ISSN: 2317-0840

Representação espectral de dados espacialmente correlacionados: relações e comparações com a Geoestatística

Edilenia Queiroz Pereira [1†], Diogo Francisco Rossoni [2]; Carla Eloize Carducci[3]

[1] *Departamento de Estatística – Universidade Estadual de Maringá – UEM.*
[2] *Departamento de Estatística – Universidade Estadual de Maringá – UEM*
[3] *Universidade Federal de Santa Catarina – Campus de Curitibanos – UFSC*

Resumo: *A Geoestatística busca detectar e explicar a dependência associada a um campo aleatório espacial contínuo. Tanto a abordagem espacial quanto a espectral, podem ser instrumentos válidos para detectar essa dependência espacial. O estudo da Geoestatística por meio do enfoque espectral busca recursos para solucionar problemas que a teoria Geoestatística enfrenta, como manipulação de grandes bancos de dados. Para este propósito, são utilizadas técnicas espectrais, sendo estas poderosas ferramentas para estudar a estrutura espacial, além de oferecerem significantes benefícios computacionais na manipulação dos dados. Constatou-se que a partir da densidade espectral foi possível obter estimativas para calcular a covariância; estando a covariância diretamente relacionada com a semivariância, pode-se obter a semivariância estimada. Além disso, mostrou-se que o tempo computacional gasto quando se trabalha com a densidade espectral permanece constante para todos os tamanhos n de amostras simuladas. Já no método clássico o tempo computacional gasto aumentou exponencialmente à medida que n aumentou.*

Palavras-chave: *Densidade Espectral, Estatística Espacial, Geoestatística, Big data*.

Abstract: *Geostatistics seeks to detect and explain the dependence associated with a continuous spatial random field. Both the spatial and spectral approaches can be valid instruments to detect this spatial dependence. The study of Geostatistics through the spectral approach seeks ways to solve problems that the Geostatistical theory faces, such as manipulation of large databases. For this purpose, spectral techniques are used, being these powerful tools to study the spatial structure, besides offering significant computational benefits in data manipulation. It was found that from the spectral density it is possible to obtain estimates to calculate the covariance; With covariance directly related to semivariance, the estimated semivariance can be obtained. In addition, it has been shown that the computational time spent when working with the spectral density remains constant for all n sizes of simulated samples. In the classical method, the computational time spent increased exponentially as n increased.*

Keywords: *Spectral density, Spatial Statistics, Geostatistics, Big Data*.

Introdução

Pesquisadores da área espacial, muitas vezes deparam-se com grandes conjuntos de dados, normalmente coletados em uma região de grande porte. A manipulação destes grandes conjuntos de dados é, de certo modo, problemática para as técnicas Geoestatística, seja por ter que lidar com a inversão de uma matriz de covariância excessivamente grande para se obter a função de probabilidade ou que exija um grande nível de processamento computacional (MATEU; JUAN; PORCU, 2007).

Os métodos espectrais estão ganhando uma importância cada vez maior na análise de dados espaciais, devido aos avanços das tecnologias computacionais (ARAÚJO; BATISTA; SCALON, 2014; KIM; FUENTES,2000). A principal vantagem do estudo sobre o domínio espectral dá-se devido aos cálculos serem realizados no domínio da frequência (VIDAKOVIC; MUELLER, 1994; MALLAT, 1989). Além disso, a função de densidade espectral e a função de covariância de um processo estocástico estacionário estão estritamente ligadas, uma vez que eles formam uma Transformada de Fourier. Assim, o estudo das propriedades de segunda ordem de um campo aleatório através da função de covariância ou da densidade espectral pode ser visto como equivalente (CASAIS, 2006).

[†]Autor correspondente: edileniaqueiroz@gmail.com.

Estudos na literatura mostram que tanto a abordagem espacial como a espectral, são instrumentos válidos capazes de detectar a dependência espacial. Para a estimativa e simulação sob uma configuração espacial são encontrados na literatura vários estudos (CHERRY, 1997; CRESSIE; HAWKINS, 1980; CRESSIE, 1985; ZIMMERMAN; ZIMMERMAN, 1991).

