EDSON ANTONIO ALVES DA SILVA

APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS MULTIVARIADOS EM PROBLEMAS DE MAPEAMENTO DE VARIÁVEIS DO SISTEMA SOLO-PLANTA

CURITIBA

JUNHO 2008

EDSON ANTONIO ALVES DA SILVA

APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS MULTIVARIADOS EM PROBLEMAS DE MAPEAMENTO DE VARIÁVEIS DO SISTEMA SOLO-PLANTA

Tese apresentada ao Curso de Pós-graduação em Métodos Numéricos em Engenharia do Setor de Tecnologia do Centro de Estudos de Engenharia Civil Professor Inaldo Ayres Vieira da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Ciências.

Orientador: Prof. PhD. Paulo Justiniano Ribeiro Jr.

CURITIBA

JUNHO 2008

TERMO DE APROVAÇÃO

EDSON ANTONIO ALVES DA SILVA

APLICAÇÃO DE MÉTODOS GEOESTATÍSTICOS MULTIVARIADOS EM PROBLEMAS DE MAPEAMENTO DE VARIÁVEIS DO SISTEMA SOLO-PLANTA

Dissertação aprovada como requisito parcial para obtenção do grau de Doutor em Ciências, pelo Programa de Pós-Graduação em Métodos Numéricos em Engenharia do Setor de Tecnologia do Centro de Estudos de Engenharia Civil Professor Inaldo Ayres Vieira da Universidade Federal do Paraná, pela seguinte banca examinadora:

Prof. PhD. Paulo Justiniano Ribeiro Jr.
Universidade Federal do Paraná

Prof. Dr. Eduardo Godoy de Souza
Universidade Estadual do Oeste do Paraná

Prof. Dr. Antonio Carlos Andrade Gonçalves
Universidade Estadual de Maringá

Prof. Dr. Julio Eduardo Arce
Universidade Federal do Paraná

Prof. Dr. Joel Mauricio Correa da Rosa
Universidade Federal do Paraná

Curitiba, 16 de junho de 2008

A meus filhos, Denise e André e a minha querida esposa Maria Elizabeth.

AGRADECIMENTOS

Agradeço ao Prof. Paulo por ter aceito a tarefa de me orientar nessa jornada e principalmente pelos seus ensinamentos e seu exemplo de atitude profissional, característica de pessoas especiais que contribuem para o desenvolvimento de toda uma sociedade.

Aos meus amigos e colegas do LEG que sempre estiveram pacientemente dispostos e presentes nas minhas necessidades e limitações. Particularmente agradeço a Ana Beatriz que trouxe sua bondade como lenimento. Poderei um dia retribuir?

A Maristela que, com sua paciência, competência e eficiência, fez-me sentir em casa.

A todos os professores do PPGMNE que, com suas competências e sensibilidades às dificuldades dos alunos, transmitiram galhardamente novos e suficientes ensinamentos.

Aos meus pais por me mostrarem que, se acreditarmos e trabalharmos com afinco, as soluções aparecem.

À Modo Battistella Reflorestamentos S/A – MOBASA, por disponibilizar dados de seus registros operacionais que enriqueceram a pesquisa.

À Capes - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior por seu apoio financeiro.

Aos meus alunos, motivo de tudo isso.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

AP	Agricultura de Precisão
GPS	*Global Positioning System*
SIG	Sistema de Informações Geográficas
MV	Máximo do logarítmo da função de verossimilhança
MVR	Máxima verossimilhança restrita
ACP	Análise de Componentes Principais
CEP	Coeficiente de efeito pepita
BLUE	*Best Linear Unbiased Estimator*
NIT	Núcleo de Inovações Tecnológicas
Unioeste	Universidade Estadual do Oeste do Paraná
COODETEC	Cooperativa Central Agropecuária de Desenvolvimento Tecnológico e Econômico Ltda
UFPr	Universidade Federal do Paraná
MOBASA	Modo Battistella Reflorestamento S/A
UTM	*Universal Transverse Mercator*
GPL	*General Public Licence*
GNU	GNU Operating System
IBGE	Instituto Brasileiro de Geografia e Estatística
CONAB	Companhia Nacional de Abastecimento
CV	Coeficiente de Variação

Resumo

Os grãos são comodities de grande importância internacional, amplamente negociadas entre importadores e exportadores. A madeira é fundamental nas atividades industriais do Brasil, seja como insumo, seja como geradora de energia. Neste início de milênio, soja, cana-de-açúcar, milho e madeira têm ocupado espaço na substituição da produção de energia de origem petroquímica e consequentemente tem aumentado sua demanda pela competição com a produção de alimentos.

Além da expansão de fronteiras agrícolas, novas tecnologias têm surgido para dar suporte ao aumento da produtividade, viabilidade econômica e preservação do habitat.

Novos conceitos vão sendo estabelecidos e a agricultura de precisão é um dos que mais se desenvolve. Ela propõe a identificação e o manejo de zonas agrícolas de característica uniforme, onde se pode dar um tratamento mais específico, evitando-se, por exemplo, subdosagens ou superdosagens de insumos. Na identificação dessas zonas de manejo, os mapas temáticos têm função de destaque. Sua elaboração requer metodologias próprias onde a geoestatística tem cumprido seu papel. Muitos estudos são realizados e importantes resultados têm levado a mapas que expressam, com qualidade, a distribuição espacial dos valores das variáveis georreferenciadas.

Nas pesquisas em que são aplicados métodos geoestatísticos é comum a coleta de um conjunto de variáveis que descrevem propriedades físicas, químicas e de produção e com posição de coleta de dados referenciadas espacialmente. Muitos trabalhos envolvem o estudo e elaboração de mapas de uma única variável por vez. Neste trabalho foram produzidos mapas em um contexto multivariado. Apesar das baixas correlações dessas variáveis reportadas na literatura, foi feita uma análise de componentes principais para a redução do conjunto de variáveis suporte à sua primeira componente, incorporando sua informação de variabilidade espacial a outra variável de interesse principal, em uma estrutura bivariada de modelo, para qual se dispunha de recursos computacionais para resolver numericamente aplicações.

Adotou-se uma formulação com base em modelos mistos. Seus parâmetros foram estimados pela otimização de funções de verossimilhança e por simulação bayesiana para a obtenção de distribuições a *posteriori*. Com esses modelos foram derivados preditores marginais e condicionais, permitindo-se estimar valores em pontos de uma estrutura compatível com a apresentação em forma de mapa.

Foram analisados dados de dois problemas distintos em sua natureza. Em um dos casos os dados reportavam variáveis de produtividade de soja associados a dados físicos e químicos, distribuídos em uma área de 1,74 ha cultivada em sistema de plantio direto. No outro, as variáveis representavam o incremento médio anual – IMA de *Pinus taeda* L. em área de 2.252,11 ha de reflorestamento e dissociadamente dados de teor de argila.

A análise revelou a capacidade do modelo bivariado em explorar a informação espacial contida nos dados quando as localizações das variáveis não eram as mesmas e a capacidade de identificar zonas de regionalização onde o modelo univariado não o fez, principalmente em se tratando de amostras pequenas da variável de interesse principal. A análise mostrou ainda que o método bayesiano, no caso univariado, define melhor as zonas diferentes quando se trata de delineamentos com poucas amostras.

Palavras-chave: Geoestatística multivariada, geoestatística bayesiana, verossimilhança, agricultura de precisão, inventário florestal.

Abstract

The grains are great international importance comodities, widely negotiated between importers and exporters. Already the wood is basic in the industrial activities of Brazil, either as manufactures material, either as generating of energy. In this beginning of milênio, soy bean, sugar cane, maize and wood have ocupaited space in the production of petrochemical energy and then it has increased its demand for the competition with the food production.

Beyond the expansion of agricultural borders, new technologies arrived to increase productivity, to guarantee economic viability and preserve the habitat.

New concepts go appearing and the Precision Farming is of that more it is developed. It propose the identification and the handling of uniform agricultural zones where it can give a more specific treatment, preventing for example, overdoses or subdoses of chemical material. In the identification of these handling zones, the thematic maps have prominence function. Its elaboration requires proper methodologies where the geostatistics fulfilled its paper. Many studies are carried through and important resulted they have taken the maps that they express, with quality, the spatial behavior of the geographically marked variables.

In the geostatistics research of agricultural problems with a set of variable that describe physical, chemical and production properties is common. Many works involve the study and elaboration of maps of one variable for each time. In this work it was produced maps in a multivariate context. Although the low correlations between agricultural variable was reported in literature, was made analysis of main components for reduction of the set of variables to its first component, having incorporated its space variability information to another variable of main interest, in a bi variated structure of model, because computational resources availability for numerical applications.

A mixed model-based formularization was adopted where its parameters had been estimated by the optimization of likelihood functions and by Bayesian simulation. With these models it was derived conditional and marginal predictors, allowing itself estimate values in localization where the structure was compatible with a map.

In this work had been analyzed data from two distinct problems in its own nature. In one of the cases the data reported variable of soy productivity associates the physical and chemical data, distributed in a 1,74 ha area cultivated in system of direct plantation. In the other, the variable represented the Annual Average Increment of *Pinus taeda* L. in area of 2.252,11 ha of reforestation and separately it Clay Concentration data.

The analysis showed the capacity of the bivariate model, where the localizations of the variable were complementary, to identify zones where the univariate model didn't make it, mainly in if treating to small samples of the variable of primary interest. The analysis showed that the Bayesian method, in the univariate case, defines the different zones better when if it

deals with samples of small size.

Key-words: Multivariate Geostatistics, Bayesian Geostatistics, Likelihood, precision agriculture, forest inventory.

1 INTRODUÇÃO

A grande explosão demográfica que acompanha o desenvolvimento da espécie humana tem exigido cada vez mais um significativo aumento na produção e distribuição de alimentos, pois saciar a fome é uma das necessidades mais primárias do ser humano. A atividade agrícola atual, em geral, não tem conseguido oferecer alimentos em quantidade suficiente e simultaneamente preservar o meio ambiente, além de ter que produzir combustíveis renováveis, devido à escassez e ao dano provocado por combustível de origem fóssil. Os resultados das pesquisas científicas não atingem, em grande escala, a consciência do produtor rural, que em muitos casos é ávido pelo lucro rápido e sem riscos econômicos.

É incorreto pensar que as fronteiras agrícolas se estabelecem nos limites de cada propriedade rural. O ecossistema é um ambiente altamente correlacionado onde os recursos disponíveis em um local específico decorrem das transformações ao longo de milhares de anos de evolução e de desenvolvimento do globo terrestre. Uma propriedade rural não representa um sistema fechado. Os insumos aplicados tendem a se distribuir além de seus limites geográficos. Os recursos naturais demandados em um dado momento, sem controle ou critério, podem levar posteriormente à sua falta ou mesmo o esgotamento definitivo, não só naquela propriedade, como também em toda uma região. Se considerarmos os recursos naturais compartilhados, como os recursos hídricos por exemplo, então um manejo isolado em uma propriedade poderá produzir consequências danosas às outras propriedades ou mesmo ao meio-ambiente local.

Tome-se como exemplo o Estado do Paraná, que tem sido historicamente um grande produtor de grãos do Brasil, com grande potencial econômico e agrícola e uma localização privilegiada em relação ao Mercado Comum do Sul - MERCOSUL. Sua região Oeste é responsável por aproximadamente um terço da produção de grãos do Estado, tendo sua economia baseada principalmente na produção de soja e trigo, com muitas propriedades disputando simultaneamente os recursos naturais da respectiva região. Outro exemplo na linha de agronegócios

se dá na região Nordeste do Estado de Santa Catarina, particularmente nos municípios de Rio Negrinho e Doutor Pedrinho onde juntos dispõem de 232 indústrias ligadas ao setor madeireiro, abastecidas por grandes áreas de reflorestamento de pinus e eucalipto desenvolvidas na região, interferindo na economia e no meio-ambiente dessas duas cidades.

Por outro lado, a globalização da economia e a grande demanda por mais alimentos exigem que a agricultura brasileira desenvolva tecnologias que possibilitem a competição de nossos produtos no mercado mundial e um aumento de produtividade para atender o crescimento populacional. Esse aumento é normalmente controlado pelo aumento do uso dos insumos agrícolas. Estes insumos compreendem principalmente os insumos biológicos, insumos mecânicos, água e insumos químicos. O uso de insumos químicos tem sido identificado como o principal fator de contaminação da água e do solo (BAKHSH et al., 1997). Deduz-se, portanto, que eles, ao mesmo tempo em que auxiliam no aumento da produtividade agrícola, apresentam grande perigo para o solo e mananciais de água.

Após o advento do plantio direto, revolucionando o manejo agrícola com o desenvolvimento de novos equipamentos, oferecendo resultados importantes para o desempenho do cultivar e para o meio-ambiente, surge o conceito de Agricultura de Precisão (AP). Inicialmente o objetivo era uniformizar manchas ou zonas de produção diferenciadas nivelando áreas pelas altas produtividades. Acreditava-se que as operações de manejo ou semeadura em taxas variáveis seria suficiente para uniformizar a produtividade no nível do rendimento máximo possível. A aplicação da AP nas propriedades agrícolas requeria o uso de tecnologias emergentes que fosse capaz de discriminar, a uma resolução refinada, a variabilidade espacial dos diversos fatores associados à produção e orientar, com o auxílio de aparelhos dotados de um Sistema de Posicionamento Global, popularmente conhecido por GPS (acrônimo do inglês *Global Positioning System*), um sistema mecanizado para aplicar insumos otimizadamente. A tecnologia SIG – Sistema de Informações Geográficas, associada com os dados agrícolas geram uma grande quantidade de informações, expressas na forma de mapas temáticos e relatórios de apoio a decisão no manejo agrícola. Mesmo assim, visando aumentos progressivos de produtividade, os agricultores utilizavam o máximo de fertilizantes e corretivos (MOLIN, 1997).

O mapa de produtividade é, ainda nessa primeira década do terceiro milênio, essencial para quem pretenda entender ou praticar as prerrogativas da AP, pois é ele quem mostra as variações de produtividade em uma área. A idéia básica continua sendo identificar zonas ou talhões de alta ou de baixa produtividade e administrar essas diferenças. O conceito atual para AP, no entanto, é o de um sistema de gestão agrícola composto de tecnologias e procedimentos que levem a uma produção otimizada (MOLIN, 2002b). Para esse autor, a melhor informação do

resultado de uma lavoura é a colheita. Esse resultado pode ser expresso pela média, usualmente expressa em toneladas por hectare, ou um mapa de produtividade que mostra, entre outras informações, a produtividade obtida em cada parte da lavoura e, no seu conjunto, a variabilidade espacial da produção.

Um dos desafios mais recentes da AP é oferecer subsídios para a definição de unidades de manejo para posterior intervenção. Essa identificação pode ser feita com informações de solo, da produtividade ou indicadores compostos. Definir unidades de manejo em talhões que mereçam tratamento diferenciado é uma tarefa pouco objetiva pois depende de fatores multivariados, como a resposta da cultura, características do solo, fatores ambientais, dentre outros (MOLIN, 2002a). De qualquer maneira, a geoestatística pode contribuir muito para se obter resultados cientificamente aceitos. A AP está fundamentada basicamente na existência da variabilidade espacial dos fatores produtivos e, portanto, da própria quantidade produzida pela cultura, constituindo a sua representação gráfica uma das mais importantes ferramentas destinadas a sua análise (BALASTREIRE; ELIAS; AMARAL, 1997).

O acompanhamento do desenvolvimento de uma cultura em tempo real e a correção dos fatores deficientes no instante que é diagnosticado foi uma das metas mais importantes e ousadas da AP (CAPELLI, 1999). O que aconteceu no final da década de 90 e no início deste segundo milênio foi que, devido aos altos preços de implementos praticados decorrentes da baixa escala de produção e pequena nacionalização industrial, a adoção do manejo a taxa variável se tornou inexpressiva. O mercado também não respondeu e atualmente não existem novidades tecnológicas em equipamentos de custo economicamente viável que levem a uma solução esperada para a variabilidade espacial dos talhões (MOLIN, 2002b).

A geoestatística se apresenta como um método que utiliza procedimentos estatísticos aplicados a problemas cujos dados provêm de fenômenos naturais e que são espacialmente distribuídos e auto correlacionados, ou seja, consideram não só o valor obtido para uma determinada variável, mas também sua posição, expressa por um sistema de coordenadas. Assim, o comportamento do evento estatístico pode ser descrito pelas diferenças entre as informações obtidas em função da distância que as separa. O valor de uma variável em uma determinada posição poderá ser estimado a partir dos valores em posições vizinhas. Atualmente a noção de métodos geoestatísticos é popular em muitas áreas das ciências e da indústria para se avaliar dados correlacionados no espaço e/ou no tempo.

Tanto experimentos baseados nos conceitos de AP quanto experimentos de outras áreas que envolvem a estatística espacial, particularmente a geoestatística, usam procedimentos univariados para a representação do comportamento de suas variáveis. Entretanto, em problemas

reais, os fenômenos frequentemente ocorrem sob circunstâncias multivariadas e espacialmente correlacionados.

Existe disponível na literatura, muitos trabalhos envolvendo métodos geoestatísticos multivariados tais como os apresentados por Einax e Soldt (1998) e Caeiro et al. (2003), mas ainda cabe investigações para se determinar as condições em que uma análise multivariada para os problemas representam um ganho efetivo na qualidade dos resultados, na confiabilidade do processo, na eficiência, sobretudo na predição. Cabe espaço também para se avaliar as características dos diferentes modelos propostos, ou seja, tanto aqueles baseados em variogramas e na estrutura da matriz de correlação como aqueles baseados em modelos de regressão. Em decorrência dessas avaliações, poderão surgir novas proposições ou recomendações de estratégias de modelagem que levem a uma viabilidade computacional, grande limitação nos métodos atuais, ou ainda, ampliar a interpretabilidade dos resultados.

O objetivo geral deste trabalho foi empregar métodos geoestatísticos univariados e multivariados em experimentos agrícolas desenvolvidos na região Oeste do Paraná e em dados de área de reflorestamento no município de Rio Negrinho-SC, incorporando técnicas que permitiram a elaboração de mapas temáticos melhores com uma redução no número de amostras de variáveis de interesse principal pela sua correlação com outras variáveis agrícolas que sejam mais facilmente disponíveis.

O capítulo 2 trata de eventos geoestatísticos univariados baseados em modelos nos quais se apresenta a geometria do espaço adotado, analisa-se os componentes do modelo, associa-se cada componente do modelo a conceitos baseados em covariograma e variograma, apresenta-se as principais funções de correlação utilizadas, deriva-se os métodos de estimação dos parâmetros do modelo baseados em maximização do logarítmo da função de verossimilhança (MV), máxima verossimilhança restrita (MVR) e da determinação empírica de distribuição a posteriori dos parâmetros baseados em inferência bayesiana. Finalmente, nesse capítulo, analisam-se dois problemas de confeção de mapas temáticos e de estimação de produção.

O capítulo 3 aborda o problema geoestatístico multivariado e particularmente o caso bivariado, possível de ser implementado computacionalmente. Inicialmente foi obtido um método bivariado sob o enfoque da geoestatística baseada em modelos, propondo-se um modelo que contemplasse a demanda surgida no problema univariado. Derivou-se também uma metodologia de predição linear multivariada, particularmente a distribuição preditiva de uma variável de interesse principal, condicionada às demais. Dois casos foram estudados, um associando uma variável de interesse principal com outra de interesse secundário, em uma mesma área de cul-

tivo e outro caso associando a variável de interesse principal com a primeira componente obtida de uma análise de componentes principais (ACP) sobre um conjunto de variáveis secundárias. Foram então analisados os dados geoestatísticos do problema abordado no capítulo 2, uma vez que cada um dos dois conjuntos de dados foram tomados em uma mesma região de manejo agrícola.

No capítulo 4 apresentam-se as conclusões gerais do trabalho e sugestões para trabalhos futuros com base nos métodos desenvolvidos. Na última parte desse trabalho encontram-se em anexo os roteiros de análise para os interessados em aplicar os mesmos procedimentos em outros estudos similares.

Os arquivos fonte da análise estatística estão disponíveis na URL (*Uniform Resource Location*) http://www.leg.ufpr.br/ como complementos para artigos e materiais do LEG (Laboratório de Estatística e Geoinformação) da UFPR (Universidade Federal do Paraná) na seção *paper companions* em Atividades e Eventos. Os dados podem ser obtidos diretamente do pacote geoR (RIBEIRO JR; DIGGLE, 2001).

2 MODELO GEOESTATÍSTICO GAUSSIANO UNIVARIADO

O modelo que se idealizou neste estudo assumiu uma variável Y observada em diferentes coordenadas de um plano cartesiano bidimensional, representando uma versão de ruído de um sinal S de um processo espacial contínuo, sendo Y condicionalmente independente de S. Estruturou-se esse modelo como um modelo linear misto, especificado com notação de modelos hierárquicos, quando necessário à sua clareza. O componente associado ao valor esperado de Y comportou uma estrutura induzida por covariáveis presentes nas coordenadas de mensuração. Como essas covariáveis atuam no componente determinístico do modelo e são obtidos nas mesmas coordenadas de mensuração de Y, o processo como um todo foi aqui considerado como um processo gaussiano univariado.

2.1 GEOMETRIA DO ESPAÇO GEOESTATÍSTICO

Neste trabalho foram considerados como dados espaciais as informações observadas de um fenômeno aleatório ocorrido em um sistema agrícola, distribuído em uma região de um espaço bidimensional. Não foram estudados dados que representassem polígonos de uma região (sub-área) e nem dados que representassem processos pontuais, como a ocorrência positiva ou negativa de um atributo. Abordou-se somente dados vinculados a um processo aleatório gaussiano de variação espacial contínua e mensurável.

O formato básico dos dados geoestatísticos univariados que se adotou foi:

$$\left\{(x_i; y_i) : x_i \in \mathbb{R}^2, y_i \in \mathbb{R}, i : 1, 2, \ldots, n\right\}$$

onde:

x_i : indica a localização espacial da i-ésima coordenada em uma região do espaço bi-dimensional ($\mathbb{R}^2$);

y_i : indica uma medida escalar da variável aleatória contínua $Y = (y_1, y_2, \ldots y_n)^T$, tomada na x_i-ésima localização.

Um particular resultado y da variável Y pode ocorrer em qualquer localização x de uma região contínua. Essas localizações $x_i : i = 1, 2, \ldots, n$ formam uma malha fixa ou estocasticamente independente de Y, onde serão obtidas as medidas de y_i.

O processo gaussiano é definido como um conjunto de n variáveis aleatórias onde a distribuição finito-dimensional de qualquer subconjunto de variáveis tomadas desse conjunto, tem distribuição gaussiana multivariada com dimensão igual ao número de variáveis do subconjunto. Assim, o conjunto $\{S(x_i) : x_i \in \mathbb{R}^2; i : 1, 2, \ldots, n\}$, foi o processo estocástico gaussiano que descreveu, de maneira teórica, o comportamento do fenômeno em uma área. Esse processo deve ter uma distribuição espacialmente contínua e o evento Y deverá ocorrer segundo a sua lei de probabilidades. O modelo geoestatístico apropriado que se adotou foi baseado em um processo estocástico espacial $S(x)$, gaussiano, contínuo, que representa o fenômeno de interesse em uma área de um espaço bidimensional ou, eventualmente, em uma reta de um espaço unidimensional. Entende-se aqui o processo estocástico gaussiano univariado como sendo um modelo probabilístico definido por um conjunto de variáveis aleatórias gaussianas $\{S(x) : x \in \mathbb{R}^2\}$ em que os $S(x_i)$ são medidas de mesma natureza, que ocorrem em diferentes locais do espaço (WALLER; GOTWAY, 1965). Assim, $Y = (y_1, y_2, \ldots, y_n)^T$ é um vetor aleatório de dimensão n contendo as medidas da realização do evento. Cada y_i é representado por uma função densidade de probabilidade gaussiana e o vetor Y tem função densidade de probabilidade conjunta dada por:

$$f_Y(y) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \ \exp\left\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\} \tag{2.1}$$

em que Σ é uma matriz não singular e de posto completo e μ um vetor de médias.

Uma realização do evento Y corresponde a um conjunto de observações em n localizações distintas e fixas, onde cada resultado é, em si, o resultado de uma variável aleatória $Y_k = Y(x_k) = y_k,\ k = 1, 2, \ldots, n$. Uma realização de Y é então a ocorrência de n variáveis aleatórias gaussianas, cada uma com uma única observação e que pode ser modelada como:

$$Y(x_i) = \mu(x_i) + S(x_i) + \varepsilon_i; \quad i = 1, \ldots, n \tag{2.2}$$

em que:

- $Y(x_i)$ é uma variável aleatória contínua com distribuição normal de média $E[Y(x_i)|S(x_i)] = \mu(x_i) + S(x_i)$ e variância condicional $Var(Y(x_i)|S(x_i)) = Var(\varepsilon_i) = \tau^2$;
- $\mu(x_i) = \beta_0 + \beta_1 d_1(x_i) + \beta_2 d_2(x_i) + \ldots + \beta_p d_p(x_i)$ que pode ser representado matricial-

mente como $D\beta$ é efeito espacial externo associado a p covariáveis $d(x_i)$, diferentes de $Y(x_i)$ mas que irão depender da localização x_i. Os coeficientes β são constantes a serem determinadas. Esse componente, também é chamado de efeito sistemático e pode tornar o modelo não estacionário;

- $S(x_i)$ é o valor, na posição x_i, do processo gaussiano multivariado $\{S(x) : x \in \mathbb{R}^2\}$, com média zero, variância σ^2 e função de correlação $\rho(u_{ij}) = Corr\{S(x_i), S(x_j)\}$ onde $u_{ij} = \|x_i - x_j\|$ é a distância euclidiana que separa duas coordenadas quaisquer x_i e x_j ;
- ε_i são erros independentes e identicamente distribuídos com distribuiccão normal de média zero e variância τ^2, ou seja, $\varepsilon_i \sim \mathrm{N}(0; \tau^2)$.

A distribuição de probabilidade da variável aleatória n-dimensional Y é então:

$$Y \sim N\left(D\beta, \sigma^2 R + \tau^2 I\right) \tag{2.3}$$

em que:

- σ^2 é a variância;
- R é uma matriz de tamanho $n \times n$ cujos elementos representam as correlações entre observações feitas em diferentes localizações;
- τ^2 representa a variância do erro ε_i e
- I a matriz identidade de tamanho $n \times n$.

2.2 COMPONENTES DO MODELO

2.2.1 Componente mensurável

A variável mensurável $Y(x)$ no modelo da equação 2.2 foi suposta com distribuição gaussiana de probabilidades. Essa suposição permitiu obter a solução analítica na estimação dos parâmetros do modelo. Entretanto, para o método geoestatístico baseado em modelos, isso não é necessário, desde que se consiga escrever uma função de verossimilhança que tenha solução. Já o método bayesiano pode facilmente resolver problemas de estimação de parâmetros para modelos lineares generalizados, sem a restrição da suposição do tipo do processo estocástico envolvido. A maioria dos trabalhos com geoestatística se valem do recurso de transformar a variável resposta para se obter a gaussianiedade.

Existem muitas razões importantes, amplamente discutidas na literatura, para se transformar dados estatísticos buscando obter uma forma de distribuição próxima da distribuição normal de probabilidades. Eventos que têm evolução não linear, representados por forte assimetria na distribuição de freqüências de seus dados, requerem tranformações logarítmicas convertendo o problema em uma escala mais aditiva, levando a distribuição em direção a um comportamento mais simétrico, próximo da distribuição gaussiana.

Box e Cox (1964) apresentam um método de transformação da família potência que basicamente consiste na adequação a uma família paramétrica numa generalização empírica do modelo gaussiano, na qual a escolha da transformação mais adequada corresponde a estimar um parâmetro de transformação λ, empregando-se para isso o método MV. Uma vez escolhido λ, procede-se com a seguinte operação nos dados observados:

$$Y^* = \begin{cases} \left(\dfrac{Y^\lambda - 1}{\lambda}\right) & \text{, se } \lambda \neq 0 \\ \log Y & \text{, se } \lambda = 0 \end{cases} \tag{2.4}$$

Na área econômica, Aguirre e Faria (1996) utilizaram a transformação de Box e Cox (1964) em uma aplicação do método dos preços hedônicos na avaliação de imóveis em estudo de viabilidade econômica do programa de canalização de córregos na implantação de vias e na recuperação ambiental e social de fundos de vales, elaborado pela prefeitura da cidade de São Paulo no período de 1993 a 1994. O método dos preços hedônicos consiste em estimar preços implícitos através de atributos ambientais característicos de bens ambientais comercializados em mercados através da observação do próprio mercado no qual os bens estão inseridos. A aplicação da transformação foi necessária pois as variáveis preços e aluguel apresentaram assimetria e altos valores de curtose, sugerindo uma distribuição log-normal para os dados. Destacam no trabalho a necessidade de, no final, transformar novamente as variáveis modificadas para a sua escala original (uma anamorfose), para que os coeficientes de regressão possam ter interpretação direta.

Frasson e Molin (2006) utilizaram a mesma transformação para confirmar a gaussianiedade na elaboração de mapa geoestatístico da produtividade de soja, com dados do ano de 2005 provenientes da Fazenda Velha Lagoa da Empresa Agropecuária Dois Irmãos no município de Campos Novos Paulista-SP, em talhões de 22,8 ha. A produtividade foi medida com monitor de colheita com detecção de massa por placa de impacto e receptor GPS com correção diferencial por algoritmo interno, instalado em colhetadeira própria.

2.2.2 Componente determinístico

Segundo Waller e Gotway (1965), dois conceitos devem ser estabelecidos antes de se modelar um processo espacial: estacionariedade e isotropia. Matematicamente um processo será estacionário quando suas propriedades forem invariantes às translações em um espaço multidimensional, ou seja, a relação entre dois eventos em um processo estacionário dependerá somente de suas posições relativas. Será isotrópico quando for invariante às rotações em torno da origem de um sistema de referência, ou seja, não deverá depender da orientação do eixo que liga suas posições no espaço.

A ausência de estacionariedade na média ocorre quando existe uma variação natural própria da área, que interfere no comportamento do processo, como por exemplo: a declividade sistemática de um solo que interfere nas características de fertilidade, umidade e compactação, importantes para se avaliar a variação da produtividade de uma área. Para estudar esse efeito, pesquisadores costumam modelar a média μ como uma função das localizações *x*, destacando os efeitos de tendência por modelos de regressão polinomial e utilizando o resíduo, para então prosseguir com a análise. Modelos assim não são cientificamente explicados pois as correlações com direções definidas não dão informações sobre o processo causador do efeito.

Esse efeito que afeta a média de um processo geoestatístico pode ser modelado relativamente às suas covariáveis. É o equivalente aos fatores em uma análise estatística tradicional. Muitas pesquisas são feitas em áreas onde existem sub-áreas de características próprias que afetam o processo em estudo. Souza, Marques JR e Pereira (2004), por exemplo, desenvolveram um trabalho em Guariba-SP com o objetivo de avaliar a variabilidade espacial do pH, cálcio (Ca), magnésio (Mg) e saturação de bases (V%) em Latossolo Vermelho entroférrico sob cultivo de cana-de-açúcar. Eles classificaram a curvatura e o perfil das formas do terreno no terço inferior da encosta em dois compartimentos, um com menor variação das formas e curvaturas, predominando a forma linear e outro com maior variação, com a presença de formas linear, côncava e convexa. Concluíram que o tipo de relevo, dentre outros resultados, condiciona uma variabilidade espacial diferenciada para os atributos químicos. Outro trabalho com o uso de covariáveis é apresentado por Carvalho e Queiroz (2002). Eles concluem que o uso da altitude como covariável para a precipitação de chuvas no Estado do Paraná define bolsões, além de evitar instabilidade numérica no sistema de equações do modelo causada pela redundância de observações da variável auxiliar. O conceito é semelhante ao delineamento de experimentos em blocos, que retira do resíduo uma fonte de variação conhecida. As informações adicionais são normalmente tomadas nas mesmas coordenadas do processo principal e não são tratadas como um segundo processo, mantendo assim o aspecto univariado da análise.

De maneira um pouco mais formal, diz-se que o processo é estacionário de primeira ordem (na média) se $\mu(x_i) = \mu,\ \forall\, x_i,\ i = 1,2,\ldots n$ e estacionário de segunda ordem (na variância) se as covariâncias para cada par de coordenadas forem função somente da distância euclidiana u_{ij} e para $u_{ij} = 0$, $\rho(u_{ij}) = \sigma^2$.

2.2.3 Componente do processo gaussiano correlacionado

Assumiu-se neste estudo que o processo $S(x)$ é desconhecido, de variação contínua, com incerteza em seus parâmetros e correlacionado na região em que ocorre e poderá ser composto por processos latentes $S_k(x)$, $k = 1,2,\ldots p,\ p \in \mathbb{N}$, escalonados por σ_k^2 (RIBEIRO JR; DIGGLE, 1999). Supôs-se ainda que o domínio D em que ocorre o processo é fixo em um espaço $\mathbb{R}^2$. O fato de D ser fixo significa que os pontos amostrais não serão aleatórios. Schabenberger e Gotway (2005) dizem ser importante associar a continuidade do processo ao domínio D e não ao atributo que está sendo medido. O fato de os dados mensuráveis serem contínuos ou discretos não determinam se são do tipo geoestatístico ou não.

Um outro aspecto importante sobre o processo gaussiano é a existência de estacionariedade na sua estrutura de correlação. Um pressuposto razoável é supor que seu valor decai a medida que a distância entre as localizações aumenta, independentemente do ângulo do eixo formado entre essas localizações. Neste caso se diz que o processo é isotrópico, caso contrário o processo é anisotrópico. Essa forma de se avaliar o comportamento das correlações é chamado de efeito direcional que na sua forma mais simples, — e talvez mais comum, é chamado anisotropia geométrica. Este tipo de anisotropia ocorre quando a estrutura de covariância apresenta alongamentos e rotações em relação aos eixos das coordenadas. Desta forma pode-se caracterizar esse efeito através de dois parâmetros: o ângulo de anisotropia ψ_A que dá a direção do efeito e a razão de anisotropia $\psi_R > 1$ que dá a relação entre o eixo maior e o eixo menor da elipse formada (JOURNEL; HUIJBREGTS, 1978).

Na prática, ψ_A e ψ_R são informações desconhecidas que podem ser convenientemente incorporadas ao modelo geoestatístico para serem estimadas a partir dos dados. Uma vez conhecida a tendência devido à anisotropia, pode-se, para efeito de análise, transformar as coordenadas. Se (a,b) for a coordenada de um ponto no plano cartesiano $\mathbb{R}^2$, poder-se-á contrair/extender e/ou rotacionar esse vetor com a transformação linear (KOLMAN, 1997):

$$(a',b') = (a,b)\begin{pmatrix} cos(\psi_A) & -sen(\psi_A) \\ sen(\psi_A) & cos(\psi_A) \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & \psi_R^{-1} \end{pmatrix}$$

Os efeitos de tendência direcional e regional têm papel fundamental na análise do processo $S(x)$, pois permitem melhorar o conhecimento subjetivo do fenômeno em estudo.

Segundo Matheron (1973), um tipo de modelo não-estacionário é o modelo intrínseco. Ele considera um caminho aleatório $S(x) = S(x-1) + Z(x)$ com $Z \sim N(0,1)$, ou seja, uma função aleatória intrínseca é um processo estocástico $S(x)$ com incrementos estacionários. Assim, o processo $S_{x'}(x) = S(x) - S(x-x')$ será dito estacionário para todo $x' \in \mathbb{R}^2$.

A principal diferença entre uma predição obtida com modelo intrínseco e modelo estacionário, é que se for usado o primeiro, a predição em uma localização x será influenciada pelo ambiente local dos dados, ou seja, por observações medidas em locais próximos. Considerar uma hipótese intrínseca para os dados significa supor que as diferenças entre os valores apresentam fraco incremento, ou seja, as diferenças serão localmente estacionárias. Com o emprego de modelos estacionários, as predições serão afetadas pelo ambiente global dos dados.

Diggle, Tawn e Moyeed (1998) mostram duas aplicações em que a metodologia geoestatística convencional para resolver problemas de valores de uma função linear de um processo estocástico espacial gaussiano baseado nas observações do modelo (Equação 2.2) é inapropriada. A estrutura teórica para o método geoestatístico baseado em modelos é condicional ao processo não observado e as observações em locações amostrais formam um modelo linear generalizado com os valores correspondentes de $S(x)$ que ficam de fora em uma ação de predição.

2.3 COVARIÂNCIA E VARIOGRAMA

Diggle e Lophaven (2006) definiram a semivariância como a função:

$$V(u) = \frac{1}{2}E\left[(Y(x) - Y(x-u))^2\right]$$

de um processo estocástico espacial estacionário.

Estimar os parâmetros de $V(u)$ pelo método dos momentos consiste em obter os valores

$$v_{ij} = \frac{1}{2}(y_i - y_j)^2 \tag{2.5}$$

provenientes de dados experimentais, agrupá-los dentro de intervalos de distâncias (e ângulos quando for o caso) e ajustar um modelo de semivariância teórica ao gráfico formado pela média dos pontos de cada intervalo locados em seu centro. O comportamento padrão de um semivariograma é dado pela Figura 2.1. Nesse gráfico, a função semivariância é uma função monótona não decrescente e depende somente do comportamento da função de correlação $\rho(u)$. O efeito pepita (*nugget*) representa a variância de pequena escala τ^2. O patamar (*sill* total) dado por $\tau^2 + \sigma^2$ representa a variância total do processo e o alcance prático de dependência espacial (*range*) é determinado por um parâmetro ϕ que controla a taxa de decaimento da função de correlação. Nessa figura nota-se que o efeito pepita (τ^2) corresponde ao valor da semivariância a distâncias nulas. Entretanto, amostras medidas exatamente na mesma posição deveriam ter o mesmo valor. Quando isso não ocorre, a diferença é ser atribuída, dentre outras razões desconhecidas, ao erro de medida amostral. A semivariância pode ainda indicar descontinuidade na origem, ou seja, a ausência de valores. Isto ocorre tanto por um planejamento amostral que não considera medidas à distância nulas (repetidas) quanto pela diferenciabilidade na origem da função semivariância na origem.

Para Journel e Huijbregts (1978) o semivariograma é um gráfico muito utilizado para representar o mecanismo de dependência espacial.

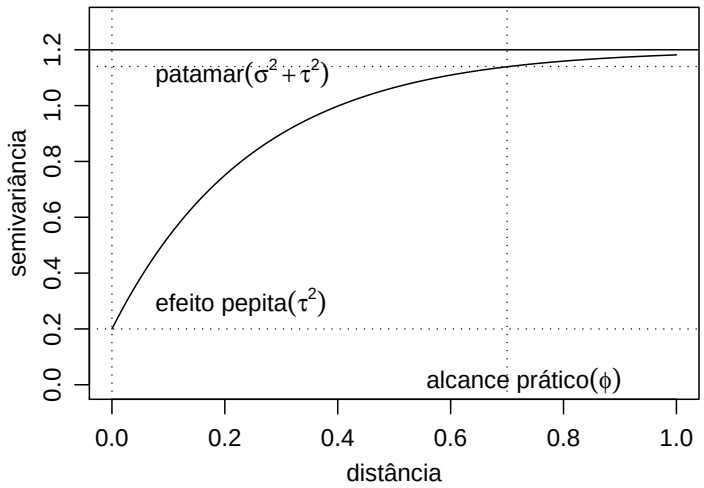

Figura 2.1: Comportamento padrão da função semivariância. Os elementos principais que a compõem são: o alcance prático proporcional a ϕ, a variância de pequena escala ou efeito pepita τ^2 e a contribuição σ^2 que corresponde à diferença entre o patamar e τ^2.

Para Tragmar, Yost e Uehara (1985) o efeito pepita mostrado na Figura 2.1 tem um papel importante na análise geoestatística, pois pode sugerir a presença de dependência espacial do processo, quando comparado ao patamar. Se os valores se aproximam, a amostra tende a não receber influência espacial. Cambardella et al. (1994) propuseram mensurar a dependência espacial pela razão percentual entre o efeito pepita e patamar ou coeficiente de efeito pepita (CEP) dado por:

$$CEP = \frac{\tau^2}{\tau^2 + \sigma^2} \times 100$$

classificando como forte para valores menores que 25%, moderada para valores entre 25% e 75% e forte para valores acima de 75%

Considerou-se o modelo dado pela Equação 2.2, supondo estacionariedade, como sendo aquele que descreve o conjunto Y das variáveis observadas de um determinado processo $S(x)$. Desta forma, y_i e y_j são observações tomadas em quaisquer duas localizações separadas por uma distância u_{ij}. Então, $Var(y_i - y_j)$ registra a variação da diferença dos valores medidos separados por essa distância. Fixando $\mu = 0$ e $\tau^2 = 0$ vem:

$$\begin{aligned} Var(y_i - y_j) &= Var(y_i) + Var(y_j) - 2\,Cov(y_i; y_j) \\ Var(y_i - y_j) &= Var(S(x_i) + \varepsilon_i) + Var(S(x_j) + \varepsilon_j) - 2\,Cov\big(S(x_i); S(x_j)\big) \end{aligned} \tag{2.6}$$

Como $S(x)$ e ε são independentes, então:

$$\begin{aligned} Var(y_i) &= Var(S(x_i) + \varepsilon_i) = Var(S(x_i)) + Var(\varepsilon_i) = \sigma^2 \\ Var(y_j) &= Var(S(x_j) + \varepsilon_j) = Var(S(x_j)) + Var(\varepsilon_j) = \sigma^2, \end{aligned}$$

e assim:

$$Var(y_i) = Var(y_j) = \sigma^2 \tag{2.7}$$

Em estatística clássica, o coeficiente de correlação de Pearson (ρ) mede o grau e a direção (positiva ou negativa) da correlação linear entre duas variáveis (MONTGOMERY; PECK, 1955). Se aplicada no contexto da geoestatística utilizando-se o resultado obtido pela Equação 2.7 tem-se:

$$\rho(u_{ij}) = \frac{Cov(y_i; y_j)}{\sqrt{Var(y_i) Var(y_j)}} = \frac{Cov(y_i; y_j)}{\sqrt{\sigma^2 \sigma^2}} = \frac{Cov(y_i; y_j)}{\sigma^2},$$

o que implica que:

$$Cov(y_i; y_j) = \sigma^2 \rho(u_{ij}) \tag{2.8}$$

Nota-se pela Equação 2.8, caso a hipótese de estacionariedade não seja rejeitada, que

a correlação entre dois valores medidos de Y irá depender somente da distância que os separa. Esta função será monótona decrescente, restrita a $\rho(0) = 1$ e $\lim_{u\to\infty} \rho(u) = 0$ para $u > 0$.

