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

Modelos geoestatísticos multivariados

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 multivariados

Orientador:
Prof. Dr. **Paulo Justiniano Ribeiro Jr.**

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

Piracicaba
2008

Dedicatória

AGRADECIMENTOS

SUMÁRIO

RESUMO

Modelos geoestatísticos multivariados

Palavras-chaves:

ABSTRACT

Multivariate geostatistic models

Keywords:

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 dado 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 devem 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 e cópulas.

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, Gelfand et al. (2005) proporam a utilização de modelos hierarquicos para o problema, Diggle e Ribeiro (2006) proporam uma abordagem utilizando a distribuição conjunta das observações. No entanto, devido a complexidade dos modelos, número elevado de parâmetros, existem 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 com essas duas abordagens, dessa forma pode-se fazer uma comparação entre as duas metodologias, e detectar quais são as vantagens, probabilísticas e computacionais, de cada método em diversas configurações paramétricas e utilizando diversas técnicas de estimação frequentistas e bayesianas. Além do estudo dos modelos exitentes na literatura, será apresentado uma adaptação modelo de Diggle e Ribeiro (2006), que originalmente não consegue capturar estruturas de correlação negativa entre as respostas.

Por último, serão apresentados resultados e análises com dados observacionais, utilizando as diferentes abordagens para os problemas geoestatísticos multivariados.

Cabe ressaltar que todas as análises foram conduzidas utilizando o ambiente R de programação e os pacotes utilizados são geoR, MASS.

2 REVISÃO DE LITERATURA

Nesta seção serão apresentadas algumas técnicas e conceitos de geoestatística, modelagem geoestatística multivariados e de técnicas de estimação.

2.1 Campos Aleatórios

Um campo aleatório ou função aleatória é algum atributo sob estudo que existe em algum espaço real d-dimensional, geralmente bi ou tri-dimensional, os valores reais da função aleatória não são conhecidos, o que caracteriza um processo latente, sendo assim, é necessário fazer uma amostragem de pontos espaciais e nas localizações amostradas o atributo de interesse será medido, com os valores observados pode-se propor algum modelo ao problema e fazer predições aos pontos não observados, abaixo segue a notação de campos aleatórios:

$$\{Z(s) : s \in G \subset R^d\}, \tag{1}$$

sendo $Z(s)$ a notação para o campo aleatório na localização s 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 pontos nas localizações s_1,s_2,...,s_n pertencentes à região G e qualquer inteiro n:

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

Uma das distribuições de probabilidades mais utilizadas na literatura é a gaussiana, que devido as suas propriedades, é relativamente, a mais fácil de estabelecer distribuições conjuntas e fazer inferências. Sendo assim, um processo espacial, como em (1) é dito ser gaussiano se $Z(s)$ segue uma distribuição Normal n-variada, logo, devido às propriedades da distribuição em questão, a cada atributo $Z(s_i)$ é associada uma distribuição normal univariada.

Considerando que um processo gaussiano segue uma distribuição normal n-variada, ele é completamente especificado pelo vetor de médias e pela matriz de variâncias e covariâncias. O vetor de médias é especificado pela presença ou ausência de covariáveis ao processo, se for considerado que não existe nenhuma tendência sobre a média do processo é dito que o processo gaussiano possui média constante, e o vetor de médias possui n valores iguais, por outro lado,

se há evidências de que existe alguma tendência na média do processo devida à presença de covariáveis, é necessário propor algum modelo estatístico para capturar essa tendência, o que aumenta o número de parâmetros a estimar. A matriz de variâncias e covariâncias deve ser positiva definida, o que não é de fácil elaboração, por conta disso, algumas funções de correlação já conhecidas, que produzem matrizes positiva definidas, são muito utilizadas na prática.

Em uma pesquisa de geoestatística, geralmente, não é possível ter mais de uma realização do processo devido aos custos envolvidos ou outros problemas,sendo assim, outras suposições devem ser impostas para que seja possível a realização de inferências. Na literatura e na prática a restrição mais utilizada é que o processo é estacionário, ou seja, a distribuição da função aleatória não depende da grandeza de escala das coordenadas, sendo assim, a distribuição conjunta dos $(Z(s_1), Z(s_2), ..., Z(s_n))$ 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 uma função aleatória $Z(s)$ é dita ter estacionariedade fraca se $E[Z(s)] = \mu$ e $Cov[Z(s), Z(s + h)] = C(h)$.

Com as definições acima, se tem que a média é igual em toda a região sob estudo e que a covariância entre atributos em locais diferentes só depende da distância entre os mesmos. 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 $E[Z(s)]=\mu$ e $Var[Z(s) - Z(s + h)] = 2\gamma(h)$, em que $\gamma(h)$ é denominado semivariograma, e a relação $\gamma = C(0) - C(h)$ é válida.

Por conta da relação entre as covariâncias e o semivariograma, a variabilidade de campos aleatórios intrinsecamente estacionários pode ser estudada por qualquer uma das medidas da relação, comumente na literatura de geoestatística utiliza-se o semivariograma.

Outra abordagem que pode ser adotada é quando o processo não possui nenhum tipo de estacionariedade, ou seja, ou a média varia ao longo da região sob estudo ou a variância

não é constante. Como dito anteriormente, quando a média não apresenta constância em toda a região sob estudo, algum modelo pode ser proposto para capturar essa variação, geralmente adota-se modelos lineares com as coordenadas como covariáveis. Com relação a variância e covariâncias não constantes no campo aleatório, pode-se tentar uma abordagem mais simples com uma transformação nos dados originais, geralmente a família de transformações de Box-Cox é utilizada. No entanto, devido a complexidade do problema, pode-se adotar outras abordagens como modelos de deformações espaciais (SAMPSON; GUTTORP, 1992, SCHMIDT; O'HAGAN, 2003) ou convoluções espaciais (HIGDON, 2002, FUENTES; SMITH, 2001). Uma outra característica que pode surgir em eventos da natureza é que o processo possui algum tipo de estacionariedade, mas mesmo assim a função de covariância depende da direção, assim a função aleatória será considerada anisotrópica, ou seja, a variabilidade é constante em todo o campo, porém há diferenças nas correlações conforme a direção em que a distância está, esse tipo de problema é muito comum em estudos de poluentes na atmosfera, onde a direção dos ventos gera uma distorção na correlação entre pontos com a mesma distância. Quando existe anisotropia nos dados, basta incluir mais parâmetros na estrutura de correlação, não há dificuldade em modelar esse problema, porém a identificação de tal padrão a partir dos dados não é fácil. A forma mais comum de tratar a anisotropia é fazer transformações nos sistemas de coordenadas, utilizando geometria para tal. Na literatura geoestatística quando um processo estacionário possui anisotropia ele é chamado de processo estacionário heterogêneo e caso contrário processo estacionário homogêneo.

2.1.1 Propriedades da função de covariância

Encontrar funções que estabeleçam a estrutura de covariância, que seja válida, não é trivial, ou seja, não é fácil achar funções de correlação que possuam um comportamento empírico, onde se espera que quanto maior a distância entre os pontos menor a correlação entre os atributos e que produzam uma matriz de covariâncias positiva definida, logo, na literatura existem algumas famílias de funções de covariâncias que são conhecidamente válidas. No entanto antes de apresentar tais famílias de funções válidas, seguem as propriedades das funções de covariância de um processo estacionário de segunda ordem:

(i) $Cov[Z(s), Z(s+0)] = Var[Z(s)] = C(0) \geq 0$;

(ii) $C(h) = C(-h)$;

(iii) $C(0) \geq \mid C(h) \mid$;

(iv) $C(h) = Cov[Z(s), Z(s+h)] = Cov[Z(0), Z(h)]$;

(v) Se C_j são funções de covariância válidas, então $\sum_j b_j \, C_j(h)$ e $\prod_j C_j(h)$ são funções válidas;

(vi) Se $C(h)$ é válida para um espaço d-dimensional, então $C(h)$ é válida para todo espaço menor que d.

A propriedade i é de interpretação trivial, onde a covariância entre um atributo medido na mesma localização espacial é simplesmente a variância do campo aleatório. A propriedade ii assegura que a covariância é igual para incrementos positivos ou negativos, no entanto em geoestatística essa característica não é muito importante, uma vez que, geralmente utiliza-se a distância euclidiana entre as localizações, logo, o incremento h sempre é positivo. A propriedade iii garante que a variância do campo aleatório sempre será maior ou igual do que as covariâncias para qualquer incremento, essa característica é uma restrição necessária para que a matriz de covariância de um campo aleatório gaussiano seja positiva definida. A propriedade iv assegura que as covariâncias só dependem dos incrementos h independentemente da grandeza das localizações. E por último, a propriedade v é importante para modelos multivariados, onde é comumente utilizadas somas de matrizes de covariâncias válidas.

2.1.2 Famílias de funções de covariância

Na prática não é fácil encontrar uma matriz de covariância válida para um campo aleatório gaussiano, dessa forma, na literatura de geoestatística existem diversas funções de correlação que só dependem dos incrementos h e que asseguram a validade das matrizes de covariância. Seguem as funções mais utilizadas recentemente na literatura:

A Família Matérn

Essa família de correlações foi proposta por Berfil Matérn (1986) e possui a seguinte função:

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

os parâmetros dessa função são $\phi > 0$ e $\kappa > 0$, que são vinculados a escala com a dimensão de distância e suavidade do processo e $K_{\kappa}(.)$ é a função Bessel de ordem κ.

A Família Exponencial Potência

$$C(h) = \exp\left(h/\phi\right)^{\kappa},$$

essa família também possui dois parâmetros, com mesmas interpretações da família Matérn, no entanto agora κ limitado no intervalo $[0, 2]$. Cabe ressaltar que a família Matérn com κ=1/2 é igual a função exponencial com κ=1.

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.

2.1.3 Modelos geoestatísticos gaussianos

Considerando que em alguma área de interese G exista um campo aleatório gaussiano Z contínuo em toda a área. No entanto, esse campo aleatório é latente, ou seja, o processo existe mas nào é observável, sendo assim, é necessário fazer uma amostragem de n localizações espaciais dentro da área G e observar valores do atributo de interesse nas posições amostradas. Logo, existe um vetor $Y(s)$ $nx1$, para $s = (s_1, s_2, ..., s_n)$, que é modelado da seguinte forma:

$$Y(s) = Z(s) + \epsilon(s),$$

sendo $Z(s)$ um campo aleatório gaussiano que possui vetor de médias μ $nx1$ e matriz de covariâncias Σ nxn que possui parâmetros provenientes das variâncias e das correlações, e $\epsilon(s)$ um vetor $nx1$ de ruídos brancos, onde esses ruídos são, por suposição, independentes e identicamente distribuidos de uma normal com média zero e variância τ.

Com a modelagem acima chega-se a distribuição de probabilidade conjunta dos valores observados, onde $Y(s)|Z(s)$ segue uma distribuição gaussiana n-variada, com vetor de médias μ $nx1$ e matriz de covariâncias $\Sigma_Y(s)$ nxn que é igual à $\Sigma + \tau I$, onde I é uma matriz identidade nxn.

2.1.4 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:

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

em que $|N(h)|$ é o número de pontos abrangidos pela distância h. Se o processo for intrinsecamente estacionário o estimador de (2) é não viesado para γ. Já se o campo aleatório possui estacionariedade fraca o estimador do semivariograma informa sobre a função de correlação do processo, ou seja, a relação $\gamma = C(0) - C(h)$ estabelece a relação entre a função de correlação e o semivariograma.

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 da função de correlação que se ajuste aos valores do semivariogramas empíricos, isso pode ser feito através do método de mínimos quadrados ponderados ou por meio de métodos "AD-HOC", no entanto, essas duas abordagens para estimar os parâmetros 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.

Por outro lado, assumindo que o campo aleatório possui estacionariedade forte, pode-se optar por estimadores de Máxima verossimilhança ou Máxima Verossimilhança Restrita, que se baseiam no logaritmo da função de verossimilhança abaixo:

$$l(\theta; Z_1, Z_2, ..., Z_n) = -(\ln(|\Sigma(\theta)|) + n\ln(2\pi) + (Z(s) - 1\mu)^t \Sigma(\theta)^{-1}(Z(s) - 1\mu))/2. \tag{3}$$

A função de verossimilhança em (3) é baseada na distribuição gaussiana associada ao campo aleatório. Os estimadores de máxima verossimilhança são muito utilizados por conta das suas propriedades assintóticas. No entanto em geoestatística, como tem-se distribuição normal multivariada associada aos campos aleatórios, a matriz de variância e covariância pode possuir dimensão muito grande, o que gera elevado tempo computacional ou até inviabiliza a

estimação dos parâmetros. Uma abordagem muito utilizada nesse contexto é a Bayesiana, o que requer métodos computacionais intensivos, como por exemplo MCMC.

