

# MODELANDO SATURAÇÃO POR BASES E PH DO SOLO: UMA APLICAÇÃO DE MODELOS GEOESTATÍSTICOS BIVARIADOS <sup>1</sup>

Bruno Henrique Fernandes Fonseca <sup>2</sup>, Paulo Justiniano Ribeiro Jr. <sup>3</sup>, Bruno Gherardi <sup>4</sup> & Luiz Eduardo Oliveira de Faria <sup>5</sup>

## RESUMO

Os modelos geoestatísticos bivariados são uma abordagem que pode ser útil em monitoramentos de solo pois existe correlação estatística entre algumas variáveis químicas do solo que possuem padrão espacial de variabilidade. Esses modelos definem funções aleatórias para dois processos estocásticos com localizações espaciais conhecidas. Pode-se adotar a suposição da existência de um campo aleatório gaussiano latente para cada variável aleatória. A suposição de gaussianidade do processo latente é conveniente para inferências sobre parâmetros do modelo e para obtenção de predições espaciais, uma vez que a distribuição de probabilidade conjunta para um conjunto de pontos do processo latente é também gaussiana. A matriz de covariância dessa distribuição deve ser positiva definida e possuir a estrutura de variabilidade espacial entre e dentre os atributos. Gelfand et al. (2004) e Diggle & Ribeiro Jr. (2007) propuseram estratégias para estruturar essa matriz, porém não existem muitos relatos sobre o uso e avaliações comparativas entre essas abordagens. Neste estudo foram utilizados dados provenientes da análise de solo de uma propriedade agrícola com 51,8ha de área, onde foram amostradas 67 localizações georeferenciadas. Foram utilizados os valores mensurados de pH e da saturação por bases do solo, que foram submetidas à análise descritiva espacial, modelagens geoestatísticas univariadas, bivariadas e predições espaciais. Para verificar vantagens quanto à adoção de modelos univariados ou bivariados, a amostra da saturação por bases, que possui coleta mais dispendiosa, foi dividida em uma subamostra de modelagem e uma subamostra de controle. A primeira foi utilizada para fazer a modelagem geoestatística e a segunda foi utilizada para comparar as precisões das predições espaciais nas localizações omitidas no processo de modelagem.

*Termos de Indexação: agricultura de precisão, estimadores de máxima verossimilhança, modelos geoestatísticos bivariados*

---

<sup>1</sup>Parte da Tese de Mestrado em Agronomia com Área de Concentração em Estatística e Experimentação Agronômica da Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo - ESALQ/USP. Pesquisa financiada pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES

<sup>2</sup>Mestrando em Agronomia pela Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo - ESALQ/USP. Av. Pádua Dias 11, CEP 13418-900 Piracicaba (SP). E-mail: bh.fonseca@hotmail.com

<sup>3</sup>Professor Adjunto da Universidade Federal do Paraná - UFPR. E-mail: paulojus@ufpr.br

<sup>4</sup>Mestrando em Agronomia pela Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo - ESALQ/USP. E-mail: bruno.gherardi@agrofficio.com.br

<sup>5</sup>Bacharel em Agronomia pela Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo - ESALQ/USP. E-mail: luiz.faria@agrofficio.com.br

## ABSTRACT

Bivariate geostatistical models are an approach that may be useful in monitoring of soil because there is statistical correlation between some chemical variables of the soil that have spatial pattern of variability. These models define random functions for two stochastic processes with known spatial locations. Existence of a Gaussian random field can be assumed for each latent random variable. This Gaussianity assumption for the latent process is a convenient one for the inferences on the model parameters and for spatial predictions once the joint distribution for a set of points is multivariate normal. The covariance matrix of this distribution should be positive-definite and to have the spatial variability structure between and among the attributes. Gelfand et al. (2004) and Diggle & Ribeiro Jr. (2007) suggested strategies for structuring this matrix, however there are few reports on comparing approaches. This work reports on a simulation study of bivariate models together with maximum likelihood estimators and spatial prediction under different sets of sampling locations space. Soil sample data from a field with 51.8 hectares is also analyzed with the two soil attributes observed at 67 spatial locations. Data on pH and base saturation were submitted to spatial descriptive analysis, univariate and bivariate modeling and spatial prediction. To check for advantages of the adoption of univariate or bivariate models, the sample of the more expensive variable was divided into a modeling and testing subsamples. The first was used to fit geostatistical models, and the second was used to compare the spatial prediction precisions in the locations not used in the modeling process.

*Index Terms: precision agriculture, maximum likelihood estimators, bivariate geostatistical models*

## INTRODUÇÃO

A modelagem geoestatística é um conjunto de técnicas que tenta encontrar uma boa função aleatória para um ou mais atributos que possuem localizações espaciais conhecidas. Sendo assim, essas ferramentas são úteis para capturar a correlação entre as observações das variáveis aleatórias de interesse, onde existe uma forte suspeita de que pontos espaciais mais próximos possuem valores dos atributos mais parecidos.

Um campo aleatório é um processo estocástico que existe em algum espaço real  $d$ -dimensional, geralmente bi ou tri-dimensional sua definição é dada por  $\{Z(s_i) : s_i \in G \subset R^d\}$ , sendo que  $Z(s_i)$  denota a variável aleatória  $Z$  na localização  $s_i$  do espaço sob estudo  $G$ .