A eficiência dos métodos espaciais contra métodos espectrais depende da geometria do domínio espacial, do método de estimação, do número de observações, e muitas outras variáveis que têm de ser levadas em consideração para realizar uma análise correta (MATEU; JUAN; PORCU, 2007).

Sendo o periodograma, um dos métodos espectrais tidos como uma ferramenta poderosa para o estudo da estrutura espacial de processos contínuos espaciais e às vezes oferecem benefícios significativos no âmbito computacional. Usando a representação espectral de um processo espacial, podemos facilmente construir funções de covariância (definida positiva) válidas e introduzir novos modelos para campos espaciais (GELFAND et al., 2010).

Deste modo, o objetivo deste trabalho consiste em verificar a aplicabilidade de técnicas espectrais para a obtenção de estimativas de semivariância e covariância; realizar comparações com estimativas obtidas pelos métodos clássicos da Geoestatística com intuito de otimização computacional; e verificar a aplicabilidade das técnicas espectrais em conjuntos reais de dados.

Estimador de densidade espectral

De acordo com Gelfand et al. (2010) o periodograma, também conhecido como densidade espectral da amostra, é um estimador não paramétrico clássico da densidade espectral, sendo a transformada de Fourier da função de covariância.

Para um processo espacial $Z(s)$ observado em um malha regular de locais igualmente espaçados, $D=\{\mathbf{s}=(s_1,s_2): s_1=0,\ldots,n_1-1, s_2=0,\ldots,n_2-1\}, D\subset R^2$, com $N=n_1n_2$ pontos. Tem-se o periodograma espacial dado por:

$$I_N(\omega_0)=\frac{\delta_1\delta_2}{(2\pi)^2 n_1n_2}\left|\sum_{\mathbf{s}\in D} Z(\Delta\mathbf{s})\exp(-i\Delta\mathbf{s}^t\omega)\right|^2 \quad (1)$$

em que: $\Delta\mathbf{s}=\delta_1 s_1\delta_2 s_2$ é o vetor de distância entre as observações próximas; ω é a frequência; n_1n_2 é o número de pontos amostrados; δ_1 é a localização em uma direção (direção do eixo x); δ_2 é a localização em outra direção (direção do eixo y).

Se a representação espectral do processo espacial $Z(s)$ é dada por:

$$Z(\mathbf{s})=\int_{R^2}\exp(i\omega^t\mathbf{s})dy(\omega) \quad (2)$$

define-se $J(\omega)$ como a transformada discreta Fourier em Z,

$$J(\omega)=\frac{1}{(\delta_1\delta_2)^{\frac{1}{2}}2\pi(n_1n_2)^{\frac{1}{2}}}\sum_{\mathbf{s}\in D} Z(\Delta\mathbf{s})\exp(-i\Delta\mathbf{s}^t\omega). \quad (3)$$

Função de densidade espectral

Nesta seção, serão apresentados o modelo exponencial quadrado da função de densidade espectral e o método de mínimo quadrados para densidade espectral.

Modelo exponencial quadrado

A função de densidade espectral de um processo espacial isotrópico com uma covariância exponencial é definida como (GELFAND et al., 2010):

$$f(\omega) = \frac{1}{2}\sigma(\pi\alpha)^{-\frac{1}{2}} \exp\left(-\frac{\omega^2}{4\alpha}\right) \tag{4}$$

sendo a covariância definida como:

$$C(h) = \sigma \exp\{-\alpha h^2\} \tag{5}$$

O parâmetro σ é a variância do processo, já o parâmetro α^{-1} explica como a correlação decai rapidamente (GELFAND et al., 2010).