2.1.5 Estimação Bayesiana em modelos geoestatísticos univariados

2.1.5.1 Amostrador de Gibbs

2.1.5.2 Metropolis Hasting

2.1.6 Krigagem

A krigagem nada mais é do que o processo de predição para os valores do campo aleatório que não foram amostrados. Ou seja, na prática os pesquisadores amostram alguns pontos dentro do espaço de interesse e realizam medição do atributo sob estudo. No entanto, existe interesse em conhecer como o atributo se comporta em todo o espaço, logo, utilizando os valores estimados para os parâmetros do modelo estabelecido é possível descrever o campo aleatório predito, contínuo no espaço. O nome krigagem é uma homenagem ao pesquisador sul-africano D.G. Krige que foi um dos pioneiros em estudos de predição espacial.

Os dois tipos de krigagem mais encontrados na literatura são a ordinária e simples, que se diferenciam quanto a pressuposição ou não de conhecimento sobre os parâmetros. Os dois tipos de krigagem são baseados nos estimadores de mínimos quadrados. Todos os valores preditos são função da média do campo aleatório com uma correção gerada pelos parâmetros de variância e covariância do campo aleatório. Segue a expressão para a krigagem simples:

ARRUMAR ESSA NOTAÇÃO...

$$\rho(z; s) = \mu_z + \sigma^2 r' \Sigma^{-1} (z - \mu_z),$$

sendo $\rho(z; s)$ o valor predito para o campo aleatório sob estudo na posição s do espaço de interesse, μ é a média de Z, σ é a variabilidade devida ao processo espacial, r' é o valor da correlação e Σ a matriz de variância e covariância do campo aleatório Z.

Diggle e Ribeiro (2006) e Elmatzoglou (2006), discutem com mais detalhes os métodos e propriedades dos processos de krigagem.

2.2 Campos aleatórios multivariados

O estudo possui, agora, mais de uma variável resposta de interesse, ou seja, existem n vetores $Y(s_i)$ de dimensão $px1$, cada um observado na posição espacial s_i, $i = 1, 2, ..., n$, do espaço de interesse d-dimensional G, a dimensão p é igual ao número de variáveis resposta sob estudo. Sendo assim, o modelo geoestatístico deve capturar a correlação entre os valores dentro de cada atributo e a correlação entre os atributos.

Um campo aleatório multivariado Z possui uma matriz de covariância Σ_Z de dimensão $npxnp$, sendo n o número de posições amostradas no espaço G, uma vez que, Σ_Z deve ser positiva definida, o maior problema nesse tipo de estudo é encontrar uma função de covariâncias $(C(s_i, s_j))_{\iota,\nu} = Cov(Z(s_i)_\iota, Z(s_j)_\nu)$ válida, sendo $Z(s_i)_\iota$ o valor do campo aleatório para a variável ι na posição s_i e $Z(s_j)_\nu$ o valor observado do campo aleatório para a variável ν na posição s_j, para todo ι, ν, s_i e s_j.

Na literatura três abordagens são utilizadas para encontrar estruturas de covariâncias cruzadas válidas para modelar campos aleatórios multivariados, são elas modelos separáveis, convolução de Kernel e convolução de modelos de covariância, a utilização dessas técnicas depende das configurações associadas ao campo aleatório, quanto a estacionariedade e isotropia.

A abordagem mais utilizada na literatura para encontrar funções de covariâncias válidas, devido a sua maior simplicidade, é de modelos separáveis, que possui, de forma geral, a seguinte expressão para a covariância entre cada par de posições:

$$C(s_i, s_j) = \rho_{s,s'}T,$$

sendo T uma matriz positiva definida de dimensão p x p e ρ uma função de correlação univariada válida. Segundo Mardia e Gooddall (1993) e Banerjee e Gelfand (2002) a expressão definida em (9) é uma matriz de covariância válida.

Como os modelos separáveis consideram estacionariedade do processo estocástico, tem-se que as associações entre os valores dentro de cada variável e entre as variáveis são atenuadas conforme a aumenta distância entre os pontos espaciais amostrados. Se, além de estacionário, o processo for isotrópico, a matriz $\Sigma_Z = R \otimes T$, sendo R uma matriz n x n com $(R)_{i,j} = \rho(s, s')$.

Detalhes sobre as técnicas de convolução para encontrar funções de correlação vál-

idas são apresentados por Gelfand et al. (2005).

Com relação a campos aleatórios que não são estacionários, Brown et al. (1994) e Sun et al. (1998) apresentam especificações para esse tipo de configuração de funções aleatórias multivariadas.

Além do problema de estruturar uma matriz de covarâncias válida, existe o problema de como será induzida na estrutura dos modelos as correlações cruzadas, na literatura existem algumas soluções, como modelos de co-regionalização, modelos hierarquicos e modelos com componente de correlação comum.

2.2.1 Modelos de co-regionalização

Como visto anteriormente, existem vetores $Y(s_i)$ de dimensão p x 1 , observados na posição espacial s_i do espaço de interesse d-dimensional G, a dimensão p é igual ao número de variáveis resposta sob estudo. Para modelar o problema Gelfand et al. (2005) propuseram o seguinte modelo:

$$Y(s) = X(s)\beta + Z(s) + \epsilon(s),$$

sendo $s = (s_1, s_2, ..., s_n)$, $X(s)$ uma matriz $npxq$ contento q possíveis covariáveis, β um vetor q x 1 de parâmetros associados as covariáveis, $Z(s)$ o campo aleatório , um processo latente, que existe mas não pode ser medido, considerando gaussianiedade, o campo aleatório possui distribuição normal np-variada com vetor de médias nulo $npx1$ e matriz de covariância Σ_Z $npxnp$ e $\epsilon(s)$ é o ruído brando associado ao processo de amostragem, geralmente esse erro aleatório é assumido como normalmente distribuído com vetor de médias nulo e matriz de covariância diagonal $npxnp$.

Dessa forma, imaginemos um modelo bivariado que foram observadas em duas localizações, empilhando os valores do campo aleatório por posição, tem-se a seguinte estrutura para a matriz de covariâncias:

$$\Sigma_Z = \begin{bmatrix} var(Z_1(s_1)) & cov(Z_1(s_1), Z_2(s_1)) & cov(Z_1(s_1), Z_1(s_2)) & cov(Z_1(s_1), Z_2(s_2)) \\ cov(Z_1(s_1), Z_2(s_1)) & var(Z_2(s_1)) & cov(Z_2(s_1), Z_1(s_2)) & cov(Z_2(s_1), Z_2(s_2)) \\ cov(Z_1(s_1), Z_1(s_2)) & cov(Z_1(s_1), Z_2(s_2)) & var(Z_1(s_2)) & cov(Z_1(s_2), Z_2(s_2)) \\ cov(Z_2(s_1), Z_1(s_2)) & cov(Z_2(s_1), Z_2(s_2)) & cov(Z_1(s_2), Z_2(s_2)) & var(Z_2(s_2)) \end{bmatrix}$$

Agora pensando que a matriz Σ_Z é paticionada em matrizes de dimensão pxp:

$$\Sigma_Z = \begin{bmatrix} \Sigma_{Z(s_1)} & \Sigma_{Z(s_1,s_2)} \\ \Sigma_{Z(s_1,s_2)} & \Sigma_{Z(s_2)} \end{bmatrix}$$

sendo que as matrizes da diagonal principal são responsáveis por capturar as covariâncias dentro de cada posição s_i e as demais matrizes capturam as covariâncias cruzadas.

A idéia do modelo de co-regionalização é decompor $Z(s_i)$:

$$Z(s_i) = A\omega(s_i),$$

sendo A uma matriz pxp de posto completo, que contem os parâmetros de variâncias e covariâncias a serem estimados e $\omega(s_i)$ um vetor aleatório $px1$, onde, para cada variável j, $\omega_j(s_i)'s$ são independentes e identicamente distribuídos e possuem média zero, e sob estacionariedade, variância 1. Dessa forma tem-se que:

$$var(Z(s_i)) = var(A\omega(s_i)) = Avar(\omega(s_i))A^{'},$$

sendo que a definição acima assegura que para todas as localizações a matriz $\Sigma_{(Z(s_i))}$ é igual, ou seja, nessa definição está incluída a estacionariedade da variância.

Retornando a simplicidade de um modelo bivariado observado em duas localizações georeferenciadas, tem-se:

$$A = \begin{bmatrix} a_{11} & 0 \\ a_{21} & a_{22} \end{bmatrix}$$

e

$$var(\omega(s_i)) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

essa matriz é identidade por conta da decomposição proposta, onde para todas variáveis j $\omega_j(s_i)'s$ são independentes e identicamente distribuídas com variância unitária.

Logo a matriz de covariância $\Sigma_{(Z(s_i))}$ para cada localização é:

$$\Sigma_{(Z(s_i))} = \begin{bmatrix} a_{11}^2 & a_{11}a_{21} \\ a_{11}a_{21} & a_{21}^2 + a_{22}^2 \end{bmatrix}$$

O próximo passo é pensar nas covariâncias cruzadas, usando a decomposição tem-se:

$$\Sigma_{Z(s_1,s_2)} = A \begin{bmatrix} cov(\omega_1(s_1), \omega_1(s_2)) & cov(\omega_1(s_1), \omega_2(s_2)) \\ cov(\omega_2(s_1), \omega_1(s_2)) & cov(\omega_2(s_1), \omega_2(s_2)) \end{bmatrix} A'$$

sendo que agora a matriz é diagonal, pois $cov(\omega_1(s_1), \omega_2(s_2)) = 0$, no entanto a diagonal principal não é mais unitária, e deve-se utilizar o resultado de que a $cov(\omega_j(s_1), \omega_j(s_2) = var(\omega_j(si))\rho_j(h)$, onde h é a distância euclidiana entre as localizações e para encontrar as correlações deve-se utilizar funções válidas, como por exemplo a Exponencial Potência. Dessa forma, tem-se:

$$\Sigma_{Z(s_1,s_2)} = \begin{bmatrix} a_{11}^2\rho_1(h) & a_{11}a_{21}\rho_1(h) \\ a_{11}a_{21}\rho_1(h) & a_{21}^2\rho_1(h) + a_{22}^2\rho_2(h) \end{bmatrix}$$

Todos os resultados acima, específicos para um modelo bivariado e com duas localizações amostradas, é facilmente extendida para n e p genéricos. Uma definição bem simples para as contas matriciais apresentadas é:

$$\Sigma_{Z(s)} = \sum_{j=1}^{p} R_j \otimes T_j,$$

sendo que cada R_j é uma matriz nxn que contem as correlações dentro de cada variável e cada $T_j = a_j a_j^{'}$ é uma matriz pxp que possui os parâmetros de covariâncias, onde a_j é a j-ésima coluna de A.

Com a estrutura do processo latente definida, o próximo passo é pensar na modelagem de toda a estrutura do modelo estatístico proposto por Gelfand et al. (2005), ou seja,

é necessário atribuir distribuição ao ruído branco. Devido as propriedades da distribuição gaussiana, ela é a mais utilizada na literatura.

Assumindo que cada vetor de ruídos brancos $\epsilon(s_i)$ possui distribuição gaussiana multivariada com vetor de médias nulo e matriz de covariância D de dimensão p x p e com todos os elementos de covariância nulos, ou seja, uma matriz diagonal, além disso, se o campo aleatório possui distribuição gaussiana multivariada com vetor de médias nulo de dimensão np x 1 e matriz de covariância cruzada $\sum_{j=1}^{p} R_j \otimes T_j$, então a distribuição de probabilidade de $Y(s)$ dado os parâmetros de média, de variância e covariância e de erro aleatório é definida como:

$$Y(s)|\theta \sim N(\mu, \sum_{j=1}^{p} R_j \otimes T_j + I_{nxn} \otimes D), \tag{4}$$

sendo θ o vetor de todos os parâmetros associados ao modelo.

Com a distribuição conjunta de $Y(s)$ estabelecida em (9), pode-se pensar agora em utilizar os estimadores de máxima verossimilhança e máxima verossimilhança restrita para fazer a estimação dos parâmetros, o que pode ser computacionalmente inviável, uma vez que, a dimensão da matriz de covariância do campo aleatório cresce exponencialmente com o aumento do número de variáveis respostas. Logo, métodos Bayesianos e métodos computacionais intensivos são muito utilizados nesse tipo de abordagem, onde utiliza-se alguma informação a priori sobre a distribuição dos parâmetros.