Segundo Schmidt & Sansó (2006) e Le & Zidek (2006), a descrição de um campo aleatório é obtida através das distribuições acumuladas finito-dimensionais  $F$ , para qualquer conjunto de localizações  $(s_1, s_2, \dots, s_n)$  pertencentes à região  $G$  e qualquer inteiro  $n$ :

Devido a sua simplicidade inferencial, a distribuição de probabilidade gaussiana é uma das mais utilizadas na literatura. Sendo assim, um campo aleatório é dito ser gaussiano se  $Z$  segue uma distribuição normal em todas as localizações do espaço sob estudo  $G$ . Para qualquer conjunto finito de localizações

$s = (s_1, s_2, \dots, s_n)$  pertencente a  $G$ ,  $Z(s)$  segue uma distribuição normal  $n$ -variada e é completamente especificado pelo vetor de média  $n \times 1$ , notado por  $\mu$ , e pela matriz de covariâncias  $n \times n$ , notada por  $\Sigma$ , que no contexto de geoestatística, em geral possui o comportamento de que quanto maior a distância euclidiana entre duas localizações  $s_l$  e  $s_k$  quaisquer, menor a correlação entre  $Z(s_l)$  e  $Z(s_k)$ . A especificação de um modelo com tal comportamento deve assegurar que a matriz de covariâncias fique positiva definida. Diggle & Ribeiro Jr. (2007) mostram algumas possíveis construções para obtenção de campos aleatórios gaussianos multivariados. Em geral não é possível ter mais de uma realização do processo, sendo assim, outras suposições devem ser impostas sobre o campo aleatório gaussiano para a realização de inferências. A restrição mais utilizada é a de que o processo estocástico é homogêneo. Um campo aleatório gaussiano é dito ser homogêneo se ele for estacionário e seu padrão de correlações não depende das direções, ou seja, isotrópico. Utilizando essa suposição, o processo estocástico, embora restritivo, consegue modelar diversos problemas naturais.

Com a suposição de homogeneidade de um campo aleatório gaussiano, é necessário estabelecer uma função matemática que dependa apenas das distâncias entre as localizações amostradas do espaço sob estudo e que estruture a matriz de covariâncias, de forma que esta seja positiva definida e com o comportamento tipicamente utilizado em geoestatística. Encontrar tais funções que sejam válidas pode ser um procedimento complexo. Na prática, o usual é adotar famílias de funções ditas permissíveis. Proposta por Matérn (1986), a família de funções de correlação de Matérn é muito utilizada devido a capacidade de produzir comportamentos distintos quanto a suavidade do processo. Além disso, sob estacionariedade fraca, essas funções possuem propriedade conhecidas e desejáveis. Schabenberger & Gotway (2005) apresentam e discutem detalhes sobre tais propriedades. Silva (2006), em um trabalho discutindo a construção de funções de correlação válidas para modelos geoestatísticos espaço temporais apresenta diversas outras funções de correlação conhecidamente válidas e bem como uma extensa lista de referências sobre o tema.

Considerando que em alguma área  $G$  exista um campo aleatório gaussiano  $Z$  latente, ou seja, o processo existe mas não é observável, é necessário fazer uma amostragem de  $n$  localizações espaciais dentro da área  $G$  e observar valores de atributo de interesse nas localizações amostradas. Sendo assim, existe um vetor  $Y(s)$  de  $n \times 1$  valores observados em  $s = (s_1, s_2, \dots, s_n)$ , que segundo Diggle & Ribeiro Jr. (2007), sob o pressuposto de gaussianidade, pode ser modelado como:

$$Y(s) = \mu + Z(s) + \epsilon, \quad (1)$$

sendo  $\mu = X\beta$ , sendo  $X$  uma matriz  $n \times q$  contendo  $q - 1$  possíveis covariáveis,  $\beta$  um vetor  $q \times 1$  de parâmetros associados a  $X$ ,  $Z(s)$  um campo aleatório gaussiano que possui vetor de médias  $n \times 1$  nulo e matriz de covariâncias  $\Sigma$  de dimensão  $n \times n$ , sendo que cada elemento  $\Sigma_{i,j}$  é igual a  $Cov(Z(s_i); Z(s_j))$ ,

para todo  $s_i$  e  $s_j$  pertencentes a  $s$ , e  $\epsilon$  um vetor  $n \times 1$  de ruídos brancos, que por suposição são independentes e identicamente distribuídos com distribuição de probabilidade normal com média zero e desvio padrão  $\tau$ . Faraco et al. (2008) mostra a aplicabilidade desse modelo.

Utilizando (1) é possível encontrar a distribuição de probabilidade de  $Y(s)$ , que é gaussiana  $n$ -variada, com vetor de médias  $X\beta$  e matriz de covariâncias  $\Sigma_Y = \Sigma + \tau^2 I$ , sendo  $I$  uma matriz identidade  $n \times n$ . Logo, existe um vetor de parâmetros  $\theta = (\beta, \sigma^2, \phi^*, \tau^2)$  a ser estimado, onde  $\phi^*$  é um vetor de parâmetros associados a função de correlação utilizada. Cabe observar que  $Cov(Z(s_i); Z(s_j)) = \sigma^2 \rho(Z(s_i); Z(s_j))$ , para todo  $s_i$  e  $s_j$  pertencentes a  $s$  e  $\sigma^2 > 0$ .

Especificado um modelo adequado para os dados em questão, o próximo passo é fazer a estimação dos parâmetros. Assumindo que o campo aleatório possui estacionariedade forte, pode-se optar por estimadores baseados na verossimilhança ou verossimilhança restrita. Esta abordagem consiste em utilizar os valores observados da variável resposta para encontrar um vetor  $\hat{\theta}$  que seja o ponto de máximo da função de verossimilhança associada a  $\theta$ . Por simplicidade matemática, normalmente utiliza-se o logaritmo da função de verossimilhança para fazer a estimação, que associada a (1) tem a seguinte forma:

$$l(\theta; Y(s)) = -0.5(n \ln(2\pi) + \ln(|\Sigma_Y|)) + (Y(s) - X\beta)^t \Sigma_Y^{-1} (Y(s) - X\beta). \quad (2)$$

