Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelos geoestatísticos gaussianos bivariados

Bruno Henrique Fernandes Fonseca

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Bruno Henrique Fernandes Fonseca

Bacharel em Estatística

Modelos geoestatísticos gaussianos bivariados

Orientador:

Prof^{a}. Dr^{a}. **Paulo Justiniano Ribeiro Jr.**

Dissertação apresentada para obtenção do título de Mestre em Agronomia. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2008

Ficha Catalográfica

Dedicatória

AGRADECIMENTOS

SUMÁRIO

RESUMO

Modelos geoestatísticos gaussianos bivariados

ABSTRACT

Bivariate gaussian geostatistic models

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

A modelagem estatística é um conjunto de ferramentas muito importante em diversos campos do conhecimento, que utilizam essas técnicas para tentar descrever o comportamento de um ou mais atributos que não possuem um modelo determinístico. De uma forma geral, os modelos estatísticos tentam explicar, o máximo possível, a variabilidade dos processos estocásticos através de uma ou mais variáveis explanatórias que possuam alguma associação ou correlação com a resposta de interesse.

Os primeiros modelos estatísticos propostos foram os lineares univariados, que assumem erros aleatórios independentes e identicamente distribuídos de uma distribuição de probabilidade gaussiana, além disso, todas as variáveis explanatórias eram consideradas fixas, ou seja, não existem distribuições de probabilidades associadas às covariáveis. No entanto, essas simplificações não são válidas na maioria dos processos naturais, logo, surgiu a necessidade de desenvolver técnicas mais sofisticadas para tentar modelar processos que possuem estruturas mais complexas de variabilidade.

Um campo de pesquisas que teve grande evolução nos últimos tempos foi a estatística espacial, que é formada por três grandes áreas de estudo: geoestatística, dados de área e processos pontuais, que são utilizadas conforme o tipo de dados em questão, neste trabalho será estudada apenas a primeira. A modelagem geoestatística é um conjunto de técnicas que tenta encontrar uma boa função matemática para um ou mais atributos que possuem localizações espaciais e pontuais conhecidas, sendo assim, essas ferramentas são úteis para capturar a correlação entre as observações dos atributos sob estudo, onde existe uma forte suspeita de que pontos espaciais mais próximos possuem valores observados dos atributos mais parecidos, ou seja, a estrutura de correlação entre as observações do processo estocástico é determinada através das distâncias entre os pontos espaciais amostrados. A abordagem geoestatística se diferencia dos modelos lineares univariados nos pressupostos, onde agora todas as observações não são independentes e existe efeito aleatório latente na parte explanatória do modelo.

Diversas pesquisas de distintas áreas podem possuir mais de uma variável resposta de interesse, ou seja, os pesquisadores possuem dois ou mais atributos que devem ser modelados, se esses atributos sob estudo forem independentes deve-se propor um modelo

estatístico para cada um deles, no entanto, se há evidências de que esses processos não sejam independentes e existindo uma explicação prática, modelos multivariados podem ser propostos, ou seja, os modelos estatísticos devem capturar ao máximo a correlação entre as variáveis respostas, para tal, algumas técnicas têm sido utilizadas, assim como, distribuições de probabilidades conjuntas.

Neste contexto, pode-se pensar em modelos geoestatísticos multivariados, ou seja, há mais de uma resposta de interesse e existe uma forte evidência de que esses processos estocásticos sejam correlacionados. Sendo assim, o modelo deve capturar a correlação entre todas as observações dentro de cada variável e entre as variáveis. Na literatura existem algumas formas distintas de estudar esse tipo de problema. No entanto, devido a complexidade dos modelos e o número elevado de parâmetros pode existir problemas para estimação dos parâmetros, além disso, em alguns casos pode ocorrer problemas com a identificabilidade do modelo, por conta disso, este trabalho, inicialmente, apresenta um estudo de simulação de modelos geoestatísticos bivariados, dessa forma pode-se fazer uma comparação entre as metodologias, e detectar quais são as vantagens, probabilísticas e computacionais, de cada método em diversas configurações paramétricas.

A agricultura de precisão é um campo de pesquisa que pode utilizar os modelos geoestatísticos e de forma bem sucinta, é o conjunto de métodos aplicados ao manejo da variabilidade. 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. Vários autores concluíram que a variabilidade espacial existe, mesmo em áreas consideradas homogêneas, inúmeros 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, sendo que estes trabalhos mostram que a variabilidade do solo não é puramente aleatória, apresentando correlação ou dependência espacial. Nesse contexto, diversas abordagens podem ser utilizadas para estruturar a variabilidade espacial das variáveis químicas do solo, em agricultura de precisão, tradicionalmente, é utilizada apenas medidas descritivas e a intuição pessoal dos pesquisadores para tal, o que pode gerar o problema de não ser possível mensurar a precisão dos resultados. Uma solução recente para esse

tipo de problema, que proporciona resultados com respaldo probabilístico, é a modelagem geoestatística em conjunto com estimação por máxima verossimilhança.

Sendo assim, além do estudo de simulação, essa pesquisa utiliza ferramentas geoestatísticas para estudar a variabilidade espacial de duas variáveis químicas do solo de uma propriedade agrícola, sendo que, a utilização de modelos geoestatísticos bivariados é possível devido a natureza dos atributos, que são fortemente correlacionados e, além disso, uma das resposta é mais dispendiosa para ser observada, sendo assim, com a estrutura conjunta de correlação espacial estabelecida, em monitoramentos futuros da mesma propriedade agrícola será possível diminuir os gastos com a coleta de informações.

2 REVISÃO DE LITERATURA

2.1 Campos aleatórios

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 $Z(s_i)$ a notação a variável aleatória Z na localização s_i do espaço sob estudo G.

Segundo Schmidt e Sansó (2006) e Le e 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,...,s_n) pertencentes à região G e qualquer inteiro n:

$$F_{S_1,S_2,...,S_n}(z_1, z_2, ..., z_n) \equiv P(Z(s_1) \leq z_1, Z(s_2) \leq z_2, ..., Z(s_n) \leq z_n)$$

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 localizações do espaço sob estudo G, logo, para qualquer conjunto finito de localizações $s = (s_1, s_2, ..., s_n)$ pertencente a G, $Z(s)$ segue uma distribuição normal n-variada e é completamente especificado pelo vetor de média $n \times 1$, notado por μ, e pela matriz de covariâncias $n \times n$, notada por Σ, que no contexto de geoestatística possui o comportamento empírico de que quanto maior a distância euclidiana entre duas localizações s_l e s_k quaisquer, menor a correlação entre $Z(s_l)$ e $Z(s_k)$, porém não é trivial encontrar uma forma para gerar esse comportamento e, ao mesmo tempo, assegurar que a matriz de covariâncias fique positiva definida. Diggle e Ribeiro (2006) mostram maiores detalhes sobre campos aleatórios gaussianos.

Em pesquisas de geoestatística, geralmente, 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 que o processo estocástico é estacionário, ou seja, a distribuição de probabilidade associada ao campo aleatório não depende da grandeza de escala das coordenadas, logo, a distribuição conjunta de $(Z(s_1), Z(s_2), ..., Z(s_n))$ é igual a distribuição conjunta de $(Z(s_1 + h), Z(s_2 + h), ..., Z(s_n+h))$, para qualquer incremento h. Outra definição menos restritiva é que a média

do campo aleatório é igual em toda a região sob estudo e a correlação entre $Z(s_l)$ e $Z(s_k)$, para quaisquer s_l e s_k, só depende da distância entre as localizações, ou seja, a grandeza de escala de Z não influencia na estrutura de correlação espacial. Esse tipo de estacionariedade é conhecido na literatura como estacionariedade fraca ou de segunda ordem, uma observação importante é que a primeira restrição implica na segunda, no entanto, o contrário não é válido, a não ser que o processo espacial seja gaussiano, que produz equivalência entre as duas restrições. No entanto, nem sempre é fácil verificar as restrições de estacionariedade forte ou fraca, logo, outra possibilidade menos restritiva é assumir que os incrementos $[Z(s)-Z(s+h)]$ possuem estacionariedade. Esta característica é denominada de estacionariedade intrínseca (SCHANBENBERGER; GOTWAY, 2005). Sendo assim, um campo aleatório é dito ser intrinsecamente estacionário se para todo s_i pertencente a G, $E[Z(s_i)]=\mu$ e $Var[Z(s_i) - Z(si + h)] = 2\gamma(h)$, sendo $\gamma(h) = Var(Z(s_i)) - Cov(Z(s_i); Z(s_i + h))$ e denomidado de semivariograma.

No entanto, em alguns casos a suposição de estacionariedade não é válida, sendo assim, diversas abordagens são propostas para contornar esse problema. Quando a média do processo estocástico não é constante na região sob estudo, geralmente, utiliza-se a inclusão de covariáveis na modelagem, onde a média é tratada como efeito fixo, e sua interpretabilidade é igual a de modelos lineares, Diggle e Ribeiro Jr. (2006) detalham melhor essa tecnicalidade. Já para problemas com variâncias e covariâncias não constantes, uma técnica mais simples é a utilização de transformação nos valores observados do campo aleatório (CHRISTENSEN; DIGGLE; RIBEIRO Jr., 2001).

Mesmo que um campo aleatório possua algum tipo de estacionariedade, o padrão espacial de correlação pode depender das distâncias e das direções envolvidas entre as localizações, a essa característica é dado o nome de anisotropia, em problemas práticos de geoestatística não é fácil identificar tal característica nos dados observacionais, no entanto, a natureza de algumas variáveis exige a utilização de tal abordagem, como por exemplo, estudos sobre poluição, onde o padrão dos ventos gera correlação espacial dependente das direções entre localizações amostradas. Silva (2006) mostra a tecnicalidade utilizada em problemas de anisotropia geométrica.

Um campo aleatório gaussiano é dito ser homogêneo se ele for estacionário e

o seu padrão de correlações não depende das direções. Utilizando essa suposição o processo estocástico fica bastante restritivo, porém 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, positiva definida, com o comportamento empírico utilizado em geoestatística. Devido a complexidade para encontrar tais funções, diversas propostas são sugeridas.

