

**ESCOLA SUPERIOR DE AGRICULTURA “LUIZ DE QUEIROZ”
UNIVERSIDADE DE SÃO PAULO
DEPARTAMENTO DE CIÊNCIAS EXATAS**

Relatório de Análise

Tiago Viana Flor de Santana

**PIRACICABA-SP
2012**

SUMÁRIO

1 INTRODUÇÃO	1
2 REVISÃO DE LITERATURA	2
3 MATERIAL E MÉTODOS	6
4 RESULTADOS E DISCUSSÃO	7
4.1 Análise espacial dos dados	7
5 CONSIDERAÇÕES FINAIS	16
REFERÊNCIAS BIBLIOGRÁFICAS	17

1 INTRODUÇÃO

Os primeiros trabalhos e estudos sobre geoestatística datam de 1950, proveniente da busca por metodologias que auxiliassem na exploração de minas de ouro. Destacam-se como estudos pioneiros na área as publicações de Daniel G. Krige, engenheiro de minas, do estatístico H. S. Sichel e, no início dos anos 60, G. Matheron que formalizou a teoria (Landim, 2006).

A metodologia de análise geoestatística difere-se da estatística clássica, pois leva em consideração na análise estatística a posição geográfica em que cada observação é tomada.

Em geral, no conjunto de dados a ser analisado, obtém-se apenas uma única observação por posição amostrada impossibilitando a análise pelos métodos estatísticos tradicionais, uma vez que o princípio da repetição não é satisfeito. Outra exigência da estatística clássica é a independência das observações, porém para muitos fenômenos físicos essa exigência não tem sentido pois, por exemplo, é impossível prever com exatidão o teor de um minério num determinado ponto de uma jazida, mas é provável que se encontre minério rico perto de minério rico.

As variáveis em um estudo geoestatístico tem duas características importantes: São aleatórias uma vez que os valores numéricos observados podem variar consideravelmente de uma posição a outra no espaço e são espaciais, pois apesar de variar no espaço preservam algum grau de dependência, no sentido de que variáveis mais próximas tendem a apresentar valores parecidos, (Guerra, 1988).

Dessa forma, o objetivo desse estudo é aplicar a metodologia de análise geoestatística em um conjunto de dados de solo e modelar o padrão espacial e descrever a distribuição espacial do solo.

2 REVISÃO DE LITERATURA

Conforme define Diggle, Ribeiro Jr. e Christensen (2003), a geoestatística é um ramo da estatística espacial na qual os dados consistem de mensurações y_1, y_2, \dots, y_n obtidas nas localizações x_1, x_2, \dots, x_n amostradas em uma região $A \subset \mathbb{R}^2$ espacialmente contínua.

As mensurações estão relacionadas a um fenômeno espacial que pode ser tratado como a realização de um processo estocástico $S(x), x \in \mathbb{R}^2$, denominado *signal* que, em geral, não é diretamente observável.

A terminologia “geoestatística baseada em modelos” foi introduzida por Diggle, Tawn e Moyeed (1998). É caracterizada pela declaração explícita e completa de um modelo, o que significa que a modelagem é realizada adotando-se um modelo paramétrico para a(s) variável(veis) resposta, e a utilização de métodos de inferência estatística baseadas na verossimilhança aplicada a problemas geoestatísticos. Assim, um modelo geoestatístico é a especificação da distribuição conjunta $[Y(x), S(x)]$, usualmente apresentada na forma fatorada

$$[S(x), Y(x)] = [S(x)][Y(x)|S(x)].$$

De acordo com Diggle, Ribeiro Jr. e Christensen (2003), no modelo geoestatístico gaussiano estacionário e isotrópico, $S(x)$ é um processo gaussiano estacionário, com $E(S(x)) = \mu$; $Var(S(x)) = \sigma^2$, e função de correlação $\rho(u_{ij}) = Corr(S(x_i), S(x_j))$ em que $u_{ij} = \|x_i - x_j\|$, $i, j = 1, \dots, n$, $\rho(u) = \rho(-u)$ e $Y(x_i)$ o valor observado na localização x_i .

Esses autores consideram plausível um modelo geoestatístico em que a distribuição de $Y(x_i)$, $i = 1, \dots, n$, condicionada à distribuição $S(\cdot)$, é gaussiana com média $S(x_i)$ e variância τ^2 ; e que $Y(x_i)$ são mutuamente independentes condicionados em $S(\cdot)$.

O modelo pode, então, ser escrito na forma

$$Y(x_i) = S(x_i) + Z(x_i) \quad i = 1, \dots, n,$$

em que $Z(x_i) \sim N(0, \tau^2)$ são erros aleatórios independentes de $S(x)$ e o valor observado y_i é uma realização de $Y(x_i)$ que é uma versão ruidosa de $S(x_i)$ no caso gaussiano.

Segundo Schabenberg e Gotway (2005), um processo estocástico $S(x)$ é estacionário se sua

distribuição espacial é invariante sob translação das coordenadas. Bailey e Gatrell (1995) mostram que isto acontece se as propriedades estatísticas $S(x)$ são independentes da localização absoluta na região de estudo $A \subset \mathbb{R}^2$. Isto implica que a média e a variância são constantes na região A e não dependem da localização x . Implica também que a covariância $Cov(S(x_i), S(x_j))$, $i \neq j$, dependerá somente das localizações relativas destes dois pontos, da distância u que as separa e da direção entre elas, e não de sua localização absoluta na região.