2.2.2 Modelos hierarquicos condicionais

Uma outra abordagem para problemas com campos aleatórios multivariados é a utilização de distribuições condicionadas entre as respostas. Para simplificar, considerando que o campo aleatório possui duas variáveis de interesse, deve-se pensar na seguinte modelagem aos dados (GELFAND et al., 2005):

$$Y_1(s) = X^T(s)\beta_1 + \sigma_1\omega_1(s) \tag{5}$$

$$Y_2(s)|Y_1(s) = X^T(s)\beta_2 + \alpha^{2|1}Y_1(s) + \sigma_2\omega_2(s) + \epsilon(s) \tag{6}$$

sendo que a primeira variável não possui ruído branco associado e explica uma parte da variabilidade da segunda variável, fato que explica a entrada da primeira variável como covariável da segunda.

Com a estrutura definida em (10) e (11), tem-se a seguinte função de verossimilhança:

$$L(\theta|Y(s)) = f(Y_1(s)|\theta_1)f(Y_2(s)|Y_1(s), \theta_2),$$

sendo θ_1 e θ_1 os vetores de parâmetros associados aos modelo para $Y_1(s)$ e $Y_2(s)|Y_1(s)$, respectivamente, e a união dos dois vetores é o vetor θ que contém todos os parâmetros do modelo.

A vantagem da abordagem condicional é que se for assumido que a distribuição a priori para θ é igual ao produto das distribuições a priori de θ_1 e θ_2, ou seja, independentes, o condicionamento resulta na fatorização de dois modelos e assim, cada um deles pode ser ajustado em separado.

2.2.3 Modelos com componente de correlação parcialmente comum

Essa modelagem para um problema de geoestatística com resposta multivariada foi proposta por Diggle e Ribeiro Jr. (2006), por simplicidade um modelo bivariado:

$$Y_1(s) = \mu_1 + Z_1(s) + \epsilon_1(s)$$

$$Y_2(s) = \mu_2 + Z_2(s) + \epsilon_2(s)$$

onde $s = (s_1, s_2, ..., s_n)$ são as localizações amostradas. Com essa modelagem, estaríamos ajustando um modelos para cada resposta, dessa forma, cada $Z_j(s)$ possui distribuição normal n-variada com vetor de médias nulo e matriz de covariância , que possui na diagonal a variância estacionária de cada variável e nas demais entradas os valores das covariâncias para cada variável.

No entanto, a intenção é induzir de alguma forma a correlação entre os dois atributos, logo, utilizando teoria de probabilidade, tem-se:

$$Z_j(s) = \sigma_j R_j(s),$$

dessa forma $R_j(s)$ possui distribuição normal n-variada com vetor de médias nulo e matriz de covariâncias igual a matriz de correlações de $Z_j(s)$, onde todos elementos da diagonal principal são iguais a 1 e as demais entradas são as correlações, $\rho_j(s_l, s_k$ para todo l diferente de k, entre cada par de valores do campo aleatório. No entanto, dessa forma ainda não induzimos a correlação entre as respostas, logo pode-se pensar em fazer mais uma separação dos campos aleatórios:

$$Z_j(s) = \sigma_{0j} R_0(s) + \sigma_j R_j(s),$$

e agora os dois campos aleatórios possuem uma componente comum $R_0(s)$ possui distribuição normal n-variada com vetor de médias nulo e matriz de covariâncias onde os elementos são correlações $\rho_0(s_l, s_k$ para todo l e k, é esse processo espacial que induzirá a correlação entre as respostas. Dessa forma chega-se ao modelo propostos:

$$Y_1(s) = \mu_1 + \sigma_{01} R_0(s) + \sigma_1 R_1(s) + \epsilon_1$$

$$Y_2(s) = \mu_2 + \sigma_{02} R_0(s) + \sigma_2 R_2(s) + \epsilon_2$$

sendo $s = (s_1, s_2, ..., s_n)$ as localizações onde observa-se as respostas $Y_1(s)$ e $Y_2(s)$, que são vetores $nx1$, μ_1 e $\mu(s)$ são vetores $nX1$ de médias, que podem possuir ou não covariáveis, $R_0(s)$, $R_1(s)$ e $R_2(s)$ são campos aleatórios gaussianos univariados mutuamente independentes com vetores de médias nulos e com matrizes de covariâncias que possuem diagonais principais unitárias e demais valores iguais a valores provinientes de funções de correlações válidas, ou seja, as matrizes de covariâncias dos campos aleatórios do modelo só possuem valores iguais ou menores que 1.

Com essa modelagem, deve-se pensar em um vetor $2nx1$ $Y(s) = (Y_1(s), Y_2(s)$, dessa forma teremos também um campo aleatório Σ_Z que possui distribuição normal $2n$-variada com vetor de médias nulo e matriz de covariância de dimensão $2nx2n$, para exemplificar imaginemos que foram amostradas duas localizações georeferenciadas, dessa forma:

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

sendo h igual a distância euclidiana entre as duas localizações amostradas, e é fácil extender esse matriz para um número maior de localizações amostradas.

Definida a estrutura paramétrica do campo aleatório associado ao modelo, o próximo passo é pensar na distribuição de probabilidade do vetor $Y(s)$, no entanto para tal é necessário fazer suposição quanto a distribuição de probabilidade dos ruídos brancos, novamente considerando gaussianiedade e independência entre os ruídos tem-se que $\epsilon(s) = (\epsilon_1(s), \epsilon_2(s))$ segue uma distribuição normal $2n$-variada com vetor de médias nulo e matriz de covariâncias diagonal $2nx2n$ D, onde as primeiras n linhas recebem a variância τ_1^2 proviniente do $\epsilon_1(s)$ e as demais linhas recebem a variância τ_1^2 proviniente do $\epsilon_2(s)$, sendo assim:

$$Y(s)|Z(s) \sim N((\mu_1, \mu_2); \Sigma_Z + D), \tag{7}$$

E agora o problema é de estimação dos parâmetros, como conhecemos a distribuição de probabilidade de $Y(s)$ podemos utilizar novamente estimadores baseados na função de verossimilhança, o que não é muito fácil de calcular devido ao número elevado de parâmetros, logo, métodos computacionais são exigidos ou pode-se pensar em técnicas bayesianas para o problema.

2.2.4 Modelos com componente de correlação negativa parcialmente comum

Na última seção foi apresentada a modelagem proposta por Diggle e Ribeiro Jr. (2006), no entanto essa abordagem não consegue capturar correlação negativa entre as respostas, ou seja, se as duas respostas de interesse possuem padrão espacial e são correlacionadas negativamente o modelo não consiguirá ajustar tal situação. Por conta dessa característica, pensamos em uma pequena alteração no modelo:

$$Y_1(s) = \mu_1 + \sigma_{01} R_0(s) + \sigma_1 R_1(s) + \epsilon_1$$

$$Y_2(s) = \mu_2 - \sigma_{02} R_0(s) + \sigma_2 R_2(s) + \epsilon_2$$

com o sinal negativo antes do parâmetro σ_{02} coseguimos induzir que a matriz de covariância Σ_Z continue gerando somente valores positivos nas partições referentes a cada variável, o que é necessário, pois não faz sentido falar em correlação negativa dentro de um mesmo atributo.

No entanto, nas partições de Σ_Z relativas as covariâncias entre as variáveis de interesse é garantido que todos elementos ficaram negativos, ou seja, conseguimos induzir o modelo a capturar correlações negativas entre as variáveis.

3 MATERIAL E MÉTODOS

Inicialmente serão conduzidos estudos de simulação para todos os modelos bivariados sob estudo, com esse estudo será possível detectar quais metodologias são mais eficientes em cada caso, onde eficiência pode ser encarado como estimação dos parâmetros próximos aos valores reais, resíduos pequenos e velocidade computacional. Para fazer as simulações e estimações foi utilizado o ambiente R de programação versão 2.7.1, sendo que além dos pacotes básicos, foi utilizado o pacote geoR e seus pacotes associados.

Após o estudo de simulação, nesse relatório serão apresentados resultados em cima de um conjunto de dados sobre qualidade de solo. Pesquisadores da ESALQ fizeram, em 2006, medição de diversos parâmetros em 67 localizações georeferenciadas para avaliar a qualidade do solo de uma fazenda em Echaporan/SP, que possui extensão de 51.8 hectares, tal fazenda possui dois tipos de manejo de solo feitos em tempos passados: pastagem feita nas regiões central e nas coordenadas com menores valores para x e cultivo de cítricos na região com valores mais elevados para as coordenadas de x, esses tipos de manejo podem afetar muito a qualidade do solo, onde se espera que a região com histórico de pastagem possua solo mais pobre.

Sendo assim, dois parâmetros levantados pelos pesquisadores serão modelados nesse trabalho: saturação por bases e Ph, 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, solos com Ph entre 5 e 7 tendem a reter mais nutrientes, sendo assim, é possível existir uma forte correlação entre essas duas respostas, o que justifica a tentativa de um modelo multivariado.

Com relação as coordenadas das localizações amostradas, os pesquisadores fizeram a transformação do sistema de latitudes e longitudes para UTM, sendo assim a distância euclidiana é uma ótima medida para modelar as funções de correlação.

Para todos os modelos propostos, a distribuição gaussiana será utilizada, devido as suas propriedades inferenciais fortes, além disso a família das funções de correlação de Matérn também serão utilizadas devido a sua capacidade de gerar funções que forneçam processos mais ou menos suaves, conforme os dados necessitem.

Dessa forma existem dois campos aleatórios gaussianos sob estudo, onde consideramos que cada um deles possui distribuição de probabilidade normal 67-variada, conforme

descrito na seção 2.1, quando utilizamos os modelos bivariados temos um vetor $134x1$ de valores observados, que segue uma distribuição normal 134-variada conforme descrito na seção 2.2.

Nos resultados serão apresentadas, inicialmente, uma breve análise descritiva dos dados, depois serão propostos modelos univariados para cada uma das respostas e por último serão expostos os modelos bivariados ajustados. E por último serão apresentados estudos de resíduos pra validação pressupostos.

4 RESULTADOS

Nesta seção serão apresentados todos os resultados dos estudos de simulação e dos estudos com dados observacionais.

4.1 Resultados das simulações

4.2 Aplicação com os dados de qualidade do solo

Antes da proposição de um modelo geoestatístico para as variáveis em questão, é necessário fazer um estudo exploratório dos dados com o objetivo de conhecer melhor o comportamento dos mesmos.

4.2.1 Análise exploratória

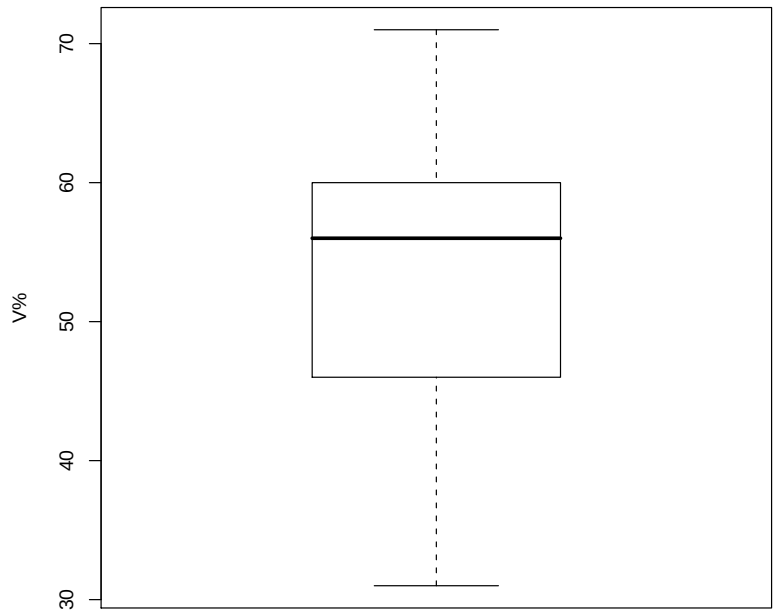
4.2.1.1 Saturação por bases do solo

Essa variável representa a capacidade do solo reter nutrientes bons para os cultivos: N, P, K, Ca e Mg, ou seja, quanto maior o valor medido da saturação, mais o solo é capaz de reter tais substâncias. Seguem as estatísticas descritivas dessa variável, desconsiderando o padrão espacial:

Mínimo	Q1	Mediana	Média	D.P.	Q3	Máximo
31.00	46.00	56.00	53.27	10.05	60.00	71.00

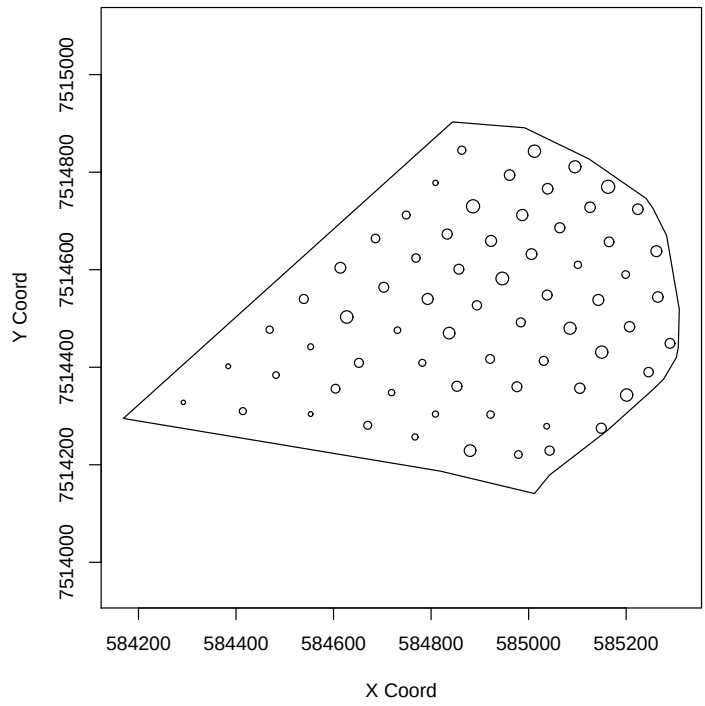
Tabela 1: Estatísticas descritivas da saturação por bases

A tabela 1 mostra que existe uma amplitude elevada dos dados, no entanto a média e a mediana se aproximam, o que revela uma certa simetria dos dados, além disso o desvio padrão é relativamente baixo com relação a média. Abaixo segue um boxplot de resumo das estatísticas descritivas:



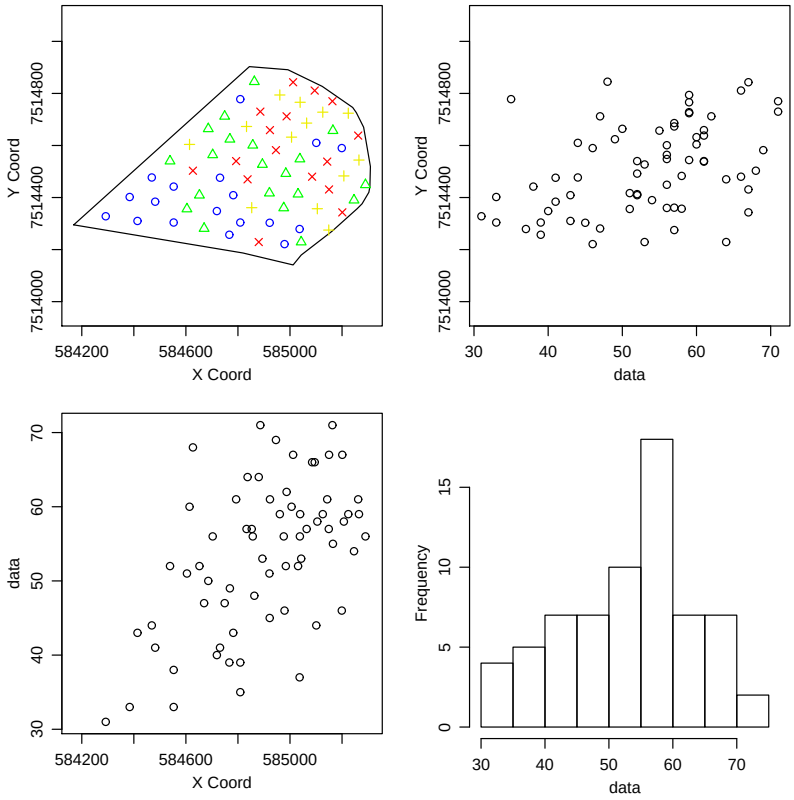
Com o gráfico acima, tem-se que, além das configurações captadas com as estatísticas descritivas, não existe pontos discordantes dos demais.

Agora o próximo passo é fazer uma análise descritiva pensando que existe um padrão espacial nos dados, segue um gráfico de círculos dos dados nas respectivas posições espaciais, onde quanto maior o círculo, maior o valor do atributo observado:



Com o gráfico acima, tem-se que aparentemente a coordenada y não interfere muito no valor médio da variável, uma vez que, existem valores distintos para todas as coordenadas, com relação a coordenada x, tem-se que, aparentemente, quanto maior o valor da coordenada x maior os valores da variável, o que leva a suspeitar da constância da média, ou seja, o campo aleatório pode não ser estacionário na média. Outra informação importante é quanto a área de manejo, a área com pastagem aparentemente possui menor valor de média.

Seguem mais gráficos exploratórios:



O primeiro e terceiro gráfico acima, corroboram a idéia de que localizações com menor valor para a coordenada x apresentam menores valores, onde as cores representam os quartis dos dados, quanto mais fria o quantil é menor. Além disso, o histograma mostra que os dados aparentemente são simétricos com relação a média. No entanto, a interpretação mais válida com relação aos gráficos é que, realmente, parece existir um padrão espacial nos dados, onde localizações mais próximas possuem valores observados parecidos, ou seja, existe uma suavidade na variação nos valores observados.

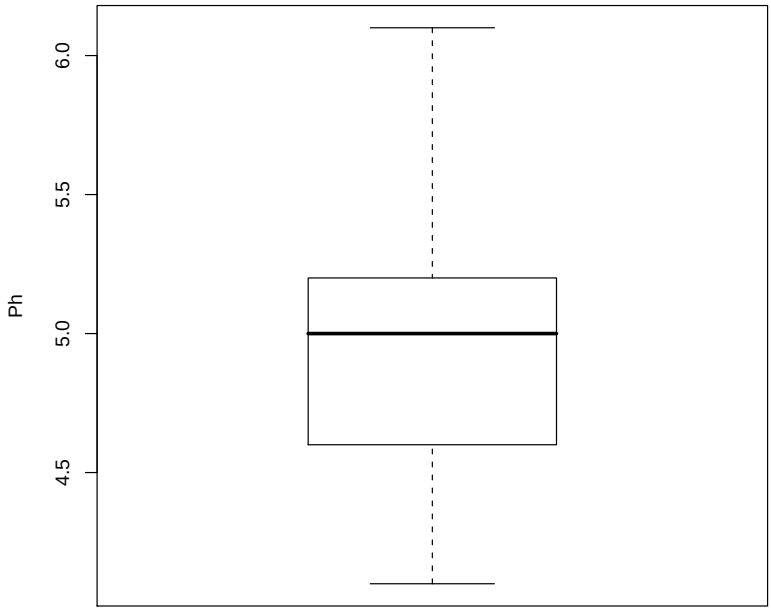
4.2.1.2 Ph do solo

Essa variável é o Ph medido no solo nas localizações amostradas, o Ph tolerado como bom para o solo é entre 5 e 7, sendo assim, abaixo seguem as análises descritivas similares a da saturação por bases:

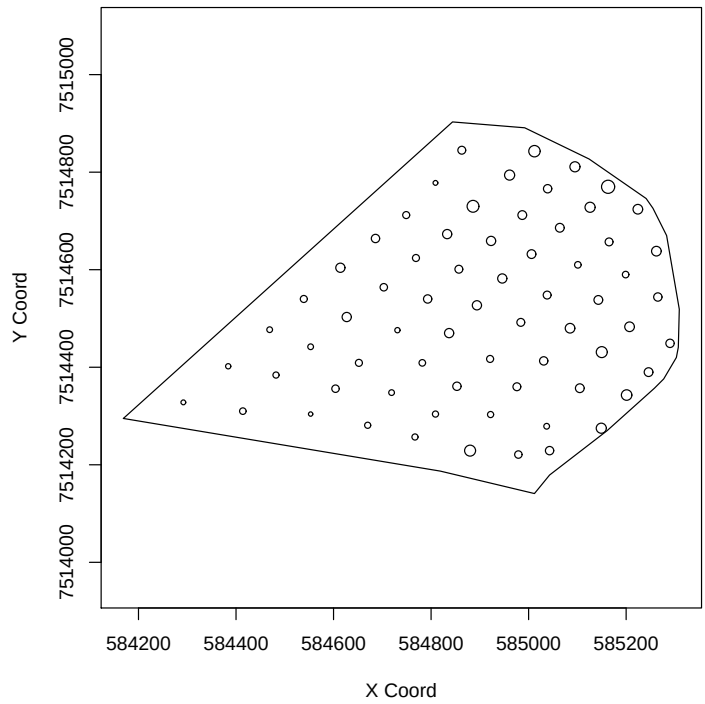
A tabela acima mostra que os dados estão poucos dispersos, a amplitude é pequena, a média se aproxima da mediana e a variabilidade é muito pequena com relação a média.

Mínimo	Q1	Mediana	Média	D.P.	Q3	Máximo
4.10	4.60	5.00	4.94	0.42	5.20	6.10

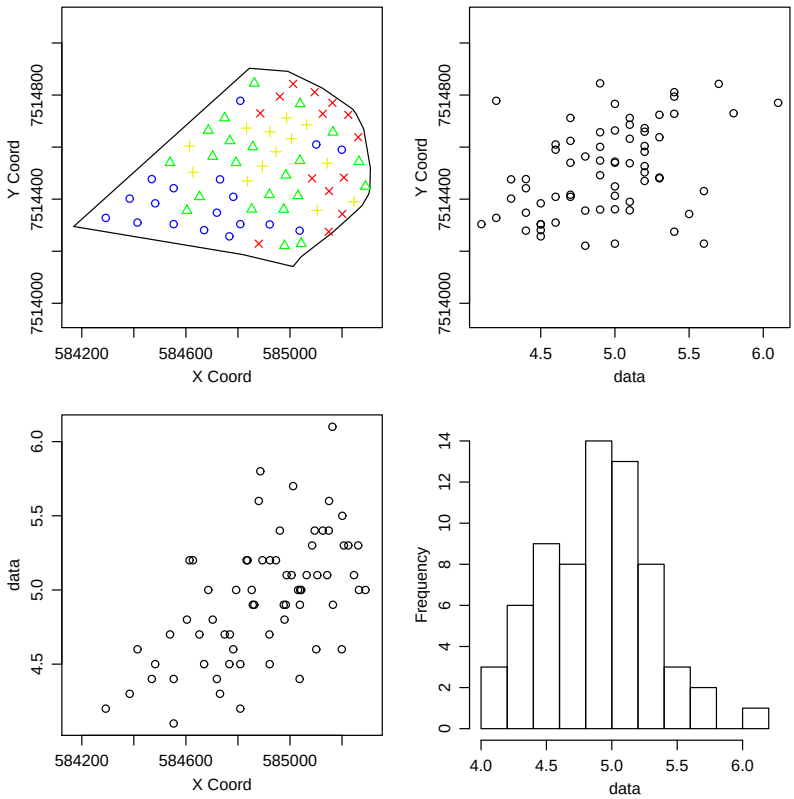
Tabela 2: Estatísticas descritivas do Ph



O boxplot acima mostra que, não existe pontos destoantes dos demais.



Com relação a tendência na



Com relação aos gráficos acima, tem-se que as interpretações são similares à da última variável.

4.2.2 Modelos Univariados

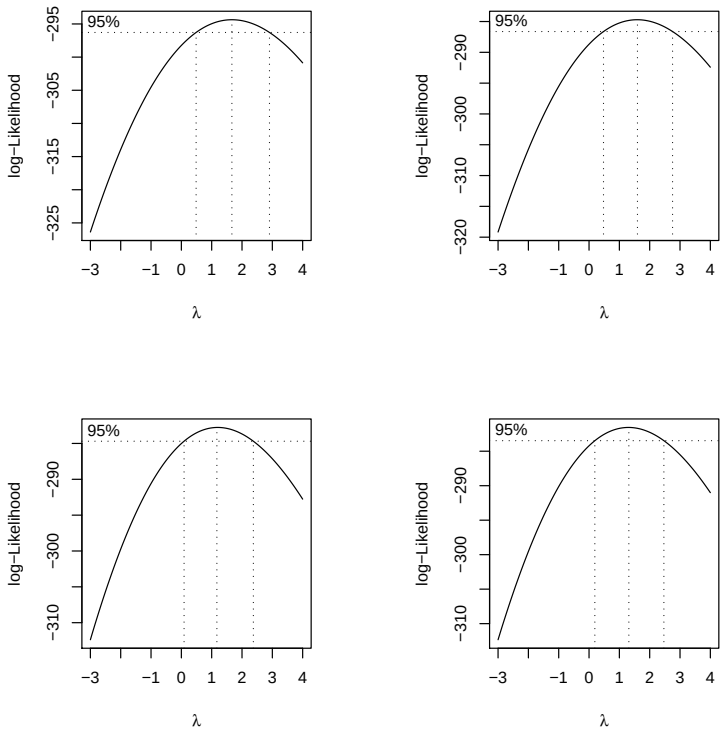
Antes de propor modelos multivariados aos dados, foram conduzidos estudos univariados, para cada atributo de interesse, onde todas as estimações feitas foram frequentistas.

