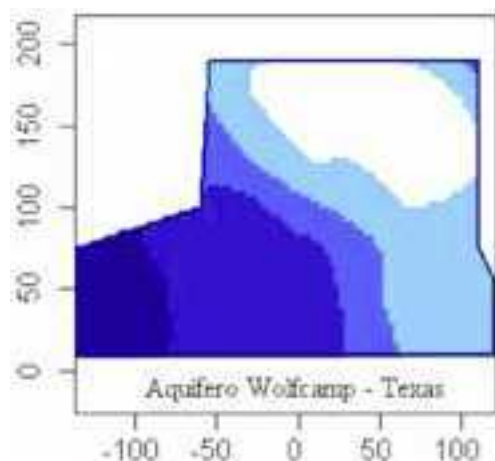


DETERMINAÇÃO DE CLASSES EM MAPA TEMÁTICO UTILIZANDO O MÉTODO DE OPTIMIZAÇÃO DE JENKS

Gustavo Henrique Dalposso¹, Marcio Paulo Oliveira¹



RESUMO

O objetivo deste trabalho é apresentar o método de otimização de Jenks como alternativa para determinar os intervalos de classe em mapas temáticos gerados por métodos geoestatísticos. As análises foram realizadas utilizando o conjunto de dados do Aquífero Wolfcamp. Devido a assimetria dos dados, realizou-se uma transformação Box-cox e utilizou-se o estimador robusto de Cressie & Hawkins para estimar o semivariograma empírico dos resíduos. Utilizando o método de estimação de mínimos quadrados ajustou-se diversos modelos teóricos de semivariograma e por meio de estatísticas de validação cruzada escolheu-se

um modelo esférico. Os parâmetros obtidos no ajuste foram utilizados para elaborar a matriz dos valores preditos por krigagem universal. Os elementos da matriz dos valores preditos foram classificados utilizando intervalos de mesma amplitude e otimização de Jenks, permitindo elaborar dois mapas temáticos. O índice de acurácia tabular (TAI) foi utilizado para identificar o melhor mapa temático por meio da classificação. O resultado mostrou que a melhor classificação foi obtida pelo método de otimização de Jenks.

Palavras-chave: geoestatística, intervalos de classe, otimização de Jenks.

1. INTRODUÇÃO

A geoestatística tem por objetivo a caracterização da dispersão espacial e espaço-temporal das grandezas que definem a quantidade e a qualidade de recursos naturais, tais como florestas, recursos geológicos, hidrológicos, ecológicos, ou outros fenômenos espaciais em que os atributos manifestam uma estrutura de dispersão no espaço ou no tempo, como a contaminação de solos e aquíferos, temperatura e pluviometria de uma região. O seu corpo metodológico consiste, basicamente, num conjunto de instrumentos estatísticos que quantificam a continuidade da grandeza

em estudo, em modelos de interpolação tendo por base a sua variabilidade estrutural (Soares, 2000).

Todo o mapa ilustra um tema e até os mapas topográficos não escapam à regra. Dessa forma, define-se como mapa temático todo o mapa que representa qualquer tema, além da representação do terreno. Os procedimentos de levantamento, redação e comunicação de informações por meio de mapas, diferem de acordo com a formação e especialização dos profissionais em cada campo, a exemplo dos geólogos, geomorfólogos, geógrafos, entre outros, que se expressam na forma gráfica (Joly, 2005).

Na análise geostatística é comum o uso de semivariogramas para descrever a estrutura de dependência espacial. Quando é identificada discrepância nos dados em relação à distribuição normal é conveniente utilizar o estimador de Cressie e Hawkins, visto que este fornece uma estimativa do semivariograma razoavelmente robusta e estável (Cressie & Hawkins, 1980).

Escolher um modelo de semivariograma apropriado não é um procedimento automático. Esta etapa da análise requer um bom julgamento baseado na experiência e uma compreensão das limitações matemáticas da função (McBratney & Webster, 1986).