Assim, substituindo-se os resultados das Equações 2.7 e 2.8 na Equação 2.6 obtém-se:

$$\begin{aligned} Var(y_i - y_j) &= (\sigma^2 + \tau^2) + (\sigma^2 + \tau^2) - 2\,\sigma^2\rho(u_{ij}) \\ Var(y_i - y_j) &= 2\,\tau^2 + 2\,\sigma^2(1 - \rho(u_{ij})) \\ Var(y_i - y_j) &= 2\left(\tau^2 + \sigma^2(1 - \rho(u_{ij}))\right) \\ \tfrac{1}{2}Var(y_i - y_j) &= \tau^2 + \sigma^2(1 - \rho(u_{ij})) \end{aligned}$$

Fazendo $\frac{1}{2}Var(Y_i - Y_j) = \gamma(u_{ij})$ a semivariância teórica fica dada por:

$$\gamma(u_{ij}) = \tau^2 + \sigma^2(1 - \rho(u_{ij})) \tag{2.9}$$

A Figura 2.2 mostra o comportamento gráfico da função semivariância onde pode-se notar o papel funtamental da função de correlação pois é ela que define no modelo, a forma com que as correlações decaem com o aumento da distância entre as coordenadas. Segundo Diggle e Ribeiro Jr (2007), sendo o processo estacionário, a função semivariância é o equivalente teórico para a função covariância, com a vantagem de oferecer elementos para a análise da estrutura espacial dos dados.

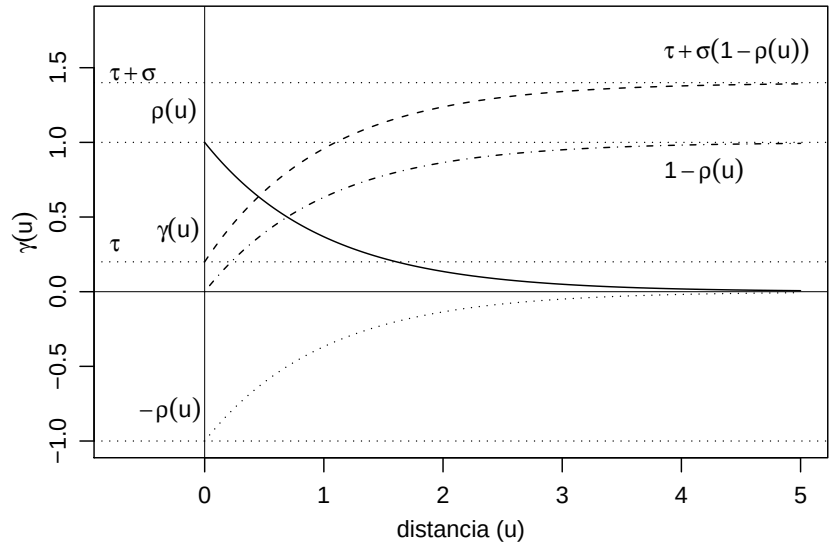

Figura 2.2: Etapas da transformação da função de correlação (linha contínua) para a função semivariograma (linha tracejada).

2.4 TIPOS DE MODELO DE CORRELAÇÃO ESPACIAL

A escolha da estrutura de correlação desempenha um papel decisivo no modelo geoestatístico pois esta escolha irá afetar diretamente a suavidade da imagem gerada. É ela que estabelece o comportamento de uma característica pontual em sua vizinhança. As medidas matemáticas aceitas para se avaliar essa suavidade são a continuidade e a diferenciabilidade. Bartlett (1955) afirma que um processo estocástico estacionário com função de correlação $\rho(u)$ será *k*-vezes diferenciável se, e somente se, $\rho(u)$ for 2 *k*-vezes diferenciável na origem.

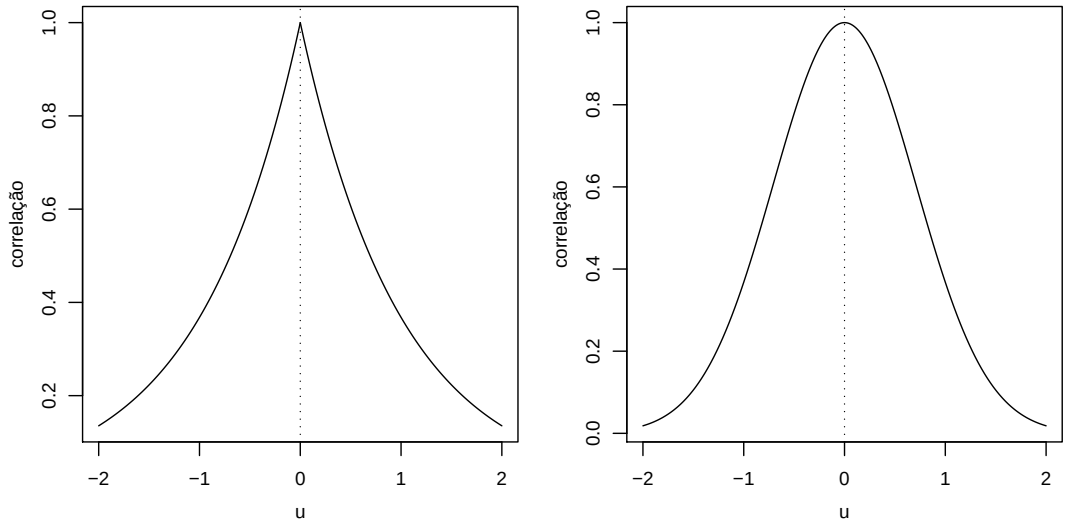

Figura 2.3: O gráfico da esquerda corresponde ao comportamento da função de correlação poder de ordem 1 ($exp(-u)$) onde a função no ponto $u = 0$ não é diferenciável. O da direita corresponde a mesma função de correlação exponencial "poder" de ordem 2 ($exp(-u^2)$), diferenciável em $u = 0$.

Na Figura 2.3 tem-se o comportamento básico da diferenciabilidade da função de correlação. Ambas as figuras ilustram o caso de funções contínuas em todo o domínio das distâncias *u*, o que é mais freqüentemente adotado, embora possam ocorrer descontinuidades na origem. A figura da esquerda apresenta um ponto "problema" que é o ponto $u = 0$ onde a função não é diferenciável, dado que da teoria do cálculo sabe-se que a derivada de uma função não existe onde a tangente ao ponto é vertical. Já a figura da direita mostra uma função contínua e diferenciável em todo o seu domínio.

O processo $S(x)$ é desconhecido e tipicamente, não observável diretamente. Assim, a experiência do pesquisador com o fenômeno estudado deve ser usada para uma boa escolha do modelo de correlação espacial. Se o evento em questão tiver variações mais abruptas, modelos com números menores de derivadas deverão ser preferidos e se tiver variações mais suaves, utiliza-se números maiores.

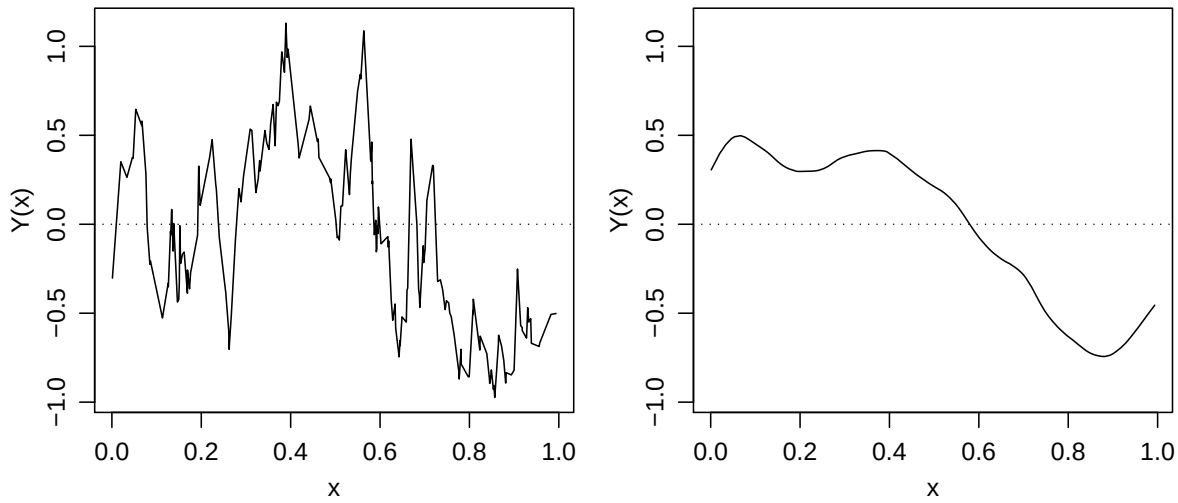

Figura 2.4: O gráfico da esquerda representa um processo de variações abruptas ao longo de uma transecção unidimensional, associada a uma função de correlação não-diferenciável. O da direita mostra um processo com variações mais suaves ao longo da mesma transecção, mas associada a uma função de correlação duas vezes diferenciável.

A Figura 2.4 ilustra-se um exemplo desse efeito a partir de simulações do processo $S(x)$. Foi gerado simulando-se 200 pontos de um processo estocástico estacionário, isotrópico, com taxas de decaimento equivalentes. Foram consideradas duas situações: função contínua não diferenciável (esquerda) onde nota-se variações bruscas da superfície gerada pelo processo e função contínua diferenciável (direita) onde as variações são mais suaves. Vale aqui salientar que o processo é o mesmo (exponencial), diferindo apenas na diferenciabilidade da função de correlação.

Deve-se lembrar que correlações com variações muito suaves perto da origem podem produzir efeitos de *quasi*-multicolinearidade na matriz de covariâncias, levando a dificuldades computacionais na solução numérica da álgebra envolvida no processo. Uma vez que se supõe diminuir a similaridade regional a longas distâncias, sendo no máximo nula, então é razoável escolher o conjunto de funções de correlações que sejam definidas positiva. Esta condição impõe restrições. Assim, para um conjunto de localizações x_i e uma constante real a_i, a condição:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_i;Y_j) \geq 0 \ \forall \ i;j$$

deve ser obedecida assegurando variância não negativa de predição e implicando que somente algumas famílias paramétrica específicas de funções de correlação, como as apresentadas a seguir, terão uso prático.

2.4.1 Função de correlação de Matèrn

Matèrn (1986) apresenta uma classe de funções de correlação que é considerada uma das mais completas, por englobar outras funções de correlação, pela simples escolha do parâmetro de diferenciabilidade. Esta é dada por:

$$\rho(u,\phi,\kappa) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{u}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{u}{\phi}\right) \tag{2.10}$$

onde $K_{\kappa}(\delta)$, $\delta = \frac{u}{\phi}$ é a função modificada de Bessel de terceiro tipo (ABRAMOWITZ; STEGUN, 1965) dada por:

$$K_{\kappa}(\delta) = \begin{cases} \left(\dfrac{\pi}{2\sin\pi\delta}\right)\{I_{-\kappa}(\delta) - I_{\kappa}(\delta)\} & \kappa \neq 0,1,2,\ldots \\ \\ \lim\limits_{p\to\kappa}\left(\dfrac{\pi}{2\sin\pi p}\right)\{I_{-\kappa}(\delta) - I_{\kappa}(\delta)\} & \kappa = 0,1,2,\ldots \end{cases}$$

sendo que:

$I_{\kappa}(\delta) = \sum\limits_{j=0}^{\infty} \dfrac{(\delta/2)^{\kappa+2j}}{j!\Gamma(\kappa+j+1)}$, para $\kappa = 0,1,2,\ldots$ e

$\Gamma(\kappa) = \int_0^{\infty} t^{\kappa-1}e^{-t}dt, \qquad \kappa > 0,$ é a função Gamma.

O parâmetro $\phi > 0$ na Equação 2.10 define a taxa na qual a função de correlação cai a zero com o aumento da distância u. O parâmetro $\kappa > 0$ é chamado de ordem do modelo de Matèrn e determina a suavidade do sinal $S(x)$. O comportamento dessa função pode ser vista na Figura 2.5.

2.4.2 Função de correlação da Família Esférica

A função de correlação dessa família é definida como:

$$\rho(u;\phi) = \begin{cases} 1 - \frac{3}{2}\left(\frac{u}{\phi}\right) + \frac{1}{2}\left(\frac{u}{\phi}\right)^3 & 0 \leq \phi \\ 0 & u > \phi \end{cases} \tag{2.11}$$

O nome desta função se deve ao fato de que $\rho(u;\phi)$ tem uma interpretação geométrica

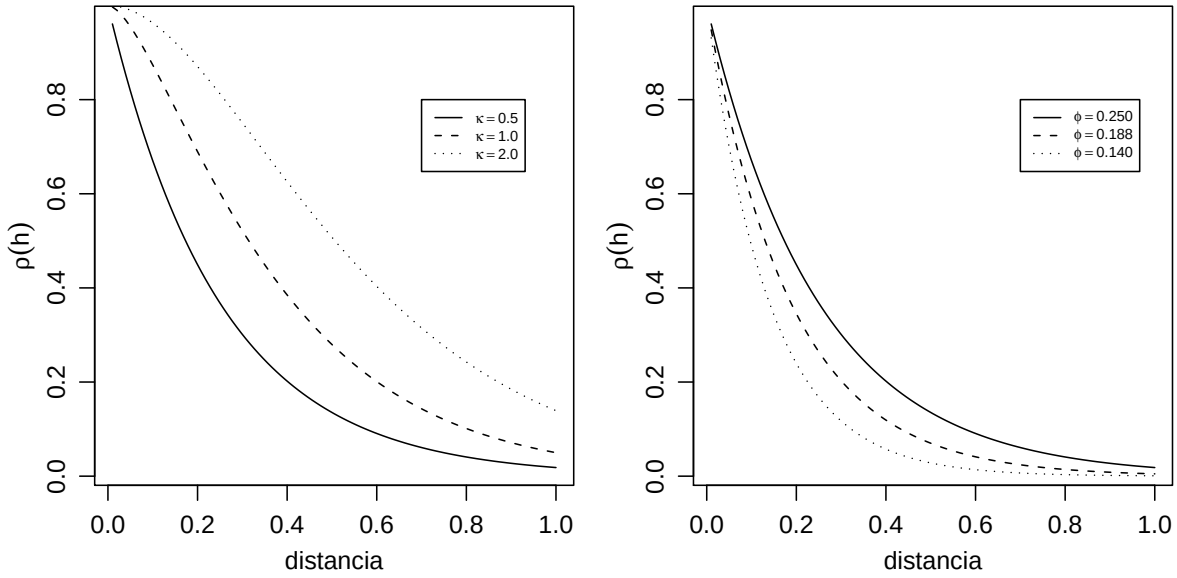

Figura 2.5: Comportamento da função de correlação de Matèrn com o parâmetro $\phi = 0,25$ fixo e diferentes valores para o parâmetro de diferenciabilidade κ (esquerda). Na mesma figura, para um mesmo valor de $\kappa = 0.5$, variou-se o parâmetro ϕ que controla a taxa de decaimento da função (direita).

como sendo o volume de interseção de duas esferas cujos centros estejam separadas de uma distância u (DIGGLE; RIBEIRO JR, 2007). Essa função de correlação tem alcance finito e depende somente do parâmetro de escala ϕ. O comportamento gráfico dessa função pode ser vista na Figura 2.6 à esquerda.

2.4.3 Função de correlação da Família "Potência" de ordem κ

A função de correlação dessa família é definida como:

$$\rho(u;\phi;\kappa) = e^{-\left(\frac{u}{\phi}\right)^{\kappa}} \quad \text{para } \phi > 0 \quad \text{e} \quad 0 < \kappa \leq 2 \tag{2.12}$$

Nesta função, se $\kappa < 2$, o processo $S(x)$ é contínuo mas não é diferenciável e se $\kappa \geq 2$ pode ser infinitamente diferenciável. Existem dois casos particulares para essa função. No caso $\kappa = 1$ a função será chamada exponencial, e para $\kappa = 2$ a função será chamada de gaussiana (Figura 2.6 à direita).

Toda a metodologia geoestatística está baseada na correlação existente entre as medidas tomadas em duas coordenadas distintas. As formas das funções apresentadas atendem ao pressuposto de que as observações mais próximas são, provavelmente, mais similares entre si do

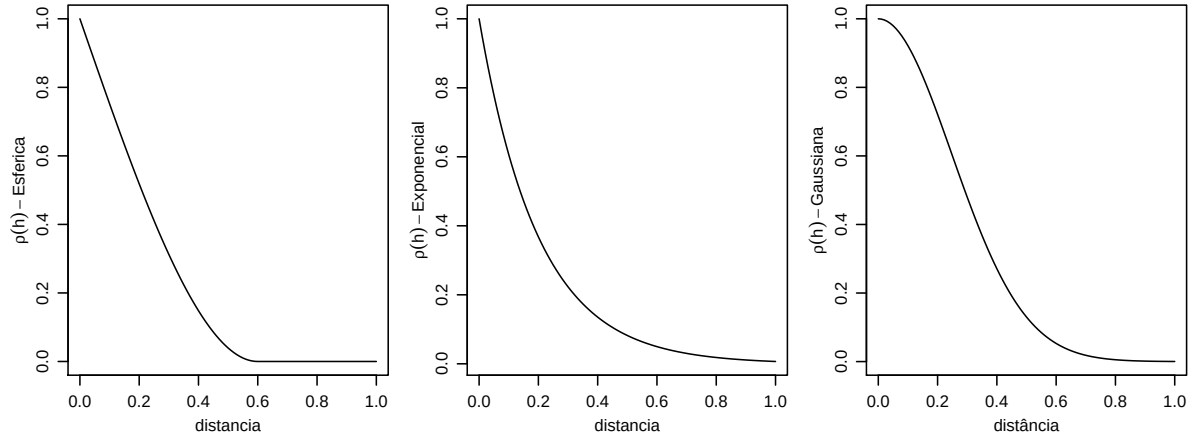

Figura 2.6: O gráfico da esquerda mostra uma função de correlação esférica com o parâmetro $\phi = 0,6$. O gráfico do centro ilustra o comportamento de uma função de correlação exponencial de ordem $\kappa = 1$ e $\phi = 0,2$. O gráfico da direita ilustra também o comportamento de uma função de correlação exponencial de ordem $\kappa = 2$ e $\phi = 0,35$, equivalente à função Gaussiana.

que aquelas muito afastadas. Isso dá o caráter regionalizado de um atributo ou uma propriedade em áreas agrícolas.

Existe na literatura outras propostas de funções de correlação, como a geométrica (WAKERNAGEL, 2003), ou mesmo aquelas que apresentam só o patamar (efeito pepita puro). Das funções apresentadas, a mais empregada é a de Matèrn pois ela permite maior flexibilidade na variação dos parâmetros por descreverem a diferenciabilidade do processo e a extensão da dependência espacial. Esta foi a família de correlações adotada no desenvolvimento deste trabalho.

2.5 ESTIMAÇÃO DE PARÂMETROS DO MODELO

2.5.1 Modelagem e estimação de parâmetros de tendência não-estacionária

No modelo geoestatístico idealizado para o processo mensurável Y, dado pela Equação 2.2 a estrutura linear para a média $\mu(x_i)$ é usualmente chamada de "tendência" e é dada de forma geral por:

$$\mu(x_i) = \beta_0 + \beta_1 d_1(x_i) + \ldots + \beta_p d_p(x_i) = \beta_0 + \sum_{j=1}^{p} \beta_j d_j(x_i) \tag{2.13}$$

sendo p o número de covariáveis presentes. Na forma matricial se representa como:

$$\mu = D\,\beta \tag{2.14}$$

onde: $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix}$; $D(x) = \begin{pmatrix} 1 & d_{11} & d_{21} & \dots & d_{p1} \\ 1 & d_{12} & d_{22} & \dots & d_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{1n} & d_{2n} & \dots & d_{pn} \end{pmatrix}$; $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_p \end{pmatrix}$

sendo a matriz D uma matriz de posto completo, ou seja, $n \geq p$. Os coeficiente são obtidos empregando-se o método dos mínimos quadrados (MONTGOMERY; PECK, 1955). Sob a hipótese de independência entre as observações, a função de mínimos quadrados para o problema pode ser escrita como:

$$MSQ(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(\mu_i - \beta_0 - \sum_{j=1}^{p} \beta_j d_j(x_i) \right)^2 \tag{2.15}$$

A solução que minimiza a Equação 2.15 em termos de β, segundo Montgomery e Peck (1955) é aquela que satisfaz:

$$\frac{\partial}{\partial \beta_0} MSQ(\beta) = -2 \sum_{i=1}^{n} \left(\mu_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j d_j(x_i) \right) = 0$$

$$\frac{\partial}{\partial \beta_j} MSQ(\beta) = -2 \sum_{i=1}^{n} \left(\mu_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j d_j(x_i) \right) d_j(x_i) = 0$$

para $j = 1, \dots, p$ e $i = 1, \dots, n$.

Expandindo-se o somatório externo e simplificando obtém-se o seguinte sistema de equações normais de mínimos quadrados:

$$\begin{cases} n\hat{\beta}_0 & +\hat{\beta}_1 \sum_{i=1}^{n} d_1(x_i) & +\hat{\beta}_2 \sum_{i=1}^{n} d_2(x_i) & +\dots & +\hat{\beta}_k \sum_{i=1}^{n} d_p(x_i) & = \sum_{i=1}^{n} \mu_i \\ \hat{\beta}_0 \sum_{i=1}^{n} d_{i1} & +\hat{\beta}_1 \sum_{i=1}^{n} d_1(x_i)^2 & +\hat{\beta}_2 \sum_{i=1}^{n} d_1(x_i) d_2(x_i) & +\dots & +\hat{\beta}_k \sum_{i=1}^{n} d_1(x_i) d_p(x_i) & = \sum_{i=1}^{n} d_{i1} \mu_i \\ \hat{\beta}_0 \sum_{i=1}^{n} d_{i2} & +\hat{\beta}_1 \sum_{i=1}^{n} d_1(x_i) d_2(x_i) & +\hat{\beta}_2 \sum_{i=1}^{n} d_2(x_i)^2 & +\dots & +\hat{\beta}_k \sum_{i=1}^{n} d_2(x_i) d_p(x_i) & = \sum_{i=1}^{n} d_{i2} \mu_i \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hat{\beta}_0 \sum_{i=1}^{n} d_{ik} & +\hat{\beta}_1 \sum_{i=1}^{n} d_p(x_i) d_1(x_i) & +\hat{\beta}_2 \sum_{i=1}^{n} d_p(x_i) d_2(x_i) & +\dots & +\hat{\beta}_k \sum_{i=1}^{n} d_p(x_i)^2 & = \sum_{i=1}^{n} d_{ik} \mu_i \end{cases}$$

Como solução dessas equações normais tem-se os estimadores de mínimos quadrados para β. Usando notação matricial, a função de mínimos quadrados para a Equação 2.14 será dada por:

$$\begin{aligned} MSQ(\beta) &= \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (\mu - D\,\beta)'(\mu - D\,\beta) \\ &= \mu'\mu - \beta' D'\mu - \mu' D\beta + \beta' D' D\beta \\ &= \mu'\mu - 2\beta' D'\mu + \beta' D' D\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \beta} MSQ(\beta) &= \frac{\partial}{\partial \beta}(\mu'\mu - 2\beta' D'\mu + \beta' D' D\beta) = \\ &= -2D'\mu + 2D' D\hat{\beta} = 0 \end{aligned}$$

Assim, $D' D\hat{\beta} = D'\mu$, e portanto, $\hat{\beta}$ pode ser estimado como:

$$\hat{\beta} = \left(D' D\right)^{-1} D'\mu$$

Considerando-se que μ_i representa a média de uma única observação na localização x_i, então esse valor coincide com o valor observado y_i e pode-se assim escrever o estimador dos coeficientes do modelo de tendência como sendo:

$$\hat{\beta} = \left(D' D\right)^{-1} D' Y$$

Se os dados não forem independentes e a matriz de covariância associada Σ do modelo (que é o caso) for conhecida, então o método será denominado mínimos quadrados generalizados. O modelo inicial será inflacionado na quantidade de parâmetros a serem estimados para definir Σ. O estimador de β será dado por:

$$\hat{\beta} = \left(D'\Sigma^{-1} D\right)^{-1} D'\Sigma^{-1} Y \tag{2.16}$$

Assumindo-se que Y tem distribuição normal multivariada, $\hat{\beta}$ será o estimador de mínimos quadrados para β, com suas importantes propriedades, coincidindo com o estimador de máxima verossimilhança.

Uma vez identificada e modelada a tendência, é possível estimar componentes residuais das observações por:

$$Y^* = Y - D\hat{\beta} \tag{2.17}$$

e que refletem apenas a estrutura de covariâncias do processo.

2.5.2 Ajuste de modelo ao semivariograma por mínimos quadrados

O semivariograma teórico trata-se do gráfico da função semivariância *versus* a distância u que separa duas posições. Como a função de correlação associada é assintoticamente decrescente, sua variação será muito pequena para grandes valores de u, podendo ser considerada estável, para efeitos práticos. Segundo Diggle e Ribeiro Jr (2007) uma convenção adotada por este modelo é considerar atingido o patamar quando, para um dado u_0, tem-se $\rho(u_0) \simeq 0,05$. Não há uma razão científica para se adotar esse valor de corte, podendo ser considerada uma quantidade numericamente razoável para a estabilização da função de correlação e, consequentemente, da função semivariância. Esse valor u_0 é denominado de alcance prático. Em termos da função semivariância, seu valor é obtido com o valor de u_0 tal que $\gamma(u_0) = \tau^2 + 0,95\ \sigma^2$.

Para a modelagem de um processo gaussiano isotrópico estacionário, o problema se reduz a definir a função de correlação mais apropriada ao fenômeno e estimar os parâmetros μ, τ^2, σ^2 e ϕ, na situação mais simples.

A estimativa de Matheron (MATHERON, 1963) para a semivariância teórica envolvendo duas medidas do processo Y é dada pela Equação 2.5, denominada semivariância experimental ou empírica. Uma área contendo n coordenadas amostrais fornecerá $\binom{n}{2}$ pares do tipo (u_{ij}, v_{ij}). Este será, dependendo do número de coordenadas amostrais, um conjunto muito grande. O seu gráfico é denominado semivariograma experimental, caracterizado por uma nuvem de pontos. Seu aspecto é mostrado pela Figura 2.7.

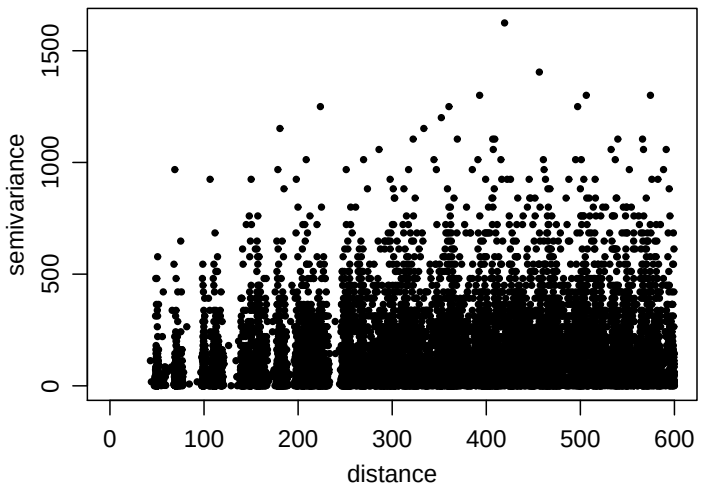

Figura 2.7: Variograma empírico de concentração de cálcio em uma área com 178 pontos amostrais, em dados de pesquisa de Oliveira (2003).

Devido ao grande número de pontos no gráfico do semivariograma empírico, bem como a forte dispersão à grandes distâncias, ele se torna uma figura de difícil interpretação, no sentido de se tornar difícil ajustar visualmente um bom modelo variográfico teórico. Diggle e

Ribeiro Jr (2007) dizem que esse comportamento errático se deve ao fato de que a distribuição amostral marginal de cada ordenada v_{ij} é proporcional a uma distribuição qui-quadrado com 1 grau de liberdade, sendo portanto, fortemente assimétrica e com alto coeficiente de variação.

Visando facilitar o aspecto computacional do processo e ter uma interpretação gráfica plausível, Pannatier (1996) sugeriu dividir em intervalos a variação das distâncias u e representar, no ponto médio de cada intervalo, o valor médio do grupo das semivariâncias relativas a esse intervalo. O semivariograma se reduz a uns poucos pontos, permitindo o ajuste de um modelo variográfico teórico, usando como critério de ajuste métodos baseados em minimizar o erro médio quadrático, dado pela diferença entre o valor médio de v para uma distância u_0 representante do intervalo e o valor teórico nessa mesma distância, ou seja, um erro do tipo $(\gamma(u_0) - v(u_0))^2$. O gráfico típico resultante desse procedimento é mostrado na Figura 2.9.

O estimador pelo método dos momentos mais utilizado para a semivariância é aquele proposto por Matheron (1962) e definido como:

$$\hat{\gamma}(u) = \frac{1}{|2N(u)|} \sum_{N(u)} \big(y(x_i) - y(x_j)\big)^2 \tag{2.18}$$

onde $N(u) = \big\{(x_i, x_j) : x_i - x_j = u; i, j = 1, 2, \ldots, n\big\}$ é o conjunto dos pares cujas distâncias é u. Para Braga (1990), se Y for uma função aleatória estacionária, então esse estimador, sob a hipótese intrínseca, é não-tendencioso e não-viciado para a média mas muito afetado por observações atípicas (outliers).

Atteia, Dubois e Webster (1994) disseram que a situação ideal em uma região homogênea segundo as observações amostrais de uma variável $Y(x)$ espacialmente distribuída, corresponderia à reta bissetriz no primeiro quadrante de um plano cartesiano, onde ficariam alocados os pontos do diagrama u-dispersão. Mas a realidade é diferente disso, apresentando pontos fora dessa reta. Esses pontos representam as diferenças de duas coordenadas quaisquer.

Pela Figura 2.8, deduz-se que:

$$\begin{aligned} \cos 45^0 &= \frac{d_\alpha}{Y(x_\alpha + u) - Y(x_\alpha)} \\ d_\alpha &= \cos 45^0 \, (Y(x_\alpha + u) - Y(x_\alpha)) \end{aligned}$$

sendo que α representa uma certa distância fixa.

A distância média quadrática $\gamma(u)$ será obtida como:

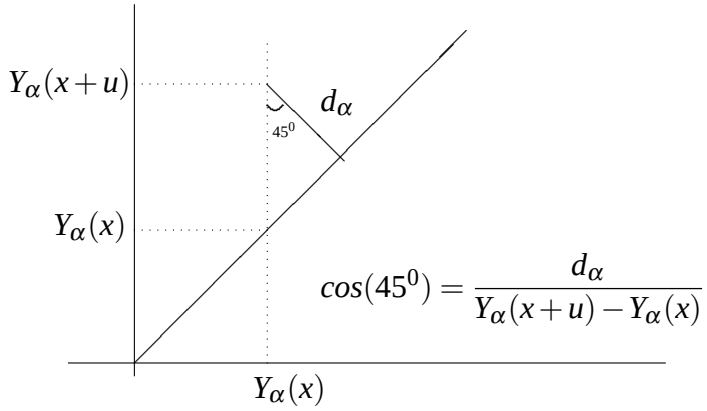

Figura 2.8: Ilustração geométrica da obtenção de um par de pontos do variograma empirico. d_α representa a distância de um ponto separado de outro por uma distância *u* até a reta bissetriz de um diagrama de dispersão u-scaterplot.

$$\gamma(u) = \frac{1}{N(u)}\sum_{\alpha=1}^{N(u)} d_\alpha^2 = \frac{\cos^2 45^0}{N(u)}\sum_{\alpha=1}^{N(u)} (Y(x_\alpha+u) - Y(x_\alpha))^2$$

$$\gamma(u) = \frac{1}{2N(u)}\sum_{\alpha=1}^{N(u)} (Y(x_\alpha+u) - Y(x_\alpha))^2$$

que corresponde à expressão do estimador do semivariograma experimental apresentada por Isaaks e Srivastava (1989), Journel e Huijbregts (1978), Pannatier (1996), Matheron (1962) entre outros.

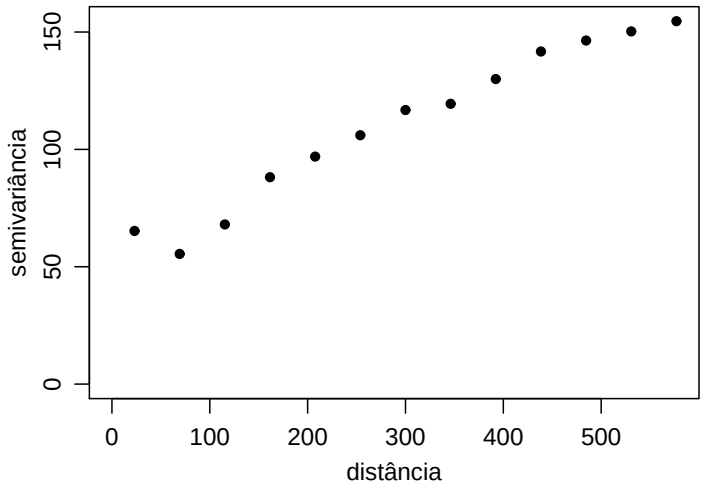

Figura 2.9: Variograma empírico agrupado em classes (“binado”) de concentração de cálcio em área com 178 pontos amostrais, em dados de pesquisa de Oliveira (2003).

Essa abordagem vem sendo adotada por diversos autores em estudos que envolvem aplicações agrícolas. Reichardt, Vieira e Libardi (1986) estudaram 50 dados de pH de solo,

de amostras coletadas com espaçamento de 1 m, em transecção de uma área de Latossolo Vermelho-escuro orto localizado em Araras-SP, cultivada com cultura de cana-de-açúcar. A técnica de autocorrelação que empregaram nos dados mostrou que observações de pH eram correlacionadas espacialmente até uma distância de 5 m. Observaram ainda que, para as amostras serem consideradas independentes e completamente casualizadas, deveriam ser espaçadas a pelo menos, 10 m. Com seu trabalho, os autores concluíram que a variabilidade espacial do solo pode ser definida corretamente e que a geoestatística era a alternativa certa às metodologias tradicionais.

Prevedello (1987) estudou a magnitude da variabilidade espacial de 47 parâmetros (físicos e químicos) de um solo com Terra Roxa Estruturada, em uma área de 4.810 m^2, em Piracicaba-SP, onde foi aplicado o manejo de uma cultura de arroz de sequeiro. O autor utilizou em seu experimento uma estrutura regular formada pelo cruzamento de 4×13 linhas, totalizando 52 pontos amostrais, separados 10 m entre si. Avaliou e discutiu a dependência espacial pela análise do autocorrelograma e do semivariograma, usando o estimador clássico de Matheron. Assim, com o emprego da teoria das variáveis regionalizadas, estabeleceu subunidades de amostragem ou de manejo individualizado, considerando-as independentes. Concluiu ainda que a área total não se mostrou homogênea para nenhum dos 47 parâmetros estudados, contrariando o que havia inicialmente suposto.

Mohamed, Evans e Shiel (1996) usaram a geoestatística para examinar a variabilidade geográfica em uma área de terra e descobrir, pela distribuição espacial a melhor densidade amostral, no sentido de obterem as propriedades de colheita e distribuição das características do solo com poucas amostras. Com o emprego do semivariograma experimental determinado pelo estimador clássico de Matheron, detectaram uma estrutura de variabilidade no solo. Com isso puderam utilizar seus parâmetros para efetuarem a interpolação de dados para produção de mapas de contornos.

Yang et al. (1998) estudaram a influência da topografia no rendimento da colheita, pela variabilidade de cinco campos em declive, da região de Palouse, em Washington-USA. Os autores desenvolveram um sistema de informações geográficas (GIS) para o manejo e análise do rendimento de trigo, juntamente com informações georreferenciadas sobre a variabilidade da topografia. Identificaram também o padrão de variabilidade do rendimento do trigo dentro de cada região plantada, para cada uma das cinco regiões estudadas e avaliaram a relação entre rendimento e atributos de topografia. Descreveram o padrão de variabilidade espacial pelo semivariograma, que mostrou claramente uma estrutura de dependência espacial justificando o emprego do manejo localizado.

2.5.3 Ajuste de modelos e estimação dos parâmetros por máxima verossimilhança

Considerando o caso estacionário do modelo geoestatístico univariado dado pela Equação 2.2, onde o processo $S(x_i)$ pode ser escrito como um conjunto de observações Y com distribuição de probabilidades de acordo com a Equação 2.3, os parâmetros gerais do modelo a serem estimados são: $\theta = (\beta, \sigma^2, \phi, \tau^2)$ onde, como já foi dito, ϕ é um parâmetro da função de correlação.

Os dadis $y = \{y_1, \ldots, y_n\}$, que representam uma realização do processo estocástico espacial em n coordenadas, possui distribuição gaussiana n-variada, ou seja, $Y \sim N_n(\mu; \Sigma)$ onde μ é um vetor de números reais, todos iguais e Σ é a matriz de variâncias e covariâncias de tamanho $n \times n$, com as propriedades de ser simétrica e definida positiva. Então, a distribuição conjunta de Y, segundo (DUDEWICZ; MISHRA, 1988) será:

$$f_Y(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(Y-\mu)'\Sigma^{-1}(Y-\mu)\right\}$$

para todo vetor Y de números reais.

Sendo Y um vetor gaussiano correlacionado, sua função de verossimilhança será composta pela sua distribuição conjunta de probabilidades dada por:

$$L(\theta) = f(y|\theta) = \frac{(|\sigma^2 R + \tau^2 I|)^{-1/2}}{(2\pi)^{n/2}} exp\left\{-\frac{1}{2}(Y-D\beta)'(\sigma^2 R + \tau^2 I)^{-1}(Y-D\beta)\right\}. \quad (2.19)$$

O logaritmo da função verossimilhança é dado por:

$$l(\theta) = -\frac{1}{2}\log(2\pi)^n - \frac{1}{2}\log(|\sigma^2 R + \tau^2 I|) - \frac{1}{2}(Y-D\beta)'(\sigma^2 R + \tau^2 I)^{-1}(Y-D\beta)$$

$$\begin{aligned} l(\theta) &= -\frac{1}{2}\left[n\log(2\pi) + \log(|\sigma^2 R + \tau^2 I|) + \right. \\ &\quad \left. + (Y-D\beta)'(\sigma^2 R + \tau^2 I)^{-1}(Y-D\beta)\right]. \end{aligned} \quad (2.20)$$

Fazendo-se $\frac{\tau^2}{\sigma^2} = \nu^2$ então $Var(Y) = \Sigma = \sigma^2 R + \tau^2 I = \sigma^2\left(R + \frac{\tau^2}{\sigma^2} I\right) = \sigma^2 V$.

Substituindo-se $\sigma^2 R + \tau^2 I$ por $\sigma^2 V$ na Equação 2.20, vem:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\sigma^2 V|) + (Y-D\beta)'(\sigma^2 V)^{-1}(Y-D\beta)\right]. \quad (2.21)$$

Agora substituindo $\sigma^2 R + \tau^2 I$ por Σ na mesma Equação (2.20), tem-se:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\Sigma|) + (Y - D\beta)'(\Sigma)^{-1}(Y - D\beta)\right] \quad (2.22)$$

Desenvolvendo-se os produtos matriciais a Equação 2.22 resulta em:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\Sigma|) + Y'\Sigma^{-1}Y - 2Y'\Sigma^{-1}D\beta + \beta' D'\Sigma^{-1}D\beta\right] \quad (2.23)$$

em que $Y'\Sigma^{-1}D\beta$ é um escalar pois $Y_{1\times n}$, $\Sigma_{n\times n}$, $D_{n\times n}$ e $\beta_{n\times 1}$.

Segundo Kolman (1997), se A é uma matriz quadrada simétrica definida positiva e $\beta = [\beta_1, \beta_2, ..., \beta_n]'$ um vetor, então:

a) $\dfrac{\partial Ax}{\partial x} = A'$ (transposta)

b) $\dfrac{\partial x'Ax}{\partial x} = 2Ax$ (forma quadrática).

Desses resultados obtém-se a derivada parcial do logaritmo da função de verossimilhança de θ com relação a β, dada por:

$$\frac{\partial l(\theta)}{\partial \beta} = -\frac{1}{2}(-2(Y'\Sigma^{-1}D)' + 2(D'\Sigma^{-1}D)\beta) = D'\Sigma^{-1}Y - D'\Sigma^{-1}D\beta.$$

Se $\dfrac{\partial l(\theta)}{\partial \beta} = 0$, então $D'\Sigma^{-1}Y - D'\Sigma^{-1}D\hat{\beta} = 0$, e assim obtém-se o estimador MV para o parâmetro β que é dado por:

$$\hat{\beta} = (D'\Sigma^{-1}D)^{-1}D'\Sigma^{-1}Y. \quad (2.24)$$

Considerando-se também que $\Sigma = \sigma^2 V$ e que $|\Sigma| = (\sigma^2)^n |V|$, então a Equação 2.21 fica:

$$l(\theta) = -\frac{1}{2}\left[n\log(2\pi) + \log(|\sigma^2 V|) + Y'(\sigma^2 V)^{-1}Y - \beta' D'(\sigma^2 V)^{-1}D\beta\right],$$

logo:

$$\begin{aligned} l(\theta) &= -\frac{1}{2}\left[n\log(2\pi) + \log[(\sigma^2)^n|V|] + \frac{Y'V^{-1}Y}{\sigma^2} - 2\frac{Y'V^{-1}D\beta}{\sigma^2} + \frac{\beta' D'V^{-1}D\beta}{\sigma^2}\right] \\ &= -\frac{1}{2}\left[n\log(2\pi) + n\log(\sigma^2) + \log|V| + \frac{(Y - D\beta)'V^{-1}(Y - D\beta)}{\sigma^2}\right] \end{aligned}$$

onde $[(Y-D\beta)'V^{-1}(Y-D\beta)]/\sigma^2$ é uma soma de quadrados ponderada pela matriz de covariâncias.