Função de Matérn

Essa família de funções de correlação foi proposta por Berfil Matérn (1986) e possui a seguinte forma:

$$\rho(h) = 2^{\kappa} - \Gamma(\kappa)^{-1}(h/\phi)^{\kappa}K_{\kappa}(h/\phi),$$

sendo h a distância euclidiana entre duas localizações quaisquer do campo aleatório, os parâmetros dessa função são $\phi > 0$ e $\kappa > 0$, sendo o primeiro vinculado ao alcance das correlações, quanto menor o parâmetro, menor o alcance das correlações, ou seja, somente observações muito próximas possuem correlação significativa. Já o segundo parâmetro é vinculado a suavidade do processo, quanto maior κ, maior a suavidade. E por último, $K_{\kappa}(.)$ é a função Bessel de ordem κ.

Função Exponencial Potência

$$\rho(h) = \exp{(h/\phi)^{\kappa}},$$

essa família possui as mesmas características da função de Matérn, no entanto, agora κ é limitado no intervalo $[0, 2]$.

Silva (2006) apresenta diversas outras funções de correlação conhecidamente válidas, no entanto, essas duas funções são muito utilizadas devido a capacidade de produzir comportamentos distintos quanto a suavidade do processo, ou seja, é possível modelar processos mais ou menos diferenciáveis. Além disso, sob estacionariedade fraca, essas funções possuem propriedade conhecidas e desejáveis, Schabenberger e Gotway (2005) apresentam e discutem detalhes sobre tais propriedades.

2.2 Modelos geoestatísticos gaussianos univariados

Considerando que em alguma área G exista um campo aleatório gaussiano Z latente, ou seja, o processo existe mas não é conhecido, sendo assim, é necessário fazer uma amostragem de n localizações espaciais dentro da área G e observar valores do atributo de interesse nas localizações amostradas. Logo, existe um vetor $Y(s)$ $n \times 1$ de valores observados em $s = (s_1, s_2, ..., s_n)$, Diggle e Ribeiro Jr (2006) utilizam a seguinte modelagem para o problema:

$$Y(s) = \mu + Z(s) + \epsilon, \tag{1}$$

sendo $\mu = X\beta$, onde X é uma matriz $n \times q$ contento $q-1$ possíveis covariáveis, β 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 Σ, de dimensão $n \times n$, onde 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 ϵ um vetor nx1 de ruídos brancos, que por suposição, são independentes e identicamente distribuidos de uma distribuição de probabilidade normal com média zero e desvio padrão τ.

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 $\Sigma_Y = \Sigma + \tau^2 I$, onde I é uma matriz identidade $n \times n$. Logo, existe um vetor de parâmetros $\theta = (\beta, \sigma^2, \phi^*, \tau^2)$ a ser estimado, onde ϕ^* é um vetor de parâmetros associados a função de correlação utilizada, cabe observar que $Cov(Z(s_i); Z(s_j)) = \rho(Z(s_i); Z(s_j))\sigma^2$, para todo s_i e s_j pertencentes a s e $\sigma^2 > 0$.

2.2.1 Estimação dos parâmetros

Estabelecidas às estruturas paramétricas, o próximo passo é fazer a estimação dos parâmetros. Se o campo aleatório é intrinsecamente estacionário pode-se trabalhar com uma estimação para o semivariograma, abaixo segue a expressão de uma estimativa empírica para o semivariograma através dos estimadores de momentos:

$$\widehat{\gamma}(h) = \frac{\sum_{|N(h)|}(Z(s_i) - Z(s_j))^2}{2|N(h)|} \tag{2}$$

em que $|N(h)|$ é o número de pontos abrangidos pela distância h.

Devido a relação entre o semivariograma empírico e as funções de correlação válidas, muitos trabalhos aplicados de geoestatística utilizam um modelo em função dos parâmetros de variabilidade e correlação que se ajuste aos valores calculados em (2), isso pode ser feito por meio de métodos "AD-HOC", no entanto, essa abordagem para estimar os parâmetros de variabilidade e da função de correlação podem não ser muito precisos, pois os valores dos semivariogramas empíricos podem se afastar muito do semivariograma real e desconhecido , devido ao tamanho e acaso amostral, sendo assim, esse tipo de estimação deve ser utilizado, na maioria dos casos, apenas como análise descritiva inicial, Diggle e Ribeiro Jr. (2006) discutem com mais detalhes esse assunto.

Por outro lado, assumindo que o campo aleatório possui estacionariedade forte, pode-se optar por estimadores de máxima verossimilhança ou máxima verossimilhança restrita, que consiste em utilizar os valores observados da variável resposta para encontrar um vetor $\widehat{\theta}$ que seja o ponto de máximo da função de verossimilhança associada a θ, no entanto, por simplicidade matemática, normalmente, utiliza-se o logaritmo da função de verossimilhança para fazer a estimação, segue a função associada a (1):

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

Maiores detalhes sobre técnicas e propriedades da estimação por máxima verossimilhança são expostas por Azzalini (1996) e Bickel e Doksum (1976). No contexto de geoestatística, Diggle e Ribeiro Jr. (2006) propõe a utilização da reparametrização $\nu = \tau/\sigma$, a qual facilita a estimação de θ. Sendo assim, agora temos o vetor de parâmetros $\theta^* = (\beta, \sigma^2, \phi^*, \nu^2)$ a ser estimado e Σ_Y pode ser escrito como $\sigma^2 V$, onde V é uma matriz $n \times n$ que depende apenas de ν e ϕ^*, logo, o logaritmo da função de verossimilhança fica da seguinte forma:

$$l(\theta^*; Y(s)) \propto -0.5(n\ln(\sigma^2) - \ln(|V|) - \sigma^{-2}(Y - X\beta)^t V^{-1}(Y - X\beta) \tag{3}$$

sendo que apenas para β e σ^2 existe forma analítica para os estimadores:

$$\widehat{\beta} = (X^t V^{-1} X)^{-1}(X^t V^{-1} Y) \quad \widehat{\sigma^2} = n^{-1}(Y - X\beta)^t V^{-1}(Y - X\beta)$$

observe que $\widehat{\beta}$ e $\widehat{\sigma^2}$ são funções dos demais parâmetros e além das formas fechadas para os estimadores, é trivial encontrar a matriz de informação de Fischer observada para os mesmos,

logo, é fácil encontrar a matriz de covariâncias associada a $(\widehat{\beta}, \widehat{\sigma^2})$ (DIGGLE; RIBEIRO Jr., 2006).

Para ϕ^* e ν^2 não existe forma analítica para os estimadores, sendo assim, utilizando $\widehat{\beta}$ e $\widehat{\sigma^2}$ em (3), tem-se o logaritmo da função de verossimilhança concentrada, que depende apenas de $\theta_c = (\nu, \phi^*)$.

Para encontrar $\widehat{\theta_c}$ podemos utilizar métodos numéricos de maximização de funções, como por exemplo, o método de Nelder e Mead (1965), o qual calcula $\widehat{\phi^*}$ e $\widehat{\nu^2}$ e a matriz Hessiana estimada, que denotaremos por H. Logo, com os parâmetros da função de máxima verossimilhança concentrada estimados, podemos encontrar, por invariância, as estimativas de β e σ^2 e τ^2.

Utilizando as propriedades assintóticas dos estimadores de máxima verossimilhança e o método Delta podemos encontrar a distribuição de probabilidade de $\widehat{\theta}$, que é $N(\theta; \Sigma_\theta)$, sendo $\Sigma_\theta = \Delta^t \Sigma_{\theta^*} \Delta$, onde a i-ésima coluna de Δ é o vetor $\frac{\partial l(\theta_i)}{\partial \theta^*}$ e:

$$\Sigma_{\theta^*} = \begin{bmatrix} \Sigma_{\beta,\sigma^2} & O \\ O^t & \Sigma_{\theta_c} \end{bmatrix}$$

sendo Σ_{β,σ^2} a matriz de covariâncias de $(\widehat{\beta}, \widehat{\sigma^2})$, que possui forma analítica, $\Sigma_{\theta_c} = -H^{-1}$ é a matriz de covariâncias de $\widehat{\theta_c}$, e O uma matriz de zeros.

No entanto, a maioria dos estudos com dados georeferenciados tem maior interesse sobre a predição espacial do campo aleatório, sendo assim, tecnicalidades de krigagens devem ser utilizadas.

2.2.2 Krigagem

A krigagem nada mais é do que o processo de predição do campo aleatório em localizações não amostradas. O nome krigagem é uma homenagem ao pesquisador sul-africano D.G. Krige que foi um dos pioneiros em estudos de predição espacial.

Antes de expor as técnicas de krigagem, é necessário utilizar algumas propriedades inferenciais da distribuição gaussiana multivariada. Suponha que o interesse é fazer predição para Z nas localizações s^*, ou seja, devemos fazer predição para $Z(s^*)$ tal que o erro quadrático médio seja mínimo, Diggle e Ribeiro Jr. (2006) mostram que a distribuição

de probabilidade de $(Z(s^*)|Y(s))$ gera as predições com melhor precisão, sendo assim:

$$E(Z(s^*)|Y(s)) = \mu_Z + \Sigma_{Z,Y}\Sigma_Y^{-1}(Y(s) - X\beta) \tag{4}$$

sendo μ_Z a média de $Z(s^*)$ e $\Sigma_{Z,Y}$ é a matriz de covariâncias cruzadas entre $Z(s^*)$ e $Y(s)$. Além disso a variância preditiva é conhecida:

$$Var(Z(s^*)|Y(s)) = \Sigma_Z - \Sigma_{Z,Y}\Sigma_Y^{-1}\Sigma_{Z,Y}^t \tag{5}$$

sendo Σ_Z a matriz de covariâncias de $Z(s^*)$. Cabe ressaltar que as expressões (4) e (5) são obtidas através das propriedades da distribuição de probabilidade gaussiana multivariada, Diggle e Ribeiro Jr. (2006) mostram mais detalhes.