No contexto de geoestatística, Diggle & Ribeiro Jr. (2007) adotam a utilização da reparametrização  $\nu = \tau/\sigma$ , a qual facilita a estimação de  $\theta$ . O vetor de parâmetros ser estimado passar a ser  $\theta^* = (\beta, \sigma^2, \phi^*, \nu^2)$  e  $\Sigma_Y$  pode ser escrita como  $\sigma^2 V$ , onde  $V$  é uma matriz  $n \times n$  que depende apenas de  $\nu$  e  $\phi^*$ . Existem formas analíticas apenas para os estimadores de  $\beta$  e  $\sigma^2$ , que são funções dos demais parâmetros e, além das formas fechadas para os estimadores, é possível encontrar a matriz de informação de Fisher observada para os mesmos e portanto a matriz de covariâncias associada a  $(\hat{\beta}, \hat{\sigma}^2)$  (DIGGLE & RIBEIRO Jr., 2007). Para  $\phi^*$  e  $\nu^2$  não existem formas analíticas para os estimadores. Utilizando  $\hat{\beta}$  e  $\hat{\sigma}^2$  em (2), tem-se o logaritmo da função de verossimilhança concentrada, que depende apenas de  $\theta_c = (\nu, \phi^*)$ .

Para encontrar  $\hat{\theta}_c$  é possível utilizar métodos numéricos de maximização de funções, como por exemplo, o método de Nelder & Mead (1965), o qual calcula numericamente  $\hat{\phi}^*$  e  $\hat{\nu}^2$  e a matriz Hessiana estimada, denotada por  $H$ . Com os parâmetros da função de máxima verossimilhança concentrada estimados, por invariância é possível encontrar as estimativas de  $\beta$  e  $\sigma^2$  e  $\tau^2$ .

Geralmente o objetivo final dos estudos com dados georeferenciados é calcular as predições espaciais em localizações não amostradas. Esse processo de predição é denominado de krigagem e é baseado nas propriedades da distribuição normal multivariada. Do ponto de vista geoestatístico, as krigagens mais frequentemente utilizadas são a simples e a ordinária, que se diferenciam quanto a suposição de

conhecimento prévio sobre os parâmetros da média do processo (DIGGLE & RIBEIRO Jr., 2007).

Quando existem dois campos aleatórios gaussianos de interesse, a intuição inicial é que seja ajustado um modelo geoestatístico univariado para cada vetor de valores observados. Porém, em alguns casos existe correlação estatística entre as variáveis aleatórias, o que leva a possibilidade de adotar modelos geoestatísticos bivariados.

Supondo a existência de dois campos aleatórios gaussianos de interesse, pode-se utilizar seguinte o modelo:

$$Y_i = \mu_i + Z_i, \quad i = 1, 2 \quad (3)$$

sendo  $Y_i$  um vetor  $n_i \times 1$  de valores observados do campo aleatório gaussiano latente  $Z_i$ , que possui vetor de médias nulo  $n_i \times 1$  e matriz de covariâncias  $\Sigma_i$   $n_i \times n_i$ ,  $\mu_i$  é um vetor, que possui os parâmetros de médias associados a  $Y_i$ . Observe que em (3) não existem termos de ruídos brancos.

A distribuição de  $Y = (Y_1, Y_2)$  é de interesse final e possui distribuição gaussiana  $n$ -variada, sendo  $n = n_1 + n_2$ , com vetor de médias  $\mu = (\mu_1, \mu_2)$  e matriz de covariâncias  $\Sigma_Y$ , positiva definida que possui o comportamento empírico de correlações utilizado em geoestatística e que pode ser particionada da forma:

$$\Sigma_Y = \begin{bmatrix} \Sigma_1 & \Sigma_{1,2} \\ \Sigma_{1,2}^t & \Sigma_2 \end{bmatrix}$$

onde  $\Sigma_i$  é uma matriz  $n_i \times n_i$  das covariâncias da variável  $Y_i$ ,  $i = 1, 2$ , e  $\Sigma_{1,2}$  uma matriz  $n_1 \times n_2$  com as covariâncias cruzadas entre  $Y_1$  e  $Y_2$ .

Nesse contexto a especificação do modelo exige uma matriz  $\Sigma_Y$  válida. A abordagem por modelos separáveis é a mais simples para resolver esse problema e utiliza decomposições os termos latentes de (3).

Diggle & Ribeiro Jr. (2007) propuseram uma abordagem para problemas geoestatísticos bivariados, a qual é abreviada nesse trabalho por BGCCM (*bivariate Gaussian common component model*) e que utiliza as seguintes decomposições dos campos aleatórios latentes de (3):

$$Z_i = \sigma_{0i}S_0 + \sigma_i S_i, \quad i = 1, 2, \quad (4)$$

sendo  $\sigma^* = (\sigma_{01}, \sigma_1, \sigma_{02}, \sigma_2)$  um vetor de parâmetros de dispersão associados a (3) e  $S_0, S_1$  e  $S_2$  são campos aleatórios gaussianos mutuamente independentes, com vetores de médias nulos, variâncias unitárias e correlações determinadas por funções de correlação válidas. As correlações destes termos, quando combinadas desta forma, vão gerar correlações cruzadas válidas entre as variáveis observáveis  $Y_1$  e  $Y_2$ , uma vez que, o campo aleatório  $S_0$  é comum as duas respostas. Portanto, sob este modelo, três funções

de correlação válidas para  $S_0$ ,  $S_1$  e  $S_2$  devem ser adotadas para gerar uma matriz  $\Sigma_Y$  válida.

Definindo  $Y_i(s_l)$  e  $Y_i(s_k)$  como observações de  $Y_i$  feitas em duas localizações quaisquer  $s_l$  e  $s_k$ , que estão separados por uma distância euclidiana  $h = h_{lk}$ ,  $l, k = 1, 2, \dots, n_i$  e  $i = 1, 2$ , tem-se que o elemento  $\Sigma_{i,(l,k)}$  é dado por  $Cov(h) = \sigma_{0i}^2 \rho_0(h) + \sigma_i^2 \rho_i(h)$ , sendo  $\rho_0$  e  $\rho_i$  as funções de correlação adotadas para  $S_i$ ,  $i = 0, 1, 2$ , respectivamente. Utilizando propriedades básicas de covariâncias, pode-se encontrar  $\Sigma_{1,2}$ , que é igual a  $\sigma_{01}\sigma_{02}R_0$ , sendo  $R_0$  uma matriz  $n_1 \times n_2$  das correlações cruzadas entre as respostas e depende da função de correlação adotada para  $S_0$ . Dessa forma  $\Sigma_Y$  fica completamente estruturada.

