariogramas teóricos. Estes são usados em procedimentos geoestatísticos para estabelecer a estrutura de correlação espacial do atributo investigado e, posteriormente, empregados para estimar e simular valores de atributos em locais não observados. A Figura 1 mostra um semivariograma teórico exponencial, ajustado sobre um empírico, e seus parâmetros, o efeito pepita C_0, o patamar $C(\mathbf{0})$ e o alcance a.

Os modelos matemáticos mais usados para ajuste de um semivariograma empírico são: Esférico, Exponencial, Gaussiano e Potencia. Por exemplo, o modelo teórico exponencial é definido pela Equação 2, onde C_1, a contribuição do modelo, é igual a $C(\mathbf{0}) - C_0$.

$$\gamma(\mathbf{h}) = \begin{cases} 0, & |\mathbf{h}| = 0 \\ C_0 + C_1\left[1 - \exp\left(-3\frac{|\mathbf{h}|}{a}\right)\right], & |\mathbf{h}| \neq 0 \end{cases} \quad (2)$$

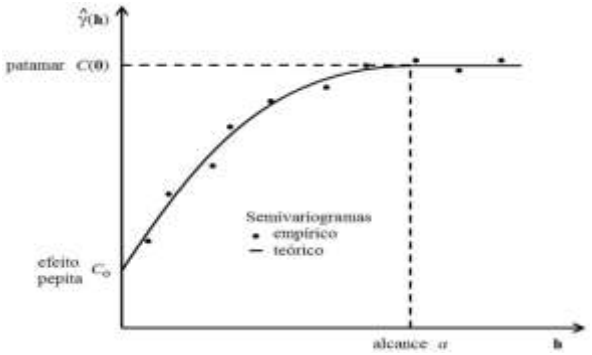


Figura 1. Ilustração de um semivariograma empírico e teórico e seus parâmetros.

2.2 Semivariogramas por Indicação

Os semivariogramas por indicação são construídos por campos por indicação derivados das amostras originais considerando um conjunto de valores de corte z_k determinados pelo usuário ou pela aplicação.

Para obter um campo binário por indicação $I(\mathbf{u};z_k)$, aplica-se a transformação da Equação 3 sobre os valores da variável $Z(\mathbf{u})$, considerando-se o valor de corte z_k.

$$I(\mathbf{u};z_k)=\begin{cases}1, & \text{se } Z(\mathbf{u})\le z_k\\ 0, & \text{se } Z(\mathbf{u})> z_k\end{cases} \tag{3}$$

Para cada campo por indicação pode-se construir um semivariograma por indicação que representa a variabilidade da variável indicadora para a região espacial considerada.

2.3 Cópulas e Semivariogramas por Indicação

Copulas são distribuições multivariadas padronizadas com marginais uniformes que podem ser utilizadas para descrever a estrutura de dependência de distribuições multivariadas separadamente de suas marginais univariadas, Bardossy (2008). Copulas Espaciais Bivariadas podem ser determinadas, a partir de conjuntos amostrais pontuais, para representar a dependência espacial de atributos, ou variáveis, regionalizados.

A copula espacial bivariada $\mathbf{C_s}$, considerando-se dois quantis quaisquer q_1 e q_2, dos valores $z(\mathbf{u})$ and $z(\mathbf{u}+\mathbf{h})$ separados por uma distancia **h,** é definida como:

$$\mathbf{C_s}(\mathbf{h},q_1,q_2)=P\left[F_z\left(Z(\mathbf{u})\right)<q_1,F_z\left(Z(\mathbf{u}+\mathbf{h})\right)<q_2\right]=C\left(F_z\left(Z(\mathbf{u})\right),F_z\left(Z(\mathbf{u}+\mathbf{h})\right)\right) \tag{4}$$

onde $Z(\mathbf{u})$ são valores de atributos nas posições espaciais **u**.