Como é grande a diversidade de modelos e métodos de ajustes de semivariogramas, a escolha pode ser feita utilizando validação cruzada, pois após validação cruzada bem sucedida do modelo do semivariograma, pode-se ter assegurado em que a predição baseada no modelo variográfico é não-viciada e o erro quadrático médio de predição é minimizado (Cressie, 1991).

O modelo escolhido fornece parâmetros necessários para a krigagem, nome dado ao interpolador geostatístico. No método da krigagem, os pesos são atribuídos de acordo com a variabilidade espacial expressa no semivariograma, no entanto, o que torna a krigagem um interpolador ótimo é a maneira como os pesos são distribuídos, não sendo tendencioso, tendo variância mínima e possibilitando que se conheça a variância da estimativa (Oliver & Webster, 1990).

O problema de especificar as amplitudes das classes em mapas temáticos tem sido foco de estudo de várias áreas das ciências e tecnologias. Geralmente recomenda-se o uso de cinco classes em mapas temáticos (Archela & Théry, 2008), no entanto, o modo de determinação dos limites dessas classes não é especificado.

Objetivou-se, no presente trabalho, realizar uma análise geostatística e utilizar os métodos de intervalos iguais e otimização de Jenks para elaborar os intervalos de classe do mapa temático. A seleção do mapa temático mais acurado considerou o índice de acurácia tabular (TAI) (Armstrong et al., 2003).

2. MATERIAL E MÉTODOS

A análise geostatística foi realizada utilizando o conjunto de dados do aquífero Wolfcamp, no Texas (Harper & Furr, 1986), que é composto de 85 medidas do nível piezométrico, em pés, acima do nível do mar (ft). A importância da análise deve-se ao fato de que na década de 80 o Departamento de Energia dos EUA estudava um depósito de resíduos nucleares no Texas e a contaminação do aquífero foi

uma preocupação, visto que muitos agricultores utilizam os poços que são mantidos pelo aquífero (Cressie, 1991).

Com objetivo de identificar o comportamento inicial dos dados realizou-se uma análise descritiva, de modo que a localização das 85 amostras não foi considerada. Após a análise inicial construiu-se um gráfico postplot para obter um indicativo do comportamento do nível piezométrico na região amostrada. Este gráfico é rico em informação, pois apresenta a malha de pontos coletados separados por cores ou símbolos identificando os quartis da distribuição dos dados.

Pela distribuição empírica dos dados observou-se que o nível piezométrico apresentou assimetria. Segundo Diggle et al. (2003), quando o ajuste de um modelo Gaussiano não parece razoável, uma alternativa é realizar uma transformação Box-Cox.

Considerando a média como um polinômio de primeira ordem sobre as coordenadas espaciais estimou-se a estrutura de dependência espacial dos resíduos utilizando o estimador de Cressie e Hawkins, apresentado na Equação (1). As semivariâncias foram calculadas utilizando um cutoff de 50% da distância máxima (Clark, 1979).

$$\hat{\gamma}(h) = \frac{1}{2} \frac{\left[\frac{1}{N(h)} \sum_{s(s+h)} |z(s+h) - z(s)|^2 \right]^2}{0,914 + \frac{0,988}{N(h)}} \quad (1)$$

Utilizando o método de ajuste dos mínimos quadrados ordinários ajustou-se ao semivariograma empírico os seguintes modelos teóricos: Exponencial, Esférico, Gaussiano, Circular, Cúbico, Wave e Gneiting.