Calculando-se a derivada de $l(\theta)$ com relação a σ^2 obtém-se:

$$\frac{\partial l(\theta)}{\partial \sigma^2} = -\frac{1}{2}\left[\frac{n}{\sigma^2} - \frac{(Y-D\beta)'V^{-1}(Y-D\beta)}{(\sigma^2)^2}\right].$$

Se $\frac{\partial l(\theta)}{\partial \sigma^2} = 0$ e considerando o vetor de parâmetros $(\beta, \sigma^2, \phi, \nu^2)'$, tem-se:

$$\begin{aligned} -\frac{n}{\hat{\sigma^2}} + \frac{(Y-D\beta)'V^{-1}(Y-D\beta)}{(\hat{\sigma}^2)^2} &= 0, \\ \frac{(Y-D\beta)'V^{-1}(Y-D\beta)}{(\hat{\sigma}^2)^2} &= n, \end{aligned}$$

logo,

$$\hat{\sigma}^2_{\phi,\nu^2} = \frac{(Y-D\beta)'V^{-1}_{\phi,\nu^2}(Y-D\beta)}{n} \tag{2.25}$$

Retomando-se a Equação (2.24) e substituindo Σ por $\sigma^2 V$ vem:

$$\begin{aligned} \beta &= \left(\frac{D'V^{-1}D}{\sigma^2}\right)^{-1}\frac{DV^{-1}Y}{\sigma^2} \\ &= (D'V^{-1}D)^{-1}\sigma^2\frac{DV^{-1}Y}{\sigma^2} \\ &= (D^{-1}V^{-1}_{\phi,\nu^2}D)^{-1}DV^{-1}_{\phi,\nu^2}Y \end{aligned} \tag{2.26}$$

que depende somente dos parâmetros ϕ e ν^2. Neste caso, a matriz de correlação será dada por:

$$V = \begin{pmatrix} 1+\nu^2 & \rho(u_{12}) & \dots & \rho_(u_{1n}) \\ \rho(u_{21}) & 1+\nu^2 & \dots & \rho(u_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(u_{n1}) & \rho(u_{n2}) & \dots & 1+\nu^2 \end{pmatrix} \tag{2.27}$$

O logaritmo da função de verossimilhança concentrada será então dada por:

$$l(\phi,\nu^2) = -\frac{1}{2}\left[n\log(2\pi)+n\log\left(\frac{(Y-D\beta)'V^{-1}_{\phi,\nu^2}(Y-D\beta)}{n}\right)+\log|V| + \frac{(Y-D\beta)'V^{-1}(Y-D\beta)}{\left(\frac{Y-D\beta)'V^{-1}(Y-D\beta}{n}\right)}\right],$$

Assim,

$$l(\phi,\nu^2) = -\frac{1}{2}\left[n\log(2\pi)+n\log\left(\frac{(Y-D\beta)'V^{-1}_{\phi,\nu^2}(Y-D\beta)}{n}\right)+\log|V|+n\right].$$

logo

$$l(\phi,\nu^2) = -\frac{1}{2}\left[n\log(2\pi)+n\log\left((Y-D\beta)'V^{-1}_{\phi,\nu^2}(Y-D\beta)\right)-n\log n + \log|V|+n\right]. \tag{2.28}$$

Para um modelo estacionário, a menos das constantes, a função $l(\phi,\nu^2)$ fica:

$$l(\phi,\nu^2) \propto -\frac{n}{2}\left((Y-\mu)'V^{-1}_{\phi,\nu^2}(Y-\mu)\right)-\frac{\log|V|}{2}. \tag{2.29}$$

Esta função recebe como argumentos, o vetor das observações do processo Y e a matriz das distâncias de cada coordenada com as demais, que permite obter V pela escolha conveniente de uma função de correlação $\rho(u_{ij})$. A maximização dessa função, segundo os parâmetros envolvidos, fornecerá a estimativa dos parâmetros do modelo de correlação espacial.

Funções côncavas são aquelas cujo gráfico está sempre acima ou sobre qualquer corda traçada numa região entre seus pontos, ou, equivalentemente, seu gráfico está abaixo da reta tangente ao seu ponto de máximo. Neste sentido, tanto a função de verossimilhança quanto o logaritmo da função de verossimilhança são funções côncavas, garantindo assim a existência de um ponto de máximo local.

Para se obter a melhor estimativa para os parâmetros, deve-se encontrar simultaneamente o valor dos parâmetros que irão maximizar essa função. Muitos programas computacionais, incluindo o geoR (RIBEIRO JR; DIGGLE, 2001), possuem algoritmos eficientes para estimar esses parâmetros. A questão importante a se destacar aqui é que esse método, usado para aderir um modelo teórico com a melhor estimativa de seus parâmetros, envolvem todas as observações amostrais, sem a necessidade dos agrupamentos feito nos ajustes através de

variogramas, evitando os erros decorrentes.

Uma restrição quanto ao uso do método da otimização do logaritmo da função de verossimilhança está relacionada à forma suave de variação de certas funções de correlação, ou seja, aquelas funções que são diferenciáveis um número grande de vezes. Nestes casos, a matriz de correlação poderá apresentar colunas muito parecidas numericamente, impossibilitando sua inversão.

Muitos pesquisadores atualmente envolvem em seus trabalhos, a escolha de modelo de correlação e ajuste dos parâmetros por este método. Oliveira (2003) o utilizou em dados experimentais coletados em levantamento detalhado de solos da Estação Experimental de Campos, Rio de Janeiro, na Fazenda Angra, em estudo pedológico onde foram avaliadas as características morfológicas, físicas e químicas dos solos, e apresentadas também, informações referentes à distribuição geográfica. No estudo geoestatístico foi considerada a variável agronômica teor de Cálcio (mmolc dm^{-3}), nas camadas de 0-20 e 20-40 cm. Dentre suas conclusões verificou que o estimador de máxima verossimilhança não foi eficiente para detectar diferenças entre os modelos com covariável.

2.5.4 Ajuste de modelos e estimação dos parâmetros por máxima verossimilhança restrita

Modelos mistos descrevem experimentos cuja estrutura linear envolve fatores fixos e fatores aleatórios, independentemente da média e do erro, exigindo uma análise separada para cada uma de suas partes. A análise da parte aleatória é feita pela estimação dos componentes da variância na presença dos efeitos fixos e a análise da parte fixa é feita pela estimativa da função que a governa e por testes de hipóteses.

Hartley e Rao (1967) apresentam em seu artigo procedimentos de estimação por MV para análise de variância para modelos mistos generalizados envolvendo qualquer combinação de fatores fixos e aleatórios e interações de qualquer ordem. O método se aplica aos casos onde as estruturas matriciais envolvidas satisfazem certas condições mostradas no seu trabalho. Os autores mostram ainda a eficiência e a consistência dos estimadores e derivam testes de hipóteses e intervalos de confiança. Já o método da máxima verossimilhança restrita é descrito por Patterson e Thompson (1971) como os procedimentos de MV modificados, extensivo a delineamentos experimentais em blocos com estruturas mais complexas para também estimar componentes de variância do modelo.

Segundo Perry e Iemma (1999) os estimadores de MVR são obtidos maximizando-se a parte da função MV que é invariante ao parâmetro de locação, ou seja, maximizando-se a função MV de um vetor de combinações lineares das observações que são invariantes a $\mu(x) = D\beta$. Considerando-se o modelo da Equação 2.2 vem:

$$KY = K\mu(x) + KS(x) + K\varepsilon \tag{2.30}$$

onde K é o tal vetor de combinações lineares. Assim:

$$Y \sim N_n\left(K\mu(x); K\left(\sigma^2 R(\phi) + \tau^2 I\right) K'\right)$$

e o logaritmo da função MVR poderá então ser escrita como:

$$\begin{aligned} -2\log L_{RE} &= \log|\sigma^2 R(\phi) + \tau^2 I| + (Y - D\hat{\beta})'(\sigma^2 R(\phi) + \tau^2 I)^{-1}(Y - D\hat{\beta}) \\ &\quad + \log|D'(\sigma^2 R(\phi) + \tau^2 I)D| + (n-k)\log 2\pi \end{aligned}$$

onde k é o posto da matriz D e:

$$\hat{\beta} = (D'(\hat{\sigma}^2 R(\hat{\phi}) + \hat{\tau}^2 I))(D'(\hat{\sigma}^2 R(\hat{\phi}) + \hat{\tau}^2 I))^{-1} Y$$

sendo $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{\tau}^2$ e $\hat{\phi}^2$ estimadores MV de β, σ^2, τ^2 e ϕ, respectivamente.

O método supõe gaussianidade dos dados, fornecendo estimativas não negativas da variância e ainda considera a perda de graus de liberdade pela presença dos fatores fixos. Apesar de ser empregado para estimar componentes da variância em dados desbalanceados (número diferentes de repetições) fornece também estimadores não viciados e de variância mínima para dados balanceados.

2.5.5 Escolha de modelos por validação cruzada

Para Cressie (1985), escolher um modelo é obter o estimador dos seus parâmetros com métodos estatísticos de otimização, e uma vez escolhido, resta saber se ele é eficiente para interpolar valores, permitindo estimativas confiáveis para a construção de mapas temáticos. Essa escolha é feita com a aplicação de métodos de validação que comparam o valor de uma variável sob um modelo geoestatístico teórico com o valor empíricos dessa variável obtido através de amostragem, em uma mesma coordenada espacial. Baseado na análise do erro de

estimação poderá ser escolhido o melhor modelo. Dentre os principais critérios para validação encontram-se o Critério de Informação de Akaike, de Filliben, da validação cruzada e o máximo valor do logaritmo da função verossimilhança (FARACO et al., 2008).

A validação cruzada é uma técnica frequentemente utilizada para se avaliar um modelo teórico idealizado para explicar um fenômeno. Após o ajuste de seus parâmetros, com base em um conjunto experimental de dados, os quais se supõe serem governados por tal modelo, testa-se o seu efeito sobre a estimação do mesmo conjunto de dados ou sobre outro conjunto conhecido e que seja supostamente governado pelo mesmo modelo. Para isso, duas principais estatégias são adotadas. Em uma delas, retira-se cada um dos pontos amostrais por vez e então o estima com o modelo ajustado ao conjunto completo de informações. Em outra estratégia, ajusta-se o modelo pretendido a um conjunto de dados experimentais e então aplica-se o modelo em outro conjunto de dados conhecidos e que também seja governado pelo mesmo modelo. A avaliação do erro de predição informará sobre a qualidade do modelo escolhido.

2.6 PREDIÇÃO LINEAR ESPACIAL UNIVARIADA

Um aspecto importante da modelagem estatística é a utilização do modelo obtido para efetuar predições. Empregar o termo predição significa fazer conjectura ou suposição sobre um resultado de Y, desconhecido, que poderá ou não acontecer. A meta é realizar boas estimativas de quantidades que variam continuamente no espaço, em função de um conjunto discreto de observações obtidas dispersamente em uma área. Esse procedimento, sob certas circunstâncias, é chamado krigagem, termo este criado por G. Matheron em reconhecimento ao trabalho do engenheiro de minas D. G. Krige (KRIGE, 1951), sendo a krigagem ordinária a mais utilizada. O método estima um valor em um ponto arbitrário de uma região fechada onde a função de correlação do processo é conhecida, empregando o conjunto de pontos amostrais conhecidos, distribuídos pela área.

Isaaks e Srivastava (1989) citam vários métodos de estimação pontual como: método poligonal de desagrupamento, método da triangulação, método do inverso do quadrado das distâncias, método dos vizinhos mais próximos. A krigagem ordinária é um método que fornece um estimador BLUE, acrônimo do inglês *Best Linear Unbiased Estimator* – melhor estimador não viciado e de variância mínima. O método é linear porque seus estimadores são feitos a partir de combinações lineares sobre as observações amostrais disponíveis, é não viciado pois o erro médio residual é zero e “melhor” porque dentre outros estimadores é o que leva à menor

variância do erro. Em uma coordenada arbitrária x_0, as estimativas serão dadas por:

$$\hat{y}(x_0) = \sum_{i=1}^{n} \omega_i y(x_i)$$

onde os x_0 são as coordenadas onde se deseja efetuar uma estimativa e ω_i é o peso associado à i-ésima observação $y(x_i)$, sujeito à restrição $\sum_{i=1}^{n} \omega_i = 1$, que garante a não tendenciosidade do preditor.

Journel e Huijbregts (1978) também salientaram que, no caso de processos não-estacionários, serão necessárias algumas condições de ausência de viés. Para eles a limitação à classe de estimadores lineares é natural, uma vez que são necessários somente os momentos de segunda ordem da função de covariância.

Schabenberger e Gotway (2005) fazem distinção entre estimação e predição, pois são procedimentos muitas vezes tidos como equivalentes. Em um modelo básico de regressão linear simples os erros não são correlacionados (são independentes) e os coeficientes são estimados por métodos de mínimos quadrados, conforme Equação 2.16 e então se prediz um valor de interesse. Não fica claro se o preditor é uma "resposta" em x_0 ou é um estimador de $E[Y(X_0)]$. Apesar da estimação de uma quantidade fixa ou predição de uma quantidade aleatória ser uma questão menor, sua importância fica clara ao se considerar uma incerteza associada a essas quantidades. No caso da geoestatística, apesar do total desconhecimento do processo $S(x)$, aplicações com predição são frequentemente mais empregadas do que aquelas que buscam a estimação de uma média.

O modelo de predição linear, sinônimo de krigagem, dependendo se a média do processo é ou não conhecida, proposto por esses autores é dado por:

$$\hat{Y}(x_0) = \hat{\mu} + r'\Sigma^{-1}(Y(x) - \hat{\mu})$$

onde $r = Cov(Y(x), Y(x_0))$ e Σ é a matriz de variâncias e covariâncias das variáveis observadas.

A variância da predição, segundo eles, será:

$$Var(\hat{Y}(x_0)) = \sigma^2 - r'\Sigma^{-1}r + \frac{(1 - \mathbf{1}'\Sigma^{-1}r)^2}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}$$

Segundo Goovaerts (1997), o estimador de krigagem é um estimador de regressão linear $\hat{S}(x)$ e é definido como:

$$\begin{aligned}\hat{S}(x) &= \mu + \sum_{i=1}^{n} \lambda_i (Y_i - \mu) \\ &= \mu + \sum_{i=1}^{n} \lambda_i Y_i - \sum_{i=1}^{n} \lambda_i \mu \\ &= \left(1 - \sum_{i=1}^{n} \lambda_i\right) \mu + \sum_{i=1}^{n} \lambda_i Y_i\end{aligned}$$

em que μ é a média do processo, mas que, em um caso geral, pode ser aplicada em seu lugar a função média $\mu(x)$. Y_i é o vetor de observações e λ a função peso.

Sendo $S(x)$ um processo estacionário e Y um vetor de variáveis aleatórias cujos valores são observáveis e T outra variável aleatória, cujo valor se deseja estimar, Y terá distribuição normal multivariada com média constante μ e variância $\sigma^2 R + \tau^2 I$ e $T = T(S)$ será a meta de predição. Se $T = S(x_0)$ então a distribuição conjunta de T e Y será normal multivariada, dada por: $(T,Y) \sim N_n \left(\mu, \begin{bmatrix} \sigma^2 & \sigma^2 r' \\ \sigma^2 r & \sigma^2 R + \tau^2 I \end{bmatrix} \right)$,

Para Diggle e Ribeiro Jr (2007) o estimador pontual $\hat{T} = E(T|Y)$ será o valor que minimiza o erro médio quadrático $MSE(\hat{T}) = E(\hat{T} - T)^2$ onde: $\hat{T}(x_0) = E(T|Y) = \mu + r'V^{-1}(Y - \mu)$ e $Var(\hat{T}(x_0)) = Var(T|Y) = \sigma^2 \left(1 - r'V^{-1}r\right)$, em que $V = \sigma^2 R + \tau^2 I$ e r é o vetor de correlação entre a posição dos valores observados e a posição do valor y_0 a ser predito.

No caso do valor de μ ser desconhecido, então ele poderá ser estimado por:

$$\hat{\mu} = (\mathbf{1}'V^{-1}\mathbf{1})^{-1}\mathbf{1}'V^{-1}Y$$

2.7 INFERÊNCIA BAYESIANA PARA MODELOS GEOESTATÍSTICOS

Autores como Lindley (1990) e Paulino, Turkman e Murteira (2003) consideram uma revolução científica a substituição do paradigma clássico pelo bayesiano. A base para essa abordagem em problemas de inferência, dizem, foi lançada por Richard Price em 1763 quando publicou a obra póstuma de Thomas Bayes, com o título *An Essay Towards Solving a Problem in the Dictrine of Chances*. Entretanto, há certo grau de discordância na literatura acerca dessa

origem, todavia o conhecido Teorema de Bayes, se aceitas as leis de probabilidades axiomáticas de Kolmogorov, é núcleo dessa abordagem.

Para Gelman et al. (2003) a análise bayesiana de dados consiste em um método para inferências que aplica modelos de probabilidades tanto para as quantidades amostrais observadas quanto para as quantidades que se deseja conhecer. A característica essencial do método, para eles, é o emprego explícito da lei de probabilidades para quantificar incertezas.

O método bayesiano para análise de dados é caracterizado pelas seguintes etapas:

- Constrói-se um modelo que comporte, na sua distribuição conjunta de probabilidades, as quantidades observáveis e as não observáveis, consistentes com o problema científico em questão;
- Condicionalmente aos dados observados, calcula-se e interpreta-se uma distribuição *a posteriori* de probabilidades para as quantidades não observáveis;
- Avalia-se o ajuste do modelo e as implicações dos resultados da distribuição a *posteriori*.

A lógica da inferência bayesiana é realizar inferência associando incerteza aos parâmetros envolvidos no modelo, tratando-os também como variável aleatória. A proposta para sua aplicação com geoestatística é a de combinar estimação e predição a partir de um conjunto de observações associadas a um processo, em um alvo de predição. O resultado obtido será uma realização de uma variável aleatória. Nela a incerteza sobre o parâmetro desconhecido, será descrita por uma distribuição de probabilidades e um conceito de predição, normalmente a média de sucessivas realizações.

2.7.1 Especificação do modelo geoestatístico bayesiano

Segundo Diggle e Ribeiro Jr (2007) um modelo geoestatístico é especificado conjuntamente para um processo espacial contínuo não observado $\{S(x) : x \in \mathbb{R}^2\}$ e para um conjunto de dados observados $Y(x)$ nas localizações x, condicionado ao processo espacial com efeito nas mesmas localizações. Sendo $\theta = \{\theta_1; \ldots; \theta_p \in \mathbb{R}\}$ o conjunto de parâmetros desconhecidos e não observáveis no modelo, então:

$$P\big(Y(x); S(x)|\theta\big) = \frac{P\big(Y(x); S(x); \theta\big)}{P(\theta)} = P\big(Y(x)|S(x); \theta\big)\, P\big(S(x)|\theta\big). \tag{2.31}$$

A distribuição preditiva do processo $S(x)$ é definida condicionalmente às observações $Y(x)$ como:

$$P(S(x)|Y(x)) = \frac{P(S(x);Y(x))}{P(Y(x))}, \tag{2.32}$$

entretanto, o cenário é constituído pela distribuição conjunta do processo $S(x)$, de observações de $Y(x)$ e de parâmetros desconhecidos θ. Para se obter a distribuição conjunta de $(S(x);Y(x))$ exigida na Equação 2.32, integra-se a distribuição de $(S(x);Y(x);\theta)$ sobre o espaço de parâmetros, ou seja:

$$P(S(x);Y(x)) = \int_\theta P(S(x);Y(x);\theta)\,d\theta = \int_\theta P(S(x)|Y(x);\theta)P(\theta|Y(x))P(Y(x))\,d\theta$$

que, substituída em 2.32 produz:

$$P\big(S(x)|Y(x)\big) = \int_\theta P\big(S(x)|Y(x);\theta\big)\,P\big(\theta|Y(x)\big)d\theta. \tag{2.33}$$

Essa distribuição corresponde a uma média ponderada pela distribuição de θ condicionada a $Y(x)$, onde os pesos $P\big(\theta|Y(x)\big)$ refletem a incerteza a *posteriori* sobre os valores dos parâmetros do modelo.

Bolstad (2004) mostra que aplicando-se o Teorema de Bayes, a distribuição a *posteriori* dos parâmetros pode ser escrita como:

$$P\big(\theta|Y(x)\big) = \frac{P\big(\theta;Y(x)\big)}{P\big(Y(x)\big)} = \frac{P\big(Y(x)|\theta\big)\,P\big(\theta\big)}{P\big(Y(x)\big)} \tag{2.34}$$

onde o termo $P\big(Y(x)\big) = \int_\theta P\big(Y(x)|\theta\big)\,P(\theta)\,d(\theta)$ é constante sob a distribuição de $[\theta|Y(x)]$. A notação $[\,\cdot\,]$ corresponde à distribuição de probabilidades do interior do colchetes. O termo $P\big(Y(x)|\theta\big)$ é a função de verossimilhança de $Y(x)|\theta$ com distribuição gaussiana multivariada e $P(\theta)$ é a distribuição a *priori* de θ que expressa o conhecimento prévio acerca da distribuição de probabilidades dos parâmetros.

Segundo Gelman et al. (2003) e Kolman (2004), excluindo-se o termo constante $P\big(Y(x)\big)$ a Equação 2.34 fica:

$$P\big(\theta|Y(x)\big) \propto P\big(Y(x)|\theta\big)\,P(\theta) \tag{2.35}$$

Destaque-se que a Equação 2.35 não é uma distribuição de probabilidades, mas $kP\big(\theta|Y(x)\big)$ o será para uma escolha adequada da constante k de proporcionalidade, obtida tanto por métodos analíticos quanto métodos numéricos.

O modelo geoestatístico dado pela Equação 2.3 pode ser especificado como um modelo hierárquico espacial misto como:

- $Y(x) = D\beta + S(x) + \varepsilon$;
- $S(x) \sim \mathsf{N}_n\big(0\,;\ \sigma^2 R(\phi)\big)$;
- $\varepsilon \overset{ind.}{\sim} \mathsf{N}_n(0;\tau^2)$;
- $\theta = (\beta;\sigma^2;\phi;\tau^2)$ que terá distribuição de probabilidades (a *priori*) a ser atribuída conforme o problema.

Na prática, a escolha da distribuição a *priori* dos parâmetros do modelo é um assunto delicado na inferência bayesiana pois ela se dá, ora por conhecimento (objetivo ou subjetivo) da sua distribuição, ora por uma conveniência que resulte em uma distribuição a *posteriori* com solução analitica. Distribuições a *priori* que resultem em uma *posteriori* da mesma família, são chamadas de *prioris* conjugadas e sua escolha se dá devido à tratabilidade analítica decorrente e a uma conveniência computacional. Se os parâmetros da distribuição a *priori* forem conhecidos, dir-se-á que ela é degenerada nos seus valores. Se o conhecimento sobre esses parâmetros é vago, dir-se-á que a distribuição *priori* não é informativa, é plana (*flat*) ou é imprópria (EHLERS, 2006).

Sendo $Y(x)$ um processo gaussiano como o definido pela Equação 2.3, e sua função de verossimilhança dada pela Equação 2.19, então a distribuição a *posteriori* dos parâmetros será dada pela Equação 2.35 como:

$$\begin{aligned} P(\beta;\sigma^2;\phi;\tau^2|Y(x)) \propto\ & |\sigma^2 R(\phi) + \tau^2 I|^{-\frac{1}{2}} \times \\ & \times exp\left\{ -\frac{1}{2}(Y(x) - D\beta)'(\sigma^2 R(\phi) + \tau^2 I)^{-1}(Y(x) - D\beta) \right\} \times \\ & \times P(\beta;\sigma^2;\phi;\tau^2) \end{aligned} \tag{2.36}$$

As distribuições a *posteriori* desses parâmetros foram obtidas considerando-se os seguintes casos:

a) Incerteza no parâmetro de média.

   Supondo-se que β tenha uma distribuição a *priori* não informativa dada por $P\left(\beta|\sigma^2;\phi\right) \propto 1$ e que σ^2 e ϕ sejam parâmetros conhecidos, utilizando-se a Equação 2.35

combinada com a Equação 2.19 tem-se:

$$\begin{aligned} P\left(\beta|Y(x);\sigma^2;\phi\right) &\propto \exp\left\{(Y(x)-D\beta)'\left(\sigma^2R(\phi)\right)^{-1}(Y(x)-D\beta)\right\} \\ &\propto \exp\Big\{Y'(x)\left(\sigma^2R(\phi)\right)^{-1}Y(x)-Y'(x)\left(\sigma^2R(\phi)\right)^{-1}D\beta \\ &\quad -\beta'D'\left(\sigma^2R(\phi)\right)^{-1}Y(x)+\beta'D'\left(\sigma^2R(\phi)\right)^{-1}D\beta\Big\} \end{aligned}$$

$$\begin{aligned} P\left(\beta|Y(x);\sigma^2;\phi\right) \propto \exp\Big\{&\beta'D'\left(\sigma^2R(\phi)\right)^{-1}D\,\beta \\ &-Y'(x)\left(\sigma^2R(\phi)\right)^{-1}D(D'\left(\sigma^2R(\phi)\right)^{-1}D)^{-1}(D'\left(\sigma^2R(\phi)\right)^{-1}D)\beta \\ &-\beta'(D'(\left(\sigma^2R(\phi)\right)^{-1}D)\,D^{-1}\left(\sigma^2R(\phi)\right)^{-1}Y(x) \\ &+Y'(x)\left(\sigma^2R(\phi)\right)^{-1}D(D\left(\sigma^2R(\phi)\right)^{-1}D) \\ &(D\left(\sigma^2R(\phi)\right)^{-1}D)^{-1}D'\left(\sigma^2R(\phi)\right)^{-1}Y(x) \\ &-Y'(x)\left(\sigma^2R(\phi)\right)^{-1}D(D\left(\sigma^2R(\phi)\right)^{-1}D) \\ &(D\left(\sigma^2R(\phi)\right)^{-1}D)^{-1}D'\left(\sigma^2R(\phi)\right)^{-1}Y(x)\Big\} \end{aligned}$$

$$\begin{aligned} P\left(\beta|Y(x);\sigma^2;\phi\right) \propto \exp\Big\{&(\beta-(D'\left(\sigma^2R(\phi)\right)^{-1}D)^{-1}D'\left(\sigma^2R(\phi)\right)^{-1}Y(x))' \\ &D'\left(\sigma^2R(\phi)\right)^{-1}D \\ &(\beta-(D'\left(\sigma^2R(\phi)\right)^{-1}D)^{-1}D'\left(\sigma^2R(\phi)\right)^{-1}Y(x))\Big\} \end{aligned}$$

De acordo com a dedução dada pela Equação 2.16,

$$\hat{\beta}=(D'\left(\sigma^2R(\phi)\right)^{-1}D)^{-1}D'\left(\sigma^2R(\phi)\right)^{-1}Y(x),$$

onde $\hat{\beta}$ é o estimador MV para β. Assim:

$$P\left(\beta|Y(x);\sigma^2;\phi\right) \propto \exp\left\{(\beta-\hat{\beta})'D'\left(\sigma^2R(\phi)\right)^{-1}D\,(\beta-\hat{\beta}).\right\} \qquad (2.37)$$

A Equação 2.37 corresponde ao núcleo de uma distribuição normal com média $\hat{\beta}$ e variância $\left(\sigma^2R(\phi)\right)^{-1}$, definindo uma distribuição a *posteriori* para β, condicionada às observações $Y(x)$ e aos parâmetros conhecidos σ^2 e ϕ:

$$\left[\beta|Y(x);\sigma^2;\phi\right]\sim N\left(\hat{\beta}\;;\;\left(\sigma^2R(\phi)\right)^{-1}\right).$$

b) Incerteza no parâmetro de escala.

Se forem conhecidos os parâmetros de média β e de correlação ϕ, Gelman et al. (2003) sugerem utilizar como uma *priori* cojungada a distribuição χ^2-inversa escalonada – χ^2_{ScI}, ou seja, $\left[\sigma^2|\beta\ ;\ \phi\right] \sim \chi^2_{ScI}(n_\sigma\ ;\ W^2)$ cuja função densidade de probabilidade é dada por:

$$P(\sigma^2) = \frac{(n_\sigma/2)^{(n_\sigma/2)}}{\Gamma(n_\sigma/2)} W^{n_\sigma}\ \left(\sigma^2\right)^{-(n_\sigma/2+1)} \exp\left\{-\frac{n_\sigma\, W^2}{2\,\sigma^2}\right\}, \qquad \sigma^2 > 0. \tag{2.38}$$

Nessa Equação 2.38, n_σ corresponde aos graus de liberdade e W a uma medida de escala.

A função de verossimilhança (Equação 2.19) pode aqui ser escrita como:

$$\begin{aligned} L(\sigma^2;\phi;\beta;Y(x)) &\propto (\sigma^2)^{-(n/2)} \exp\left\{-\frac{1}{2\,\sigma^2}(Y(x)-D\beta)'(R(\phi))^{-1}(Y(x)-D\beta)\right\} \\ &\propto (\sigma^2)^{-(n/2)} \exp\left\{-\frac{n\,\hat{\sigma}^2}{2\,\sigma^2},\right\} \end{aligned} \tag{2.39}$$

onde $\hat{\sigma}^2 = \frac{1}{n}(Y(x)-D\beta)'(R(\phi))^{-1}(Y(x)-D\beta)$ é o estimador MV para σ^2.

Tomando-se a *priori* conjugada sugerida tem-se:

$$P(\sigma^2|\beta;\phi) \propto (\sigma^2)^{-\frac{n_\sigma}{2}+1} \exp\left\{-\frac{n_\sigma\, W^2}{2\,\sigma^2}\right\}. \tag{2.40}$$

Substituindo-se a Equação 2.39 e a Equação 2.40 na Equação 2.35 vem:

$$\begin{aligned} P(\sigma^2|Y(x);\beta;\phi) &\propto \left(\sigma^2\right)^{-\frac{n+n_\sigma}{2}+1} \exp\left\{-\frac{n\,\hat{\sigma}^2+n_\sigma W^2}{2\,\sigma^2}\right\} \\ &\propto \left(\sigma^2\right)^{-\frac{n+n_\sigma}{2}+1} \exp\left\{-\frac{(n+n_\sigma)\dfrac{n\,\hat{\sigma}^2+n_\sigma W^2}{n+n_\sigma}}{2\,\sigma^2}.\right\} \end{aligned} \tag{2.41}$$

Portanto, $\left[\sigma^2|Y(x);\beta;\phi\right] \sim \chi^2_{ScI}\left(n+n_\sigma\ ;\ \dfrac{n\,\hat{\sigma}^2+n_\sigma W^2}{n+n_\sigma}\right)$.

c) Incerteza nos parâmetro de média e de escala.

Neste caso considera-se desconhecido os parâmetros da média β e de escala σ^2 e conhecido o parâmetro ϕ da função de correlação. Ehlers (2006) sugere especificar uma *priori* conjugada em duas etapas. Na primeira considera-se σ^2 fixo e se utiliza o resultado obtido na Equação 2.37 e na segunda etapa combina-se a função de verossimilhança da Equação 2.19 com uma distribuição a *priori* para σ^2, conforme a Equação 2.41, usando-se o fato de que $P(\beta;\sigma^2) = P(\beta|\sigma^2)\ P(\sigma^2)$. Esquematicamente o procedimento fica:

- Especifica-se as *priori* marginais, supondo independência entre elas;
- Obtém-se a *priori* conjunta;

- Obtém-se a *posteriori* conjunta através do produto da *priori* conjunta pela função de verossimilhança;
- Obtém-se as *posteriori* marginais por integração.

Seja:

$$P\big(\beta\ ;\ \sigma^2|Y(x)\big) \propto P\big(\beta\ ;\ \sigma^2\big)\ P\big(Y(x)|\beta\ ;\ \sigma^2\big) \tag{2.42}$$

a distribuição a posteriori conjunta para β e σ^2 desejada e seja:

$$\big[Y(x)|\beta\ ;\ \sigma^2\big] \sim N_n\big(D\,\beta\ ;\ \sigma^2\ R(\phi)\big)$$

a função de verossimilhança. Então:

$$P(\beta\ ;\ \sigma^2|Y(x)) \propto \big(\sigma^2\big)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\,\sigma^2}(Y(x)-D\,\beta)'\big(R(\phi)\big)^{-1}(Y(x)-D\,\beta)\right\}$$

Conforme solução da Equação 2.37 tem-se:

$$\begin{aligned}\big(Y(x)-D\,\beta\big)'\big(R(\phi)\big)^{-1}\big(Y(x)-D\,\beta\big) &= \big(\beta-\hat{\beta}\big)'\big(D'\ \big(R(\phi)\big)^{-1}D\big)(\beta-\hat{\beta}) \\ &+ \ (Y(x)-D\,\hat{\beta})'\big(R(\phi)\big)^{-1}(Y(x)-D\,\hat{\beta})\end{aligned}$$

Gelman et al. (2003) sugerem utilizar a distribuição Normal χ^2_{ScI} para a distribuição conjunta de β e σ^2, ou seja,

$$\big[\beta;\ \sigma^2|\ \phi\big] \sim N\big(m_\beta;\sigma^2V_\beta\big)\chi^2_{ScI}\big(n_\sigma;W^2_\sigma\big) \tag{2.43}$$

Combinando-se a função de verosimilhança com a *priori* dada pela Equação 2.43 obtém-se a *posteriori*:

$$\big[\beta;\ \sigma^2|\ Y(x);\phi\big] \sim N\big(\hat{\beta}_N;\sigma^2V_{\hat{\beta}_N}\big)\chi^2_{ScI}\left(n_\sigma+n;\frac{W^2_1}{n_\sigma+n}\right) \tag{2.44}$$

onde:

$$\begin{aligned}W^2_1 &= n_\sigma W^2_\sigma + n\ \hat{\sigma}^2 + \hat{\beta}'V^{-1}_{\hat{\beta}}\hat{\beta} + m'_\beta V^{-1}_\beta m_\beta \\ &\quad -\big(V^{-1}_{\hat{\beta}}\hat{\beta}+V^{-1}_\beta m_\beta\big)'V_{\hat{\beta}_N}\big(V^{-1}_{\hat{\beta}}\hat{\beta}+V^{-1}_\beta m_\beta\big)\end{aligned}$$

Como:

- $P\big(\beta;\ \sigma^2|Y(x);\phi\big) = P\big(\beta|Y(x);\ \sigma^2;\phi\big)\ P\big(\sigma^2|Y(x);\phi\big)$,
- $\big[\beta|Y(x);\sigma^2;\phi\big] \sim N\big(\hat{\beta}_N;\sigma^2V_{\hat{\beta}_N}\big)$ e
- $\big[\sigma^2|Y(x);\phi\big] \sim \chi^2_{ScI}\big(n_\sigma+n;W^2_1\big)$

então:

$$\left[\beta;|Y(x);\phi\right] \sim t_{n_\sigma+n}\left(\hat{\beta}_N; W_1^2 V_{\hat{\beta}_N}\right)$$

d) Incerteza nos parâmetro de média, de escala e de correlação.

Além dos parâmetros de média e de escala serem desconhecidos, o acréscimo do parâmetro de correlação ϕ, também desconhecido, presente na matriz de correlações espaciais do modelo, será feito com a suposição de que modelo seja isotrópico e sem efeito pepita. O acréscimo de um efeito de anisotropia exigirá somente o acréscimo de mais dois parâmetros na mesma matriz de correlações, sendo um para controlar uma razão de anisotropia e outro para controlar um ângulo de inclinação de uma elipse associada ao efeito. Assim, estamos supondo que o parâmetro $\psi_R = 1$ e $\psi_A = 0$.

A *priori* conjunta $\left(\beta;\sigma^2;\phi\right)$ terá distribuição de probabilidades dada por:

$$P\left(\beta;\sigma^2;\phi\right) = P\left(\beta;\sigma^2|\,\phi\right)\,P\left(\phi\right)$$

e uma distribuição a *posteriori* dada por:

$$P\left(\beta;\sigma^2;\phi|\,Y(x)\right) = P\left(\beta;\sigma^2|\,Y(x);\phi\right)P\left(\phi|\,Y(x)\right)$$

onde $P\left(\beta;\sigma^2|\,Y(x);\phi\right)$ tem distribuição Normal-χ^2_{ScI} conforme visto na Equação 2.44 e $P\left(\phi|\,Y(x)\right) \propto P\left(Y(x)|\,\phi\right)P\left(\phi\right)$.

Para Ribeiro Jr e Diggle (1999), usando uma *priori* imprópria $P(\beta;\sigma^2|\phi) \propto \sigma^{-2}$ a *posteriori* para o parâmetro de correlação fica:

$$P\left(\phi|Y(x)\right) \propto P\left(\phi\right)|V_{\hat{\beta}}|^{\frac{1}{2}}|R_y|^{-\frac{1}{2}}\left(W^2\right)^{-\frac{n-p}{2}} \tag{2.45}$$

o que não define uma distribuição de probabilidades conhecida. Uma solução sugerida pelos autores é a utilização de inferência por simulação, discretizando a distribuição de $\left(\phi|Y(x)\right)$ e adotanto uma distribuição uniforme discreta para ϕ.

2.7.2 Predição linear espacial bayesiana

Segundo Ribeiro Jr e Diggle (1999), em problemas de variáveis espacialmente correlacionadas, frequentemente o interesse é a predição de uma variável $Y_0(x')$ em um novo conjunto de localizações x' para a elaboração de mapas temáticos e então o modelo deverá incluir essa nova variável. Entretanto, a questão de predição refere-se a afirmações sobre $Y_0(x')$ depois de

ter sido observada a amostra $Y(x)$, significando que o que se deseja de fato é $P(Y_0(x')|Y(x))$. DeGroot (1989) afirma que o preditor $E(Y_0(x')|Y(x))$ é um preditor ótimo pois minimiza o erro quadrático médio da predição.

O cenário geoestatístico bayesiano de interesse é aquele formado pela distribuição conjunta de $Y(x)$ e $Y_0(x')$, cujo modelo é escrito como:

$$(Y(x);Y_0(x')|\,\theta) \sim N\left(\begin{bmatrix} D \\ D_0 \end{bmatrix}\beta\,;\ \tau^2 I + \begin{bmatrix} V(\sigma^2;\phi) & v(\sigma^2;\phi) \\ v(\sigma^2;\phi)' & V_0(\sigma^2;\phi) \end{bmatrix}\right) \tag{2.46}$$

onde $V(\sigma^2;\phi)$ é a matriz de covariâncias formada pelas amostras de $Y(x)$ nas coordenadas x, $V_0(\sigma^2;\phi)$ é a matriz do covariâncias formada pelas estimativas de $Y_0(x')$ nas coordenadas x' e $v(\sigma^2;\phi)$ é a matriz de covariâncias cruzadas formada pelas coodenadas de $Y(x)$ e $Y_0(x')$. Sob uma distribuição gaussiana, considerando-se os parâmetros de θ conhecidos, a obtenção da distribuição conjunta dada pela Equação 2.46 será simples, tanto quanto será a distribuição marginal de $[Y(x)]$ e a condicional $\left[Y_0(x')|Y(x)\right]$. Todavia os parâmetros não são conhecidos e necessitam então ser estimados, sob o enfoque não bayesiano ou se deve integrar a preditiva sob a posteriori dos parâmetros no estimador bayesiano.

As predições de $Y_0(x')$ condicionalmente às observações de $Y(x)$ serão obtidas como:

$$\begin{aligned} P\big(Y_0(x')|Y(x)\big) &= \int P\big(Y_0(x');\theta|Y(x)\big)\,d\theta = \int \frac{P\big(Y_0(x');Y(x);\theta\big)}{P\big(Y(x)\big)}\,d\theta \\ &= \int \frac{P\big(Y_0(x')|Y(x);\theta\big)\,P\big(Y(x);\theta\big)}{P\big(Y(x)\big)}\,d\theta \\ &= \int \frac{P\big(Y_0(x')|Y(x);\theta\big)\,P\big(\theta|Y(x)\big)}{P\big(Y(x)\big)}\,P\big(Y(x)\big)\,d\theta \\ &= \int P\big(Y_0(x')|Y(x);\theta\big)\,P\big(\theta|Y(x)\big)\,d\theta \end{aligned} \tag{2.47}$$

que representa uma média ponderada de $(Y_0(x)|Y(x);\theta)$ sobre o espaço dos parâmetros $\theta = (\beta;\sigma^2;\phi;\tau^2)$ onde os pesos são determinados pela distribuição a *posteriori* conjunta $\left[\theta|Y(x)\right]$. Considerando as propriedades da distribuição gaussiana, e um modelo sem o efeito pepita, ou seja, $\tau^2 = 0$, a Equação 2.47 terá uma distribuição de probabilidades dada por:

$$\begin{aligned} \left[Y_0(x')|Y(x);\beta;\sigma^2;\phi\right] \sim N_n\Big(& D_0\beta + r'(\phi)\big(R(\phi)\big)^{-1}\left(Y(x) - D\beta\right)\,; \\ & \sigma^2\left(R_0(\phi) - r'(\phi)(R_Y(\phi))^{-1}\,r(\phi)\right)\Big) \end{aligned} \tag{2.48}$$

onde $\sigma^2 R_0(\phi)$ representa a variância sem se levar em consideração a informação da amostra e

$\sigma^2 r'(\phi)R_Y^{-1}(\phi)\, r(\phi)$ é a redução na variância devido à informação da amostra, cuja intensidade irá depender da configuração das localizações *x* (RIBEIRO JR; DIGGLE, 1999).