A função covariância é denominada isotrópica na ausência de dependência da direção, ou melhor, quando depende somente da distância absoluta entre os pares de pontos. O processo espacial é isotrópico se, em acréscimo à estacionariedade, a covariância depender somente da distância entre dois pontos, e não da direção nos quais estão separados (BAILEY, GATRELL, 1995).

Na estrutura do modelo proposto, a família *Matérn* de funções de correlação, apresentada em Diggle e Ribeiro Jr. (2007), é uma importante função paramétrica de correlação com expressão analítica dada por

$$\rho(u, k, \phi) = \frac{1}{2^{k-1} \Gamma(k)} \left(\frac{u}{\phi} \right)^k K_k \left(\frac{u}{\phi} \right),$$

em que $K_k(\cdot)$ denota a função de Bessel modificada de ordem k , $\phi > 0$ é um parâmetro de escala, associado à extensão de dependência espacial (alcance) e $k > 0$ é um parâmetro de forma que determina a suavidade analítica do processo $S(x)$, interpretado como uma medida da diferenciabilidade do processo.

Devido à dificuldade de identificação de todos os parâmetros do modelo, na prática, os valores de k podem ser escolhidos dentre os três valores $\{0,5; 1,5; 2,5\}$ correspondendo, respectivamente, à não diferenciabilidade e a processos estocásticos uma e duas vezes diferenciáveis na origem.

Nessa família Matérn, fazendo $k = 0,5$, obtém-se a função de correlação exponencial

$$\rho(u) = \exp \left(-\frac{u}{\phi} \right), \quad u_0 = 3\phi \quad \text{e} \quad \lim_{k \rightarrow \infty} \rho(u) = \exp \left\{ -\left(\frac{u}{\phi} \right)^2 \right\},$$

obtendo-se a função de correlação gaussiana para a qual $u_0 \simeq \sqrt{3}\phi$.

Quando não existe estacionariedade na média, a situação mais comum é que $\mu(x)$, denominada *superfície de tendência*, seja escrita como um modelo de regressão polinomial usando potências e produtos cruzados das coordenadas cartesianas de x como variáveis explicativas.

Uma superfície de tendência linear pode ser escrita como

$$\mu(x) = \beta_0 + \sum_{j=1}^p \beta_j d_j(x),$$

em que $d_j(x)$ são variáveis explicativas, dependentes ou não das coordenadas.

Quando a variável observada $Y(x)$ não segue uma distribuição gaussiana ou em processos de variância não estacionária, o valor do erro quadrático médio mínimo do preditor é afetado, com a consequência de se obter aproximações ruins. Em muitos casos, através de transformações, é possível que os dados passem a seguir uma distribuição gaussiana.

Box e Cox (1964) propõem a família de transformações

$$Y^*(x) = \begin{cases} \frac{Y^\lambda(x) - 1}{\lambda}, & \lambda \neq 0; \\ \log(Y(x)), & \lambda = 0, \end{cases}$$

em que o parâmetro λ introduz flexibilidade ao modelo. Alguns valores interpretáveis para esse parâmetro são:

$$\begin{aligned} \lambda = 1,0 & : \text{ sem transformação;} \\ \lambda = 0,5 & : \text{ transformação raiz quadrada;} \\ \lambda = -1,0 & : \text{ transformação recíproco.} \end{aligned}$$

Os métodos usuais de estimação dos parâmetros de um modelo gaussiano linear são o método dos mínimos quadrados e o método da máxima verossimilhança. Utilizando o método da máxima verossimilhança, pode-se obter estimativas consistentes dos parâmetros de média e covariância maximizando a função de log-verossimilhança.

Admitindo uma superfície de tendência polinomial para $\mu(x)$, tem-se

$$Y(x) \sim N_n(\mathbf{D}\boldsymbol{\beta}, \boldsymbol{\sigma}^2 \mathbf{R} + \boldsymbol{\tau}^2 \mathbf{I})$$

em que \mathbf{D} é uma matriz $n \times p$ de covariáveis, $\boldsymbol{\beta}$ é o vetor de parâmetros da regressão e a matriz de correlação \mathbf{R} depende de ϕ e da distância, os parâmetros da função de correlação exponencial. A função de verossimilhança de $Y(x)$ é dada por:

$$L(Y(x)) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\sigma}^2 \mathbf{R} + \boldsymbol{\tau}^2 \mathbf{I}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (Y(x) - \mathbf{D}\boldsymbol{\beta})' (\boldsymbol{\sigma}^2 \mathbf{R} + \boldsymbol{\tau}^2 \mathbf{I})^{-1} (Y(x) - \mathbf{D}\boldsymbol{\beta}) \right\}$$

Outra etapa das análises geoestatísticas, consta da seleção do modelo, que se julga satisfatório para proceder à interpolação, após o procedimento de ajuste. Além da apreciação visual do ajuste da curva sobre as semivariâncias, existem índices e procedimentos que se constituem em critérios que possibilitam uma avaliação sem a subjetividade observada na avaliação visual. É o caso do Critério de Informação de Akaike (AIC - Akaike Information Criterion) e da Validação Cruzada.

O AIC (Akaike, 1994) foi estabelecido a partir da distância de Kulback e Leibler (1951), que corresponde a uma medida da discrepância entre as curvas de um modelo verdadeiro e de um modelo aproximado.

$$AIC = -2\log(L) + 2p_i,$$

sendo L a função de verossimilhança e p o número de parâmetros ajustados.