Do ponto de vista geoestatístico, as krigagens mais utilizadas são a simples e a ordinária, que utilizam (3) e (4), mas que se diferenciam quanto a suposição de conhecimento sobre os parâmetros (DIGGLE; RIBEIRO Jr., 2006).

2.3 Modelos geoestatísticos gaussianos bivariados

Em muitos estudos o interesse não é sobre um único atributo como, como por exemplo, avaliação da qualidade do solo, que geralmente, utiliza um processo de amostragem de localizações pertencentes a uma certa região sob estudo para observar diversos atributos relativos ao solo, os quais são importantes para nortear tomadas de decisão quanto ao manejo e trato do solo. Sendo assim, existe mais de um campo aleatório a passar pelo processo de modelagem e krigagem, a intuição inicial é que seja feito esse processo para cada atributo individualmente. Porém, pode haver correlação estatística entre alguns atributos, o que leva a possibilidade de adotar modelos geoestatísticos multivariados, contudo, em muitos casos, somente com essa justificativa estatística não é vantajoso utilizar tal abordagem, é necessário que existam justificativas e vantagens práticas para esse aumento de complexidade dos modelos geoestatísticos, que agora devem capturar a correlação entre e dentro dos atributos.

Por simplicidade, iremos supor a existência de dois campos aleatórios gaussianos sob estudo, que podem ser modelados da seguinte forma:

$$Y_i = \mu_i + Z_i, \ i = 1, 2 \tag{6}$$

sendo Y_i um vetor $n_i \times 1$ de valores observados do campo aleatótio gaussiano latente Z_i, que possui vetor de médias nulo $n_i \times 1$ e matriz de covariâncias Σ_i $n_i \times n_i$, μ_i é um vetor, que possui os parâmetros de médias associados a Y_i. Contudo, o interesse é sob o vetor $Y = (Y_1, Y_2)$, que possui distribuição gaussiana n-variada, sendo $n = n_1 + n_2$, com vetor de médias $\mu = (\mu_1, \mu_2)$ e matriz de covariâncias Σ_Y, positiva definida que possui o comportamento empírico de correlações utilizado em geoestatística e que pode ser particionada:

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

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

Nesse contexto, a maior dificuldade é propor Σ_Y válida, algumas abordagens são propostas na literatura, modelos separáveis é a mais simples e se baseia na suposição de isotropia dos campos aleatórios e utiliza as propriedades das funções de correlação para estruturar a matriz de covariâncias válida para Y.

Na literatura de geoestatística algumas propostas utilizam modelos separáveis para estruturar Σ_Y, a maioria utiliza decomposições dos termos Z_1 e Z_2 de (6) para tal.

2.3.1 Modelo gaussiano bivariado com componente de correlação parcialmente comum

Diggle e Ribeiro Jr. (2006) proproram essa abordagem para problemas geoestatísticos bivariados, a qual chamaremos de BGCCM, que utiliza as seguintes decomposições dos campos aleatórios latentes de (6):

$$Z_i = \sigma_{0i} S_{0i} + \sigma i S_i, \ i = 1, 2, \tag{7}$$

sendo $\sigma^* = (\sigma_{01}, \sigma_1, \sigma_{02}, \sigma_2)$ um vetor de parâmetros de dispersão associados a (6) e S_{01}, S_1, S_{02} e S_2 campos aleatórios gaussianos mutuamente independentes, com vetores de médias nulos, variâncias unitárias e correlações determinadas por funções de correlação válidas quaisquer, as quais vão gerar empiricamente as correlações cruzadas entre Y_1 e Y_2, uma vez que, as funções de correlação adotadas para S_{01} e S_{02} devem ser idênticas, ou seja, existe

uma componente de correlaçào comum as duas respotas, logo, estamos adotando três funções de correlação válidas e utilizando somas das mesmas para gerar uma matriz Σ_Y válida.

Logo, (6) fica da seguinte maneira:

$$\begin{cases} Y_1 = \mu_1 + \sigma_{01}S_{01} + \sigma_1 S_1 \\ Y_2 = \mu_2 + \sigma_{02}S_{02} + \sigma_2 S_2 \end{cases}$$

Definindo $Y_i(s_l)$ e $Y_i(s_k)$ como observações feitas do atributo Z_i em duas localizações quaisquer s_l e s_k, que estão separados pela distância euclidiana h, para todo $l, k = 1, 2, ..., n_i$ e $i = 1, 2$, tem-se que o elemento $\Sigma_{i,(l,k)}$ é a $Cov(h) = \sigma_{0i}^2\rho_0(h) + \sigma_i^2\rho_i(h)$, sendo ρ_0 e ρ_i as funções de correlação adotadas para S_{0i} e S_i, respectivamente. Utilizando propriedades básicas de covariâncias, pode-se encontrar $\Sigma_{1,2}$, que é igual a $\sigma_{01}\sigma_{02}R_0$, onde R_0 é uma matriz $n_1 \times n_2$ das correlações cruzadas entre as respostas e depende da função de correlação adotada para S_{01} e S_{02}. Dessa forma Σ_Y fica completamente estruturada, por simplicidade, suponha que existam apenas duas localizações amostradas, logo:

$$\Sigma_Y = \begin{bmatrix} \sigma_{01}^2 + \sigma_1^2 & \sigma_{01}^2\rho_0(h) + \sigma_1^2\rho_1(h) & \sigma_{01}\sigma_{02} & \sigma_{01}\sigma_{02}\rho_0(h) \\ \sigma_{01}^2\rho_0(h) + \sigma_1^2\rho_1(h) & \sigma_{01}^2 + \sigma_1^2 & \sigma_{01}\sigma_{02}\rho_0(h) & \sigma_{01}\sigma_{02} \\ \sigma_{01}\sigma_{02} & \sigma_{01}\sigma_{02}\rho_0(h) & \sigma_{02}^2 + \sigma_2^2 & \sigma_{02}^2\rho_0(h) + \sigma_2^2\rho_2(h) \\ \sigma_{01}\sigma_{02}\rho_0(h) & \sigma_{01}\sigma_{02} & \sigma_{02}^2\rho_0(h) + \sigma_2^2\rho_2(h) & \sigma_{02}^2 + \sigma_2^2 \end{bmatrix}$$

Sendo assim, a distribuição de probabilidade do vetor $Y = (Y_1, Y_2)$ está estabelecida e depende do vetor de parâmetros $\theta = (\beta^*, \sigma^*, \phi_0^*, \phi_1^*, \phi_2^*)$, onde β^* é um vetor de parâmetros associado a μ e ϕ_j^* é um vetor de parâmetros associado a escolha da função ρ_j, para todo $j = 0, 1, 2$.

Novamente técnicas de estimaçào por máxima verossimilhança podem ser utilizadas para encontrar $\widehat{\theta}$, o qual deve ser plugado nas expressões de predição espacial, que são idênticas as utilizadas no modelo geoestatístico univariado.

2.3.2 Modelo bivariado de co-regionalização

Essa abordagem foi proposta inicialmente por Matheron (1982), chamaremos de BCRM e é uma abordagem concorrente ao BGCCM para propor uma estrutura

paramétrica válida para Σ_Y, agora os componentes aleatórios de (6) são decomposto de outra maneira:

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

sendo S_{11}, S_{12} e S_{22} campos aleatórios gaussianos mutuamente independentes, com vetores de médias nulo, variância unitária e com correlações determinadas pela escolha de funções de correlação conhecidamente válidas, sendo que a escolha para S_{11} e S_{12} deve ser idêntica, ou seja, agora a variável Y_1 possui apenas um termo de correlação, que é comum as duas variáveis, essa abordagem gera um número menor de parâmetros associados as funções de correlação adotadas, apenas duas, e a variabilidade espacial que agora é $\sigma^* = (\sigma_{11}, \sigma_{12}, \sigma_{22})$. Gelfand et al. (2005) mostram mais detalhes sobre essa abordagem do ponto de vista de modelos geoestatísticos.

Utilizando o BCRM temos que (6) fica da seguinte forma:

$$\begin{cases} Y_1 = \mu_1 + \sigma_{11} S_{11} \\ Y_2 = \mu_2 + \sigma_{12} S_{12} + \sigma_{22} S_{22} \end{cases}$$

Definindo $Y_i(s_l)$ e $Y_i(s_k)$ como observações feitas do atributo Z_i em duas localizações quaisquer s_l e s_k, que estão separadas pela distância euclidiana h, para todo $l, k = 1, 2, ..., n_i$ e $i = 1, 2$, tem-se que o elemento $\Sigma_{1,(l,k)}$ é a $Cov(h) = \sigma_{11}^2 \rho_1(h)$ e o elemento $\Sigma_{2,(l,k)}$ é a $Cov(h) = \sigma_{12}^2 \rho_1(h) + \sigma_{22}^2 \rho_2(h)$ sendo ρ_1 e ρ_2 as funções de correlação adotadas. Utilizando propriedades básicas de covariâncias, pode-se encontrar $\Sigma_{1,2}$, que é igual a $\sigma_{11}\sigma_{12}R_1$, onde R_1 é uma matriz $n_1 \times n_2$ das correlações cruzadas entre as respostas e depende apenas da função de correlação adotada para Y_1. Dessa forma Σ_Y fica completamente estruturada, por simplicidade, suponha que existam apenas duas localizações amostradas, logo:

$$\Sigma_Z = \begin{bmatrix} \sigma_{11}^2 & \sigma_{11}^2\rho_1(h) & \sigma_{11}\sigma_{12} & \sigma_{11}\sigma_{12}\rho_1(h) \\ \sigma_{11}^2\rho_1(h) & \sigma_{11}^2 & \sigma_{11}\sigma_{12}\rho_1(h) & \sigma_{11}\sigma_{12} \\ \sigma_{11}\sigma_{12} & \sigma_{11}\sigma_{12}\rho_1(h) & \sigma_{12}^2 + \sigma_{22}^2 & \sigma_{12}^2\rho_1(h) + \sigma_{22}^2\rho_2(h) \\ \sigma_{11}\sigma_{12}\rho_1(h) & \sigma_{11}\sigma_{12} & \sigma_{12}^2\rho_1(h) + \sigma_{22}^2\rho_2(h) & \sigma_{12}^2 + \sigma_{22}^2 \end{bmatrix}$$

Sendo assim, a distribuição de probabilidade do vetor $Y = (Y_1, Y_2)$ está estabelecida e depende do vetor de parâmetros $\theta = (\beta^*, \sigma^*, \phi_1^*, \phi_2^*)$, onde β^* é um vetor de parâmetros associado a μ e ϕ_j^* é um vetor de parâmetros associado a escolha da função ρ_j, para todo $j = 1, 2$. Logo, novamente o problema é fazer a estimação dos parâmetros por e depois calcular as predições espaciais.