Busca-se na maneira bayesiana, incorporar incertezas na forma de uma distribuição de probabilidades, em todas as predições e na a distribuição a posteriori $\left[Y_0(x_0)\right]$. Considerou-se aqui todas as circunstâncias abordadas na Seção 2.7.1.

a) Predição com incerteza no parâmetro β

A distribuição a posteriori da predição de Y_0 considerando-se incerteza no parâmetro β e a Equação 2.47 será dada por:

$$\begin{aligned} P\big(Y_0(x')|\,Y(x);\sigma^2;\phi\big) &= \int_\beta P\big(y_0;\beta|\,y;\sigma^2;\phi\big)d\,\beta \\ &= \int_\beta P\big(y_0|y;\beta;\sigma^2;\phi\big)P\big(\beta|\,y;\sigma^2,\phi\big)\,d\,\beta \end{aligned}$$

O termo dentro da integral é uma expressão de uma normal bivariada, então, a integral irá resultar também uma distribuição normal na forma:

$$\left[Y_0(x')|\,Y(x);\sigma^2;\phi\right] \sim N\big(\mu_1;\sigma^2\Sigma_1\big) \tag{2.49}$$

onde:

$$\begin{aligned} \mu_1 &= E\left[Y_0(x')|\,Y(x);\sigma^2;\phi\right] \\ &= \big(D_0 - r'(\phi)R_Y^{-1}(\phi)D\big)\big(V_\beta^{-1} + D'R_Y^{-1}(\phi)D\big)^{-1}V_\beta^{-1}\mu_\beta \\ &+ \left[r'(\phi)R_Y^{-1}(\phi) + \big(D_0 - r'(\phi)R_Y^{-1}(\phi)D\big)\big(V_\beta^{-1} + D'R_Y^{-1}(\phi)D\big)^{-1}D'R_Y^{-1}(\phi)\right]Y(x) \end{aligned}$$

e

$$\begin{aligned} \Sigma_1 &= Var\big(Y_0(x')|\,Y(x);\sigma^2;\phi\big) \\ &= R_0^{-1}(\phi) - r'R_Y^{-1}(\phi)r \\ &+ \big(D_0 - r'R_Y^{-1}(\phi)D\big)'\big(V_\beta^{-1} + \big(D'R_Y^{-1}(\phi)D\big)^{-1}D\big)^{-1}\big(D_0 - r'R_Y^{-1}(\phi)D\big) \end{aligned}$$

b) Predição com incerteza no parâmetro σ^2

Neste caso, a distribuição preditiva para $Y_0(x')$ será dada por:

$$\begin{aligned} P\big(Y_0(x')|\,Y(x);\beta;\phi\big) &= \int_{\sigma^2} P\big(y_0;\sigma^2|\,y;\beta;\phi\big)d\,\sigma^2 \\ &= \int_{\sigma^2} P\big(y_0|y;\beta;\sigma^2;\phi\big)\,P\big(\sigma^2|\,y;\beta,\phi\big)\,d\,\sigma^2. \end{aligned}$$

Substituindo-se o termo da verossimilhança (com distribuição normal) e o termo da *priori* (com distribuição χ^2_{ScI}) a solução analítica resultará em uma distribuição a *posteriori* para $Y_0(x')$ com distribuição $t-student$ dada por:

$$\left[Y_0(x')|Y(x);\beta;\phi\right] \sim t_{n_\sigma+n}\left(\mu_0; Q_0\,\Sigma_1\right) \tag{2.50}$$

onde:

$$\begin{aligned}
\mu_0 &= D_0\beta + r'(\phi)R_Y^{-1}(\phi)(y-D\beta),\\
Q_0 &= \frac{n_\sigma W_\sigma + n\hat{\sigma}^2}{n_\sigma+n},\\
\Sigma_0 &= R_Y^{-1}(\phi) - r'(\phi)R_Y^{-1}(\phi)r(\phi).
\end{aligned}$$

Assim:

$$\begin{aligned}
E\left[Y_0(x')|Y(x)\right] &= \mu_0;\\
Var\left[Y_0(x')|Y(x)\right] &= \left(\frac{n_\sigma+n}{n_\sigma+n-2}\right) Q_0\,\Sigma_0.
\end{aligned}$$

c) Predição com incerteza nos parâmetros β e σ^2

Neste caso, a distribuição preditiva para $Y_0(x')$ será dada por:

$$\begin{aligned}
P\left(Y_0(x')|Y(x);\phi)\right) &= \int_{\sigma^2}\int_{\beta} P\left(y_0;\beta;\sigma^2|y;\phi\right)d\beta\,d\sigma^2\\
&= \int_{\sigma^2}\int_{\beta} P(y_0;\beta|y;\sigma^2;\phi)P(\sigma^2|y;\phi)\,d\beta\,d\sigma^2.
\end{aligned}$$

Integrando-se em relação a β vem:

$$P\left(Y_0(x')|Y(x);\phi\right) = \int_{\sigma^2} P(y_0|y;\sigma^2;\phi)P(\sigma^2|y;\phi)\,d\sigma^2.$$

O primeiro termo da integral corresponde à distribuição preditiva dada pela Equação 2.49 e o segundo termo corresponde a distribuição a *posteriori* marginal $P\left(\sigma^2|Y(x)\right)$. Para as *prioris* aqui adotadas, a distribuição a posteriori ficará:

$$\left[Y_0(x')|Y(x);\phi\right] \sim t_{n_\sigma+n}\left(\mu_{1N}; W_1^2\,\Sigma_{1N}\right) \tag{2.51}$$

onde μ_{1N} e Σ_{1N} dependerão da escolha para a priori de β, e então:

$$\begin{aligned}
E\left[Y_0(x')|Y(x)\right] &= \mu_{1N};\\
Var\left[Y_0(x')|Y(x)\right] &= \left(\frac{W_1^2\,\Sigma_{1N}}{n_\sigma+n-2}\right).
\end{aligned}$$

d) Predição com incerteza nos parâmetros β, σ^2 e ϕ

Neste caso, a distribuição preditiva para $Y_0(x')$ será dada por:

$$\begin{aligned} P\big(Y_0(x')|Y(x)\big) &= \int_\phi \int_\beta \int_{\sigma^2} P\big(y_0;\beta;\sigma^2;\phi|y\big)\, d\beta\, d\sigma^2\, d\phi \\ &= \int_\phi \left[\int_\beta \int_{\sigma^2} P\big(y_0;\beta;\sigma^2|y;\phi\big) d\beta\, d\sigma^2\right] P\big(\phi|Y(x)\big) d\phi \\ &= \int_\theta P\big(y_0;|y;\phi\big) P\big(\phi|Y(x)\big)\, d\theta \end{aligned}$$

onde $\big(Y_0(x')|Y(x)\big) \sim t_{n_\sigma+n}\big(\mu_{1N};W_1^2\,\Sigma_{1N}\big)$ conforme Equação 2.51 e $\big(\phi|Y(x)\big)$ é obtido conforme Equação 2.45.

Para a inclusão de um efeito pepita (*nugget*), escreve-se a matriz de correlação na forma:

$$R\big(\phi;\nu\big) = \sigma^2\Big[R(\phi) + \nu^2 I\Big] \tag{2.52}$$

onde $\nu^2 = \dfrac{\tau^2}{\sigma^2}$ corresponde ao efeito pepipa relativo e assim, a distribuição a *posteriori* para esse parâmetro é obtida discretizando-se $\big(\phi;\nu^2\big)$

2.8 APLICAÇÃO DO MODELO GEOESTATÍSTICO UNIVARIADO

2.8.1 Estudo de caso

Caso 1: Produtividade de Soja em área comercial

Aqui foram utilizados os dados de pesquisa do Núcleo de Inovações Tecnológicas – NIT, da Universidade Estadual do Oeste do Paraná – Unioeste. A área de Latossolo vermelho distrófico, com declividade média de 0,19%, em área de 1,74 ha, está localizada localizado no Centro de Pesquisa Eloy Gomes, da Cooperativa Central Agropecuária de Desenvolvimento Tecnológico e Econômico Ltda. – COODETEC, situada na BR 467, km 98, em Cascavel-PR. Nessa área, no final do ano de 1997, cultivou-se soja em sistema de semeadura direta. Em abril

de 1998, a produção de cada parcela foi colhida e pesada. Simultaneamente foram tomadas, em cada parcela, amostras do solo para a análise química.

Para modelar a estrutura de variabilidade espacial e correlacioná-la com uma variável da cultura implantada na área, foram utilizados os atributos químicos: pH, Matéria Orgânica – MO (%), Potássio – K ($Cmol_c\ dm^{-3}$), Fósforo – P ($mg\ dm^{-3}$) e Índice de Saturação de Bases – SB (%) e o atributo físico resistência mecânica à penetração no solo, expressa pelo índice de cone – iCone ($Kg\ cm^{-2}$)

As amostras foram obtidas com 7 cm de diâmetro e 15 cm de profundidade dentro de cada uma das 256 parcelas, estruturadas em um grid de 7,20 $\times$ 7,20 m, com carreador de 2,4 m em uma das direções, usando-se o sistema desalinhado, sistemático estratificado de Wollenhaupt e Wolkowski (1994) e adaptados por Souza et al. (1999). Para a produtividade, foram colhidas e identificadas as parcelas de 5,0 $\times$ 5,0 m, excluídos bordaduras e carreador (Figura 2.10).

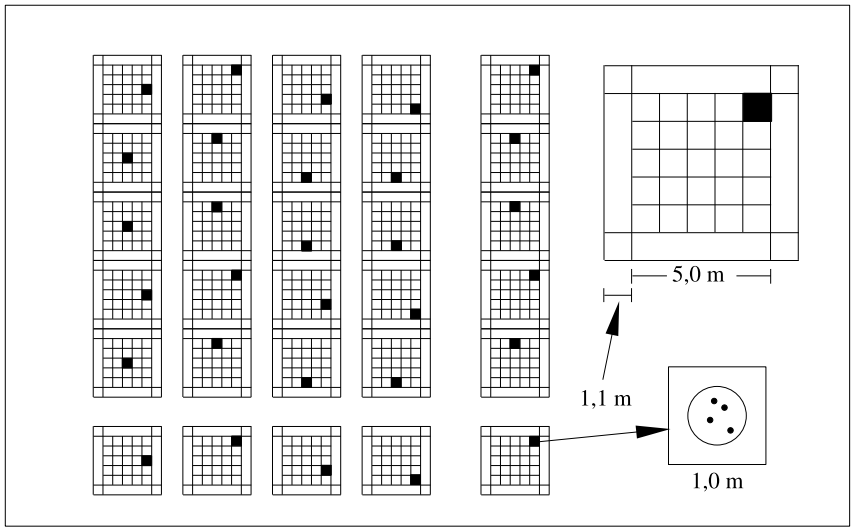

Figura 2.10: Esquema de mostragem com locação das parcelas e pontos amostrais em sistema desalinhado, sistemático estratificado porposto por Wollenhaupt e Wolkowski (1994) e adaptados por Souza et al. (1999).

Caso 2: Rendimento de *Pinus taeda* L. em área de reflorestamento comercial

Foi utilizado também um segundo conjunto de dados provenientes de banco de dados cartográficos gerado em padrões e formatos ArcGis, disponíveis no setor de informações geográficas da Universidade Federal do Paraná - UFPr e da empresa Modo Battistella Reflorestamento S/A – MOBASA. O estudo foi desenvolvido em parcelas de inventários florestais contínuos com plantio de Pinus da espécie *P. Taeda* L. em fazendas situadas no município de Rio Negrinho no Estado de Santa Catarina sob domínio das bacias e coberturas sedimentares na

região do patamar oriental da Bacia do Paraná e na unidade do patamar de Mafra-SC. Trata-se de uma área de 2.252 ha localizada no Norte do Estado de Santa Catarina onde o relevo é quase plano, com cotas altimétricas diminuindo de leste para oeste, atingindo valores entre 650 a 740 m. A geologia é representada pelo grupo Itararé compreendendo todo o pacote de sedimentos de origem glacial e periglacial relacionado ao carbonífero superior e permiano inferior. O experimento foi em delineamento completamente casualizado e dados não balanceados em que as idades de coleta das informações das parcelas de inventário florestal contínuo foram consideradas como repetições do modelo. Os três fatores qualitativos considerados foram os municípios (três), as fazendas (sete) e as parcelas, com idades de 11, 12, 13, 14 e 15 anos (BOGNOLA, 2007).

Nessa área foram efetuados levantamentos pedológicos com prospecção por tradagem e em perfis em barrancos de estrada, acompanhada de coleta de amostras para análises químicas de: $pH(CaCl_2)$, Fósforo disponível (P), Potássio disponível (K), Al^{3+}, Carbono orgânico, $H + Al^{3+}$, Soma de bases (SB), capacidade de troca catiônica (CTC) e Saturação por bases ($V\%$) e análises granulométricas de: areia, silte e argila.

Foram ainda analisadas 18 árvores com idades que variavam de 11 a 15 anos, nas quais foram medidas o diâmetro (cm) a 1,3 m de altura, a altura média (m) das árvores da parcela, o número de árvores por hectare, a altura dominante (m) das 10 maiores árvores, a área basal (m^3), o volume médio ($m^3\ ha^{-1}$) e incremento médio anual - IMA (m^3). Essas foram consideradas as variáveis principais por estarem relacionadas com algum interesse econômico.

O delineamento geoestatístico foi feito em pontos aleatórios da área registrando-se os pontos amostrais em coordenadas ortogonais UTM (*Universal Transverse Mercator*) com auxílio de aparelho de posicionamento por satélite GPS e anotando-se, para cada localização, a análise do material geológico, as medições das árvores, a profundidade efetiva do perfil do solo (horizontes A + B), altura estimada do lençol freático, a posição na encosta, o percentual de declividade e a altitude.

2.8.2 Recursos computacionais

Foram adotados ao longo das análises estatísticas deste estudo, recursos computacionais baseados em programas livres, suportados pela licença internacional GPL – *General Public Licence*. O sistema operacional foi o GNU/Linux e o pacote estatístico foi o R (R Development Core Team, 2008) e módulo geoestatístico geoR (RIBEIRO JR; DIGGLE, 2001) na

versão 1.6-20. Com esse pacote geoestatístico foi possível ajustar modelos teóricos univariados e bivariados válidos, ajustar modelos lineares de corregionalização, efetuar estimação linear e simulações por krigagem e cokrigagem e produzir gráficos e mapas temáticos. O pacote estatístico, além do suporte às funções dos pacotes geoestatísticos, permitiu a análise convencional dos dados bem como a ACP.

2.8.3 Análise geoestatística dos dados de produtividade de soja

Foi dado inicialmente um enfoque estatístico tradicional para a análise do conjunto de variáveis aleatórias medidas na área. Empregou-se uma análise descritiva, visando identificar e avaliar a estrutura de pontos discrepantes, verificar homogeneidade e tendência direcional, bem como obter indicadores de atendimento aos pressupostos de um modelo geoestatístico e para referências exploratórias no ajuste de parâmetros. Em seguida foram ajustados e analisados modelos teóricos para o processo geoestatístico gaussiano que determinou a produtividade de soja na área. Com o modelo foi possível construir mapas temáticos procurando identificar classes diferenciadas de produtividade em função dos resultados obtidos. Foi feita também uma análise descritiva das predições obtidas nos mapas visando confrontar os resultados com o valor da produtividade conhecida na área a partir de 256 amostras. A densidade amostral foi de 157 pontos por hectare, valor esse considerado bem acima do usualmente praticado no manejo agrícola pelo alto custo que representa.

Análise descritiva das amostras de produtividade de soja e de variáveis Físicas e Químicas do solo

As amostras foram analisadas em três enfoques sendo o primeiro constituído de todos os 256 pontos amostrais disponibilizados pelo experimento original, o segundo por 128 parcelas tomadas aleatóriamente das 256 coordenadas originais e um terceiro com 64 parcelas também tomadas aleatóriamente das 256 coordenadas originais. As localizações das amostras está representada na Figura 2.11.

A soma das áreas das parcelas totalizou 6.400 m^2, correspondendo a 36,7% da área total cultivada. A colheita resultou então em 1,758 t de soja que, projetada para a área total cultivada resultou em 4,789 t. Essa produção corresponde a 2,75 t ha^{-1} que é compatível com os 2,74 t ha^{-1} registrados pelo IBGE na safra de 2003 e com as 2,55 t ha^{-1} da safra 97/98 do

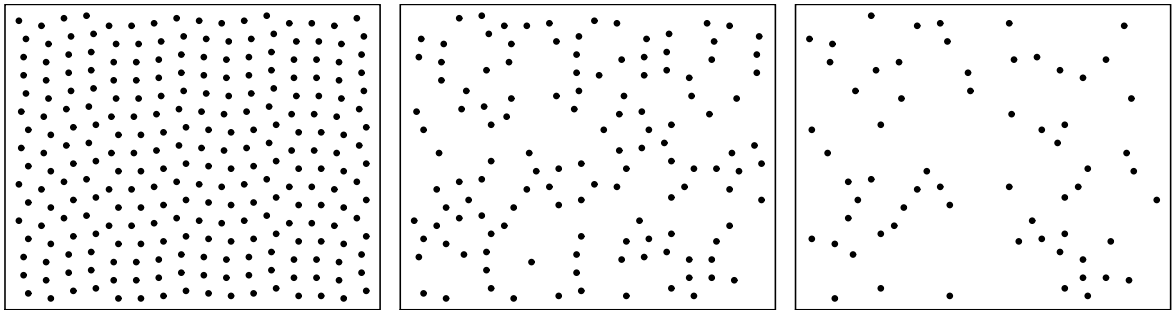

Figura 2.11: Localização das amostras na área de cultivo. Os delineamentos amostrais comportam, da esquerda para a direita, 256 pontos origalmente estruturados pelo sistema sistemático desalinhado estratificado, 128 e 64 pontos sorteados dos 256 pontos originais. O eixo horizontal corresponde a distância total de 141,2 m e o eixo vertical 115,2 m.

Paraná segundo levantamento feito pela CONAB em abril de 2006. Tomou-se então como valor referência a produtividade de 2,75 t ha^{-1}, projetada na área como única informação disponível.

A Tabela 2.1 mostra os principais resultados descritivos sobre o resultado da colheita. Nota-se que os valores da média de produtividade aumentam quando se diminui o tamanho da amostra. Considerando ensaios agrícolas de campo, Gomes (1963) classificou os experimentos com base no coeficiente de variação – CV, onde valores superiores a 20% já indicam perda de precisão. Para a produtividade de soja o CV se manteve abaixo de 20%, valor este sugerindo homogeneidade das medidas. Os intervalos de confiança de 95% para o parâmetro de transformação λ de normalidade de Box-Cox (BOX; COX, 1964) incluem o valor unitário. Assim, optou-se por não transformar a variável resposta e considerou-se os dados de produtividade compatíveis com uma distribuição normal de probabilidades. Os perfis de verossimilhança para o parâmetro λ de transformação podem ser vistos na Figura 2.12.

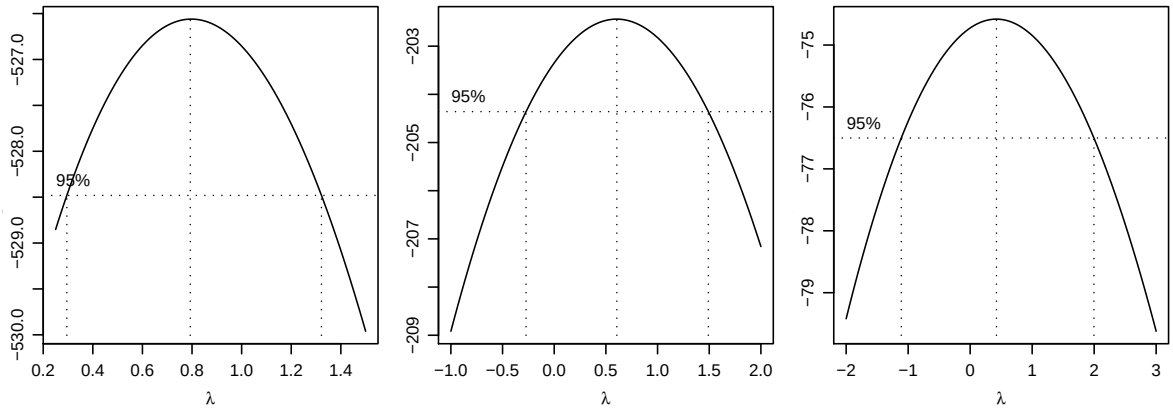

Figura 2.12: Perfil do log-verossimilhança para o parâmetro λ de transformação de Box-Cox. Intervalo de 95% de confiança que contenha o valor unitário implica em normalidade da distribuição dos dados. Da esquerda para a direita as figuras representam o log da função de verossimilhança para o parâmetro λ com relação aos delineamentos amostrais de soja em 256 pontos estruturados, 128 e 64 pontos sorteados.

Tabela 2.1: Estatística descritiva da variável soja medida em 256 pontos estruturados (Soja256), 128 e 64 pontos sorteados (Soja128 e Soja64, respectivamente) dentre os 256 pontos originais disponíveis.

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.%	LI-Bx	LS-Bx
Soja256	1,190	2,7464	4,140	0,4904	17,85	0,3	1,3
Soja128	1,710	2,7544	4,140	0,4328	15,71	-0,3	1,3
Soja64	1,930	2,8080	3,700	0,4057	14,45	-1,1	2,0

Medidas em ton ha^{-1}. D.P.: Desvio Padrão, C.V.: Coeficiente de Variação, LI-Bx e LS-Bx: Respectivamente o limite inferior e limite superior do intervalo de 95% de confiança para o parâmetro λ de Box-Cox.

As variáveis de natureza física e química do solo (secundárias), apresentadas na Tabela 2.2, mostram resultados compatíveis com a literatura. Excetuando-se MO e pH, as variáveis apresentam coeficiente de variação acima de 20%, evidenciando a homogeneidade das medidas. As variáveis P, pH e K não apresentam o valor unitário no intervalo de confiança de Box-Cox, sugerindo a necessidade de transformação em aplicações que exijam o pressuposto de normalidade. As variáveis MO, SB e iCone contém o valor unitário no intervalo, sugerindo normalidade nas medidas, dispensando transformações.

Tabela 2.2: Estatísticas descritivas das variáveis secundárias P, pH, K, MO, SB e iCone, todas tomadas nos mesmos 150 pontos aleatórios, selecionados dos 256 disponíveis.

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.%	LI-Bx	LS-Bx
P	2,000	4,047	13,100	1,439	36,60	-1,2	-0,3
pH	4,300	5,145	6,600	0,505	9,80	-3,5	-0,5
K	0,170	0,331	0,550	0,085	25,70	-0,6	0,6
MO	36,55	52,680	68,350	6,392	12,13	0,3	2,6
V%	14,89	55,470	84,030	12,896	23,20	0,4	1,4
iCone	10,50	20,300	31,100	4,163	20,50	0,3	1,6

P em mg dm^{-3} , pH sem unidade de medida, K em cmolc dm^{-3}, MO em g dm^{-3}, SB em porcentagem e iCone em kg cm^{-2}. D.P.: Desvio Padrão, C.V.: Coeficiente de Variação, LI-Bx e LS-Bx: Respectivamente o limite inferior e limite superior do intervalo de 95% de confiança para o parâmetro λ de Box-Cox.

Análise espacial das amostras de produtividade de soja

Devido a dificuldade de se ajustar um modelo teórico confiável ao semivariograma experimental pelo método dos mínimos quadrados ou suas variações (Figura 2.7), optou-se nesse trabalho por se obter uma estimativa pontual para os parâmetros do modelo geoestatístico por MV, MVR e por métodos bayesianos. A Tabela 2.3 mostra os resultados obtidos por MV. Nela nota-se que a estimativa para o parâmetro β, que corresponde à produtividade média, é próximo

do valor de referência de 2,75 t ha^{-1}, ou seja, o erro relativo, dado pela diferença entre o valor observado nas amostras e o valor médio estimado pelo método não superou 8%. Destaca-se também nessa tabela que o CEP apresentou, para amostras de 256 e 128 pontos, moderada dependência espacial e para amostras de 64 pontos, uma fraca dependência espacial, segundo classificação de Cambardella et al. (1994). Foi adotado no modelo, a função de correlação de Matèrn com $\kappa = 0,5$ e média constante.

Tabela 2.3: Estimação dos parâmetros do modelo geoestatístico por MV.

Delineamento	β	τ^2	σ^2	ϕ	CEP (%)	$-\log L$
Soja256	2,6583	0,1889	0,0725	63,8	72	168
Soja128	2,7220	0,1359	0,0620	50,0	69	67
Soja64	2,8083	0,1588	0,0033	38,6	98	33

β: parâmetro do efeito sistemático do modelo, σ^2 e ϕ: parâmetros da função de correlação, τ^2: parâmetro do erro, CEP: coeficiente de efeito pepita e $\log L$: valor de MV. A função de correlação adotada foi a de Matèrn com $\kappa = 0,5$.

A Tabela 2.4 mostra os resultados obtidos pelo método MVR. Nela a estimativa para o parâmetro β, que corresponde à produtividade média, segue a tendência observada na Tabela 2.3. O parâmetro ϕ de alcance forneceu valores muito maiores que a distância máxima da área. Como os valores máximos do logarítimo da função de verosimilhança são equivalentes tanto para o método MV quanto no método MVR, esse último não foi empregado nas análises seguintes e na produção de mapas.

Tabela 2.4: Estimação dos parâmetros do modelo geoestatístico pelo método MVR.

Delineamento	β	τ^2	σ^2	ϕ	CEP (%)	$-\log L$
Soja256	2,5454	0,1904	0,4495	$\gg 177$	30	165
Soja128	2,6649	0,1389	0,2440	$\gg 177$	36	66
Soja64	2,8160	0,1535	4,1380	$\gg 177$	4	32

β: parâmetros do efeito sistemático do modelo, σ^2 e ϕ: parâmetros da função de correlação, τ^2: parâmetro do erro, κ: parâmetro de diferenciabilidade da função de correlação, CEP:coeficiente de efeito pepita e $\log L$: valor MVR

Para se avaliar o comportamento espacial de uma variável é usual fazer a sua representação através de um mapa dos resultados obtidos, locados em suas coordenadas de coleta. O mapa da produtividade de soja, ilustrado na Figura 2.13, foi classificado segundo os quantis 20%, 40%, 60% e 80%. Os retângulos mais claros correspondem às baixas produtividades e os mais escuros às altas produtividades. As classes corresponderam a valores abaixo de 2,34 t ha^{-1}, representado por seu ponto médio de 1,8 t ha^{-1}, valores entre 2,34 e 2,61 t ha^{-1},

representados por deu ponto médio de 2,5 t ha^{-1}, valores entre 2,61 e 2,85 t ha^{-1}, representados por seu ponto médio de 2,7 t ha^{-1}, valores entre 2,85 e 3,16 t ha^{-1}, representados por se ponto médio de 3,0 t ha^{-1} e valores acima de 3,16 t ha^{-1}, representados por seu ponto médio de 3,7 t ha^{-1}. Essa classificação não seguiu um critério agronômico, servindo apenas para identificar zonas diferenciadas de produtividade.

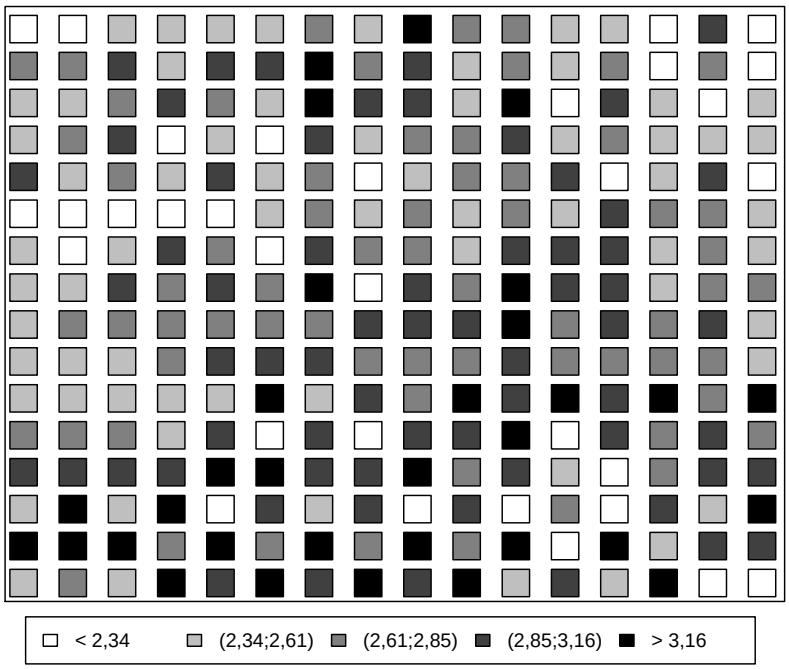

Figura 2.13: Gráfico de padrões de intensidade por parcela colhida classificado pelos quantis de produtividade 20, 40, 60 e 80%. A largura da figura corresponde a uma distância de 141,2 m e a altura 115,2 m. Cada retângulo corresponde a uma área de 25 m^2.

Avaliou-se os modelos geoestatísticos com parâmetros otimizados por MV através da validação cruzada. No caso da amostra de 256 coordenadas utilizou-se a estatégia de retirar "uma amostra por vez" e estimá-la com o modelo ajustado. Nas amostras de 128 e 64 coordenadas sorteadas do conjunto de 256 pontos originais, utilizou-se a estratégia de estimar os valores no conjunto complementar (onde as amostras não foram utilizadas). Os resultados estão apresentados no anexo A. Nessas figuras, a validação resultou em erros de predição com uma distribuição gaussiana de probabilidades em torno do zero com uma variância de predição menor que a variância dos dados. Não se identifica também um padrão de regionalização dos erros, como por exemplo uma concentrações de valores de superestimação ou subestimação.

Na Figura 2.14 ilustra-se a distribuição espacial da produtividade de soja estimada por krigagem a partir de modelo ajustado por MV, em 690 pontos de uma malha regular. Os pontos

amostrais têm dimensão 5 $\times$ 5 m, compatível com o suporte de medida de 25 m^2 adotados na colheita das parcelas. Os resultados também foram classificados segundo os quantis 20, 40, 60 e 80%, observado nos dados, permitindo assim a comparação com o mapa da Figura 2.13. O mapa à esquerda corresponde às predições com base nos 256 pontos amostrais originais, onde, conforme Tabela 2.6, somente 0,7% das predições ficaram abaixo de 2,34 t ha^{-1} e nenhuma acima de 3,16 t ha^{-1}, entretando identificou-se zonas de produtividade diferenciadas. O mapa central corresponde às predições com base em 128 pontos amostrais, onde, conforme a mesma tabela, apresentou uma concentração maior dos valore nas classes centrais. Nenhum resultado ficou abaixo de 2,34 t ha^{-1} e nenhum acima de 3,16 t ha^{-1}. Esse mapa também identificou zonas diferenciadas. No mapa da direita, as predições foram efetuadas com base em 64 pontos amostrais e ficaram todas na classe central, entre 2,61 t ha^{-1} e 2,85 t ha^{-1}. O mapa correspondente não reproduziu um padrão espacial capaz de identificar zonas de produtividade diferenciadas segundo esses critérios de classificação.

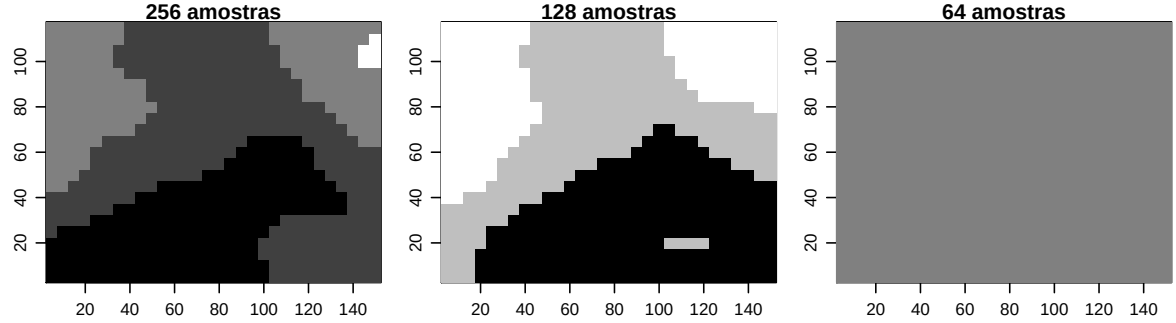

Figura 2.14: Mapas de produtividade de soja estimados por krigagem convencional a partir de modelo ajustado por MV, em uma malha regular de 690 pontos a partir de 256 (esquerda), 128 (centro) e 64 (direita) pontos amostrais. Os pontos brancos correspondem às produtividades abaixo de 2,34 t ha^{-1} e os pontos pretos às produtividades acima de 3,16 t ha^{-1}. Os pontos em escalas cinza correspondem às produtividades intermediárias.

A Tabela 2.5 mostra os resultados descritivos da krigagem convencional em 690 pontos regularmente distribuídos pela área, conforme justificado anteriormente. Nota-se nela que a média dos valores estimados ficou próximo do valor de referência de 2,75 t ha^{-1} nos três delineamentos amostrais, mantendo um erro relativo absoluto de predição abaixo de 2,1 %. Houve também uma uniformidade nos resultados, mantendo o CV abaixo de 8%. Nota-se que a krigagem recuperou a informação da produtividade média da área, diferentemente da descrição da variabilidade espacial mostrada pelo respectivos mapas.

Na tabela 2.6 se apresenta a porcentagem dos pontos estimados, incidentes em cada intervalo de classificação, segundo três tipos de amostragem utilizados no modelo univariado. Nela verifica-se, quando comparada com a distribuição percentual dos dados originais,

Tabela 2.5: Estatística descritiva das predições por krigagem convencional da produtividade de soja medida em uma malha de 690 pontos, com base em amostras de 256, 128 e 64 pontos.

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.(%)	ER(%)
Soja256	2,3285	2,7422	3,1365	0,1784	6,5	-0,27
Soja128	2,3946	2,7562	3,0566	0,1736	6,3	0,24
Soja64	2,7792	2,8078	2,8360	0,0161	0,6	2,12

Medidas em t ha^{-1}. D.P.: Desvio Padrão, C.V.: Coeficiente de Variação, ER(%): erro em relação ao valor referência de 2,75 t ha^{-1}.

a tendência do número de pontos preditos se concentrarem na classe central, tanto na predição através da utilização dos 256 pontos amostrados quanto na predição que envolveu uma redução no número de amostras.

Tabela 2.6: Porcentagem dos pontos estimados por método de krigagem com modelo univariado, incidentes em cada intervalo de classificação, segundo três tipos de amostragem.

Classe	PM	S256	S128	S64	S256d
1,19 a 2,34	1,8	0,7	0,0	0,0	20,3
2,34 a 2,61	2,5	22,9	25,2	0,0	19,9
2,61 a 2,85	2,7	46,2	37,4	100,0	20,7
2,85 a 3,16	3,0	29,1	37,4	0,0	19,1
3,16 a 4,14	3,7	0,0	0,0	0,0	19,9

PM: ponto médio da classe, S256: soja em 256 amostras, S128: soja em 128 amostras, S64: soja em 64 amostras, IC: índice de cone, CP1: primeira componente principal, S256d: dados originais de soja classificados.

Análise espacial bayesiana das amostras de produtividade de soja

Para inferência bayesiana sobre os parâmetros do modelo foi considerado um modelo isotrópico, sem tendência direcional ou efeito sistemático e função de correlação de Matèrn com parâmetro de diferenciabilidade $\kappa = 0.5$. Adotou-se as distribuições a *priori* plana ("flat") para o parâmetro β, recíproca para σ^2 e uniforme discreta para ϕ. Considerou-se ainda o valor de τ^2 relativo igual a zero, ou seja, ausência do efeito de pequena escala. A opção por tais procedimentos se deu por conveniência computacional, cabendo uma exploração mais detalhada sobre o assunto em trabalhos futuros. A Figura 2.15 mostra a distribuição a *posteriori* para os parâmetros β e σ^2 no caso em que se utilizou a amostra de produtividade de soja em 256 pontos amostrais. As Figuras 2.16 e 2.17 correspondem às distribuições a *posteriori* para os parâmetros β e σ^2 no caso em que se utilizou amostra de produtividade de soja em 128 e 64 pontos amostrais, respectivamente. A Tabela 2.7 apresenta as estimativas pontuais dos parâmetros, obtidos pela média de 1.000 simulações bayesianas obtidas pelas respectivas distribuições a *posteriori*. As estimativas dos parâmetros β nessa tabela são compatíveis com

os resultados obtidos por MV mas as estimativas dos parâmetros σ^2 foram aumentadas devido à incorporação de incertezas no modelo, próprio do método bayesiano.

Tabela 2.7: Média da posteriori dos parâmetros do modelo geoestatístico obtido por inferência bayesiana.

Delineamento	β	σ^2	ϕ
Soja256	2,7203	0,3072	7,20
Soja128	2,7604	0,2108	7,06
Soja64	2,8126	0,1945	6,78

β é o parâmetro do efeito sistemático do modelo, σ^2 e ϕ são parâmetros da função de correlação, todos obtidos pela média de 1.000 aproximações numéricas pelo método bayesiano. Neste processo, τ^2 é o parâmetro do erro, fixado em zero.

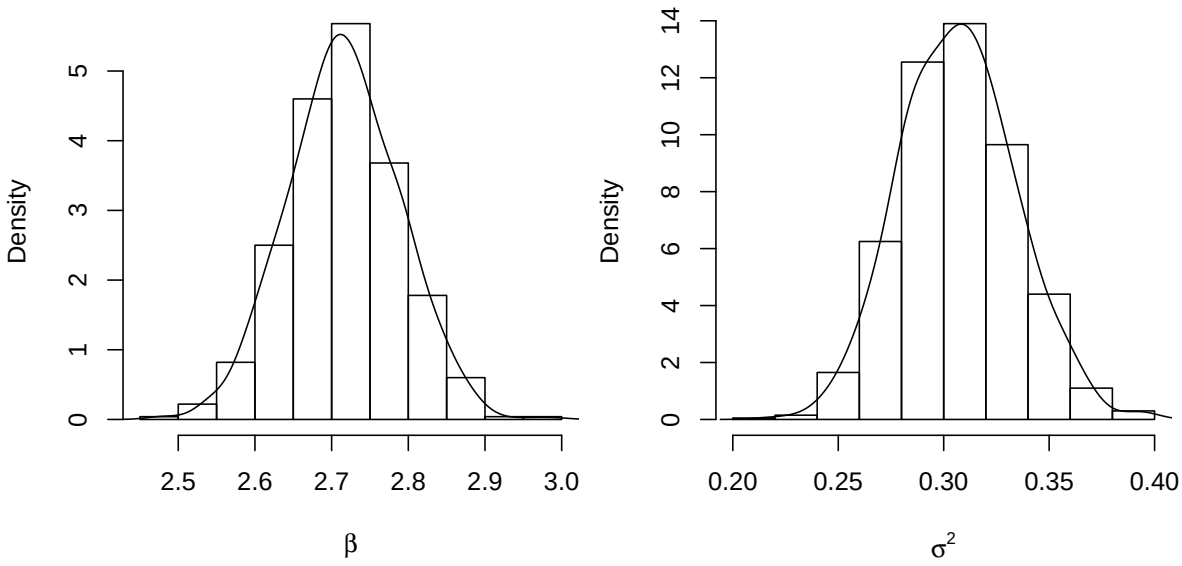

Figura 2.15: Distribuição a *posteriori* para os parâmetros β e σ^2 com 50 níveis de ϕ e 1.000 aproximações numéricas a partir de um grupo de 256 amostras.

A Figura 2.18 ilustra a distribuição espacial da predição da produtividade de soja obtida pela média de 1.000 simulação preditivas em uma malha de 690 pontos, com amostras de 256, 128 e 64 pontos, apresentadas da esquerda para a direita, respectivamente, com resultados classificados pelos quantis 20, 40, 60 e 80% observado nos dados. Os pontos em tons de cinza correspondem a diferentes classes de produtividade. Nessa figura, no caso de 256 amostras, observa-se um padrão da distribuição espacial semelhante ao apresentado na Figura 2.13. Esse resultado diferiu do obtido pelo método de krigagem a partir de valores pontuais dos parâmetros, estimados por MV. Perdeu-se a definição de zonas diferenciadas de produtividade, como o apresentado na Figura 2.14, contudo apresentou um padrão de variabilidade espacial, em todos os tamanhos de amostras.