Gelfand et al. (2004) propôs uma abordagem para encontrar uma estrutura paramétrica válida para  $\Sigma_Y$ , que é abreviada aqui por BCRM (*bivariate coregionalisation model*). Nesta abordagem os termos latentes de (3) são decompostos da seguinte forma:

$$\begin{cases} Z_1 = \sigma_{11}S_1 \\ Z_2 = \sigma_{12}S_1 + \sigma_{22}S_2 \end{cases}$$

sendo  $S_1$  e  $S_2$  campos aleatórios gaussianos mutuamente independentes, com vetores de médias nulo, variância unitária e com correlações determinadas pela escolha de funções de correlação conhecidamente válidas. Como o campo aleatório  $S_1$  é comum as duas variáveis, é induzida a correlação empírica entre as respostas. Gelfand et al. (2004) mostram mais detalhes sobre essa abordagem do ponto de vista de modelos geoestatísticos.

Utilizando o BCRM, o desenvolvimento probabilístico para encontrar a estrutura de  $\Sigma_Y$  é similar ao exposto para o BGCCM. O vetor de parâmetros associados a distribuição de probabilidade de  $Y$  é  $\theta = (\beta^*, \sigma_{11}, \sigma_{12}, \sigma_{22}, \phi_1^*, \phi_2^*)$ , sendo  $\beta^*$  um vetor de parâmetros associado a  $\mu$  e  $\phi_j^*$  um vetor de parâmetros associado a escolha da função de correlação  $\rho_j$ , para  $j = 1, 2$ .

Com  $\Sigma_Y$  especificada sob o BGCCM ou o BCRM, o próximo passo é estimar o vetor de parâmetros  $\theta$  e realizar a krigagem. As técnicas são similares às adotadas no modelo univariado, no entanto, nesse caso é necessário um número maior de reparametrizações para fazer as estimações.

A agricultura de precisão é, de forma sucinta, um conjunto de métodos aplicados ao manejo da variabilidade, sendo um campo de pesquisa que pode utilizar os modelos geoestatísticos para melhor extrair informação de dados. SCHUELLER (1992) definiu como um método de administração cuidadosa e detalhada do solo e da cultura para adequar as diferentes condições encontradas em cada pedaço de lavoura, tendo em vista a desuniformidade intrínseca dos solos.

Alguns autores concluíram que a variabilidade espacial existe, mesmo em áreas consideradas homogêneas. Trabalhos de campo têm mostrado a importância do estudo das variações das condições do solo como aspecto fundamental para se implementar uma agricultura mais eficiente e rentável. Estes

trabalhos mostram que a variabilidade do solo não é puramente aleatória, apresentando correlação ou dependência espacial. (OLIVEIRA et al., 1999; CARVALHO et al., 2003; SILVA et al., 2003). Uma possível abordagem para esse tipo de problema, que proporciona resultados com respaldo probabilístico, é a modelagem geoestatística conjunta de atributos de solo e em particular, em conjunto com métodos de análise baseados em verossimilhança.

Esse trabalho utiliza ferramentas geoestatísticas para estudar a variabilidade espacial de duas variáveis químicas do solo de uma propriedade agrícola. A utilização de modelos geoestatísticos bivariados é possível devido a natureza dos atributos, que são fortemente correlacionados e, além disso, uma das respostas é considerada ser mais dispendiosa para ser observada. Os resultados fornecem mapas de interesse prático descrevendo o comportamento dos atributos na região, com mensurações de incertezas associadas. Além disto, com a estrutura conjunta de correlação espacial estabelecida, em monitoramentos futuros da mesma propriedade agrícola, será possível avaliar possíveis diminuições nos gastos com a coleta de informações.

## MATERIAL E MÉTODOS

Os dados analisados com as técnicas apresentadas foram obtidos por meio de uma pesquisa realizada em julho 2006 na fazenda Tupã, localizada no município de Echaporã/SP, que possui 51,8ha de área, solo argissolo de textura média e dois históricos de manejo distintos, soja numa região e pastagem na outra. Para analisar as variáveis químicas do solo, foram amostradas 67 localizações, com grid regular a cada hectare, georeferenciadas no sistema Universal Transverse Mercator (UTM). Com a utilização de um aparelho GPS foram medidos os atributos de interesse.

Dois parâmetros químicos do solo levantados pelos pesquisadores foram modelados nesse trabalho: saturação por bases e pH do solo, sendo que, existe uma forte correlação entre essas duas respostas, o que justifica a tentativa de modelagem bivariada.

Do ponto de vista geoestatístico, o pH e a saturação por bases são dois campos aleatórios gaussianos e latentes. A intuição inicial é que seja ajustado individualmente o modelo (1) para cada vetor de valores observados, que foram notados por  $Y_1$  e  $Y_2$  para a saturação por base e o pH, respectivamente. Então  $Y_i$  segue uma distribuição gaussiana 67-variada, com matriz de covariâncias  $\Sigma_i$  e vetor de médias  $\mu_i = X_i\beta_i^*$ , onde  $X_i$  é uma matriz  $67 \times p$  com o intercepto e  $p - 1$  possíveis covariáveis e  $\beta_i^*$  um vetor  $p \times 1$  com os parâmetros associados a  $\mu_i$ .