Para avaliar que modelo melhor se ajustou ao semivariograma empírico utilizou-se validação cruzada, uma técnica de avaliação de erros de estimativas que permite comparar os valores previstos com os amostrados (Isaaks & Srivastava, 1989). A Equação (2) apresenta o erro médio reduzido (SME), que deve ser próximo a zero, a Equação (3) apresenta a raiz quadrada do erro quadrático médio reduzido (RMNSE), que deve ser próximo a 1 (um) e a Equação (4) apresenta o erro absoluto (AE), que deve ser o menor.

$$SME = \frac{1}{l} \cdot \sum_{j=1}^l \frac{\hat{z}(s_j) - z^*(s_j)}{\hat{\sigma}_j} \quad (2)$$

$$RMNSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^l \left[\frac{\hat{z}(s_j) - z^*(s_j)}{\hat{\sigma}_j} \right]^2} \quad (3)$$

$$AE = \frac{1}{l} \sum_{j=1}^l |\hat{z}(s_j) - z^*(s_j)| \quad (4)$$

Com a escolha do melhor modelo, utilizou-se krigagem universal (Journel & Huijbregts, 1978) para obter o nível piezométrico em regiões não amostradas. Para determinar as classes do mapa temático foram utilizadas duas metodologias: intervalos de mesma amplitude e otimização de Jenks.

Intervalos de mesma amplitude são constantemente usados na apresentação de mapas temáticos. A vantagem deste método deve-se ao fato da simplicidade de implementação. Uma desvantagem diz respeito à distribuição estatística dos valores krigados, pois se a distribuição for muito concentrada de um lado, este corte deixa apenas um número muito pequeno de áreas nas classes da perna mais longa da distribuição e, como resultado, a maior parte das áreas será alocada a uma ou duas cores (Druck et al., 2004).

O método de otimização de Jenks ou intervalos naturais (Jenks, 1967) visa minimizar a variância intraclasses e maximizar a variância interclasses, produzindo assim classes homogêneas. Assim, este método forma grupos que são homogêneos internamente e assegura a heterogeneidade entre classes (Dent, 1993).

Após elaborar os dois mapas temáticos, foi calculado o índice de acurácia tabular (TAI) (Armstrong, 2003), apresentado na Equação (5), para identificar o melhor mapa temático por meio da classificação.

$$TAI = 1 - \frac{\sum_{j=1}^k \sum_{i=1}^N |z_{ij} - \bar{z}_j|^2}{\sum_{j=1}^N |z_j - \bar{z}|} \quad (5)$$

A análise foi realizada utilizando o software R (R Development Core Team, 2011) sendo que para a análise geoestatística utilizou-se o pacote geoR (Ribeiro Jr. & Diggle, 2001) e para a classificação de Jenks utilizou-se o pacote classInt (Bivand et al., 2009).

3. RESULTADOS E DISCUSSÃO

A tabela 1 apresenta as estatísticas descritivas do nível piezométrico (ft).

Destaca-se que, em média, o nível piezométrico é de 2002 pés, porém, o coeficiente de variação indica uma alta heterogeneidade dos dados (CV > 30%). O valor da mediana menos o mínimo, 773, difere do valor do máximo menos a mediana, 1774, evidenciando uma distribuição assimétrica, fato este corroborado pelo alto valor do coeficiente de assimetria de Person (1,01). A figura 1 apresenta o gráfico boxplot do nível piezométrico (ft) no aquífero Wolfcamp.

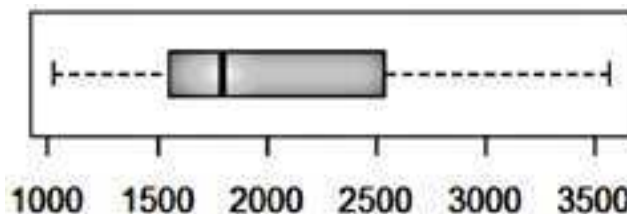


Fig. 1 - Boxplot do nível piezométrico (ft).