Estimativa do domínio espectral por mínimos quadrados

Para predição espacial, o comportamento do processo em altas freqüências é mais relevante. Pode ser obtida (com $N \to \infty$) ótima predição quando a densidade espectral em freqüências curtas é mal especificada (FUENTES, 2002). Uma aproximação da densidade espectral da classe de *Matérn* para valores de alta frequência pode ser obtida pela seguinte equação:

$$f(\omega) = \phi \|\omega\|^{(-2\nu - d)} \tag{6}$$

possibilitando que $\|\omega\| \to \infty$.

em que ν e ϕ são os graus de suavização e d é o número de dimensões em $\square^d$.

Uma alternativa ao modelo de alta frequência dado na equação (6) é ajustada por uma escala de log do modelo linear utilizando o método de mínimos quadrados ordinais:

$$\log(f(\omega)) = \beta_0 + \beta_1 x \tag{7}$$

em que $x = \log(\omega)$, $\beta_0 = \log(\phi)$ e $\beta_1 = 2\left(-\nu - \frac{d}{2}\right)$.

Relação entre a covariância e a semivariância

Em Geoestatítica, as principais ferramentas para compreender e modelar o componente aleatório e espacialmente correlacionado são a covariância, a correlação e a semivariância (YAMAMOTO; BARBOSA, 2013; JOURNEL, 1978).

Dentre estas, a semivariância é uma ferramenta crítica para os estudos de Geoestatística: (1) é uma ferramenta para analisar e quantificar a variabilidade espacial do fenômeno em estudo; (2) a maioria dos métodos de estimação Geoestatística e algoritmos de simulação requerem um modelo teórico ajustado a uma semivariância empírica (GRINGARTEN; DEUTSCH, 2001).

Tem-se que quando a estacionariedade de segunda ordem dos dados é estabelecida, é possível estabelecer uma relação direta entre a covariância e a semivariância (MELLO et al., 2005). Sendo, essa relação dada por:

$$\gamma(h) = \text{cov}(0) - \text{cov}(h). \tag{8}$$

A covariância pode ser obtida através das funções de densidade espectrais.

Material e métodos

Para o desenvolvimento da análise utilizou-se o software estatístico livre R (R CORE TEAM, 2015) e o pacote *fractaldim* (SEVCIKOVA; PERCIVAL; GNEITING, 2014).

Simulação para aplicação teórica

Os dados experimentais foram obtidos por meio de simulação no software estatístico R (R CORE TEAM, 2015). Para isso, utilizou-se o pacote *RandomFields* (SCHLATHER et al., 2015), o qual permite a análise e simulação de processos aleatórios espaciais. Esse pacote apresenta vários métodos, os quais permite estimar a dimensão fractal para dados de uma ou duas dimensões. Utilizou-se a função *RPspectral* para que os dados a serem gerados apresentassem função de covariância exponencial com variância igual a dois. Foram gerados também dois vetores x e y de comprimento igual a quatrocentos valores variando entre zero e dez. E por fim, utilizou-se a função *RFsimulate* de acordo com o modelo de covariância obtido anteriormente para a geração de 160.000 observações em um grid regular.

Simulação para cálculo computacional

Para calcular o tempo computacional gasto foi realizado uma simulação baseada no modelo exponencial com variância igual a dois, sendo este modelo o mais comum quando se estuda dependência espacial (CARDUCCI et al., 2014c) e a variância é definida de forma arbitrária. Esse processo foi repetido para vários tamanhos de amostras, no qual todas as amostras foram feitas utilizando grids regulares. Os valores dos tamanhos das amostras podem ser observados na Tabela 1.