Foram consideradas três formas diferentes para  $X_i$ , média constante, média com tendência induzida pela área de manejo e média com tendência induzida pela coordenada oeste-leste das localizações amostradas. Foi utilizada a família de funções de correlação de Matérn para estruturar a matriz  $67 \times 67$  de correlações, considerando o parâmetro de suavidade  $\kappa$  fixo nos valores 0,5, 1, 1.5, 2, 2.5, logo,  $\phi_i^* = \phi_i$ . Combinando as diferentes escolhas de  $\kappa$  e  $X_i$  foram ajustados diversos modelos. As estima-

tivas paramétricas foram calculadas por máxima verossimilhança, onde foi utilizado a reparametrização  $\nu_i = \tau_i/\sigma_i$  e o método numérico de Nelder & Mead (1965) para encontrar as estimativas  $\hat{\phi}_i$  e  $\hat{\nu}_i$  que maximizam a função de máxima verossimilhança concentrada a partir das quais são encontradas as demais estimativas de parâmetros de interesse.

Os valores dos máximos estimados da função de máxima verossimilhança concentrada e o Critério de Informação de Akaike (AIC) foram utilizados para selecionar o modelo final de cada atributo. Com os parâmetros estimados dos modelos finais foram conduzidas as krigagens ordinárias para as variáveis químicas do solo. Por último, uma análise de resíduos foi realizada para verificação de pressupostos dos ruídos brancos, os quais se comportaram conforme o esperado.

Após estudo individual de cada variável química, foram ajustados modelos bivariados para o vetor  $Y = (Y_1, Y_2)$ , utilizando as proposições de Diggle & Ribeiro Jr. (2007) e de Gelfand et al. (2004). Nas duas abordagens foi utilizado  $\mu = X\beta^*$ , onde  $X$  é uma matriz  $134 \times p$  contendo um intercepto para cada resposta e  $p - 2$  possíveis covariáveis e  $\beta^*$  é um vetor  $p \times 1$  com os parâmetros de média associados a  $Y$ . Para estruturar  $\Sigma_Y$  somente funções de correlação da família de Matérn foram utilizadas, sendo que o parâmetro  $\kappa$  foi sempre considerado fixo em valores similares aos modelos univariados.

Diversos modelos foram ajustados para combinações de diferentes escolhas de  $\kappa$  e  $X$ . Em todos os casos o vetor de parâmetros  $\theta$  foi estimado por máxima verossimilhança. No BGCCM utilizou-se as reparametrizações  $\sigma = \sigma_{01}$ ,  $\eta = \sigma_{02}/\sigma_{01}$ ,  $\nu_1 = \sigma_1/\sigma_{01}$  e  $\nu_2 = \sigma_2/\sigma_{01}$  e no BCRM das reparametrizações  $\sigma = \sigma_1$ ,  $\nu_1 = \sigma_{12}/\sigma_{11}$  e  $\nu_2 = \sigma_{22}/\sigma_{11}$ . Nas duas abordagens existe forma analítica para os estimadores de  $\beta^*$  e  $\sigma^2$ , que são função dos demais parâmetros, os quais foram estimados pelo método de Nelder & Mead (1965). Utilizando a propriedade de invariância dos estimadores foi possível encontrar estimativas para todos os parâmetros dos modelos. Quanto a seleção de modelos e a krigagem, as mesmas ferramentas dos modelos univariados foram utilizadas.

Para exemplificar o possível ganho com adoção de modelos bivariados, a amostra da saturação por bases foi dividida em uma subamostra de modelagem e uma subamostra de controle, que contêm 47 e 20 observações, respectivamente. O grupo de modelagem foi utilizado para modelar novamente a saturação por bases sob as abordagens uni e bivariadas. Cabe ressaltar que nos modelos bivariados foi utilizada a informação completa do pH. Com os parâmetros estimados foram calculadas as previsões da saturação por bases nas localizações do grupo de controle. Para comparar as metodologias foram calculados os erros de krigagem produzidos em cada caso.

Todas as análises e resultados foram obtidos através do ambiente R de programação (R Development Core Team, 2006), utilizando o pacote *geoR* (Ribeiro Jr. & Diggle, 2001).

## RESULTADOS E DISCUSSÃO

A Figura 1 mostra os gráficos de círculos das variáveis sob estudo, onde os círculos representam os valores mensurados dos atributos nas 67 localizações amostradas da propriedade agrícola. Quanto maior o círculo mais elevado é o valor observado do atributo. Essa primeira análise gráfica evidencia que existe um possível padrão espacial nas variáveis, onde há uma suavidade dos processos ao longo da fazenda. A linha no meio dos gráficos separa as áreas com históricos de manejo distintos: cultivo de soja e pastagem, para as regiões a direita e a esquerda, respectivamente. A primeira área possui valores mais elevados, o que leva a possibilidade de ser considerada essa informação como covariável no processo de modelagem. Além disso, os gráficos são bem parecidos para as duas variáveis, o que corrobora a idéia de que os dados são fortemente correlacionados.

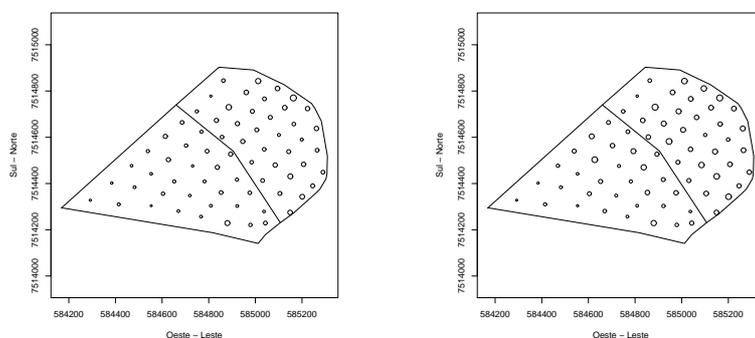


FIGURA 1: Gráficos de círculos - o gráfico a esquerda é para a variável pH e o da direita é da saturação por bases

Estatísticas descritivas do pH e da saturação por bases foram calculadas. Nos dois casos a mediana é próxima da média e existe uma variabilidade relativamente pequena dos dados. De uma maneira geral os atributos de solo não mostram nítidos desvios de Gaussianidade havendo uma simetria dos dados com relação a média amostral e não existem valores observados muito discrepantes. Não existe muita variabilidade nas duas variáveis químicas, o que pode ser atribuído ao tamanho da área da fazenda Tupã, que é relativamente pequena e homogênea. Mesmo assim parece existir padrão espacial de variabilidade nos atributos de solo.