Tab. 1 – Estatísticas descritivas do nível piezométrico (ft) do aquífero Wolfcamp

N	85
Mínimo	1024
Média	2002
Primeiro Quartil	1548
Mediana	1797
Terceiro Quartil	2540
Máximo	3571
Desvio Padrão	611,34
Coefficiente de Variação (%)	30,53
Coefficiente de Assimetria	1,01

O boxplot do nível piezométrico não apresenta outliers, porém, fica claramente evidenciada uma assimetria. A figura 2 apresenta o gráfico espacial da área em estudo, representando o nível piezométrico separado por quartis.

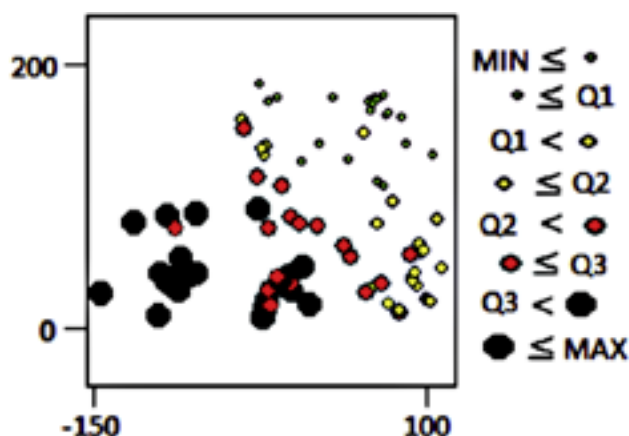


Fig. 2 – Postplot do nível piezométrico (ft).

Observa-se no gráfico espacial que o nível piezométrico apresenta uma tendência de distribuição, pois os valores mais elevados encontram-se no sudoeste e conforme as amostras se distanciam para o nordeste os valores diminuem. O teste de normalidade de Shapiro-Wilk apresentou um nível descritivo (p-valor) de 0,0016 indicando não normalidade dos dados. Diante do fato de que a distribuição dos dados não é gaussiana optou-se por realizar uma transformação boxcox, sendo $\lambda = 0,5$ o valor do lambda ótimo estimado.

A figura 3 apresenta o semivariograma empírico omnidirecional dos resíduos que foi obtido utilizando o estimador de Cressie e Hawkins.

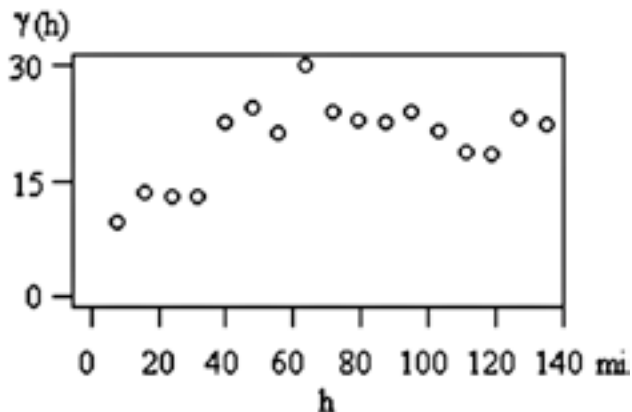


Fig. 3 – Semivariograma omnidirecional.

Destaca-se que a primeira semivariância encontra-se na distância de 5 milhas, indicando que não se tem conhecimento do nível piezométrico a menores distâncias. A maior semivariância encontra-se na distância de 135 milhas, que corresponde a 50% da distância máxima entre amostras. Na maioria das situações práticas, a representatividade de $\gamma(h)$ começa a ser questionável para valores de $|h|$ superiores a metade da área em estudo (Soares, 2000). Cada semivariância foi calculada considerando um mínimo de 30 pares (Guerra, 1988).

A tabela 2 apresenta as estimativas dos parâmetros referentes aos ajustes dos modelos teóricos de semivariograma. Observa-se que o maior raio de dependência espacial é obtido pelo modelo Cúbico com, aproximadamente, 75 milhas. O uso do coeficiente de efeito pepita relativo é mencionado em Isaaks & Srivastava (1989) sendo este utilizado para medir o grau de dependência espacial.

