



**Universidade Federal de Pernambuco
Centro de Ciências Exatas e da Natureza
Programa de Pós-Graduação em Estatística**

EDUARDO LUCAS ENSSLIN DOS SANTOS

**INFLUÊNCIA LOCAL EM MODELOS ESPACIAIS LINEARES GAUSSIANOS,
UTILIZANDO O MÉTODO DE MÁXIMA VEROSSIMILHANÇA RESTRITA**

Recife

2019

EDUARDO LUCAS ENSSLIN DOS SANTOS

**INFLUÊNCIA LOCAL EM MODELOS ESPACIAIS LINEARES GAUSSIANOS,
UTILIZANDO O MÉTODO DE MÁXIMA VEROSSIMILHANÇA RESTRITA**

Dissertação apresentada ao Programa de Pós-Graduação em Estatística do Centro de Ciências Exatas e da Natureza da Universidade Federal de Pernambuco, como requisito parcial à obtenção do título de mestre em Estatística.

Área de Concentração: Estatística Aplicada

Orientadora: Profa. Dra. Fernanda De Bastiani

Coorientador: Prof. Dr. Miguel Angel Uribe-Opazo

Recife

2019

Catálogo na fonte
Bibliotecária Mariana de Souza Alves CRB4-2106

S237i Santos, Eduardo Lucas Ensslin dos
Influência local em modelos espaciais lineares gaussianos,
utilizando o método de máxima verossimilhança restrita – 2019.
83f.: il., fig., tab.

Orientadora: Fernanda de Bastiani
Dissertação (Mestrado) – Universidade Federal de
Pernambuco. CCEN, Estatística. Recife, 2019.
Inclui referências e apêndices.

1. Estatística Aplicada. 2. Diagnóstico de influência. 3.
Geoestatística. 4. Krigagem. I. Bastiani, Fernanda de
(orientadora). II. Título.

310

CDD (22. ed.)

UFPE-MEI 2019-140

EDUARDO LUCAS ENSSLIN DOS SANTOS

**INFLUÊNCIA LOCAL EM MODELOS ESPACIAIS LINEARES GAUSSIANOS
UTILIZANDO O MÉTODO DE MÁXIMA VEROSSIMILHANÇA RESTRITA**

Dissertação apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Estatística.

Aprovada em: 29 de julho de 2019.

BANCA EXAMINADORA

Prof.^(a) FERNANDA DE BASTIANI
UFPE

Prof.^(a) Audrey Helen Mariz de Aquino Cysneiros
UFPE

Prof.^(o) Eufrásio de Andrade Lima Neto
UFPB

À minha família:

Sergio, Lisete, Sérgio, Malu e Anabeth.

AGRADECIMENTOS

À Anabeth, pela cumplicidade, carinho e amor. Por fazer de mim a minha melhor versão e por estar ao meu lado em todos os momentos, me apoiando e encorajando sempre. Por fazer meus dias mais completos. A jornada é melhor com você do meu lado.

Aos meus pais, pela confiança que depositaram em mim desde muito cedo, por acreditarem que sou capaz. Pelos valores passados a mim, por todo suporte e apoio incondicional.

Aos amigos e colegas do PPGE, pelas trocas contínuas de conhecimento e cultura. Em particular, aos que me ajudaram e apoiaram nessa jornada. Também aos amigos da graduação e do tempo de colégio, pela forte amizade de longa data. Em especial, ao irmão que Guiné-Bissau me deu, por me mostrar outra visão do mundo, com certeza graças a ti hoje sou uma pessoa melhor. À Ascension, por todas as amizades e momentos de descontração.

Agradeço aos meus mentores. À minha orientadora Prof. Fernanda De Bastiani, pelos ensinamentos, paciência e por me mostrar que posso ir além dos meus limites. Ao meu coorientador, Prof. Miguel Uribe-Opazo, pela parceria e as suas valiosas contribuições. Agradeço também a todos os professores por tonar a aprendizagem uma experiência ímpar.

Ao CNPq, pelo apoio financeiro.

RESUMO

O conhecimento da Geoestatística fundamenta-se no estudo de variáveis aleatórias indexadas pela sua localização geográfica. A análise de métodos de diagnóstico de influência é frequentemente utilizada para indentificar a presença de observações potencialmente influentes nos dados e/ou no modelo, e avaliar a distorção que estas observações podem causar nos resultados das análises estatísticas. Nesse contexto, considera-se o estudo em modelos espaciais lineares Gaussianos, que levam em conta a dependência espacial das variáveis em estudo. Nesse caso a modelagem da estrutura de dependência espacial faz-se necessária para a definição dos parâmetros que a descrevem. O modelo também é considerado nas técnicas de interpolação, como a krigagem, que é utilizada para a predição de valores em locais não amostrados. No entanto, a estimação dos parâmetros podem ser distorcidas pela existência de observações influentes. Nesse contexto, a principal contribuição desta dissertação é propor um esquema de perturbação na variável resposta para investigar a influência local em modelos espaciais lineares Gaussianos, considerando o método de máxima verossimilhança restrita para a estimação dos parâmetros. Realizaram-se estudos de simulação computacional e aplicação a dois conjuntos de dados reais.

Palavras-chave: Diagnóstico de influência. Geoestatística. Krigagem. Variabilidade espacial.

ABSTRACT

The knowledge of Geostatistics is based on the study of random variables indexed by their geographic location. The analysis of influence diagnosis methods is often used to identify the presence of potentially influential observations in the data and / or the model and to assess the distortion that these observations may cause in the results of the statistical analyzes. In this context, the study is considered in Gaussian spatial linear models, which take into account the spatial variability of the variables under study. In this case the modeling of the spatial dependence structure is necessary for the definition of the parameters that describe it. The model is also considered in interpolation techniques, such as kriging, which is used to predict values in non-sampled locations. However, the estimates of the parameters can be distorted by the existence of influential observations. In this context, the main purpose of this dissertation is to propose a perturbation scheme in the response variable to investigate the local influence in linear Gaussian models, considering the method of restricted maximum likelihood for the estimation of parameters. Simulation and application studies were performed on two real data sets.

Keywords: Geostatistic. Influence diagnostic. Kriging. Spatial variability.

LISTA DE FIGURAS

Figura 1 – Dados simulados apresentando os comportamentos de isotropia e anisotropia.	19
Figura 2 – Relação entre a semivariância $\gamma(\delta)$ e a covariância espacial $C(\delta)$ (DE BAS- TIANI, 2012).	23
Figura 3 – Gráfico da função de semivariância para o modelo esférico.	24
Figura 4 – Gráfico da função de semivariância para o exponencial.	25
Figura 5 – Gráfico da função de semivariância para o modelo Gaussiano.	26
Figura 6 – Gráfico da função de semivariância para o modelo da família Matérn para diferentes valores de κ e ϕ_3	27
Figura 7 – Gráfico de Influência (VERBEKE; MOLENBERGHS, 2009).	38
Figura 8 – Malha amostral utilizada para realização das simulações.	43
Figura 9 – Malha amostral 10×10 utilizada para realização das simulações.	44
Figura 10 – Malha amostral dos dados de Jones.	47
Figura 11 – Gráfico <i>boxplot</i> dos dados de Jones.	49
Figura 12 – Mapas de krigagem para o modelo Gaussiano utilizando os métodos da ML e REML para os dados de Jones.	51
Figura 13 – Gráficos de influência para o modelo Gaussiano utilizando o método da REML para os dados de Jones.	52
Figura 14 – Mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de Jones.	54
Figura 15 – Demais mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de Jones.	55
Figura 16 – Malha amostral dos dados de produtividade de soja.	57
Figura 17 – Box-plot dos dados de produtividade de soja.	58
Figura 18 – Post plot dos dados de produtividade de soja.	59
Figura 19 – Semivariogramas direcionais dos dados de produtividade de soja.	59
Figura 20 – Mapas de krigagem para o modelo Gaussiano utilizando o método da REML e ML para os dados de produtividade de soja.	63
Figura 21 – Gráficos de influência para o modelo Gaussiano utilizando o método da REML para os dados de produtividade de soja.	65
Figura 22 – Mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de produtividade de soja.	67

Figura 23 – Demais mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de produtividade de soja.	68
---	----

LISTA DE TABELAS

Tabela 1 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0$, $\phi_2 = 1$ e $\phi_3 = 75$ para a malha dos dados de produtividade de soja	43
Tabela 2 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0.1$, $\phi_2 = 1$ e $\phi_3 = 75$ para a malha dos dados de produtividade de soja	44
Tabela 3 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0$, $\phi_2 = 1$ e $\phi_3 = 0.3$ para a malha regular 10×10	45
Tabela 4 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0.1$, $\phi_2 = 1$ e $\phi_3 = 0.3$ para a malha regular 10×10	45
Tabela 5 – Estatísticas descritivas dos dados de Jones	48
Tabela 6 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para os dados de Jones	49
Tabela 7 – Critérios para seleção do melhor ajuste entre os métodos da máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para os dados de Jones	50
Tabela 8 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança restrita (REML) para os dados de Jones	53
Tabela 9 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança restrita (REML) para os dados de Jones	53
Tabela 10 – Matriz de erros da quantidade de área em km^2 para os dados de Jones	56
Tabela 11 – Estatísticas descritivas dos dados de produtividade de soja	58
Tabela 12 – Resumo das informações sobre o semivariograma em uma distância δ , semi-variância $\gamma(\delta)$ e número de pares $N(\delta)$	60
Tabela 13 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança (ML) para os dados de produtividade de soja	61

Tabela 14 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança restrita (REML) para os dados de produtividade de soja . . .	62
Tabela 15 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança (ML) para os dados de produtividade de soja	62
Tabela 16 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança restrita (REML) para os dados de produtividade de soja . . .	63
Tabela 17 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para os dados de produtividade de soja	63
Tabela 18 – Matriz de erros da quantidade de área em <i>ha</i> para os dados de produtividade de soja	64
Tabela 19 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança restrita (REML) sem considerar os pontos influentes para os dados de produtividade de soja	66
Tabela 20 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança restrita (REML) para os dados de produtividade de soja . . .	66
Tabela 21 – Matriz de erros da quantidade de área em <i>ha</i> para os dados de produtividade de soja	69

SUMÁRIO

1	INTRODUÇÃO	14
1.1	MOTIVAÇÃO	14
1.2	OBJETIVOS	15
1.3	ESTRUTURA DO DOCUMENTO	15
2	GEOESTATÍSTICA	17
2.1	TEORIA DAS VARIÁVEIS REGIONALIZADAS	18
2.2	MODELOS ESPACIAIS LINEARES	19
2.2.1	Modelos Teóricos	23
2.2.1.1	Modelo Esférico	24
2.2.1.2	Modelo Exponencial	25
2.2.1.3	Modelo Gaussiano	25
2.2.1.4	Modelo Família Matérn	26
2.3	MÉTODOS DE ESTIMAÇÃO DOS PARÂMETROS	27
2.3.1	Método da Máxima Verossimilhança	28
2.3.1.1	Matriz de Informação Observada	29
2.3.1.2	Matriz de Informação Esperada	29
2.3.2	Método da Máxima Verossimilhança Restrita	29
2.3.2.1	Matriz de Informação Observada	30
2.3.2.2	Matriz de Informação Esperada	30
2.4	KRIGAGEM	30
2.4.1	Krigagem Ordinária	31
2.4.2	Krigagem Universal	34
2.5	MÉTODOS DE SELEÇÃO DE MODELOS	35
2.5.1	Validação Cruzada	35
3	DIAGNÓSTICOS DE INFLUÊNCIA	37
3.1	INFLUÊNCIA LOCAL	37
3.1.1	Influência Local na Variável Resposta	40
3.1.1.1	Perturbação de Zhu	40
4	SIMULAÇÕES	42
4.1	CONSIDERAÇÕES GERAIS	46
5	APLICAÇÃO A DADOS REAIS	47

5.1	DADOS JONES	47
5.1.1	Ajuste do Modelo	48
5.1.1.1	Critérios de Validação Cruzada	49
5.1.1.2	Diagnóstico de Influência Local	51
5.1.2	Ajuste do Modelo sem Observações Influentes	52
5.1.2.1	Critérios de Validação Cruzada	53
5.1.3	Considerações Gerais	56
5.2	DADOS DE PRODUTIVIDADE DE SOJA	57
5.2.1	Ajuste do Modelo	60
5.2.1.1	Critérios de Validação Cruzada	61
5.2.1.2	Diagnóstico de Influência Local	64
5.2.2	Ajuste do Modelo sem Observações Influentes	65
5.2.2.1	Critérios de Validação Cruzada	65
5.2.3	Considerações Gerais	69
6	CONCLUSÕES	71
	REFERÊNCIAS	73
	APÊNDICE A – CONTAS	78
	APÊNDICE B – PROVA	83

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

Uma das principais práticas na estatística atualmente é a construção de modelos estatísticos que possam descrever de forma realística dados em estudo (SEBER; LEE, 2012). Em particular a análise de regressão é uma das técnicas mais utilizadas para explorar dados em diversas áreas. Esta metodologia consiste em utilizar uma equação matemática para expressar a relação entre a variável de interesse (resposta) com um conjunto de variáveis de predição (explicativas) (MONTGOMERY *et al.*, 2012). No entanto, esses modelos descrevem processos que podem envolver dados de natureza complexa, implicando em uma interpretação que não seja totalmente correta, ou seja, em um modelo mal ajustado (COOK, 1986). Um dos motivos que podem influenciar no ajuste do modelo é a presença de dados discrepantes, que são observações que não seguem o padrão da maioria dos dados (ROUSSEEUW; ZOMEREN, 1990). Diversas técnicas têm sido desenvolvidas a fim de identificar a existência de observações influentes em conjuntos de dados e como esses valores podem influenciar na interpretação dos resultados.

As técnicas de regressão clássicas comumente utilizadas são muito úteis para descrever uma grande variedade de fenômenos naturais, entretanto, elas não consideram a informação espacial dos dados. No contexto da estatística espacial, a Geoestatística proporciona uma maneira de descrever como os valores são relacionados com a posição espacial que eles ocupam, e fornece adaptações de técnicas de regressão clássicas para aproveitar essa informação (ISAKS; SRIVASTAVA, 1989).

A Geoestatística pode ser considerada como uma coleção de ferramentas estatísticas que incorporam as características de atributos espaciais, também conhecidas como variáveis regionalizadas, no processo de análise dos dados (GOOVAERTS *et al.*, 1997; OLEA, 2012). É utilizada em diversas áreas do conhecimento para avaliar dados que provêm de fenômenos naturais e que são espacialmente distribuídos e autocorrelacionados. Para a interpretação desses dados leva-se em conta, além do valor obtido para essa variável, a sua posição, que geralmente é expressa por um sistema de coordenadas.

Diversos trabalhos vem sendo propostos na literatura para analisar a estrutura espacial

dos dados, que pode ser realizada por meio de modelos espaciais. A estimação dos parâmetros e a escolha do modelo é de fundamental importância para se obter uma interpretação adequada. No entanto, mesmo com um modelo bem ajustado, o resultado pode sofrer efeito de pontos influentes. Dessa forma é necessário a utilização de alguma ferramenta que identifique a sensibilidade das estimativas dos parâmetros do modelo. O método de diagnóstico de influência conhecido como influência local consiste em verificar a presença de observações atípicas que podem causar distorções nos resultados. Nesse contexto, utilizar técnicas de diagnóstico é essencial para se obter um modelo que represente os dados de forma mais confiável.

1.2 OBJETIVOS

Devido à importância de se avaliar os parâmetros que compõem a estrutura de dependência espacial dos modelos espaciais, o objetivo geral deste trabalho é:

Utilizar a metodologia de influência local em modelos espaciais lineares Gaussianos, considerando o método da máxima verossimilhança restrita.

Para alcançar o objetivo geral proposto, os objetivos específicos pontuados nessa dissertação são:

1. Realizar uma revisão da literatura do estado da arte sobre a Geoestatística, apresentando definições e conceitos da área;
2. Apresentar uma revisão sobre os modelos teóricos utilizados na Geoestatística para descrever a estrutura espacial dos dados bem como os métodos de estimação dos parâmetros dos modelos considerados neste trabalho;
3. Apresentar uma revisão sobre o método de interpolação (krigagem) e técnicas utilizadas para diagnósticos de influência;
4. Realizar estudos de simulação computacional para avaliar a influência local em modelos utilizando o método da máxima verossimilhança restrita;
5. Aplicar a abordagem desenvolvida para influência local a dois conjuntos de dados reais, comparando os métodos de estimação da máxima verossimilhança restrita e máxima verossimilhança para evidenciar a vantagem da abordagem proposta.

1.3 ESTRUTURA DO DOCUMENTO

O presente trabalho está estruturado da seguinte forma. No Capítulo 2, é apresentada a formalização teórica da Geoestatística, a Teoria das Variáveis Regionalizadas, apresentação dos modelos espaciais lineares de forma geral e alguns modelos específicos. Apresenta-se os

métodos de estimação dos parâmetros, com ênfase ao método utilizado no contexto do trabalho. Ainda no Capítulo 2, é apresentado o método de interpolação para obtenção dos mapas temáticos, conhecido como krigagem.

No Capítulo 3, são descritos os métodos de diagnósticos de influência local, em especial com o esquema de perturbação na variável resposta. Esse esquema considerando o método de máxima verossimilhança restrita para a estimação dos parâmetros é a contribuição desse trabalho. Nos Capítulos 4 e 5, são apresentados, respectivamente, os resultados de simulações computacionais e aplicação a dois conjuntos de dados reais distintos considerando os estimadores de máxima verossimilhança restrita para o modelo espacial linear Gaussiano.

Finalmente, no Capítulo 6, são apresentadas as conclusões do presente trabalho. O Apêndice A apresenta as contas desenvolvidas das derivadas das matrizes e no Apêndice B são exibidos relações de cálculo matricial que foram utilizadas no desenvolvimento das contas do Apêndice A.

2 GEOESTATÍSTICA

No presente capítulo, é apresentado o enquadramento teórico sobre a Geoestatística, bem como sua interpretação matemática.

Conforme Cressie (1993) o estudo Geoestatística surgiu na África do Sul, em 1951, com a pretensão de encontrar estimativas para o cálculo de reservas minerais. Esta primeira abordagem à Geoestatística foi proposta por Daniel G. Krige e H.S. Sichel. Já na década seguinte, na França, Matheron (1962, 1963) fez importante colaboração, onde expandiu o tema sobre a Geoestatística e formalizou a Teoria das Variáveis Regionalizadas.