Bardossy (2006) mostra que há uma relação entre as copulas e os semivariogramas por indicação. Para cada valor de corte β o variograma por indicação γ_β é dado por:

$$\gamma_\beta(\mathbf{h})=F_z(\beta)-\mathbf{C_s}(\mathbf{h},F_z(\beta),F_z(\beta)) \tag{5}$$

e, o variograma cruzado por indicação γ_{β_1,β_2}correspondente aos cortes β_1 e β_2 é:

$$\gamma_{\beta_1,\beta_2}(\mathbf{h})=\frac{1}{2}\min\left(F_z(\beta_1),F_z(\beta_2)\right)-\mathbf{C_s}(\mathbf{h},F_z(\beta_1),F_z(\beta_2)) \tag{6}$$

A demonstração dessas equações pode ser encontrada em Bardossy (2006). A presente abordagem é diferente da apresentada por Journel e Deutsch (1996) por não usar um único semivariograma para descrever a dependência média no espaço, mas usar cópulas para quantificar a dependência em toda a gama de quantis (Bardossy, 2006)

2.4 Procedimento Geoestatístico por Indicação

Abordagens geoestatísticas por indicação possibilitam inferir os modelos de incertezas de variáveis aleatórias, condicionados a um conjunto de amostras pontuais, em locais não observados. Esses modelos são utilizados para se estimar e simular valores das variáveis aleatórias utilizando-se de procedimentos geoestatísticos conhecidos como krigeagem e simulação por indicação respectivamente.

A krigeagem de variáveis aleatórias por indicação RV $I(\mathbf{u};z)$ é usada para estimativa, de mínima variância e não tendenciosa, para o valor esperado de $I(\mathbf{u};z)$. Esse valor esperado é igual a *função de distribuição acumulada condicionada* - *fdac* - na posição espacial **u**, condicionado às *n* amostras, como definido na Equação 7.

$$\begin{aligned} \mathrm{E}\{I(\mathbf{u};z_k)\,|\,(n))\} &= 1\cdot \mathrm{Prob}\{I(\mathbf{u};z_k)=1\,|\,(n)\} + 0\cdot \mathrm{Prob}\{I(\mathbf{u};z_k)=0\,|\,(n)\} \\ &= 1\cdot \mathrm{Prob}\{I(\mathbf{u};z_k)=1\,|\,(n)\} = F(\mathbf{u};z_k\,|\,(n)) \end{aligned} \tag{7}$$

Assim a krigeagem por indicação fornece o modelo de incerteza, a *fdac*, local sobre $z(\mathbf{u})$. Utilizando-se um conjunto de *K* valores de corte z_k é possível se obter uma aproximação do modelo completo da fdac para $z(\mathbf{u})$. Esse modelo é usado para se inferir parâmetros estatísticos da variável aleatória $Z(\mathbf{u})$, tais como, a média μ e a variância σ^2, a partir das Equações 8 e 9, Goovaerts (1997):

$$\mu = [z(\mathbf{u})]^*_{\mathrm{E}} = \int_{-\infty}^{\infty} z\cdot \mathrm{d}\,F(\mathbf{u};z\,|\,(n)) \approx \sum_{k=1}^{\mathrm{K+1}} z'_k \cdot [F(\mathbf{u};z_k\,|\,(n)) - F(\mathbf{u};z_{k-1}\,|\,(n))] \tag{8}$$

$$\sigma^2 = Var[z(\mathbf{u})]^* \approx \sum_{\mathrm{k=1}}^{\mathrm{K+1}} (z'_{\mathrm{k}} - [z(\mathbf{u})]^*_E)^2 \cdot [F(\mathbf{u};z_{\mathrm{k}}\,|\,(n)) - F(\mathbf{u};z_{\mathrm{k-1}}\,|\,(n))] \tag{9}$$

Além disso a *fdac* obtida pode ser utilizada para se estimar intervalos de confiança baseados no seu desvio padrão, raiz quadrada da variância, e em valores de quantis de $Z(\mathbf{u})$. Os quantis são inferidos por interpolações entre os K valores $F(\mathbf{u};z_{\mathrm{k}}\,|\,(n))$, Felgueiras (1999).

A simulação por indicação possibilita, ainda, a inferência de *modelos de incertezas globais* condicionados aos valores das amostras observadas e aos valores pré-realizados. Goovaerts (1997) mostra e detalha esse tipo de procedimento conhecido como *simulação sequencial por indicação*.