Tab. 2 - Modelos ajustados e parâmetros

Modelo	α	$\phi 1$	$\phi 2$	$\phi 1 + \phi 2$	E(%)
Exp	58,91	0,98	21,90	22,88	4,28
Esf	63,49	5,77	17,05	22,82	25,27
Gaus	54,04	8,39	14,44	22,83	36,74
Circ	54,13	5,62	17,09	22,72	24,74
Cúb	75,07	8,30	14,52	22,81	36,35
Wa	43,94	8,74	13,64	22,38	39,02
Gn	53,15	8,39	14,43	22,82	36,74

α : alcance; $\phi 1$: efeito pepita; $\phi 2$: contribuição; $\phi 1 + \phi 2$: patamar; E: $100 \cdot (\phi 1 / (\phi 1 + \phi 2))$ efeito pepita relativo; Exp: Exponencial; Esf: Esférico; Gaus: Gaussiano; Circ: Circular; Cúb: Cúbico; Wa: Wave e Gn: Gneiting

Segundo Cambardella et al. (1994) o modelo exponencial apresenta forte dependência espacial ($E \leq 25\%$) e os demais modelos apresentam moderada dependência espacial ($25\% < E \leq 75\%$). Como se têm vários modelos, é necessário utilizar técnicas que permitam escolher o melhor ajuste, como por exemplo, a validação cruzada, apresentada na tabela 3.

Tab. 3 – Validação cruzada

Modelo	SME	RMNSE	AE
Exp	0,0634	1,1930	11922,63
Esf	0,0379	1,1021	11709,5
Gaus	0,0326	1,1167	11858,3
Circ	0,0398	1,1130	11762,94
Cúb	0,0324	1,1142	11810,97
Wa	0,0284	1,1716	11809,03
Gn	0,0325	1,1167	11843,14

SME: Erro médio reduzido; RMNSE: raiz quadrada do erro quadrático médio reduzido; AE: Erro absoluto

Observa-se que o modelo Wave apresentou o erro médio reduzido mais próximo de zero e o modelo Esférico apresentou o erro quadrático médio reduzido mais próximo de 1 e o menor erro absoluto e sendo assim, escolhe-se o modelo Esférico como melhor ajuste. A figura 4 apresenta o semivariograma empírico juntamente com o modelo ajustado.

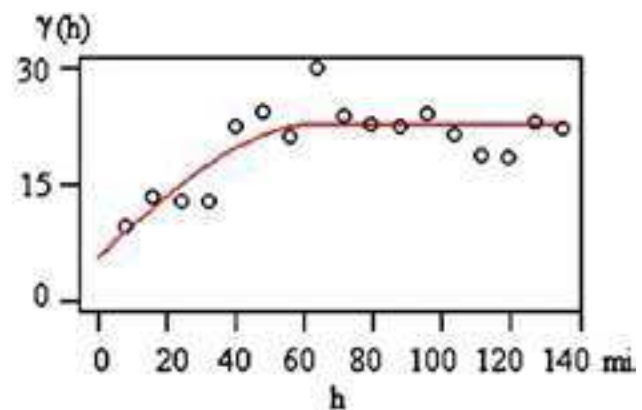


Fig. 4 – Modelo Esférico ajustado.

Destaca-se o raio de dependência espacial de aproximadamente 63,5 milhas. Os parâmetros deste modelo foram utilizados para realizar uma interpolação por krigagem universal. Os dados obtidos foram classificados utilizando intervalos de mesma amplitude e o método de otimização de Jenks. Em ambas as classificações padronizou-se um número de 5 classes e os mapas temáticos do nível piezométrico do aquífero Wolfcamp são apresentados na figura 5.

Observa-se que os mapas são parecidos e refletem o comportamento do nível piezométrico já detectado pelo gráfico postplot, em que a região noroeste apresenta os níveis mais altos e a região nordeste os níveis mais baixos.