Segundo Chilès e Delfiner (1999), a Geoestatística tem como objetivo fornecer descrições quantitativas de variáveis naturais distribuídas no espaço, ou seja, busca descrever o comportamento espacial dos dados, modelando a variabilidade espacial de variáveis georreferenciadas (ASSUMPÇÃO *et al.*, 2014). Esta metodologia pode ser utilizada em diversas áreas, como: geologia (GOOVAERTS *et al.*, 1997), ciências do solo (TRANGMAR *et al.*, 1986), meteorologia (GNEITING, 2002), socioeconomia (ANSELIN, 2001), hidrologia (BÁRDOSSY, 2006), processamento de imagens (CHICA-OLMO; ABARCA-HERNANDEZ, 2000) ou qualquer área que trabalhe com dados georeferenciados. O prefixo "geo" enfatiza o aspecto espacial do problema. Consequentemente, como a metodologia se comunica com muitas outras áreas de pesquisa, diversas abordagens tem sido desenvolvidas na literatura. Journel e Huijbregts (1978) apresentaram uma abordagem em aplicações de mineração. Kitanidis e VoMvoris (1983) tratam do tema no estudo de lençóis freáticos. Goovaerts (2001) apresenta a questão da modelagem da incerteza sobre o valor de atributos do solo. Borssoi *et al.* (2009, 2011a, 2011b) apresentaram técnicas de diagnóstico de influência local na análise espacial da produtividade de soja. Uribe-Opazo *et al.* (2012) desenvolveram métodos de diagnóstico de influência local no caso do estudo de modelos espaciais lineares com distribuição gaussiana. De Bastiani *et al.* (2015a, 2015b, 2018) apresentaram diversos estudos de diagnósticos de influência considerando modelos lineares espaciais.

Cressie (1989) sustenta a teoria em que os dados espaciais são caracterizados pelo espaço e tempo atribuídos a eles. O autor sugere que os dados mais próximos são geralmente mais correlacionados entre si do que aqueles que estão mais distantes, conforme a Primeira Lei da Geografia: *"Todas as coisas estão relacionadas, mas coisas mais próximas são mais parecidas do que as coisas mais distantes"* (TOBLER, 1970). Segundo Isaaks e Srivastava (1989) a Geoestatística oferece técnicas para descrever a continuidade espacial que é uma característica

presente nos fenômenos naturais. A Geoestatística pode ser considerada como uma coleção de técnicas, utilizadas para modelar as incertezas atribuídas as variáveis espaciais (OLEA, 2012).

2.1 TEORIA DAS VARIÁVEIS REGIONALIZADAS

Matheron (1963) formalizou, de forma independente, a mesma idéia do estudo da Geoestatística e a denominou de "Teoria das Variáveis Regionalizadas". Esta teoria, define uma variável regionalizada como sendo uma função real, que toma um valor definido em cada ponto do espaço (uma variável indexada a uma localização). Esta variável pode apresentar continuidade espacial, e isotropia, quando a variabilidade espacial apresenta comportamento semelhante em todas as direções. Em caso contrário, denominamos o fenômeno por anisotropia. Nessa situação identificamos a presença de uma direção privilegiada dos dados.

Atualmente a Geoestatística é um reconhecido tópico da estatística aplicada que trata de problemas referentes às variáveis regionalizadas (LANDIM, 2006). A notação mais usual que define este processo estocástico é dada por (SCHABENBERGER; PIERCE, 2001):

$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}, \quad (2.1)$$

em que $Z(\mathbf{s})$ é um vetor aleatório na localização $\mathbf{s}$, sendo $\mathbf{s}$ um par de coordenadas, que representa a localização espacial da variável, D um subconjunto fixo de $\mathbb{R}^d$ e $\mathbb{R}^d$ um espaço Euclidiano d -dimensional ($d \geq 1$) (neste trabalho, consideramos $d = 2$). Considera-se que $z(\mathbf{s}_i)$ seja o valor de uma variável Z no ponto $\mathbf{s}_i$, $\mathbf{s}_i = (x_i, y_i)$, $\mathbf{s}_i \in D \subset \mathbb{R}^2$, com $i = 1, 2, \dots, n$, $n \in \mathbb{N}$. Pode-se representar a variabilidade da função $f(z(\mathbf{s}_i))$ no espaço, com Z variando dependentemente do local em que foi observado. Cada valor do atributo em estudo está relacionado a uma localização $\mathbf{s}_i$, então, $z(\mathbf{s}_i)$ representa uma realização de uma variável aleatória $Z(\mathbf{s}_i)$. Quando se considera n variáveis em estudo, nas localizações $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ na região $D \subset \mathbb{R}^2$, em que $\mathbf{s}$ varia continuamente sobre D , obtemos a realização de um processo estocástico, que pode ser denotado por:

$$Z(\mathbf{s}) = \{Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)\}. \quad (2.2)$$

Na prática, $\mathbf{s}$ é definida como a localização da variável no espaço e é representada por um par de coordenadas (latitude, longitude). Dessa forma, caracterizamos a variável regionalizada. As variáveis regionalizadas possuem características diretamente associadas à estrutura do fenômeno natural que elas representam, que são a localização, continuidade e isotropia. Essas características podem ser descritas como:

- a) Localização - Determinar a malha amostral onde será realizado o estudo é de fundamental importância. Para esta tarefa o conhecimento do pesquisador é necessário, uma vez que nela serão coletadas as variáveis de interesse. A determinação da malha deve também ser feita com cuidado e respeitando certas condições de homogeneidade física;
- b) Continuidade - Fundamenta-se que a esperança da média amostral, para amostras próximas no espaço, sejam similares entre si, perdendo a similaridade em função da distância. Denomina-se efeito pepita puro quando a continuidade espacial não é identificada;
- c) Isotropia - Quando verificamos que não há uma direção privilegiada nos dados, dizemos que os dados apresentam isotropia. A dependência espacial neste caso é uma função que depende apenas da distância. Quando a não indentificamos tal característica, ou seja, identificamos um comportamento distinto para as diferentes direções, assumimos o cenário de anisotropia. Geralmente, estuda-se as direções 0° , 45° , 90° e 135° . A Figura 1 apresenta dados simulados isotrópicos, conforme Figura 1a, e dados que apresentam anisotropia na direção de 45° , conforme a Figura 1b.

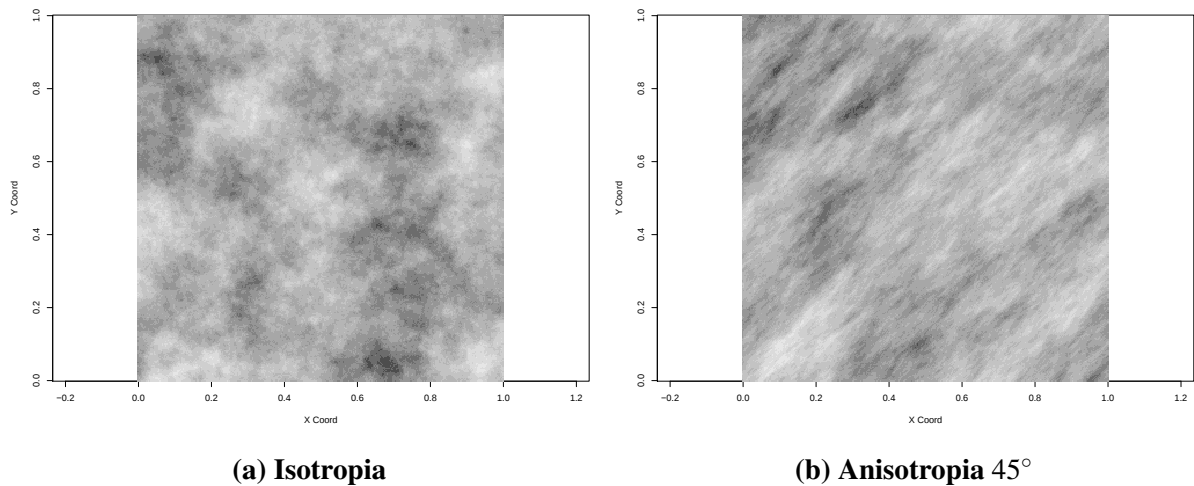


Figura 1 – Dados simulados apresentando os comportamentos de isotropia e anisotropia.

2.2 MODELOS ESPACIAIS LINEARES

Para modelar um conjunto de dados na qual sua estrutura possua dependência espacial, segundo Mardia e Marshall (1984), consideramos um processo estocástico $Z = \{Z(\mathbf{s}_i), \mathbf{s}_i \in D\}$, em que D é um subconjunto de $\mathbb{R}^2$, o espaço euclidiano. Considerando $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$ as observações do processo estocástico de localizações conhecidas distin-

tas $\mathbf{s}_i$, em que $\mathbf{s}_i$ é um vetor d -dimensional. Temos que $Z(\mathbf{s}_i)$ são geradas pelo modelo

$$Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \quad (2.3)$$

em que ambos os termos, $\mu(\mathbf{s}_i)$ o termo determinístico e $\varepsilon_i = \varepsilon(\mathbf{s}_i)$ o termo estocástico, dependem da localização espacial em que $Z(\mathbf{s}_i)$ é avaliada. Assume-se que o erro estocástico tem média zero, isto é, $E[\varepsilon(\mathbf{s}_i)] = 0$, e que a variação espacial entre as observações é determinada por uma função de covariância $C(\mathbf{s}_i, \mathbf{s}_j) = \text{cov}[\varepsilon(\mathbf{s}_i), \varepsilon(\mathbf{s}_j)]$, $i, j = 1, 2, \dots, n$.

Assume-se que para algumas funções conhecidas de $\mathbf{s}_i$, $x_1(\mathbf{s}_i), \dots, x_p(\mathbf{s}_i)$, a média do processo estocástico é dada por

$$\mu(\mathbf{s}_i) = \sum_{j=1}^p x_j(\mathbf{s}_i) \beta_j, \quad (2.4)$$

em que $\beta_1, \dots, \beta_p$ são os parâmetros desconhecidos a serem estimados. Além disso, a função de covariância $C(\mathbf{s}_i, \mathbf{s}_j)$ é caracterizada por um vetor de parâmetros q -dimensional $\boldsymbol{\phi} = (\phi_1, \dots, \phi_q)^\top$ (neste trabalho, consideramos $q = 3$). Sua notação é explicitada por:

$$\mathbf{Z} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad (2.5)$$

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \quad (2.6)$$

em que $\mathbf{X}$ é uma matriz $n \times p$ com i -ésima linha $\mathbf{x}_i^\top = (x_1(\mathbf{s}_1), \dots, x_p(\mathbf{s}_i))$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^\top$ e $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$. A matriz $\boldsymbol{\varepsilon}$ é dada por $\boldsymbol{\Sigma} = [(\sigma_{ij})]$, em que $\sigma_{ij} = C(\mathbf{s}_i, \mathbf{s}_j)$.

A matriz de covariância $\boldsymbol{\Sigma}$ também pode ser expressa por

$$\boldsymbol{\Sigma} = \phi_1 \mathbf{I}_n + \phi_2 \mathbf{R}(\phi_3), \quad (2.7)$$

em que $\mathbf{R}(\phi_3) = [(r_{ij})]$ é uma matriz $n \times n$ simétrica com elementos da diagonal $r_{ii} = 1$, para $i = 1, \dots, n$, com elementos $r_{ij} = (1/\phi_2)\sigma_{ij}$ para $i \neq j = 1, \dots, n$ e $\phi_2 \neq 0$, e $r_{ij} = 0$ para $i \neq j = 1, \dots, n$ e $\phi_2 = 0$.

Os parâmetros que definem a estrutura de dependência espacial mostrados na Equação (2.7) anterior são respectivamente o efeito pepita (ϕ_1), contribuição (ϕ_2), o alcance que é dado por uma função de ϕ_3 ($a = g(\phi_3)$) e o patamar ($P = \phi_1 + \phi_2$).

Segundo De Bastiani (2012) o parâmetro de efeito pepita representa a descontinuidade do semivariograma para pequenas distâncias, menores do que a menor distância entre os elementos amostrais. Segundo Isaaks e Srivastava (1989), a causa da descontinuidade na origem do semivariograma pode ser explicada por erros na amostragem ou da variabilidade de pequena escala que não é captada. Cressie (1993), também destaca que os motivos que acarretam

na descontinuidade na origem do semivariograma são causados pelos erros de medição ou da variabilidade em microescala.

Segundo De Bastiani (2012) a contribuição conhecida como variância de dispersão, representa as diferenças espaciais entre os valores de uma variável tomada em dois pontos separados por distâncias cada vez maiores. É também a diferença entre o patamar (que será apresentado a seguir) e o efeito pepita. O alcance é a distância dentro da qual as amostras apresentam autocorrelação espacial, ou seja, determina o raio de dependência espacial, distância a partir da qual uma variável não exerce mais influência sobre a variável vizinha. O alcance é função do parâmetro ϕ_3 ($a = g(\phi_3)$). O patamar é uma função da variância total da variável em estudo. É o valor máximo do semivariograma correspondente ao seu alcance, a partir do qual o semivariograma se estabiliza. Deste ponto em diante, consideramos que não existe mais dependência espacial entre as amostras.

Considerando que o processo estocástico $Z(\mathbf{s}_i)$ tem valor esperado e variâncias definidas respectivamente por

$$E[Z(\mathbf{s}_i)] = \mu(\mathbf{s}_i) \quad \text{e} \quad E[Z(\mathbf{s}_i + \delta)] = \mu(\mathbf{s}_i), \quad (2.8)$$

$$Var[Z(\mathbf{s}_i)] = C(0) \quad \text{e} \quad Var[Z(\mathbf{s}_i + \delta)] = C(0), \quad (2.9)$$

para as localizações $\mathbf{s}_i$, $\mathbf{s}_i + \delta$, em que δ é um vetor qualquer de distâncias. Assim, a função de covariância espacial é dada por

$$C(\mathbf{s}_i, \mathbf{s}_i + \delta) = C(\delta) = E[Z(\mathbf{s}_i)Z(\mathbf{s}_i + \delta)] - E[Z(\mathbf{s}_i)]E[Z(\mathbf{s}_i + \delta)]. \quad (2.10)$$

e a função de semivariância é dada por:

$$\gamma(\delta) = \gamma(\mathbf{s}_i, \mathbf{s}_i + \delta) = \frac{1}{2}E[Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \delta)]^2. \quad (2.11)$$

Segundo Oliver e Webster (2015), assumimos que a variável regionalizada seja estacionária, isto é, esta suposição nos permite considerar os dados como se tivessem o mesmo grau de variação sobre a região de interesse. Sob a hipótese de estacionaridade de segunda ordem Isaaks e Srivastava (1989), Wackernagel (2013) demonstraram a partir das equações (2.10) e (2.11) que

$$\gamma(\delta) = C(0) - C(\delta), \quad (2.12)$$

em que $C(0) = \phi_1 + \phi_2$ representa a variância do processo.

A função de semivariância apresenta algumas propriedades, que são listadas a seguir:

- i) $\gamma(0) = 0$;
- ii) $\gamma(\delta) = \gamma(-\delta) \geq 0$, a função de semivariância sempre produz valores positivos;
- iii) $\lim_{|\delta| \rightarrow \infty} \frac{\gamma(\delta)}{|\delta|^2} = 0$, ou seja, $\gamma(\delta)$ cresce mais lentamente que δ^2 .

Quando $C(0) \neq 0$ na Equação (2.12), temos que:

$$\rho(\delta) = \frac{C(\delta)}{C(0)} = 1 - \frac{\gamma(\delta)}{C(0)}, \quad (2.13)$$

em que $\rho(\delta)$ é função de correlação espacial. Sabe-se que $\rho(\delta)$ é função monótona não crescente, ou seja, a correlação entre duas observações decresce ou permanece constante a medida que a distância aumenta. Temos também que, quando $\delta \rightarrow \infty$, $\rho(\delta) \rightarrow 0$.

Ainda, a variância é definida por

$$\text{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \delta)] = 2\gamma(\mathbf{s}_i, \mathbf{s}_i + \delta). \quad (2.14)$$

No entanto, para que seja uma função de variância, deve satisfazer a propriedade de ser condicionalmente negativa definida. Uma função de variância $2\gamma(\mathbf{s}_i, \mathbf{s}_i + \delta)$ é negativa definida se, e somente se, para um número finito de localizações $\{\mathbf{s}_i : i = 1, \dots, n\}$ e números reais $\{a_i : i = 1, \dots, n\}$ tal que $\sum_{i=1}^n a_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n a_i a_j 2\gamma(s_i - s_j) \leq 0. \quad (2.15)$$

Assim, qualquer função que satisfaz

$$\text{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \delta)] = 2\gamma(\mathbf{s}_i, \mathbf{s}_i + \delta), \quad (2.16)$$

é condicional negativa definida.

Quando o processo estocástico $Z(\mathbf{s}_i)$ é intrinsecamente estacionário, temos que

$$\text{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \delta)] = E[Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \delta)]^2. \quad (2.17)$$

A análise geoestatística envolve estimativa e modelagem da dependência espacial, que pode realizada através das covariâncias ou semivariâncias. Normalmente utiliza-se as semivariâncias, o que gera o semivariograma. O semivariograma traça as semivariâncias em função das distâncias (BIVAND *et al.*, 2008).

Assim, podemos desenvolver o gráfico do semivariograma, que é gerado pela $\gamma(\delta)$ *versus* o valor correspondente de δ . A função de semivariância mede a variância entre as variáveis separadas por uma distância δ . Assim, pontos mais próximos, por estarem correlacionados, terão

essa variância pequena e aumentam a medida que a distância cresce. O contrário acontece com a função de covariância, que mede a correlação entre os pontos, que é grande para pequenas distâncias e diminui a medida que a distância aumenta (YAMAMOTO, 2001). A Figura 2 abaixo apresenta a relação entre a covariância espacial e a semivariância.

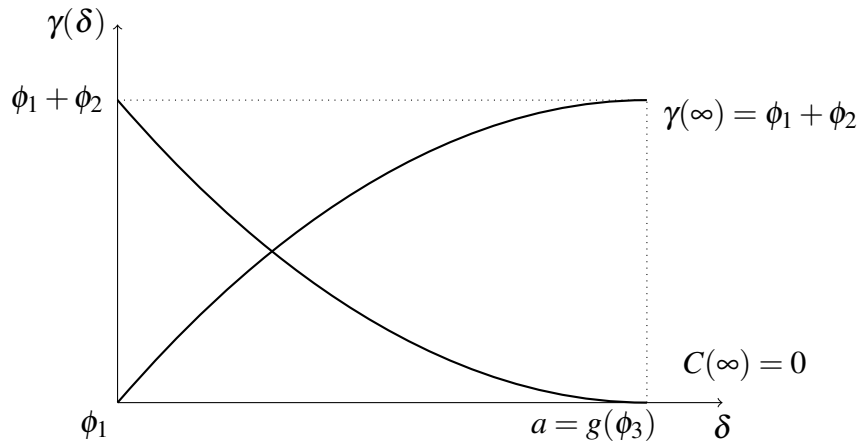


Figura 2 – Relação entre a semivariância $\gamma(\delta)$ e a covariância espacial $C(\delta)$ (DE BASTIANI, 2012).