4.2.2.1 Saturação por bases do solo

Com relação a esta variável, da análise exploratória inicial, suspeita-se que existe um padrão espacial nos dados, além disso, suspeita-se que a média do processo, aparentemente, é influenciada ou pela coordenada x ou pela área de manejo. Sendo assim, serão propostos modelos que consideram estacionariedade da função de correlação, mas com diferentes tendências para as médias, sendo assim, o modelo pode ter mais ou menos parâmetros relativos a média.

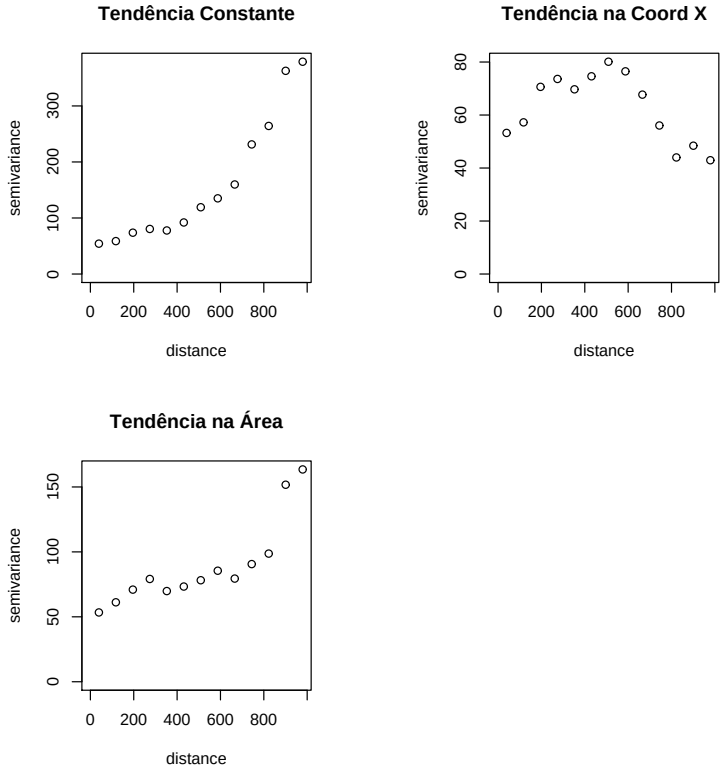
Para toda a modelagem foi utilizada a família Matérn de funções de correlações válidas, essa escolha foi feita por conta dessa família possuir funções deriváveis e não deriváveis em todo o domínio, ou seja, essa família engloba funções suaves e não suaves para as correlações, e essa suavidade do processo é determinada através do parâmetro κ da função, kappas maiores que 1.5 são as funções deriváveis.

No entanto, antes de propor alguma modelagem, é atribuído ao campo aleatório e ao ruído branco distribuições gaussianas, além disso é suposto estacionariedade das variâncias e covariâncias, conforme visto na revisão bibliográfica, sendo assim, esses pressupostos devem ser testados, seguem os gráficos dos λ's estimados para a transformação da família de Box-Cox para cada tendência estudada:



Os gráficos acima representam os intervalos de confiança para os lambdas estimados para a transformação de Box-Cox, sendo que o primeiro não considera tendência alguma, o segundo considera tendência na área de manejo, o terceiro considera tendência na coordenada X e o último considera tendência na área e na coordenada X. Como todos os intervalos de confiança contem λ=1, não será conduzida nenhuma transformação nos dados.

Agora o próximo passo é fazer a estimação dos parâmetros para alguns modelos, para tal será utilizado o método da máxima verossimilhança, no entanto, devido a complexidade do sistema de derivadas que deve ser resolvido, esse método utiliza métodos numéricos para calculas as estimativas, e como todos métodos numéricos precisam de um valor inicial para começar as iterações serão apresentados gráficos de semivariogramas empíricos, os quais serão utilizados para dar chutes iniciais aos parâmetros, cabe ressaltar que foram fixados alguns κ's distintos de forma que as funções de correlações englobadas na modelagem sejam mais ou menos suvaes:



Os gráficos acima mostram que, para cada tendência considerada os valores de semivariograma empírica são bem distintos, além disso, tem-se que para distância grande entre as localizações os valores de semivariograma empírico se comportam de forma estranha, nos casos de tendência na área de manejo e tendência constante, o semivariograma não pára de crescer, ou seja, se fosse utilizado esse método para estimação dos parâmetros, deveria ser considerando um alcance prático de forma que a partir de uma certa distância seria considerado que as localizações não possuem mais correlação. No caso da tendência em x, o comportamento é ainda mais estranho, pois o ruído branco é maior que o sinal e tem-se ainda que, o semivariograma vai aumentando conforme a distância aumenta e em um certo ponto a estatística cai novamente, ou seja, para valores mais distantes a correlação entre as observações volta a crescer, essa característica destoa totalmente dos pressupostos da função de covariância, que quanto maior as distâncias menor a correlação entre os valores do campo aleatório. No entanto, o semivariograma empírico não é uma boa medida para estabelecer os parâmetros estimados, ou seja, não é muito adequado tentar ajustar um modelo aos valores do semivariograma empírico e considerar que esse ajuste são as estimativas para os parâmetros envolvidos nos modelos, esse método não deve ser utilizado por conta do acaso amostral ou pelo tamanho da amostra, pois se existem poucas observações,

alguns semivariogramas serão calculados com poucas observações que estarão dentro da distância considerada. Sendo assim, os gráficos acima têm caráter exploratório e serão utilizados para dar os valores iniciais para os estimadores de máxima verossimilhança.

A tabela abaixo refere-se aos parâmetros estimados dos modelos com todas tendências levadas em consideração, com todos os κ's utilizados e os valores maximizados dos logaritmos das funções de verossimilhança:

β	τ^2	σ^2	ϕ	κ	log-verossim.
48.53	59.57	120.62	625.58	1	-239.8
47.98	62.40	124.31	516.78	1.5	-239.8
47.27	63.29	149.69	489.58	2	-239.7
47.32	63.61	122.37	326.39	3	-239.7

Tabela 3: Estimativas de Máxima Verossimilhança - Tendência constante

Com os resultados acima, tem-se que independente dos valores fixados para κ as estimativas de verossimilhança se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 2 e 3 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
47.92	8.83	38.78	39.28	59.18	1	-238.00
47.92	8.86	43.35	34.69	50.33	1.5	-237.98
47.92	8.88	45.53	32.47	44.34	2	-237.96
47.44	4.85	64.96	34.57	190.39	3	-239.27

Tabela 4: Estimativas de Máxima Verossimilhança - Tendência na área de manejo

Agora com tendência na área de manejo, tem-se que os resultados acima, independente dos valores fixados para κ, se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 1.5 e 2 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

Com tendência na coordenada x, tem-se que os resultados acima, independente dos valores fixados para κ, se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 2 e 3 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
-14300.66	0.025	27.56	41.46	45.42	1	-234.72
-14300.91	0.025	33.53	35.53	39.51	1.5	-234.70
-14300.17	0.025	36.40	32.68	35.33	2	-234.68
-14297.97	0.025	39.14	29.97	29.74	3	-234.66

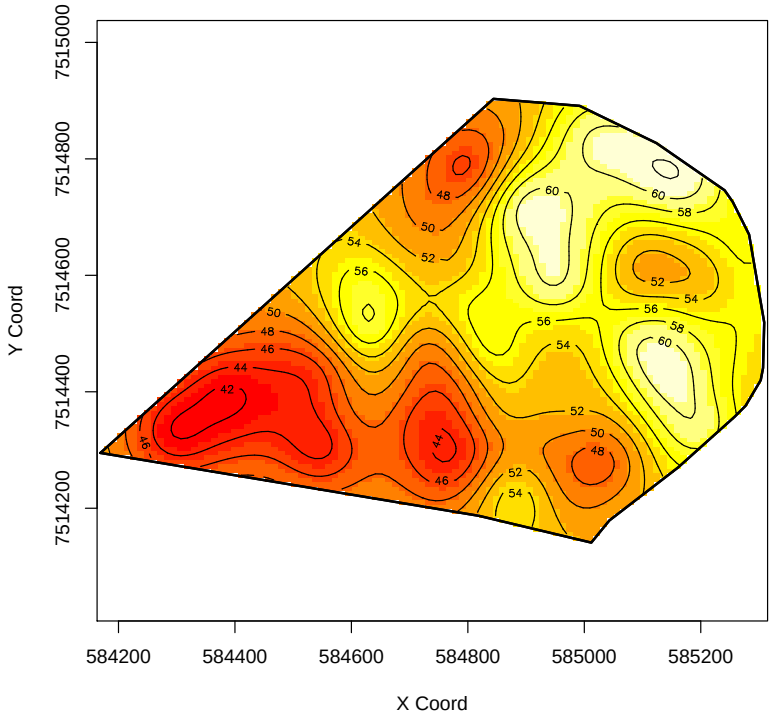
Tabela 5: Estimativas de Máxima Verossimilhança - Tendência na coordenada x

O próximo passo é escolher entre os modelos com mais ou menos parâmetros na média, ou seja, devemos fazer a seleção de covariáveis importantes ao modelo. Para tal, não se pode comparar os máximos das funções de verossimilhança, uma vez que, os valores das mesmas são alterados conforme o número de parâmetros no modelo, sendo assim, como os modelos possuem números de parâmetros de média distintos, se deve utilizar outro critério para seleção, sendo assim, será utilizado o critério da informação de Akaike, esse critério faz uma ponderação entre a explicação do modelo e o número de parâmetros usados, ou seja, esse critério é uma espécie de punição ao modelo pelo número de parâmetros utilizados para explicar uma determinada variabilidade, logo, quanto menor o valor da estatística melhor o modelo:

Tendência	κ	AIC
$Constante$	2	487.499
$Constante$	3	487.328
$Area$	1.5	485.9588
$Area$	2	485.922
$Coord.X$	2	479.361
$Coord.X$	3	479.321

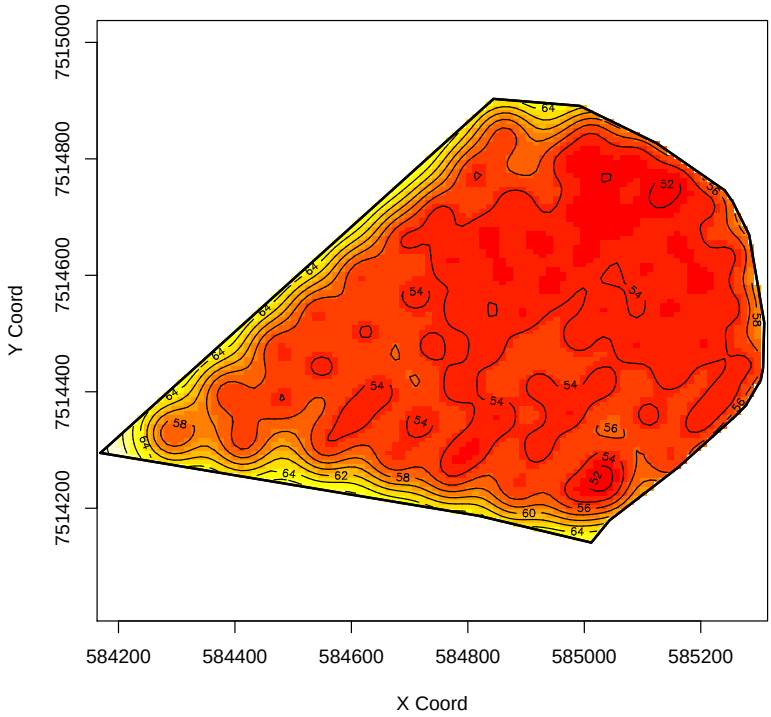
Tabela 6: Critério de informação de Akaike