2.2 Metodologia de trabalho

A metodologia deste trabalho segue as etapas apresentadas na sequência abaixo:

1. Definir a imagem e a banda(s) a serem analisadas;
2. Sortear um conjunto amostral de pontos representativos da imagem;
3. Definir o conjunto de valores de cortes da variável aleatória ;
4. Para cada valor de corte, definir os semivariogramas empíricos a partir de copulas bivariadas;
5. Ajustar os semivariogramas empíricos por teóricos e;
6. Utilizar os semivariogramas teóricos nos procedimentos geoestatísticos por indicação

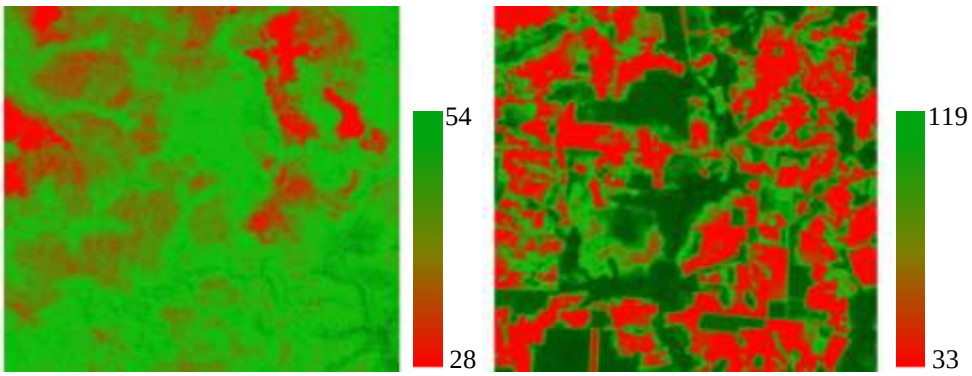
3. Resultados e Discussões

3.1 Regiões de Estudo

Neste trabalho foram consideradas duas pequenas regiões da banda 3, de uma imagem CCD CBERS, mostradas nas Figuras 2 e 3. Estas imagens são partes da cena de órbita/ponto

171/114 desse sensor adquirida em 01 de setembro de 2008 . Essa cena foi importada para o software SPRING, Camara et al. (1996), onde essas regiões foram cortadas e salvas.

A Figura 2(a) mostra a imagem da região 1 que representa uma área típica de floresta natural e a Figura 2(b) mostra a imagem da região 2 que representa uma área parcialmente desmatada. As coordenadas, em graus decimais, do retângulo envolvente da região 1 são: longitude 60.2146, latitude 12.3786 e longitude 60.1218, latitude 12.2873; enquanto que as da região 2 são: longitude 60.4195, latitude 13.0131 e longitude 60.3262, latitude 12.9218. Cada imagem tem resolução espacial de 20m x 20m e tamanho de 500 linhas x 500 colunas.



(a) (b)

Figura 2. Bandas 3 da imagem CBERS: (a) da região 1 (floresta natural) e (b) da região 2 (floresta parcialmente desmatada)

3.2 Conjunto Amostral e Estatísticas Básicas

Para cada região, apresentadas na Figura 2, foi realizada uma amostragem aleatória de tamanho de 10000 amostras. Felgueiras (2013) mostra que este tamanho de amostra é suficiente para representar a variabilidade espacial das regiões 1 e 2 para as bandas 4 dessas imagens. Aqui considera-se que o mesmo resultado é válido para as bandas 3.

Os valores de análise das estatísticas básicas de refletâncias são apresentados na Tabela 1.

Tabela 1. Estatísticas básicas dos valores de refletâncias das bandas 3 das regiões 1 e 2

Estatísticas Básicas	Região 1	Região 2
Valor Mínimo	28	33
Valor Máximo	54	119
Média	34.08	49.99
Variância	2.93	125.1
Desvio Padrão	1.71	11.19
Coeficiente de Variação	0.05	0.22
Assimetria	2.18	0.33
Achatamento	14.16	2.35
Quartil Inferior	33	39
Mediana	34	50
Quartil Superior	35	60

3.3 Semivariogramas por indicação

Os semivariogramas por indicação foram obtidos por Copulas segundo a Equação 5.