2.2.1 Modelos Teóricos

Os modelos teóricos são utilizados para avaliar a estrutura de variabilidade espacial. Os modelos mais encontrados na literatura são os modelos esférico, exponencial, Gaussiano e da família Matérn. A grande notoriedade na utilização desses modelos pode ser explicada pelo fato de que suas características descrevem e ajustam de maneira mais eficiente a natureza dos dados. Os modelos citados acima e demais modelos não explicitados aqui, podem ser encontrados em (ISAAKS; SRIVASTAVA, 1989; CRESSIE, 1993; DIGGLE; RIBEIRO, 2007).

Os modelos teóricos são classificados em dois grupos: modelos transitivos, quando possuem patamar, e não transitivos, quando não possuem patamar. Segundo Journel e Huijbregts (1978), os modelos que consideram a variável em estudo são, por natureza, isotrópicos, isto é, possuem um mesmo padrão de continuidade espacial em todas as direções. Em consequência disso, o semivariograma deve ter características estruturais semelhantes em todas as direções do espaço.

A seguir, serão apresentados os principais modelos teóricos.

2.2.1.1 Modelo Esférico

O modelo esférico tem como característica um rápido crescimento na origem e atinge seu patamar a 2/3 do alcance. O modelo é válido em $\mathbb{R}^2$ e tem função de semivariância expressa por (SCHABENBERGER; PIERCE, 2001)

$$\gamma(\delta) = \begin{cases} 0, & \text{se } \delta = 0, \\ \phi_1 + \phi_2 \left[\frac{3}{2} \left(\frac{\delta}{\phi_3} \right) - \frac{1}{2} \left(\frac{\delta}{\phi_3} \right)^3 \right], & \text{se } 0 < \delta \leq \phi_3, \\ \phi_1 + \phi_2, & \text{se } \delta > \phi_3. \end{cases} \quad (2.18)$$

Sua função de covariância é dada por:

$$C(\delta) = \begin{cases} \phi_1 + \phi_2, & \text{se } \delta = 0, \\ \phi_2 \left[1 - \frac{3}{2} \left(\frac{\delta}{\phi_3} \right) + \frac{1}{2} \left(\frac{\delta}{\phi_3} \right)^3 \right], & \text{se } 0 < \|\delta\| \leq \phi_3, \\ 0, & \text{se } \|\delta\| > \phi_3, \end{cases} \quad (2.19)$$

e a função de correlação é determinada por:

$$\rho(\delta) = \begin{cases} 1, & \text{se } \delta = 0, \\ 1 - \frac{3}{2} \left(\frac{\delta}{\phi_3} \right) + \frac{1}{2} \left(\frac{\delta}{\phi_3} \right)^3, & \text{se } \delta > 0. \end{cases} \quad (2.20)$$

Na Figura 3 é apresentado o comportamento da função de semivariância, quando considerado o modelo esférico para realização do ajuste do conjunto de dados.

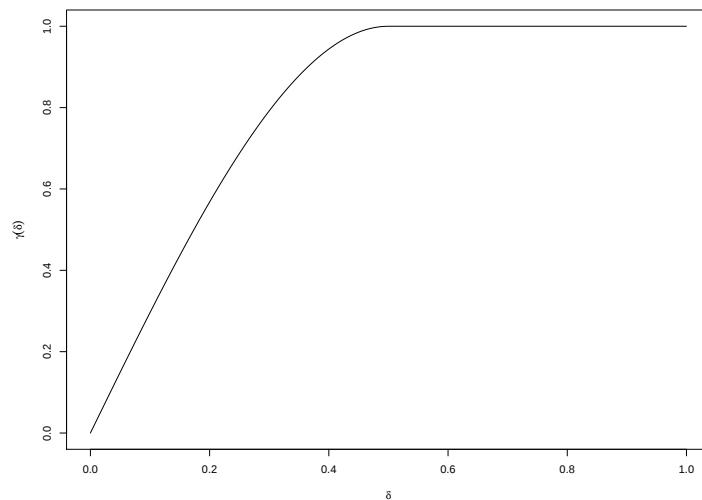


Figura 3 – Gráfico da função de semivariância para o modelo esférico.
Fonte: Do autor.

2.2.1.2 Modelo Exponencial

O modelo exponencial apresenta um comportamento aproximadamente linear na origem e atinge o patamar assintoticamente. O alcance é dado por $a = 3\phi_3$. Este modelo é válido em $\mathbb{R}^2$ e tem função de semivariância expressa por:

$$\gamma(\delta) = \begin{cases} 0, & \text{se } \delta = 0, \\ \phi_1 + \phi_2 \left[1 - e^{\left(-\frac{\delta}{\phi_3}\right)} \right], & \text{se } \delta > 0. \end{cases} \quad (2.21)$$

A função de covariância é dada por:

$$C(\delta) = \begin{cases} \phi_1 + \phi_2, & \text{se } \delta = 0, \\ \phi_2 e^{\left(-\frac{\delta}{\phi_3}\right)}, & \text{se } \delta > 0, \end{cases} \quad (2.22)$$

e a função de correlação é determinada por:

$$\rho(\delta) = \begin{cases} 1, & \text{se } \delta = 0, \\ e^{\left(-\frac{\delta}{\phi_3}\right)}, & \text{se } \delta > 0. \end{cases} \quad (2.23)$$

A Figura 4 é apresentado o comportamento da função de semivariância, quando considerado o modelo exponencial para realização do ajuste do conjunto de dados.

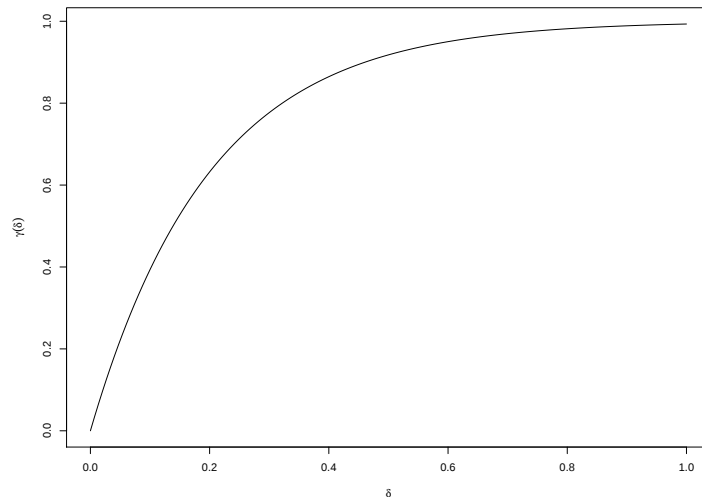


Figura 4 – Gráfico da função de semivariância para o exponencial.
Fonte: Do autor.

2.2.1.3 Modelo Gaussiano

O modelo Gaussiano apresenta um comportamento parabólico na origem e atinge o patamar assintoticamente. Seu alcance é dado por $a = \sqrt{3}\phi_3$. O modelo é válido em $\mathbb{R}^2$ e tem

função de semivariância expressa por:

$$\gamma(\delta) = \begin{cases} 0, & \text{se } \delta = 0, \\ \phi_1 + \phi_2 \left[1 - e^{-\left(\frac{\delta}{\phi_3}\right)^2} \right], & \text{se } \delta > 0. \end{cases} \quad (2.24)$$

A função de covariância é dada por:

$$C(\delta) = \begin{cases} \phi_1 + \phi_2, & \text{se } \delta = 0, \\ \phi_2 \left[1 - e^{-\left(\frac{\delta}{\phi_3}\right)^2} \right], & \text{se } \delta > 0. \end{cases} \quad (2.25)$$

e a função de correlação é determinada por:

$$\rho(\delta) = \begin{cases} 1, & \text{se } \delta = 0, \\ 1 - e^{-\left(\frac{\delta}{\phi_3}\right)^2}, & \text{se } \delta > 0. \end{cases} \quad (2.26)$$

A Figura 5 é apresentado o comportamento da função de semivariância, quando considerado o modelo Gaussiano para realização do ajuste do conjunto de dados.

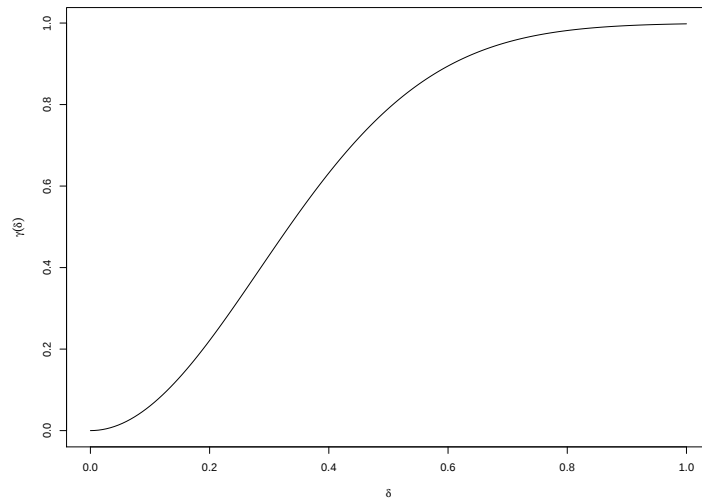


Figura 5 – Gráfico da função de semivariância para o modelo Gaussiano.
Fonte: Do autor.

2.2.1.4 Modelo Família Matérn

Matérn (1986) apresentou a função conhecida atualmente como família Matérn. Sua função de semivariância é expressa por:

$$\gamma(\delta) = \begin{cases} 0, & \text{se } \|\delta\| = 0, \\ \phi_1 + \phi_2 \left[1 - (2^{\kappa-1} \Gamma(\kappa))^{-1} \left(\frac{\|\delta\|}{\phi_3} \right)^{\kappa} K_{\kappa} \left(\frac{\|\delta\|}{\phi_3} \right) \right], & \text{se } \|\delta\| > 0. \end{cases} \quad (2.27)$$

Em que K_κ , é a função de Bessel modificada de terceira ordem κ , sendo

$$K_\kappa(u) = \frac{1}{2} \int_0^\infty x^{\kappa-1} e^{-\frac{1}{2}u(x+\frac{1}{x})} dx. \quad (2.28)$$

Esta função é válida para $\phi_3, \kappa > 0$.

O parâmetro de forma κ determina a suavização analítica do processo $Z(\mathbf{s})$. A função de covariância é definida por:

$$C(\delta) = \begin{cases} 0, & \text{se } \|\delta\| = 0, \\ \phi_2 \left[(2^{\kappa-1} \Gamma(\kappa))^{-1} \left(\frac{\|\delta\|}{\phi_3} \right)^\kappa K_\kappa \left(\frac{\|\delta\|}{\phi_3} \right) \right], & \text{se } \|\delta\| > 0. \end{cases} \quad (2.29)$$

A função de correlação é expressa por:

$$\rho(\delta) = \begin{cases} 1, & \text{se } \|\delta\| = 0 \\ (2^{\kappa-1} \Gamma(\kappa))^{-1} \left(\frac{\|\delta\|}{\phi_3} \right)^\kappa K_\kappa \left(\frac{\|\delta\|}{\phi_3} \right), & \text{se } \|\delta\| > 0. \end{cases} \quad (2.30)$$

A Figura 6 é apresentado o comportamento da função de semivariância, quando considerado o modelo da família Matérn para realização do ajuste do conjunto de dados, para diferentes valores de κ e ϕ .

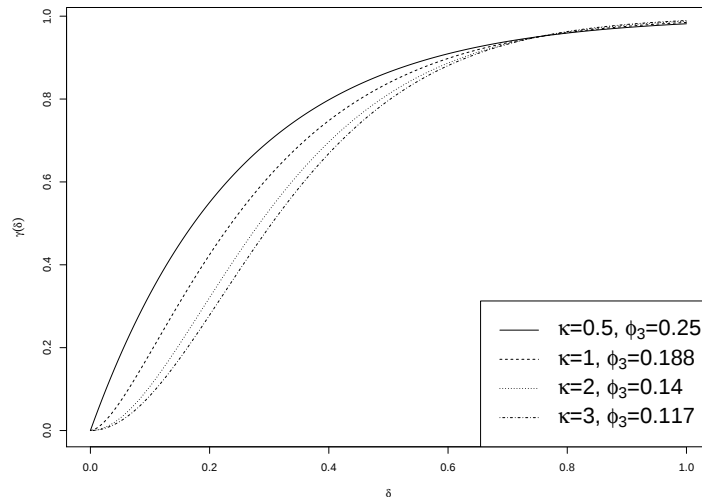


Figura 6 – Gráfico da função de semivariância para o modelo da família Matérn para diferentes valores de κ e ϕ_3 .

Fonte: Do autor.

2.3 MÉTODOS DE ESTIMAÇÃO DOS PARÂMETROS

Como temos interesse em obter a estimação dos parâmetros do semivariograma $(\phi_1, \phi_2 \text{ e } a = g(\phi_3))$, é essencial o uso de algum método estatístico. Na estatística, existem

diversos métodos para obter a estimação dos parâmetros, como o método dos momentos, estimadores de máxima verossimilhança, estimadores de Bayes (CASELLA; BERGER, 2002). Vamos destacar aqui, o método da máxima verossimilhança (*maximum likelihood* - *ML*) que é usado extensivamente para a estimação de parâmetros desconhecidos de modelos (DOWD; PARDO-IGÚZQUIZA, 2002). Este estimador é apresentado no contexto da Geoestatística em (MARDIA; MARSHALL, 1984), e é muito utilizado na literatura pois possui boas propriedades assintóticas (DIGGLE; RIBEIRO, 2007).

Outro método que pode ser utilizado para estimar os parâmetros da função de covariância, é o método da máxima verossimilhança restrita (*restricted maximum likelihood* - *MLE*). Proposto por Patterson e Thompson (1971) na análise de variância, foi introduzido também no contexto de funções espaciais por Kitanidis (1983).

A seguir, serão apresentados estes dois métodos estatísticos de estimação de parâmetros, que serão utilizados neste trabalho.

2.3.1 Método da Máxima Verossimilhança

Introduzida por Mardia e Marshall (1984) para o caso em que a função aleatória $Z(x)$ é normalmente distribuída (MCBRATNEY; WEBSTER, 1986). A estimação dos parâmetros por meio da máxima verossimilhança é um método já consolidado, com boas propriedades assintóticas e possui variância mínima (KITANIDIS, 1983; DIGGLE; RIBEIRO, 2007). Cox e Hinkley (1979) apresentam mais detalhadamente as condições para que as propriedades do estimador de máxima verossimilhança sejam satisfeitas.

Considera-se um modelo $Z \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, em que $\mathbf{X}$ é uma matriz de covariáveis, $\boldsymbol{\beta}$ o vetor de parâmetros de regressão e matriz de covariância $\boldsymbol{\Sigma} = \phi_1 \mathbf{I}_n + \phi_2 \mathbf{R}$, que depende apenas do vetor de parâmetros $\boldsymbol{\phi}$ (DIGGLE; RIBEIRO, 2007). O logaritmo da função de verossimilhança é dado por (CASTRILLON-CANDÁS *et al.*, 2016):

$$l(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (Z - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (Z - \mathbf{X}\boldsymbol{\beta}), \quad (2.31)$$

em que $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top)^\top$. Por meio da maximização da função escore são geradas as estimativas de máxima verossimilhança dos parâmetros do modelo.

2.3.1.1 Matriz de Informação Observada

A matriz de informação observada é definida por $I(\theta) = -\ddot{L}(\theta)$ e deve ser avaliada em $\theta = \hat{\theta}$. Sua forma particionada pode ser encontrada a partir de $-I(\theta) = \ddot{L}(\theta)$, sendo

$$\ddot{L}(\theta) = \ddot{L}_{\theta\theta} = \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^\top} = \begin{pmatrix} \ddot{L}_{\beta\beta} & \ddot{L}_{\beta\phi} \\ \ddot{L}_{\phi\beta} & \ddot{L}_{\phi\phi} \end{pmatrix}. \quad (2.32)$$

2.3.1.2 Matriz de Informação Esperada

Para distribuição normal, matriz de informação esperada é dada por (URIBE-OPAZO *et al.*, 2012):

$$K(\theta) = E[I(\theta)] = \begin{pmatrix} E[-\ddot{L}_{\beta\beta}] & E[-\ddot{L}_{\beta\phi}] \\ E[-\ddot{L}_{\phi\beta}] & E[-\ddot{L}_{\phi\phi}] \end{pmatrix} = \begin{pmatrix} K(\beta) & 0 \\ 0 & K(\phi) \end{pmatrix}, \quad (2.33)$$

em que $K(\beta) = X^\top \Sigma^{-1} X$ e $K(\phi) = [(k_{ij}(\phi))]$, em que ij -ésimo elemento de k é dado por:

$$k_{ij}(\phi) = \frac{1}{2} \text{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \right), \quad (2.34)$$

com $i, j = 1, 2, \dots, q$. Neste trabalho $q = 3$.

2.3.2 Método da Máxima Verossimilhança Restrita

O método de máxima verossimilhança restrita tem como propósito estimar o vetor de parâmetros $\theta = (\phi_1, \phi_2, \phi_3)$ de uma função de covariância (ZIMMERMAN, 1989; KITANIDIS, 1987). Assumindo que $E[Z] = \mathbf{X}\beta$, pode-se transformar linearmente os dados para $Z^* = AY$, em que a distribuição não Z^* não depende diretamente de β , já que a informação do parâmetro β é incorporada pela transformação. O método tem como princípio estimar os parâmetros que determinam a estrutura de covariância dos dados, aplicando o método de máxima verossimilhança nos dados transformados Z^* (DIGGLE; RIBEIRO, 2007). Por exemplo, a matriz de projeção dos resíduos dos mínimos quadrados ordinários é dado por (DUTTA *et al.*, 2016):

$$A = X(X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1}, \quad (2.35)$$

sendo A uma matriz de projeção, logo essa matriz é simétrica.

Sob hipótese de normalidade, em que $Z \sim N_n(\mathbf{X}\beta, \Sigma)$, os parâmetros desconhecidos do modelo podem ser estimados através do logaritmo da função de verossimilhança restrita, dada por (MARCHANT; LARK, 2007):

$$l_R(\phi) = -\frac{n-p}{2} + \frac{1}{2} \log(|\mathbf{X}^\top \mathbf{X}|) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} \log|\mathbf{X}^\top \Sigma^{-1} \mathbf{X}| - \frac{1}{2} \mathbf{Z}^\top \mathbf{P} \mathbf{Z}, \quad (2.36)$$

em que Σ é a matriz de covariância dada na Equação (2.7), $P = \Sigma^{-1}(\mathbf{I} - \mathbf{A})$ é a inversa da matriz de covariância que multiplica a matriz identidade ($\mathbf{I}$) e a matriz de projeção $\mathbf{A}$, em que $\mathbf{A} = \mathbf{X}(\mathbf{X}^\top \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Sigma^{-1}$.