Destaca-se claramente que o mapa, utilizando intervalos de mesma amplitude, apresenta uma região mais heterogênea, compreendida entre a terceira e quarta classes, quando comparada com a mesma área no mapa utilizando o método de Jenks. O índice de acurácia tabular calculado para o mapa com classes de mesma amplitude foi de 0,80

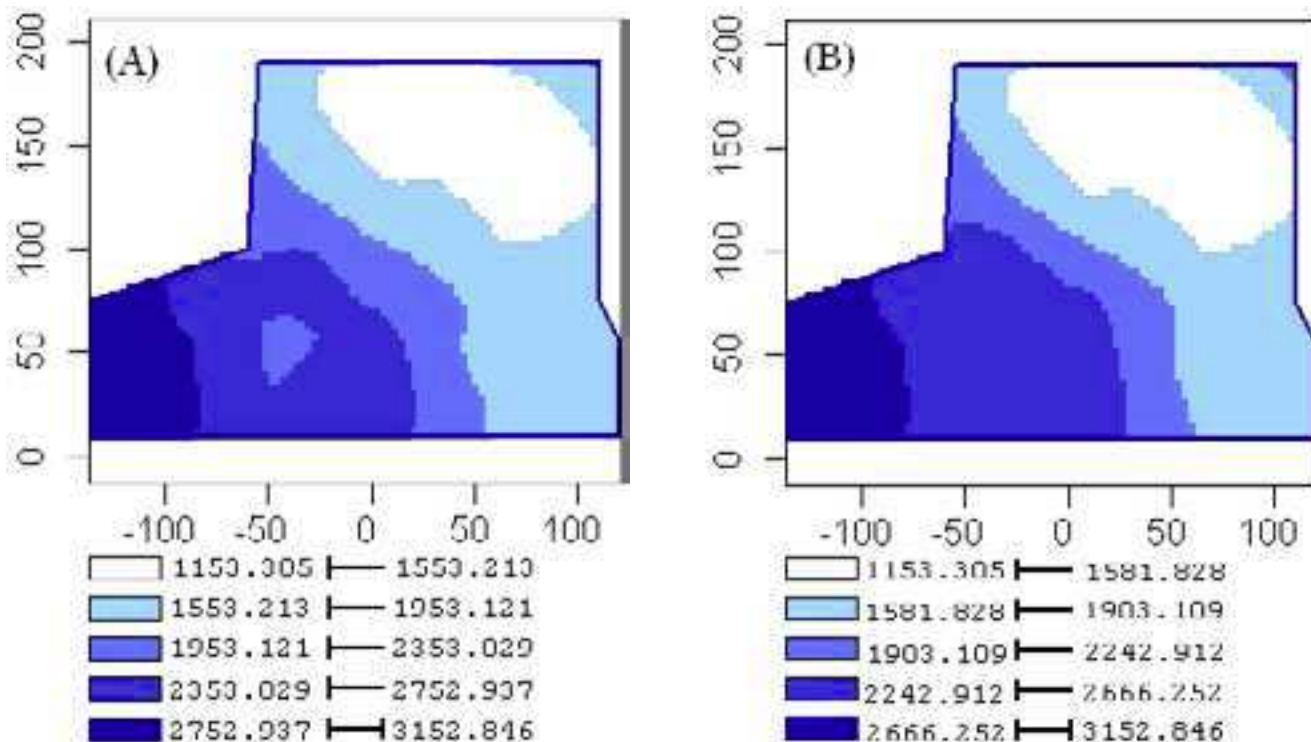


Fig. 5 – Mapa temático do nível piezométrico (ft). (A) utilizando intervalos de mesma amplitude e (B) utilizando o método de otimização de Jenks.

e o mesmo índice para o mapa com classes obtidas por otimização de Jenks foi 0,82 logo, este mapa é mais adequado para a representação da região.