3 MATERIAL E MÉTODOS

Alguns trabalhos vêm utilizando modelos geoestatísticos bivariados para estruturar a dependência espacial entre e dentro atributos de interesse. No entanto, não existem pesquisas comparando o BGCCM e o BCRM sob o enfoque frequentista. Sendo assim, esse trabalho concentra-se em utilizar tais modelos e verificar suas vantagens e desvantagens em análise de dados.

Com o intuito de verificar a qualidade das estimativas por máxima verossimilhança e das krigagens utilizando os modelos bivariados em questão, inicialmente conduzimos um estudo de simulação com diversas configurações para o BGCCM e o BCRM.

Após o estudo de simulação analisamos detalhadamente dados observacionais provinientes de uma pesquisa sobre qualidade do solo de uma propriedade agrícola.

Cabe ressaltar que todas as análises e resultados foram obtidos através do ambiente R de programação (R Development Core Team, 2005), sendo que o pacote mais utilizado foi o *geoR* (Ribeiro Jr. e Diggle, 2001), no entanto, para os modelos bivariados a maioria das tecnicalidades não está implementada nos pacotes do R, sendo assim, seguem, em anexo, as programações utilizadas.

3.1 Estudo de Simulação

Essa fase do trabalho é muito importante para detectar se existem possíveis problemas com os modelos geoestatísticos bivariados e verificar as vantagens e desvantagens de cada abordagem considerada. Nesse contexto, selecionamos 3 configurações de cada abordagem bivariada e simulamos 1000 conjuntos de dados para cada uma dentro de um quadrado de lado igual a 1 e com grid irregular, sendo que, a cada vetor $Y = (Y_1, Y_2)$ de dados simulados separamos 20 observações de cada resposta e suas respectivas localizações espaciais para análise de predição, as demais observações foram utilizadas no processo de estimação dos parâmetros por máxima verossimilhança.

Para simular observações com o BGCCM, consideramos os parâmetros de média e variabilidade sempre fixos em $\mu_1 = 150$, $\mu_2 = 60$, $\sigma_{01} = 8$, $\sigma_1 = 4$, $\sigma_{02} = 5$ e $\sigma_2 = 2$, logo, $Var(Y_1) = \sigma_{01}^2 + \sigma_1^2 = 80$, $Var(Y_1) = \sigma_{02}^2 + \sigma_2^2 = 29$. Adotamos a proposta de Matérn para as três funções de correlação, sendo que, $\phi_0 = 0,25$, $\phi_1 = 0,2$, $\phi_2 = 0,2$, $\kappa_0 = 0,5$, $\kappa_1 = 0,5$

e $\kappa_2 = 0,5$. Utilizando essa configuração paramétrica temos os dois campos aleatórios são homogêneos, além disso, os parâmetros escolhidos para as estruturar as correlações geram alcanes práticos, distância h em que as $\rho(h) = 0.05$, aproximadamente iguais a $0,75$, $0,60$ e $0,60$ para ρ_0, ρ_1 e ρ_2, respectivamente. Já para a amostragem de localizações, consideramos três configurações distintas: na primeira utilizamos $n_1 = n_2 = 100$ e as localizações das duas respostas completamente coincidentes, dados co-locados; na segunda utilizamos utilizamos dados completamente co-locados, porém desbalanceados, sendo $n_1 = 100$ e $n_2 = 50$; e na última configuração consideramos n_1=$n_2 = 100$ porém apenas metade dos dados co-locados.

Com o BCRM, utilizamos as configurações das localizações e dos parâmetros de média identicas as do BGCCM e os demais parâmetros foram fixos em $\sigma_1 = 9$, $\sigma_{12} = 5$, $\sigma_2 = 2$, $\phi_1 = 0,25$, $\phi_2 = 0,2$, $\kappa_1 = 0,5$ e $\kappa_2 = 0,5$. Tais configurações foram utilizadas para gerar características similares as da outra abordagem, sendo assim, fica mais fácil fazer algumas comparações entre as propostas.

Estimamos os parâmetros por máxima verossimilhança para cada uma das observações de Y utilizando o mesmo modelo do qual a observação é proviniente e o modelo concorrente. Sendo que, para facilitar a estimação de θ utilizamos reparametrizações dos modelos, no BGCCM definimos $\sigma = \sigma_{01}$, $eta = \sigma_{02}/\sigma_{01}$, $\nu_1 = \sigma_1/\sigma_{01}$ e $\nu_2 = \sigma_2/\sigma_{01}$, já no BCRM usamos $\sigma = \sigma_{11}$, $\nu_1 = \sigma_{12}/\sigma_{11}$ e $\nu_2 = \sigma_{22}/\sigma_{11}$, logo, nas duas abordagens podemos escrever $\Sigma_Y = \sigma^2 V$, sendo assim, existe forma analítica para $\widehat{\mu}$ e $\widehat{\sigma}$ e, novamente, V é função dos demais parâmetros, os quais foram estimados maximizando a função de máxima verossimilhança concentrada pelo método de Nelder e Mead (1965), as tecnicalidades para encontrar $\widehat{\theta}$ e sua distribuição de probabilidade assintótica são similares as utilizadas nos modelos geoestatísticos univariados, onde utilizamos a propriedade de invariância para calcular todas as estimativas e o método Delta para encontrar a matriz de variância de $\widehat{\theta}$. Cabe ressaltar que, em nenhum caso nós estimamos os parâmetros de suavidade das funções de correlação, os mesmos foram considerados fixos.

Para avaliar a qualidade das estimativas, intervalos de confiança marginais para cada parâmetro foram calculados, os quais devem conter em sua maioria os valores dos parâmetros utilizados para fazer a simulação, no entanto, essa ferramenta não é muito precisa, pois temos um problema multiparamétrico e não é trivial encontrar uma superfície

de confiança que seja precisa, sendo assim, também utilizamos erros absolutos médios e erros quadráticos médios na análise das estimativas.

Como para cada conjunto de dados simulados nós armazenamos 20 localizações espaciais e seus respectivos valores observados, utilizamos cada vetor de parâmetros estimados para fazer as krigagens dos campos aleatórios nessas localizações não utilizadas no processo de modelagem. Sendo assim, para cada modelo ajustado, temos 20 valores observados de cada resposta e suas predições espaciais, logo, novamente analisamos intervalos de confiança marginais e medidas de erros.

3.2 Dados sobre a qualidade do solo

Os dados utilizados foram obtidos por meio de uma pesquisa realizada em julho 2006 na fazenda Tupã, localizada no município de Echaporan/SP, que possui 51,8ha de área, solo argissolo de textura média e dois históricos de manejo distintos, soja numa região e pastagem em outra. Para fazer observação das variáveis químicas do solo, foram amostradas 67 localizações, com grid regular a cada hectare, georeferenciadas no sistema Universal Transverse Mercatur (UTM) e então utilizando um aparelho GPS foram coletadas as informações de interesse.

Sendo assim, dois parâmetros levantados pelos pesquisadores serão modelados nesse trabalho: saturação por bases e pH do solo, a primeira variável é uma medida de capacidade do solo reter bons nutrientes N, P, K, Ca, Mg e a segunda variável mede a acidez do solo, sendo que, existe uma forte correlação entre essas duas respostas, o que justifica a tentativa de modelagem bivariada.

Uma análise exploratória inicial dos dados foi conduzida utilizando gráficos e medidas descritivas, os quais detectaram padrão espacial das variáveis químicas. Do ponto de vista geoestatístico, o pH e a saturação por bases são dois campos aleatórios gaussianos e latentes, logo, a intuição inicial é que seja ajustado individualmente o modelo (1) para Y_1 e Y_2, notação para os dois vetores $67x1$ de valores observados da saturação por base e do pH, repectivamente. Sendo assim Y_i, para todo $i = 1, 2$, segue uma distribuição gaussiana 67-variada, com matriz de covariâncias Σ_i e vetor de médias $\mu_i = X_i\beta_i^*$, onde X_i é uma matriz $67 \times p$ com o intercepto e possíveis $p - 1$ covariáveis e β_i^* um vetor $p \times 1$ com os

parâmetros associados a μ_i.

Utilizando os resultados da análise exploratória, consideramos 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, além disso, foi utilizada a família de funções de correlação de Mátern para estruturar a matriz 67×67 de correlações, considerando o parâmetro de suavidade κ fixo nos valores 0,5, 1, 1.5, 2, 2.5, logo, ϕ_i^* depende apenas de ϕ_i, proviniente da função de correlação escolhida. Ou seja, ajustamos diversos modelos combinando as diferentes escolhas de κ e de X_i, sendo que em todos os casos foi utilizada estimação por máxima verossimilhança para os parâmetros, onde utilizamos a reparametrizaçao $\nu_i = \tau_i/\sigma_i$ e o método numérico de Nelder e Mead (1965) para encontrar as estimativas $\widehat{\phi_i}$ e $\widehat{\nu_i}$ que maximizam a função de máxima verossimilhança concentrada, logo, utilizando a propriedade de invariância dos estimadores em questão encontramos as demais estimativas de interesse.