3 MATERIAL E MÉTODOS

Os dados experimentais, que fazem parte deste trabalho, foram coletados do levantamento detalhado de solos da Estação Experimental de Campos, Fazenda Angra (Capeche et al., 1997), por pesquisadores da PESAGRO e Embrapa Solos. No estudo pedológico foram avaliadas as características morfológicas, físicas e químicas dos solos, e apresentadas também, informações referentes à distribuição geográfica. Para o estudo geoestatístico foi considerada a variável agrônômica teor de Areia (%), Argila (%) e Silte (%), na camada de 0-20cm.

A área compreendida pelo levantamento situa-se ao Norte do estado do Rio de Janeiro, à margem esquerda do Rio Paraíba do Sul, no município de Campos dos Goytacazes, entre os paralelos $21^{\circ}44'47''$ e $41^{\circ}18'24''$ WGr.

A Estação Experimental de Campos da PESAGRO-RIO abrange uma área de 105 hectares, e faz parte de uma das áreas mais tradicionais do cultivo da cana-de-açúcar no Brasil. Esta cultura se desenvolveu em áreas denominadas de tabuleiros: Podzólico Amarelo e Latossolo Amarelo e Baixada: associações e complexos de solos Hidromórficos e Cambissolos desenvolvidos de substratos colúvio-aluvionais (Embrapa Solos, 1997). Devido às enchentes do Rio Paraíba do Sul, à esquerda da fazenda e à direita pelas enchentes das Lagoas dos Prazeres e Taquaraçú, ocorreu a deposição de sedimentos, que contribuiu para a grande variação nas propriedades avaliadas no solo associadas às adubações realizadas na área experimental.

O objetivo principal do levantamento foi o planejamento adequado do uso dos solos possibilitando um manejo racional e sustentável, evitando o processo de erosão dando subsídios em campos experimentais na condução de pesquisas que são extrapoladas para áreas edafoclimáticas similares.

As amostras de solo foram coletadas com um trado tipo holandês, em uma malha regular, sob duas coordenadas geográficas: leste e norte, distando nas duas direções 900 metros. Esta malha foi realizada com espaçamento de 50 metros, totalizando 273 amostras, na camada de 0 - 20cm.

4 RESULTADOS E DISCUSSÃO

4.1 Análise espacial dos dados

A distribuição granulométrica do solo, bem como macro e micronutrientes são fatores importantes para o desenvolvimento de qualquer cultura. Neste estudo observou-se que o teor de areia, silte e argila (%) variou grandemente na região selecionada para a implantação de várias culturas, numa região predominantemente com plantação de cana-de-açúcar, há cem anos. Esta variação se deu nos valores de 10 a 45 no que se refere a variável silte. Os valores mais baixos foram encontrados na região superior à esquerda, região esta, um pouco acidentada e com constantes enchentes nas cheias, pois ao norte desta região está próxima de lagoas.

Inicialmente será realizada uma análise descritiva dos dados da para todas as variáveis em estudo, descrito na Tabela 1.

TABELA 1: Estatísticas descritivas do teor de silte, areia, argila, cálcio que compõem o solo em estudo, dentre os 273 pontos originais.

Variável	Silte(%)	Areia(%)	Argila	Cálcio
Numero de observações	273	273	273	273
Média	30,4	17,5	52,14	43,1
1o Quartil	26,0	4,0	46,0	31,0
2o Quartil	34,0	9,0	56,0	45,0
3o Quartil	37,0	21,0	62,0	56,0
Mínimo	4,0	2,0	10,0	7,0
Máximo	57,0	86,0	80,0	94,0
Desvio padrao	9,93	20,14	13,65	9,9
Coefficiente de Assimetria	-0,97	1,71	-0,98	-0,98
Coefficiente de Curtose	3,26	5,01	3,57	3,26
Coefficiente de Variação (%)	32,70	115,07	26,17	32,70

A análise descritiva fornece uma idéia geral do comportamento dos dados, mas não considera a posição geográfica dos pontos amostrados.

Sendo assim, uma primeira análise espacial será realizada levando-se em conta a localização dos dados no espaço. Deseja-se que o gráfico das coordenadas apresente padrões de cores e formatos a fim de que eventualmente haja correlação espacial entre os dados.

Tomaremos como base para aplicação dos métodos geoestatísticos e obtenção dos resultados a variável silte(%) na camada 0 – 20 cm do solo, e essa aplicação será análoga a demais variáveis.

A primeira fase de análise exploratória de dados espaciais é simplesmente traçar a resposta dos dados em relação às suas posições. A inspeção cuidadosa da área pode revelar *outliers* espaciais, as respostas que aparecem grosseiramente discordantes dos seus vizinhos espaciais, ou tendências espacial que pode sugerir a necessidade de incluir um modelo de superfície de tendência para uma média variando espacialmente, ou talvez qualitativamente comportamentos diferentes em diferentes sub-regiões.