2.3.2.1 Matriz de Informação Observada

A matriz de informação observada é definida por $I(\theta) = -L(\theta)$ e deve ser avaliada em $\theta = \hat{\theta}$. Sua forma particionada pode ser encontrada a partir de $-I(\theta) = L(\theta)$, sendo

$$L(\theta) = L_{\phi\phi} = \frac{\partial^2 L(\phi)}{\partial \phi_i \partial \phi_j}, \quad (2.37)$$

e, utilizando os resultados de cálculo matricial, obtém-se para os modelos espaciais com distribuição Gaussiana, dado por (URIBE-OPAZO *et al.*, 2012):

$$L_{\phi\phi} = \frac{1}{2} \text{tr} \left[P \left(\frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} - \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} \right) \right] + \frac{1}{2} Z^\top P \left[\frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} - \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} - \frac{\partial \Sigma}{\partial \phi_i} P \frac{\partial \Sigma}{\partial \phi_j} \right] P Z. \quad (2.38)$$

2.3.2.2 Matriz de Informação Esperada

A matriz de informação esperada neste caso, é dada por:

$$K(\theta) = E[-\ddot{L}_{\phi\phi}] = K(\phi), \quad (2.39)$$

em que, neste caso $K(\phi) = [k_{ij}(\phi)]$, e o ij -ésimo elemento de k é dado por:

$$k_{ij}(\phi) = \frac{1}{2} \text{tr} \left\{ \Sigma \left[P \left(\frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} - 2 \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} \right) P \right] \right\} + \frac{1}{2} Z^\top \left[P \left(\frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} - 2 \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} \right) P \right] Z, \quad (2.40)$$

com $i, j = 1, 2, 3$.

Os resultados dos cálculos matriciais são apresentados no Apêndice A. Para o desenvolvimento dos cálculos foram utilizados os seguintes materiais como referência (NEL, 1980; LUTKEPOHL, 1997; PETERSEN; PEDERSEN, 2012).

2.4 KRIGAGEM

O termo "*kriging*" (em inglês), krigagem (em português) que dá nome ao método, foi uma homenagem de Matheron (1963) em reconhecimento ao pioneirismo de Daniel Krige (1951), que já utilizava a metodologia para estimação de ouro e outros metais. A krigagem é uma ferramenta indispensável para o estudo da variabilidade espacial da variável de interesse.

A krigagem tem como função a estimação do valor de atributos em locais ainda não amostrados. Para isso, se beneficia dos pontos já conhecidos, ou seja, utiliza-se da informação já disponibilizada pelas variáveis amostradas para estimar o valor das variáveis vizinhas. Segundo Vieira *et al.* (2000) a krigagem usa a dependência espacial entre as amostras vizinhas, que são expressas no semivariograma, para estimar valores em qualquer posição do campo de amostragem, sem tendência e com variância mínima. É útil mencionar a necessidade que exista uma certa continuidade espacial. Segundo Landim (2000), por ser um método que proporciona estimativas lineares, não viesadas e variância mínima de seus erros de estimação, é dito ser um interpolador BLUE - *best linear unbiased estimator* - melhor estimador não viesado (KITANIDIS; SHEN, 1996).

Nessa etapa, realiza-se a construção de mapas que permitem descrever e analisar comportamento da variabilidade espacial. A construção dos mapas de krigagem pode ser feita através dos métodos de interpolação geoestatísticos. Segundo Oliver e Webster (2015), nenhum método fornece boas estimativas dos erros em suas interpolações, exceto a krigagem.

A seguir, serão apresentados os métodos de krigagem ordinária e krigagem universal que foi proposta por Matheron (1969) para fornecer estimativas lineares de uma variável, mesmo quando identificada a presença de tendência ou desvio.

2.4.1 Krigagem Ordinária

A krigagem ordinária foi a primeira técnica de estimação que forneceu o cálculo do erro de estimativa pela krigagem ou estimação da variância (ROCHA; YAMAMOTO, 2000). A krigagem ordinária não requer que a média da variável estimada seja constante conhecida, desde que permaneça constante (JOURNEL, 1989). Ainda de acordo com Journel (1986), a krigagem ordinária é constituída na idéia de regressão linear, na qual a estimação de qualquer atributo desconhecido $Z_0 = Z(\mathbf{s}_0)$, em que $Z(\mathbf{s}_0)$ é uma observação na localização não amostrada, $\mathbf{s}_0 \in D$, é obtido por meio de uma combinação linear de n valores conhecidos $Z(\mathbf{s}_i)$. O estimador é dado por:

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^n \lambda_i Z(\mathbf{s}_i), \quad (2.41)$$

em que $\hat{Z}(\mathbf{s}_0)$ é o valor a ser estimado na posição não amostrada $\mathbf{s}_0$, λ_i são os pesos associados entre as distâncias entre os pontos vizinhos ao ponto estimados e $Z(\mathbf{s}_i)$ são os valores conhecidos da variável nos locais já amostrados.

Este preditor, apresentado na Equação (2.41), visa minimizar o erro quadrático médio da predição, sob a condição de que não seja viesado (SCHABENBERGER; PIERCE, 2001). A condição é satisfeita, quando

$$E[Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0)] = 0 \Rightarrow E[Z(\mathbf{s}_0)] = E[\hat{Z}(\mathbf{s}_0)]. \quad (2.42)$$

Segundo Schabenberger e Pierce (2001), na krigagem ordinária a única restrição colocada nos pesos de krigagem (λ_i), é que:

$$\sum_{i=1}^n \lambda_i = 1. \quad (2.43)$$

que garante a condição de que não seja viesado.

Dada as condições anteriores, espera-se que a variância do erro seja mínima, isto é

$$Var[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = E[(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2] - \{E[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)]\}^2, \quad (2.44)$$

em que o primeiro termo corresponde ao erro quadrático médio e o segundo termo ao viés. Assim, como o preditor é não viesado, ou seja, o viés é igual a zero, a Equação (2.44) pode ser reescrita por

$$Var[\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0)] = E[(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2]. \quad (2.45)$$

Por fim, temos que a expressão anterior é dada por

$$E[(\hat{Z}(\mathbf{s}_0) - Z(\mathbf{s}_0))^2] = 2 \sum_{i=1}^n \lambda_i \gamma(\delta_{i0}) - \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j \gamma(\mathbf{s}_{ij}). \quad (2.46)$$

Minimizar a expressão da Equação (2.46) pode ser uma tarefa complicada, no entanto, este problema pode ser resolvido utilizando o multiplicador de Lagrange (α) (ISAACS; SRIVASTAVA, 1989; SCHABENBERGER; PIERCE, 2001), tornando um problema de minimização restrita em um problema de minimização irrestrita:

$$\begin{cases} \sum_{i=1}^n \lambda_i \gamma_{ij} + \alpha &= \gamma_{i0} \\ \sum_{i=1}^n \lambda_i &= 1, \end{cases} \quad (2.47)$$

em que $\gamma_{ij} = \gamma(\mathbf{s}_{ij})$, sendo $\mathbf{s}_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$. O sistema de krigagem ordinária pode ser escrito em forma matricial como $\boldsymbol{\gamma} \boldsymbol{\lambda} = \mathbf{A}$ (ISAACS; SRIVASTAVA, 1989; CRESSIE, 1993), com solução

$$\boldsymbol{\lambda} = \boldsymbol{\gamma}^{-1} \mathbf{A}, \quad (2.48)$$

em que

$$\boldsymbol{\gamma} = \begin{bmatrix} \gamma(s_{11}) & \gamma(s_{12}) & \dots & \gamma(s_{nn}) & 1 \\ \gamma(s_{21}) & \gamma(s_{22}) & \dots & \gamma(s_{2n}) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(s_{n1}) & \gamma(s_{n2}) & \dots & \gamma(s_{nn}) & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} \gamma(s_{10}) \\ \gamma(s_{20}) \\ \vdots \\ \gamma(s_{n0}) \\ 1 \end{bmatrix} \text{ e } \boldsymbol{\lambda} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \alpha \end{bmatrix}. \quad (2.49)$$

Obtemos os coeficientes $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)^\top$ da Equação (2.48), que são dados por

$$\boldsymbol{\lambda}^\top = \left(\boldsymbol{\gamma} + \mathbf{1} \frac{\mathbf{1} - \mathbf{1}^\top \mathbf{A}^{-1} \boldsymbol{\gamma}}{\mathbf{1}^\top \mathbf{A}^{-1} \mathbf{1}} \right) \mathbf{A}^{-1} \quad (2.50)$$

e

$$\boldsymbol{\alpha} = - \left(\frac{\mathbf{1} - \mathbf{1}^\top \mathbf{A}^{-1} \boldsymbol{\gamma}}{\mathbf{1}^\top \mathbf{A}^{-1} \mathbf{1}} \right), \quad (2.51)$$

em que $\boldsymbol{\gamma} \equiv (\gamma(s_0 - s_1), \dots, \gamma(s_0 - s_n))^\top$ e $\mathbf{A}$ uma matriz $n \times n$ na qual o (i, j) -ésimo elemento é $\gamma(\mathbf{s}_i - \mathbf{s}_j)$.

Ainda, a variância da krigagem associada a uma estimação por krigagem ordinária, segundo Cressie (1993), é dada por:

$$\sigma_0^2 = \sum_{i=1}^n \lambda_i \gamma_{i0} + \boldsymbol{\alpha}. \quad (2.52)$$

A variância da krigagem, também pode ser escrita em termos da função de covariância, representada por:

$$\sigma_0^2 = C(0) - \boldsymbol{\lambda}^\top \mathbf{C}_0 + \boldsymbol{\alpha}, \quad (2.53)$$

em que $\mathbf{C}_0 = (C(\delta_{10}), \dots, C(\delta_{n0}))^\top$ e $\boldsymbol{\lambda}^\top$ é dado por:

$$\boldsymbol{\lambda}^\top = \left(\mathbf{C}_0 + \mathbf{1} \frac{\mathbf{1} - \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{C}_0}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}} \right) \boldsymbol{\Sigma}^{-1}, \quad (2.54)$$

e $\boldsymbol{\alpha}$ é representado por:

$$\boldsymbol{\alpha} = - \left(\frac{\mathbf{1} - \mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{c}}{\mathbf{1}^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}} \right). \quad (2.55)$$

De acordo com Wackernagel (2013), a krigagem ordinária é um interpolador exato, no sentido que, se um ponto a ser amostrado é idêntico a localização dos dados, o valor estimado será idêntico ao valor dos dados nesse ponto. Para ser mais específico, a krigagem se preocupa com a previsão de uma parte de um processo estocástico a partir de observações de outras partes (CRESSIE, 1990).

2.4.2 Krigagem Universal

Assim como na krigagem ordinária, a média do campo aleatório não é conhecida, mas pode ser expressa por um modelo linear (SCHABENBERGER; PIERCE, 2001). A krigagem universal é um interpolador utilizado para produzir mapas a partir de pontos que podem estar espaçados irregularmente, levando em conta a tendência (WARNES, 1986), e é também um método de interpolação linearmente não viesado (MATHERON, 1969). Da mesma maneira, utiliza-se para predições de valores de Z em locais s_0 não amostrados, no entanto, o que distingue os métodos são os tratamentos para $\mu(s)$. Na krigagem ordinária, $\mu(s)$ é uma constante simples desconhecida enquanto na krigagem universal a média varia, assim, utiliza-se uma função para estimar $\mu(s)$ e os coeficientes da equação de regressão que descreve esta tendência são desconhecidos e diferentes para cada ponto de predição. Quando identificada uma tendência (*drift*), que é uma variação no espaço, a suposição de estacionaridade não é válida, ou seja, a média não é constante. A técnica da krigagem universal incorpora funções simples das coordenadas que levam em consideração a tendência (OLIVER; WEBSTER, 2015).

Como citado anteriormente, na krigagem ordinária assume-se média constante, representada por $E[Z(s)] = \mu$. Contudo, na krigagem universal a média depende da localização espacial (SHERMAN, 2011). Então, segundo Christensen *et al.* (1992, 1993) a média é dada por:

$$\mu(s) = \sum_{k=1}^p \beta_k f_k(s), \quad (2.56)$$

em que β_k é a tendência local, $f_k(s)$ são funções do local de coordenadas, e $s \in D \subset \mathbb{R}^d (d \geq 1)$. Obtém-se o valor estimado em s_0 , a partir de

$$\hat{Z}(s_0) - \sum_{k=1}^p \beta_k f_k(s_0) = \sum_{u=1}^n \lambda_u \left[Z(s_u) - \sum_{k=1}^p \beta_k f_k(s_u) \right], \quad (2.57)$$

ou seja,

$$\hat{Z}(s_0) = - \sum_{u=1}^n \lambda_u Z(s_u) + \sum_{k=1}^p \beta_k \left[f_k(s_0) - \sum_{u=1}^n \lambda_u f_k(s_u) \right]. \quad (2.58)$$

No entanto, obtém-se os coeficientes e derivadas de β_k , sob a restrição de que

$$\sum_{u=1}^n \lambda_u f_k(s_u) = f_k(s_0), \quad k = 1, 2, \dots, p. \quad (2.59)$$

Conforme Armstrong (1984) a krigagem universal baseia-se na decomposição do fenômeno em estudo que leva em consideração uma tendência e suas flutuações correlacionadas.

Ela depende apenas da existência da função de covariância ou do variograma (CHRISTENSEN, 1990).

2.5 MÉTODOS DE SELEÇÃO DE MODELOS

Após a realização das análises, a verificação da bondade de ajuste dos modelos faz-se necessária. Os métodos de seleção dos modelos comparam valores obtidos na amostragem com os valores teóricos do modelo espacial escolhido. Entre os principais critérios para seleção de modelos está a validação cruzada, apresentada a seguir.

2.5.1 Validação Cruzada

O método de validação cruzada é uma técnica que permite comparar valores estimados e verdadeiros, ou seja, avalia erros de estimativas, utilizando somente as informações disponíveis no conjunto de dados da amostra (ISAAKS; SRIVASTAVA, 1989). A função da validação cruzada é evitar erros e destacar dados potencialmente problemáticos (CRESSIE, 1993). Na prática, o método consiste em avaliar cada ponto da amostra individualmente, seleciona-se um ponto em uma localização $Z(\mathbf{s}_i)$ que é temporariamente descartado do conjunto de dados, e estima-se o seu valor por krigagem utilizando os demais valores conhecidos. Após, compara-se o valor estimado com o valor verdadeiro e repete-se este procedimento para todos os valores da amostra.

Conforme McBratney e Webster (1986) e Cressie (1993) pode-se utilizar as medidas de erro médio, erro médio reduzido, desvio-padrão dos erros médios, desvio-padrão dos erros médios reduzidos e do erro absoluto, como ferramentas para avaliar modelos. Essas medidas podem ser obtidas através das seguintes equações (FARACO *et al.*, 2008).

O erro médio (EM) por validação cruzada é dado por:

$$EM = \frac{1}{n} \sum_{i=1}^n (Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_{(i)})), \quad (2.60)$$

em que n é o numero de amostras, $Z(\mathbf{s}_i)$ o valor observado no ponto $\mathbf{s}_i$ e $\hat{Z}(\mathbf{s}_{(i)})$ o valor predito por krigagem no ponto $\mathbf{s}_i$, sem considerar o valor conhecido de $Z(\mathbf{s}_i)$.

O erro médio reduzido ($\overline{ER}$) é determinado por:

$$\overline{ER} = \frac{1}{n} \sum_{i=1}^n \frac{(Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_{(i)}))}{\sigma(\hat{Z}(\mathbf{s}_{(i)}))}, \quad (2.61)$$

em que $\sigma(\hat{Z}(\mathbf{s}_{(i)}))$ é o desvio-padrão da krigagem no ponto $\mathbf{s}_i$, sem considerar o valor conhecido de $Z(\mathbf{s}_i)$. O desvio-padrão dos erros reduzidos é denotado por:

$$S_{ER} = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_{(i)}))^2}{\sigma(\hat{Z}(\mathbf{s}_{(i)}))}}. \quad (2.62)$$

O erro absoluto é definido por:

$$\widehat{EA} = \sum_{i=1}^n |Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_{(i)})|. \quad (2.63)$$

Segundo Faraco *et al.* (2008) e Mello *et al.* (2005), satisfazendo a condição de não tendenciosidade, os valores para o erro médio reduzido devem ser zero e do desvio-padrão do erro reduzido iguais a 1. O erro absoluto é definido a partir do conhecimento dos valores medidos $Z(\mathbf{s}_i)$ e preditos por krigagem $\hat{Z}(\mathbf{s}_{(i)})$. O erro absoluto é a diferença entre o valor conhecido e o estimado. Deseja-se o modelo que apresentar os menores valores para as medidas de erro médio, erro médio reduzido e desvio padrão do erro médio. Considerando a medida de desvio padrão do erro médio reduzido, espera-se que tenha o valor próximo de um. Outro critério que pode ser utilizado para validação do modelo é o máximo valor do logaritmo da função de verossimilhança (LMV). Nesse caso deseja-se o modelo que apresentar o maior valor.

3 DIAGNÓSTICOS DE INFLUÊNCIA

Uma das etapas no decorrer da análise do processo de ajustes de modelos é identificar possíveis observações influentes, que podem causar alguma distorção nos resultados do ajuste (ORTEGA *et al.*, 2003). Estas observações influentes, podem afetar negativamente como os dados se encaixam em um modelo. Para realizar este procedimento, existem diferentes métodos que avaliam a presença e influência que estas observações podem causar na análise de um conjunto de dados e nos parâmetros estimados do modelo. Uma das técnicas mais conhecidas, é a retirada destas observações influentes. Desse modo, podemos avaliar o impacto causado nas estimativas dos parâmetros.

A "*distância de Cook*" desenvolvida por (COOK, 1986), tornou-se uma ferramenta de diagnóstico muito popular para identificar conjuntamente observações influentes na regressão linear e não linear, tendo como principal característica o fato de poder ser aplicada a qualquer modelo paramétrico (ASSUMPÇÃO *et al.*, 2014). No entanto, a retirada individualizada de observações pode acarretar em um problema, resultando na não detecção de pontos conjuntamente discrepantes. Em virtude disso, existe alguns procedimentos que visam a retirada de múltiplos pontos, porém são menos convencionais devido ao alto custo computacional. A influência local, proposta por Cook (1986) na área de diagnósticos, tem por objetivo avaliar a influência conjunta das observações mediante pequenas perturbações no modelo, ao invés da avaliação pela retirada individual ou conjunta de observações.

3.1 INFLUÊNCIA LOCAL

Um dos métodos de diagnósticos mais utilizados na literatura é o método de influência local proposto por Cook(1986). Este método tem como propósito a verificação de pequenas perturbações que podem causar grandes distorções nos resultados do ajuste (GARCIA-PAPANI *et al.*, 2018). Considera-se $\mathbf{Z} = (z_1, z_2, \dots, z_n)^\top$ um vetor de observações geradas por um modelo. A perturbação no modelo é dada através de um vetor de perturbações ω , e como consequência, considera-se o modelo perturbado.