Utilizamos os valores dos máximos estimados da função de máxima verossimilhança concentrada e o Critério de Informação de Akaike (AIC) para selecionar o modelo final de cada atributo. Então foram utilizados os parâmetros estimados dos modelos finais para fazer as krigagens dos campos aleatórios. 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, ajustamos modelos bivariados para o vetor $Y = (Y_1, Y_2)$, utilizando as proposições de Diggle e Ribeiro Jr. (2006) e de Matheron (1982). Nas duas abordagens foi utilizado $\mu = X\beta^*$, sendo X uma matriz $134 \times p$ contento um intercepto para cada campo aleatório e $p-2$ possíveis covariáveis e β^* um vetor $p \times 1$ com os parâmetros de média associados a Y, para estruturar Σ_Y somente funções de correlação da família de Matérn foram utilizadas, logo no BGCCM foram utilizadas três funções de correlação de Matérn e para no BCRM duas, todos os parâmetros κ foram considerados fixos em valores similares aos modelos univariados. Sendo assim, com intuito de selecionar o melhor modelo para cada abordagem, algumas combinações de escolhas de κ e X foram utilizadas. Então em cada caso o vetor de parâmetros θ foi estimado por máxima verossimilhança, onde foram utilizadas as mesmas tecnicalidades dos modelos univariados,

sendo que no BGCCM utilizamos 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 as as reparametrizoções $\sigma = \sigma_1$, $\nu_1 = \sigma_{12}/\sigma_{11}$ e $\nu_2 = \sigma_{22}/\sigma_{11}$, sendo assim, nos dois casos existe forma analítica para os estimadores de β^* e σ^2, que são função dos demais parâmetros, para estimar os demais parâmetros foi utilizado o métodos de Nelder e Mead (1965), que maximiza a função de máxima verossimilhança concentrada. E novamente, utilizando a propriedade de invariância dos estimadores foi possível encontrar estimativas para todos os parâmetros dos modelos. Quanto a seleção de modelos, a krigagem e os testes de pressupostos, as mesmas tecnicalidades dos modelos univariados foram adotadas.

A abordagem de modelos geoestatísticos bivariados é uma possibilidade para esse conjunto de dados, uma vez que, os dois atributos químicos em questão são altamente correlacionados e, além disso, a coleta da informação sobre a saturação por bases é mais dispendiosa, e como a fazenda continuará sendo monitorada periodicamente, é interessante estruturar a correlação espacial entre os dois atributos, pois nas próximas análises de solo é possível amostrar menos pontos georeferenciados para observar a saturação por bases e utilizar a informação do pH para fazer inferências. Logo, para exemplificar tal tecnicalidade, por último, retiramos do conjunto de dados 20 observações da saturação por bases e conduzimos novamente a modelagem univariada desse atributo e as modelagens bivariadas utilizando a informaçào completa do pH. Sendo que os parâmetros estimados foram utilizados para fazer as krigagens da saturação por bases nas localizações espaciais omitidas. Então, temos os valores observados e não utilizados na modelagem, suas predições utilizando somente os demais valores da saturação por bases e suas predições utilizando as duas abordagens de modelos bivariados, os quais apresentaram krigagens mais precisas.

4 RESULTADOS E DISCUSSÃO

4.1 Estudo de simulação

Tabela 1 - Médias, medianas e desvios padrões do pH e da saturação por bases

Variável	Mediana	Média	D.P.
Saturação	56,00	53,27	10,05
pH	5,00	4,94	0,42

4.2 Estudo observacional sobre qualidade do solo

Com o intuito de apresentar as tecnicalidades da geoestatística aplicadas à agricultura de precisão, esse tópico da pesquisa traz uma análise geoestatística detalhada do conjunto de dados provinientes da fazenda Tupã.

4.2.1 Análise exploratória

A Figura 1 mostra os gráficos de círculos das variáveis sob estudo, onde os círculos representam os valores mensurados dos atributos nas 67 localização amostrada da propriedade agrícola, ou seja, quanto maior o círculo mais elevado é o valor observado do atributo. Sendo assim, essa primeira análise gráfica nos permite suspeitar que existe uma 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 esquerda, respectivamente, sendo que, a primeira possui valores mais elevados, o que nos leva a possibilidade de considerar 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. No entanto, antes de dar continuidade a análise espacial desses dados, iremos mostrar uma breve análise descritiva.

A Tabela 1 mostra as estatísticas descritivas do pH e da saturação por bases, 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, estão bem comportados, onde há uma simetria moderada dos dados com relação a média amostral e não existem valores observados muito discrepantes.

Como uma análise exploratória espacial dos dados, analisamos alguns gráfi-

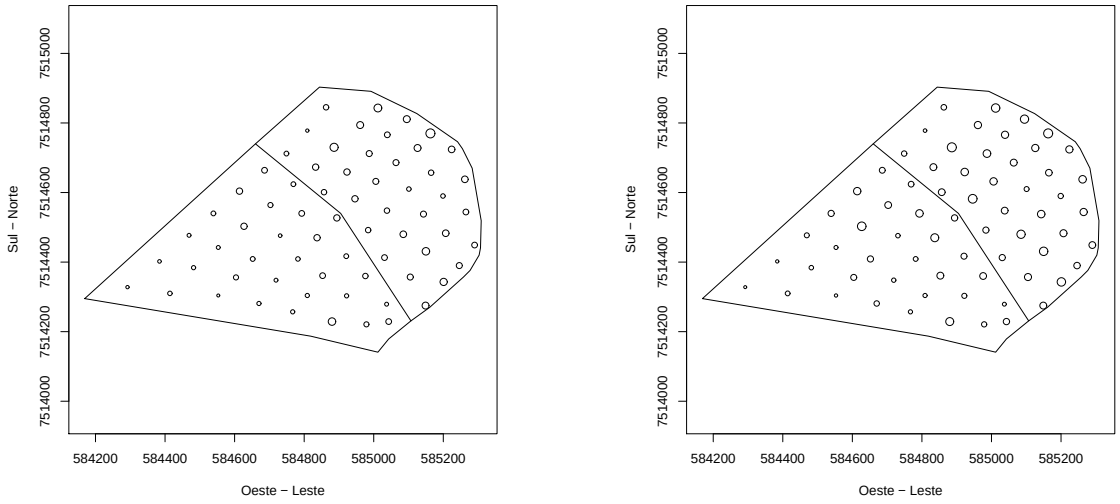

Figura 1 - Gráficos de círculos - o gráfico a esquerda é para a variável pH e o da direita é da saturação por bases

cos, sendo que os gráficos dos cantos superiores esquerdos das Figuras 2 e 3 categorizam os dados nos quartis amostrais das observações, onde quanto mais fria a cor menor o quartil amostral, essas imagens corroboram a idéia que de existe padrão espacial nos dados, uma vez que, existem conglomerados das categorias, já os gráficos dos cantos inferiores esquerdos mostram que quanto maior o valor da coordenada x, oeste-leste, maior os valores observados dos atributos, logo, iremos considerar essas coordenadas como possível covariável no processo de modelagem. Além disso, os gráficos dos cantos inferiores direitos mostram as densidades amostrais dos dados desconsiderando o possível padrão espacial, onde existe uma pequena fulga na distribuição gaussiana. Sendo assim, conduzimos análises de possíveis transformações de variáveis, a Figura 4 mostra quais transformações de Box-Cox maximizam a probabilidade dos dados serem provinientes da distribuição normal, logo, como os intervalos de confiança para λ contemplam o valor 1, não transformamos as variáveis, logo, aceitamos a suposição de campos aleatórios gaussianos homogêneos.

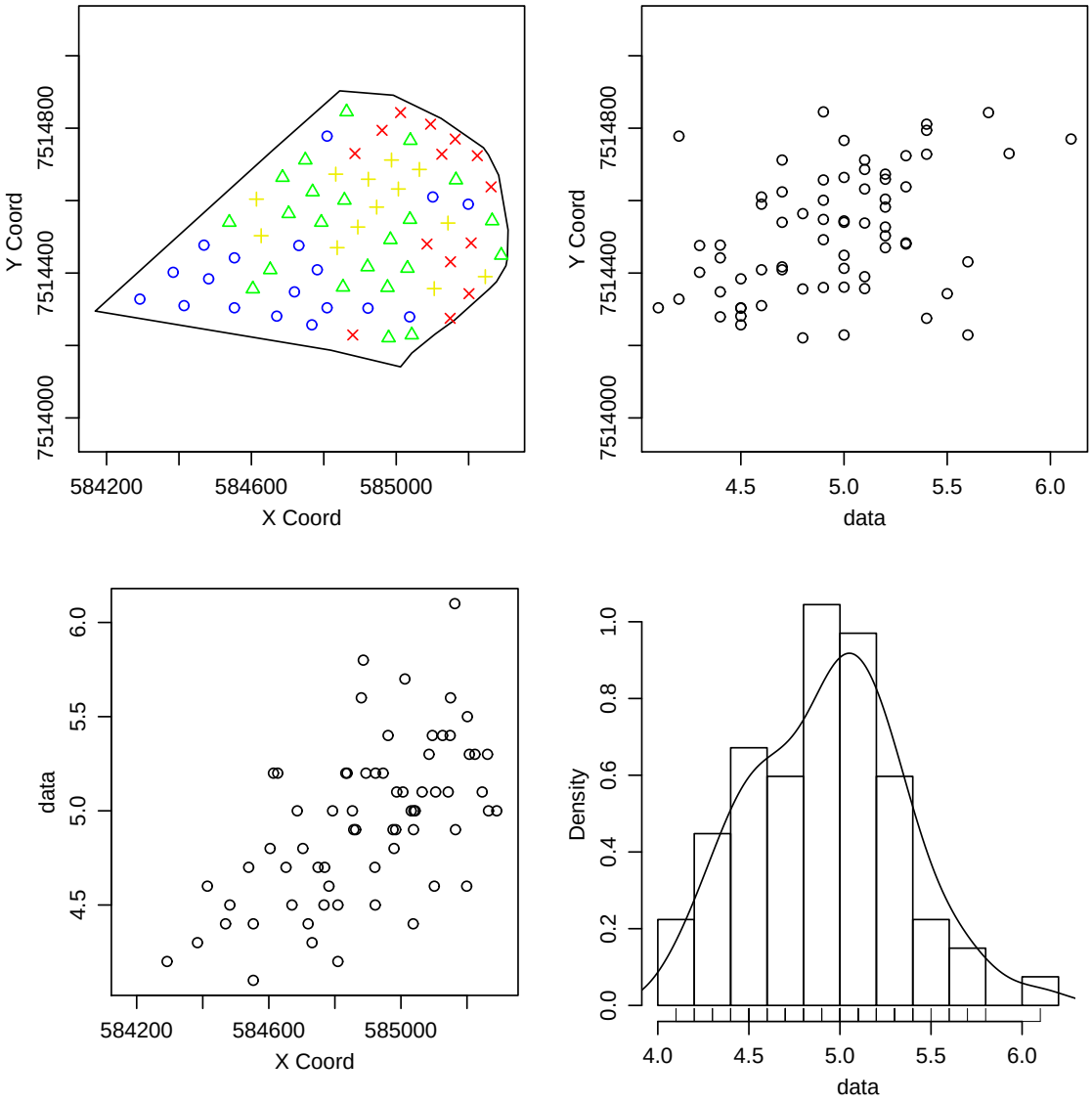

Figura 2 - Gráficos descritivos do padrão espacial do pH

4.2.2 Modelagem univariada do pH

As Tabelas 2, 3 e 4 mostram as estimativas de máxima verossimilhança para os parâmetros dos modelos univariados ajustados para o pH considerando as três possíveis formas para a matriz de delineamento X: média constantes, média com tendência induzida pela área de manejo e média induzida pela coordenada oeste-leste. Utilizando cada abordagem para a média, as diferentes escolhas para κ não produzem muita diferença nas estimativas