Tabela 1: Tamanho das amostras utilizadas no cálculo do tempo computacional

Modelo	Tamanho da amostra (n)	Tamanho da amostra (n)
E	100	3600
X	200	4225
P	484	4900
O	529	6400
N	625	8100
E	900	10000
N	1024	12100
C	1225	14400
I	1600	16900
A	2025	22500
L	2500	40000
	3025	

Aplicação em dados reais

Com o intuito de aplicar a metodologia proposta em dados reais foram utilizados dados de solo sob plantações de café. O estudo foi realizado em uma área de $52ha$ localizada no município de São Roque de Minas – MG, com coordenadas 20º11'31" S e 46º22'07" W e altitude de 826m. O solo foi classificado como Latossolo Vermelho Amarelo distrófico cambissólico (LVDa) (CARDUCCI et al., 2014a, 2014b, 2014c, 2015). A variável em estudo foi a densidade do solo medida em uma profundidade de 0,20-0,34m. Para a obtenção da densidade do solo foram utilizadas imagens tomográficas com 550 secções variando de 0 a 35mm. O periodograma foi estimado pela equação (1). Procedeu-se com o ajuste dos modelos através da equação (4) e da equação (7).

Diversos critérios para seleção de modelos são apresentados na literatura (Bozdogan, 1987; BURNHAM, 2004). Neste trabalho foram utilizados o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC).

O critério de informação de Akaike (AIC) foi proposto por Akaike (1974) é definido por ele como:

$$AIC = -2\log\left(\mathrm{L}\left(\hat{\theta}\right)\right) + 2p \ , \qquad (1.9)$$

em que, $\log\left(L\left(\hat{\theta}\right)\right)$ é a log-verossimilhança estimada e p é o número de parâmetros a serem estimados no modelo.

O Critério Bayesiano de Schwarz (BIC) foi proposto por Schwarz (1978) e é definido por:

$$BIC = -2\log\left(\mathrm{L}\left(\hat{\theta}\right)\right) + 2p\log(n) \ , \qquad (1.10)$$

em que, $\log\left(L\left(\hat{\theta}\right)\right)$ é a log-verossimilhança estimada, p é o número de parâmetros a serem estimados no modelo e n é o tamanho da amostra. Sendo o modelo preferível aquele que apresentar a menor medida de *AIC* e de *BIC*.

Resultados e discussão

Aplicação teórica

Com o auxílio do software estatístico R e do pacote *fractaldim*, utilizando-se os dados simulados, estimou-se o periodograma para 150 distâncias h, por meio da função *fd.estim.periodogram*. A distribuição empírica do periodograma espacial é apresentada na Figura 1.

Para ajustar o modelo exponencial teórico – equação (4) – ao periodograma espacial, foi utilizada a função *nls* do software R. A Tabela **2** apresenta as estimativas dos parâmetros do modelo exponencial e a Soma Quadrática do Resíduo (SQR).

Tabela 2: Estimativas dos parâmetros da função de densidade exponencial.

σ	α	*SQR*
52.78	*0.51*	*227.1*

Com essas estimativas tem-se que a função de densidade espectral estimada de um processo espacial isotrópico com uma covariância exponencial é:

$$\hat{f}(\omega) = \frac{1}{2}0.51(\pi 52.78)^{-\frac{1}{2}}\exp\left(-\frac{\omega^2}{4(52.78)}\right) \qquad (11)$$

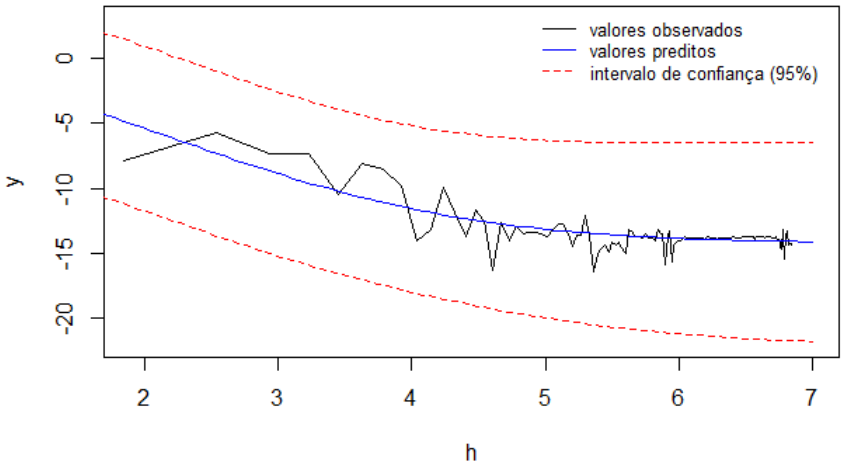

Figura 1: Periodograma espacial e modelo exponencial ajustado.