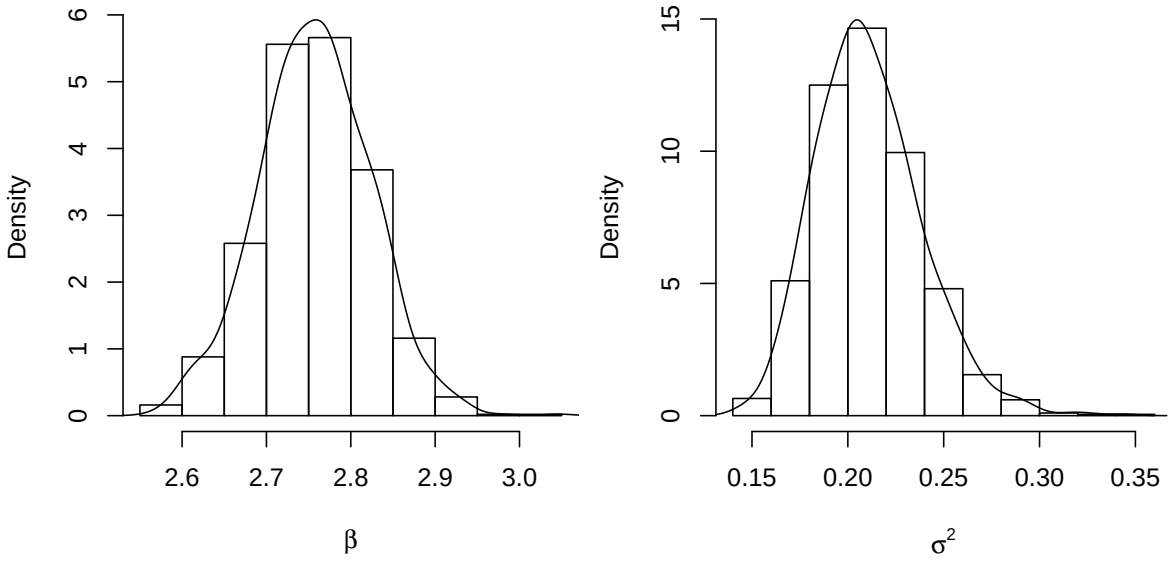

Figura 2.16: Distribuição a *posteriori* para os parâmetros β e σ^2 com 50 níveis de ϕ e 1.000 aproximações numéricas a partir de um grupo de 128 amostras.

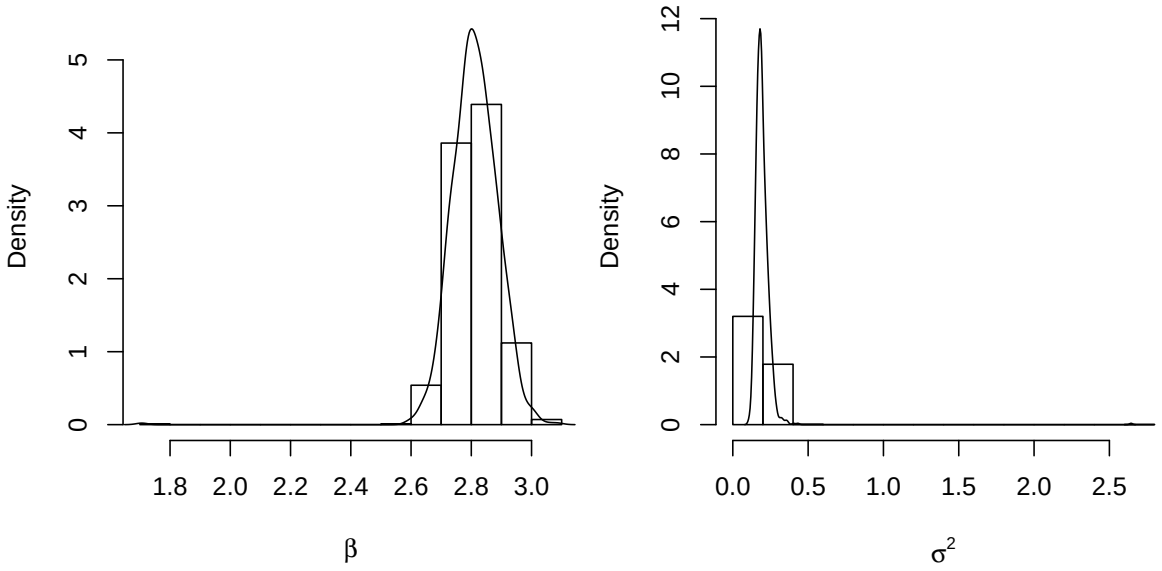

Figura 2.17: Distribuição a *posteriori* para os parâmetros β e σ^2 com 50 níveis de ϕ e 1.000 aproximações numéricas a partir de um grupo de 64 amostras.

A Tabela 2.8 apresenta resultados descritivos da predição de soja por simulação bayesiana com base nos três diferentes delineamentos amostrais já mencionados. Esses resultados indicaram a manutenção no erro relativo comparado com o resultado obtido pela predição por krigagem convencional, conforme visto na Tabela 2.5 e coeficiente de variação da mesma ordem de grandeza, garantindo-se a uniformidade dos resultados.

Na Tabela 2.9 apresenta-se a porcentagem dos pontos estimados por método bayesi-

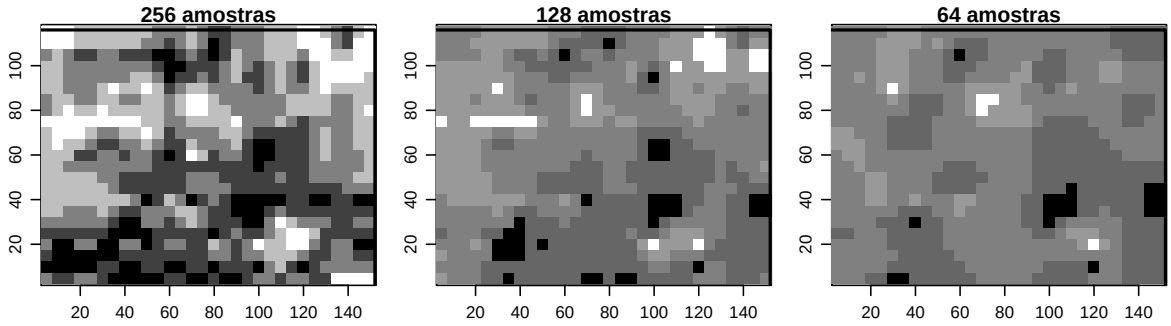

Figura 2.18: Mapas de produtividade de soja estimados por inferência bayesiana em uma malha regular de 690 pontos com base em 256 (esquerda) 128 (centro) e 64 (direita) pontos amostrais. Em cada mapa a largura corresponde a 141,2 m e a altura 115,2 m. Os pontos brancos correspondem às produtividades abaixo de 2,4075 t ha^{-1}, os cinza às produtividades entre 2,4075 e 3,045 t ha^{-1} e os pretos ‘as produtividades acima de 3,045 t ha^{-1}.

Tabela 2.8: Estatísticas descritivas das predições bayesianas da produtividade de soja medida em uma malha de 1.131 pontos, com base em amostras de 256 pontos estruturados (Soja256), 128 e 64 pontos aleatórios (Soja128 e Soja64, respectivamente) tomados dos 256 disponíveis

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.%	ER(%)
Soja256	1,6757	2,7460	3,7000	0,3186	11,6	-0,13
Soja128	1,9606	2,7616	3,6255	0,2461	8,9	0,44
Soja64	2,1471	2,8069	3,5022	0,1675	6,0	2,08

Medidas em t ha^{-1}. D.P.: Desvio Padrão, C.V.: Coeficiente de Variação, ER(%): erro em relação ao valor referência de 2,7496 t ha^{-1}.

ano, incidentes em cada intervalo de classificação, segundo três tipos de amostragem utilizados no modelo univariado. Nela verifica-se, quando comparada com a distribuição percentual dos dados originais segundo os intervalos de classe, a tendência do número de pontos preditos se concentrarem na classe central, tanto na predição através da utilização dos 256 pontos amostrados quanto na predição que envolveu uma redução no número de amostras. Todavia o efeito é menos acentuado quando comparado aos resultados da Tabela 2.6.

2.8.4 Análise geoestatística dos dados rendimento de *P. Taeda L.*

Análogo ao caso dos dados de produção de soja, visto na Seção 2.8.3, aqui foi dado inicialmente um enfoque estatístico tradicional pela análise descritiva básica, visando-se também identificar e avaliar a estrutura de pontos discrepantes, homogeneidade e tendência direcional, bem como obter indicadores de atendimento aos pressupostos de um modelo geoestatístico e às referências exploratórias na obtenção de estimativas dos parâmetros. Foram ajustados e analisados modelos teóricos para o processo geoestatístico supostamente gaussiano que determinou

Tabela 2.9: Porcentagem dos pontos estimados por método bayesiano com modelo univariado, incidentes em cada intervalo de classificação, segundo três tipos de amostragem.

Classe	PM	S256	S128	S64	S256d
$[1,19\,;\,2,34)$	1,8	9,1	3,8	0,7	20,3
$[2,34\,;\,2,61)$	2,5	23,5	22,9	10,1	19,9
$[2,61\,;\,2,85)$	2,7	31,0	38,0	54,1	20,7
$[2,85\,;\,3,16)$	3,0	26,2	30,3	32,3	19,1
$[3,16\,;\,4,14)$	3,7	10,1	5,1	2,8	19,9

PM: ponto médio da classe, S256: soja em 256 amostras, S128: soja em 128 amostras, S64: soja em 64 amostras, IC: índice de cone, CP1: primeira componente principal, S256d: dados originais de soja classificados.

o rendimento de madeira. Com o modelo escolhido foi possível construir mapas temáticos procurando identificar zonas diferenciadas. Também foi feita uma análise descritiva das predições obtidas nos mapas. A densidade amostral foi de 0,008 pontos por ha^{-1}, valor esse considerado inexpressivo.

Análise descritivas das amostras de argila e IMA

As amostras foram analisadas em 2 enfoques. O primeiro constituído de 18 pontos amostrais com medidas de IMA e Teor de Argila. O segundo pelo conjunto de 555 parcelas onde foi medido somente o Teor de Argila. O layout está ilustrado na Figura 2.19.

Pela Tabela 2.10 o valor médio do IMA, avaliado em 18 árvores, foi de 26,2 m^3 ha^{-1} ano^{-1}, com CV de 24%, o que, segundo Gomes (1963), compromete a precisão da medida de tal variável. Mainardi, Schneider e Finger (1996) obtiveram um valor médio de IMA de 28,24 m^3 ha^{-1} ano^{-1} em árvores aos 12 anos e 32,94 m^3 ha^{-1} ano^{-1} em árvores aos 15 anos tomados em 7 diferentes sítios na região de Cambará do Sul no Estado do Rio Grande do Sul. Carvalho et al. (1999) apresentaram em estudo dos efeitos das características do solo na capacidade produtiva um resultado do IMA de 25,3 m^3 ha^{-1} ano^{-1} em 16 amostras de árvores com 15 anos. Na mesma tabela encontra-se o valor médio obtido de 32,8% do teor de argila em 18 amostras e 25,2% para 555 amostras, indicando em um solo de textura média, adequado ao crescimento radicular (ROSOLEN et al., 1999).

A Figura 2.20 ilustra o perfil da função log da verossimilhança para o parâmetro λ de

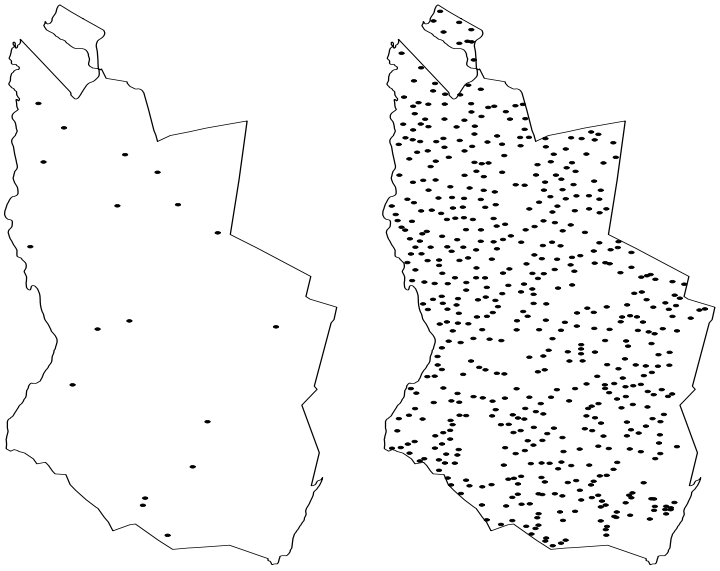

Figura 2.19: Localização das amostras na área de reflorestamento da fazenda MOBASA em Rio Pedrinho-SC. Os 18 pontos amostrais na figura à esquerda representam as coordenadas de dados de análises físicas e químicas e os 555 pontos na figura à direita representam as análises físicas.

Tabela 2.10: Estatísticas descritivas das amostras de IMA em 18 localizações e de Teor de Argila em 18 e 555 localizações.

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.%
IMA18	14,09	26,2	37,02	6,546	24,96
Argila18	9,00	32,8	64,20	12,542	38,19
Argila555	8,00	25,2	58,00	7,776	30,87

IMA: m^3 ha^{-1} ano^{-1}; Teor de Argila: porcentagem; D.P.: Desvio Padrão, C.V.: Coeficiente de Variação.

transformação de Box-Cox da variável IMA em 18 amostras. Foi obtido o intervalo de 95% de confiança (0,9; 2,9) para esse parâmetro, o qual incluindo a unidade, não sugere a necessidade de transformação, indicando válida a hipótese de gaussianiedade das medidas.

Ajuste dos parâmetros do modelo univariado para IMA

A Tabela 2.11 mostra os resultados da estimação pontual e médias dos parâmetros do modelo gaussiano, obtidos por MV e inferência bayesiana. Na obtenção da verossimilhança o Teor de Argila foi considerado covariável para o efeito sistemático. No método bayesiano, foram adotadas as *priori* plana ("flat") para β, recíproca para σ^2 e uniforme discreta para ϕ e o efeito de pequena escala e/ou erro aleatório foi considerado nulo. Na tabela as distribuições a

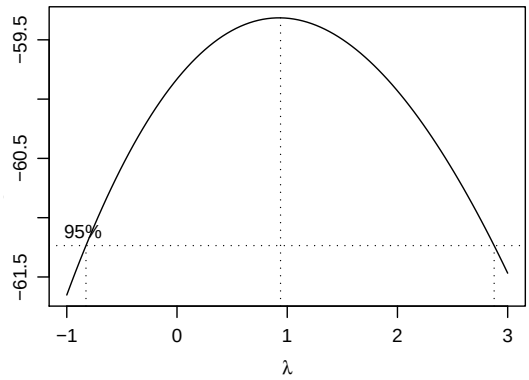

Figura 2.20: Perfil do log-verossimilhança para o parâmetro λ de transformação de Box-Cox da variável IMA.

posteriori dos parâmetros σ^2 e β são, respectivamente χ^2_{ScI} e Normal e seus gráficos, obtidos a partir de simulações estão apresentadas na Figura 2.21.

Tabela 2.11: Parâmetros do modelo geoestatístico estimado por MV e inferência bayesiana. O modelo de função de correlação usado foi o de Matèrn com parâmetro de diferenciabilidade $\kappa = 0,5$.

IMA18	β_0	β_1	τ^2	σ^2	ϕ
MV	35,9	-0,29	0,0	26,7	191,5
BAYES	25,5	–	0,0	242,9	5.732,3

β_0 e β_1 são parâmetros do efeito sistemático do modelo, σ^2 e ϕ são parâmetros da função de correlação, τ^2 é o parâmetro do erro.

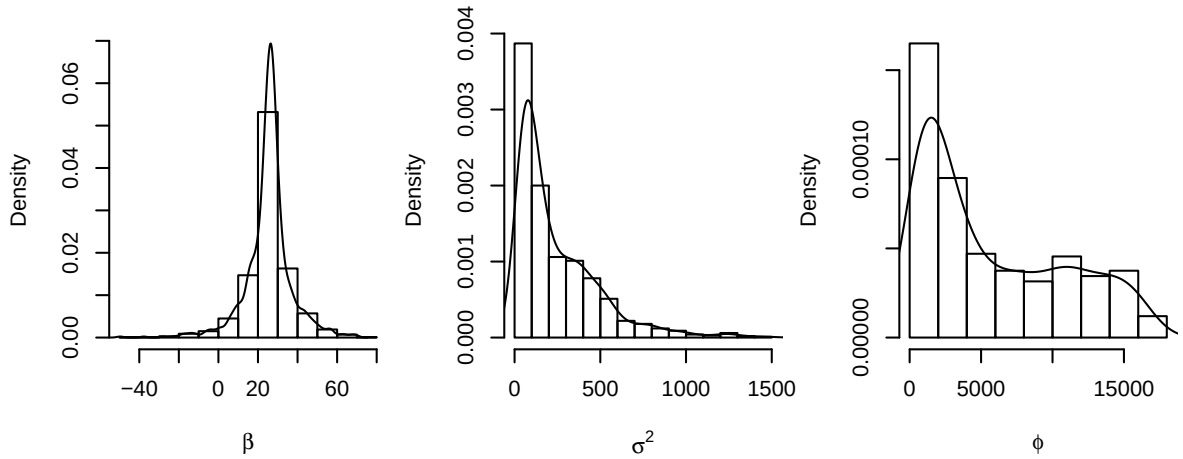

Figura 2.21: Distribuição a *posteriori* para os parâmetros β, σ^2 e ϕ a partir de 1.000 aproximações numéricas da variável IMA tomada em 18 pontos amostrais

Estatísticas descritivas da predição univariada de IMA

A Tabela 2.12 mostra estatísticas descritivas da predição de IMA com base em uma amostra de 18 pontos tomados casualmente na área de reflorestamento, empregando-se krigagem convencional com parâmetros do modelo estimados por MV e predição bayesiana. Nota-se nessa tabela que os resultados são muito próximos, exceto por uma variância um pouco maior no caso bayesiano refletindo nessa maior variabilidade o fato da incerteza sobre os parâmetros ser considerada nessa abordagem. Entretanto, não se pode estabelecer uma diferença de predição de um método em relação ao outro.

Tabela 2.12: Estatísticas descritivas das predições do IMA através da krigagem convencional e por estimativa bayesiana.

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.%
Clássico	17,48	26,53	36,09	0,747	2,8
Bayesiano	14,69	26,02	36,72	5,100	19,6

IMA: $m^3\ ha^{-1}\ ano^{-1}$; D.P.: Desvio Padrão, C.V.: Coeficiente de Variação.

Na Tabela 2.13 se apresenta a porcentagem dos pontos de IMA estimados por método bayesiano, incidentes em cada intervalo de classificação. Nela verifica-se a tendência do número de pontos preditos se concentrarem nas classes centrais, tanto na predição através da krigagem quanto na predição que envolveu o método bayesiano. Todavia o efeito é menos acentuado no método bayesiano.

Tabela 2.13: Porcentagem dos pontos estimados de IMA por método bayesiano com modelo univariado, incidentes em cada intervalo de classificação.

Classe	PM	MV18	Bayes18
$[14,09\ ;\ 20,50)$	17,3	0,8	4,6
$[20,50\ ;\ 23,43)$	22,0	3,6	13,0
$[23,43\ ;\ 28,32)$	25,9	84,6	54,4
$[28,32\ ;\ 32,50)$	30,4	10,7	19,5
$[32,50\ ;\ 37,02)$	34,8	0,4	3,6

PM: ponto médio da classe, MV18: IMA em 18 amostras estimado por MV, Bayes18: IMA em 18 amostras estimado por método bayesiano.

Mapa de predição univariada de IMA

Para a produção dos mapas temáticos os resultados foram classificados segundo os quantis 20, 40, 60 e 80% de predição de IMA. A Figura 2.22 a esquerda ilustra a distribuição espacial de IMA estimada por krigagem convencional. Nela, segundo a Tabela 2.13 0,8% da área representa IMA abaixo de 20,50 m^3 ha^{-1} ano^{-1}, 3,6% representa IMA entre 20,50 e 23,42 m^3 ha^{-1} ano^{-1}, 84,6% representa IMA entre 23,42 e 28,33 m^3 ha^{-1} ano^{-1}, 10,7% entre 28,33 e 32,50 m^3 ha^{-1} ano^{-1} e 0,4% acima de 32,50 m^3 ha^{-1} ano^{-1}. Já a figura da direita ilustra a distribuição espacial estimada por métodos bayesianos. Nela, 4,6% da área representa IMA abaixo de 20,50 m^3 ha^{-1} ano^{-1}, 13% representa IMA entre 20,50 e 23,42 m^3 ha^{-1} ano^{-1}, 54,4% representa IMA entre 23,42 e 28,33 m^3 ha^{-1} ano^{-1}, 19,5% entre 28,33 e 32,50 m^3 ha^{-1} ano^{-1} e 3,6% acima de 32,50 m^3 ha^{-1} ano^{-1}. Isso indica, neste caso, que o método bayesiano concentrou menos predições que a krigagem baseada em MV, mostrando áreas maiores de rendimento diferenciado.

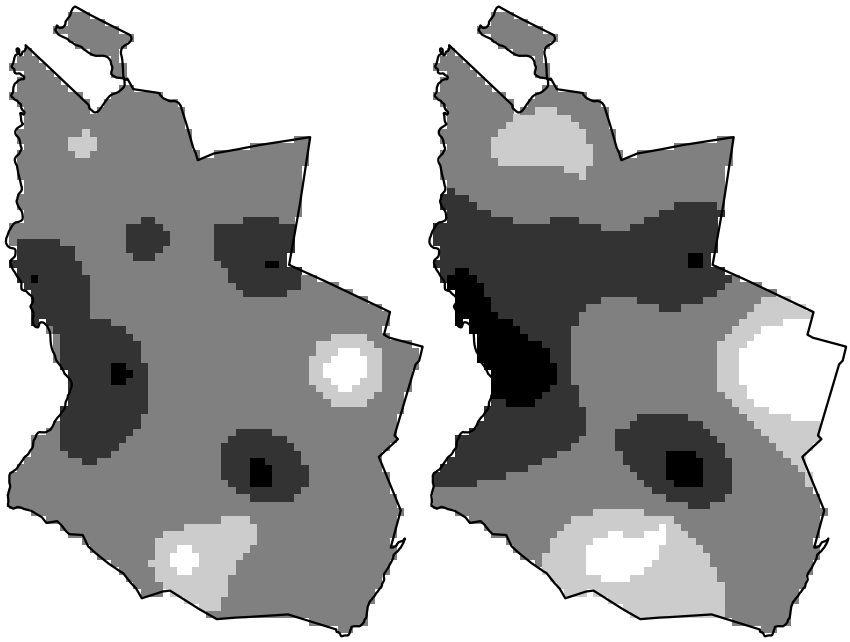

Figura 2.22: Mapa de predição de IMA com 18 amostras, classificada pelos quartis. A figura da esquerda foi obtida por krigagem convencional e a da direita por predição bayesiana.

2.8.5 Conclusões sobre o método univariado

- O método da máxima verossimilhança foi capaz de estimar a produtividade de soja com erro relativo abaixo de 2,1% e estimar a média de IMA com erro relativo abaixo de 1,2%, considerando-se os valores conhecidos de 18 pontos amostrais;
- O método bayesiano foi capaz de estimar a produtividade de soja com erro relativo abaixo de 2,1%, e estimar a média de IMA com erro relativo abaixo de 1,0%;
- O mapa da distribuição espacial da média de produtividade de soja, quando feito pelo método MV, perde identificabilidade de zonas de manejo com a diminuição do número de amostras;
- O método bayesiano, aplicado tanto no estudo de produtividade de soja quanto no de IMA foi capaz de identificar zonas de manejo para amostras de tamanho onde o método da MV não o foi, além de ser mais sensível na identificação das variações a pequenas distâncias.
- A krigagem baseada em MV e a predição bayesiana induziram a uma concentração de valores preditos em torno da média, inversamente proporcional ao tamanho da amostra tomada considerada nos métodos.

3 MODELO GEOESTATÍSTICO GAUSSIANO MULTIVARIADO

3.1 INTRODUÇÃO

Para Ver Hoef e Cressie (1993), em ciências da terra é freqüente o interesse em predizer conjuntamente uma grande quantidade de variáveis. Normalmente se prediz uma variável por vez, usando dados de um mesmo tipo (krigagem) ou utilizando informações adicionais de outra variável tomada nas mesmas coordenadas (krigagem com covariável). O modelo bivariado mostrou que a predição de uma variável com base em uma outra variável correlacionada, mas em locais diferentes (cokrigagem) resultou em predições precisas. Predições espaciais multivariadas permitem construir regiões de predição multivariada. Esses autores relacionam e comparam predições baseadas no variograma cruzado, predições espaciais multivariadas e estimação de parâmetros por mínimos quadrados generalizados.

Os modelos geoestatísticos multivariados dizem respeito a um conjunto de variáveis aleatórias gaussianas dadas por:

$$\left\{Y_1(x), Y_2(x), \ldots, Y_p(x) : Y_k(x) \in S_k(x); x_i \in \mathbb{R}^2; i = 1, 2, \ldots, n\right\} \tag{3.1}$$

Essas varíaveis são georreferenciadas em uma mesma região, todas com igual interesse científico. É uma situação pouco realística pois esta descrição não leva a uma interpretação física no sentido prático, entretanto o será se puder ser escrita a distribuição condicional de uma das variáveis, eleita de interesse primário, condicionada a uma ou mais variáveis espacialmente localizadas. Neste caso, exige-se que todas as variáveis sejam tomadas nas mesmas posições geográficas e que haja uma certa correlação entre elas. Outra situação prática ocorre quando a variável primária for de difícil aquisição, então, pode-se formar um conjunto das variáveis restantes, supostamente de fácil observação, como o conjunto de variáveis preditoras que, modeladas adequadamente, permitirão fazer estimativas da variável primária em locais onde foram obtidas as demais variáveis. Neste caso, as variáveis podem ser em quantidades, tipos e localizações diferentes. Pretendeu-se aqui abordar ambos os casos e ainda utilizar o

suporte da ACP para a redução do número de variáveis envolvidas no problema. Apresentou-se o problema geoestatístico multivariado envolvendo duas variáveis, sendo uma a principal e a outra, secundária.

3.2 MODELO GEOESTATÍSTICO MULTIVARIADO

O conjunto dado pela expressão 3.1 é uma coleção p-dimensional de variáveis aleatórias. A matriz Σ de covariâncias desse conjunto é dada por:

$$\Sigma = \begin{pmatrix} Cov(Y_1;Y_1) & Cov(Y_1;Y_2) & \dots & Cov(Y_1;Y_p) \\ Cov(Y_2;Y_1) & Cov(Y_2;Y_2) & \dots & Cov(Y_2;Y_p) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(Y_p;Y_1) & Cov(Y_p;Y_2) & \dots & Cov(Y_p;Y_p) \end{pmatrix}$$

sendo a diagonal a autocorrelação de cada variável $Y_k : k = 1,2,\dots,p$ do conjunto. Os elementos fora da diagonal representam a matriz correlação cruzada para cada combinação de pares de variáveis. Essa matriz é uma extensão daquela matriz para o caso univariado dado pela Equação 2.27. Ela deve ser uma matriz quadrada, simétrica, definida positiva e passível de decomposição, para se obter sua inversa.

Para qualquer par de variáveis, por exemplo $\big(Y_c(a);Y_d(b)\big)$, $Cov\big(Y_c(a);Y_d(b)\big) = \sigma_{ab}^{cd}$. Neste caso, afirma-se que a covariância (e a correlação) se estabelece entre a variável $Y_c(x)$ tomada na coordenada a e a variável $Y_d(x)$ tomada na coordenada b. Uma propriedade imediata, é a sua natureza simétrica, ou seja, $\sigma_{ab}^{cd} = \sigma_{ba}^{dc}$.

A matriz de correlação será dada por $R(u)$, cujos elementos serão:

$$\rho_{ab}^{cd} = \frac{\sigma_{ab}^{cd}}{\sqrt{\sigma_a^2 \sigma_b^2}} = \frac{\sigma_{ab}^{cd}}{\sigma_a \sigma_b}$$

Quando $a = b$, a função $\rho^{aa}(u) = \rho^{bb}(u)$ corresponderá à função de correlação do processo univariado $Y_a(x)$ e $\rho^{aa}(-u) = \rho^{aa}(u)$. Se $a \neq b$ a função $\rho^{ab}(u)$ será chamada função de correlação cruzada de $Y_a(x)$ e $Y_b(x)$, mas não será necessariamente simétrica na matriz $R(u)$, mas ainda assim satisfará a condição de que $\rho^{ab}(u) = \rho^{ba}(-u)$ (DIGGLE; RIBEIRO JR, 2007).

Pebesma e Wesseling (1998) apresentam um modelo de predição multivariada envolvendo variáveis cruzadas correlacionadas. O modelo utilizado para cada variável $Y_k; k = 1,2,\dots,p$ é aquele definido pela Equação 2.2, portanto, um processo não estacionário. O mo-

delo multivariado, neste caso de envolvimento de todas as variáveis do conjunto, é dado por:

$$Y = D\,\beta + S(x) + \varepsilon$$

onde $D\,\beta$ corresponde à matriz de tendência externa do modelo aplicada às respectivas variáveis. As matrizes envolvidas são:

$$D = \begin{pmatrix} D_1 & 0 & \dots & 0 \\ 0 & D_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D_p \end{pmatrix} \text{ onde } D_k = \begin{pmatrix} 1 & d_{11}^{(k)} & d_{21}^{(k)} & \dots & d_{p1}^{(k)} \\ 1 & d_{12}^{(k)} & d_{22}^{(k)} & \dots & d_{p2}^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & d_{1n_k}^{(k)} & d_{2n_k}^{(k)} & \dots & d_{pn_k}^{(k)} \end{pmatrix},$$

para n_k representando o número de coordenadas relativas à k-ésima variável externa d e p é o número de variáveis externas associada a um particular processo Y_k. $S(x) = \{S_1(x), S_2(x), \dots, S_p(x)\}$.

O melhor preditor linear não viciado será:

$$\hat{Y}(x_0) = d(x_0)\,\hat{\beta} + r'V^{-1}Y(x) - D\,\hat{\beta}$$

no qual

$$d(x_0) = \begin{pmatrix} d_1\big(x(0)\big) & 0 & \dots & 0 \\ 0 & d_2\big(x(0)\big) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_p\big(x(0)\big) \end{pmatrix}$$

e $d_k(x_0)$ correspondendo à linha da matriz D que contém o valor correspondente de $Y(x_0)$, r a matriz de correlações de cada variável com o ponto x_0 a ser estimado e V a matriz de correlações obtidas à partir da matriz de covariâncias multivariadas dada pela equação 3.2.

A variância do erro de predição será dada por:

$$Var\big(\hat{Y}(x_0)\big) = \hat{\beta} - r'\,V^{-1}\,r + \big(d(x_0) - r'V^{-1}D\big)\big(D^{-1}V^{-1}D\big)^{-1}\big(d(x_0) - r'V^{-1}r\big)^{-1}$$

Para Ver Hoef e Cressie (1993) este modelo não impõe restrições ao número de variáveis e cada variável pode ter um número diferente de localizações.

O pantanal matogrossense, segundo Couto e Cunha (2002), apresenta muitas unidades de pedopaisagens com áreas periodicamente inundáveis, onde a amostragem é difícil devido a elevada variabilidade espacial inter e intra estratos. Para sua pesquisa coletaram e analisaram cento e onze amostras sistemáticas com cinco atributos físicos e quinze atributos químicos em

três ecossistemas. Efetuaram uma análise de componentes principais e fatorial. Das amostras restaram quatro componentes que explicaram 77% da variância total e dois fatores que mostraram a melhor separação entre as pedopaisagens. Utilizaram, nas estimativas dos semivariogramas para os componentes principais os softwares GS+ produzido e comercializado pela empresa *Gamma Design Software* (www.gammadesign.com) e o software Surfer produzido pela empresa *Golden Software, Inc.* (www.goldensoftware.com). Relatam os aspectos da análise multivariada como apoio às aplicações geoestatísticas mas não explicitam o emprego da geoestatística multivariada.

Filzmoser e Reimann (2002) discutem e comparam métodos e propriedades da análise de componentes principais e da análise fatorial. Eles expõem as vantagens em se aplicar métodos multivariados robustos em geoestatística. Ilustram porém, com aplicações a um conjunto de dados geoquímicos, aplicações da geoestatística univariada.

3.3 MODELO GEOESTATÍSTICO BIVARIADO

Considerou-se o seguinte processo gaussiano estacionário bivariado:

$$\left\{S(x) = \left(S_1(x_i), S_2(x_j)\right) : S_1(x), S_2(x) \in \mathbb{R}; x_i, x_j \in \mathbb{R}^2; i = 1,2,\ldots,r; j = 1,2,\ldots,s\right\},$$

com $E(S_1(x_i)) = 0$; $E(S_2(x_i)) = 0$ e $Var(S_1(x_i)) = \sigma_1^2$ e $Var(S_2(x_i)) = \sigma_2^2$.

A matriz de covariância associada é dada por:

$$\Sigma = \left(\begin{array}{c|c} Cov(S_1;S_1) & Cov(S_1;S_2) \\ \hline Cov(S_2;S_1) & Cov(S_2;S_2) \end{array}\right) = \begin{pmatrix} \sigma_{i,j}^{(1,1)} & \sigma_{i,j}^{(1,2)} \\ \sigma_{i,j}^{(2,1)} & \sigma_{i,j}^{(2,2)} \end{pmatrix}$$

sendo,

- $\sigma_{(i,j)}^{1,1}$ $i = 1,\ldots,r$ e $j = 1,\ldots,r$;
- $\sigma_{(i,j)}^{1,2}$ $i = 1,\ldots,r$ e $j = 1,\ldots,s$;
- $\sigma_{(i,j)}^{2,1}$ $i = 1,\ldots,s$ e $j = 1,\ldots,r$;
- $\sigma_{(i,j)}^{2,2}$ $i = 1,\ldots,s$ e $j = 1,\ldots,s$;

Expandindo essa matriz vem:

$$\Sigma = \left(\begin{array}{cccc|cccc} \sigma_{1,1}^{(1,1)} & \sigma_{1,2}^{(1,1)} & \dots & \sigma_{1,r}^{(1,1)} & \sigma_{1,1}^{(1,2)} & \sigma_{1,2}^{(1,2)} & \dots & \sigma_{1,s}^{(1,2)} \\ \sigma_{2,1}^{(1,1)} & \sigma_{2,2}^{(1,1)} & \dots & \sigma_{2,r}^{(1,1)} & \sigma_{2,1}^{(1,2)} & \sigma_{2,2}^{(1,2)} & \dots & \sigma_{2,s}^{(1,2)} \\ \dots & \dots & & \dots & \dots & \dots & & \dots \\ \sigma_{r,1}^{(1,1)} & \sigma_{r,2}^{(1,1)} & \dots & \sigma_{r,r}^{(1,1)} & \sigma_{r,1}^{(1,2)} & \sigma_{r,2}^{(1,2)} & \dots & \sigma_{r,s}^{(1,2)} \\ \hline \sigma_{1,1}^{(2,1)} & \sigma_{1,2}^{(2,1)} & \dots & \sigma_{1,r}^{(2,1)} & \sigma_{1,1}^{(2,2)} & \sigma_{1,2}^{(2,2)} & \dots & \sigma_{1,s}^{(2,2)} \\ \sigma_{2,1}^{(2,1)} & \sigma_{2,2}^{(2,1)} & \dots & \sigma_{2,r}^{(2,1)} & \sigma_{2,1}^{(2,2)} & \sigma_{2,2}^{(2,2)} & \dots & \sigma_{2,s}^{(2,2)} \\ \dots & \dots & & \dots & \dots & \dots & & \dots \\ \sigma_{s,1}^{(2,1)} & \sigma_{s,2}^{(2,1)} & \dots & \sigma_{s,r}^{(2,1)} & \sigma_{s,1}^{(2,2)} & \sigma_{s,2}^{(2,2)} & \dots & \sigma_{s,s}^{(2,2)} \end{array}\right)$$

Nesta matriz, o bloco superior esquerdo representa as autocovariâncias da variável Y_1 e o bloco inferior direito as autocovariâncias das variável Y_2. Os blocos superior direito e inferior esquerdo representam as covariâncias cruzadas entre as variáveis Y_1 e Y_2. Os índices no expoente dos elementos da matriz representam as variáveis envolvidas e os índices abaixo correspondem às localizações. Assim, o elemento $\sigma_{i,j}^{(2,1)}$ representa a covariância entre a variável Y_2 medida na localização x_i e Y_1 medida na localização x_j. De uma forma geral, essa matriz de covariâncias não estabelece que as coordenadas devam ser totalmente ou parciamente coincidentes. Considerou-se aqui a notação x_i para a i-ésima coordenada da variável Y_1 e x'_j a j-ésima coordenada da variável Y_2.

Modelos bivariados podem ser escritos como uma junção de modelos univariados como:

$$\begin{cases} Y_{1,i} = \mu_1(x_i) + S_1(x_i) + \varepsilon_i, & i = 1,\dots,r. \\ Y_{2,j} = \mu_2(x_j) + S_2(x_j) + \varepsilon_j, & j = 1,\dots,s. \end{cases}$$

onde:

a) $\mu_1(x_i) = D_1\beta^{(1)}$ e $\mu_2(x_j) = D_2\beta^{(2)}$ são componentes determinísticos do modelo associados a p_1 covariáveis em D_1 e a p_2 covariáveis em D_2;

b) $S_1(x_i)$ e $S_2(x_j)$ são variáveis aleatórias espacialmente correlacionadas onde $S_1(x_i) \sim N_n\big(0;\sigma_1^2 R(\phi_1)\big)$ e $S_2(x_j) \sim N_n\big(0;\sigma_2^2 R(\phi_2)\big)$ com $\sigma_1^2 R(\phi_1)$ e $\sigma_2^2 R(\phi_2)$ representando as autocorrelações de $S_1(x_i)$ e $S_2(x_j)$, respectivamente;

c) $\varepsilon_i \sim N\big(0;\tau_1^2\big)$ e $\varepsilon_j \sim N\big(0;\tau_2^2\big)$ são os erros aleatórios independentes.

As variáveis Y_1 e Y_2 não precisarão ser co-localizadas e nem medidas o mesmo número de vezes, ou seja, podem ou não serem coincidentes na área (Figura 3.1).

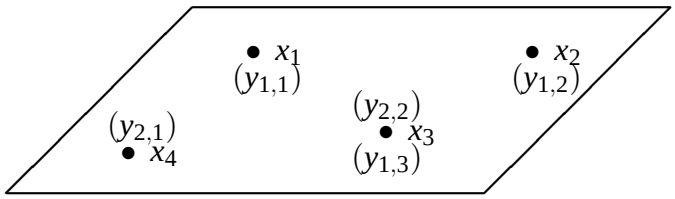

Figura 3.1: Representação de uma área típica com processos geoestatísticos bivariados contendo quatro localizações amostrais, onde as variáveis não são co-localizadas e nem oferecem o mesmo número de observações.

Deve-se aqui considerar quatro possibilidades distintas para modelos assim especificados, considerando as características de seus elementos, assumindo que Y_1 e Y_2 ocorrem simultaneamente em uma mesma área de um espaço bidimensional:

a) Sendo $\tau_1 = \tau_2 = 0$ e $S_1(x_i) \sim \mathsf{N}_n\big(0; \sigma_1^2 R(\phi_1)\big)$ independente de $S_2(x_j) \sim \mathsf{N}_n\big(0; \sigma_2^2 R(\phi_2)\big)$, então Y_1 será independente de Y_2, ou seja, não serão correlacionados. Um problema escrito desta maneira, exigirá a estimação de quatro parâmetros: σ_1, σ_2, ϕ_1 e ϕ_2.

b) Sendo $\tau_1 = \tau_2 = 0$ e $S_1(x_i) \sim \mathsf{N}_n\big(0; \sigma^2 R(\phi)\big)$ idêntico a $S_2(x_j) \sim \mathsf{N}_n\big(0; \sigma^2 R(\phi)\big)$, então Y_1 será perfeitamente correlacionado com Y_2. Um problema escrito desta maneira, exigirá a estimação de dois parâmetros: σ e ϕ.

c) Sendo $\tau_1 \neq \tau_2$ e $S_1(x_i) \sim \mathsf{N}_n\big(0; \sigma^2 R(\phi)\big)$ idêntico a $S_2(x_j) \sim \mathsf{N}_n\big(0; \sigma^2 R(\phi)\big)$ então Y_1 será parcialmente correlacionado com Y_2, provocando uma dispersão difusa, dependendo da variância σ^2. Um problema modelado desta maneira, exigirá a estimação de quatro parâmetros: τ_1, τ_2, σ, ϕ.

d) Sendo $\tau_1 = \tau_2 = 0$, $S_0(x_i) \sim \mathsf{N}_n\big(0; \sigma_{0;1}^2 R(\phi_0)\big)$ idêntico a $S_0(x_j) \sim \mathsf{N}_n\big(0; \sigma_{0;2}^2 R(\phi_0)\big)$, mas escalonado por σ^2 e $S_1(x_i) \sim \mathsf{N}_n\big(0; \sigma_1^2 R(\phi_1)\big)$ diferente de $S_2(x_j) \sim \mathsf{N}_n\big(0; \sigma_2^2 R(\phi_2)\big)$ então Y_1 será parcialmente correlacionado com Y_2. Um problema concebido dessa maneira exigirá a estimação de sete parâmetros: $\sigma_{0;1}^2$, $\sigma_{0;2}^2$, σ_1^2, σ_2^2, ϕ_0, ϕ_1 e ϕ_2. Incluindo-se $\tau_1 \neq \tau_2$ o número de parâmetros a serem estimados poderá chegar a nove.

Como exemplo de uma situação mais realística, ilustrando a situação contemplada em (d), considerando-se estacionariedade na média e $\varepsilon = 0$ (sem perda de generalidade) o modelo bivariado pode ser escrito como:

$$\begin{cases} Y_{1,j} = \mu_1 + S_0\big(x_i\big) + S_1\big(x_i\big) \\ Y_{2,j} = \mu_2 + S_0\big(x_j\big) + S_2\big(x_j\big) \qquad i = 1,2,\ldots,m; \quad j = 1,2,\ldots,n \end{cases} \tag{3.2}$$

sendo $S_0(x_i)$, $S_1(x_i)$, $S_0(x_j)$ e $S_2(x_j)$ processos univariados independentes e com a mesma funcional de correlação espacial.