4. CONCLUSÕES

A teoria geoestatística foi capaz de detectar e descrever a variabilidade espacial do nível piezométrico no aquífero Wolfcamp.

O método de otimização de Jenks proporcionou uma divisão em que os dados ficam homogêneos nas classes e heterogêneos interclasses. Esta classificação é adequada para a construção de mapas utilizados em agricultura de precisão.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Archela, R. S. e Théry, H. Orientação metodológica para construção e leitura de mapas temáticos. Retirado em 23 junho de 2008, Disponível em World Wide Web: <http://confins.revues.org/index3483.html>

Armstrong, M.P.; Xiao, N.; Bennett, D.A. Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps. *Annals of the Association of American Geographers*, 93(3):595–623, 2003.

Bivand, R.; Ono, H.; Dunlap, R. classInt: Choose univariate class intervals. Retirado em 06 de junho de 2009, disponível em World Wide Web: <http://cran.r-project.org/web/packages/classInt/index.html>.

Cambardella, C.A.; Moorman, T.B; Novack, J.M; Parkin, T.B; Karlen, D.L; Turco R.F.; Knopka, A.E. Field-scale variability of soil properties in central Iowa soils. *Soil Science Society America Journal*, Madison, 58:1240-1248, 1994.

Clark, I. *Practical Geostatistics*. 1979. Essex, Applied Science Publishers, 129p.

Cressie, N. *Statistics for spatial data*. 1991. New York, John Wiley, 900p.

Cressie, N.; Hawkins, D.M. Robust estimation of the variogram. *Mathematical Geology*. New York, 12(2): 241-8, 1987.

Dent, D.B. 1993. *Cartography: Thematic Map Design*. WCB, England.

Diggle, P. J.; Ribeiro Jr, P. J.; Christensen, O.F. Na introduction to model-based geostatistics, in Moller (ed.), *Spatial Statistics and Computational Methods*, Springer, pp. 43-86, 2003.

Druck, S.; Carvalho, M.S.; Câmara, G.; Monteiro, A.V.M. *Análise Espacial de Dados Geográficos*. Brasília, EMBRAPA, 2004.

Guerra, P.A.G. *Geoestatística Operacional*. 1988. Brasília, Ministério das Minas e Energia, 145p.

Harper, W.V and Furr, J.M. *Geostatistical analysis of potentiometric data in the Wolfcamp Aquifer of the Palo Duro Basin, Texas*. Technical Report BMI/ONWI-587, Bettelle Memorial Institute, Columbus, OH, 1986.

Isaaks, E.H.; Srisvastava, R.M. Na introduction to applied geostatistics. 1989. New York, Oxford University Press, 560p.

Jenks, George F. 1967. “The Data Model Concept in Statistical Mapping”, *International Yearbook of Cartography* 7: 186-190.

Joly, F. *A Cartografia Editora Papirus*, São Paulo, 2005 (8ª edição).

Journel, A.; Huijbregts, C. 1978. *Mining geostatistics*. London, Academic Press.

McBratney, A.B.; Webster, R. Choosing functions for semivariograms of soil properties and fitting them to sample estimates. *Journal of Soil Science*, Oxford, 37: 617-39, 1986.

Oliver, M. A.; Webster, R. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, Hants, 4(3): 313-32, 1990.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr., P.J. & Diggle, P.J. 2001. *geoR: A package for geostatistical analysis*. *R NEWS*, Vol 1, No 2, 15-18.

Soares, A. 2000. *Geoestatística para as Ciências da Terra e do Ambiente*. Coleção Ensino da Ciência e Tecnologia, nº 9, Lisboa, ISTPress. 206p.

AGRADECIMENTOS

A CAPES e ao CNPq.

- (1) Matemático, Mestre em Engenharia Agrícola, Universidade Tecnológica Federal do Paraná, Campus Toledo. Brasil.