Com os resultados acima, tem-se que o modelo com tendência na coordenada X e κ igual a 3 é o que melhor se ajustou aos dados, sendo assim, o próximo passo é fazer a predição ou krigagem para todo o espaço da fazenda, sendo assim, com a estimação dos parâmetros feita, é possível prever o valor do campo aleatório para localizações não amostradas, essa predição é feita através da média estimada para a localização ponderada pelos valores estimados para a variância e covariância do campo aleatório. Segue o gráfico da krigagem:



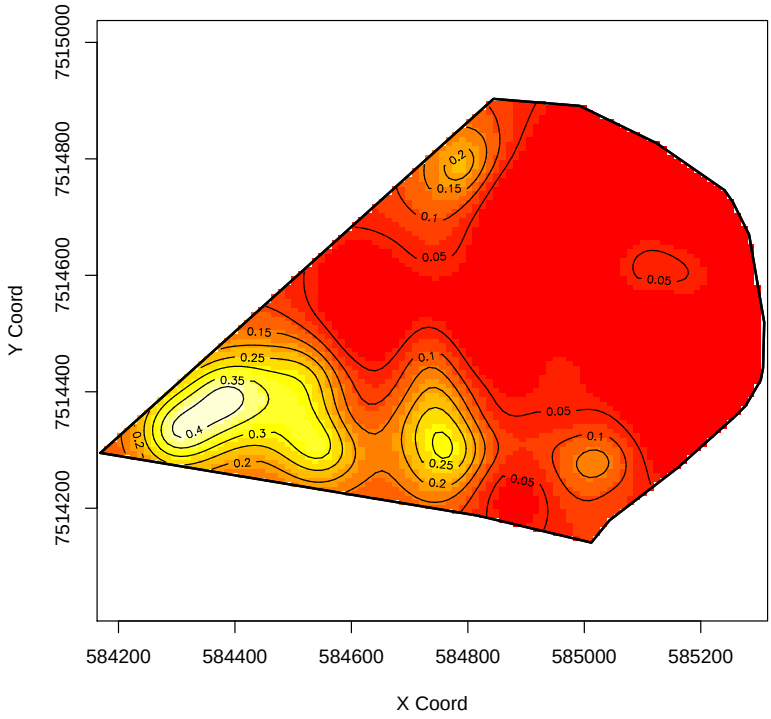
No gráfico acima as cores mais próximas do branco indicam valores mais elevados para a saturação por bases e conseqüentemente cores próximas do vermelho indicam valores menores para a saturação. Analisando os valores observados nas localizações amostradas, tem-se que as predições se aproximaram refletiram os valores reais e suavizou para o restante do espaço o campo aleatório.

O próximo passo é analisar o gráfico das variâncias das estimativas do campo aleatório, segue o gráfico:



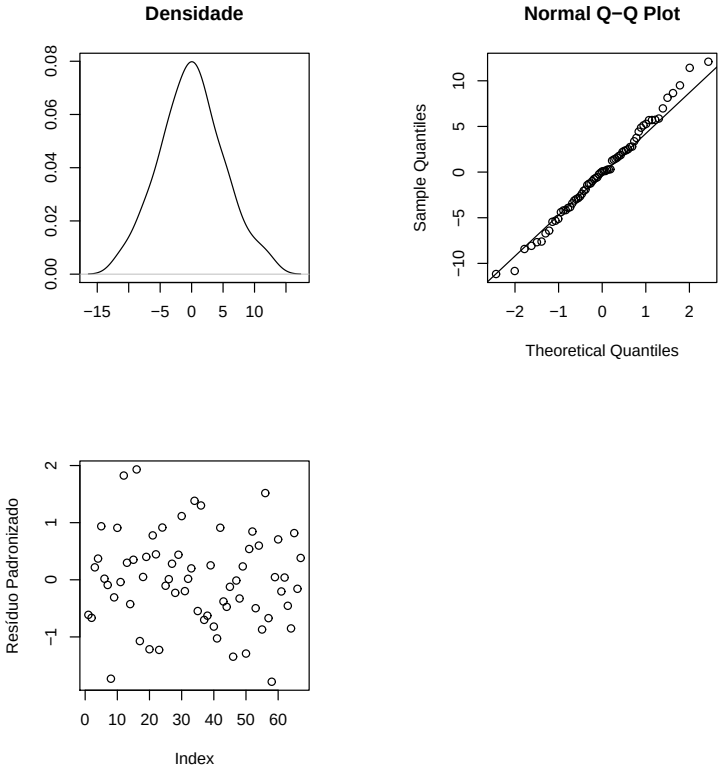
Com o gráfico acima tem-se que existe o padrão espacial esperado para as variâncias, ou seja, as variâncias mudam no espaço todo, mas essa mudança é suave, o que de certa forma corrobora a idéia de estacionariedade da função de covariância.

Com a definição dos parâmetros estimados, e com os valores preditos para cada posição do grid definido, é possível fazer análises a partir da distribuição normal de cada posição onde foi feita predição, assim é possível calcular a probabilidade prevista de que a saturação por bases seja menor que 40 em cada posição predita, ou seja, se pode fazer um gráfico das probabilidades de cada valor predito ser inferior a 40, que é um valor abaixo do esperado para uma boa qualidade do solo da fazenda sob estudo, abaixo segue o gráfico:



Com o gráfico acima tem-se que valores preditos em regiões com menor valor para a coordenada X possuem maior probabilidade da saturação ser inferior a 40, informação que corrobora a análise descritiva onde se tem a clara impressão de que a média da saturação é inferior em regiões com menores coordenadas X.

Uma outra análise que pode ser conduzida é a validação do pressuposto de normalidade univariada do ruído branco, abaixo seguem análises gráficas do tipo:



O primeiro gráfico acima mostram que a densidade dos resíduos brancos estimados se aproximam de uma distribuição normal, além disso o qqplot mostra que existe uma pequena fuga da normalidade na cauda superior dos resíduos, mas nada que afete o pressuposto e por último não existe valores para os resíduos padronizados fora do intervalo de [-3,3], o que indica não existir valores discrepantes dos erros brancos. Após a análise gráfica o teste de normalidade de Shapiro-Wilk foi conduzido, o qual gerou um p-valor de 0.9674, logo a normalidade não é rejeitada e esse pressuposto está atendido.

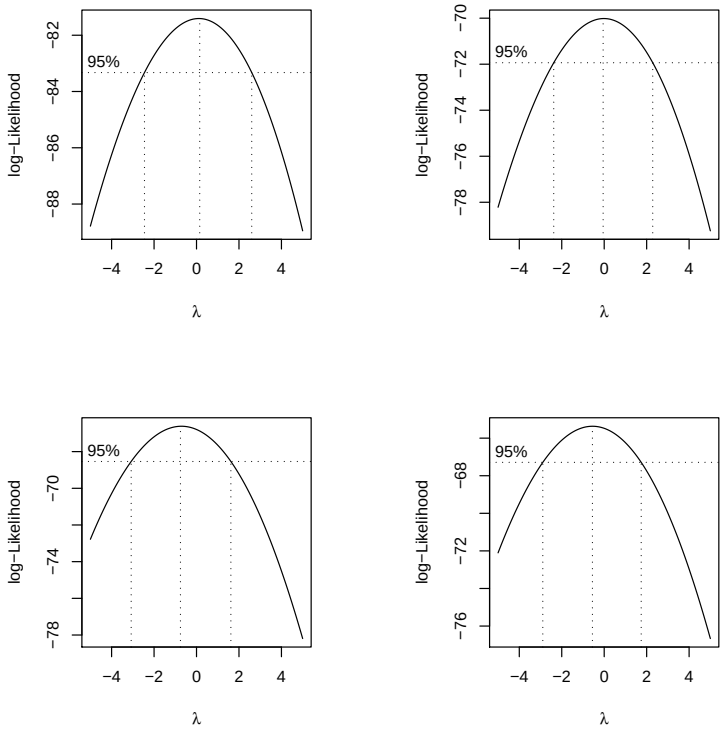
4.2.2.2 Ph do solo

Com relação a esta variável, da análise exploratória inicial, suspeita-se que existe um padrão espacial nos dados, além disso, suspeita-se que a média do processo, aparentemente, é influenciada ou pela coordenada x ou pela área de manejo. Sendo assim, serão propostos modelos que consideram estacionariedade da função de correlação, mas com diferentes tendências para as médias, sendo assim, o modelo pode ter mais ou menos parâmetros relativos a média.

Para toda a modelagem foi utilizada a família Matérn de funções de correlações

válidas, essa escolha foi feita por conta dessa família possuir funções deriváveis e não deriváveis em todo o domínio, ou seja, essa família engloba funções suaves e não suaves para as correlações, e essa suavidade do processo é determinada através do parâmetro κ da função, kappas maiores que 1.5 são as funções deriváveis.

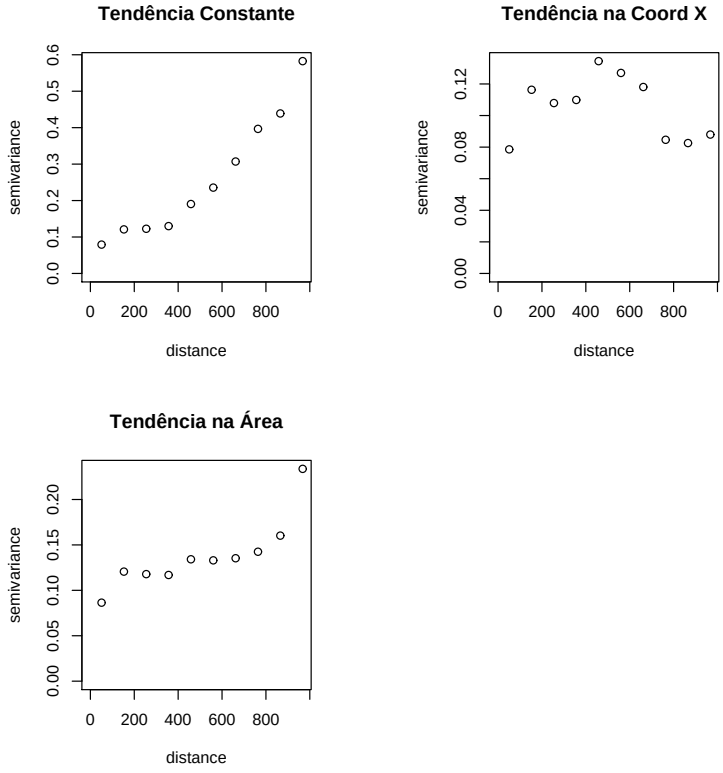
No entanto, antes de propor alguma modelagem, é atribuído ao campo aleatório e ao ruído branco distribuições gaussianas, além disso é suposto estacionariedade das variâncias e covariâncias, conforme visto na revisão bibliográfica, sendo assim, esses pressupostos devem ser testados, seguem os gráficos dos λ's estimados para a transformação da família de Box-Cox para cada tendência estudada:



Os gráficos acima representam os intervalos de confiança para os lambdas estimados para a transformação de Box-Cox, sendo que o primeiro não considera tendência alguma, o segundo considera tendência na área de manejo, o terceiro considera tendência na coordenada X e o último considera tendência na área e na coordenada X. Como todos os intervalos de confiança contem λ=1, não será conduzida nenhuma transformação nos dados.

Agora o próximo passo é fazer a estimação dos parâmetros para alguns modelos,

para tal será utilizado o método da máxima verossimilhança, no entanto, devido a complexidade do sistema de derivadas que deve ser resolvido, esse método utiliza métodos numéricos para calculas as estimativas, e como todos métodos numéricos precisam de um valor inicial para começar as iterações serão apresentados gráficos de semivariogramas empíricos, os quais serão utilizados para dar chutes iniciais aos parâmetros, cabe ressaltar que foram fixados alguns κ's distintos de forma que as funções de correlações englobadas na modelagem sejam mais ou menos suaves:



Os gráficos acima mostram que, para cada tendência considerada os valores de semivariograma empírica são bem distintos, além disso, tem-se que para distância grande entre as localizações os valores de semivariograma empírico se comportam de forma estranha, nos casos de tendência na área de manejo e tendência constante, o semivariograma não pára de crescer, ou seja, se fosse utilizado esse método para estimação dos parâmetros, deveria ser considerando um alcance prático de forma que a partis de uma certa distância seria considerado que as localizações não possuem mais correlação. No caso da tendência em X, o comportamento é ainda mais estranho, pois o ruído branco é maior que o sinal e tem-se ainda que, o semivariograma vai aumentando conforme a distância aumenta e em um certo ponto a estatística cai novamente, ou seja, para

valores mais distantes a correlação entre as observações volta a crescer, essa característica destoa totalmente dos pressupostos da função de covariância, que quanto maior as distâncias menor a correlação entre os valores do campo aleatório. No entanto, o semivariograma empírico não é uma boa medida para estabelecer os parâmetros estimados, ou seja, não é muito adequado tentar ajustar um modelo aos valores do semivariograma empírico e considerar que esse ajuste são as estimativas para os parâmetros envolvidos nos modelos, esse método não deve ser utilizado por conta do acaso amostral ou pelo tamanho da amostra, pois se existem poucas observações, alguns semivariogramas serão calculados com poucas observações que estarão dentro da distância considerada. Sendo assim, os gráficos acima têm caráter exploratório e serão utilizados para dar os valores iniciais para os estimadores de máxima verossimilhança.