A covariância entre duas variáveis aleatórias Y_1 e Y_2 é definida como:

$$Cov\big(Y_1(x);Y_2(x)\big) = E\left(\big(Y_1(x)-\mu_{Y_1}\big)\big(Y_2(x)-\mu_{Y_2}\big)\right)$$

em que $\mu_{Y_1} = E\big(Y_1(x)\big)$ e $\mu_{Y_2} = E\big(Y_2(x)\big)$ e define o coeficiente de correlação entre elas como sendo:

$$\rho_{Y_1;Y_2} = \frac{Cov\big(Y_1(x);Y_2(x)\big)}{\sigma_{Y_1}\sigma_{Y_2}}$$

em que $\sigma^2_{Y_1} = Var\big(Y_1(x)\big)$ e $\sigma^2_{Y_2} = Var\big(Y_2(x)\big)$.

Goovaerts (1997), define uma função que estima a correlação entre duas variáveis Y_1 e Y_2 (nesta ordem), separadas por uma mesma distância $h_k; k = 1,2,\ldots,s$ (s a quantidade de pares que correspondem essa distância) como:

$$C_{1;2}(h) = \frac{1}{N(h)}\sum_{k=1}^{N(h)} y_1(x_k)y_2(x'_k) - \hat{\mu}_1\hat{\mu}_2$$

onde $\hat{\mu}_1 = \dfrac{1}{N(h)}\displaystyle\sum_{k=1}^{N(h)} y_1(x_k)$, $\hat{\mu}_2 = \dfrac{1}{N(h)}\displaystyle\sum_{k=1}^{N(h)} y_2(x'_k)$ e $N(h)$ é o número de pares pertencentes à mesma classe de distâncias e direção. Os estimadores $\hat{\mu}_1$ e $\hat{\mu}_2$ são, respectivamente, os estimadores das médias μ_1 de Y_1 e μ_2 de Y_2 nas suas respectivas coordenadas do conjunto formado pelas distâncias h. Um exemplo é apresentado na Figura 3.2 para a distância h_1 (fixa) em que $\hat{\mu}_1$ seria a média das observações $y(x_i)$ (círculos) do conjunto dessas distâncias e $\hat{\mu}_2$ a média das observações $y(x'_i)$ (estrelas) do mesmo conjunto.

A covariância obtida para essas diferentes distâncias é chamada de função covariância cruzada experimental. De maneira geral $C_{(1;2)}(h) \neq C_{(1;2)}(-h)$.

A estimativa do correlograma cruzado será dada por:

$$\rho_{1;2}(h) = \frac{C_{(1;2)}(h)}{\sqrt{\sigma_1^2\sigma_2^2}}$$

em que, $\sigma_1^2 = \dfrac{1}{N(h)}\displaystyle\sum_{k=1}^{N(h)}(y_1(x_k)-\hat{\mu}_1)^2$ e $\sigma_2^2 = \dfrac{1}{N(h)}\displaystyle\sum_{k=1}^{N(h)}(y_2(x'_k)-\hat{\mu}_2)^2$ sendo que σ_1^2 e σ_2^2 são as variâncias de Y_1 e Y_2 nas suas respectivas coordenadas do conjunto formado pelas distâncias h.

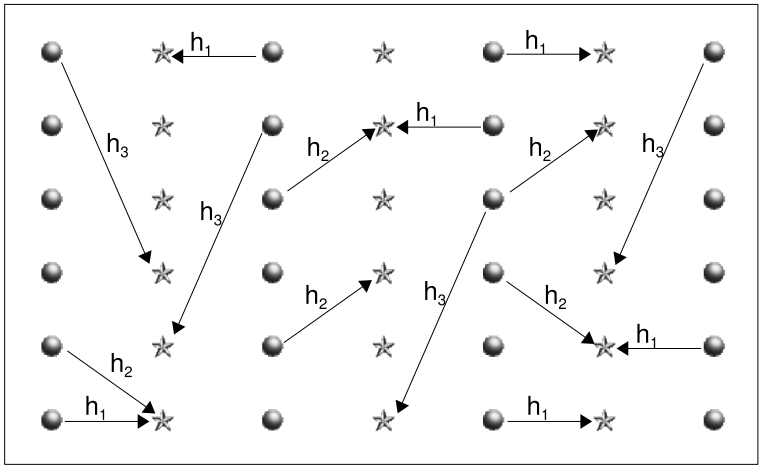

Figura 3.2: Grid regular com locação amostral de duas variáveis com círculos representando a primeira e estrelas a segunda. As setas estabelecem a direção das correlações e os *h*, através de seus índices indicam o grupo de correlações entre variáveis separadas por uma mesma distância.

Considerando-se a Equação 3.2 a covariância entre $Y(x_i)$ e $Y(x_j)$ pode ser expressa como:

$$\begin{aligned} Cov\big(Y_1(x_i);Y_2(x_j)\big) &= Cov\big(\mu_1+S_0(x_i)+S_1(x_i)\ ;\ \mu_2+S_0(x_j)+S_2(x_j)\big) \\ &\overset{\text{ind.}}{=} Cov\big(S_0(x_i);S_0(x_j)\big) \end{aligned} \tag{3.3}$$

De forma semelhante ao resultado obtido na Equação 2.8, a Equação (3.3) fica:

$$Cov\big(Y_1(x_i);Y_2(x_j)\big) = \sigma_{01}\sigma_{02}\rho(\phi_0) \tag{3.4}$$

Análogamente as autocovariâncias de Y_1 e de Y_2 serão dadas por:

$$Cov\big(Y_1(x);Y_1(x)\big) = \sigma_0^2\rho(\phi_0)+\sigma_1^2\rho(\phi_1) \tag{3.5}$$

$$Cov\big(Y_2(x);Y_2(x)\big) = \sigma_0^2\rho(\phi_0)+\sigma_2^2\rho(\phi_2) \tag{3.6}$$

Fazendo-se as parametrizações $\sigma_{01}=\sigma;\ \ \sigma_{02}=\eta\sigma;\ \ \sigma_1=\nu_1\sigma$ e $\sigma_2=\nu_2\sigma$ e substituindo nas Equações 3.4, 3.5 e 3.6 tem-se:

$$\begin{aligned} Cov\big(Y_1(x);Y_2(x')\big) &= \sigma\eta\sigma\rho(\phi_0)=\sigma^2\eta\rho(\phi_0) \\ Var\big(Y_1(x)\big) &= \sigma^2+\nu_1^2\sigma^2=\sigma^2[\rho(\phi_0)+\nu_1^2\rho(\phi_1)] \\ Var\big(Y_2(x)\big) &= \eta^2\sigma^2+\nu_2^2\sigma^2=\sigma^2[\eta^2\rho(\phi_0)+\nu_2^2\rho(\phi_2)] \end{aligned}$$

Logo:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathsf{N}_n \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \sigma^2 \begin{bmatrix} V(\nu_1;\phi_0;\phi_1) & V(\eta;\phi_0) \\ V'(\eta;\phi_0) & V(\eta;\nu_2;\phi_0;\phi_2) \end{bmatrix} \right) \tag{3.7}$$

A literatura geralmente sugere que a estimativa dos parâmetros de um modelo geoestatístico bivariado seja feita através do ajuste de certas funções de correlação, como as já mencionadas no Capítulo 2, aos semivariogramas relativos a Y_1 e Y_2 bem como ao semivariograma cruzado. Neste trabalho empregamos método baseado em verossimilhança aplicados á distribuição conjunta de Y_1 e Y_2.

3.4 PREDIÇÃO LINEAR ESPACIAL BIVARIADA

Isaaks e Srivastava (1989) apresentam a cokrigagem como um método de estimação, envolvendo a correlação cruzada entre variáveis secundárias e uma variável primária. A grande utilidade do método, alegada pelos autores, é que as variáveis secundárias podem apresentar características favoráveis à sua obtenção, como baixo custo, fácil acesso, dentre outras, que podem ser utilizadas para estimar variáveis primárias sujeitas a subamostragem.

Consideram-se aqui dois processos estocásticos Y_1 e Y_2 distintos, mas ocorrendo simultaneamente em uma região. Por conveniência de notação, Y_1 foi a variável primária. No caso de uma única variável, para alguma coordenada onde não se tenha um valor medido, este poderá ser estimado por krigagem usando uma combinação linear com pesos w associados a valores conhecidos, tal como: $\hat{y}_0 = \sum_{i=1}^{n} w_i y_i$ onde y_i é o valor medido na i-ésima coordenada x. No caso de duas variáveis, a estimativa por cokrigagem, com um modelo linear de corregionalização, será obtida por uma combinação linear das duas variáveis, como:

$$\hat{y}_1(x_0) = \sum_{i=1}^{n} a_i y_1(x_i) + \sum_{j=1}^{m} b_j y_2(x_j) \tag{3.8}$$

onde $\hat{y}_1(x_0)$ é a estimativa da variável primária em uma particular localização x_0 não amostrada; $y_1(x) = \big(y_1(x_1), y_1(x_2), \ldots y_1(x_n)\big)$ são os dados da variável primária observados em n localizações da área, $y_2(x_i) = (y_2(x_1), y_2(x_2), \ldots y_2(x_m))$ são os dados da variável secundária observados em m localizações da mesma área, que podem ser parcialmente ou totalmente coincidentes ou isoladas com relação às localizações da variável primária; $a_1, a_2, \ldots, a_n$ e $b_1, b_2, \ldots, b_m$ são, respectivamente, os pesos de krigagem a serem determinados associados às

observações $y_1(x_i)$ e $y_2(x_j)$, $i = 1,\ldots,n;\ j = 1,\ldots,m$.

Sendo $\big(\hat{y}_1(x_0) - y_1(x_0)\big)$ o erro de predição na coordenada x_0 então:

$$Var\big(\hat{y}_1(x_0) - y_1(x_0)\big) = w'C\,w \tag{3.9}$$

em que:

a) $w' = (a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m, -1)$,

b) $Y^* = \big(Y_1(x_1), \ldots, Y_1(x_n), Y_2(x_1), \ldots, Y_2(x_m), Y_1(x_0)\big)$,

c) C é a matriz de covariância de Y^*.

Desenvolvendo o lado direito da Equação 3.8 vem:

$$\begin{aligned}
Var(\hat{y}_1(x_0) - y_1(x_0)) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_1(x_i); Y_1(x_j)) + \\
&+ \sum_{i=1}^{m}\sum_{j=1}^{m} b_i b_j Cov(Y_2(x_i); Y_2(x_j)) + \\
&+ 2\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(Y_1(x_i); Y_2(x_j)) - \\
&- 2\sum_{i=1}^{n} a_i Cov(Y_1(x_i); Y_1(x_0)) - \\
&- 2\sum_{j=1}^{m} b_j Cov(Y_2(x_j); Y_1(x_0)) + \\
&+ Cov(Y_1(x_0); Y_1(x_0))
\end{aligned} \tag{3.10}$$

As condições a que os pesos de krigagem devem satisfazer são de que:

a) Devem levar a uma estimativa não viciada, o que ocorrerá se $\sum_{i=1}^{n} a_i = 1$ e $\sum_{j=1}^{m} b_j = 0$, o que pode ser comprovado aplicando-se a definição de estimador não-viciado dada por Mood, Graybill e Boes (1974). De fato:

$$E\big(\hat{y}_1(x_0)\big) = E\left(\sum_{i=1}^{n} a_i y_1(x_i) + \sum_{j=1}^{m} b_j y_2(x_j)\right) = \mu_1 \sum_{i=1}^{n} a_i + \mu_2 \sum_{j=1}^{m} b_j = \mu_1;$$

b) A variância do erro dado pela equação 3.10 deverá ser a menor possível, para escolhas convenientes dos pesos.

Minimizar a variância implica em igualar n derivadas parciais a zero, levando a um sistema de n equações e n incógnitas. Devido a condição de não tendenciosidade, que irá gerar mais uma equação sem parâmetros, o sistema será ampliado para $n+1$ equações e n incógnitas, cuja solução não é direta. Os multiplicadores de Lagrange são aplicados para converter esses problemas de minimização restrita em um problema irrestrito. Isaaks e Srivastava (1989) introduzem os multiplicadores de Lagrange ϑ_1 e ϑ_2 na Equação 3.10, o que resulta em:

$$Var(\hat{y}_1(x_0) - y_1(x_0)) = w'C\,w + 2\vartheta_1\left(\sum_{i=1}^{n} a_i - 1\right) + 2\vartheta_2\left(\sum_{j=1}^{n} b_i\right) \tag{3.11}$$

Sob a condição dada pelo item (a), a expressão 3.11 não muda. Ela poderá ser minimizada derivando-se a equação em relação a cada um dos pesos, inclusive os multiplicadores de Lagrange e igualando-se a zero, o que resulta em:

$$\begin{aligned}
\frac{\partial}{\partial a_k}\left(Var(\hat{y}_1(x_0) - y_1(x_0))\right) &= 2\sum_{i=1}^{n} a_i Cov(Y_1(x_i); Y_1(x_k)) \\
&+ 2\sum_{i=1}^{n} b_i Cov(Y_1(x_i); Y_2(x_k)) \\
&- 2Cov(Y_1(x_0); Y_1(x_k)) + 2\vartheta_1 = 0 \text{ para } k = 1, 2, \ldots, n
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial b_k}\left(Var(\hat{y}_1(x_0) - y_1(x_0))\right) &= 2\sum_{i=1}^{n} a_i Cov(Y_1(x_i); Y_2(x_k)) \\
&+ 2\sum_{i=1}^{n} b_i Cov(Y_1(x_i); Y_2(x_k)) \\
&- 2Cov(Y_1(x_0); Y_2(x_k)) + 2\vartheta_2 = 0 \text{ para } k = 1, 2, \ldots, m
\end{aligned}$$

$$\frac{\partial}{\partial \vartheta_1}\left(Var(\hat{y}_1(x_0) - y_1(x_0))\right) = 2\left(\sum_{i=1}^{n} a_i - 1)\right) = 0$$

$$\frac{\partial}{\partial \vartheta_2}\left(Var(\hat{y}_1(x_0) - y_1(x_0))\right) = 2\sum_{i=1}^{n} b_i = 0$$

A variância do erro fica:

$$\begin{aligned}
Var(\hat{y}_1(x_0) - y_1(x_0)) &= Cov\big(Y_1(x_0); Y_1(x_0)\big) - \vartheta_1 - \sum_{i=1}^{n} a_i Cov\big(Y_1(x_i); Y_1(x_0)\big) \\
&\quad - \sum_{j=1}^{m} b_j Cov(Y_2(x_j); Y_1(x_0))
\end{aligned}$$

O método de cokrigagem poderá ser escrito em termos de semivariogramas desde que as covariâncias cruzadas sejam simétricas. A continuidade espacial será modelada utilizando semivariogramas posteriormente convertidos para as covariâncias equivalentes pela transformação:

$$C_{Y_1;Y_2}(u) = \gamma_{Y_1;Y_2}(\infty) - \gamma_{Y_1;Y_2}(u)$$

e empregados na matriz de krigagem dada por $C\,w = D$ onde:

$$C = \left(\begin{array}{c|c|c} C(Y_1;Y_1) & C(Y_1;Y_2) & 1 \\ \hline C(Y_2;Y_1) & C(Y_2;Y_2) & 1 \\ \hline 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right);\quad w = \left(\begin{array}{c} a \\ \hline b \\ \hline -\vartheta_1 \\ -\vartheta_2 \end{array}\right);\quad D = \left(\begin{array}{c} C(Y_1;Y_0) \\ \hline C(Y_2;Y_0) \\ \hline 1 \\ 0 \end{array}\right)$$

Este sistema de equações para a cokrigagem é válida somente para estimação pontual. Poderão ser estimadas as médias em um número suficientemente grande de pontos em coordenadas de uma região e então se obter a média destas estimativas. Isaaks e Srivastava (1989) alertam que, para que a solução das equações existam e sejam únicas, o conjunto das autocorrelações e das correlações cruzadas devem formar matrizes que sejam definidas positivas. Dizem ainda os autores que, se as variáveis forem obtidas nas mesmas coordenadas, as estimativas por cokrigagem e krigagem ordinária serão idênticas.

A condição necessária que garantirá que a matriz de correlação seja definida positiva é dada por:

$$\omega' C \omega = \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i \omega_j C_{(i;j)} > 0$$

onde $\omega = (\omega_1, \omega_2, \ldots, \omega_n)'$ é o vetor de pesos de krigagem e pelo menos um de seus elementos deve ser diferente de zero.

Essa condição garante que a variância de qualquer variável aleatória formada pela combinação linear ponderada pelos pesos ω de outras variáveis aleatórias será positiva, ou seja, tem-se a garantia de que a variância do erro de estimação dada por $\left(\hat{Y}(x_0) - Y(x_0)\right)$ será positiva.

Ver Hoef e Barry (1998) usam o termo cokrigagem para referir-se a uma predição de uma variável primária em uma específica localização x_0 a partir de um conjunto multivariado de dados e o termo predição espacial quando se deseja predizer um vetor de variáveis aleatórias (de diferentes tipos) também em uma específica localização x_0. Eles destacam três problemas com a aplicação da cokrigagem tradicional. O primeiro surge quando se pretende minimizar o erro médio quadrático de predição usando o semivariograma cruzado, na forma porposta por

Journel e Huijbregts (1978). O procedimento será viável, segundo eles, quando a função de covariância cruzada for uma função par e de reflexão simétrica, ou seja, $C_{(i;j)}(h) = C_{(i;j)}(-h)$. A condição de simetria é muito restritiva e pode tornar questionável o uso da função tradicional do semivariograma cruzado.

O segundo problema será o de estimar o semivariograma cruzado quando os dados de ambas as variáveis envolvidas forem tomados nas mesmas coordenadas. Os autores propõem uma adaptação do que chamaram pseudo-variograma cruzado, dado por:

$$2\gamma_{(k;m)}(x_i;x_j) \equiv Var\left(Y_k(x_i) - Y_m(x_j)\right)$$

o que elimina a necessidade das variáveis estarem localizadas nas mesmas coordenadas.

O terceiro problema é a dificuldade em produzir modelos de semivariogramas cruzados válidos que sejam consistentes com os modelos de semivariogramas conhecidos. Por modelos válidos os autores se referem àqueles cuja variância de predição se mantém positiva.

Uma outra forma de se derivar resultados de predição para os modelos de corregionalização aqui discutidos pode ser pelas propriedades conhecidas da distribuição gaussiana multivariada. Assim, considerando-se a intenção de se predizer Y_1 em uma coordenada x_0 $\left(Y_1(x_0) = Y_0\right)$ no modelo bivariado dado pela Equação 3.7, utiliza-se a distribuição conjunta dada por:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ Y_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{bmatrix}; \sigma^2 \begin{bmatrix} \Sigma_{00} & \Sigma_{01} & \Sigma_{02} \\ \Sigma_{10} & \Sigma_{11} & \Sigma_{12} \\ \Sigma_{20} & \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

como $\mu_{12} = \mu_1 + \Sigma_{12}{\Sigma_{22}}^{-1}(Y_2 - \mu_2)$ e $\Sigma_{12} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ então o preditor será dado por:

$$\left[Y_0|Y_1;Y_2\right] \sim \mathsf{N}_n\left(\mu_{0|12};\Sigma_{0|12}\right) \tag{3.12}$$

onde:

$$\mu_{0|12} = \mu_0 + \begin{bmatrix} \Sigma_{01} & \Sigma_{02} \end{bmatrix}' \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \left(\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) \qquad \text{e} \qquad \Sigma_{0|12} = \Sigma_{00} - \Sigma_{012}$$

3.5 REDUÇÃO DO NÚMERO DE VARIÁVEIS AOS COMPONENTES PRINCIPAIS

A análise de componentes principais – ACP é amplamente utilizada em pesquisas e mais recentemente, vem sendo aplicada a conjuntos de variáveis com dados autocorrelacionados em modelos geoestatísticos.

Este tipo de análise estatística de dados tem a finalidade de transformar linearmente variáveis correlacionadas em seus componentes principais não correlacionados e organizar esses componentes em ordem decrescente de suas variâncias. A idéia é reduzir a quantidade de dados aos componentes que retêm a maior parte da variância total do conjunto de variáveis. Deve-se aqui atentar para o fato de que, na presença de valores discrepantes (*outliers*) a variabilidade dos dados poderá ser comprometida, alterando o papel da variável portadora desses valores no processo de análise dos componentes do conjunto. Para processos gaussianos, os componentes escolhidos podem ser tidos como fatores.

Seja $Y = (Y_1, Y_2, \ldots, Y_p)$ um processo estocástico p-dimensional onde cada variável Y_k ($k = 1, 2, \ldots, p$) segue o modelo definido pela Equação 2.3. É importante explicar a estrutura de covariância desse processo para a redução de seu número de variáveis devido a redundâncias ou de uma interpretação correlacional (JOHNSON; WICHERN, 1992). Esse tipo de análise de dados é tido como um processo intermediário para investigações mais amplas como regressão múltipla ou análise de agrupamentos.

Considera-se o vetor Y e a partir dele, constrói-se a matriz de covariâncias

$$\Sigma = E\left((Y-\mu)(Y-\mu)'\right)$$

sendo $\mu = (\mu_1, \mu_2, \ldots, \mu_p)'$ o vetor das médias relativas a cada variável do vetor Y. Já os elementos da matriz de covariâncias amostrais são:

$$s_{kk'} = \frac{1}{n-1}\sum_{i=i}^{n}(y_{ik} - \bar{y_k})(y_{ik'} - \bar{y_{k'}}) \qquad k, k' = 1, 2, \ldots, p. \tag{3.13}$$

ou, em forma matricial, $S = (n-1)^{-1}(Y - \bar{Y})(Y - \bar{Y})'$.

Decompondo-se a matriz de covariâncias obtém-se os p pares de autovalores e autovetores associados $(\lambda_k; e_k)$, tais que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

Para Kolman (1997) sendo Y um conjunto de vetores em um mesmo espaço vetorial,

então um outro vetor Cp, nesse mesmo espaço vetorial será uma combinação linear dos vetores de Y se existirem números reais $a_1,\ldots,a_p$ tais que $Cp = a_1Y_1, a_2Y_2, \ldots a_pY_p$. Assim, segundo Reis (1997), pode-se escrever o vetor Y como uma combinação de seus elementos como:

$$\begin{aligned} Cp_1 &= a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1p}Y_p \\ Cp_2 &= a_{21}Y_1 + a_{22}Y_2 + \ldots + a_{2p}Y_p \\ &\vdots \\ Cp_p &= a_{p1}Y_1 + a_{p2}Y_2 + \ldots + a_{pp}Y_p \end{aligned}$$

sendo Cp_k a k-ésima componente principal (não correlacionada) cuja variância seja a maior possível, ou seja:

- $Var(Cp_k) = e_k'\Sigma e_k = \lambda_k$;
- $Cov(Cp_j; Cp_k) = e_j'\Sigma e_k = 0 \qquad j \neq k$;
- $\sum_{k=1}^{p} Var(Y_i) = \sum_{k=1}^{p} Var(Cp_i) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$.

A porcentagem de contribuição de cada componente é determinada como:

$$\%CCp_k = \lambda_k \left(\sum_{j=1}^{p} \lambda_j \right)^{-1}. \tag{3.14}$$

Desta forma, aquelas primeiras m variáveis Y_k que acumularem maior porcentagem, poderão ser substituídas pelas m componentes principais, reduzindo assim, o número de variáveis sem grande perda na variabilidade do processo.

Segundo Johnson e Wichern (1992) o coeficiente de correlação entre as componentes e as variáveis primárias Y_k é dado por:

$$\rho(Cp_i; Y_k) = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{VarY_k}} = \frac{e_{ik}\sqrt{\lambda_i}}{\sigma_{ii}} \qquad \text{onde } i;k = 1,2,\ldots,p.$$

Apesar da correlação entre as p variáveis e seus componentes principais ajudar a interpretar o papel dos componentes, esta mede somente a contribuição univariada de um particular Y_k para formar a componente Cp_k e não a sua importância na presença das demais. Quando as variáveis Y_k forem processos medidos em escalas diferentes, recomenda-se utilizar a sua padronização para que possam ser comparáveis. O processo de seleção de componentes principais a partir de variáveis padronizadas Z_k se dá a partir da matriz de correlações R obtida

como:

$$R = \left(V^{-\frac{1}{2}}\right)^{-1} \Sigma \left(V^{-\frac{1}{2}}\right)^{-1} \tag{3.15}$$

onde Σ é a matriz de covariâncias e:

$$V^{-\frac{1}{2}} = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 & \dots & 0 \\ 0 & \sqrt{\sigma_{22}} & \dots & 0 \\ \vdots & \vdots & & 0 \\ 0 & 0 & \dots & \sqrt{\sigma_{pp}} \end{pmatrix}.$$

Decompondo a matriz de correlações R obtém-se também os p pares de autovalores e autovetores ordenados $(\lambda_k; e_k)$. A k-ésima componente principal padronizada será:

$$Cp_k = \sum_{k=1}^{p} e_k Z_k \qquad \text{onde} \qquad Z_k = \frac{Y_k - \overline{Y_k}}{\sqrt{s_{kk}}}. \tag{3.16}$$

A porcentagem de variação explicada por cada componente será dada pela Equação 3.14 e a correlação entre componente e variável padronizada será dada por:

$$\rho(Cp_k, Z_l) = e_{kl}\sqrt{\lambda_k}; \qquad k; l : 1, 2, \dots, p. \tag{3.17}$$

Wackernagel (1998) diz que é de vital importância tal tipo de análise para verificar se os dados são intrinsecamente correlacionados, senão o método geoestatístico multivariado poderá gerar resultados viesados.

Para se detectar uma correlação intrínseca em dados autocorrelacionados no espaço, há a necessidade de se verificar se os dados seguem um modelo de correlação intrínseca. Nesse modelo, todo autovariograma de duas variáveis Y_i, Y_j serão proporcionais a um variograma geral $\gamma(u)$, ou seja,

$$\gamma_{ij} = b_{ij}\gamma(u) \qquad \text{para } i, j = 1, 2, \dots, n$$

onde os b_{ij} são coeficientes.

Uma corregionalização (um conjunto de variáveis espacialmente correlacionadas) é intrinsecamente correlacionada quando o quociente:

$$\frac{\gamma_{ij}(u)}{\sqrt{\gamma_{ii}\gamma_{jj}(u)}} = \frac{b_{ij}}{\sqrt{b_{ii}b_{jj}}} = r_{ij}$$

é constante para qualquer distância u. Notar que a correlação entre duas variáveis não depende de u, diferentemente da autocorrelação de cada uma das variáveis separadamente.

A correlação intrínseca pode ser avaliada determinando suas componentes principais para a seguir determinar o variograma cruzado entre os primeiros componentes principais. No caso de existência de correlação intrínseca, o variograma cruzado resultante será nulo, caso contrário, as componentes serão correlacionadas espacialmente em alguma região do espaço e então o modelo deverá ser preterido a favor de outros modelos de corregionalização.

3.6 APLICAÇÃO COM MODELOS GEOESTATÍSTICOS MULTIVARIADOS

3.6.1 Dados da Pesquisa

Os dados utilizados para a análise multivariada foram os mesmos apresentados na Seção 2.8.1. Entretanto o arquivo de dados foi formatado às necessidades requeridas para esse tipo de análise. A forma pela qual foi feita essa formatação está descrita em código próprio nos anexos desse trabalho.

3.6.2 Recursos computacionais

Nesta fase da análise, foi empregada a função `likfitBGCCM`, implementada no pacote `geoR`, que estima, pelo método da máxima verossimilhança os parâmetros dos componentes do modelo geoestatístico bivariado, conforme a Equação 3.2. A função e seus principais argumentos padrão são dados por:

```
likfitBGCCM(geodata1, geodata2,
      ini.sigmasq, ini.phi,
      cov0.model="matern", cov1.model="matern",cov2.model="matern",
      kappa0=0.5, kappa1=0.5, kappa2=0.5,
      fc.min = c("optim", "nlminb"), ...)
```

onde `geodata1` e `geodata2` representam, respectivamente os objetos associados às variáveis Y_1 e Y_2 do modelo bivariado. A variável Y_2 foi considerada a variável secundária no caso bivariado

ou a componente representante de um conjunto de variáveis, determinada pela ACP no caso multivariado. Maiores detalhes sobre os demais argumentos da função pode ser obtida diretamente na documentação desta função, disponível no pacote geoestatístico geoR (RIBEIRO JR; DIGGLE, 2001).

A ACP foi desenvolvida diretamente no ambiente do R (R Development Core Team, 2008) utilizando-se comandos fundamentais de operações com matrizes. Os roteiros das análises estão apresentadas pelo código fonte em anexo.

3.6.3 Análise geoestatística dos dados de produtividade de soja

Análise dos Componentes Principais

Na área de plantio, juntamente com a produtividade de soja, foram coletadas 256 amostras de solo. Foram medidas as variáveis químicas P, H, K, MO e SB. Considerou-se nesta análise 150 observações aleatoriamente amostradas das 256 disponíveis de cada uma das cinco variáveis. As respectivas matrizes de média e covariâncias resultaram em:

$$\bar{X} = \begin{pmatrix} 4,168 \\ 5,121 \\ 0,334 \\ 52,685 \\ 55,029 \end{pmatrix} \qquad S = \begin{pmatrix} 1,99 & -0,16 & 0,03 & 1,99 & -3,58 \\ -0,16 & 0,23 & -0,00 & -1,81 & 5,28 \\ 0,03 & -0,00 & 0,01 & 0,14 & -0,01 \\ 1,99 & -1,81 & 0,14 & 41,97 & -42,25 \\ -3,58 & 5,28 & -0,01 & -42,25 & 151,79 \end{pmatrix}$$

Na diagonalização da matriz de correlação os autovalores Λ e os autovetores EIG resultantes foram:

$$\Lambda = \begin{pmatrix} 166,453 \\ 27,622 \\ 1,870 \\ 0,038 \\ 0,006 \end{pmatrix} \qquad EIG = \begin{pmatrix} 0,024 & -0,028 & 0,999 & -0,014 & -0,014 \\ -0,034 & 0,000 & -0,013 & -0,998 & 0,053 \\ 0,000 & -0,005 & 0,014 & 0,053 & 0,998 \\ 0,322 & -0,946 & -0,035 & -0,011 & -0,004 \\ -0,946 & -0,323 & 0,015 & 0,031 & -0,003 \end{pmatrix}$$

A porcentagem de explicação de cada componente resultou, respectivamente em

84,93% 14,09% 0,95% 0,02% e 0,00%. Adotou-se então a primeira componente principal – CP1, para representar o variável secundária no modelo geoestatístico bivariado. Essa componente ficou composta da seguinte maneira:

$$CP_1 = 0,024\,P - 0,034\,PH + 0,322\,MO - 0,946\,SB$$

Ajuste dos parâmetros do modelo bivariado

A Tabela 3.1 apresenta as estimativas por MV para o modelo bivariado dado pela Equação 3.2. Nessa tabela, μ_1 e μ_2 são estimativas do efeito sistemático do modelo, presentes respectivamente nas variáveis primária (Soja) e secundária (iCone ou CP1), σ^2_{01}, σ^2_1, σ^2_{02} e σ^2_2 são, respectivamente, parâmetros de escala associados as matrizes de correlação $R_0(\phi_0)$ das variáveis primária e secundária, $R_1(\phi_1)$ da variável primária e $R_2(\phi_2)$ da variável secundária. ϕ_0, ϕ_1 e ϕ_2 são parâmetros da função de correlação que, neste estudo foi adotada a de Matèrn com $\kappa = 0,5$, que equivale a uma função exponencial.

Tabela 3.1: Parâmetros estimados para os modelos geoestatísticos bivariados por MV.

Modelo	μ_1	μ_2	σ^2_{01}	σ^2_1	σ^2_{02}	σ^2_2	ϕ_0	ϕ_1	ϕ_2
S128iCone	2,7640	21,9	0,14	0,32	9,32	31,27	14,7	24,6	20,3
S64iCone	2,8861	21,9	0,14	0,33	7,92	33,52	21,8	24,7	19,6
S128CP1	2,7576	0,3	0,00	0,19	0,01	0,01	6,3	4,7	52,0
S64CP1	2,8078	0,3	0,00	0,16	0,00	0,01	8,9	0,1	39,9
IMAArg	25,500	24,0	9,31	50,48	103,30	0,10	722,6	1.488,7	1.675,2

Obs.: μ_1 e μ_2: parâmetros do efeito sistemático, σ^2_{01}, σ^2_1, σ^2_{02} e σ^2_2: parâmetros de escala associados a $R_0(\phi_0)$, $R_1(\phi_1)$ e $R_2(\phi_2)$ respectivamente. S128 e S64: soja tomadas em 128 e 64 parcelas, iCone: Índice de Cone, CP1: primeira componente principal do conjunto de variáveis P, PH, K, MO e SB, Arg: Teor de Argila.

Estatísticas descritivas da predição bivariada de produtividade de soja suportada por iCone e CP1

A Tabela 3.2 mostra a média das predições multivariadas de soja utilizando uma amostra aleatória de 128 parcelas de produtividade conhecida na área como variável primária e dois cenários para a variável secundária. O primeiro utiliza 150 amostras aleatórias de iCone na mesma área e o segundo utiliza 150 amostras aleatórias da variável CP1. Essa Tabela mostra ainda a média das predições multivariadas de soja utilizando uma amostra aleatória de 64 parcelas de produtivade conhecida na área como variável primária e os dois cenários para variável

secundária apresentada no caso anterior. Observa-se que são iguais às estimativas para Soja128 nas duas concepções de variável secundária bem como o são no caso da Soja64.

Tabela 3.2: Estatísticas descritivas das predições da produtividade de soja medida em 128 e 64 pontos aleatórios condicionadas às observações de iCone e da CP1 e predições de IMA condicionadas a 555 observações de Teor de Argila.

Delineamento	Mínimo	Média	Máximo	D.P.	C.V.(%)	ER(%)
S128iCone	1,8626	2,7708	3,7627	0,3135	11,3	0,8
S64iCone	2,0243	2,8049	3,6266	0,2672	9,5	2,0
S128CP1	2,0280	2,7583	3,4739	0,1949	7,1	0,3
S64CP1	2,7943	2,8084	2,8244	0,0043	0,2	2,1
IMAArg	17,740	25,750	37,150	3,3100	12,9	–

Estimativas de soja em t ha^{-1} e estimativas de IMA em m^3. D.P.: Desvio Padrão, C.V.: Coeficiente de Variação, ER(%): erro em relação ao valor referência de 2,75 t ha^{-1}.

Na tabela 3.3, a porcentagem dos pontos estimados por método de krigagem com modelo bivariado, incidentes em cada intervalo de classificação segundo os diferentes modelos, mostra a tendência do número de pontos preditos se concentrarem na classe central, quando comparada com a distribuição percentual dos dados originais. No caso em que se empregou CP1 como informação adicional no modelo esse efeito fica mais evidente, chegando a concentrar todos os pontos na classe de 2,61 a 3,16 t ha^{-1}, o que não permitiu representar um padrão espacial pelo respectivo mapa.

Tabela 3.3: Porcentagem dos pontos estimados com modelo bivariado, incidentes nos respectivos intervalo de classificação.

Classe	PM	S128 e IC	S64 e IC	S128 e CP1	S64 e CP1	S256d
1,19 a 2,34	1,8	8,3	4,2	2,3	0,0	20,3
2,34 a 2,61	2,5	22,8	18,3	17,5	0,0	19,9
2,61 a 2,85	2,7	28,4	34,2	51,9	100,0	20,7
2,85 a 3,16	3,0	28,8	33,6	25,7	0,0	19,1
3,16 a 4,14	3,7	11,7	9,7	2,6	0,0	19,9

PM: ponto médio da classe, S128: soja em 128 amostras, S64: soja em 64 amostras, IC: iCone, CP1: primeira componente principal, S256d: dados originais de soja classificados.

Mapa da média da predição bivariada de produtividade de soja

A Figura 3.3 apresenta os mapas de produtividade de soja estimados por métodos de krigagem condicional em modelos bivariados em uma malha de 690 pontos. No mapa à esquerda o modelo aplicado utilizou informações de soja em 128 pontos e iCone em 150 pontos

da variável secundária. No mapa da direita, o modelo utilizou informações em 64 pontos sorteados da variável principal (soja) e 150 pontos da variável secundária (iCone). Os pontos brancos da figura correspondem às produtividades abaixo de 2,34 t ha^{-1} e os pontos pretos correspondem às produtividades acima de 3,16 t ha^{-1}. Os demais pontos variam seu tom de cinza proporcionalmente às classes que representam. Nessas figuras nota-se que os mapas que se utilizaram de informações de uma segunda variável indicaram um padrão de variabilidade espacial semelhante ao descrito na Figura 2.13 da página 53, entretanto não definiram visualmente zonas de produtividade de amplas dimensões.

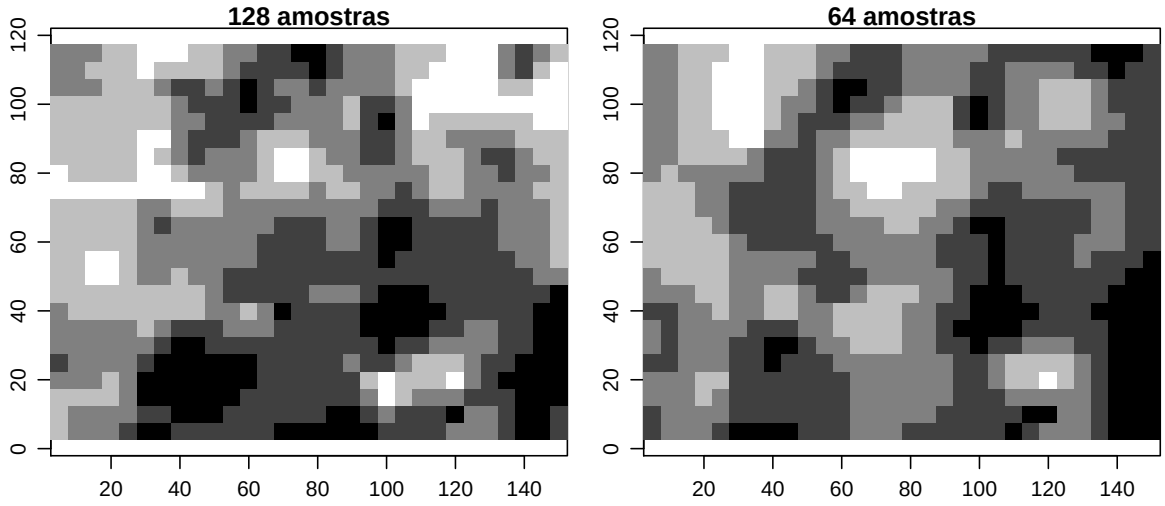

Figura 3.3: Mapas de produtividade de soja em modelos bivariados em uma malha regular de 690 pontos. No modelo do mapa da esquerda utilizou-se 128 amostras de soja e no da direita, 64. A variável secundária foi 150 amostras de iCone

A Figura 3.4 também apresenta mapas de produtividade de soja estimados por métodos de krigagem condicional em modelos bivariados em uma malha de 690 pontos. Na construção da figura da esquerda o modelo utilizou informações em 128 pontos sorteados da variável principal (soja) e 150 pontos da variável secundária (CP1). No mapa da direita, o modelo utilizou informações em 64 pontos sorteados da variável principal (soja) e 150 pontos da variável secundária (CP1). Os pontos brancos da figura correspondem às produtividades abaixo de 2,34 t ha^{-1} e os pontos pretos correspondem às produtividades acima de 3,16 t ha^{-1}. Os demais pontos variam seu tom de cinza proporcionalmente às classes que representam. Tanto nessa figura como na Tabela 3.3 nota-se que o modelo que se utilizou de 128 amostras de soja na variável principal, foram preditos poucos pontos que correspondem a valores baixos (2,3% do total) e poucos pontos que correspondem altos (2,6% do total), levando a um mapa que ilustrou um padrão espacial mais concentrado nas classes centrais. O efeito dessa concentração foi maior no modelo que se utilizou de 64 amostras de soja. Todos os pontos foram estimados dentro do intervalo central de 2,61 a 2,85 t ha^{-1}, intervalo esse que contém a produtividade média de 2,75

t ha^{-1} obtida de fato na área. Tal resultado não permitiu contruir um mapa capaz de identificar zonas diferenciadas de produtividade.

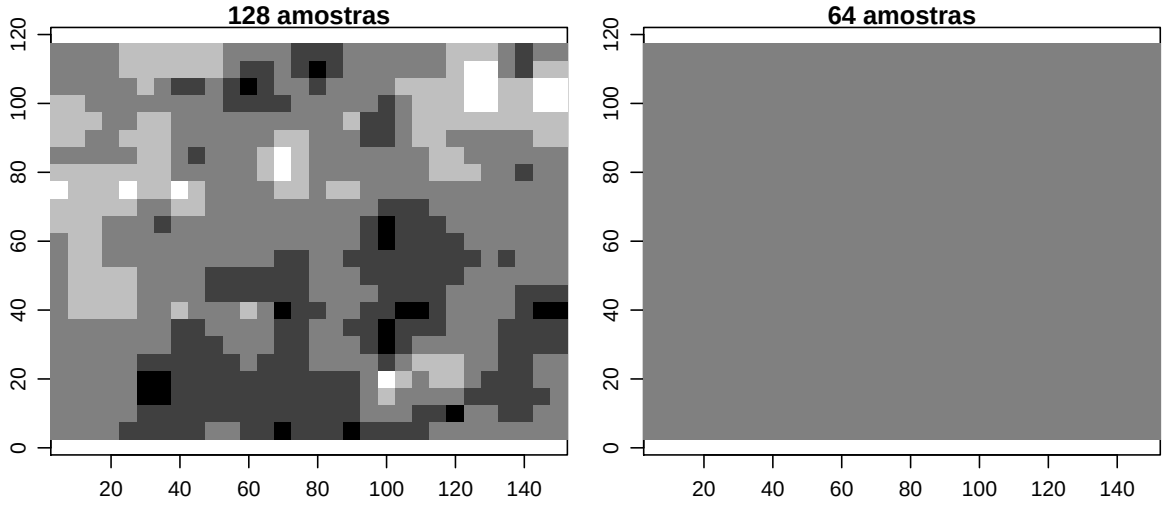

Figura 3.4: Mapas de produtividade de soja estimados por MV em modelos bivariados em uma malha regular de 690 pontos. No modelo do mapa da esquerda utilizou-se 128 amostras de soja e no da direita, 64. A variável secundária foi 150 amostras da CP1

3.6.4 Análise geoestatística dos dados rendimento de *P. Taeda* L.

Os dados a que se refere essa pesquisa envolvem duas variáveis, o IMA e o Teor de Argila. A variável IMA, de interesse principal, foi obtida de uma amostragem de 18 localizações e a variável Teor de Argila (secundária) foi coletada em 555 localizações.