Sendo a covariância estimada definida como:

$$\hat{C}(h) = 52.78 \exp\left\{-0.51h^2\right\}. \qquad (12)$$

Conhecendo a covariância pode-se agora encontrar o semivariograma estimado pela equação(8). A Figura 2 apresenta a covariância estimada pela equação (12) e sua respectiva semivariância.

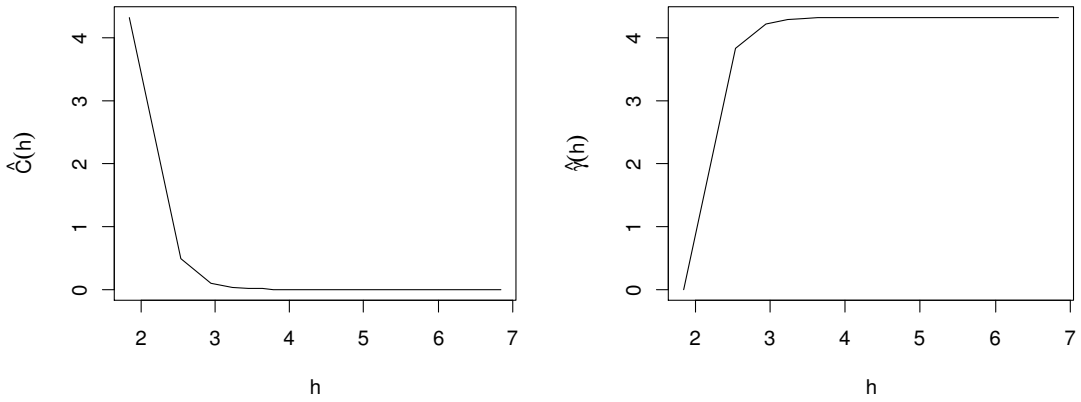

Figura 2: Covariância e semivariograma estimado.

Comparação de tempo computacional

Ao comparar o tempo computacional entre o método de estimação clássico do semivariograma e o método espectral para todas as configurações da Tabela 1, o método espectral mostrou-se mais eficiente.

Observa-se na Figura 3 a representação do tempo computacional gasto com relação ao tamanho da amostra. Verifica-se que para amostras maiores que 20100, o tempo computacional gasto não apresenta uma diferença significativa, ou seja, permanece aproximadamente constante. Por outro lado, o tempo do método de estimação clássico aumenta à medida que o tamanho da amostra aumenta.

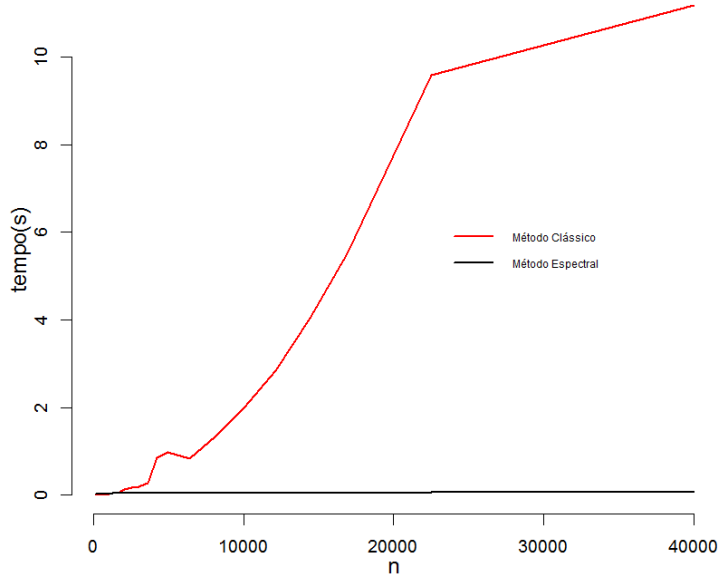

Figura 3: Comparação do tempo computacional entre o método espectral e o método clássico (estimador clássico de semivariância).