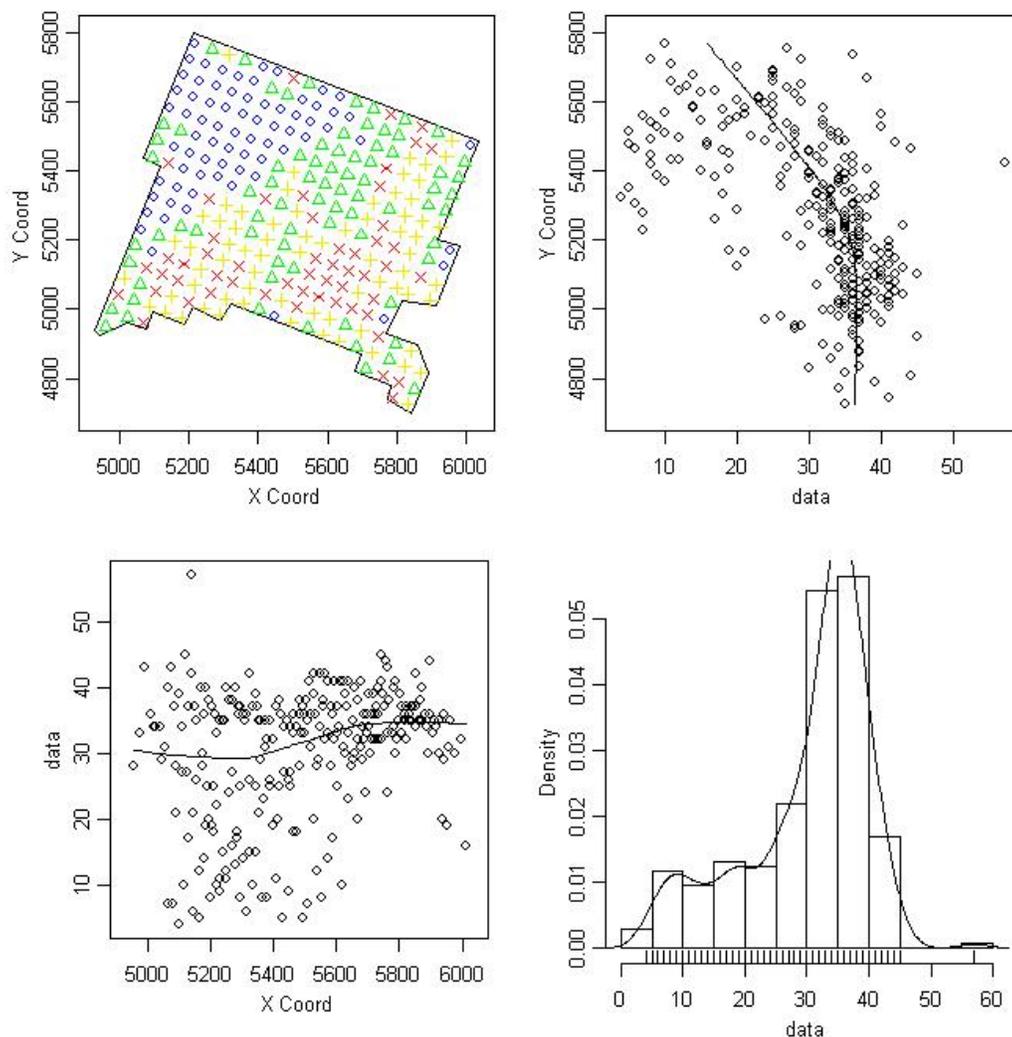


FIGURA 1: Gráfico das localizações dos pontos (superior esquerdo), os valores observados contra coordenadas (canto superior direito e parte inferior esquerda) e histograma (parte inferior direita) do teor de silte (%) medido.

Através do histograma da (Figura 2) e do valor do coeficiente de curtose (Tabela 1), maior que zero há evidências de valores discrepantes, à esquerda, em relação à distribuição dos dados. Para verificar a necessidade de transformação nos dados, a estimativa do parâmetro de transformação Box-Cox indica os possíveis valores do parâmetro lambda a ser utilizado. Os intervalos de confiança de 95% para o parâmetro de transformação λ de normalidade de Box-Cox (BOX; COX, 1964) não incluem o valor unitário. Assim, optou-se por transformar a variável silte, o valor sugerido para lambda é 2.2 (Figura 4).

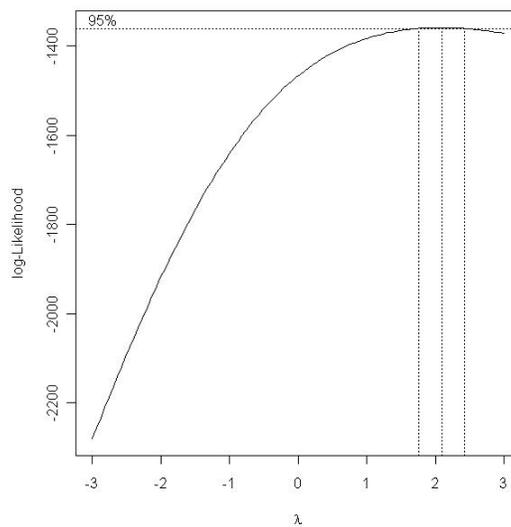


FIGURA 2: Perfil do log-verossimilhança para o parâmetro λ de transformação de Box- Cox.

A característica mais evidente da figura 3 é a preponderância de grandes valores de resposta para o extremo leste da região de estudo. Isto sugere que uma superfície de tendência no modelo pode ser apropriada. Em alguns aplicações, o contexto específico dos dados poderia sugerir que há algo de especial na direção norte-leste, por exemplo. Caso contrário, nossa vista seria que, se uma superfície de tendência é para ser incluída no modelo em tudo, então ambas as coordenadas espaciais devem contribuir para isso porque a orientação da região de estudo é essencialmente arbitrária.