A Figura 3 apresenta os semivariogramas empíricos e teóricos, exponenciais, para três valores de corte que representam os quartis da distribuição dos valores da região 1 em estudo. A Tabela 2 reporta os parâmetros dos semivariogramas teóricos ajustados aos empíricos.

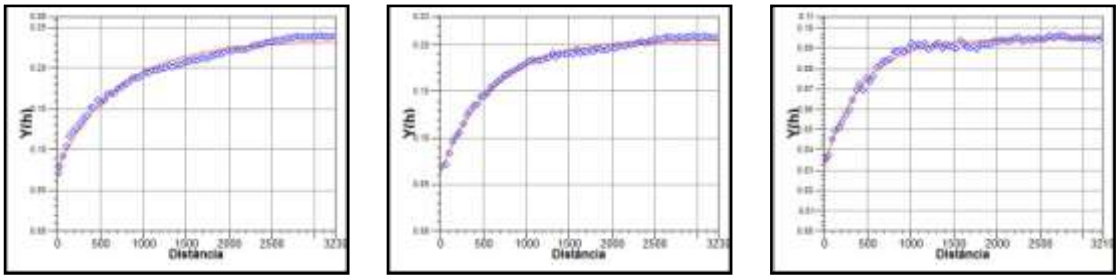


Figura 3. Semivariogramas por Indicação da região 1 para os valores de corte 33, 34 e 35 (respectivamente da esquerda para a direita)

Tabela 2. Parâmetros dos semivariogramas teóricos ajustados para os cortes da região 1

Valor de Corte	Efeito Pepita	Contribuição	Alcance
33	0.074	0.159	2124.95
34	0.059	0.146	1778.54
35	0.030	0.064	1312.87

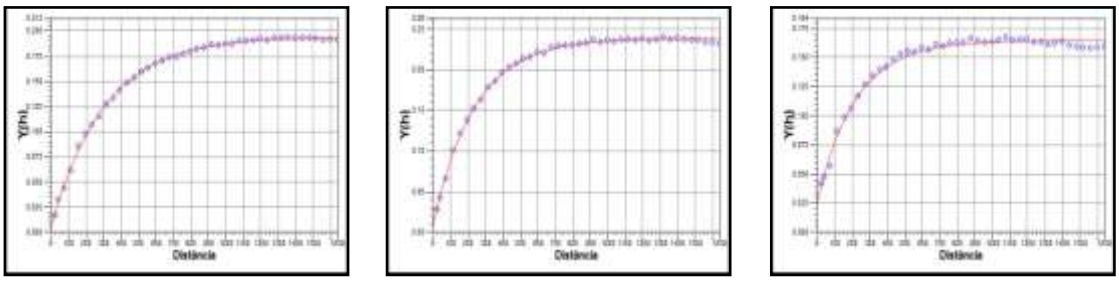


Figura 4. Semivariogramas por Indicação da região 2 para os valores de corte 39, 50 e 60 (respectivamente da esquerda para a direita)

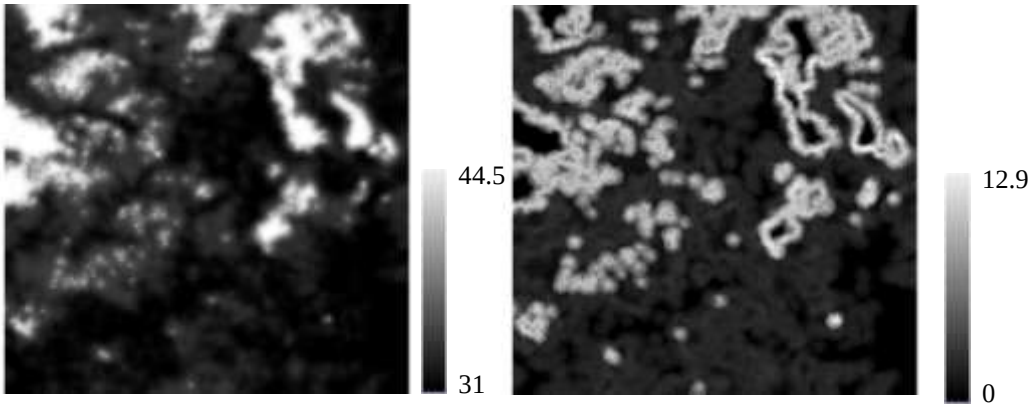
Tabela 3. Parâmetros dos semivariogramas teóricos ajustados para os cortes da região 2