Pelos resultados obtidos observou-se que, para amostras de tamanho até 1024, o método clássico foi mais rápido, sendo que o tempo computacional dispendido ficou entre 33% e 67% do tempo gasto pelo método espectral. Já para amostras de tamanho entre 1024 a 1600 é possível observar tempos iguais para ambos os métodos. Para tamanho de amostras maiores que 1600 observações, o tempo do método clássico aumentou gradativamente, conforme pode-se verificar pela Figura 4. Para uma amostra de tamanho 40000, o método clássico gastou 18650% a mais de tempo do que o método espectral.

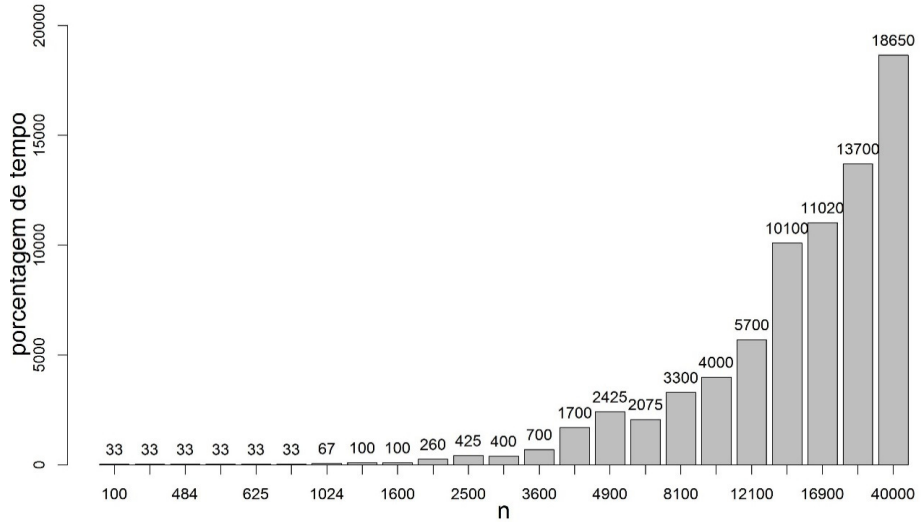

Figura 4: Tempo computacional gasto pelo método de estimação clássico para diferentes tamanho de amostras.

Aplicação em dados reais

A Figura 5 apresenta o boxplot e o histograma da densidade do solo. A densidade máxima foi de $851hu$, mínima de $761,9hu$, com média de $809,2hu$ e desvio padrão de 21,58. A Figura 6a apresenta a variação da densidade em função das 550 secções com o lag variando de 0 a 35 mm, em que o lag é a tolerância correspondente a um intervalo de distância fixado. Vale salientar que há grande variabilidade da densidade em função do lag, conforme pode se observa pela Figura 6a.