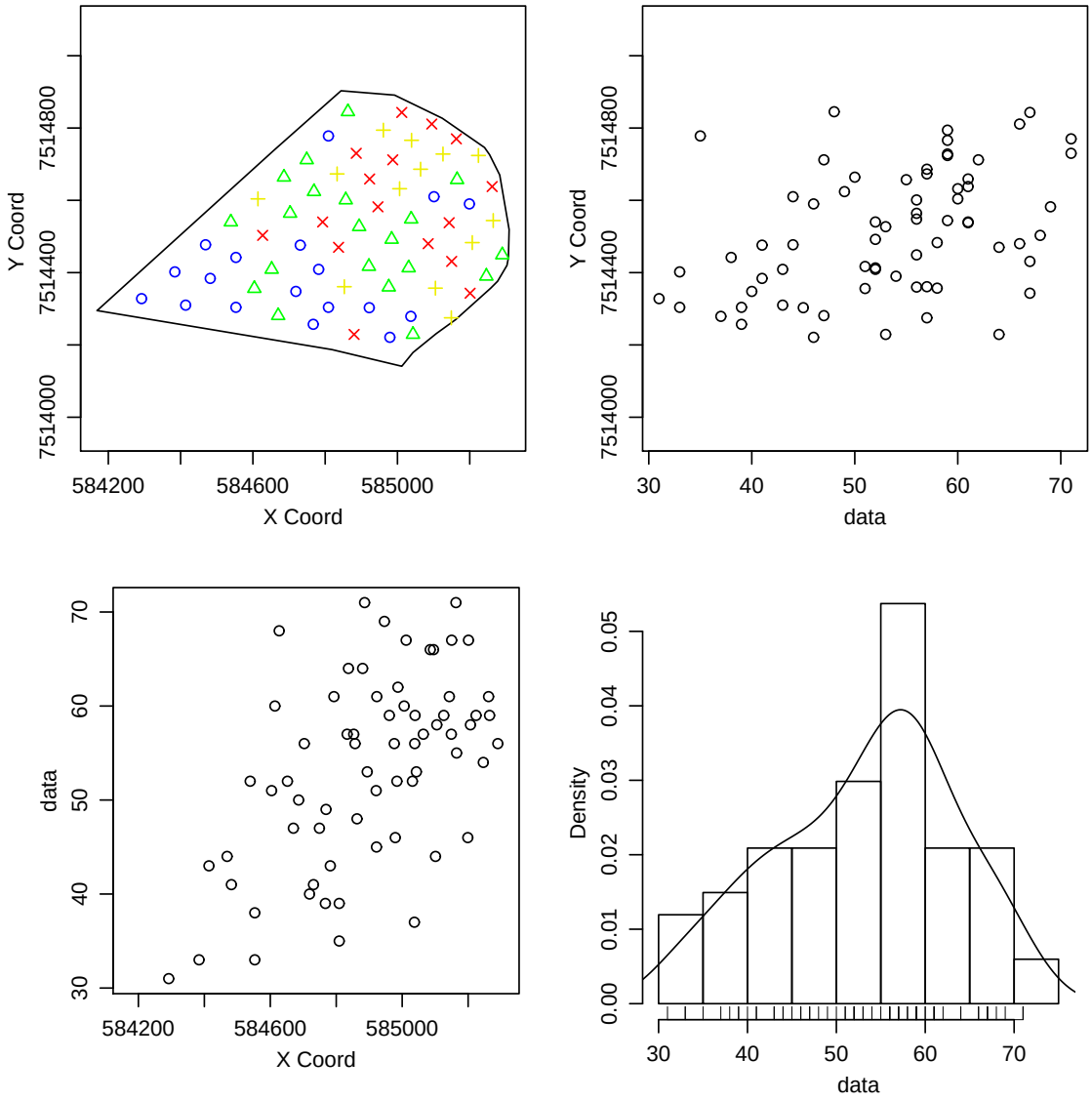

Figura 3 - Gráficos descritivos do padrão espacial da saturação por bases

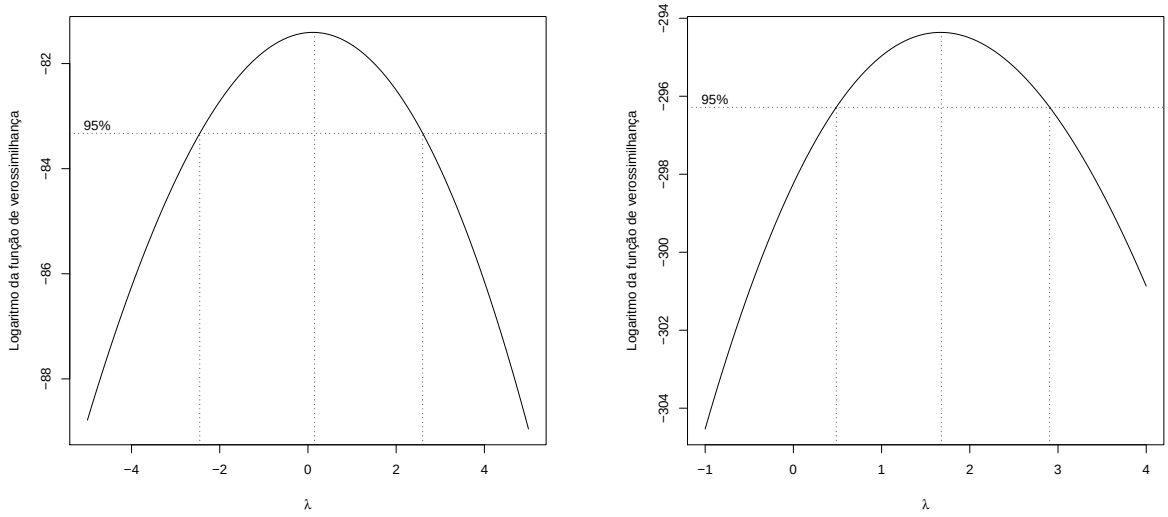

Figura 4 - Gráficos de possíveis transformações de variáveis de Box-Cox, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

dos parâmetros de média e de variabilidade, no entanto o parâmetro ϕ é mais afetado, o que é normal, uma vez que, os parâmetros provinientes da função de correlação Mátern não são ortogonais, fato que gera alcances práticos das correlações parecidos para essas estimativas. Com a inclusão de covariáveis no modelo muda um pouco a relação entre σ^2 e τ^2, onde os dois valores ficaram bem próximos, além disso, a estimativa ϕ indica que o alcance das correlações é menor.

Analisando os valores estimados dos máximos do logaritmo da função de verossimilhança e os seus respectivos AIC, foi selecionado o modelo final, que é o com tendência induzida pela coordenada oeste-leste e $\kappa = 2,5$, uma vez que, esse ajuste gerou o maior máximo da função alvo e o menor valor do critério de penalidade por número de parâmetros associado ao modelo.

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, o padrão dos dados é completamente aleatório, sendo que, sobre esse enfoque o AIC ficou igual a 49,8, o que afirma a idéia de que o pH possui variabilidade espacial significativa.

Tabela 2 - Estimativas de máxima verossimilhança dos parâmetros associados ao modelo para o pH com média constantes

β	τ^2	σ^2	ϕ	κ	$l(\theta)$	AIC
$4,907$	$0,077$	$0,163$	560	$0,5$	$-25,49$	$58,98$
$4,905$	$0,101$	$0,149$	$510,6$	1	$-25,69$	$59,38$
$4,904$	$0,106$	$0,251$	$650,5$	$1,5$	$-25,57$	$59,15$
$4,903$	$0,106$	$0,288$	$569,7$	2	$-25,45$	$58,90$
$4,903$	$0,106$	$0,297$	$491,1$	$2,5$	$-25,36$	$58,73$

Tabela 3 - Estimativas de máxima verossimilhança dos parâmetros associados ao modelo para o pH com tendência na média induzida pela área de manejo

β_0	β_1	τ^2	σ^2	ϕ	κ	$l(\theta)$	AIC
$4,725$	$0,397$	0	$0,125$	$49,19$	$0,5$	$-23,44$	$56,89$
$4,723$	$0,402$	0	$0,125$	$33,42$	1	$-23,35$	$56,70$
$4,723$	$0,403$	$0,014$	$0,111$	$28,43$	$1,5$	$-23,33$	$56,66$
$4,779$	$0,227$	$0,110$	$0,061$	$349,3$	2	$-24,86$	$59,73$
$4,786$	$0,204$	$0,109$	$0,086$	$353,1$	$2,5$	$-24,83$	$59,65$

Tabela 4 - Estimativas de máxima verossimilhança dos parâmetros associados ao modelo para o pH com tendência na média induzida pelas coordenadas oeste-leste