Valor de Corte	Efeito Pepita	Contribuição	Alcance
39	0.003	0.191	915.46
50	0.005	0.233	714.59
60	0.023	0.141	655.66

Os resultados das tabelas e figuras refletem o fato de que a região 1 tem variância menor, é mais homogênea, que a região 2. Isto significa que os pixels da região 1 estão mais concentrados em torno do seu valor médio. Além disso, observa-se que a diferença entre os valores mínimo e máximo é maior para a região 2. Por isso as informações de refletância da imagem da região 1 foram quantizadas com um número menor de valores digitais. Por ser mais homogênea, a região 1 tem seu semivariograma com valores de alcance maiores e com valores de contribuição menores do que os da região 2. Os efeitos pepitas maiores na região 1 podem ser explicados pela quantização com uma quantidade menor de níveis digitais.

3.4 Estimativas da Krigeagem por Indicação

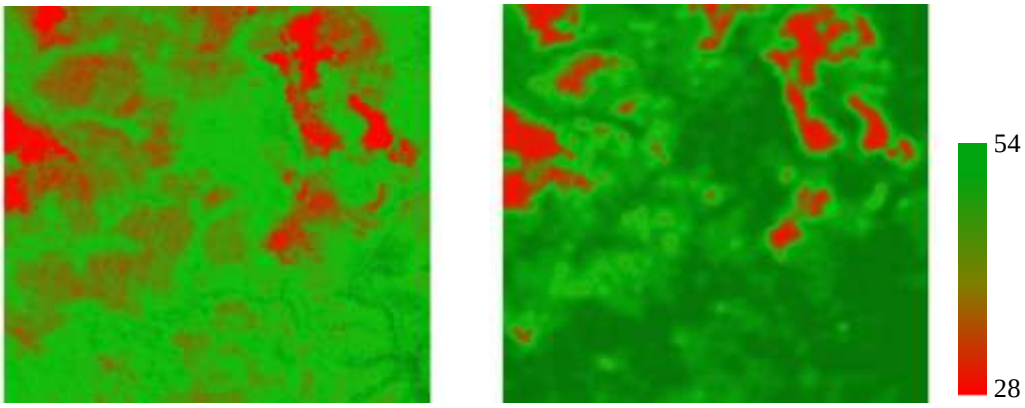
Utilizando-se do conjunto de 10000 amostras e os semivariogramas teóricos da Figura 3 aplicou-se o procedimento de krigeagem por indicação para a região 1. Os resultados das estimativas dos valores médios das *fdacs* locais são apresentados no mapa da Figura 5(a). O mapa da Figura 5(b) mostra informações de incertezas, associadas aos valores estimados, baseadas nos desvios padrões dos modelos de incertezas locais. Essas incertezas são utilizadas para qualificar os resultados obtidos. As incertezas menores ocorrem em regiões homogêneas onde as amostras contem níveis digitais parecidos. Em áreas mais heterogêneas e nas transições entre áreas homogêneos as incertezas são maiores.



(a) (b)

Figura 5. Imagem dos valores médios (a) e das incertezas por desvios padrões (b) obtidos por krigeagem por indicação sobre as amostras e variogramas dos cortes da região 1

A imagem original, mostrada na Figura 6(a), pode ser comparada com a imagem estimada por krigeagem por indicação da Figura 6 (b).



(a) (b)

Figura 6. Imagem original (a) e imagem estimada por krigeagem por indicação (b) da região 1

Qualitativamente, observa-se que a imagem estimada possui informações de tendências da imagem original. Isto porque nesta interpolação foram considerados apenas 10000 amostras. O resultado qualitativo pode ser melhorado com o aumento do número de elementos no conjunto amostral.