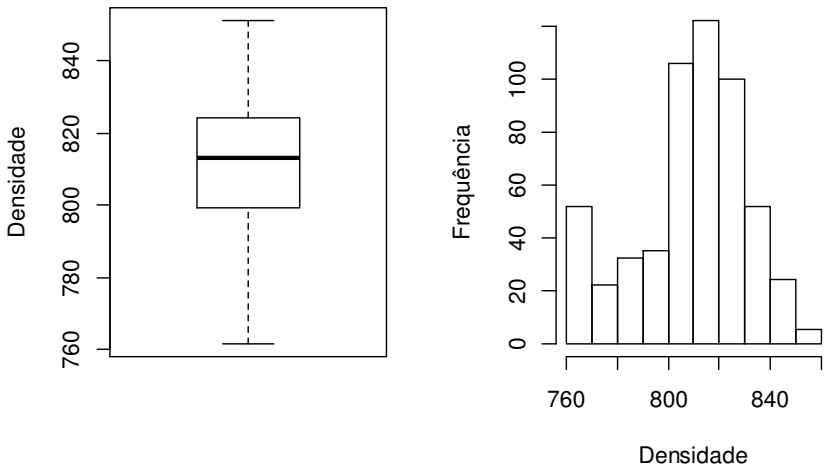

Figura 5: Boxplot e histograma da densidade do solo

Assim, realizou-se a estimativa da densidade espectral a partir do periodograma. A Figura 6b apresenta o periodograma espacial ajustado pelo modelo $f(\omega) = \phi \|\omega\|^{(-2\nu-d)}$. O referido modelo apresentou um AIC de 1195,74 e BIC de 1206,73 ante um AIC de 4695,77 e BIC de 4706,76 do modelo $f(\omega) = \frac{1}{2}\sigma(\pi\alpha)^{-\frac{1}{2}} \exp\left(-\frac{\omega^2}{4\alpha}\right)$. O erro do modelo selecionado pode ser considerado normal pelo teste de Shapiro-Wilk a 5% de significância.

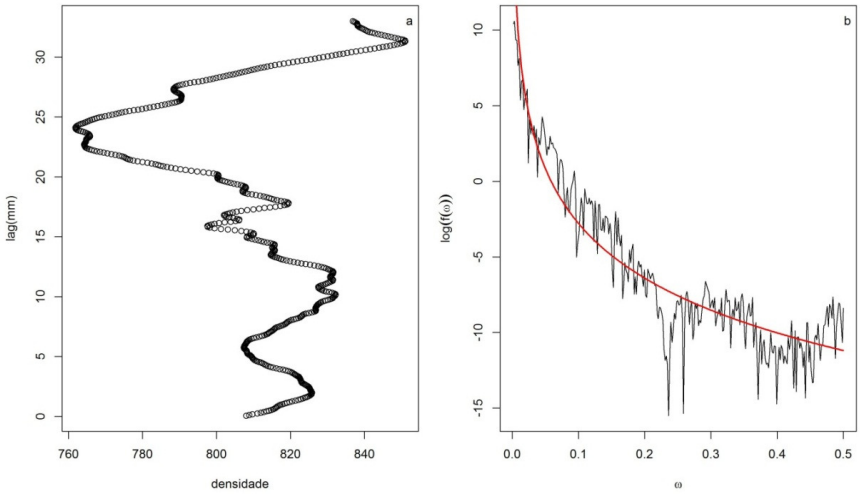

Figura 6: (a) Densidade do solo em função do lag; (b) Periodograma espacial

Conclusões

Com as estimativas obtidas no ajuste da função de densidade espectral, mostrou-se que foi possível estimar a covariância, a qual está diretamente relacionada com a semivariância. Ao comparar o tempo computacional de estimação entre o método clássico e o método espectral, observou-se que o método espectral foi mais eficiente à medida que o tamanho da amostra aumenta (n maior que 1600). Com a aplicação em dados reais, mostrou-se que através do periodograma espacial é possível fazer a representação espectral de um processo espacial. Dois modelos de densidade espectral foram ajustados, sendo que o modelo linear foi o que melhor se ajustou aos dados segundo os critérios de AIC e BIC.

Referências

ARAÚJO, E. S. B.; BATISTA, L. S.; SCALON, J. D. Análise espacial espectral bidimensional em partículas de material compósito Al/SiC. **TEMA (São Carlos)**, v. 15, n. 1, p. 73-81, 2014.

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**., Boston, v.19, n.6, p.716-723, Dec. 1974.

BOZDOGAN, Hamparsum. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. **Psychometrika**, v. 52, n. 3, p. 345-370, 1987.