β_0	β_1	τ^2	σ^2	ϕ	κ	$l(\theta)$	AIC
$-607,48$	$0,001$	0	$0,111$	$49,32$	$0,5$	$-19,66$	$49,31$
$-608,51$	$0,001$	0	$0,112$	$34,28$	1	$-19,44$	$48,89$
$-608,33$	$0,001$	0	$0,112$	$27,58$	$1,5$	$-19,39$	$48,77$
$-608,36$	$0,001$	$0,008$	$0,104$	$24,53$	2	$-19,36$	$48,73$
$-608,46$	$0,001$	$0,015$	$0,097$	$22,46$	$2,5$	$-19,35$	$48,70$

Tabela 5 - Estimativas de máxima verossimilhança dos parâmetros associados ao modelo para a saturação por bases com média constantes

β	τ^2	σ^2	ϕ	κ	$l(\theta)$	AIC
$49,53$	$47,66$	$105,33$	$700,00$	$0,5$	$-239,53$	$487,05$
$48,53$	$59,57$	$120,62$	$625,58$	1	$-239,76$	$487,51$
$47,98$	$62,40$	$124,31$	$516,78$	$1,5$	$-239,79$	$487,57$
$47,27$	$63,29$	$149,69$	$489,58$	2	$-239,75$	$487,50$
$47,18$	$63,51$	$139,18$	$400,00$	$2,5$	$-239,70$	$487,40$

Tabela 6 - Estimativas de máxima verossimilhança dos parâmetros associados ao modelo para a saturação por bases com tendência na média induzida pela área de manejo

β_0	β_1	τ^2	σ^2	ϕ	κ	$l(\theta)$	AIC
$47,91$	$8,80$	$24,35$	$53,68$	$73,68$	$0,5$	$-238,05$	$486,10$
$47,92$	$8,83$	$38,78$	$39,28$	$59,18$	1	$-238,01$	$486,01$
$47,92$	$8,86$	$43,35$	$34,69$	$50,33$	$1,5$	$-237,98$	$485,96$
$47,92$	$8,88$	$45,53$	$32,47$	$44,34$	2	$-237,96$	$485,92$
$47,93$	$8,89$	$46,79$	$31,18$	$39,98$	$2,5$	$-237,95$	$485,90$

4.2.3 Modelagem univariada da saturação por bases

Para essa variável foram consideradas as mesmas tecnicalidade aplicadas ao pH, as Tabelas 5, 6 e 7 mostram os resultados, sendo que, novamente o melhor modelo foi o que considera tendência na média induzida pela coordenada oeste-leste e $\kappa = 2,5$. Além disso, novamente foi considerado um modelo de regressão linear simples com a coordenada oeste-leste como covariável, o qual gerou um AIC de 480,2, sendo assim, para essa variável também existe padrão espacial de variabilidade significativo.

Tabela 7 - Estimativas de máxima verossimilhança dos parâmetros associados ao modelo para a saturação por bases com tendência na média induzida pelas coordenadas oeste-leste

β_0	β_1	τ^2	σ^2	ϕ	κ	$l(\theta)$	AIC
$-14297,01$	$0,025$	$8,98$	$59,98$	$54,51$	$0,5$	$-234,75$	$479,50$
$-14300,66$	$0,025$	$27,56$	$41,46$	$45,42$	1	$-234,72$	$479,44$
$-14300,92$	$0,025$	$33,53$	$35,53$	$39,51$	$1,5$	$-234,70$	$479,39$
$-14300,17$	$0,025$	$36,39$	$32,68$	$35,33$	2	$-234,68$	$479,36$
$-14298,98$	$0,025$	$38,06$	$31,03$	$32,19$	$2,5$	$-234,67$	$479,34$

4.2.4 Krigagens

Com os modelos finais estabelecidos, foram calculadas as predições espaciais para as duas variáveis químicas do solo em uma malha de 10000 localizações espaciais, sendo que, as estimativas paramétricas foram plugadas nas fórmulas de krigagem e variância preditiva. A Figura 5 mostra os mapas preditivos dos campos aleatórios, onde, 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 relativamenete pequenos com relação a grandeza de escala das variáveis, sendo assim, a precisão das krigagens ficaram em um patamar aceitável.

4.2.5 Análise de resíduos

A Figura 6 mostra que o comportamento dos resíduos do modelo ajustado para o pH aparentemente foge um pouco da suposição de gaussianiedade, no entanto, o teste de normalidade de Shapiro-Wilk resultou em um p-valor de 0,22, ou seja, não existe evidências suficientes contra a suposição feita.

A Figura 7 mostra que para a saturação por bases os resíduos do modelos final estão bem próximos da suposição imposta, característica corroborada pelo teste de

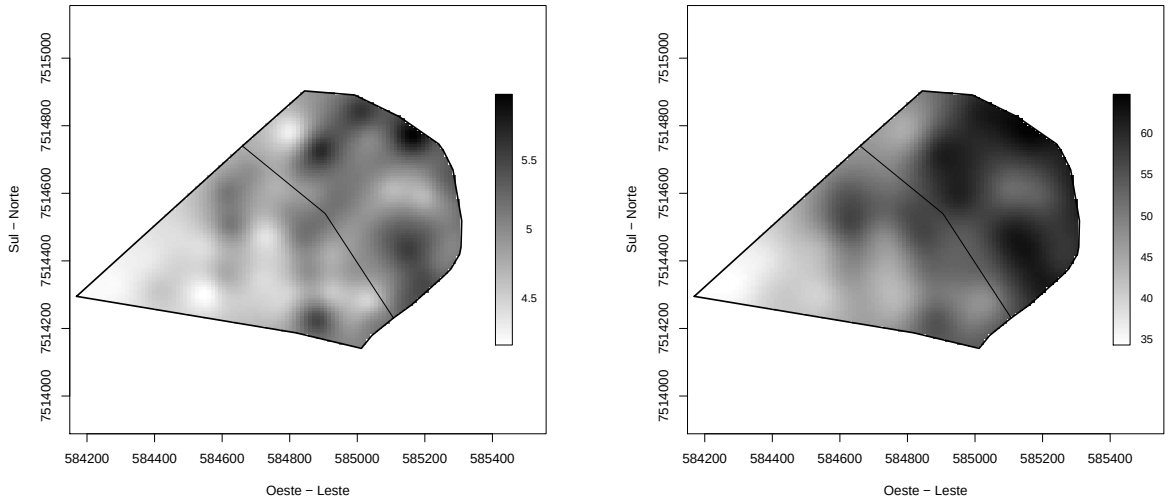

Figura 5 - Gráficos de predições espacias, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

normalidade de Shapiro-Wilk, que gerou um p-valor de 0,97.

4.2.6 Abordagem bivariada para o pH e a saturação por bases

Com a modelagem univariada concluída, o próximo passo foi ajustar modelos bivariados para as respostas, em caratér descritivo foi calculada a correlação de Pearson entre os atributos, a estatística resultante ficou em 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, utilizamos as estimativas dos modelos univariados para atribuir valores iniciais ao método numérico de maximização. Sendo que 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, as quais combinadas com diversas escolhas para os parâmetros de suavidade das correlações geraram diversas estimativas diferentes para os parâmetros dos modelos. Sendo assim, inicialmente, selecionamos as melhores estimativas em cada abordagem para a média.

A Tabela 8 mostra os resultados do BGCCM, onde não há muita diferença entre

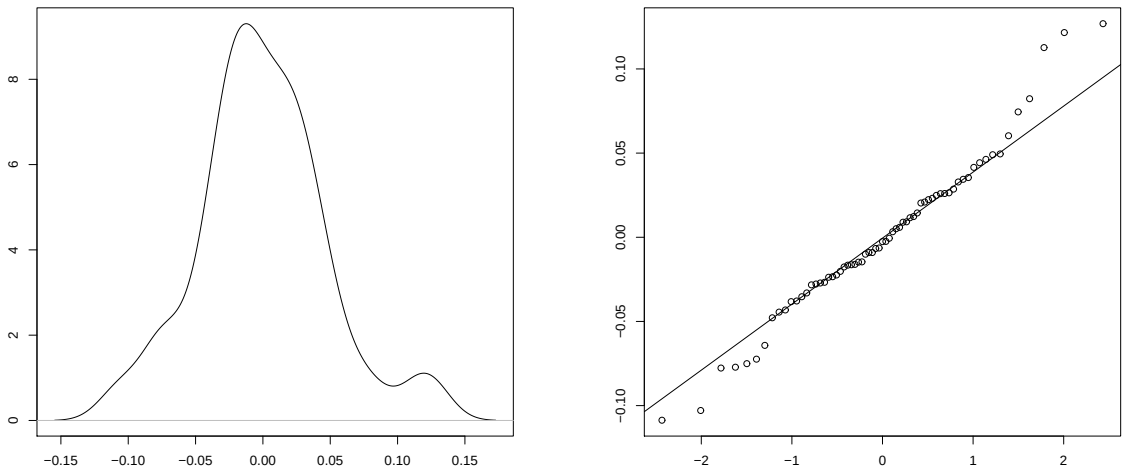

Figura 6 - Densidade e gráfico de quartis dos resdíduos para o modelo final do pH

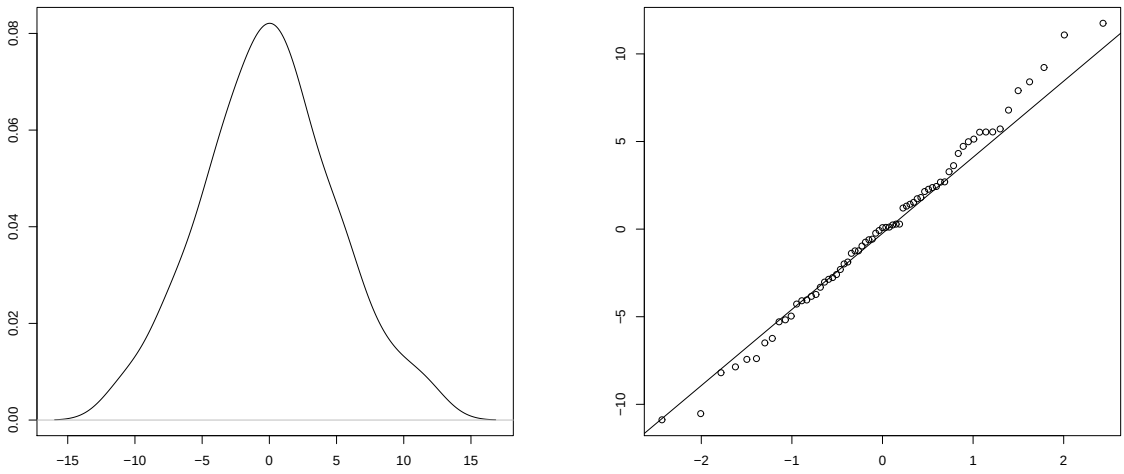

Figura 7 - Densidade e gráfico de quartis dos resdíduos para o modelo final da saturação por bases

os parâmetros de variabilidade e correlação estimados para cada abordagem de tendência, a não ser pelo parâmetros ϕ_0, que para a tendência na área de manejo ficou bem menor que nas outras duas, fato explicado pela adoção de um κ_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 9 mostra os resultados da estimação de máxima verossimilhança utilizando o BCRM. Para essa modelagem bivariada, com a inclusão da covariável coordenada leste-oeste ocorreu inflacionamento dos parâmetros de variabilidade. Com relação as demais estimativas, os resultados são parecidos com os do BGCCM.