Considerando as diferentes abordagens para a matriz de delineamento  $X$  e as diferentes escolhas para  $\kappa$ , inicialmente foram ajustados diversos modelos geoestatísticos univariados para o pH. Foram utilizados os valores estimados dos máximos do logaritmo da função de verossimilhança para selecionar o melhor modelo dentro de cada metodologia para a tendência na média. A Tabela 1 mostra os resultados do melhor modelo univariado ajustado para o pH em cada uma das formas da matriz  $X$ . Os parâmetros de média e de variabilidade não são muito afetados com as diferentes modelagens, já o parâmetro  $\phi$

apresentou resultados com maior diferença entre as abordagens, fato explicado, em parte, pela não ortogonalidade entre os parâmetros da função de correlação de Matérn. Para selecionar o modelo final foram utilizados os valores estimados do AIC, o modelo que considera tendência induzida pela coordenada oeste-leste foi o escolhido, pois foi o menos penalizado. O modelo final gera um alcance prático estimado de 132,93, ou seja, a partir dessa distância qualquer par de localizações possui correlação inferior a 0,05.

Tendência	$\beta_0$	$\beta_1$	$\tau^2$	$\sigma^2$	$\phi$	$\kappa$	$l(\theta)$	AIC
Constante	4,903	—	0,106	0,297	491,1	2,5	-25,36	58,73
Área de Manejo	4,723	0,403	0,014	0,111	28,43	1,5	-23,33	56,66
Oeste-Leste	-608,5	0,001	0,015	0,097	22,46	2,5	-19,35	48,70

TABELA 1: Estimativas de máxima verossimilhança dos parâmetros associados à modelagem do pH

Além de selecionar o modelo final dentre os que consideram o padrão espacial, foi conduzida uma regressão linear simples considerando a coordenada oeste-leste como covariável, ou seja, considerando que o padrão dos dados é completamente aleatório. Sobre esse enfoque o AIC ficou igual a 49,8, o que afirma a idéia de que o pH possui variabilidade espacial significativa.

Para modelar a saturação por bases foi conduzida a mesma metodologia aplicada ao pH. O modelo final também inclui a coordenada leste-oeste como covariável e possui  $\kappa = 2.5$ . As estimativas foram  $\beta_0 = -14298,98$ ,  $\beta_1 = 0,025$ ,  $\tau^2 = 38,06$ ,  $\sigma^2 = 31,03$  e  $\phi = 32,19$ . Para essa variável o alcance prático estimado ficou em 190,52. Além disso, o valor do AIC com o modelo linear simples para a saturação foi maior do que o modelo geoestatístico.

Com os modelos finais estabelecidos, foram calculadas as previsões espaciais para as duas variáveis químicas do solo em uma malha de 10000 localizações espaciais, sendo que, as estimativas paramétricas foram substituídas nas fórmulas de krigagem e variância preditiva. A Figura 2 mostra os mapas preditivos dos campos aleatórios. As imagens ficaram condizentes com os dados observados, ou seja, localizações com menores valores para a coordenada oeste-leste possuem, em sua maioria, valores preditos menores. Além disso, as variâncias preditivas da saturação por bases estão todas no intervalo  $[12, 35; 30, 21]$ , já para o pH os valores máximo e mínimo das variâncias de krigagens ficaram iguais a 0,10 e 0,01, respectivamente. Esses valores das variâncias preditivas são relativamente pequenos com relação a grandeza de escala das variáveis, sendo assim, a precisão das krigagens ficaram em um patamar aceitável.

Para validar os pressupostos feitos sob os ruídos brancos dos modelos univariados, foram utilizados testes de normalidade de Shapiro-Wilk, que geraram p-valores de 0,22 e 0,97 para os erros do pH e da saturação por bases, respectivamente. Logo, esse pressuposto está garantido.

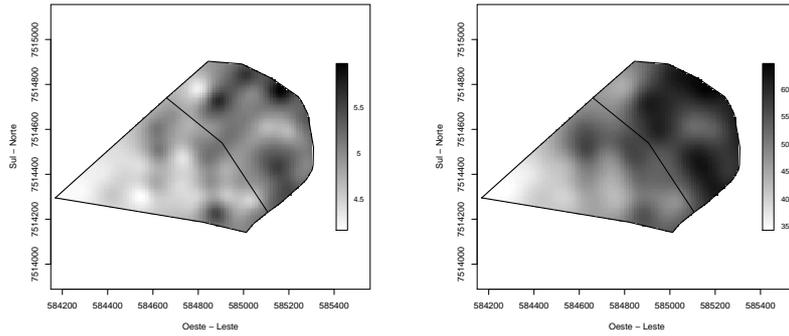


FIGURA 2: Gráficos de predições espaciais, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

Concluída a modelagem univariada, o próximo passo foi ajustar modelos bivariados para as respostas. A correlação de Pearson entre os atributos é 0,92, o que indica uma correlação positiva e quase perfeita entre as duas variáveis químicas em questão. Para ajustar o BGCCM e o BCRM, foram utilizadas as estimativas dos modelos univariados para atribuir valores iniciais ao método numérico de maximização. Três abordagens para modelar as médias dos campos aleatórios foram consideradas: médias constantes, médias com tendências induzidas pelas coordenadas oeste-leste e médias com tendências induzidas pela área de manejo do solo. Tais possibilidades combinadas com diversas escolhas para os parâmetros de suavidade das correlações geraram diversas estimativas diferentes para os parâmetros dos modelos. Inicialmente, foram selecionadas as melhores estimativas em cada abordagem para a média. A Tabela 2 mostra os resultados do BGCCM, onde não há muita diferença entre os parâmetros de variabilidade e correlação estimados para cada abordagem de tendência, exceto para o parâmetro  $\phi_0$ , que para a tendência na área de manejo ficou bem menor do que as demais. Esse fato pode ser explicado pela adoção de um  $\kappa_0$  maior, logo, combinando essas duas informações, tem-se que os alcances de correlações são bem parecidos para as três abordagens de média.