Ajuste dos parâmetros do modelo bivariado (IMA e Argila)

Na Tabela 3.1 encontam-se as estimativas por MV para o modelo bivariado. Nessa tabela, μ_1 e μ_2 são estimativas do efeito sistemático do modelo presentes, respectivamente, nas variáveis primária (IMA) e secundária (Argila); σ^2_{01}, σ^2_1, σ^2_{02} e σ^2_2 são, respectivamente, parâmetros de escala associados às matrizes de correlação $R_0(\phi_0)$ das variáveis primária e secundária, $R_1(\phi_1)$ da variável primária e $R_2(\phi_2)$ da variável secundária. ϕ_0, ϕ_1 e ϕ_2 são parâmetros da função de correlação que, neste estudo, foi adotada a de Matèrn com $\kappa = 0,5$. Nesse modelo, considerou-se o efeito pepita relativo igual a zero.

Estatísticas descritivas da predição bivariada do IMA suportada pelo Teor de Argila

Na última linha da Tabela 3.2 encontra-se as medidas descritivas da média de predição de IMA, onde a média obtida de 25,73 m^3 ha^{-1} ano^{-1} é compatível, em ordem de grandeza,

com o valor de 26,53 m^3 ha^{-1} ano^{-1} obtidos pela krigagem convencional com modelo ajustado pelo método MV em 18 amostras. Esses resultados são ainda comparáveis com a literatura, conforme Mainardi, Schneider e Finger (1996) e Carvalho et al. (1999).

Mapa da média da predição bivariada IMA suportada pelo Teor de Argila

Na Figura 3.5 apresenta-se o mapa de predição bivariada de IMA condicionada ao Teor de Argila. Nesse mapa, 4,5% das predições de IMA ficaram abaixo de 20,5 m^3 ha^{-1} ano^{-1}, 17,7% ficaram entre 20,5 e 23,4 m^3 ha^{-1} ano^{-1}, 55,6% ficaram entre 23,4 e 28,3 m^3 ha^{-1} ano^{-1}, 19,6% entre 28,3 e 32,5 m^3 ha^{-1} ano^{-1} e 2,7% acima de 32,5 m^3 ha^{-1} ano^{-1}. Esses resultados são muito próximos aos obtidos por predição bayesiana univariada (Tabela 2.13), entretanto o mapa desta figura definiu mais zonas de rendimento de IMA quando comparada coma a Figura 2.22, tanto naquela com predições por krigagem baseada em MV (esquerda) quanto aquela obtida por predição bayesiana (direita). Os resultados dos agrupamento em classes indicaram ainda uma concentração das predições bivariadas em direção às classes mais centrais, efeito esse não caracterizado pelo mapa.

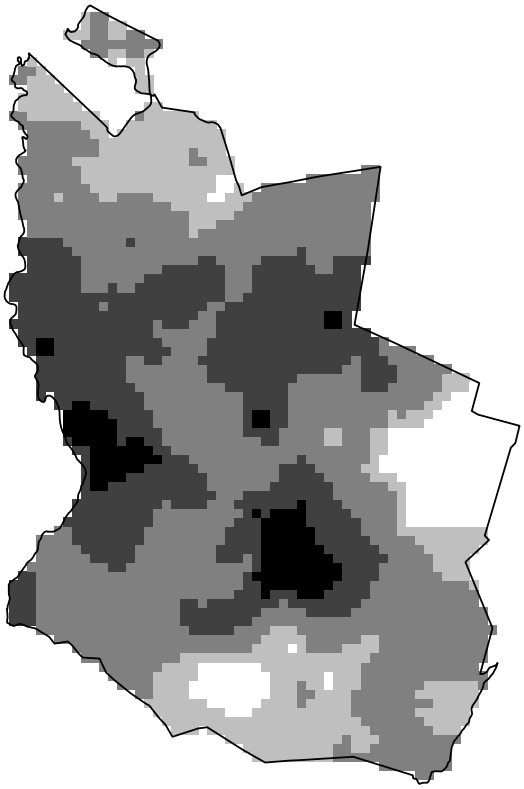

Figura 3.5: Mapa de predição de IMA classificada pelos quartis, usando krigagem convencional e Teor de Argila como variável secundária no modelo bivariado.

3.6.5 Conclusões sobre o método multivariado

Conclusões sobre a estimação bivariada da produtividade de soja

- Usar uma variável simples (iCone) ou uma variável composta (CP1) como suporte para a descrição da variabilidade espacial de uma variável de interesse primário (soja) em 128 amostras, não produziu diferenças nem na média de predição nem nos aspectos dos respectivos mapas temáticos. No caso de predições baseadas em 64 amostras o método produziu uma concentração na classe central quando se utilizou CP1 como variável secundária.

- O tamanho da amostra da variável de interesse primário (soja) produziu diferenças no aspecto dos mapas temáticos. A amostra de tamanho menor induziu um mapa com agrupamento maior de zonas de classificação, comparado com os mapas com amostra de tamanho maior.

Conclusões sobre IMA e Teor de Argila

- A média dos valores obtidos nos mapas, comparada a média dos valores nos mapas do caso univariado não foram conclusivas.

- O mapa da distribuição espacial bivariada de IMA condicionada à distribuição espacial do Teor de Argila definiu melhores zonas de manejo viáveis quando comparada com os mapas produzidos em dois métodos univariados.

4 CONCLUSÕES E SUGESTÕES DE TRABALHOS FUTUROS

Conclusões

- A abordagem fundamentada na declaração explícita de modelos associados a métodos formais de inferência baseados na informação contida na função de verossimilhança mostrou-se adequada em análise dos dados, bem como aponta caminhos para generalizações de procedimentos de análise tais como extensões multivariadas de modelos.
- A adoção de métodos bayesianos, que incorporam a incerteza associada aos parâmetros nos procedimentos de predição permitiu, nos exemplos analisados, uma melhor definição e melhor caracterização da incerteza sobre zonas viáveis de manejo em experimentos agronômicos que envolvem a produção de mapas temáticos, sobretudo quando se dispõe de amostra pequena da variável de interesse. Permite estimar a média do processo gaussiano subjacente que reflete a produtividade total quando comparado a uma estimação por krigagem convencional.
- Métodos bivariados produzem mapas temáticos com variações espaciais mais definidas e maior quantidade de zonas de manejo quando comparadas com krigagem convencional univariada ou métodos geoestatísticos bayesianos univariados.
- A formulação e adoção de métodos multivariados permite melhor uso da informação disponível bem como abre possibilidade para a investigação de planos amostrais específicos ao problema em questão que combinem dados de diferentes naturezas.

Proposição de trabalhos futuros decorrente dos resultados deste trabalho

- Escrever e implementar computacionalmente uma solução bayesiana para o modelo bivariado proposto neste trabalho.

- Comparar diferentes formulações de modelos bivariados e multivariados e seum desempenho em análise de dados reais.
- Comparar o efeito de diferentes tamanhos de amostras da variável principal e da variável secundária em modelos bivariados com base em dados reais e dados simulados.
- Qualificar as zonas de manejo produzidas por diferentes métodos de estimação geoestatística.
- Criar critérios de classificação baseados em conceitos agronômicos.

Referências Bibliográficas

ABRAMOWITZ, M.; STEGUN, I. *Handbook of Mathematical Functions*. Ninth. New York: Dover, 1965.

AGUIRRE, A.; FARIA, D. M. C. P. de. *A utilização dos "preços hedônicos"na avaliação social de projetos*. [S.l.], 1996. Disponível em: <http://ideas.repec.org/p/cdp/texdis/td103.html>.

ATTEIA, O.; DUBOIS, J. P.; WEBSTER, R. Geostatistical analysis of soil contamination in the swiss jura. *Environmental Pollution*, n. 86, p. 315–327, 1994.

BAKHSH, A. et al. Spatial distribution of soil attributes affecting crop yield. *ASAE*, n. 97, p. 1032, 1997.

BALASTREIRE, L. A.; ELIAS, A. I.; AMARAL, J. R. Agricultura de precisão: Mapeamento da produtividade da cultura de milho. *Revista de Engenharia Rural*, n. 8, p. 97–111, 1997.

BARTLETT, M. *Stochastic Process*. USA: Cambridge University Press, 1955.

BOGNOLA, I. A. *Unidades de manejo para* Pinus Taeda *L. no planalto norte catarinense, com base em características do meio físico*. 180 p. Tese (Tese de Doutorado) — Universidade Federal do Paraná, Curitiba, 2007. Programa de Pós-graduação em Engenharia Florestal.

BOLSTAD, W. M. *Introduction to bayesian statistics*. USA: John Wiley & Sons, 2004.

BOX, G.; COX, D. An analysis of transformation. *JRSSB*, n. 26, p. 211–252, 1964.

BRAGA, L. P. V. Geoestatística e aplicações. *Anais do IX Simpósio Brasileiro de Probabilidade e Estatística do IME/USP - São Paulo*, p. 36p, 1990.

CAEIRO, S. et al. Delineation of estuarine management areas using multivariate geostatistic: the case of sado estuary. *Environmental Science e Tecnology*, v. 37, n. 18, p. 4052–4054, 2003.

CAMBARDELLA, C. A. et al. Field-scale variability of soil properties in central iowa soil. *Soil Science Society of Ameriaca Journal*, v. 58, p. 1501 – 1511, 1994.

CAPELLI, N. L. Agricultura de precisão: Novas tecnologias para o processo produtivo. *LIE/DMAQAG/ FEAGRI/UNICAMP*, 1999.

CARVALHO, A. P. de et al. *Efeitos de Características de Solo sobre a Capacidade Produtiva de Pinus taeta*. Colombo, 1999.

CARVALHO, J. R. P.; QUEIROZ, E. *Uso da cokrigagem colocalizada na determinação da distribuição espacial de precipitação*. Campinas, 2002. Comunicado Técnico.

COUTO, E. G.; CUNHA, C. N. Application of multivariate geostatistics to identify soil landscapes in the Pantanal of Mato Grosso - Brazil. *Revista Agricultura Tropical*, v. 6, n. 1, p. 48–65, 2002.

CRESSIE, N. Fitting variogram models by weighted least squares. *Mathematical Geology*, v. 17, n. 4, p. 563–586, 1985.

DEGROOT, M. *Probability and Statistics*. [S.l.]: Addilson-Wesley, 1989.

DIGGLE, P. J.; LOPHAVEN, S. Bayesian geostatistical design. *Scandinavian Journal of Statistics*, v. 33, p. 55–64, 2006.

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

DIGGLE, P. J.; TAWN, J. A.; MOYEED, R. A. Model-based geostatistics. *Royal Statistical Society*, v. 43, n. 3, p. 299–350, 1998.

DUDEWICZ, E.; MISHRA, S. *Modern Mathematical Statistics*. Singapore: John Wiley & Sons, 1988.

EHLERS, R. S. *Introdução à Inferência Bayesiana*. Curitiba: Universidade Federal do Paraná - Departamento de Estatística, 2006. Disponível em: <http://leg.est.ufpr.br/ ehlers/notas/bayes.pdf>.

EINAX, J. W.; SOLDT, U. Multivariate geostatistical analysis of soil contamination. *Freseniu's Journal of Analytical Chemistry*, v. 361, n. 1, p. 10–14, april 1998.

FARACO, M. A. et al. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade de soja. *Revista Brasileira de Ciências do Solo*, v. 32, n. 2, p. 463–476, 2008.

FILZMOSER, P.; REIMANN, C. Robust multivariate methods in geostatistics. *W. Gaul and G. Ritter, editors, Classification, Automation, and New Media*, p. 429–436, 2002.

FRASSON, F. R.; MOLIN, J. P. Análise da variabilidade espacial da produtividade de soja utilizando recursos do software r. In: USP/ESALQ (Ed.). *Congresso Brasileiro de Agricultura de Precisão*. São Pedro-SP: USP/ESALQ, 2006. p. 1–10.

GELMAN, A. et al. *Bayesian Data Analysis*. Second. Singapore: Chapman & Hall, 2003.

GOMES, F. P. *Estatística experimental*. Piracicaba: ESALQ, 1963.

GOOVAERTS, P. *Geostatistics for Natural Resources Evaluation*. Oxford: Oxford University Press, 1997.

HARTLEY, H. O.; RAO, J. N. K. Maximun likelihood estimation for the mixed analysis of variance model. *Biometrika*, v. 54, p. 93–108, 1967.

ISAAKS, E. H.; SRIVASTAVA, R. M. *Applied Geostatistics*. New York: Oxford University, 1989.

JOHNSON, R. A.; WICHERN, D. W. *Applied Multivariate Statistical Analysis*. Third. USA: Prentice-Hall, 1992.

JOURNEL, A. G.; HUIJBREGTS, C. J. *Mining Geostatistics*. London: Academic Press, 1978.

KOLMAN, B. *Introductory Linear Algebra with Applications*. USA: Prentice Hall, 1997.

KOLMAN, B. *Bayesian statistics: an introduction*. Third. New York: Hodder Arnold, 2004.

KRIGE, D. G. *A Statistical Approach to Some Mine Valuations and Allied Problems at Witwatersrand*. Tese (Master's thesis) — University of Witwatersrand, 1951.

LINDLEY, D. V. The 1988 world memorial lectures: The present position in bayesian statistics. *Statistical Science*, n. 5, p. 44–89, 1990.

MAINARDI, G. L.; SCHNEIDER, P. R.; FINGER, C. A. G. aes. Produção de *Pinus taeda L.* na região de Cambará do Sul, rs. *Revista Ciência do Solo*, v. 6, n. 1, p. 39–52, 1996.

MATÈRN, B. *Spatial Variation Analysis*. Second. Berlin: Springer Verlag, 1986.

MATHERON, G. Traite de geostatistique appliquee. *Bureau de Recherches Geologiques et Minieres*, v. 14, p. 1246–1266, 1962.

MATHERON, G. Principles of geostatistics. *Economic Geology*, v. 58, p. 1246–1266, 1963.

MATHERON, G. The intrinsic random function and their application. *Advances in Applied Probability*, n. 5, p. 508–541, 1973.

MOHAMED, S. B.; EVANS, E. J.; SHIEL, R. S. Mapping techniques and intensity of soil sampling for precision farming. *Proceedings of the 3rd International Conference on Precision Agriculture*, p. 217–226, 1996.

MOLIN, J. P. Agricultura de precisão: mais um desafio para o agricultor brasileiro. *Plantio Direto*, n. 39(3), p. 26–27, 1997.

MOLIN, J. P. Definição de unidades de manejo a partir de mapas de produtividade. *Engenharia Agrícola*, Jaboticabal, v. 22, n. 1, p. 83–92, 2002.

MOLIN, J. P. Fatores restritivos à adoção da agricultura de precisão. *Anais do II Congresso Brasileiro de Soja*, Embrapa Soja, Londrina, v. 180, n. 2, p. 221–229, 2002.

MONTGOMERY, D. C.; PECK, E. A. *Introduction to Linear Regression Analysis*. USA: Cambridge University Press, 1955.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. *Introduction to the Theory of Statistics*. Third. Singapore: McGraw-Hill, 1974.

OLIVEIRA, M. C. N. *Métodos de estimação de parâmetros em modelos geoestatísticos com diferentes estruturas de covariâncias: uma aplicação ao teor de cálcio no solo.* Tese (Tese de Doutorado) — Universidade de São Paulo/ESALQ, 2003.

PANNATIER, Y. *Variowin 2.2: Software for Epatial Data Analysis in 2D*. New York: Springer, 1996.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when blocks sizes are inequal. *Biometrika*, v. 58, p. 545–554, 1971.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística Bayesiana*. Lisboa: Fundação Calouste Gulbenkian, 2003. 446p.

PEBESMA, E. J.; WESSELING, C. G. Gstat: a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences*, v. 1, n. 24, p. 17–31, 1998.

PERRY, S. H. V.; IEMMA, A. F. Procedimento “mixed´´ do sas para análise de modelos mistos. *Scientia Agricola*, v. 56, n. 4, p. 959–967, 1999.

PREVEDELLO, B. M. S. *Variabilidade espacial de parâmetros de solo e planta*. 166 p. Tese (Tese de Doutorado) — Universidade de São Paulo, São Paulo, 1987. Curso do Pós-graduação em Nutrição de Plantas.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

REICHARDT, K.; VIEIRA, S. R.; LIBARDI, P. L. Variabilidade espacial de solos e experimentação de campo. *Revista Brasileira de Ciência do Solo*, v. 10, n. 1, p. 1–6, 1986.

REIS, E. *Estatística Multivariada Aplicada*. Lisboa: Silabo, 1997.

RIBEIRO JR, P. J.; DIGGLE, P. J. *Bayesian inference in gaussian model-based geostatistics*. Lancaster-UK, 1999. Maths. and Stats. Dept.

RIBEIRO JR, P. J.; DIGGLE, P. J. geoR: A package for geostatistical analysis. *R-NEWS*, v. 01, n. 2, 2001. ISSN 1609-3631.

ROSOLEN, C. A. et al. Crescimento radicular de plântulas de milho afetado pela resistência do solo à penetração. *Pesquisa Agropecuária Brasileira*, v. 34, n. 5, p. 821–828, 1999.

SCHABENBERGER, O.; GOTWAY, C. A. *Statistical Methods for Spatial Data Analysis*. New York: Chapman-Hall, 2005.

SOUZA, E. G. et al. Variabilidade espacial dos atributos químicos do solo em um latossolo roxo distrófico da região de Cascavel-PR. *Engenharia Agrícola*, v. 18, n. 3, p. 80–92, 1999.

SOUZA, M. Z.; MARQUES JR, J.; PEREIRA, G. T. Variabilidade espacial do pH, Ca, Mg e V% do solo em diferentes formas do relevo sob cultivo de cana-de-açúcar. *Revista Ciência Rural*, v. 34, n. 6, p. 1763–1771, nov-dez 2004.

TRAGMAR, B. B.; YOST, R. S.; UEHARA, G. Application of geostatistics to spatial studies of soil properties. *Advances in Agronomy*, v. 38, p. 45 – 94, 1985.

VER HOEF, J. M.; BARRY, R. P. Constructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, v. 69, p. 275–294, 1998.

VER HOEF, J. M.; CRESSIE, N. Multivariate spatial prediction. *Mathematical Geology*, v. 25, n. 2, p. 219–240, 1993.

WACKERNAGEL, H. Principal component analysis for autocorrelated data: a geoestatistical perspective. *Techinical Report 22/98-G - Centre de Géostatistique - Ecole des Mines de Paris*, 1998. Disponível em: <http://cg.ensmp.fr>.

WAKERNAGEL, H. *Multivariate geostatistics: an introduction with applications*. Third. Germany: Springer, 2003.

WALLER, L. A.; GOTWAY, C. A. *Applied spatial statistics for public health data*. USA: John Wiley & Sons, 1965. (Wiley series in probability and statistics).

WOLLENHAUPT, N. C.; WOLKOWSKI, R. P. Grid soil sampling. better crops with plant food. *NORCROSS*, v. 78, n. 4, p. 6–9, 1994.

YANG, C. et al. Spatial variability of field topography and wheat yield in the palouse region of the pacific northwest. *American Society of Agricultural Engineers*, v. 41, n. 1, p. 17–27, 1998.

ANEXO A – Figuras: Validação Cruzada

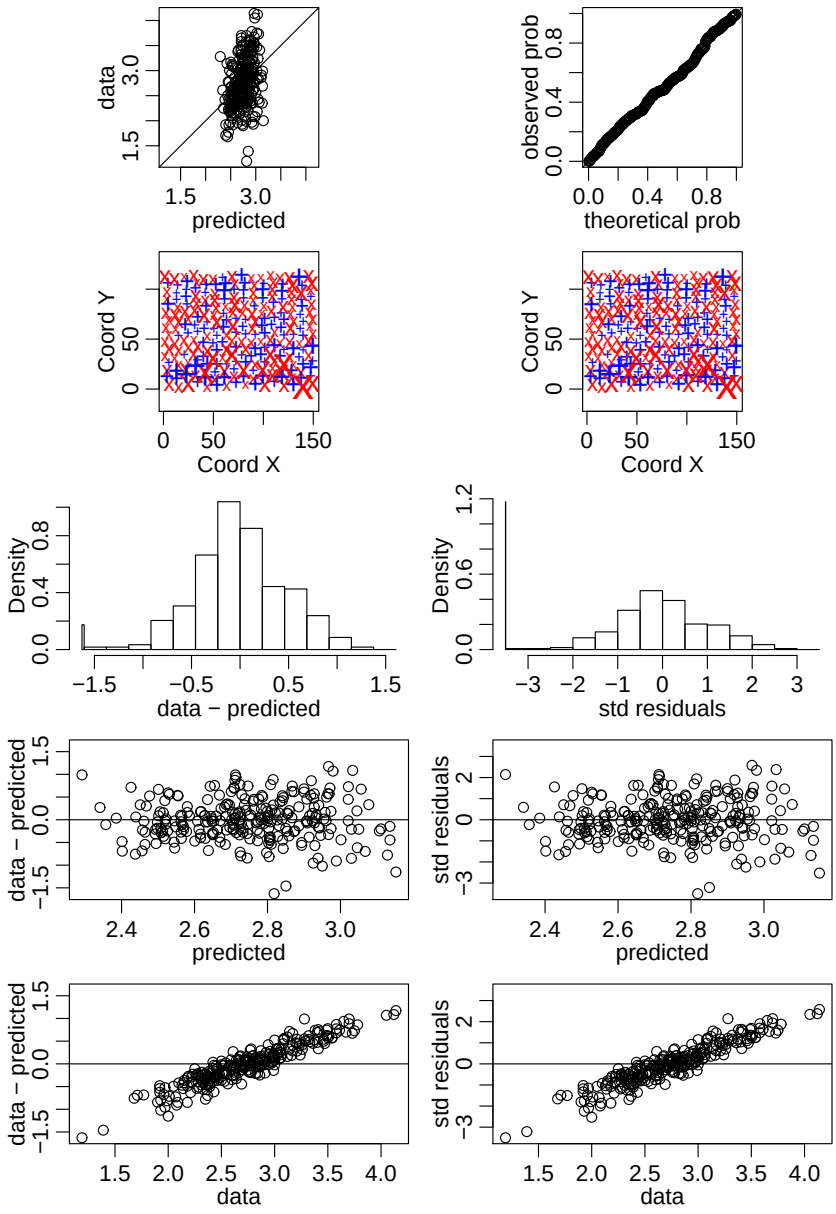

Figura A.1: Erros de predição por Validação cruzada. Predição nas mesmas coordenadas da malha de dados de estimação do modelo com a estratégia de retirar um ponto por vez e estimá-lo com o modelo.

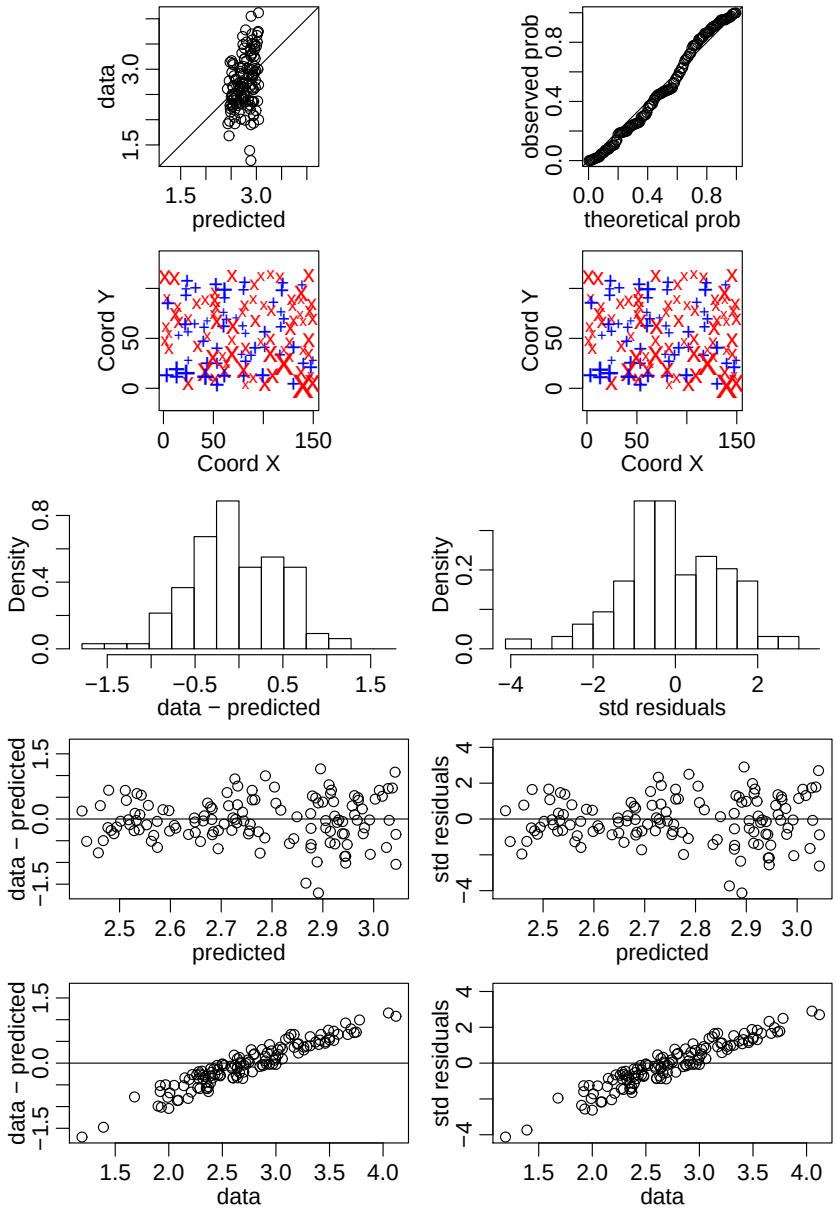

Figura A.2: Erros de predição por Validação cruzada. Predição em 128 coordenadas externas à malha de dados de estimação do modelo.

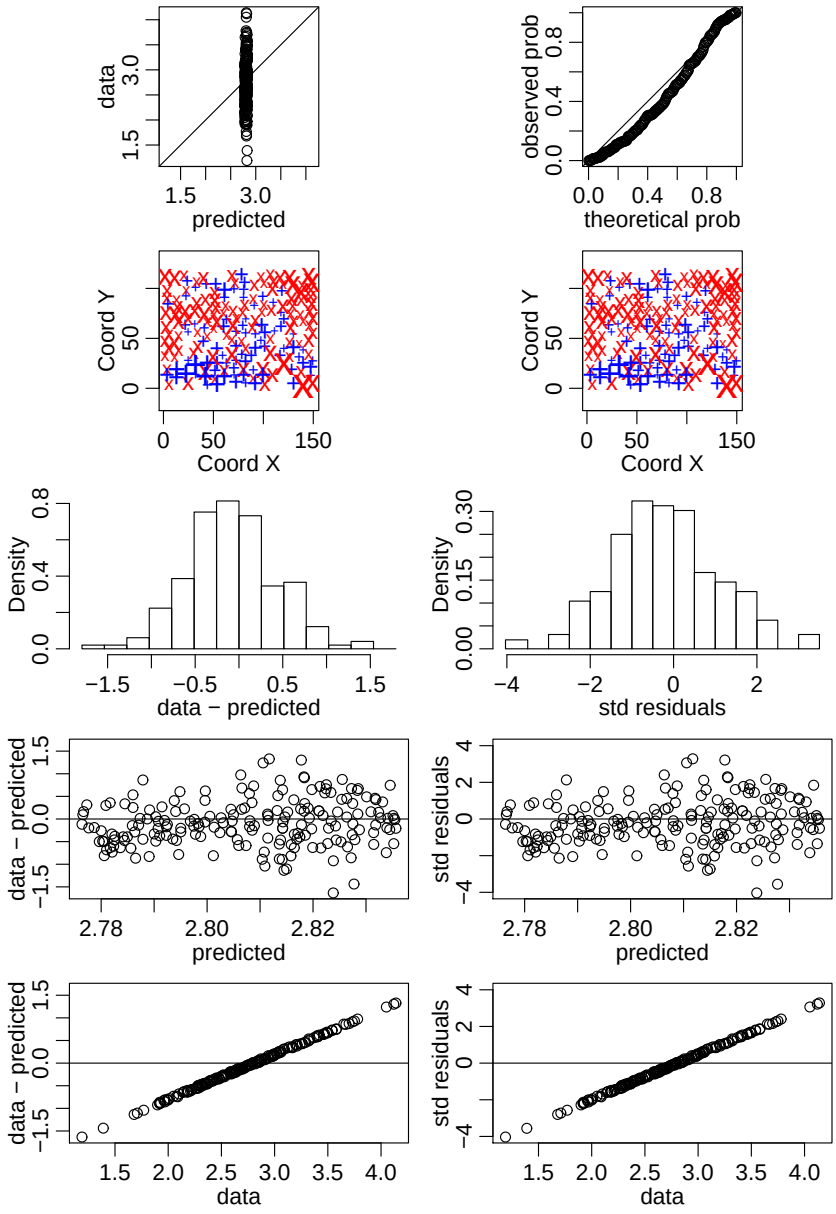

Figura A.3: Erros de predição por Validação cruzada. Predição em 192 coordenadas externas à malha de dados de estimação do modelo.