Com a retirada do ponto discrepante e o uso da transformação sugerida podemos obter uma nova representação gráfica para os dados, mostrada na Figura 3.

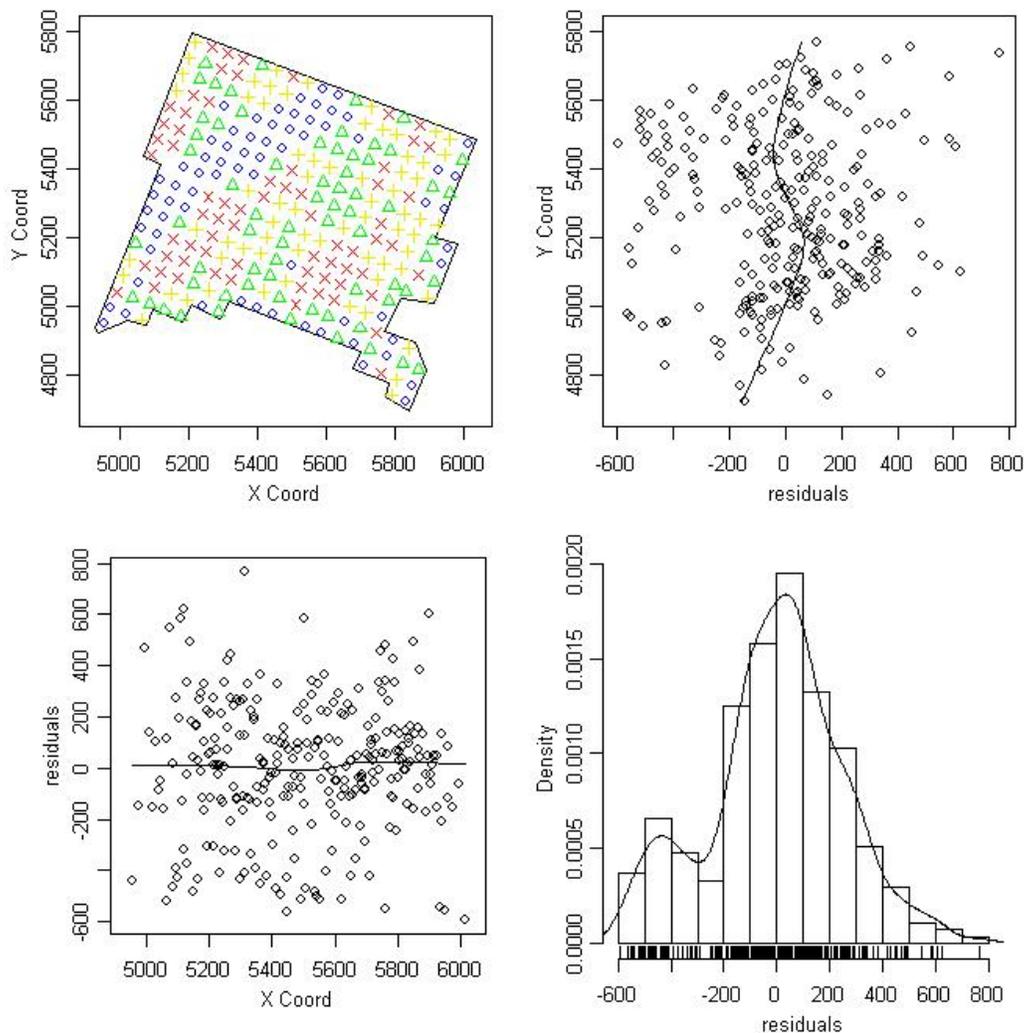


FIGURA 3: Gráfico das localizações dos pontos (superior esquerdo), os valores observados contra coordenadas (canto superior direito e parte inferior esquerda) e histograma (parte inferior direita) do teor de silte (%) medido para os dados transformados.

Uma maneira de diagnosticar a dependência é através dos variogramas empíricos. Analisando a expressão da função semivariância, pode-se imaginar que quanto mais próximos estiverem os pontos amostrados, maior será a semelhança entre eles e, portanto, menor a semivariância. O semivariograma apresentados na Figura 6 indica estacionaridade de segunda ordem para a variável, porque apresenta patamar claro e bem definido.

Se o semivariograma for constante e igual ao patamar para qualquer valor de h , temos o efeito pepita puro e, neste caso, temos a ausência total de dependência espacial, ou seja, a dependência espacial, se existir, será manifestada à distância ou tempo menor do que o menor espaçamento entre

amostras.