Para avaliar a influência local, Cook (1986) utiliza a medida de afastamento da verossimilhança, que é dada por:

$$LD(\omega) = 2\{L(\hat{\theta}) - L(\hat{\theta}|\omega)\}, \quad (3.1)$$

em que $LD(\omega) \geq 0$ e $\hat{\theta}$ e $\hat{\theta}_\omega$ são as estimativas de máxima verossimilhança dos modelos não perturbado $L(\theta)$ e perturbado $L(\theta|\omega)$, respectivamente. Observe que existe ω_0 tal que $LD(\omega_0) = 0$.

A fim de avaliar essa medida, verifica-se a construção de um gráfico de $LD(\omega)$ versus ω , que dispõe informações essenciais sobre a influência do esquema de perturbação empregado (GALEA *et al.*, 2003). A análise do gráfico ocorre por meio de uma superfície geométrica $(q+1)$ -dimensional formada pelo vetor

$$\alpha(\omega) = \begin{pmatrix} \omega \\ LD(\omega) \end{pmatrix}, \quad (3.2)$$

quando ω varia em Ω . Denomina-se essa superfície como gráfico de influência.

O estudo da influência local baseia-se em analisar como a superfície $\alpha(\omega)$ desvia-se de seu plano tangente T_0 em torno do ponto $\omega_0(T_0)$. Esta análise consiste em estudar as curvaturas das seções normais da superfície $\alpha(\omega)$ em ω_0 , conhecidas por curvaturas normais. Obtém-se as seções normais por meio das intersecções entre a superfície $\alpha(\omega)$ e os planos contendo o vetor normal ao seu plano tangente em ω_0 , conforme Figura 7.

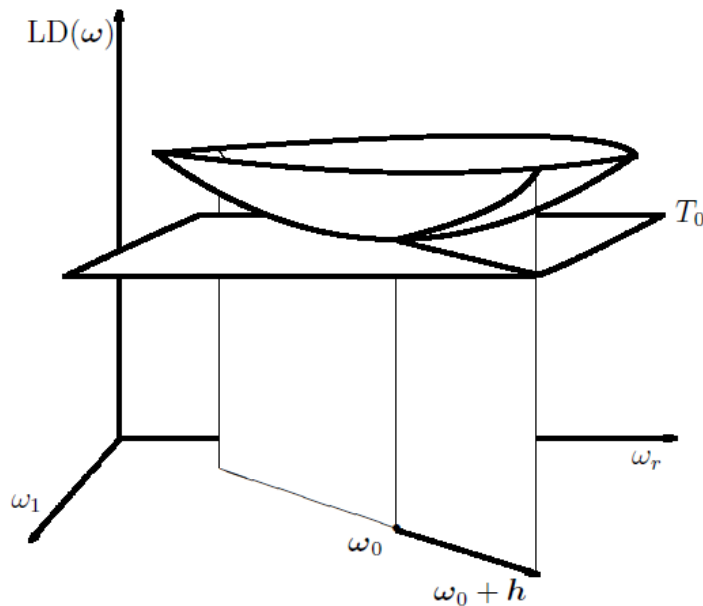


Figura 7 – Gráfico de Influência (VERBEKE; MOLENBERGHS, 2009).

Para gerar uma curvatura normal, tem-se um vetor l de comprimento unitário em uma direção qualquer em $\mathbb{R}^q$, e uma reta em $\Omega \subset \mathbb{R}^q$ que passa por ω_0 . Assim, temos:

$$\omega(a) = \omega_0 + al, \quad \text{com } a \in \mathbb{R}, \quad (3.3)$$

dessa forma, tem-se uma linha projetada sobre a superfície $\alpha(\omega)$ que passa pelo ponto $\alpha(\omega_0)$. Essa linha pode também ser obtida pelo gráfico de $LD(\omega_0 + al)$ versus $a \in \mathbb{R}$. Assim, a curvatura da normal da linha projetada, indicada por C_l , é definida como sendo a curvatura de $(a, LD\{\omega(a)\})$, sendo $a = 0$, em que $\omega(a) = \omega_0 + al$, ou seja, C_l é a curvatura da normal da superfície $\alpha(\omega)$ em torno de ω_0 e do vetor unitário l . Recomenda-se considerar a direção l_{max} correspondente à maior curvatura C_l .

Para analisar a influência local, Cook (1986) propôs o uso da curvatura normal na direção do vetor unitário l , dado por:

$$C_l(\theta) = 2|l^\top \ddot{F} l|, \quad (3.4)$$

em que $\ddot{F}$ é uma matriz quadrada de dimensão $q \times q$, e os elementos são dados por:

$$\ddot{F} = \left. \frac{\partial^2 L(\hat{\theta}|\omega)}{\partial \omega \partial \omega^\top} \right|_{\omega=\omega_0}.$$

Em notação matricial, $\ddot{F}$ pode ser escrita como:

$$\ddot{F} = J^\top \ddot{L} J, \quad (3.5)$$

em que $-\ddot{L}$ é a matriz de informação observada do modelo postulado avaliada em $\hat{\theta}$ e ω_0 , J é uma matriz $p \times q$, onde os elementos são dados por:

$$J_{ij} = \left. \frac{\partial L(\hat{\theta}_{i\omega})}{\partial \omega_j} \right|_{\omega=\omega_0},$$

sendo $\hat{\theta}_{i\omega}$ a i -ésima componente do vetor $\hat{\theta}_\omega$. Temos que J é dado por:

$$J = -(\ddot{L})^{-1} \Delta,$$

sendo Δ uma matriz $(p+q) \times n$, definida por

$$\Delta = \frac{\partial^2 L(\theta|\omega)}{\partial \theta_j \partial \omega_i},$$

avaliados em $\theta = \hat{\theta}$ e $\omega = \omega_0$, com $j = 1, 2, \dots, p$ e $i = 1, 2, \dots, q$, respectivamente.

Assim, substituindo J na Equação 3.5, obtém-se:

$$\begin{aligned} \ddot{F} &= J^\top \ddot{L}^{-1} J \\ &= (-\Delta \ddot{L}^{-1}) \ddot{L} (-\ddot{L}^{-1} \Delta) \\ &= \Delta^\top \ddot{L}^{-1} \Delta. \end{aligned} \quad (3.6)$$

Aplicando o resultado da Equação (3.6) na Equação (3.4), obtém-se a seguinte forma da curvatura normal para o gráfico de influência $\alpha(\omega)$, na direção de um vetor unitário l ,

$$C_l(\theta) = 2|l^\top \Delta^\top \ddot{L}^{-1} \Delta l|, \quad (3.7)$$

em que $\|l\| = 1$.

Segundo Cook (1986), os componentes individuais de (3.7) podem ser obtidos de maneira simples uma vez que o esquema de perturbação tenha sido definido. Cook (1986) ainda ressalta que existem diversas maneiras de estudar α_ω na prática. Entre elas, os extremos $C_{max} = \max C_l$ e $C_{min} = \min C_l$, para $l = 1, 2, \dots, n$, que correspondem aos autovalores absolutos máximo e mínimo respectivamente. Deseja-se investigar a direção que produza a maior influência local.

Geralmente se utiliza a máxima curvatura C_{max} e L_{max} para identificar as observações localmente influentes, e o vetor L_{max} possibilita identificar os fatores mais influentes. O gráfico de C_i é um indicador útil para problemas locais, mas pode não detectar informações globais, portanto, uma inspeção no gráfico de L_{max} pode ser necessária (COOK, 1986).

3.1.1 Influência Local na Variável Resposta

Nesta seção será apresentada a principal contribuição deste trabalho, o resultado que será tratado aqui ainda não se encontra na literatura. Foi considerado como método para avaliar a influência local na variável resposta um esquema de perturbação baseado no esquema de perturbação de Zhu *et al.* (2007), que será apresentado a seguir.

3.1.1.1 Perturbação de Zhu

Nosso interesse é encontrar o resultado da matriz de perturbação utilizando a função de verossimilhança restrita, a fim de verificar se este resultado mostra-se mais adequado do que quando consideramos a função de máxima verossimilhança. Resultado que já é conhecido na literatura e que a sua utilização pode ser vista por exemplo em Uribe-Opazo *et al.* (2012).

Para isso precisamos encontrar a segunda derivada da Equação 3.8. Abaixo então é apresentado o logaritmo da função de verossimilhança perturbada restrita.

Seja o logaritmo da função de verossimilhança perturbada restrita dado por:

$$l(\theta|\omega) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma| - \frac{1}{2} (u_\omega), \quad (3.8)$$

em que $u_{\omega} = (\mathbf{Z}_{\omega} - \mathbf{X}\boldsymbol{\beta})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{Z}_{\omega} - \mathbf{X}\boldsymbol{\beta})$, $\mathbf{Z}_{\omega} = \mathbf{Z} + \boldsymbol{\Sigma}^{1/2} \boldsymbol{\omega}$ com $\boldsymbol{\omega} = (\omega_1, \omega_2, \dots, \omega_n)^{\top}$ o vetor de perturbação das respostas pertencente a um espaço de perturbações $\Omega (\boldsymbol{\omega} \in \Omega)$.

Obter o resultado final da Equação 3.9 não é uma tarefa simples, uma vez que é necessário o conhecimento de algumas importantes regras de diferenciação de matrizes para que se possa encontrar a solução correta.

Sendo assim, a matriz Δ_{ϕ} é obtida através da solução da diferenciação:

$$\Delta_{\phi} = \frac{\partial^2 L(\theta|\omega)}{\partial \phi \partial \omega^{\top}}. \quad (3.9)$$

Dessa forma, a matriz Δ de perturbação encontrada baseada na perturbação de Zhu *et al.* (2007) é dada por:

$$\Delta_{\phi_j} = (\mathbf{Z}_{\omega} - \mathbf{X}\boldsymbol{\beta})^{\top} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}^{1/2}}{\partial \phi_j} - (\mathbf{Z}_{\omega} - \mathbf{X}\boldsymbol{\beta})^{\top} \boldsymbol{\Sigma}^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \phi_j} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^{1/2}, \quad (3.10)$$

sendo Δ_{ϕ_j} elementos da matriz Δ_{ϕ} .

Encontrado o resultado da Equação 3.10, foi realizado a implementação deste resultado no Software R para investigar a melhoria nos resultados quando comparado a utilização da máxima verossimilhança restrita *versus* máxima verossimilhança.

A seguir no Capítulo 4, é apresentado um estudo de simulação mostrando os bons resultados quando utilizamos a matriz Δ da Equação 3.10, motivando o uso da máxima verossimilhança restrita. Após, no Capítulo 5, apresentamos os resultados comparativos quando usa-se máxima verossimilhança e máxima verossimilhança restrita.

4 SIMULAÇÕES

Para avaliar o desempenho da abordagem apresentada neste trabalho na detecção de observações potencialmente influentes, foram realizados estudos de simulação. Assim, para a simulação, foi considerada a distribuição normal multivariada e os parâmetros foram estimados via método de estimação de máxima verossimilhança restrita. Foram considerados os modelo exponencial, Gaussiano e da família Matérn com diferentes valores de κ para realizar o ajuste. Os estudos de simulação foram realizados utilizando o *software* R (R Core Team, 2018) e as análises geoestatísticas utilizando o pacote geoR (Ribeiro Jr; DIGGLE, 2016).

Sendo assim, o esquema de simulação foi realizado da seguinte maneira:

1. Foram geradas n observações de uma normal multivariada considerando média zero e a estrutura espacial para a matriz Σ , representada na Equação (2.7), considerando os parâmetros ϕ_1 , ϕ_2 e ϕ_3 .
2. As observações geradas foram alocadas em coordenadas anteriormente já definidas pela malha que foi utilizada para gerar os cenários.
3. Foi montado um esquema de perturbação no qual a observação gerada de maior valor Z recebeu uma $Z = Z + \xi \times \text{sd}(Z)$, variando o parâmetro de perturbação em $\xi = 0, 2, 4$.
4. Assim, os parâmetros da matriz de covariância σ foram estimados via máxima verossimilhança restrita. Após esta etapa, foi verificado se o gráfico de influência local identificou a observação que foi perturbada.
5. O processo foi repetido $N = 1000$ vezes, e verificou-se quantas vezes o gráfico de influência local identificou a observação perturbada.

Inicialmente, foi considerada a grade amostral dos dados de produtividade de soja, que será apresentado no Capítulo 5, conforme Figura 8. Neste caso, foram geradas $n = 98$ observações nas coordenadas já existentes, conforme detalha o esquema de perturbação.

Na Tabela 1 são apresentados os resultados da simulação considerando o primeiro cenário, em que o parâmetro de efeito pepita $\phi_1 = 0$, contribuição $\phi_2 = 1$ e alcance $\phi_3 = 75$, o valor do parâmetro de alcance ϕ_3 , foi escolhido a partir do conhecimento prévio baseado na análise dos dados reais que serão apresentados no Capítulo 5. Para este cenário foram considerados os modelos exponencial, Gaussiano e da família Matérn para diferentes valores de κ . Pode-se verificar que, quando foi considerado o parâmetro de não perturbação $\xi = 0$, a taxa de acerto foi baixa principalmente quando considerado o modelo da família Matérn. À medida que o parâmetro de perturbação foi aumentado, a taxa de acerto foi maior do que 90% na maioria

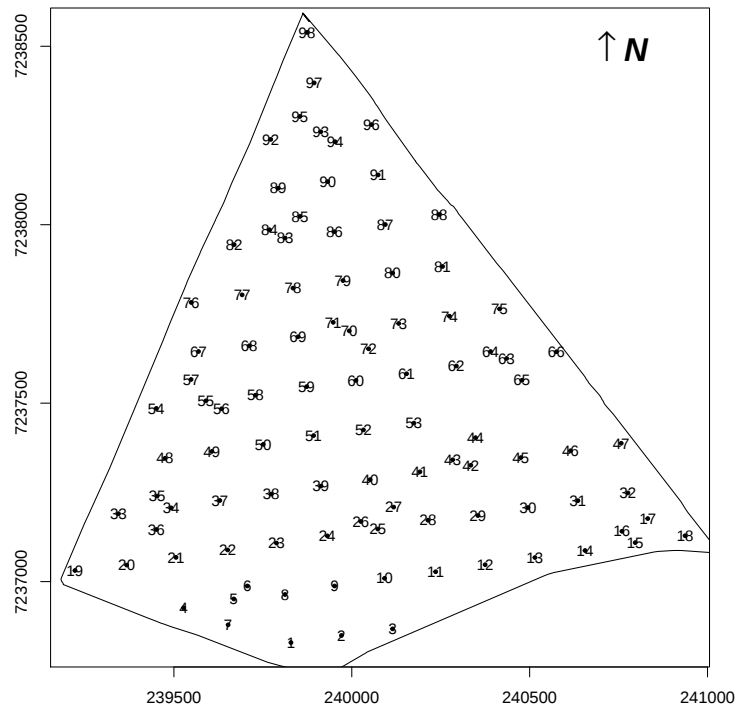


Figura 8 – Malha amostral utilizada para realização das simulações.

dos casos.

Tabela 1 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0$, $\phi_2 = 1$ e $\phi_3 = 75$ para a malha dos dados de produtividade de soja

Modelo	Parâmetro de Perturbação (ξ)		
	0	2	4
Exponencial	80.8%	97.3%	97.5%
Gaussiano	78.4%	99.4%	99.6%
Matérn ($\kappa = 0.7$)	67.4%	97.4%	97.9%
Matérn ($\kappa = 1$)	43.5%	94.1%	97.1%
Matérn ($\kappa = 1.5$)	19.2%	72.0%	87.5%

A Tabela 2 apresenta os resultados considerando um segundo cenário em que foi adicionado um valor para o parâmetro de efeito pepita, $\phi_1 = 0.1$. Os modelos exponencial e Gaussiano apresentaram os melhores resultados. No entanto, percebe-se que ao considerar o parâmetro de perturbação $\xi = 4$, a taxa de acerto teve um pequeno decréscimo.

Tabela 2 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0.1$, $\phi_2 = 1$ e $\phi_3 = 75$ para a malha dos dados de produtividade de soja

Modelo	Parâmetro de Perturbação (ξ)		
	0	2	4
Exponencial	81.6%	99.6%	99.5%
Gaussiano	83.3%	99.6%	99.5%
Matérn ($\kappa = 0.7$)	71.6%	97.8%	97.7%
Matérn ($\kappa = 1$)	54.7%	95.8%	97.6%
Matérn ($\kappa = 1.5$)	27.0%	82.9%	91.9%

Para o estudo de simulação, também foi considerado uma malha regular 10×10 . Onde, neste caso, foram geradas $n = 100$ observações de uma normal multivariada para serem alocadas em cada ponto da malha. A malha é apresentada na Figura 9.

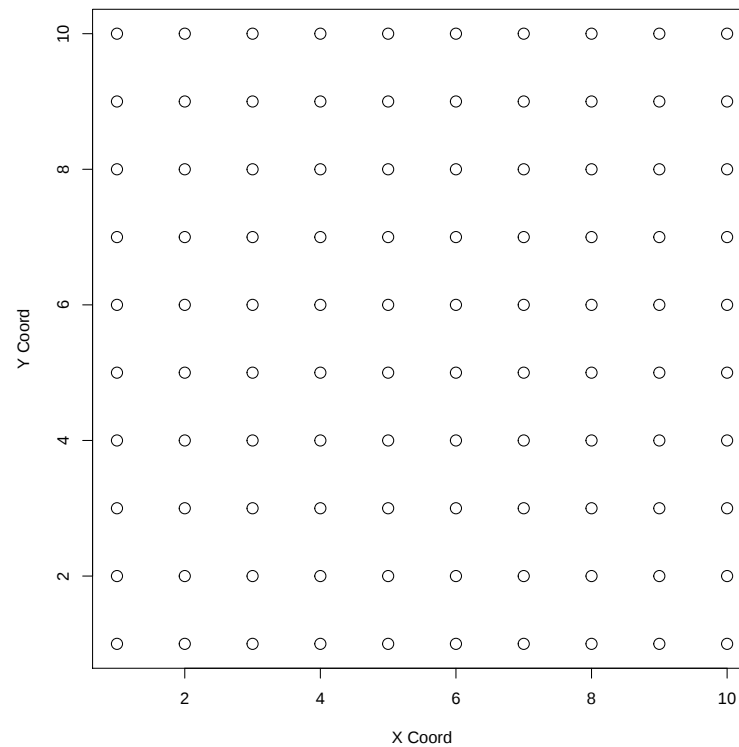


Figura 9 – Malha amostral 10×10 utilizada para realização das simulações.