ANEXO B – Código fonte R das análises estatísticas

Listagem B.1: Análise univariada de dados de soja

```
## ##########################################################
##
## Analise Geoestatistica Univariada -- SOJA
##
## Aplicacao de metodos geoestatisticos multivariados
## em problemas de mapeamento de variaveis do sistema
## solo-planta.
##
## Dados: Soja 98
## Manejo convencional em 256 pontos amostrais
##
## Edson Antonio Alves da Silva, Prof. MSc.
##
## Orientador: Paulo Justiniano Ribeiro Jr., Prof. PhD.
##
## Criado em  : 16 de Marco de 2008
## Revisado em: 05 de julho de 2008
##
## ##########################################################
##
## Limpando a workspace
## ----------------------------------------------------------
##
objects()
rm(list=ls(all=TRUE))
oldpar <- par(no.readonly=TRUE)
##
##
## FUNCAO DE CLASSIFICACAO
## Classifica em 5 grupos separados por percentis
## Quantiles: 0,2 -- 0,4 -- 0,6 -- 0,8
## ----------------------------------------------------------
##
Class <- function(x)
    {
    if (x<2.341) x=1.8 # < 20%
       else
       {if (x<2.611) x=2.5 # < 40%
          else
            {if (x<2.851) x=2.7 # < 60%
               else
                 {if (x<3.161) x=3.0 else 3.7}
```

```
            }

        }
    }
## ------------------------------------------------------------------
## Entrada de dados (requer conexao com a internet)
##
Dados256 <- read.csv("http://wiki.leg.ufpr.br/data/media/pessoais/edson/
    tese/dados/coodetec/soja98d.csv", head=T, sep="", dec=",")
##
## numerando a sequencia de coordenadas
## ------------------------------------------------------------------
Dados256 <- cbind(Dados256,Ordem=seq(1:256))
##
## S E P A R A N D O   A M O S T R A S
## ------------------------------------------------------------------
##
## Criando indice para intercalar as parcelas
Parcelas <- rep(c(rep.int(0:1,8),rep.int(1:0,8)),8)
Dados.Parcelas <- cbind(Dados256,Parcelas)
##
## Amostra de tamanho 150
set.seed(1500) # necessaria para obter sempre a mesma amostra
Dados150S <- Dados256[sample(Dados256$Ordem,150,rep=FALSE),]
## ------------------------------------------------------------------
## Amostra de tamanho 128 (estruturado)
##
Dados128CV <- Dados.Parcelas[Dados.Parcelas$Parcelas==1,]
##
## Extraindo o Conjunto complementar
Dados128AP <- Dados.Parcelas[Dados.Parcelas$Parcelas==0,]
## ------------------------------------------------------------------
## Amostra de tamanho 128 (aleatoria)
set.seed(1956) # necessaria para obter sempre a mesma amostra
Dados128S <- Dados256[sample(Dados256$Ordem,128,rep=FALSE),]
##
## Extraindo o conjunto complementar
Dados128S.comp <- Dados256
Dados128S.comp <- Dados128S.comp[-Dados128S$Ordem,]
##
## ------------------------------------------------------------
## Amostra de tamanho 64 (aleatoria)
set.seed(1956) # necessaria para obter sempre a mesma amostra
```

```
Dados64S <- Dados256[sample(Dados256$Ordem,64,rep=FALSE),]
##
## Extraindo o conjunto complementar
Dados192S.comp <- Dados256
Dados192S.comp <- Dados192S.comp[-Dados64S$Ordem,]
##
## ------------------------------------------------------
## Ilustrando os grids amostrais para SOJA
## ------------------------------------------------------
## Grid.256  <- 256 parcelas originais
## Grid.128  <- 128 parcelas intercaladas
## Grid.128S <- 128 parcelas sorteadas
## Grid.064  <-  64 parcelas sorteadas
## Grid.150  <- 150 parcelas sorteadas
##
## ++++++++++++++++++++++++++++++++++
##
## Para salvar o grafico no formato Postscript,
## descomente a linha seguinte e a ultima linha (dev.off)
##
## postscript('layout.ps',horizontal=FALSE, pointsize=12,  width=10,height
   =2)
## par(mfrow=c(1,3),mgp=c(2,0.8,0),mar=c(0,0.5,0,0.5))
## plot(Dados256[,1:2],xlab="",ylab="",axes=FALSE,pch=20)
## box()
## plot(Dados128CV[,1:2],xlab="",ylab="", axes=FALSE, pch=20)
## box()
## plot(Dados128S[,1:2],xlab="",ylab="", axes=FALSE, pch=20)
## box()
## plot(Dados64S[,1:2],xlab="",ylab="", axes=FALSE,pch=20)
## box()
## plot(Dados150S[,1:2],xlab="",ylab="", axes=FALSE, pch=20)
## box()
## dev.off()
## par(mfrow=c(1,1))
##
## Examinando a correlacao entre as variaveis
## ------------------------------------------------------
##
## mat.X <- matrix(c(Dados256$P,
##                   Dados256$PH,
##                   Dados256$K,
##                   Dados256$MO,
```

```
##                     Dados256$SB,
##                     Dados256$iCone,
##                     Dados256$PROD),
##                     byrow=F, ncol=7)
##
## nc <- ncol(mat.X)                                    # Nr colunas de
    X
## nr <- nrow(mat.X)                                    # Nr linhas de
    X
## muX <- round(sapply(1:nc, function(p) mean(mat.X[,p])),3) # Vetor de
    medias
## X1 <- sapply(1:nc, function(p) mat.X[,p]-mean(mat.X[,p]))
## S <- round((1/(nr-1))*t(X1)%*%X1,3)
## DP <- diag(1,nc)*(diag(S))^(-.5)                     # Matriz desvio padrao
## R <- round(DP%*%S%*%DP,2)                            # Matriz correlacao
## R
##
## Analise exploratoria da produtividade de soja
## ------------------------------------------------------------
summary(Dados256$PROD); sd(Dados256$PROD); 100*sd(Dados256$PROD)/mean(
    Dados256$PROD)
summary(Dados128S$PROD); sd(Dados128S$PROD); 100*sd(Dados128S$PROD)/mean(
    Dados128S$PROD)
summary(Dados64S$PROD); sd(Dados64S$PROD); 100*sd(Dados64S$PROD)/mean(
    Dados64S$PROD)
##
## Analise exploratoria das variaveis secundarias
## ------------------------------------------------------------
summary(Dados150S$P); sd(Dados150S$P); 100*sd(Dados150S$P)/mean(Dados150S$P)
summary(Dados150S$PH); sd(Dados150S$PH); 100*sd(Dados150S$PH)/mean(Dados150S$
    PH)
summary(Dados150S$K); sd(Dados150S$K); 100*sd(Dados150S$K)/mean(Dados150S$K)
summary(Dados150S$MO); sd(Dados150S$MO); 100*sd(Dados150S$MO)/mean(Dados150S$
    MO)
summary(Dados150S$SB); sd(Dados150S$SB); 100*sd(Dados150S$SB)/mean(Dados150S$
    SB)
summary(Dados150S$iCone); sd(Dados150S$iCone); 100*sd(Dados150S$iCone)/mean(
    Dados150S$iCone)
##
## A N A L I S E   G E O E S T A T I S T I C A
## ------------------------------------------------------------
## ATENCAO: O suporde das amostras de producao
## correspondem a uma area de 25 m2 (5x5) que sera
```

```
## o tamanho minimo do ponto a ser estimado pela krigagem
##
## Conversao dos objetos SOJA para o formato geodata
## ------------------------------------------------------------
Soja256.geo <- as.geodata(Dados256,coords.col=11:12, data.col=10)
## Soja128.geo <- as.geodata(Dados128CV,coords.col=11:12,data.col=10)
## Soja128.comp.geo <- as.geodata(Dados128AP,coords.col=11:12, data.col=10)
Soja128S.geo <- as.geodata(Dados128S,coords.col=11:12, data.col=10)
## Soja128S.comp.geo <- as.geodata(Dados128S.comp,coords.col=11:12, data.
    col=10)
Soja64S.geo <- as.geodata(Dados64S,coords.col=11:12, data.col=10)
## Soja192S.comp.geo <- as.geodata(Dados192S.comp,coords.col=11:12,data.col
    =10)
##
## Conversao dos objeto das variaveis secundarias
## para o formato geodata
## ------------------------------------------------------------
## P150S.geodata <- as.geodata(Dados150S, coords.col=1:2, data.col=3)
## PH150S.geodata <- as.geodata(Dados150S, coords.col=1:2, data.col=4)
## K150S.geodata <- as.geodata(Dados150S, coords.col=1:2, data.col=5)
## MO150S.geodata <- as.geodata(Dados150S, coords.col=1:2, data.col=6)
## SB150S.geodata <- as.geodata(Dados150S, coords.col=1:2, data.col=7)
## iCone150S.geodata <- as.geodata(Dados150S, coords.col=1:2, data.col=8)
##
## Criando contorno da area
## ------------------------------------------------------------
##
border <- cbind(east=c(0,0,152,152),north=c(0,116,116,0))
##
## Adicionando as informacoes de contorno nos objetos geodata
## ------------------------------------------------------------
Soja256.geo$borders <- border
## Soja128.geo$borders <- border
## Soja128.comp.geo$borders <- border
Soja128S.geo$borders <- border
## Soja128S.comp.geo$borders <- border
Soja64S.geo$borders <- border
## Soja192S.comp.geo$borders <- border
##
## iCone150S.geodata$borders <- border
##
## Analise exploratoria
## ------------------------------------------------------------
```

```
## summary(Soja256.geo)
## round(sd(Soja256.geo$data),4)
## round(sd(Soja256.geo$data)/mean(Soja256.geo$data)*100,2)
##
## summary(Soja128.geo)
## max(Soja128.geo$data)
## round(sd(Soja128.geo$data),4)
## round(sd(Soja128.geo$data)/mean(Soja128.geo$data)*100,2)
##
## summary(Soja128.comp.geo)
## round(sd(Soja128.comp.geo$data),4)
## round(sd(Soja128.comp.geo$data)/mean(Soja128.comp.geo$data)*100,2)
##
## summary(Soja128S.geo)
## round(sd(Soja128S.geo$data),4)
## round(sd(Soja128S.geo$data)/mean(Soja128S.geo$data)*100,2)
##
## mean(Soja128S.comp.geo$data)
## round(sd(Soja128S.comp.geo$data),4)
## round(sd(Soja128S.comp.geo$data)/mean(Soja128S.comp.geo$data)*100,2)
##
## mean(Soja64S.geo$data)
## round(sd(Soja64S.geo$data),4)
## round(sd(Soja64S.geo$data)/mean(Soja64S.geo$data)*100,2)
##
## mean(Soja192S.comp.geo$data)
## round(sd(Soja192S.comp.geo$data),4)
## round(sd(Soja192S.comp.geo$data)/mean(Soja192S.comp.geo$data)*100,2)
##
##
## Verificando a necessidade de transformacao
## ------------------------------------------------------
## lambda = 1    --> nao transformar
## lambda = 0.5  --> transformacao raiz quadrada
## lambda = 0.0  --> transformacao log
##
## require(MASS)
## postscript('boxcox.ps', horizontal=FALSE, pointsize=12, width=10, height
   =3)
## par(mfrow=c(1,3), mgp=c(2,0.8,0), mar=c(5,2,1,1))
## boxcox(Soja256.geo, lambda=seq(0.25,1.5,len=20))
## boxcox(Soja128.geo, lambda=seq(-0.5,1.5,len=20))
## boxcox(Soja128.comp.geo, lambda=seq(0,2.5,len=20))
```

```
## boxcox(Soja128S.geo,lambda=seq(-1,2,len=20))
## boxcox(Soja128S.comp.geo,lambda=seq(0,2,len=20))
## boxcox(Soja64S.geo,lambda=seq(-2,3,len=20))
## boxcox(Soja192S.comp.geo,lambda=seq(0,1.5,len=20))
## dev.off()
##
## boxcox(P150S.geodata, lambda=seq(-1.2,0,len=20))
## boxcox(PH150S.geodata, lambda=seq(-4,0,len=20))
## boxcox(K150S.geodata, lambda=seq(-0.8,1,len=20))
## boxcox(MO150S.geodata, lambda=seq(-0.4,3,len=20))
## boxcox(SB150S.geodata, lambda=seq(0.35,1.5,len=20))
## boxcox(iCone150S.geodata, lambda=seq(0,1.8,len=20))
##
## Spatial-plot (antigo post-plot)
## ------------------------------------------------------
Soja256.geo.class <- Soja256.geo
Soja256.geo.class$data <- sapply(Soja256.geo$data,Class)
round(prop.table(table(Soja256.geo.class$data))*100,1)
postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/SpatialPlot.
    ps',
           horizontal=FALSE,
           pointsize=10,
           width=6,height=6)
par(mfrow=c(1,1),
    mgp=c(0,0,0),
    mar=c(0,.1,0,0.1),
    bty='n',xaxt='n',yaxt='n')
points(Soja256.geo,
       pt.div="equal",
       axes=FALSE,
       ylab="",xlab="",
       cex.min=3.5,cex.max=3.5,
       col=gray(seq(1,0,l=5)),
       asp=1, pch=22)
legend(4, -3, c("< 2,34",
                "(2,34;2,61)",
                "(2,61;2,85)",
                "(2,85;3,16)",
                "> 3,16" ),
       horiz=T,
       fill=gray(seq(1,0,l=5)),
       cex=1.0)
 dev.off()
```

```
##
## ------------------------------------------------------------
## Plot Geodata
## ------------------------------------------------------------
## plot(Soja256.geo)
## plot(Soja128.geo)
## plot(Soja128.comp.geo)
## plot(Soja128S.geo)
## plot(Soja128S.comp.geo)
## plot(Soja64S.geo)
## plot(Soja192S.comp.geo)
##
## ------------------------------------------------------------
## Ajuste de modelos por maxima verossimilhanca
## ------------------------------------------------------------
##
## Global
Soja256.lik <- likfit(Soja256.geo,
                      ini=c(0.25,80),
                      cov.model="matern",
#                       fix.nugget=TRUE,
#                       nugget=0,
#                       method.lik='REML',
                      kappa=0.5)
##
Soja256.lik
##
## Amostra sistematica de 128 pontos
## Soja128.lik <- likfit(Soja128.geo,
##                       ini=c(0.25,80),
##                       cov.model="matern",
##                       kappa=0.5)
##
## Amostra sistematica de 128 pontos complementares
## Soja128.comp.lik <- likfit(Soja128.comp.geo,
##                            ini=c(0.25,80),
##                            cov.model="matern",
##                            kappa=0.5)
##
## Amostra aleatoria 128 pontos
Soja128S.lik <- likfit(Soja128S.geo,
                       ini=c(0.25,80),
                       cov.model="matern",
```

```
#                               fix.nugget=TRUE,
#                               nugget=0,
#                               method.lik='REML',
                              kappa=0.5)
##
Soja128S.lik
## ------------------------------------------------------------
## Amostra aleatoria 128 pontos complementares
## Soja128S.comp.lik <- likfit(Soja128S.comp.geo,
##                                 ini=c(0.25,80),
##                                 cov.model="matern",
##                                 kappa=0.5)
##
## ------------------------------------------------------------
## Amostra aleatoria 64 pontos
Soja64S.lik <- likfit(Soja64S.geo,
                          ini=c(0.25,80),
                          cov.model="matern",
#                          fix.nugget=TRUE,
#                          nugget=0,
#                          method.lik='REML',
                          kappa=0.5)
##
Soja64S.lik
## ------------------------------------------------------------
## Amostra aleatoria 64 pontos
## Soja192S.comp.lik <- likfit(Soja192S.comp.geo,
##                                 ini=c(0.25,80),
##                                 cov.model="matern",
##                                 kappa=0.5)
##
## ------------------------------------------------------------
##
##          VALIDACAO CRUZADA
##
## ------------------------------------------------------------
Soja256.valid <- xvalid(Soja256.geo,
                            model=Soja256.lik)
par.ori <- par(no.readonly=T)
postscript('/media/KINGSTON/TESE/FINAL/256Valid.ps',
            horizontal=FALSE,
            pointsize=22,
            width=24,height=45)
```

```
par(mfcol=c(5,2), mar=c(3,3,.2,.2), mgp=c(1.5, 0.6, 0))
plot(Soja256.valid)
dev.off()
par(par.ori)
##
## ------------------------------------------------------------
Soja128S.valid <- xvalid(Soja128S.geo,
                         model=Soja128S.lik,
                         locations.xvalid=Soja128S.comp.geo$coords,
                         data.xvalid=Soja128S.comp.geo$data)
## postscript('/media/KINGSTON/TESE/FINAL/128SValid.ps',
            horizontal=FALSE,
            pointsize=22,
            width=24,height=45)
par(mfcol=c(5,2), mar=c(3,3,.2,.2), mgp=c(1.5, 0.6, 0))
plot(Soja128S.valid)
## dev.off()
##
## ------------------------------------------------------------
Soja64S.valid <- xvalid(Soja64S.geo,
                          model=Soja64S.lik,
                         locations.xvalid=Soja192S.comp.geo$coords,
                         data.xvalid=Soja192S.comp.geo$data)
## postscript('/media/KINGSTON/TESE/FINAL/64SValid.ps',
            horizontal=FALSE,
            pointsize=22,
            width=24,height=45)
par(mfcol=c(5,2), mar=c(3,3,.2,.2), mgp=c(1.5, 0.6, 0))
plot(Soja64S.valid)
## dev.off()
##
## ------------------------------------------------------------
##
##          P R E D I C A O   E S P A C I A L
##       (K R I G A G E M   U N I V A R I A D A )
##
##        SUPORTE: Parcelas com 25 m2 (5 x 5 m)
##        Eixo horizontal (0 -- 150): 30 partes de 5 m
##        Eixo vertical   (0 -- 115): 23 partes de 5 m
##        Malha com 690 unidades de 5 x 5
## ------------------------------------------------------------
##
gr <- expand.grid(X=seq(5,150,by=5), Y=seq(5,115,by=5))
```

```
##
## ------------------------------------------------------------
KC256 <- krige.control(obj=Soja256.lik)
Soja256.KC <- krige.conv(Soja256.geo,
                         loc=gr,
                         borders=Soja256.geo$borders,
                         krige=KC256)
##
summary(Soja256.KC$predict, digits=5)
mean(Soja256.KC$predict)
round(sd(Soja256.KC$predict),4)
round(sd(Soja256.KC$predict)*100/mean(Soja256.KC$predict),1)
round((mean(Soja256.KC$predict)-2.7496)*100/2.7496,2)
##
## ------------------------------------------------------------
## KC128 <- krige.control(obj=Soja128.lik)
## Soja128.KC <- krige.conv(Soja128.geo, loc=gr, borders=Soja128.geo$
    borders, krige=KC128)
##
## mean(Soja128.KC$predict)
## sd(Soja128.KC$predict)
## sd(Soja128.KC$predict)*100/mean(Soja128.KC$predict)
## (mean(Soja128.KC$predict)-2.7496)*100/2.7496
##
## ------------------------------------------------------------
KC128S <- krige.control(obj=Soja128S.lik)
Soja128S.KC <- krige.conv(Soja128S.geo,
                          loc=gr,
                          borders=Soja128S.geo$borders,
                          krige=KC128S)
##
summary(Soja128S.KC$predict, digits=5)
mean(Soja128S.KC$predict)
round(sd(Soja128S.KC$predict),4)
round(sd(Soja128S.KC$predict)*100/mean(Soja128S.KC$predict),1)
round((mean(Soja128S.KC$predict)-2.7496)*100/2.7496,2)
##
## ------------------------------------------------------------
KC64S <- krige.control(obj=Soja64S.lik)
Soja64S.KC <- krige.conv(Soja64S.geo,
                         loc=gr,
                         borders=Soja64S.geo$borders,
                         krige=KC64S)
```

```
##
summary(Soja64S.KC$predict, digits=5)
mean(Soja64S.KC$predict)
round(sd(Soja64S.KC$predict),4)
round(sd(Soja64S.KC$predict)*100/mean(Soja64S.KC$predict),1)
round((mean(Soja64S.KC$predict)-2.7496)*100/2.7496,2)
##
## ------------------------------------------------------------
##          MAPA  DE  PREDICAO
##              UNIVARIADA (Classificado)
##
## ------------------------------------------------------------
##
## extraindo as predicoes
## ------------------------------------------------------------
## Global (256)
S256K.pred  <- Soja256.KC$predict # Verossimilhanca
##
## Intercalado 128
## S128K.pred <- Soja128.KC$predict # Verossimilhanca
##
## Amostrado 128
S128SK.pred <- Soja128S.KC$predict # Verossimilhanca
##
## Amostrado 64
S64SK.pred <- Soja64S.KC$predict # Verossimilhanca
## summary(S64SK.pred)
## round(prop.table(table(cut(S64SK.pred,c(0,2.4075,3.045,3.4)))),3)
##
## Aplicando a Funcao de classificacao
## ------------------------------------------------------------
## 0.0000 -- 2.4075 --> 2.4075 (BAIXA)
## 2.4076 -- 3.0450 --> 3.0460 (MEDIA)
## 3.0451 >         --> 4.0000 (ALTA)
##
## Global (256)
S256K.class <- sapply(S256K.pred,Class) # verossimilhanca
round(prop.table(table(S256K.class))*100,1)
##
## Intercalado 128
## S128K.class <- sapply(S128K.pred,Class) # verossimilhanca
##
## Amostrado 128
```

```
S128SK.class <- sapply(S128SK.pred,Class) # verossimilhanca
round(prop.table(table(S128SK.class))*100,1)
##
## Amostrado 64
S64SK.class <- sapply(S64SK.pred,Class) # verossimilhanca
round(prop.table(table(S64SK.class))*100,1)
##
## devolvendo as predicoes classificadas Global (4466) krigado
## ------------------------------------------------------------
##
## Global
Soja256.KC$predict <- S256K.class # verossimilhanca
##
## Amostrado 128
## Soja128.KC$predict <- S128K.class # verossimilhanca
##
## Amostrado 128
Soja128S.KC$predict <- S128SK.class # verossimilhanca
##
## Amostrado 64
Soja64S.KC$predict <- S64SK.class # verossimilhanca
##
## ------------------------------------------------------------
## MAPA: Verossimilhanca
## ------------------------------------------------------------
postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/
    MapaUnivarClass.ps',
            horizontal=FALSE,
            pointsize=12,
            width=10, height=2.5)
##
par(mfrow=c(1,3),
    mgp=c(2,0.8,0),
    mar=c(5,2,1,1))
##     pty='s',
##     bty='n',xaxt='n',yaxt='n'
##
image(seq(5,150,by=5),
      seq(5,115,by=5),
      t(matrix(Soja256.KC$predict, byrow=T, ncol=30)),
      col=gray(seq(1,0,l=5)),
      ann=FALSE)
##
```

```
title(main="256 amostras")
##
## ------------------------------------------------------------
## 128 amostras
image(seq(5,150,by=5),
      seq(5,115,by=5),
      t(matrix(Soja128S.KC$predict, byrow=T, ncol=30)),
      col=gray(seq(1,0,l=5)),
      ann=FALSE)
##
title(main="128 amostras")
##
## ------------------------------------------------------------
## 64 amostras
image(seq(5,150,by=5),
      seq(5,115,by=5),
      t(matrix(Soja64S.KC$predict, byrow=T, ncol=30)),
      col=gray(seq(1,0,l=5)),
      ann=FALSE)
title(main="64 amostras")
##
## legend(0, -5, c("< 2,34",
##                 "(2,34;2,61)",
##                 "(2,61;2,85)",
##                 "(2,85;3,16)",
##                 "> 3,16" ),
##        horiz=T,
##        fill=gray(seq(1,0,l=5)),
##        cex=1.0)
## ------------------------------------------------------------
dev.off()
##
##
##
## I N F E R E N C I A   B A Y E S I A N A
##
## ########################################################
##
## Definindo o modelo para Soja256
## - Kappa = 1,6
## - Funcao de correlacao de Matern
## - Sem tranformacao nos dados (lambda =1)
## - Prioris: phi <- discreta uniforme
```

```
##             beta <- flat
##             sigmasq <- reciproca
##             tausq.rel <- fixa em zero
##
## Grid de predicao -- 690 pontos
## ------------------------------------------------------------
GP <- gr
##
## Estimacao (krigagem bayesiana)
## ------------------------------------------------------------
Soja256.kb <- krige.bayes( Soja256.geo,loc=GP)
Soja256.kb$posterior$beta$summary
Soja256.kb$posterior$sigmasq$summary
Soja256.kb$posterior$phi$summary
##
Soja128S.kb <- krige.bayes(Soja128S.geo,loc=GP)
Soja128S.kb$posterior$beta$summary
Soja128S.kb$posterior$sigmasq$summary
Soja128S.kb$posterior$phi$summary
##
Soja64S.kb  <- krige.bayes( Soja64S.geo,loc=GP)
Soja64S.kb$posterior$beta$summary
Soja64S.kb$posterior$sigmasq$summary
Soja64S.kb$posterior$phi$summary
##
## grafico das distribuicoes a posteriori
## ------------------------------------------------------------
## jpeg(filename='Posterior256.jpg', width=600,height=250, pointsize=12,
    quality=150)
postscript('Posterior256.ps',horizontal=FALSE, width=10,height=4, pointsize
    =12)
par(mfrow=c(1,2),mar=c(5,4,1,0.5))
hist(Soja256.kb) # beta e sigmasq
## dev.off()
##
## postscript('Posterior128.ps', horizontal=FALSE, width=10,height=4,
    pointsize=12)
par(mfrow=c(1,2),mar=c(5,4,1,0.5))
hist(Soja128.kb) # beta e sigmasq
## dev.off()
##
## postscript('Posterior64.ps', horizontal=FALSE,width=10,height=4,
    pointsize=12)
```

```
par(mfrow=c(1,2),mar=c(5,4,1,0.5))
hist(Soja64.kb) # beta e sigmasq
## dev.off()
##
## Mapas por predicao bayesina (classificados)
## ------------------------------------------------------
Soja256.kb.class <- Soja256.kb
Soja256.kb.class$predictive$mean <- sapply(Soja256.kb.class$predictive$mean
    , Class)
##
round(prop.table(table(Soja256.kb.class$predictive$mean))*100,1)
##
Soja128S.kb.class <- Soja128S.kb
Soja128S.kb.class$predictive$mean <- sapply(Soja128S.kb.class$predictive$
    mean, Class)
##
round(prop.table(table(Soja128S.kb.class$predictive$mean))*100,1)
##
Soja64S.kb.class <- Soja64S.kb
Soja64S.kb.class$predictive$mean  <- sapply(Soja64S.kb.class$predictive$
    mean, Class)
##
round(prop.table(table(Soja64S.kb.class$predictive$mean))*100,1)
##
##
postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/
    MapaBayesianoClass.ps',
             horizontal=FALSE,
             pointsize=12,
             width=10,height=2.5)
##
par(mfrow=c(1,3),
    mgp=c(2,0.8,0),
    mar=c(5,2,1,1))
##
image(Soja256.kb.class,col=gray(seq(1,0,l=5)), xlab="", ylab="")
title(main="256 amostras")
image(Soja128S.kb.class,col=gray(seq(1,0,l=11)), xlab="", ylab="")
title(main="128 amostras")
image( Soja64S.kb.class,col=gray(seq(1,0,l=11)), xlab="", ylab="")
title(main="64 amostras")
dev.off()
##
```

```
## ------------------------------------------------------------------------
##
summary(Soja256.kb$predictive$mean, digits=5)
sd(Soja256.kb$predictive$mean)
100*sd(Soja256.kb$predictive$mean)/mean(Soja256.kb$predictive$mean)
(mean(Soja256.kb$predictive$mean) -2.7496)*100/2.7496
##
summary(Soja128S.kb$predictive$mean, digits=5)
sd(Soja128S.kb$predictive$mean)
100*sd(Soja128S.kb$predictive$mean)/mean(Soja128S.kb$predictive$mean)
(mean(Soja128S.kb$predictive$mean) -2.7496)*100/2.7496
##
summary(Soja64S.kb$predictive$mean, digits=5)
sd(Soja64S.kb$predictive$mean)
100*sd(Soja64S.kb$predictive$mean)/mean(Soja64S.kb$predictive$mean)
(mean(Soja64S.kb$predictive$mean) -2.7496)*100/2.7496
##
# Box Cox da estimativa de beta
XX <- data.frame(X=GP[,1], Y=GP[,2],ZZ= Soja256.kb$predictive$mean)
XX.geo <- as.geodata(XX, coords.col=1:2, data.col=3 )
boxcox(XX.geo, lambda=seq(0.5,2,len=20) )
##
## ------------------------------------------------------------------------
##  F I M
## ------------------------------------------------------------------------
```

Listagem B.2: Análise univariada de dados de IMA

```
## ######################################################
##
## Analise Geoestatistica Univariada - Pinus taeda L.
##
## Titulo: Aplicacao de metodos geoestatisticos multivariados
##         em problemas de mapeamento de variaveis do sistema
##         solo-planta.
##
## Dados: Fazenda Mobasa
## Area de reflorestamento de P. Taeda L.
##
## Edson Antonio Alves da Silva, Prof. MSc.
##
## Orientador: Paulo Justiniano Ribeiro Jr., Prof. PhD.
##
##   Criado em: 21 de Marco de 2008
## Revisado em: 05 de Julho de 2008
##
## ######################################################
##
## IMA + Argila (Azul) --> 18 pontos
## ARGILA (Vermelho)   --> 555 pontos
##
## ------------------------------------------------------
## Limpando a workspace
##
oldpar <- par(no.readonly=TRUE)
rm(list=ls(all=TRUE))
objects()
##
## ------------------------------------------------------
## FUNCAO DE CLASSIFICACAO
## Classifica em 5 grupos separados por percentis
## Quantiles: 0,2 -- 0,4 -- 0,6 -- 0,8
##
Class <- function(x)
   {
   if (x<20.504) x=17.3 # < 20%
      else
      {if (x<23.432) x=22.0 # < 40%
         else
           {if (x<28.332) x=25.9 # < 60%
```

```
                else
                  {if (x<32.504) x=30.4 else 34.8}
45              }

        }
    }
## ------------------------------------------------------------
##  Leitura do arquivo de dados
##
IMA18 <- read.table("/home/edson/DOCUMENTOS/DADOS/MOBASA/ima_arg_azul.csv",
                    head=T,
                    sep="",
55                  dec=",",
                    col.names=c("X","Y","IMA","ARG1","Arg2"))
##
## IMA18 <- read.table("http://wiki.leg.ufpr.br/data/media/projetos/mobasa/
    artigo/ima_arg_azul.csv", head=T, sep="", dec=",",col.names=c("X","Y","
    IMA","ARG1","Arg2"))
##
60 Arg555 <- read.table("/home/edson/DOCUMENTOS/DADOS/MOBASA/arg_vermelho.csv"
    , head=T, dec=",", sep="")
##
## Arg555 <- read.table("http://wiki.leg.ufpr.br/data/media/projetos/mobasa
    /artigo/arg_vermelho.csv", head=T, dec=",", sep="")
##
## ------------------------------------------------------------
65 ## Lendo o arquivo de bordas
##
borda <- read.table("/home/edson/DOCUMENTOS/DADOS/MOBASA/contorno.txt",
    head=T, dec=".", sep="")
##
## borda <- read.table("http://wiki.leg.ufpr.br/data/media/projetos/mobasa/
    artigo/contorno.txt", head=T, dec=".", sep="")
70 ## ------------------------------------------------------------
##
## Analise Descritiva
##
## ------------------------------------------------------------
75 summary(IMA18$IMA)
sd(IMA18$IMA)
sd(IMA18$IMA)*100/mean(IMA18$IMA)
##
summary(IMA18$ARG1)
```

```
sd(IMA18$ARG1)
sd(IMA18$ARG1)*100/mean(IMA18$ARG1)
##
summary(Arg555$Arg)
sd(Arg555$Arg)
sd(Arg555$Arg)*100/mean(Arg555$Arg)
##
quantile(IMA18$IMA)
##
## ---------------------------------------------------------
##
##                    Layout
##
## ---------------------------------------------------------
## Para salvar o grafico no formato postscript (ps ou eps),
## descomente a linha seguinte e a ultima linha (dev.off)
##
## postscript('LayoutMobasa.ps',horizontal=FALSE)
par(mfrow=c(1,2),mgp=c(2,0.8,0),mar=c(0,0.5,0,0.5))
##
plot(borda, xlab="", ylab="", axes=FALSE, type='l')
points(IMA18[,1:2], xlab="", ylab="", pch=20)
##
plot(borda, xlab="", ylab="", axes=FALSE, type='l')
points(Arg555[,1:2], xlab="", ylab="", pch=20)
## dev.off()
##
## ---------------------------------------------------------
##  Gerando os arquivos geodata
require(geoR)
IMA18.geo <- as.geodata(IMA18,coords.col=1:2, data.col=3,covar.col=4)
Arg555.geo <- as.geodata(Arg555,coords.col=1:2, data.col=3)
##
names(IMA18.geo)
##
## Examinando os dados
plot(Arg555.geo,borders=borda )
plot(IMA18.geo,borders=borda)
##
## ---------------------------------------------------------
##  B O X   C O X
## ---------------------------------------------------------
## Verificando a necessidade de transformacao
```

```
require(MASS)
##
## postscript('boxcoxIMA.eps',
##                  horizontal=FALSE,
##                  width=4, height=3.2,
##                  pointsize=10)
par( mgp=c(2,0.8,0), mar=c(5,2,1,1))
boxcox(IMA18.geo,lambda=seq(-1,3,len=20))
## dev.off()
##
## ------------------------------------------------------------
## MAXIMA  VEROSSIMILHANCA
## ------------------------------------------------------------
## Ajustando um modelo exponencial por maxima verossimilhanca
## Usando a argila como co-variavel
IMA18.lik <- likfit(IMA18.geo, fix.nugget = TRUE, nugget = 0, ini=c
    (50,2007),cov.model="matern", kappa=0.5)
## IMAeARG.lik <- likfit(IMA18.geo, fix.nugget = TRUE, nugget = 0, ini=c
    (50,2007),cov.model="matern", kappa=0.5, trend=~ARG1)
## IMAeARG.lik
##
## ------------------------------------------------------------
##  KRIGAGEM
## ------------------------------------------------------------
## Construindo um grid de predicao
gr <- pred_grid(borda,by=120)
##
## Estabelecendo os controles de krigagem
KC.IMA18 <- krige.control(obj.model=IMA18.lik)
## KC.IMA18 <- krige.control(obj.model=IMAeARG.lik)
##
## Krigando .........
IMA.k <- krige.conv(IMA18.geo,loc=gr,krige=KC.IMA18)
##
## estatisticas descritivas da predicao
summary(IMA.k$predict)
sd(IMA.k$predict)
sd(IMA.k$predict)*100/mean(IMA.k$predict)
mean(IMA.k$krige.var) # variancia da krigagem
##
## ------------------------------------------------------------
## Estimacao  bayesiana do IMA
## ------------------------------------------------------------
```

```
##
## Definindo o modelo para IMA
## - Kappa = 0,5
## - Funcao de correlacao de Matern
## - Sem tranformacao nos dados (lambda =1)
## - Prioris: phi <- discreta uniforme
##            beta <- flat
##            sigmasq <- reciproca
##            tausq.rel <- fixa em zero
## GP <- pred_grid(borda,by=120) # (5.046 pontos)
##
## ########################
##     A T E N C A O
## DEMORA MAIS DE DEZ HORAS
## ########################
## IMA18.kb <- krige.bayes(IMA18.geo,loc=GP)
IMA18.kb
## Preservando o resultado para uso futuro
## save(IMA18.kb,file='IMABayes.Rdata')
objects()
load('IMABayes.Rdata')
##
## ------------------------------------------------------------
## estatisticas descritivas da predicao bayesiana
## ------------------------------------------------------------
names(IMA18.kb)
names(IMA18.kb$predictive)
summary(IMA18.kb$predictive$mean)
(SD <- sqrt(mean(IMA18.kb$predictive$mean)))
SD*100/mean(IMA18.kb$predictive$mean)
##
## media da variancia da predicao bayesiana
## diferente da variancia da media de predicao
mean(IMA18.kb$predictive$variance)
##
## ------------------------------------------------------------
## grafico das distribuicoes a posteriori
## ------------------------------------------------------------
## X11(height=3,width=9)
## postscript('PosteriorIMA18.ps',horizontal=FALSE, width=6,height=2.4,
    pointsize=12)
par(mfrow=c(1,3),mar=c(5,4,1,0.5))
hist(IMA18.kb) # beta, sigmasq e phi
```

```
## dev.off()
##
## ------------------------------------------------------------
## M A P A S
## ------------------------------------------------------------
##
##  Mapa da krigagem e estimacao bayesiana
##  univariada e classificada - Pinus Taeda L.
## ------------------------------------------------------------
## classificando os dados estimados pela krigagem
## ------------------------------------------------------------
## Classificando a predicao por krigagem
IMA.k.class <- IMA.k
IMA.k.class$predict  <- sapply(IMA.k.class$predict, Class)
round(prop.table(table(IMA.k.class$predict))*100,1)
##
## Classificando a predicao bayesiana
IMA18.kb.class <- IMA18.kb # predicao original
IMA18.kb.class$predictive$mean <- sapply(IMA18.kb$predictive$mean, Class)
round(prop.table(table(IMA18.kb.class$predictive$mean))*100,1)
##
## ..... fazendo o mapa
## ------------------------------------------------------------
postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/
   MapaKigeBayesUniIMAClass.ps',
               horizontal=FALSE,
               width=9,height=9,
               pointsize=12)
X11(height=9,width=9)
par(mfrow=c(1,2),mgp=c(0,0,0),mar=c(0,0,0,0), bty='n',xaxt='n',yaxt='n')
par(mfrow=c(1,2))
##
## Gerando a imagem com krigagem
## ------------------------------------------------------------
image(IMA.k.class,loc=gr,col=gray(seq(1,0,l=11)),
      xlab="", ylab="", axes=F, borders=borda)
##
## Gerando a imagem bayesiana
## ------------------------------------------------------------
image(IMA18.kb.class,border=borda,col=gray(seq(1,0,l=11)),
            xlab="", ylab="", axes=F)
dev.off()
##
```

```
## ------------------------------------------------------------------
jpeg('/home/edson/DOCUMENTOS/DEFESA/IMA18.jpg',
            width = 600,height=600,
            units = "px",
            quality = 150)
par(mfrow=c(1,1),mgp=c(0,0,0),mar=c(0,0,0,0), bty='n',xaxt='n',yaxt='n')
image(IMA.k.class,loc=gr,col=gray(seq(1,0,l=11)),
      xlab="", ylab="", axes=F, borders=borda)
dev.off()
##
## ------------------------------------------------------------------
##      F I M
## ------------------------------------------------------------------
```

Listagem B.3: Análise bivariada de dados de soja

```
## ######################################################
##
## Analise Geoestatistica Multivariada
##
## Dados: Soja 98
##
## Aluno: Edson Antonio Alves da Silva, Prof. MSc.
## Orientador: Paulo Justiniano Ribeiro Jr., Prof. PhD.
##
## Criado em    : 23 de Marco de 2008
## Revisado em : 05 de Julho de 2008
##
## ######################################################
##
##   Modelo:(Y1,Y2) onde
##       Y1: Produtividade
##       Y2: Indice de Cone (Resist. a Penetracao)
##
## ------------------------------------------------------
rm(list=ls(all=TRUE))
objects()
oldpar <- par(no.readonly=TRUE)
## ------------------------------------------------------
##
## Funcao de classificacao dos dados de soja
## ------------------------------------------------------
## Classifica em 5 grupos separados por percentis
## Quantiles: 0,2 -- 0,4 -- 0,6 -- 0,8
##
Class <- function(x)
   {
   if (x<2.341) x=1.8 # < 20%
      else
      {if (x<2.611) x=2.5 # < 40%
         else
           {if (x<2.851) x=2.7 # < 60%
              else
                {if (x<3.161) x=3.0 else 3.7}
           }

      }
   }
```

```
## ------------------------------------------------------------
##
Dados256 <- read.csv("http://wiki.leg.ufpr.br/data/media/pessoais/edson/
    tese/dados/coodetec/soja98d.csv", head=T, sep="", dec=",")
##
## numerando a sequencia de coordenadas
Dados256 <- cbind(Dados256,Ordem=seq(1:256))
Dados256 <- cbind(Dados256,
                  CP1=round((0.019*Dados256$P-0.034*Dados256$PH+
                  0.326*Dados256$MO-0.945*Dados256$SB +68)/100,3))
##
## Separando as amostras
## ------------------------------------------------------------
## Amostra de tamanho 150 (aleatoria)
set.seed(1500) # necessaria para obter sempre a mesma amostra
Dados150S <- Dados256[sample(Dados256$Ordem,150,rep=FALSE),]
##
## Amostra de tamanho 128 (aleatoria)
set.seed(1956) # necessaria para obter sempre a mesma amostra
Dados128S <- Dados256[sample(Dados256$Ordem,128,rep=FALSE),]
##
## Amostra de tamanho 64  (aleatoria)
set.seed(1956) # necessaria para obter sempre a mesma amostra
Dados64S <- Dados256[sample(Dados256$Ordem,64,rep=FALSE),]
##
##       Analise de componentes principais
## ------------------------------------------------------------
MatX <- matrix(
        c(Dados150$P,Dados150$PH,Dados150$K,Dados150$MO,Dados150$SB),
        ncol=5)
head(MatX)

nc <- ncol(MatX)

nr <- nrow(MatX)
##
## vetor de medias
(muX <- round(sapply(1:nc, function (p) mean(MatX[,p])),3))

(mean(MatX[,3]))
## matriz de observacao excluida a de media
MatX.c <- sapply(1:nc, function (p) MatX[,p]-mean(MatX[,p]))
MatX.c
```

```
85
##
## matriz de covariancias
MatS <- round((1/(nr-1))*t(MatX.c)%*%MatX.c,4)
##
90 ## autovalor e autovetor
EIG <- eigen(MatS)
EIG.l <- round(EIG$values,3) # autovalores
EIG.v <- round(EIG$vectors,3)# autovetores
##
95 ## porcentagem de explicacao de cada componente
VCP <- round(sapply(1:nc, function (p) EIG$values[p]/sum(EIG$values))*
    100,2)
VCP
##
## Determinacao com variaveis padronizandas
100 ## ------------------------------------------------------------
DP <- diag(1,nc)*diag(MatS)^(-0.5)
##
## matriz de correlacao
MatR <- round(DP%*%DP,2)
105 ##
## autovalor e autovetor
EIGr <- eigen(MatR)
EIGr.l <- round(EIGr$values,3)
EIGr.v <- round(EIGr$vectors,3)
110 ##
## porcentagem de explicacao de cada componente (padronizado)
VCPr <- round(sapply(1:nc, function (p) EIGr$values[p]/sum(EIGr$values)),4)
VCPr
##
115 ## ######################################################
##
##      A N A L I S E   G E O E S T A T I S T I C A
##
## Conversao dos objetos para o formato geodata
120 ## ------------------------------------------------------------
require(geoR)
##
## SOJA
Soja256.geo <- as.geodata(Dados256,coords.col=11:12, data.col=10)
125 Soja128S.geo <- as.geodata(Dados128S,coords.col=11:12, data.col=10)
Soja64S.geo <- as.geodata(Dados64S,coords.col=11:12, data.col=10)
```

```
##
## iCone
iCone150S.geo <- as.geodata(Dados150S, coords.col=11:12, data.col=8)
##
## CP1 (Feito diretamente na tabela de dados)
CP1s150.geo <- as.geodata(Dados150S, coords.col=11:12, data.col=14)
##
## Criando contorno da area
## ------------------------------------------------------------
border <- cbind(east=c(0,0,152,152),north=c(0,116,116,0))
##
## Adicionando as informacoes de contorno nos objetos geodata
##
Soja256.geo$border <- border
Soja128S.geo$borders <- border
Soja64S.geo$borders <- border
##
iCone150S.geo$borders <- border
##
CP1s150.geo$borders <- border
##
## Ajuste de Modelo Bivariado por MV
## ------------------------------------------------------------
## Modelo 1 -- Y1: Soja 128 (aleatoria), Y2: iCone150
##
S128S_IC <- likfitBGCCM(Soja128S.geo,iCone150S.geo)
save(S128S_IC,file='S128IC.Rdata')
S128S_IC
##
## Modelo 2 -- Y1: Soja 64 (aleatoria), Y2: iCone150
S64S_IC <- likfitBGCCM(Soja64S.geo,iCone150S.geo)
save(S64S_IC,file='S64IC.Rdata')
S64S_IC
##
## Modelo 3 -- Y1: Soja 128 (aleatoria), Y2: CP1
S128S_CP1 <- likfitBGCCM(Soja128S.geo,CP1s150.geo)
save(S128S_CP1,file='S128CP.Rdata')
S128S_CP1
##
## Modelo 4 -- Y1: Soja 64 (aleatoria), Y2: CP1
S64S_CP1 <- likfitBGCCM(Soja64S.geo,CP1s150.geo)
save(S64S_CP1,file='S64CP.Rdata')
S64S_CP1
```

```
##
## ######################################################
##
## Predicoes de Y1 condicionadas a Y2
##
## grid de predicao (4.466 pontos)
gr <- expand.grid(X=seq(5,150,by=5), Y=seq(5,115,by=5))
##
## Modelo 1 -- Y1: Soja 128, Y2: iCone 150
S128S_IC.pred <- predict(S128S_IC, loc=gr)
summary(S128S_IC.pred$predict, digits=5)
sd(S128S_IC.pred$predict)
sd(S128S_IC.pred$predict)*100/mean(S128S_IC.pred$predict)
(mean(S128S_IC.pred$predict)-2.7496)*100/2.7496
##
## Modelo 2 -- Y1: Soja 64, Y2: iCone 150
S64S_IC.pred <- predict(S64S_IC, loc=gr)
summary(S64S_IC.pred$predict, digits=5)
sd(S64S_IC.pred$predict)
sd(S64S_IC.pred$predict)*100/mean(S64S_IC.pred$predict)
(mean(S64S_IC.pred$predict)-2.7496)*100/2.7496
##
## Modelo 3 -- Y1: Soja 128, Y2: CP1
S128S_CP1.pred <- predict(S128S_CP1, loc=gr)
summary(S128S_CP1.pred$predict, digits=5)
sd(S128S_CP1.pred$predict)
sd(S128S_CP1.pred$predict)*100/mean(S128S_CP1.pred$predict)
(mean(S128S_CP1.pred$predict)-2.7496)*100/2.7496
##
## Modelo 4 -- Y1: Soja 64, Y2: CP1
S64S_CP1.pred <- predict(S64S_CP1, loc=gr)
summary(S64S_CP1.pred$predict, digits=5)
sd(S64S_CP1.pred$predict)
sd(S64S_CP1.pred$predict)*100/mean(S64S_CP1.pred$predict)
(mean(S64S_CP1.pred$predict)-2.7496)*100/2.7496
##
## ------------------------------------------------------
##
## Classificando as predicoes em classes
S128S_IC.pred$predict <- sapply(S128S_IC.pred$predict, Class)
round(prop.table(table(S128S_IC.pred$predic))*100,1)
##
S64S_IC.pred$predict <- sapply(S64S_IC.pred$predict, Class)
```

```
round(prop.table(table(S64S_IC.pred$predic))*100,1)
##
S128S_CP1.pred$predict <- sapply(S128S_CP1.pred$predict,Class)
round(prop.table(table(S128S_CP1.pred$predic))*100,1)
##
S64S_CP1.pred$predict <- sapply(S64S_CP1.pred$predict,Class)
round(prop.table(table(S64S_CP1.pred$predic))*100,1)
##
##
## ######################################################
##
##        MAPA  DE  PREDICAO
##
##    Predicoes classificadas por intervalo de classe
##
## ######################################################
##
postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/MapaSojaBiIC.
    ps',
           horizontal=FALSE,
           pointsize=10,
           width=10, height=3.8)
##
par(mfrow=c(1,2),
    mgp=c(2,0.8,0),
    mar=c(5,2,1,1))
##
## ------------------------------------------------------------
## Soja e ICone
## ------------------------------------------------------------
## 128 amostras
image(S128S_IC.pred,
      loc=gr,
      col=gray(seq(1,0,l=5)),
      ann=FALSE,
      ylab="", xlab="")
##
title(main="128 amostras")
## ------------------------------------------------------------
## 64 amostras
image(S64S_IC.pred,
      loc=gr,
      col=gray(seq(1,0,l=5)),
```

```
255         ann=FALSE,
            ylab="", xlab="")
    ##
    title(main="64 amostras")
    ##
260 dev.off()
    ##
    ## ------------------------------------------------------------
    ##   Soja e CP1
    ## ------------------------------------------------------------
265 postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/MapaSojaBiCP.
        ps',
                horizontal=FALSE,
                pointsize=10,
                width=10, height=3.8)
    ##
270 par(mfrow=c(1,2),
        mgp=c(2,0.8,0),
        mar=c(5,2,1,1))
    ##
    ## ------------------------------------------------------------
275 ## 128
    image(S128S_CP1.pred,
            loc=gr,
            col=gray(seq(1,0,l=5)),
            ann=FALSE,
280         ylab="", xlab="")
    ##
    title(main="128 amostras")
    ##
    ## ------------------------------------------------------------
285 ## 64
    image(S64S_CP1.pred,
            loc=gr,
            col=gray(seq(1,0,l=5)),
            ann=FALSE,
290         ylab="", xlab="")
    ##
    title(main="64 amostras")
    ##
    dev.off()
295 ##
    ## ------------------------------------------------------------
```

```
##   F  I  M
## ————————————————————————————————————————————————————————————————————————————————
```

Listagem B.4: Análise bivariada de dados de IMA

```
## ######################################################
##
## Analise Geoestatistica Multivariada
##
## Pinus Taeda L.
## Dados: Mobasa
##
## Aluno: Edson Antonio Alves da Silva, Prof. MSc.
## Orientador: Paulo Justiniano Ribeiro Jr., Prof. PhD.
##
##   Criado em  : 23 Marco de 2008
## Revisado em  : 06 Julho de 2008
##
##   Modelo:(Y1,Y2) onde
##        Y1: IMA
##        Y2: Teor de Argila
##
## ############################################################
##
rm(list=ls(all=TRUE))
objects()
## ------------------------------------------------------------
## FUNCAO DE CLASSIFICACAO
## Classifica em 5 grupos separados por percentis
## Quantiles: 0,2 -- 0,4 -- 0,6 -- 0,8
##
Class <- function(x)
   {
   if (x<20.504) x=17.3 # < 20%
      else
      {if (x<23.432) x=22.0 # < 40%
         else
           {if (x<28.332) x=25.9 # < 60%
              else
                {if (x<32.504) x=30.4 else 34.8}
           }

      }
   }
##
## ------------------------------------------------------------
##  Entrada de dados
```

```
##
IMA18 <- read.table("http://wiki.leg.ufpr.br/data/media/projetos/mobasa/
    artigo/ima_arg_azul.csv", head=T, sep="", dec=",",col.names=c("X","Y","
    IMA","ARG1","Arg2"))
##
Arg555 <- read.table("http://wiki.leg.ufpr.br/data/media/projetos/mobasa/
    artigo/arg_vermelho.csv", head=T, dec=",", sep="")
##
borda <- read.table("http://wiki.leg.ufpr.br/data/media/projetos/mobasa/
    artigo/contorno.txt", head=T, dec=".", sep="")
##
## ###########################################################
##
##      A N A L I S E   G E O E S T A T I S T I C A
##
## ###########################################################
##
## Conversao dos objetos para o formato geodata
## ------------------------------------------------------------
require(geoR)
IMA18.geo <- as.geodata(IMA18,coords.col=1:2, data.col=3)
IMA18.geo$border <- borda
Arg555.geo <- as.geodata(Arg555,coords.col=1:2, data.col=3)
Arg555.geo$border <- borda
##
## Ajuste de Modelo Bivariado por MV
## Modelo 1 -- Y1: IMA 18,  Y2: Argila 555
## ------------------------------------------------------------
##
IMA18Arg <- likfitBGCCM(IMA18.geo,Arg555.geo)
## save(IMA18Arg,file='IMAArgMulti.Rdata')
## load('IMAArgMulti.Rdata')
##
## Predicoes de Y1 condicionadas a Y2
## ------------------------------------------------------------
##
## grid de predicao ( 5.046 pontos)
gr <- pred_grid(borda,by=120)
## ------------------------------------------------------------
## Modelo 1 -- Y1: IMA (18), Y2: Argila (555)
IMA18Arg.pred <- predict(IMA18Arg, loc=gr)
summary(IMA18Arg.pred$predict)
sd(IMA18Arg.pred$predict)
```

```
sd(IMA18Arg.pred$predict)*100/mean(IMA18Arg.pred$predict)
(mean(S128_IC.pred$predict)-2.7496)*100/2.7496
##
## Classificando as predicoes em classes
## ------------------------------------------------------------
IMA18Arg.pred$predict <- sapply(IMA18Arg.pred$predict, Class)
round(prop.table(table(IMA18Arg.pred$predict))*100,1)
##
## fazendo o mapa
## ------------------------------------------------------------
##
postscript('/home/edson/DOCUMENTOS/TESE/PosDEFESA/VersaoFinal/IMA18ArgClass
    .ps',
                horizontal=FALSE,
                width=9,height=9,
                pointsize=12)
par(mfrow=c(1,1),
    mgp=c(0,0,0),
    mar=c(0,0,0,0),
    bty='n',xaxt='n',yaxt='n')
image(IMA18Arg.pred,
      loc=gr,
      col=gray(seq(1,0,l=5)),
      xlab="", ylab="", axes=F,
      borders=borda)
dev.off()
##
## ------------------------------------------------------------
## F I M
## ------------------------------------------------------------
```