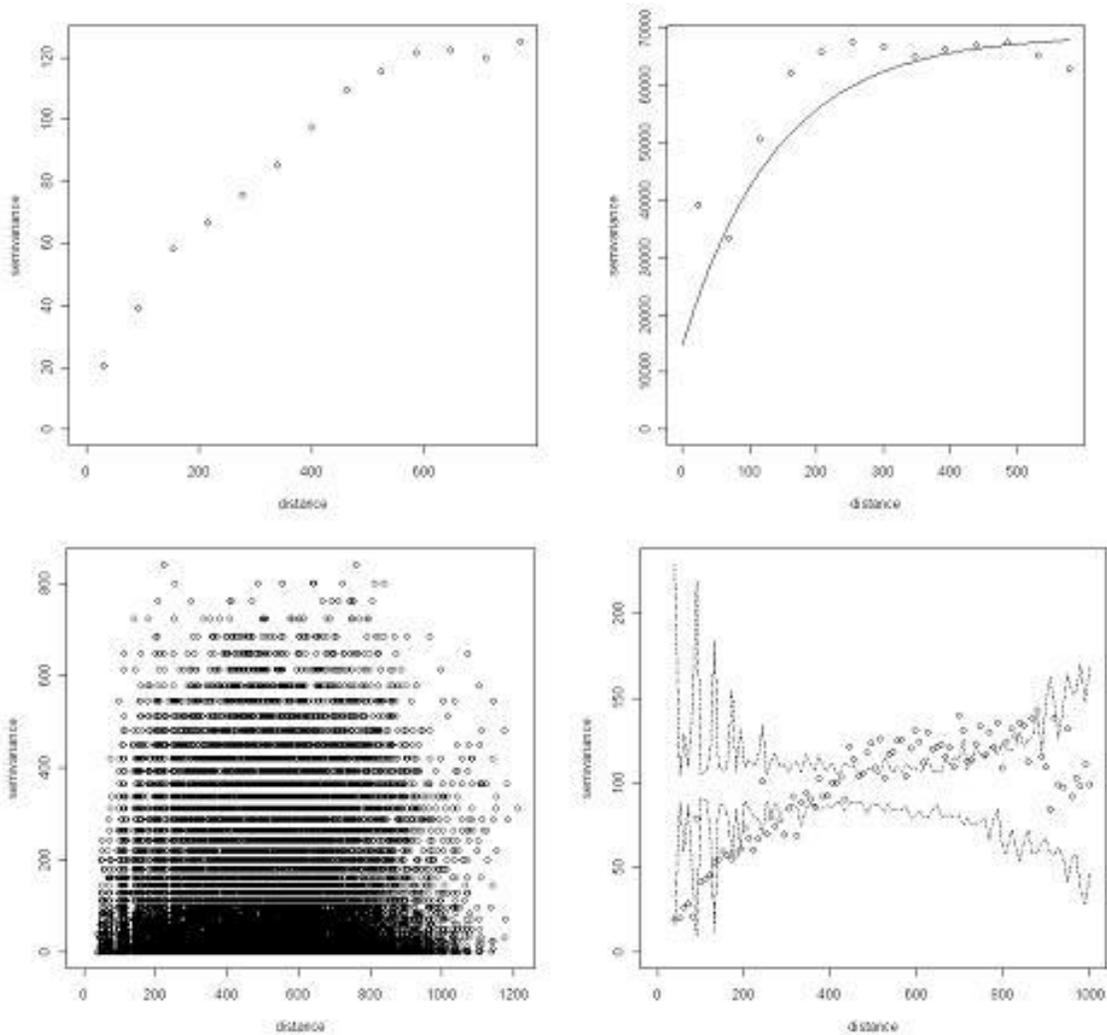


FIGURA 4: Variogramas e envelope simulando indicador de dependência espacial

Também é possível verificar a existência de dependência espacial é através do diagnóstico gráfico utilizando envelopes simulados e variogramas empíricos (Figura 7). Para que haja dependência, deve haver ao menos um ponto do variograma fora do envelope simulado. Deste modo, se isso ocorrer rejeita-se a hipótese nula, de que não há dependência espacial. Isso mostra que a tendência crescente no variograma empírico é estatisticamente significativa, confirmando a presença de correlação espacial positiva.

Devido a dificuldade de se ajustar um modelo teórico confiável ao semivariograma experimental, optou-se nesse trabalho por se obter uma estimativa pontual para os parâmetros do modelo

geoestatístico pelo método da máxima verossimilhança (MV). A Tabela 2 mostra os resultados obtidos por MV.

TABELA 2: Estimação dos parâmetros do modelo geoestatístico por MV, sendo β : parâmetro do efeito sistemático do modelo, σ^2 e ϕ : parâmetros da função de correlação, τ^2 : parâmetro do erro, AIC: Informação de Akaike e logL: valor de MV..

Silte(%)	β	τ^2	σ^2	ϕ	AIC	-logL
kappa 0.5	579,5	4589,5	171398,7	300,0	1671,0	831,4
kappa 1.5	268,1	22652,4	1573306,8	299,9	1691,0	841,6
kappa 2.5	-671,8	26949,1	17858714,6	299,8	1723	859,6

É possível observar as curvas ajustadas da função de correlação de Matèrn com diferentes valores para o parâmetro de diferenciabilidade. De acordo com os resultados da Tabela 2 foi adotado no modelo, a função de correlação de Matèrn com $k = 1,5$ e média constante, pois apresentou menor AIC.

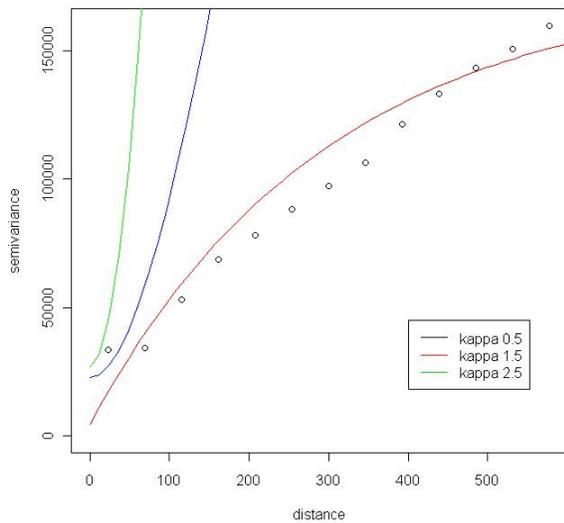


FIGURA 5: Comportamento da função de correlação de Matèrn com diferentes valores para o parâmetro de diferenciabilidade k (esquerda).