Nas duas abordagens, o melhor ajuste foi o que considera a área de manejo da propriedade como covariável, os quais geraram AIC igual a 437,35 e 432,96 para o BGCCM e BCRM, respectivamente. Sendo assim, as estimativas finais foram estabelecidas e podemos pensar na krigagem ordinária, para tal, foi utilizado o mesmo grid de pontos utilizados na modelagem univariada.

O processo de krigagem utilizando as estimativas do BGCCM gerou variâncias preditivas relativamente pequenas, sendo que, 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 (8) e (9) ilustram os mapas preditivos utilizando os valores estimados em cada abordagem de modelos bivariados. Não existe muita diferença entre as predições utilizando modelos univariados ou bivariados, no entanto, como nas abordagens mais complexas foi considerada a covariável área de manejo nos modelos finais, o mapa preditivo tem um salto nos valores, quando passamos de uma região de manejo para outra.

Sendo assim, tem-se que os modelos geoestatísticos bivariados estão ajustando bem os dados observacionais, sendo que o BCRM se mostrou mais eficiente para fazer estimação, onde em todos os casos o AIC é menor com essa abordagem, já do ponto de vista preditivo, o BGCCM apresentou uma menor variabilidade preditiva. Finalmente, não existe muita diferença entre os dois modelos tratados nesse estudo, os quais forneceram bons resultados preditivos para o pH e a saturação por bases da fazenda Tupã.

Tabela 8 - Estimativas de máxima verossimilhança dos parâmetros associados ao BGCCM para a saturação por bases e o pH

θ	Constantes	Oeste-leste	Área de manejo
β_{01}	$52,57$	$1,54e-6$	$47,83$
β_{02}	$4,92$	$1,45e-7$	$4,72$
β_1	$-$	$8,98e-5$	$9,46$
β_2	$-$	$8,41e-6$	$0,40$
σ_{01}	$8,51$	$8,35$	$6,85$
σ_1	$3,47$	$3,46$	$3,44$
σ_{02}	$0,38$	$0,38$	$0,31$
σ_2	$7e-5$	$5e-4$	$1e-5$
ϕ_0	$59,37$	$57,79$	$26,73$
ϕ_1	$46,35$	$48,15$	$49,08$
ϕ_2	$66,82$	$71,59$	$86,95$
κ_0	1	1	$1,5$
κ_1	$0,5$	$0,5$	$0,5$
κ_2	$0,5$	$0,5$	$0,5$

Tabela 9 - Estimativas de máxima verossimilhança dos parâmetros associados ao BCRM para a saturação por bases e o pH

θ	Constantes	Oeste-leste	Área de Manejo
β_{01}	$51,76$	$1,51e-6$	$47,93$
β_{02}	$4,89$	$1,43e-7$	$4,72$
β_1	$-$	$8,85e-5$	$9,12$
β_2	$-$	$8,36e-6$	$0,40$
σ_{11}	$8,26$	$99,53$	$7,82$
σ_{12}	$0,31$	$3,67$	$0,28$
σ_{22}	$0,13$	$1,54$	$0,14$
ϕ_1	$100,63$	$100,21$	$53,32$
ϕ_2	$21,43$	$21,38$	$19,34$
κ_1	$0,5$	$0,5$	$0,5$
κ_2	$2,5$	$2,5$	$2,5$

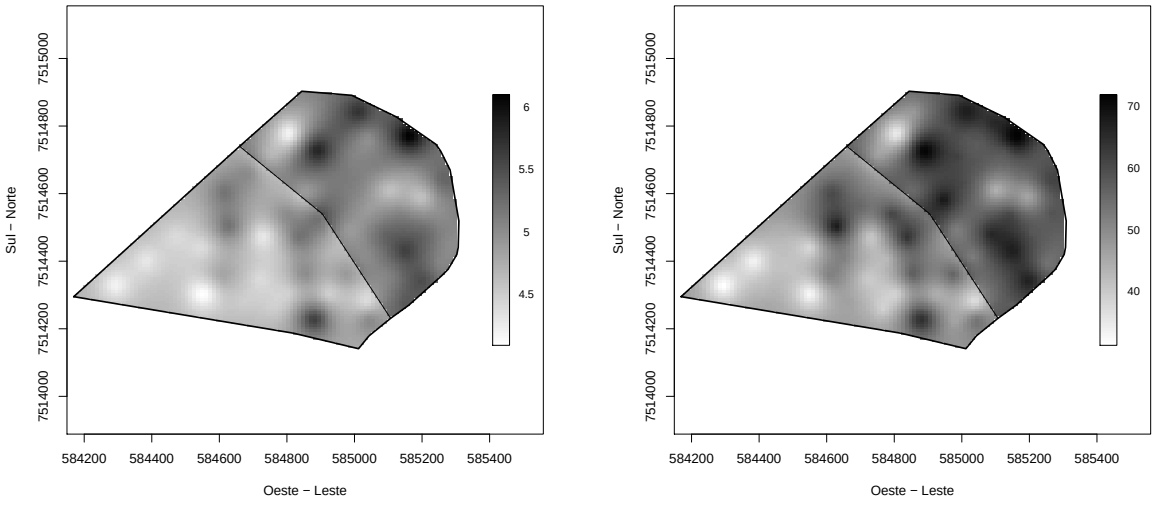

Figura 8 - Gráficos de predições espacias com o BGCCM, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

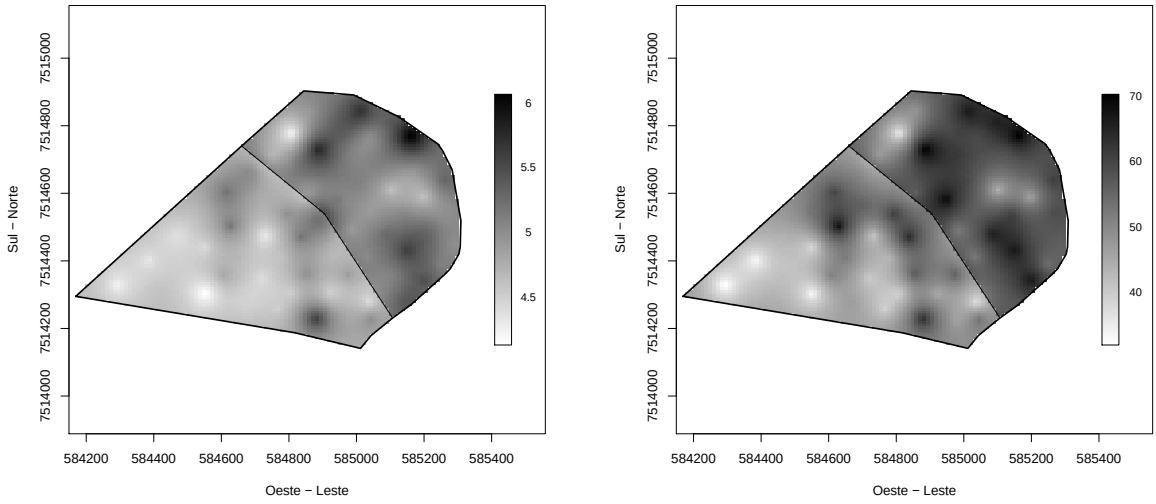

Figura 9 - Gráficos de predições espacias com o BCRM, o gráfico da esquerda é relativo ao pH e o gráfico da direita é relativo à saturação por bases

4.2.7 Estudo empírico de comparação entre as abordagens univariada e bivariadas

Além da fundamentação estatística para considerar modelos bivariados para esse problema, existe a justificativa prática para adoção de tal abordagem, sendo que, 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 amostrada um número menor de localizações espaciais para observação da saturação por bases, e mesmo assim é viável ter boas análises para esse atributo, uma vez que, pode-se utilizar a informação contida no pH para fazer inferências sobre a saturação por bases.

Para ilustrar essa tecnicalidade, foram omitidas do conjunto de dados 20 localizações espaciais da saturação por bases. Estimando novamente, em todas as abordagens, os parâmetros dos modelos finais selecionados, tem-se que não existe muita diferença entre essas estimativas e as com informação completa da saturação por bases. O próximo passo foi fazer a krigagem ordinária da saturação por bases nas 20 localizações retiradas do conjunto de dados, para tal, utilizamos os parâmetros estimados com o modelo univariado, com o

Tabela 10 - Médias e desvios padrões dos erros de krigagem da saturação por bases nas 20 localizações omitidas no processo de estimação

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

BGCCM e com o BCRM.

A Tabela 10 mostra os resultados descritivos das diferenças entre os verdadeiros valores observados e os valores preditos nas três abordagens de modelagem, sendo que, 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, realmente é melhor utilizar a informaçàò do pH para fazer as inferências sobre a saturação por bases.