A Tabela 3 mostra os resultados da estimação por máxima verossimilhança utilizando o BCRM. Com esse modelo, a inclusão da covariável coordenada leste-oeste gerou inflacionamento dos parâmetros de variabilidade. Com relação às demais estimativas, os resultados são parecidos com os do BGCCM.

As duas abordagens de modelos bivariados se mostraram estáveis para fazer estimação dos parâmetros, uma vez que, as estimativas dos parâmetros de variabilidade e de correlação apresentaram resultados que geram características similares às encontradas nos modelos univariados. O melhor ajuste foi o que considera a área de manejo da propriedade como covariável, os quais geraram AIC iguais a 437,35 e 432,96 para o BGCCM e BCRM, respectivamente. O BCRM foi menos penalizado pelo AIC, esse modelo se comportou melhor para estimação paramétrica.

$\theta$	Constantes	Oeste-leste	Área de manejo
$\beta_{01}$	52,57	$1,54e - 6$	47,83
$\beta_{02}$	4,92	$1,45e - 7$	4,72
$\beta_1$	—	$8,98e - 5$	9,46
$\beta_2$	—	$8,41e - 6$	0,40
$\sigma_{01}$	8,51	8,35	6,85
$\sigma_1$	3,47	3,46	3,44
$\sigma_{02}$	0,38	0,38	0,31
$\sigma_2$	$7e - 5$	$5e - 4$	$1e - 5$
$\phi_0$	59,37	57,79	26,73
$\phi_1$	46,35	48,15	49,08
$\phi_2$	66,82	71,59	86,95
$\kappa_0$	1	1	1,5
$\kappa_1$	0,5	0,5	0,5
$\kappa_2$	0,5	0,5	0,5

TABELA 2: Estimativas de máxima verossimilhança dos parâmetros associados ao BGCCM para a saturação por bases e o pH

$\theta$	Constantes	Oeste-leste	Área de Manejo
$\beta_{01}$	51,76	$1,51e - 6$	47,93
$\beta_{02}$	4,89	$1,43e - 7$	4,72
$\beta_1$	—	$8,85e - 5$	9,12
$\beta_2$	—	$8,36e - 6$	0,40
$\sigma_{11}$	8,26	99,53	7,82
$\sigma_{12}$	0,31	3,67	0,28
$\sigma_{22}$	0,13	1,54	0,14
$\phi_1$	100,63	100,21	53,32
$\phi_2$	21,43	21,38	19,34
$\kappa_1$	0,5	0,5	0,5
$\kappa_2$	2,5	2,5	2,5

TABELA 3: Estimativas de máxima verossimilhança dos parâmetros associados ao BCRM para a saturação por bases e o pH

O processo de krigagem utilizando as estimativas do BGCCM gerou variâncias preditivas relativamente pequenas, para a saturação por bases os valores mínimos e máximos são 0,47 e 58,40, respectivamente. Já para o pH esses valores ficaram iguais a 0,0001 e 0,096. Utilizando o BCRM as variâncias preditivas da saturação por bases e do pH ficaram dentro dos intervalos (2;60,35) e (0,002;0,098), respectivamente. As Figuras 3 e 4 ilustram os mapas preditivos utilizando os valores estimados em cada abordagem de modelos bivariados. Não existe muita diferença entre as predições espaciais utilizando modelos univariados ou bivariados, mas o BGCCM gerou menores variâncias de krigagem.

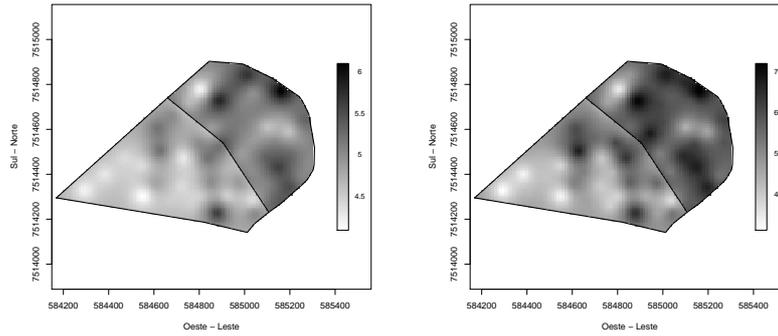


FIGURA 3: Gráficos de predições espaciais com o BGCCM, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

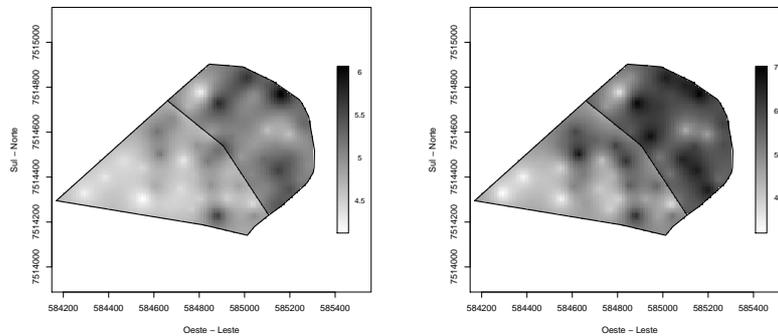


FIGURA 4: Gráficos de predições espaciais com o BCRM, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

Além da fundamentação estatística para considerar modelos bivariados para o problema dos dados do solo, existe a justificativa prática para adoção de tal abordagem. A coleta de informação da saturação por bases é mais dispendiosa e como a propriedade agrícola continuará sendo monitorada futuramente, é possível que nas próximas avaliações do solo seja amostrado um número menor de localizações espaciais para observação da saturação por bases, pois se pode utilizar a informação contida no pH para fazer inferências sobre a saturação por bases. Ou seja, é possível diminuir gastos em monitoramentos de solos se forem utilizados os modelos geoestatísticos bivariados para problemas com variáveis químicas correlacionadas.