Uma vantagem desse procedimento é inferir valores onde existem problemas com falta ou com erros nas informações da imagem. A falta pode ser por falha dos sensores e os erros de imageamento podem ocorrer em regiões com nuvens ou com de sombras de nuvens, por exemplo. Nestes casos pode-se utilizar a metodologia deste trabalho para inferência dessas informações utilizando-se os semivariogramas da amostragem geral, como os das Figuras 3 e

4, e uma amostragem local, vizinhança próxima a essas regiões. Essa metodologia pode ser também aplicada para propósitos de refinamento da resolução espacial das imagens digitais.

4. Conclusões

Este artigo mostrou ser possível construir modelos de incerteza para dados de imagens de sensoriamento remoto, de regiões de florestas, usando abordagens geoestatísticas por indicação baseadas em copulas bivariadas. O uso dessas copulas tem a vantagem de representar a dependência espacial em toda a gama de quantis enquanto que semivariogramas tradicionais descrevem a dependência espacial com valores médios, a dependência média.

A metodologia proposta foi ilustrada com um estudo de caso com bandas de imagens CCD CBERS. Os resultados obtidos mostraram ser possível modelar a incertezas com um subconjunto amostral aleatório das imagens. Isto permite a análise de dependência espacial por semivariogramas sem o uso de todas as informações da imagem, o que é muito custoso temporal e computacionalmente. As incertezas são usadas para estimativas de valores em locais não conhecidos, com erros ou faltantes, e ainda qualificar os resultados das estimativas.

Embora não explorada neste trabalho, as inferências dos modelos de incertezas podem ser também obtidas por simulação sequencial por indicação com os mesmos semivariogramas utilizados pela krigeagem por indicação. A vantagem de se utilizar a simulação, em lugar da krigeagem, é a de se obter modelos de incertezas globais que representam melhor a covariança do conjunto amostral de entrada. No futuro pretende-se explorar essa metodologia para outras imagens de sensoriamento remoto e a simulação sequencial por indicação.

Referências Bibliográficas

Bárdossy, A. Copula-based geostatistical models for groundwater quality parameters, **Water Resources Research**, 42, W11416, doi:10.1029/2005WR004754, 2006.

Bárdossy, A.; Li, J. Geostatistical interpolation using copulas. **Water Resources Research**, 44, W07412, doi:10.1029/2007WR006115, 2008

Burgess, T. M.; Webster, R., "Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging". In: **Journal of Soil Science**, Vol 31:315-31, 1980.

Camara G., Souza, RCM, Freitas UM and Garrido J. SPRING, Integrating Remote Sensing and GIS by object-oriented data modeling. **Computer & Graphics**, v. 20, n. 17, p.395-403, 1996.

Camargo, E.C.G.; Felgueiras, C.A. Monteiro, A.M. A Importância da Modelagem da Anisotropia na Distribuição Espacial de Variáveis Ambientais Utilizando Procedimentos Geoestatísticos. In **Anais** do X Simpósio Brasileiro de Sensoriamento Remoto SBSR, Foz do Iguaçu, p. 395-402, INPE, Abril, 2001.

Deutsch, C.V.; Journel, A.G. **GSLIB: geostatistical software library and user's guide**. New York: Oxford University Press, 1998. 369p.

Felgueiras, C.A. **Modelagem ambiental com tratamento de incertezas em sistemas de informação geográfica: o paradigma geoestatístico por indicação**. 1999. 165p. PhD Thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos. 1999.

Felgueiras, C. A.; Camargo, E. C. G.; Rennó, C. D.; Ortiz, J. O.Spatial Variability Analysis of CBERS CCD Images in Forest Regions. In **Anais** do XVI Simpósio Brasileiro de Sensoriamento Remoto, SBSR2013, Foz do Iguaçu, PR. 2013.

Goovaerts, P. **Geostatistics for natural resources evaluation**. New York: Oxford University Press, 1997. 483p.

Heuvelink, G. B. M. **Error Propagation in Environmental Modeling with GIS**. Brystol: Taylor and Francis Inc, 1998. 345p.

Isaaks, E. H; Srivastava, R.M. **An introduction to applied geostatistics**. New York: Oxford University Press, 1989. 561p.

Kazianka, H.; Pilz, J. Copula-based geostatistical modeling of continuous and discrete data including covariates. **Stochastic Environmental Research and Risk Assessment.** v. 24, p.661-673, 2010.