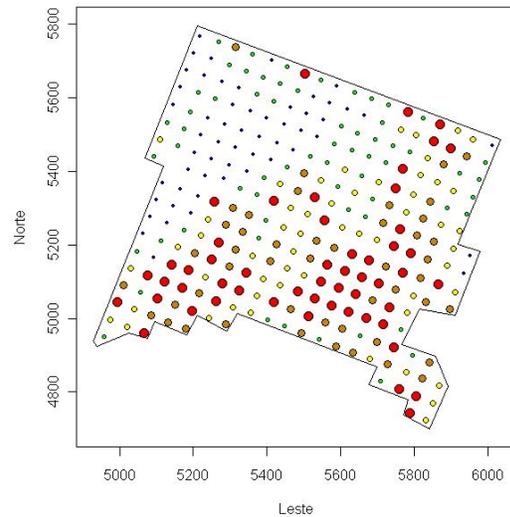


FIGURA 6: Gráfico de padrões de intensidade por parcela colhida classificado pelos quantis (direita).

Observou-se que existe tendência linear mediana da dependência espacial de silte(%) na direção norte na camada 0-20 cm não muito expressiva e negativamente e entre silte(%) e a direção leste

a dispersão dos dados foi ao acaso. Por meio do gráfico das coordenadas é possível verificar que existe um certo padrão de cores indicando que há evidências de correlação espacial.

Para a predição do subjacente, vamos ilustrar aqui o mais simples de todos os métodos geostatísticos: krigagem simples. Em nossos termos, krigagem simples é a previsão de erro mínimo quadrado médio no âmbito do modelo Gaussiano estacionário, mas isto é, ignorando a incerteza de parâmetros, as estimativas de todos os parâmetros do modelo estão ligados em equações de predição, como se fossem os verdadeiros valores do parâmetro.

Após a definição da dependência espacial do teor de silte (%), realizou-se a krigagem ordinária dos dados para obtenção dos mapas temáticos dessas seções transversais, estimando valores em locais não medidos e com menor variância da estimada, o que é uma característica desse método. A krigagem foi estimada partir de modelo ajustado por MV, em 273 pontos de uma malha regular.

Foi obtido no estudo o gráfico das probabilidades do teor de silte serem maior que 30(%), onde observa-se que 80% da área tem teores em torno de 30.

O preditor de krigagem é uma combinação linear de valores amostrados pelo peso atribuído a cada vizinho. Além disso, o estimador linear apresenta as características desejáveis de ser não viesado e de variância mínima. Nota-se porém, que o estimador pressupõe ausência de erros de medida e variogramas corretos. Essas figuras mostram alterações significativas no comportamento da variável, no perfil do solo.

TABELA 3: Estatística descritiva das predições por krigagem convencional do teor de silte(%)do solo medido em uma malha de 273 pontos.