A tabela abaixo refere-se aos parâmetros estimados dos modelos com todas tendências levadas em consideração, com todos os κ's utilizados e os valores maximizados dos logaritmos das funções de verossimilhança:

β	τ^2	σ^2	ϕ	κ	log-verossim.
4.905	0.1006	0.1493	510.58	1	-25.69
4.904	0.1055	0.2513	650.50	1.5	-25.57
4.903	0.1061	0.2880	569.74	2	-25.45
4.902	0.1063	0.2961	430.36	3	-25.31

Tabela 7: Estimativas de Máxima Verossimilhança - Tendência constante

Com os resultados acima, tem-se que independente dos valores fixados para κ as estimativas de verossimilhança se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 2 e 3 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores, no entanto, quanto maior o κ mais o máximo da verossimilhança está aumentando, o que pode indicar que existe alguma covariável não considerada.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
4.7094	0.4421	0.1220	0.0000	0.0000	1	-24.60
4.7094	0.4421	0.1220	0.0000	0.0000	1.5	-24.60
4.7794	0.2272	0.1100	0.1100	349.35	2	-24.86
4.7880	0.1948	0.1092	0.1092	327.03	3	-24.79

Tabela 8: Estimativas de Máxima Verossimilhança - Tendência na área de manejo

Agora com tendência na área de manejo, tem-se que os resultados acima mostram não existir padrão espacial for considerada essa tendência na média, pois os melhores modelos foram os com variabilidade e parâmetro de correlação do campo aleatório igual a zero. Sendo assim esse tipo de tendência será desconsiderado do estudo.

β_0	β_1	τ^2	σ^2	ϕ	κ	log-verossim.
-584.69	0.001	0.1124	0.0000	0.0000	1	-21.86
-586.88	0.001	0.1050	0.0104	171.76	1.5	-21.71
-608.32	0.001	0.0081	0.1040	24.508	2	-19.36
-584.69	0.001	0.1124	0.0001	822.95	3	-21.89

Tabela 9: Estimativas de Máxima Verossimilhança - Tendência na coordenada X

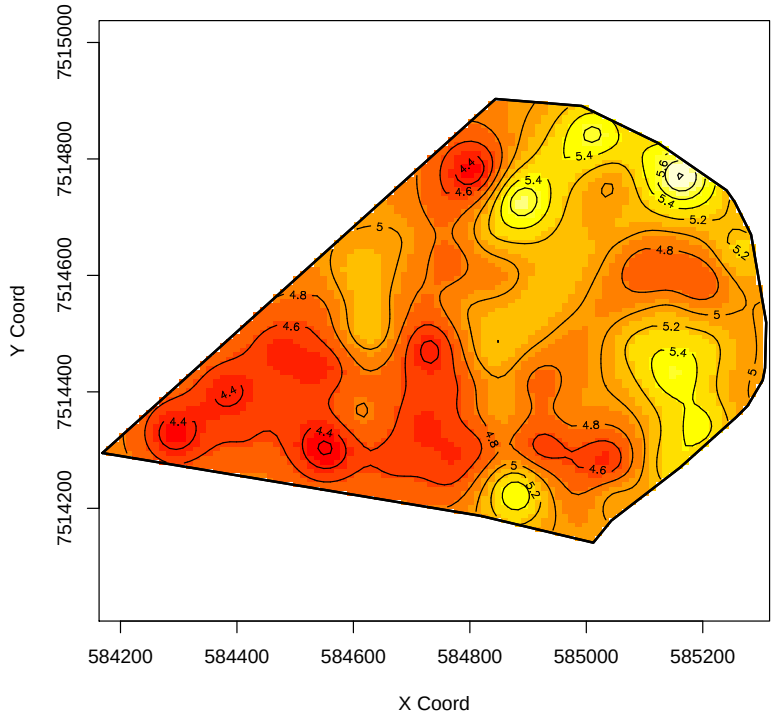
Com tendência na coordenada X, tem-se que os resultados acima, independente dos valores fixados para κ, se aproximaram bastante, logo, devemos escolher entre um desses modelos, os com κ igual a 1.5 e 2 são os melhores e se assemelham muito, uma vez que os máximos de verossimilhança são maiores.

O próximo passo é escolher entre os modelos com mais ou menos parâmetros na média, ou seja, devemos fazer a seleção de covariáveis importantes ao modelo. Para tal, não se pode comparar os máximos das funções de verossimilhança, uma vez que, os valores das mesmas são alterados conforme o número de parâmetros no modelo, sendo assim, como os modelos possuem números de parâmetros de média distintos, se deve utilizar outro critério para seleção, sendo assim, será utilizado o critério da informação de Akaike, esse critério faz uma ponderação entre a explicação do modelo e o número de parâmetros usados, ou seja, esse critério é uma espécie de punição ao modelo pelo número de parâmetros utilizados para explicar uma determinada variabilidade, logo, quanto menor o valor da estatística melhor o modelo:

Tendência	κ	AIC
$Constante$	1.5	59.147
$Constante$	2	58.902
$Constante$	3	58.612
$Coord.X$	1	53.722
$Coord.X$	1.5	53.429
$Coord.X$	2	48.727

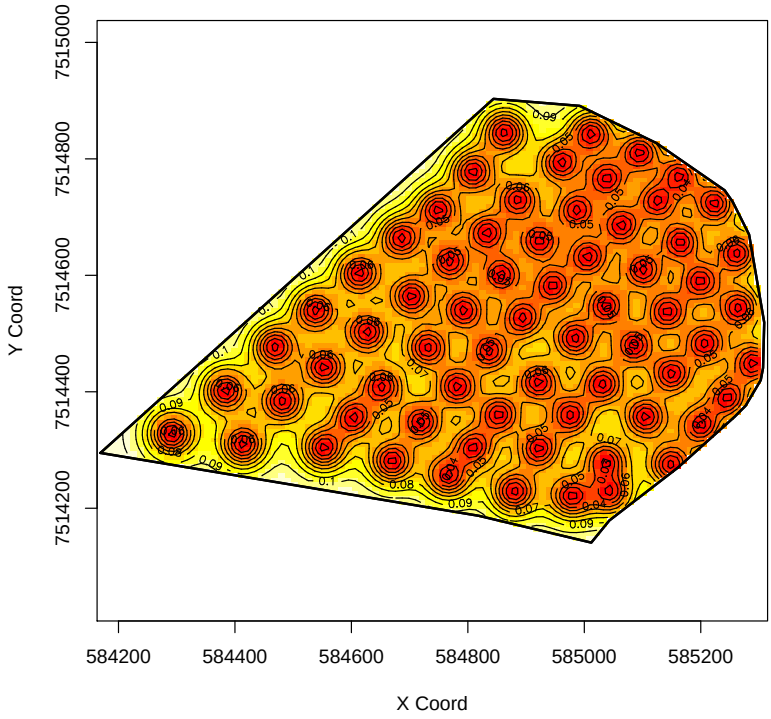
Tabela 10: Critério de informação de Akaike

Com os resultados acima, tem-se que todos os modelos com tendência a coordenada X foram melhores, pois obtiveram menor AIC, do que os modelos sem tendência na média, logo, o modelo com tendência na coordenada X e κ igual a 2 é o que melhor se ajustou aos dados, sendo assim, o próximo passo é fazer a predição ou krigagem para todo o espaço da fazenda, sendo assim, com a estimação dos parâmetros feita, é possível prever o valor do campo aleatório para localizações não amostradas, essa predição é feita através da média estimada para a localização ponderada pelos valores estimados para a variância e covariância do campo aleatório. Segue o gráfico da krigagem:



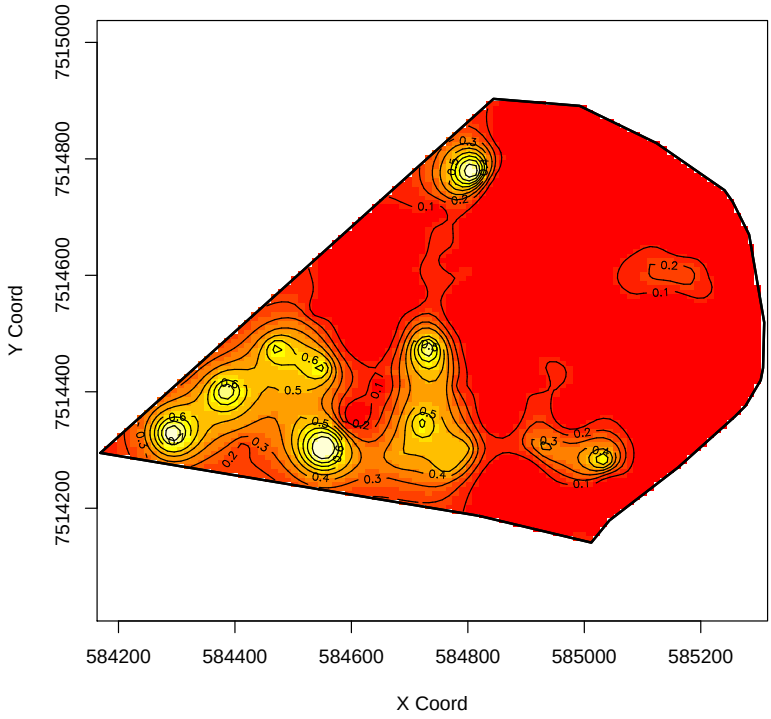
No gráfico acima as cores mais próximas do branco indicam valores mais elevados para a saturação por bases e conseqüentemente cores próximas do vermelho indicam valores menores para a saturação. Analisando os valores observados nas localizações amostradas, tem-se que as predições se aproximaram refletiram os valores reais e suavizou para o restante do espaço o campo aleatório.

O próximo passo é analisar o gráfico das variâncias das estimativas do campo aleatório, segue o gráfico:



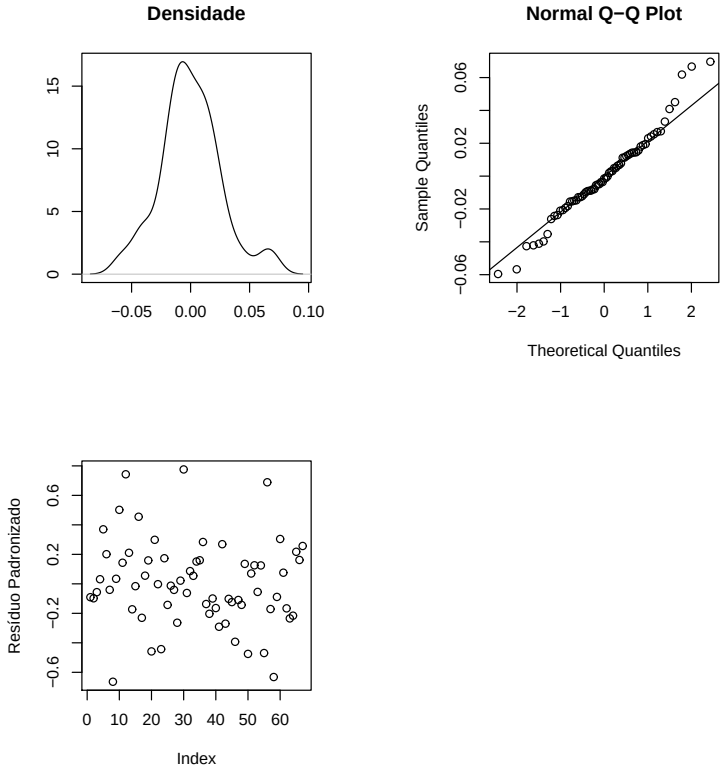
Com o gráfico acima tem-se que existe o padrão espacial esperado para as variâncias, ou seja, as variâncias mudam no espaço todo, mas essa mudança é suave, o que de certa forma corrobora a idéia de estacionariedade da função de covariância, além disso fica bem claro que valores de variância onde foi feita amostra são menores do que localizações não amostradas.