Para ilustrar esse fato, foram omitidas do conjunto de dados 20 localizações espaciais da saturação por bases. Foram estimados novamente, em todas as abordagens, os parâmetros dos modelos finais selecionados. Ocorreu uma diferença moderada entre os resultados das estimativas calculadas com a omissão de localizações e os das calculadas com a informação completa da saturação por bases. No

entanto, o objetivo é avaliar a qualidade preditiva dos modelos univariados e bivariados.

Com intuito de verificar se existe vantagem utilizando os modelos bivariados, foi conduzida a krigagem ordinária da saturação por bases nas 20 localizações retiradas do conjunto de dados sob os três modelos reajustados. A Tabela 4 apresenta os resultados descritivos das diferenças entre os valores observados e os valores preditos nas três abordagens de modelagem. Os modelos bivariados se mostraram mais eficientes para fazer a krigagem, uma vez que, os desvios padrões dos erros de krigagem são menores do que o valor gerado pela abordagem univariada, ou seja, há evidências de que é melhor utilizar a informação do pH para fazer as inferências sobre a saturação por bases com menos observações.

Estatística	Univariado	BGCCM	BCRM
Média	-0,34	-0,50	-0,50
Desvio Padrão	7,52	3,02	2,84

TABELA 4: Médias e desvios padrões dos erros de krigagem da saturação por bases nas 20 localizações omitidas no processo de estimação

## CONCLUSÕES

Esse trabalho teve o objetivo de avaliar o comportamento de modelos geoestatísticos bivariados, em conjunto com estimação por máxima verossimilhança e krigagem ordinária. Foram feitas implementações computacionais e comparações entre dois modelos bivariados, bem como procurou verificar a aplicabilidade desses modelos com dados observacionais e suas vantagens com relação aos modelos univariados.

Os modelos geoestatísticos univariados e bivariados em conjunto com estimadores por máxima verossimilhança se mostraram eficientes para gerar mapas preditivos do pH e da saturação por bases da fazenda Tupã.

A análise dos dados mostrou que os modelos geoestatísticos bivariados são uma ferramenta útil para a agricultura de precisão, pois sua utilização pode baratear os custos dos monitoramentos de solos e ao mesmo tempo gerar mapas preditivos com alta qualidade.

São disponibilizados em [www.leg.ufpr.br/papercompanions](http://www.leg.ufpr.br/papercompanions) os códigos, resultados e dados utilizados aqui, em que corresponde ao complemento online deste artigo.

## LITERATURA CITADA

CARVALHO, M.P. et al. Variabilidade espacial de atributos de um solo sob videira em Vitória Brasil (SP). R. Bras. Ci. Solo, Viçosa, 27:695-703, 2003.

DIGGLE, P.J. & RIBEIRO Jr., P.J. Model-Based geostatistics. New York, Springer, 2007. 230p.

FARACO, M.A.; URIBE-OPAZO, M.A.; SILVA E.A.A.; JOHANN J.A. & BORSSOI J.A. Nonstationary

multivariate process modeling through spatially varying coregionalization. *R. Bras. Ci Solo, Viçosa*, 32:463-476, 2008.

GELFAND, A.E.; SCHMIDT, A.M.; BANERJEE S. & SIRMANS, C.F. Nonstationary multivariate process modeling through spatially varying coregionalization. *Test, Valencia*, 13:263-312, 2004.

LE, D.N. & ZIDEK, J.V. *Statistical analysis of environmental space-time processes*. New York, Springer, 2006. 327p.

MATÈRN, B. *Spatial variation*. Berlin, Springer, 1986. 365 p.

NELDER, J.A. & MEAD R. A simplex method for function minimization. *Comput. Journal, Oxford*, 7:308-313, 1965.

OLIVEIRA, J.J. et al. Variabilidade espacial de propriedades químicas em um solo salino-sódico. *R. Bras. Ci. Solo, Viçosa*, 23:783-789, 1999.

R: A Language and Environmental for Statistical Computing, R Development Core Team, ISBN 3-900051-07-0. Disponível em: <<http://www.R-project.org>>. Acesso em 28 mai 2007.

RIBEIRO Jr, P.J. & DIGGLE, P.J. geoR: A package for geostatistical analysis. *R-NEWS, Vienna*, 1:14-18, 2001. ISSN 1609-3631.

SCHABENBERGER, O. & GOTWAY, C.A. *Statistical methods for spatial data analysis*. Boca Raton, Chapman and Hall / CRC, 2005. 488p.

SCHIMIDT, A.M. & SANSÓ, B. Modelagem bayesiana da estrutura de covariância de processos espaciais e espaço-temporais. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 14, Caxambú, 2006. Minicurso... São Paulo: Associação Brasileira de Estatística, 2006. 151 p.

SCHULLER, J.K. A review and integrating analysis of spatially-variable crop control of crop production. *Fert. Res., Amsterdam*, 33:1-34, 1992.

SILVA, A.S. Modelos gaussianos geoestatísticos espaço-temporais e aplicações, 2006. 70 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz" da Univeridade de São Paulo, 2006.

SILVA, V.R. et al. Variabilidade espacial das características químicas do solo e produtividade de milho em um Argissolo Vermelho-Amarelo distrófico arênico. *R. Bras. Ci. Solo, Viçosa*, 27:1013-1020, 2003.