Variável	Mínimo	Média	Máximo
Silte	-0,3514	-0,0009	0,3596

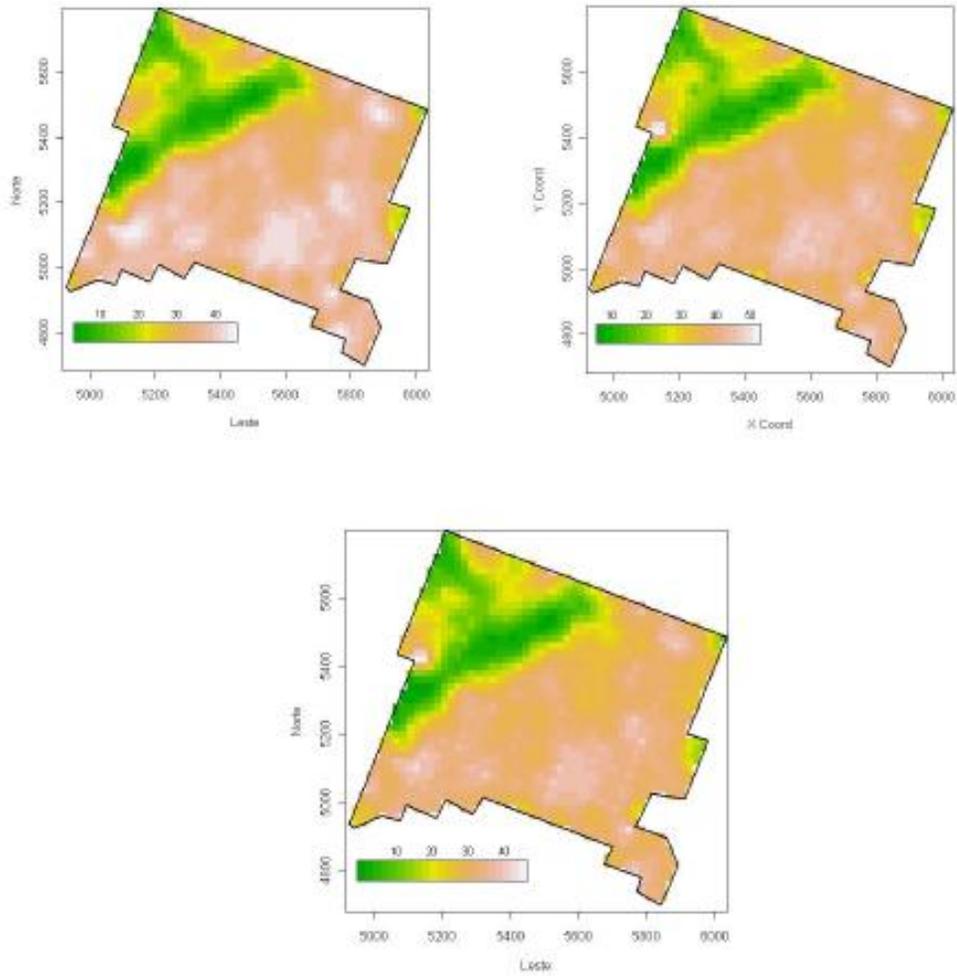


FIGURA 7: Gráfico das localizações dos pontos (superior esquerdo), os valores observados contra coordenadas (canto superior direito e parte inferior esquerda) e histograma (parte inferior direita) do teor de silte (%) medido para os dados transformados.

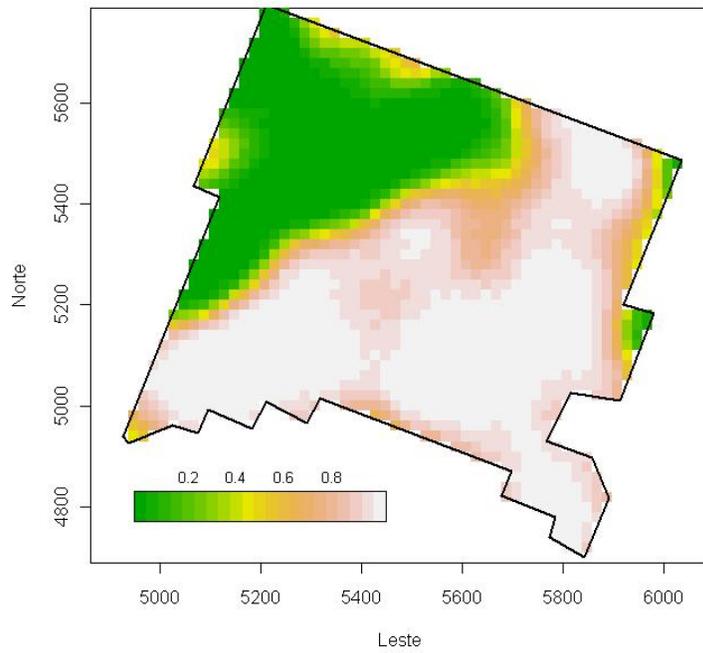


FIGURA 8: Gráfico das probalidades do teor de silte.

5 CONSIDERAÇÕES FINAIS

Considerando-se as condições de realização deste trabalho pode-se concluir que o diagnóstico exploratório é importante para detectar onde estão os padrões espaciais de continuidade e os outliers; o modelo subjetivo ou empírico para estudo geoestatístico permite detectar a dependência espacial; o modelo subjetivo ou empírico permite calcular os valores iniciais dos parâmetros; o estudo georreferenciado é importante nos casos onde o recurso para pesquisa é limitado e dando subsídios para a instalação de futuros experimentos.

REFERÊNCIAS BIBLIOGRÁFICAS

AITCHISON, J. **The statistical analysis of compositional data**. Royal Statistical Society, Series B, v. 44, 1982.

AITCHISON, J. (Ed.). **The statistical analysis of compositional data**. New Jersey: The Blackburn Press, 1986.

BAILEY, T. C., GATRELL, A. C. **Interactive spatial data analysis**. Harlow: Longman, 1995.

BOX, G., COX, D. An analysis of transformation. **Journal of the Royal Statistical Society, Seires B**. p. 211-252, 1964.

DIGGLE, P. J., TAWN, J. A., MOYEED, R. A. Model-based geostatistics. **Applied Statistics**. v. 47, n. 3, p. 299-350, 1998. Disponível em: <citeseer.ist.psu.edu/diggle98modelbased.html>

DIGGLE, P. J., RIBEIRO JR., P. J., CHRISTENSEN, O. F. **An introduction to model-based geostatistics**. In: MØLLER, J. (Ed.). **Spatial Statistics and Computational Methods**. New York: Speinger, 2003. p. 43-86.

DIGGLE, P. J., RIBEIRO JR., P. J. **Model-based geostatistics**. USA: Springer Series in Statistics, 2007.

GUERRA, P. A. G. **Geoestatística operacional**. Brasília: Ministério das Minas e Energia, 1988. 145p.

LANDIM, P. M. B. **Introdução à geoestatística**. Rio Claro: UNESP, 1998, 226p.

MARTINS, A. B. T. **Análise geoestatística para dados composicionais**. Tese (Dourado), Universidade Federal do Paraná, Curitiba, 2010.

OBAGE, S. C. **Uma análise bayesiana para dados composicionais**. 2005. 69p. Dissertação (Mestrado em Estatística) - Universidade Federal de São Carlos, São Carlos.

PAWLOWSKY-GLAHN, V., OLEA, R. A. **Geostatistical analysis of compositional data**. New York: Oxford University Press, Inc., 2004. 181p.

SCHABENBERGER, O., GOTWAY, C. A. **Statistical Methods for Spatial Data Analysis**. Boca Raton: Chapman and Hall, 2005.