Na Tabela 3 são apresentados os resultados da simulação considerando o primeiro cenário para o grille de amostragem 10×10 . Nesse cenário foram considerados os seguintes

valores para os parâmetros: $\phi_1 = 0$, $\phi_2 = 1$ e $\phi_3 = 0.3$. Novamente o modelo Gaussiano apresentou a melhor taxa de acerto entre os modelos ajustados. No entanto, novamente, quando o parâmetro de perturbação $\xi = 4$, obteve-se um pequeno decréscimo na taxa de acerto.

Tabela 3 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0$, $\phi_2 = 1$ e $\phi_3 = 0.3$ para a malha regular 10×10

Modelo	Parâmetro de Perturbação (ξ)		
	0	2	4
Exponencial	89.4%	94.2%	94.6%
Gaussiano	97.4%	98.4%	98.3%
Matérn ($\kappa = 0.7$)	91.2%	96.2%	96.0%
Matérn ($\kappa = 1$)	91.0%	97.2%	97.7%
Matérn ($\kappa = 1.5$)	87.1%	97.1%	98.1%

A Tabela 4 apresenta os resultados de simulação para o segundo cenário da malha regular, em que $\phi_1 = 0.1$, $\phi_2 = 1$ e $\phi_3 = 0.3$. Mais uma vez, o modelo Gaussiano foi o que apresentou a melhor taxa de cobertura considerando todos os valores do parâmetro de perturbação ξ .

Tabela 4 – Esquema de perturbação para avaliar a taxa de identificação de observações influentes para os modelos exponencial, Gaussiano e Matérn com os parâmetros $\phi_1 = 0.1$, $\phi_2 = 1$ e $\phi_3 = 0.3$ para a malha regular 10×10

Modelo	Parâmetro de Perturbação (ξ)		
	0	2	4
Exponencial	89.2%	94.6%	95.2%
Gaussiano	97.4%	98.2%	98.7%
Matérn ($\kappa = 0.7$)	91.0%	95.9%	96.8%
Matérn ($\kappa = 1$)	91.6%	97.8%	97.2%
Matérn ($\kappa = 1.5$)	87.5%	97.6%	98.3%

Também foram consideradas malhas de tamanho 8×8 e 12×12 para o estudo de simulação. No entanto, como estes cenários obtiveram resultados semelhantes ao da malha 10×10 , não serão apresentados.

4.1 CONSIDERAÇÕES GERAIS

Ao realizar o estudo de simulação, pôde-se verificar que os modelos, considerando o método de estimação de máxima verossimilhança restrita, apresentam uma boa taxa de acerto na identificação de possíveis pontos influentes. Porém, em alguns casos, percebemos que ao considerar o parâmetro de perturbação $\xi = 4$, houve um pequeno decréscimo na taxa de identificação. Há diversos fatores que podem causar este comportamento, um deles pode ser o fato de que essa perturbação esteja causando uma grande dispersão nos dados. No entanto, é importante considerar outros cenários.

5 APLICAÇÃO A DADOS REAIS

No presente capítulo, serão apresentadas as análises de dois conjuntos de dados distintos, com o propósito de avaliar as técnicas de diagnóstico desenvolvidas neste trabalho. Inicialmente, foi realizada uma análise descritiva dos dados e em seguida realizou-se o ajuste dos modelos, bem como o uso das técnicas de diagnóstico para detecção de pontos influentes. Finalmente, elaborou-se a construção dos mapas temáticos. Posteriormente, foram realizadas as análises dos dados sem considerar as observações influentes. Todas as análises apresentadas neste trabalho foram realizadas utilizando o *software* R (R Core Team, 2018) e para as análises geoestatísticas foi utilizado o pacote *geoR* (Ribeiro Jr; DIGGLE, 2016).

5.1 DADOS JONES

Nesta seção, foi considerado um conjunto de dados com observações de 93 poços em um único aquífero próximo a Saratoga Valley, Wyoming. As alturas das águas, z , são dadas em metros acima do nível médio do mar e o tamanho do aquífero é de aproximadamente 1.300 km^2 . As coordenadas x e y são dadas em quilômetros. Este conjunto de dados pode ser encontrado em Jones (1989). A utilização deste conjunto de dados se dá pelo fato de já ser bastante explorado anteriormente na literatura, vide exemplo em Uribe-Opazo *et al.* (2012). Desse modo, a análise sobre o conjunto de dados será apresentada mais brevemente.

Na Figura 10 é apresentada a grade amostral e a distribuição dos pontos ao longo da área considerada.

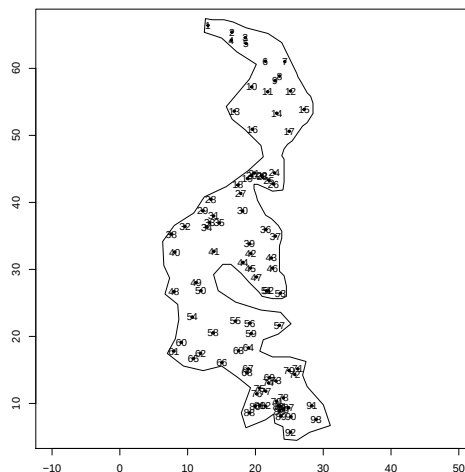


Figura 10 – Malha amostral dos dados de Jones.

Na Tabela 5 são apresentadas as estatísticas descritivas dos dados das alturas das águas dos poços no aquífero, que são dados em metros de um conjunto total de 93 observações ($n = 93$). O valor mínimo da altura dos poços é de 2001.60 m e a altura máxima é de 2254.30 m, a média igual 2126.94 m. Segundo o coeficiente de variação ($CV = 2.77\%$), de acordo com a classificação proposta por Neto (2002), dados que apresentam $CV \leq 20\%$ são poucos dispersos, portanto apresentam homogeneidade. Os coeficientes de assimetria e curtose mostram que a variável altura dos poços apresenta características levemente assimétrica positiva e platicúrtica.

Tabela 5 – Estatísticas descritivas dos dados de Jones

n	93
Mínimo.	2001.60
1º Qu.	2073
Mediana	2112.30
Média	2126.94
3º Qu.	2173.50
Max.	2254.30
Desvio Padrão	58.96
Coef. de Variação (%)	2.77
Coef. de Assimetria	0.13
Coef. Curtose	-0.98

Na Figura 11 é apresentado o gráfico *boxplot*, que pode ser utilizado para auxiliar na detecção de pontos discrepantes. No entanto, para estes dados, não foi possível identificar nenhum *outlier* utilizando apenas este gráfico como referência.

5.1.1 Ajuste do Modelo

Após a análise descritiva dos dados, foi realizado o ajuste do modelo considerando os métodos de estimação dos parâmetros de máxima verossimilhança e máxima verossimilhança restrita.

A Tabela 6 apresenta as estimativas dos parâmetros considerando o método da máxima verossimilhança (ML) e máxima verossimilhança restrita (REML). Para obtenção dos parâmetros estimados foi utilizado o modelo Gaussiano. Quando considerado o método da máxima verossimilhança, os valores estimados de efeito pepita ($\hat{\phi}_1$) foi igual a 25.7923, contribuição ($\hat{\phi}_2$) igual a 950.4528 e o alcance prático $\hat{a} = (g(\hat{\phi}_3))$ igual a 13.4121. E quando considerado o método da máxima verossimilhança restrita, obteve-se efeito pepita ($\hat{\phi}_1$) =

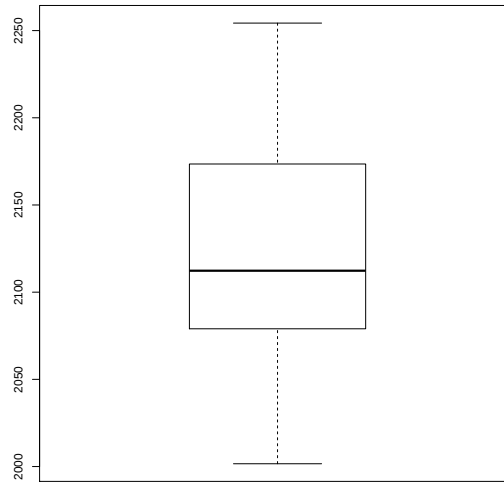


Figura 11 – Gráfico *boxplot* dos dados de Jones.

26.3557, contribuição $(\hat{\phi}_2) = 1145.6370$ e o alcance prático $\hat{a} = (g(\hat{\phi}_3)) = 13.9723$.

Pode-se observar ainda na Tabela 6 que os valores estimados por ML e REML de $\hat{\phi}_1$, $\hat{\phi}_2$ e $\hat{\phi}_3$ apresentaram resultados com valores próximos. No entanto, quando foi considerado o método da máxima verossimilhança restrita (REML), todos os valores dos parâmetros estimados tiveram um pequeno aumento. Em ambos os casos verificou-se forte dependência espacial, segundo a escala de classificação de Cambardella *et al.* (1994) onde o efeito pepita relativo (EPR) é dado por $\frac{\hat{\phi}_1}{\hat{\phi}_2 + \hat{\phi}_1}$. Considera-se dependência espacial forte quando a medida apresenta valores igual a $EPR \leq 0.25$, dependência espacial moderada para $0.25 < EPR \leq 0.75$ e dependência espacial fraca para $EPR > 0.75$.

Tabela 6 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para os dados de Jones

Modelo (Método de Estimação)	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{a} = g(\hat{\phi}_3)$	EPR
Gaussiano (ML)	2248.2280 (30.9242)	-0.6881 (1.2133)	-3.0336 (0.4974)	25.7923 (5.3874)	950.4528 (320.2882)	7.7435 (0.7143)	13.4121 (1.7320)	0.0264
Gaussiano (REML)	—	—	—	26.3557 (5.4426)	1145.6370 (392.3155)	8.0669 (0.7279)	13.9723 (1.7320)	0.0225

5.1.1.1 Critérios de Validação Cruzada

Para a escolha do melhor modelo, serão considerados os seguintes critérios de validação cruzada: erro médio (EM), erro médio reduzido ($\overline{ER}$), desvio-padrão do erro médio (SEM), desvio-padrão do erro médio reduzido (SE_R), erro absoluto (EA), logaritmo da função de

verossimilhança (LMV) e critério de informação de Akaike (AIC). Para comparar dois ou mais modelos para o mesmo conjunto de dados, utiliza-se o critério de AIC (critério de informação de Akaike), que penaliza o número de parâmetros no logaritmo da função de verossimilhança dado pela expressão:

$$AIC = 2l(\theta) + 2k \quad (5.1)$$

em que k é o número de parâmetros do modelo ajustado. Deseja-se o modelo que apresentar o menor AIC .

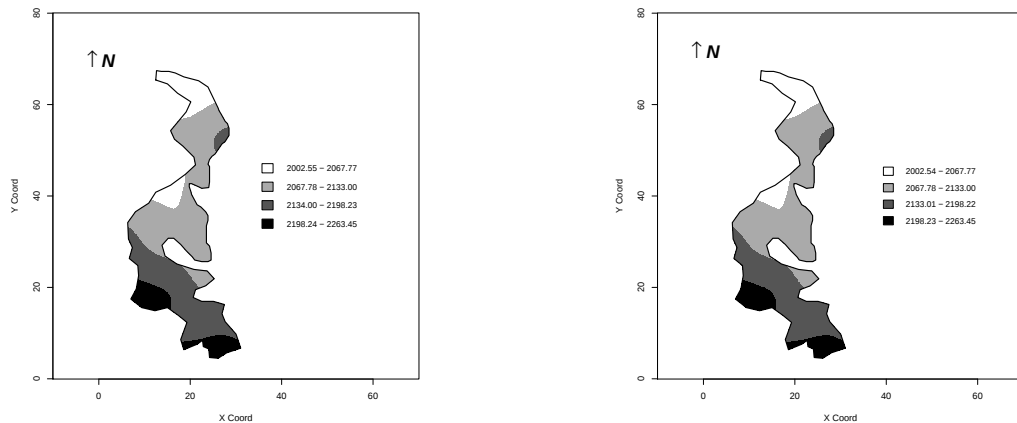
De acordo com os critérios citados anteriormente, o ajuste que obteve o melhor resultado foi quando considerado o método da máxima verossimilhança restrita. Na Tabela 7 são apresentados os critérios de validação cruzada para os ajustes, em que o estimador de máxima verossimilhança restrita apresentou melhores resultados.

Tabela 7 – Critérios para seleção do melhor ajuste entre os métodos da máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para os dados de Jones

Modelo (Método de Estimação)	EM	$\overline{ER}$	S_{EM}	S_{ER}	EA	LMV	AIC
Gaussiano (ML)	0.5502	0.0272	7.5084	1.0077	476.9456	-346.8	705.6
Gaussiano (REML)	0.5379	0.0257	7.5816	1.0054	478.1453	-330.8	673.7

Escolhido o melhor modelo ajustado, considerando o estimador de máxima verossimilhança restrita, foi gerado o mapa temático através da técnica de krigagem.

A Figura 12 apresenta os mapas temáticos para os dados de Jones para a distribuição Gaussiana, utilizando o método da máxima verossimilhança e máxima verossimilhança restrita restrita. Foi considerado quatro classes de mesma amplitude para ambos os mapas. Os mapas não apresentam diferenças perceptíveis visualmente, mas foi possível verificar que o uso do método da REML é mais adequado.



(a) Mapa de krigagem para os dados de Jones via ML (b) Mapa de krigagem para os dados de Jones via REML

Figura 12 – Mapas de krigagem para o modelo Gaussiano utilizando os métodos da ML e REML para os dados de Jones.

5.1.1.2 Diagnóstico de Influência Local

Após a escolha do melhor modelo, é necessário verificar a presença de observações atípicas que podem influenciar no resultado do ajuste. Para isso foi utilizado o método de diagnóstico de influência local, perturbação de Zhu.

Na Figura 13 são apresentados os gráficos de influência para a perturbação de Zhu considerando o modelo gaussiano. O gráfico C_i vs ordem da Figura 13a destacou as observações #49, #50, #69 e #81 que correspondem respectivamente as profundidades dos poços nos valores de 2158.9, 2121.4, 2163.8 e 2174.7 como possíveis pontos influentes na resposta. Já o gráfico $|L_{max}|$ vs ordem da Figura 13b identificou as observações #49, #50, #68, #76 e #81 que correspondem respectivamente as profundidades dos poços nos valores de 2158.9, 2121.4, 2166.8, 2155.5 e 2174.7 como influentes na resposta. A linha representada no gráfico significa um ponto de corte para a identificação de pontos influentes (ZHU; LEE, 2001).

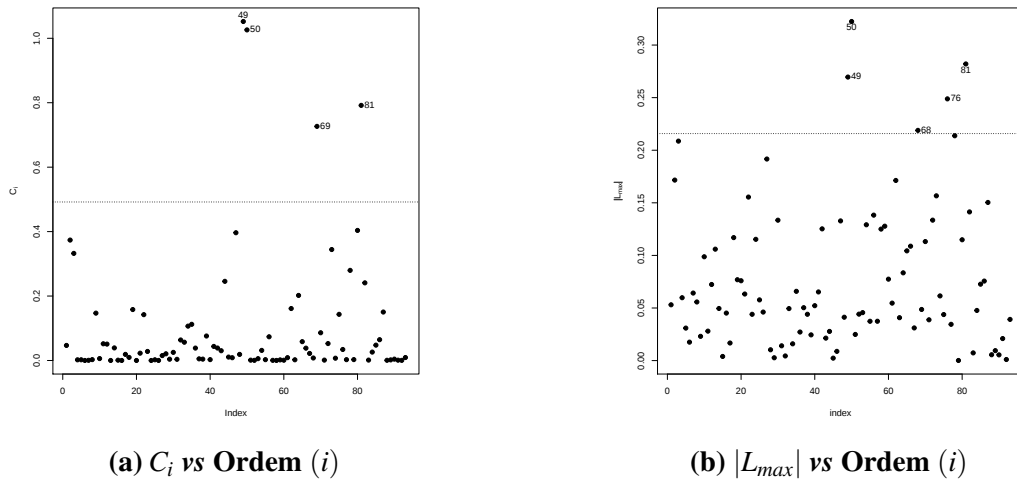


Figura 13 – Gráficos de influência para o modelo Gaussiano utilizando o método da REML para os dados de Jones.

5.1.2 Ajuste do Modelo sem Observações Influentes

Uma vez que identificamos a presença de alguns pontos influentes através do método de influência local utilizado, é necessário considerar um novo ajuste sem essas observações. Foram realizados diversos ajustes considerando a retirada desses pontos influentes, apenas os com resultados mais relevantes serão apresentados a seguir.

A análise do conjunto de dados foi realizada sem considerar as observações #49, #50, #69 e #81 para apresentar o que pode acontecer sem estes pontos na análise, e as estimativas dos parâmetros obtidas por REML são apresentadas na Tabela 8. Inicialmente realizou-se a análise sem considerar individualmente cada ponto que foi detectado como influente. Posteriormente retirou-se todos eles, e por fim, fez-se a retirada apenas dos pontos #49 e #50. Pôde-se observar, que ao eliminar todas as observações, a estimativa do parâmetro de contribuição ϕ_2 sofreu maior variação em relação as demais análises. Todos os casos apresentaram dependência espacial forte ($EPR \leq 0.25$).

Tabela 8 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança restrita (REML) para os dados de Jones

Método de Estimação	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{a} = g(\hat{\phi}_3)$	EPR^1
REML	26.3557 (5.4426)	1145.6370 (392.3155)	8.0669 (0.7279)	13.9723 (2.1837)	0.0225
REML (−#49)	19.3156 (4.1808)	1107.2460 (365.0879)	7.7247 (0.6495)	13.3796 (1.9485)	0.0171
REML (−#50)	20.0247 (4.2673)	1188.9770 (397.5703)	7.9921 (0.6763)	13.8427 (2.0289)	0.0167
REML (−#69)	23.1856 (4.9204)	1136.4800 (382.4840)	7.9068 (0.6920)	13.6950 (2.076)	0.0200
REML (−#81)	23.7326 (4.9998)	1158.1170 (392.9884)	8.0181 (0.7050)	13.8878 (2.115)	0.0201
REML (−#49, −#50)	19.6532 (4.2223)	1152.5320 (386.4998)	7.9169 (0.6787)	13.7125 2.0361	0.0168
REML (−#49, −#50, −#69, −#81)	17.9949 (3.7098)	2857.6320 (1044.6480)	9.9896 (0.7780)	17.3025 (2.334)	0.0006

$$^1 EPR = \frac{\phi_1}{\phi_1 + \phi_2}$$

5.1.2.1 Critérios de Validação Cruzada

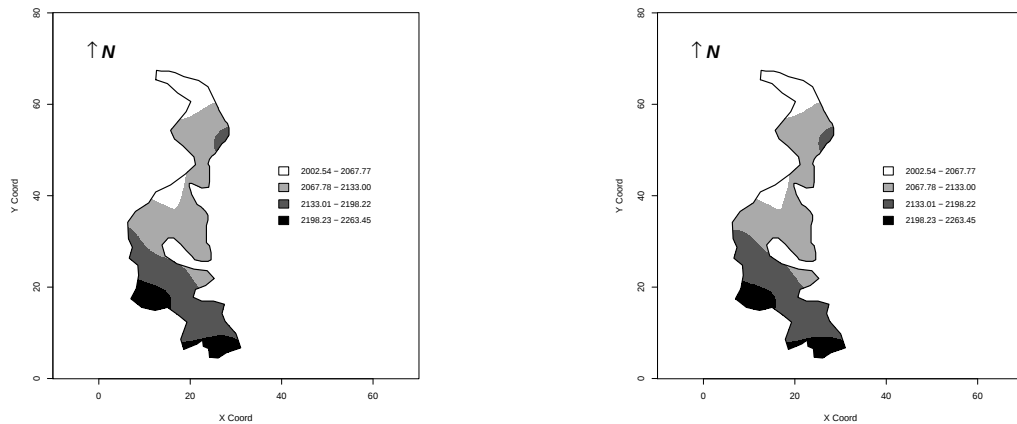
Na Tabela 9 são apresentados os critérios de validação cruzada para a escolha do melhor ajuste. Quando foi considerado a retirada da observação #49, este cenário apresentou melhores resultados para a maioria dos critérios como o erro médio, erro médio reduzido, desvio padrão do erro médio e erro absoluto.