Com a definição dos parâmetros estimados, e com os valores preditos para cada posição do grid definido, é possível fazer análises a partir da distribuição normal de cada posição onde foi feita predição, assim é possível calcular a probabilidade prevista de que a saturação por bases seja menor que 4.5 em cada posição predita, ou seja, se pode fazer um gráfico das probabilidades de cada valor predito ser inferior a 4.5, que é um valor abaixo do esperado para uma boa qualidade do solo da fazenda sob estudo, abaixo segue o gráfico:



Com o gráfico acima tem-se que valores preditos em regiões com menor valor para a coordenada X possuem maior probabilidade do ph ser inferior a 4.5, informação que corrobora a análise descritiva onde se tem a clara impressão de que a média da saturação é inferior em regiões com menores coordenadas X.

Uma outra análise que pode ser conduzida é a validação do pressuposto de normalidade univariada do ruído branco, abaixo seguem análises gráficas do tipo:



O primeiro gráfico acima mostram que a densidade dos resíduos brancos estimados se aproximam de uma distribuição normal, além disso o qqplot mostra que existe uma pequena fuga da normalidade na cauda superior dos resíduos, mas nada que afete o pressuposto e por último não existe valores para os resíduos padronizados fora do intervalo de [-1,1], o que indica não existir valores discrepantes dos erros brancos. Após a análise gráfica o teste de normalidade de Shapiro-Wilk foi conduzido, o qual gerou um p-valor de 0.2185, logo a normalidade não é rejeitada e esse pressuposto está atendido.

4.2.3 Modelos bivariados

Nas subseções anteriores foram apresentadas as abordagens univariadas, no entanto, como mencionado nos materiais e métodos, a medição da saturação por bases é muito mais cara do que a medição do Ph, e além disso há suspeita de que as variáveis são fortemente correlacionadas, pois valores de Ph entre 5 e 7 fornecem uma maior captação de substâncias boas pelo solo, que pode ser medida através da saturação por bases, e como na amostra não foi detectados valores de Ph acima de 7, não existe distorção e a interpretação é que quanto maior o Ph maior é saturação

por bases do solo. A correlação de Pearson entre essas duas variáveis é aproximadamente igual a 0,92.

Logo, confirmada a correlação entre as variáveis, um modelo bivariado pode ser proposto para identificar a estrutura de correlação entre as respostas, o que em medições futuras do mesmo terreno podem diminuir os gastos de coleta com saturação por bases, que cada vez pode ser coletada em menos pontos, pois a predição pode ser feita através da amostragem maior da variável Ph.

Nesse contexto as análises univariadas serão utilizadas como exploração dos dados e valores iniciais para procedimentos numéricos.

4.2.3.1 Modelo de co-regionalização

4.2.3.2 Modelo hierarquico condicional

4.2.3.3 Modelos com componente de correlação parcialmente comum

ARRUMAR ESTE TÓPICO

Sendo assim, estimadores de máxima verossimilhança serão utilizados na estimação dos parâmetros, na abordagem de distribuições conjuntas, conforme visto na revisão bibliográfica, a abordagem condicional fica de tópico para estudos futuros. Outro tópico que fica para estudos futuros é a seleção de modelos, uma vez que, os softwares estatísticos ainda não possuem muitos recursos para modelagem espacial bivariada, algumas pressuposições não são verificadas e além disso ainda não é possível colocar tendências nas médias dos processos ou ainda usar abordagens bayesianas de estimação. Abaixo segue um modelo estimado, o que gerou maior valor para o máximo da função de verossimilhança, cabe ressaltar que, os κ's foram fixados em 1 para a parte da função de correlação comum aos dois modelos, em 3 para a função de correlação da saturação por bases e 2 para a função de correlação do Ph:

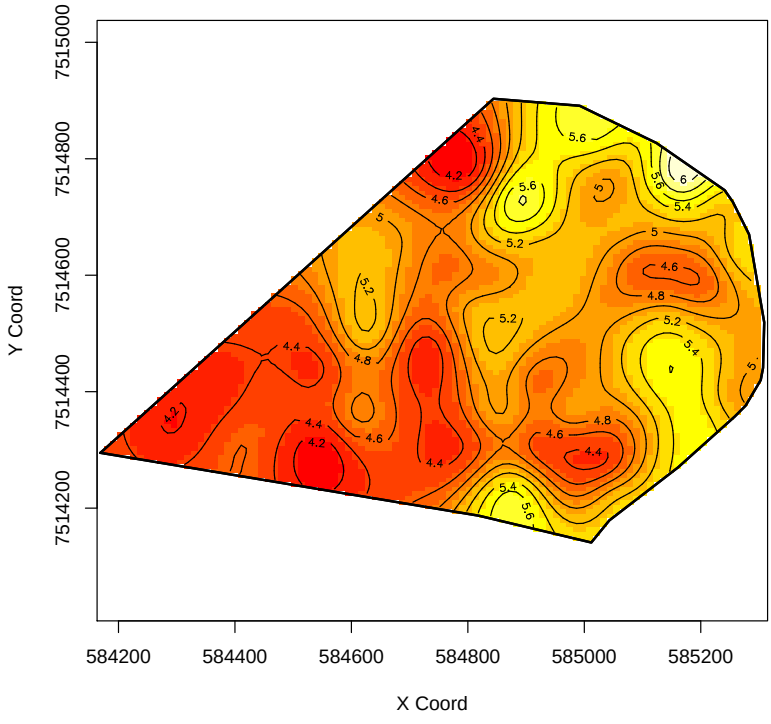
Os valores acima mostram que as médias convergiram para valores similares aos da abordagem univariada, com relação a variabilidade a maior parte da variabilidade estão sendo explicada pelo fator comum das matrizes de correlação, o que corrobora a idéia de que a correlação entre as variáveis é muito elevada, no entanto, os valores estimados para os σ's são mais inflacionados do que o esperado. Outra característica importante é que diferentes chutes inici-

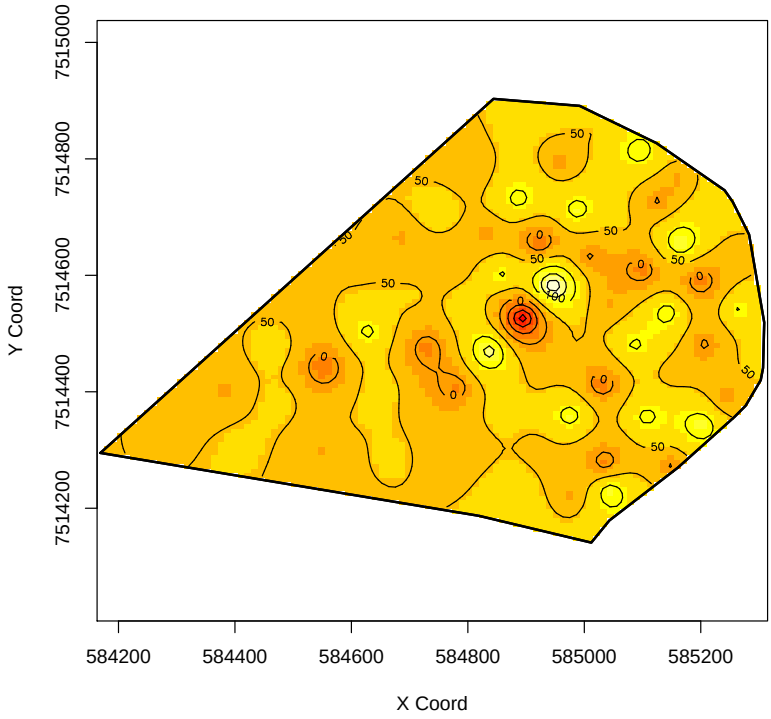
Parâmetro	Estimativa
$\mu do Ph$	4.961
$\mu da Sat.$	52.96
σ_{01}	0.5581
σ_1	0.0029
σ_{02}	257.89
σ_2	57.387
ϕ_0	79.948
ϕ_1	136.144
ϕ_2	41.761

Tabela 11: Estimativas de máxima verossimilhança

ais ao parâmetros estão gerando valores estimados, em alguns casos, muito distintos, isso pode ocorrer devido a não baixa identificabilidade do modelo, ou muitos parâmetros, ou falta de utilização de mais métodos de otimização ou ainda deveria ser tentada a abordagem condicional de modelagem.

Mesmo com os problemas citados acima, tem-se mapas de krigagem para as variáveis marginalmente e uma comparação com a abordagem univariada é pertinente, segue:





Os gráficos acima mostram que para o Ph quase não houve distorção dos resultados obtidos na abordagem univariada, no entanto, com relação a variável de saturação por bases no solo, a distorção foi elevada, onde a maior parte do campo recebeu previsão próxima a 50 e ainda existe um valor previsto acima de 100, o que não pode existir nesse tipo de variável, fato que corrobora a variabilidade estimada inflacionada, essa distorção pode ser reflexo de alguns pares de pontos que não estão de acordo com a correlação observada no conjunto de dados, todas essas características revelam que esse modelo não está bem ajustado e estudos mais aprofundados devem ser levados em consideração.

5 CONSIDERAÇÕES FINAIS

BANERJEE, S.; GELFAND, A.E. Predict, Interpolation and regression for spatial misaligned data points. **Sankhya**, v.64, p.227-245, 2002.

BROWN, P.J.; LE, N.D.; ZIDEK, J.V. Multivariate spatial interpolation and exposure to air pollutants. **The Canadian Journal of Statistics**, v.22, p.489-509, 1994.

CRESSIE N.; HUANG, H-C. Classes of Non-Separable, Spatio-temporal stationary covariance functions. **Journal of the American Statistical Association**, Alexandria, v.94, p.1330-1340, 1999.

DIGGLE, P.J.; RIBEIRO Jr., P.J. **Model-Based geostatistics**. New York: Springer, 2006. 230p.

ELMATZOGLOU, I. **Spatio-temporal geostatistical models, with an aplication in fish stock**, 2006. 53 p. Submited for the degree of (Master in statistics) - Lancaster University, Lancaster, 2006.

FERNANDES, M.V.M. **Modelos para Processos Espaço-Temporais Inflacionados de Zeros**, 2006. 128 p. Dissertação (Mestrado em Estatística) - Instituto de Matemática da Univeridade Federal do Rio de Janeiro, 2006.

FUENTES, M.; SMITH, R.L. **A new class of stationary spatial models**. North Caroline: Department of Statistics, North Caroline State University, 2001. Technical (Report, 2534).

GELFAND, A.E; SCHMIDT, A.M.; BANERJEE S.; SIRMANS, C.F. Nonstationary multivariate process modeling through spatially varying coregionalization. **Sociedad Española de Estadística e Investigación Operativa - Test**, v.13, p.263-312, 2005.

GOMES, E.M.C. **Modelos de regressão com resposta bivariada através de cópulas: Análise de sensibilidade e resíduo**, 2007. 101 p. Dissertação (Mestrado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura ″Luiz de Queiroz” da Univeridade

de São Paulo, 2007.

GNEITING, T. Nonseparable, Stationary Covariance Functions for Space-Time Data. **Journal of the American Statistical Association**, Alexandria, v.97, p.590-600, 2002.

HIGDON, D. **Quantitative methods for current environmental issues**, Chichester: Wiley, 2002. 185 p.

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

MARDIA, K.V.; GOODALL, C.R. Spatial-temporal analysis of multivariate environmental monitoring data. In: G. P. PATIL and C. R. Rao, eds., **Multivariate Environmental Statistics**, p.347-386, 1993.

MATÉRN, B. **Spatial variation**. Verlag, Berlin: Spinger, 1986. 365 p.

SAMPSON P.D.; GUTTORP, P. Nonparametric estimation of nonstationary spatial covariance structure. **Journal of American Statistical Association**, Alexandria, v.87, p.108-119, 1992.

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

SCHMIDT, A.M.; O'HAGAN, A. Bayesian inference for nonstationary spatial covariance structure via spatial deformations, **Journal of Royal Statistical Society**, Oxford: v.65, p.743-758, 2003.

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

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.

SUN, W.; LE, N.D.; ZIDEK, J.V.; BURNETT, R. Assessment of a bayesian multivariate interpolation approach for health impact studies. **Environmetrics**, Washington, v.9, p.565-586, 1998.

ANEXOS