BURNHAM, Kenneth P.; ANDERSON, David R. Multimodel inference understanding AIC and BIC in model selection. **Sociological methods & research**, v. 33, n. 2, p. 261-304, 2004

CARDUCCI, C. E. et al. **Scaling of pores in 3D images of Latosols (Oxisols) with contrasting mineralogy under a conservation management systemSoil Research**, 2014a. Disponível em: <http://dx.doi.org/10.1071/SR13238>

CARDUCCI, C. E. et al. Distribuição espacial das raízes de cafeeiro e dos poros de dois Latossolos sob manejo conservacionista. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 18, n. 3, p. 270–278, mar. 2014b.

CARDUCCI, C. E. et al. Spatial variability of pores in oxidic latosol under a conservation management system with different gypsium doses. **Ciência e Agrotecnologia**, v. 38, n. 5, p. 445–460, out. 2014c.

CARDUCCI, C. E. et al. Gypsum effects on the spatial distribution of coffee roots and the pores system in oxidic Brazilian Latosol. **Soil and Tillage Research**, v. 145, p. 171–180, jan. 2015.

CASAIS, R. M. C. **CONTRIBUTIONS TO SPECTRAL SPATIAL STATISTICS**. Santiago de Compostela: Universidade de Santiago de Compostela, 2006.

CHERRY, S. **Nonparametric estimation of the sill in geostatistics**. [s.l.] Environmetric, 1997.

CRESSIE, N. A. C. **Fitting variogram models by weighted least squares**. [s.l: s.n.], 1985.

CRESSIE, N. A. C.; HAWKINS, D. M. **Robust estimation of the variogram: Mathematical Geology.** [s.l: s.n.]. v. 12

FUENTES, M. Spectral methods for nonstationary spatial processes. **Biometrika**, v. 89, n. 1, p. 197–210, 2002.

GELFAND, A. et al. **Handbook os Spatial Statistics**. London: CRC Press, 2010.

GRINGARTEN, E.; DEUTSCH, C. V. Teacher's Aide Variogram Interpretation and Modeling. **Mathematical Geology**, v. 33, n. 4, p. 507–534, 2001.
JOURNEL, Andre G.; HUIJBREGTS, Ch J. **Mining geostatistics**. Academic press, 1978.

KIM, H.; FUENTES, M. Spectral Analysis with Spatial Periodogram and Data Tapers. In: **Proceedings Joint Statistical Meeting**. 2000.

MATEU, J; JUAN, P; PORCU, E. Geostatistical Analysis Through Spectral Techniques: Some Words of Caution. **Communications in Statistics - Simulation and Computation**, v. 36, n. 5, p. 1035–1051, 2007.

MALLAT, S. G. Multiresolution approximations and wavelet orthonormal bases of l2(r). Transactions of the American mathematical society, 1989.

MELLO, JM de et al. Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de Eucalyptus grandis. Scientia Forestalis, v. 69, n. 1, 2005.

R CORE TEAM. **R: A language and environment for statistical computing**Vienna, AustriaR Foundation for Statistical Computing, 2015.
VIDAKOVIC, B.; MUELLER, P. **Wavelets for kids.** Instituto de Estadística, Universidad de Duke, 1994.

SCHLATHER, M. et al. Analysis, Simulation and Prediction of Multivariate Random Fields with Package RandomFields. **Journal of Statistical Software**, v. 63, n. 1, p. 1–25, 2015.

SCHWARZ, Gideon et al. Estimating the dimension of a model. **The annals of statistics**, v. 6, n. 2, p. 461-464, 1978.

SEVCIKOVA, H.; PERCIVAL, D.; GNEITING, T. **fractaldim: Estimation of fractal dimensions**, 2014. Disponível em: <http://cran.r-project.org/package=fractaldim>

YAMAMOTO, Jorge Kazuo; LANDIM, Paulo M. Barbosa. **Geoestatística: conceitos e aplicações**. Oficina de Textos, 2013.

ZIMMERMAN, D. L.; ZIMMERMAN, M. B. **A Monte Carlo comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors**. 33. ed. [s.l.] Technometrics, 1991.