Tabela 9 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança restrita (REML) para os dados de Jones

Modelo (Método de Estimação)	EM	$\overline{ER}$	S_{EM}	S_{ER}	EA	LMV	AIC
Gaus (REML)	0.5379	0.0257	7.5816	1.0054	478.1453	−330.8	673.7
Gaus (REML) (−#49)	0.5103	0.0256	6.6968	0.9684	418.1480	−321.4	654.8
Gaus (REML) (−#50)	0.5399	0.0280	6.8497	0.9768	435.3742	−321.2	654.5
Gaus (REML) (−#69)	0.6002	0.0305	7.4355	1.0122	453.9374	−325.3	662.6
Gaus (REML) (−#81)	0.5986	0.0302	7.4845	1.0152	461.4937	−325.4	662.8
Gaus (REML) (−#49, −#50)	0.5356	0.0273	6.6278	0.9659	413.8841	−317.2	646.4
Gaus (REML) (−#49, −#50, −#69, −#81)	0.6188	0.0304	7.5462	1.0238	435.1979	−308.5	628.9

Considerando o melhor modelo ajustado retirando as observações influentes, foi gerado um novo mapa temático, a fim de comparação com o mapa gerado anteriormente. A

Figura 14 apresenta os mapas temáticos para a distribuição Gaussiana considerando quatro classes de mesma amplitude. Na Figura 14a é apresentado o mapa temático com todos os pontos e na Figura 14b, o mapa temático sem o ponto #49. Ambos os mapas foram construídos com a mesma escala, e visualmente é possível notar as diferenças, ainda que pequenas.

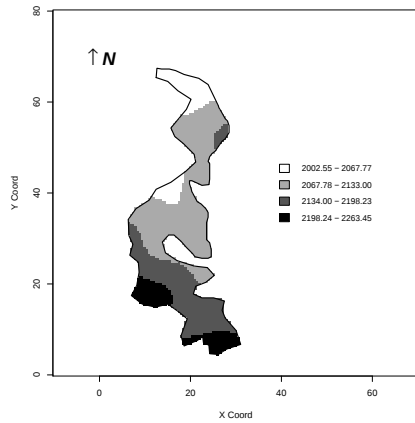


(a) Mapa de krigagem REML

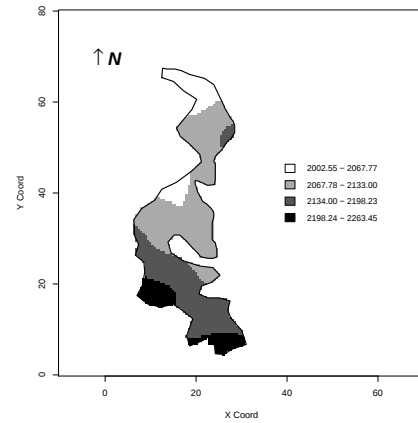
(b) Mapa de Krigagem REML (-#49)

Figura 14 – Mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de Jones.

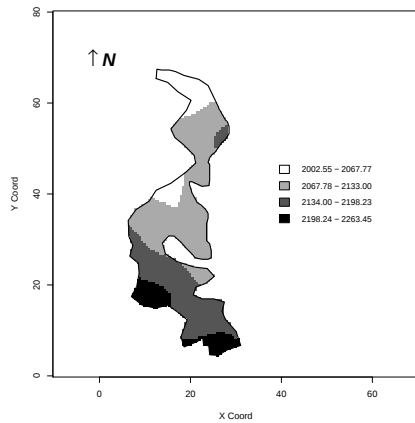
Na Figura 15 são apresentados os demais mapas de krigagem referentes aos diferentes modelos ajustados considerando a retirada dos diferentes pontos potencialmente influêntes que foram identificados pelos gráficos de influência mostrados na Figura 13, e posteriormente submetidos aos critérios de seleção do melhor ajuste na Tabela 9.



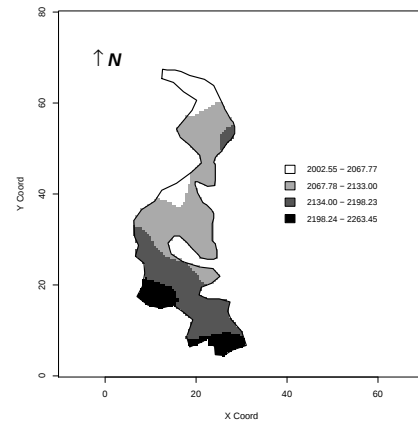
(a) Mapa de krigagem REML (-#50)



(b) Mapa de Krigagem REML (-#69)



(c) Mapa de Krigagem REML (-#81)



(d) Mapa de Krigagem REML (-#49, -#50)

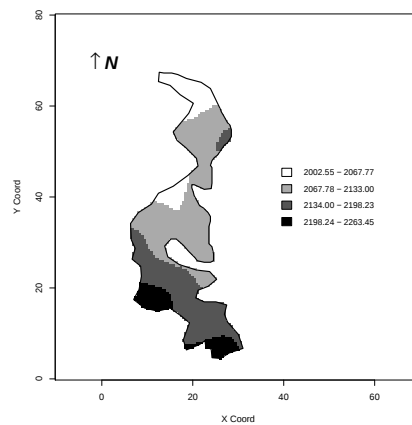
(e) Mapa de Krigagem REML
(-#49, -#50, -#69, -#81)

Figura 15 – Demais mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de Jones.

Na Tabela 10 é apresentada a matriz de erros que fornece a quantidade de área em km^2 de cada mapa temático construído apresentados na Figura 14, bem como as áreas coincidentes para a comparação dos mapas temáticos com e sem o ponto influente. Pôde-se

observar que a segunda classe foi a que apresentou maior quantidade de área na mesma região entre os mapas, seguido da terceira classe com 583.52 km^2 e 378.68 km^2 , respectivamente.

Tabela 10 – Matriz de erros da quantidade de área em km^2 para os dados de Jones

Mapa com todos os pontos	Mapa sem o ponto #49				Total
	2002.54 – 2067.77	2067.78 – 2133.00	2133.01 – 2198.22	2198.23 – 2263.45	
2002.54 – 2067.77	156.85	4.12	0	0	160.97
2067.78 – 2133.00	0.44	583.52	21.33	0	605.29
2133.01 – 2198.22	0	1.77	378.68	0.04	380.49
2198.23 – 2263.45	0	0	5.93	147.32	153.25
Total	157.29	589.41	405.94	147.36	1.300

Para medir a semelhança entre os mapas temáticos dos dados de Jones, foi considerado o cálculo do índice de exatidão Kappa (K) (COHEN, 1960). Segundo Krippendorff (2004), como o índice apresentou o valor de $K = 0.9612$, os mapas apresetam alta exatidão. O intervalo de confiança (IC) de 95% obtido para essa medida foi de $[0.9594, 0.9630]$.

5.1.3 Considerações Gerais

De acordo com as análises realizadas utilizando métodos geoestatísticos foi possível verificar a variabilidade espacial dos dados em estudo. Inicialmente consideramos os estimadores de máxima verossimilhança e máxima verossimilhança restrita para estimar os parâmetros do modelo Gaussiano considerando todas as observações. Observamos que os valores dos parâmetros estimados, para ambos métodos de estimação, apresentaram valores muito semelhantes, conforme pode ser verificado na Tabela 6. Porém, ao realizar o método de validação cruzada, o modelo considerando o estimador de máxima verossimilhança restrita apresentou melhor ajuste.

Aplicando o método de influência local, observamos a presença de quatro observações influentes no modelo, as observações #49, #50, #69 e #81. Ao realizar o ajuste sem considerar estas observações verificamos que o modelo retirando apenas a observação #49 apresentou o melhor ajuste. A fim de comparação, apresentamos os mapas temáticos do modelo considerando todas as observações e do modelo retirando a observação #49.

5.2 DADOS DE PRODUTIVIDADE DE SOJA

Os dados que compõem o estudo, nessa seção, são referentes a produtividade de soja no ano agrícola de 2016/2017 em uma propriedade localizada na cidade de Cascavel-PR, Brasil. A medida das variáveis é representada por toneladas por hectare (*ton/ha*). A grade amostral da área em estudo é de 167.35 *ha*, com coordenadas geográficas aproximadas de latitude 24°57'21"S e longitude 53°34'32"W, coletadas por meio de um receptor de sinal GPS (*Global Positioning System*), com *datum* WGS84. Na Figura 16 é apresentada a grade amostral e a distribuição dos pontos ao longo da área considerada, em que as observações circuladas representam as observações de menor valor, identificadas como observações discrepantes no conjunto de dados, conforme pode ser visto na Figura 17.

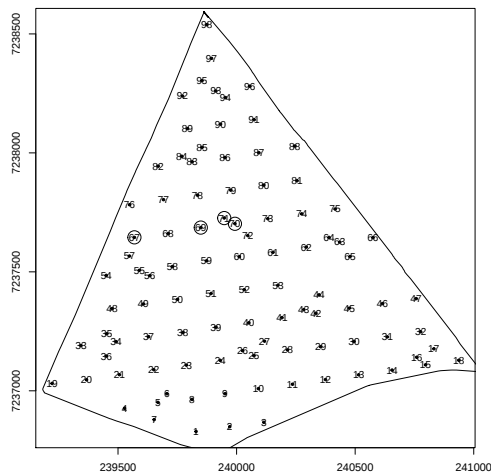


Figura 16 – Malha amostral dos dados de produtividade de soja.

Na Tabela 11 são apresentadas estatísticas descritivas dos dados da produtividade de soja, em que $n = 98$ corresponde ao número de observações. O valor mínimo é de 1.52 (*ton/ha*) e o máximo de 4.20 (*ton/ha*), a média igual a 3.12 (*ton/ha*). Conforme o coeficiente de variação ($CV=17.34\%$), de acordo com a classificação proposta por Neto (2002), dados que apresentam $CV \leq 20\%$ são poucos dispersos, portanto apresentam homogeneidade. Segundo os coeficientes de assimetria e curtose a variável produtividade têm características levemente assimétrica negativa e leptocúrtica.

Tabela 11 – Estatísticas descritivas dos dados de produtividade de soja

n	98
Mínimo	1.52
1º Quartil.	2.85
Mediana	3.20
Média	3.12
3º Quartil	3.46
Máximo	4.20
Desvio Padrão	0.54
Coef. de Variação (%)	17.34
Coef. de Assimetria	−0.73
Coef. Curtose	0.55

Na Figura 17 é apresentado o gráfico *boxplot*, onde podemos observar a presença de quatro observações discrepantes, sendo elas as observações #67, #69, #70 e #71. Verificou-se que o menor destes é o elemento amostral #70 sendo seu valor medido igual a 1.518 (mínimo), e de acordo com o gráfico da malha apresentado na Figura 16, ele está cercado pelos pontos #69, #71, #72 e #73. Estes pontos apresentaram valores de produtividade de soja variando de 1.5 à 2.1 (*ton/ha*). Podemos considerar que estes pontos são candidatos a serem potencialmente influentes.

**Figura 17 – Box-plot dos dados de produtividade de soja.**

Na Figura 18 é apresentado o gráfico de *post plot* para os dados de produtividade de

soja. Nota-se que próximo ao centro do mapa, há uma concentração observações com menores valores. Ainda, pode-se observar na região leste do mapa, a presença de valores mais elevados para a variável produtividade de soja.

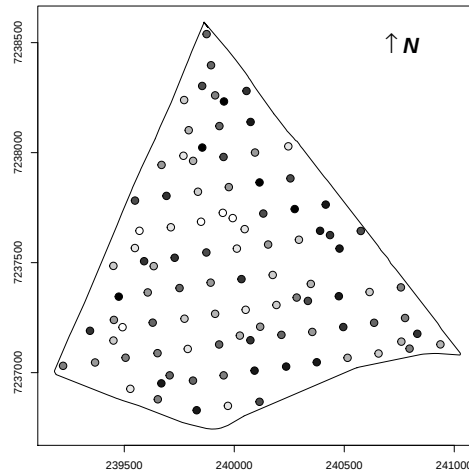


Figura 18 – Post plot dos dados de produtividade de soja.

Para verificar a anisotropia, foram construídos semivariogramas nas direções: 0° , 45° , 90° e 135° (GUEDES *et al.*, 2008). Considera-se a metade da maior distância, isto é, um *cutoff* de 50% (CLARK, 1979). Como o conjunto amostral apresentou distância máxima apresentada de 1765.957 m, a distância a ser considerada neste caso é de aproximadamente 883 m. Na Figura 19 é apresentado o gráfico com os semivariogramas direcionais. Pôde-se verificar um comportamento distinto nas direções de 0° e 90° , indicando a possibilidade da existência de uma tendência para estas direções na grade amostral.

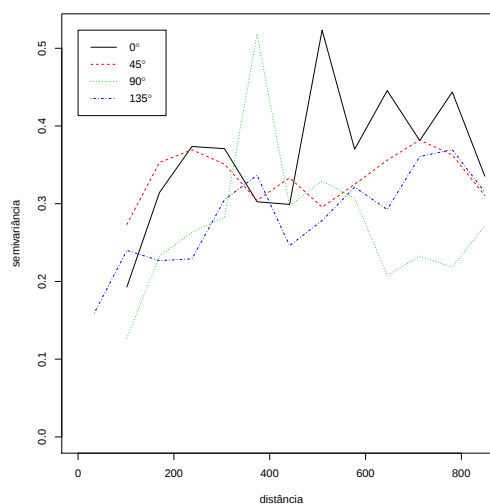


Figura 19 – Semivariogramas direcionais dos dados de produtividade de soja.

Na Tabela 12 é apresentado um resumo das informações sobre a função de semivariância $\gamma(\delta)$ em uma distância δ , e número de pares $N(\delta)$. Pode-se notar que o menor número de pares foi de 69 e o maior foi 478, sendo respectivamente, das distâncias 89 e 890 m. O número de pares com que foi calculado o semivariograma experimental, segue as recomendações de (JOURNEL; HUIJBREGTS, 1978), de um número mínimo de 30 pares. Para a escolha do *lag* fez-se um estudo com diversos valores, e o número de *lags* igual a 10 mostrou-se o mais adequado.

Tabela 12 – Resumo das informações sobre o semivariograma em uma distância δ , semivariância $\gamma(\delta)$ e número de pares $N(\delta)$.

	δ	$\hat{\gamma}(\delta)$	$N(\delta)$
1	89	0.2175438	69
2	178	0.2847493	317
3	267	0.3178154	217
4	356	0.3412497	399
5	445	0.2970399	426
6	534	0.3433789	388
7	623	0.3140708	478
8	712	0.3424597	423
9	801	0.3402491	398
10	890	0.3102631	385

5.2.1 Ajuste do Modelo

Após realizar uma análise descritiva dos dados, foi realizado o ajuste do modelo considerando os métodos de estimação de máxima verossimilhança e máxima verossimilhança restrita.

A Tabela 13 apresenta as estimativas dos parâmetros considerando o método da máxima verossimilhança. Para a estimação dos parâmetros, foram considerados os modelos Gaussiano, exponencial e da família Matérn. Para o modelo Matérn variou-se os valores de k em $k = 0.7$, $k = 1.0$ e $k = 1.5$. As estimativas dos parâmetros de efeito pepita e contribuição sofreram as menores variações, indo de 0 à 0.0845, enquanto o parâmetro de alcance sofreu grande variação, seu maior valor igual a 105.1861 quando considerado o modelo Gaussiano, e menor valor igual a 40.4732 quando considerado o modelo Matérn com $k = 1.5$. Com exceção do modelo Gaussiano, que apresentou moderada dependência espacial $0.25 < EPR \leq 0.75$, os

demais modelos apresentaram forte dependência espacial $EPR \leq 0.25$, segundo a escala de classificação de Cambardella *et al.* (1994).

Tabela 13 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança (ML) para os dados de produtividade de soja

Modelo (Método de Estimação)	$\hat{\beta}$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{a} = g(\hat{\phi}_3)$	EPR^1
Gaus (ML)	3.1367 (0.0694)	0.0845 (0.0610)	0.1927 (0.0734)	105.1861 (24.1016)	182.1877 (41.7452)	0.3048
Exp (ML)	3.1402 (0.0766)	0.0000 (0.1744)	0.2772 (0.1820)	71.1512 (37.8414)	213.4536 (113.5242)	0.000
Matérn $k = 0.7$ (ML)	3.1396 (0.0746)	0.0000 (0.1491)	0.2775 (0.1574)	58.3087 (26.3039)	201.1650 (90.7484)	0.000
Matérn $k = 1.0$ (ML)	3.1389 (0.0722)	0.0000 (0.1304)	0.2777 (0.1392)	46.9260 (18.3784)	187.7040 (73.5136)	0.0000
Matérn $k = 1.5$ (ML)	3.1385 (0.0719)	0.0335 (0.1034)	0.2442 (0.1130)	40.4732 (14.2670)	192.2477 (67.7682)	0.1206

A Tabela 14 apresenta as estimativas dos parâmetros considerando o método da máxima verossimilhança restrita. Para estimação dos parâmetros, também foram considerados os modelos Gaussiano, exponencial e Matérn. Para o modelo Matérn variou-se os valores de k em $k = 0.7$, $k = 1.0$ e $k = 1.5$. Da mesma forma, quando foi utilizado o método de máxima verossimilhança, os parâmetros de efeito pepita e contribuição sofreram menores variações nas estimativas para os diferentes modelos. Novamente, o parâmetro de alcance sofreu a maior variações entre os modelos, sendo seu maior valor igual a 107.6469 quando considerado o modelo Gaussiano e seu menor valor igual a 42.2765 quando considerado o modelo da família Matérn com $k = 1.5$. Assim como no caso anterior, com exceção do modelo Gaussiano, que apresentou moderada dependência espacial $0.25 < EPR \leq 0.75$, os demais modelos apresentaram forte dependência espacial $EPR \leq 0.25$.

5.2.1.1 Critérios de Validação Cruzada

Após realizar os ajustes dos modelos considerando os métodos de estimação dos parâmetros de máxima verossimilhança e máxima verossimilhança restrita, é necessário verificar qual modelo apresentou melhor ajuste. Para isso, os critérios de validação cruzada foram utilizados.

Na Tabela 15 são apresentados os critérios de validação cruzada para a escolha do melhor ajuste considerando o método de máxima verossimilhança. De acordo com os critérios

Tabela 14 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança restrita (REML) para os dados de produtividade de soja

Modelo (Método de Estimação)	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{a} = g(\hat{\phi}_3)$	EPR^2
Gaus (REML)	0.0867 (0.0599)	0.1950 (0.0729)	107.6469 (24.1766)	186.4499 (41.8951)	0.3078
Exp (REML)	0.0000 (0.1643)	0.2830 (0.1725)	74.5180 (38.2529)	223.5540 (114.7587)	0.0000
Matérn $k = 0.7$ (REML)	0.0000 (0.1414)	0.2829 (0.1503)	60.4390 (26.3300)	208.5146 (90.8385)	0.0000
Matérn $k = 1.0$ (REML)	0.0079 (0.1214)	0.2749 (0.1309)	49.3396 (18.8069)	197.3584 (75.2276)	0.0279
Matérn $k = 1.5$ (REML)	0.0393 (0.0975)	0.2435 (0.1078)	42.2765 (14.4962)	200.8134 (68.8569)	0.1390

estabelecidos, escolheu-se o modelo Gaussiano, o qual apresentou melhores resultados nos critérios de desvio-padrão do erro médio, erro absoluto, LMV e AIC.

Tabela 15 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança (ML) para os dados de produtividade de soja

Modelo (Método de Estimação)	EM	$\overline{ER}$	S_{EM}	S_{ER}	EA	LMV	AIC
Gaus (ML)	0.0050	0.0052	0.4668	0.9817	37.4381	-71.56	151.1
Exp (ML)	0.0041	0.0042	0.4771	0.9802	37.6294	-72.26	152.5
Matérn $k = 0.7$ (ML)	0.0047	0.0050	0.4735	0.9813	37.5661	-71.99	152.0
Matérn $k = 1.0$ (ML)	0.0054	0.0057	0.4708	0.9830	37.5425	-71.86	151.7
Matérn $k = 1.5$ (ML)	0.0053	0.0057	0.4700	0.9828	37.5012	-71.80	151.6

Na Tabela 16 são apresentados os critérios de validação cruzada para a escolha do melhor ajuste para o método da máxima verossimilhança restrita. De acordo com os critérios estabelecidos, escolheu-se o modelo Gaussiano, o qual apresentou melhores valores desejados nos valores de erro médio, erro médio reduzido, LMV e AIC.

Na Tabela 17 são apresentados os critérios de validação cruzada dos melhores ajustes para o método da máxima verossimilhança e máxima verossimilhança restrita, a fim de comparação. De acordo com os critérios estabelecidos, pôde-se verificar, que quando comparado os melhores ajustes de cada método, o método da máxima verossimilhança restrita apresentou melhores resultados.

Após a escolha do melhor ajuste considerando os diferentes métodos de estimação,

Tabela 16 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança restrita (REML) para os dados de produtividade de soja

Modelo (Método de Estimação)	EM	$\overline{ER}$	S_{EM}	S_{ER}	EA	LMV	AIC
Gaus (REML)	0.0049	0.0051	0.4668	0.9788	37.4600	-71.01	150.0
Exp (REML)	0.0042	0.0043	0.4766	0.9779	37.5920	-71.60	151.2
Matérn $k = 0.7$ (REML)	0.0048	0.0051	0.4729	0.9790	37.5368	-71.36	150.7
Matérn $k = 1.0$ (REML)	0.0054	0.0057	0.4706	0.9802	37.5090	-71.26	150.5
Matérn $k = 1.5$ (REML)	0.0053	0.0056	0.4698	0.9800	37.4674	-71.21	150.4

Tabela 17 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança (ML) e máxima verossimilhança restrita (REML) para os dados de produtividade de soja

Modelo (Método de Estimação)	EM	$\overline{ER}$	S_{EM}	S_{ER}	EA	LMV	AIC
Gaus (ML)	0.0050	0.0052	0.4668	0.9817	37.4381	-71.56	151.1
Gaus (REML)	0.0049	0.0051	0.4668	0.9788	37.4600	-71.01	150.0

foram gerados os mapas temáticos. A Figura 20 apresenta estes mapas para os dados de produtividade de soja dos modelos selecionados através dos critérios de validação cruzada. Foram consideradas quatro classes de mesma amplitude. A Figura 20a apresenta o mapa de krigagem do modelo Gaussiano utilizando o método da máxima verossimilhança restrita, e na Figura 20b o mapa de krigagem do modelo Gaussiano utilizando o método da máxima verossimilhança. Ambos os mapas foram construídos considerando a mesma escala, e visualmente não é possível notar a diferença entre os mapas.

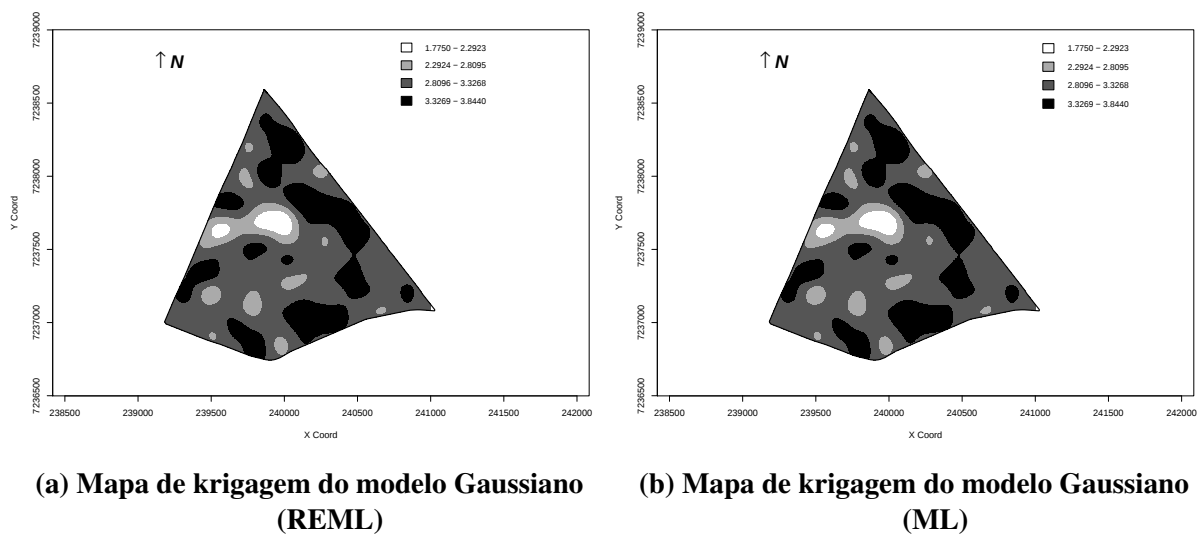


Figura 20 – Mapas de krigagem para o modelo Gaussiano utilizando o método da REML e ML para os dados de produtividade de soja.

Na Tabela 18 é apresentada a matriz de erros que fornece a quantidade de área em hectares (*ha*) da produtividade de soja (*ton/ha*) de cada mapa temático, bem como as áreas que coincidem para a comparação dos mapas considerando os métodos de máxima verossimilhança (ML) e máxima verossimilhança restrita (REML). Pôde-se observar que a terceira classe sofreu menor mudança entre as demais classes, com cerca de 96 *ha*.

Tabela 18 – Matriz de erros da quantidade de área em *ha* para os dados de produtividade de soja

Mapa utilizando o método de REML	Mapa utilizando o método de ML				Total
	1.7750 – 2.2923	2.2924 – 2.8095	2.8096 – 3.3268	3.3269 – 3.8440	
1.7750 – 2.2923	3.9726	0.4013	0.0016	0	4.3799
2.2924 – 2.8095	0.0216	15.7168	0.1304	0	15.8688
2.8096 – 3.3268	0	0.2212	96.8946	0.0802	97.1960
3.3269 – 3.8440	0	0	1.0835	48.8218	49.9053
Total	3.9942	16.3393	98.1001	48.9020	167.35

Para medir a semelhança entre os mapas temáticos dos dados de produtividade de soja considerando os diferentes métodos de estimação ML e REML, foi considerado o cálculo do índice Kappa (K). A medida apresentou como resultado o valor de $K = 0.9794$, o que indica alta exatidão entre os mapas. O intervalo de confiança de 95% obtido para essa medida foi de $[0.9789, 0.9800]$.

5.2.1.2 Diagnóstico de Influência Local

A fim de verificar a presença de observações atípicas que podem influenciar no resultado do ajuste do modelo, foi utilizado o método de influência local perturbação de Zhu. Na Figura 21 são apresentados os gráficos de influência para a perturbação de Zhu considerando o modelo Gaussiano, anteriormente escolhido através dos critérios de validação cruzada. O gráfico C_i vs ordem da Figura 21a destacou as observações #34, #55, #67, #69, #70, #71 e #85 como possíveis pontos influentes na resposta. Já o gráfico $|L_{max}|$ vs ordem da Figura 21b identificou as observações #29, #35, #56 e #93 como influentes na resposta.

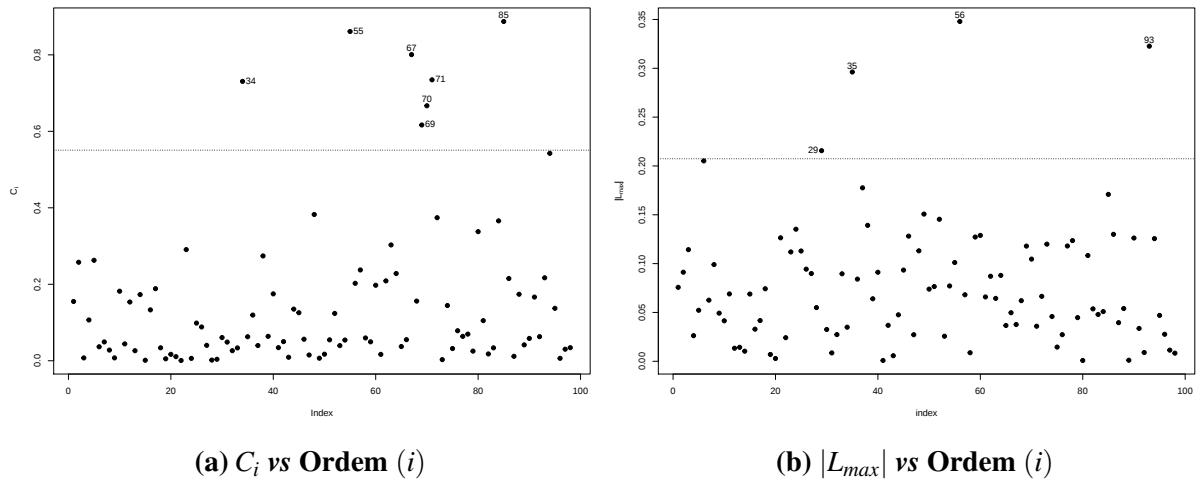


Figura 21 – Gráficos de influência para o modelo Gaussiano utilizando o método da REML para os dados de produtividade de soja.

5.2.2 Ajuste do Modelo sem Observações Influentes

A análise do conjunto de dados foi realizada sem considerar as observações #34, #65, #67, #69, #70, #71 e #85 individualmente, e algumas combinações destes pontos. As estimativas dos parâmetros obtidas para os diferentes modelos são apresentadas na Tabela 19. Inicialmente, realizou-se a análise sem considerar individualmente cada ponto que foi detectado como influente através dos gráficos de influência.

5.2.2.1 Critérios de Validação Cruzada

Pôde-se observar que ao retirar alguns conjuntos de observações, obteve-se um melhor ajuste, segundo os critérios de validação cruzada. Na Tabela 20 são apresentados os critérios de validação cruzada para a escolha do melhor ajuste. De acordo com os critérios estabelecidos, escolheu-se o modelo quando foi considerado a retirada das observações #34, #55, #67, #69 e #85. Neste caso o modelo apresentou melhores resultados para os critérios de desvio-padrão do erro médio, erro absoluto, LMV e AIC.

Tabela 19 – Estimativas e erros-padrão (entre parênteses) dos parâmetros via máxima verossimilhança restrita (REML) sem considerar os pontos influentes para os dados de produtividade de soja

Modelo	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$	$\hat{a} = g(\hat{\phi}_3)$	EPR
Gaus	0.0867 (0.0599)	0.1950 (0.0729)	107.6469 (24.1766)	186.4499 (41.8951)	0.3078
Gaus (−#34)	0.0931 (0.0566)	0.1802 (0.0691)	116.5564 (26.5096)	201.8816 (45.9160)	0.3406
Gaus (−#55)	0.0605 (0.0534)	0.2194 (0.0689)	106.1089 (20.6835)	183.7860 (35.8249)	0.2161
Gaus (−#67)	0.1073 (0.0627)	0.1578 (0.0725)	112.8390 (29.9030)	195.4429 (51.7935)	0.4048
Gaus (−#69)	0.0727 (0.0667)	0.1903 (0.0776)	95.4906 (24.1468)	165.3946 (41.8235)	0.2764
Gaus (−#85)	0.0802 (0.0494)	0.1916 (0.0641)	117.3133 (23.8710)	203.1926 (41.3458)	0.2951
Gaus (−#67, −#85)	0.0936 (0.0516)	0.1017 (0.0633)	121.2927 (28.3019)	210.0851 (49.0203)	0.4793
Gaus (−#34, −#67, −#85)	0.0923 (0.0465)	0.1552 (0.0586)	131.9701 (30.4834)	228.5789 (52.7988)	0.3729
Gaus (−#34, −#55, −#67, −#85)	0.0578 (0.0392)	0.1884 (0.0552)	123.9766 (22.2720)	214.7338 (38.5762)	0.2348
Gaus (−#34, −#55, −#67, −#69, −#85)	0.0501 (0.0420)	0.1758 (0.0551)	111.8213 (21.5502)	193.6802 (37.3260)	0.2218

Tabela 20 – Critérios para seleção do melhor ajuste utilizando o método da máxima verossimilhança restrita (REML) para os dados de produtividade de soja

Modelo	EM	$\overline{ER}$	S_{EM}	S_{ER}	EA	LMV	AIC
Gaus	0.0049	0.0051	0.4668	0.9788	37.4600	−71.01	150.0
Gaus (−#34)	0.0018	0.0017	0.4614	0.9851	36.7577	−68.36	144.7
Gaus (−#55)	0.0069	0.0078	0.4527	0.9725	35.8406	−68.92	145.8
Gaus (−#67)	0.0027	0.0028	0.4676	0.9880	36.6536	−68.28	144.6
Gaus (−#69)	0.0058	0.0063	0.4566	0.9748	36.4358	−68.14	144.3
Gaus (−#85)	0.0066	0.0071	0.4504	0.9820	35.9681	−67.07	142.1
Gaus (−#67, −#85)	0.0041	0.0045	0.4501	0.9908	35.1248	−64.29	136.6
Gaus (−#34, −#67, −#85)	0.0015	0.0015	0.4407	0.9981	34.2188	−61.08	130.2
Gaus (−#34, −#55, −#67, −#85)	0.0025	0.0030	0.4262	0.9963	32.9166	−58.48	125.0
Gaus (−#34, −#55, −#67, −#69, −#85)	0.0032	0.0041	0.4182	0.9888	32.3223	−55.84	119.7

Após a escolha do melhor modelo ajustado, foi gerado um novo mapa temático sem considerar o grupo de observações influentes, a fim de comparar com o mapa do modelo ajustado com todas as observações. A Figura 22 apresenta os mapas temáticos para os dados de produtividade de soja utilizando o método da máxima verossimilhança restrita. Na Figura 22a é apresentado o mapa de krigagem considerando todos os pontos, e na Figura 22b o mapa de krigagem quando considerado a retirada do conjunto de observações atípicas. Ambos os mapas foram construídos considerando a mesma escala, e visualmente é possível notar a diferença entre eles.

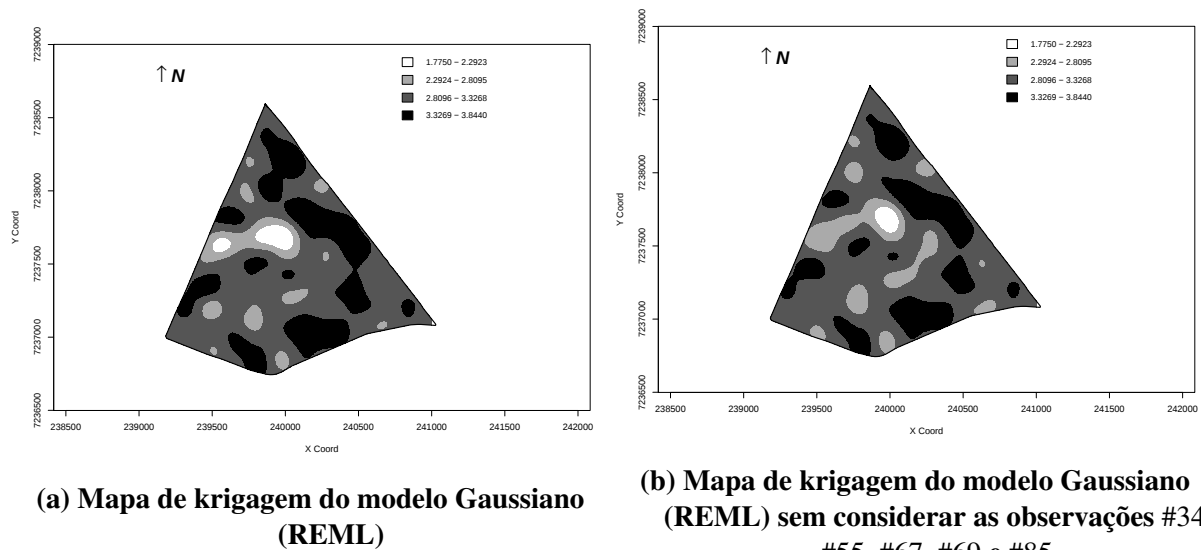


Figura 22 – Mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de produtividade de soja.

Na Figura 23 são apresentados os demais mapas de krigagem referentes aos diferentes modelos ajustados considerando a retirada dos diferentes pontos potencialmente influentes que foram identificados pelos gráficos de influência mostrados na Figura 21, e posteriormente submetidos aos critérios de seleção do melhor ajuste na Tabela 20.

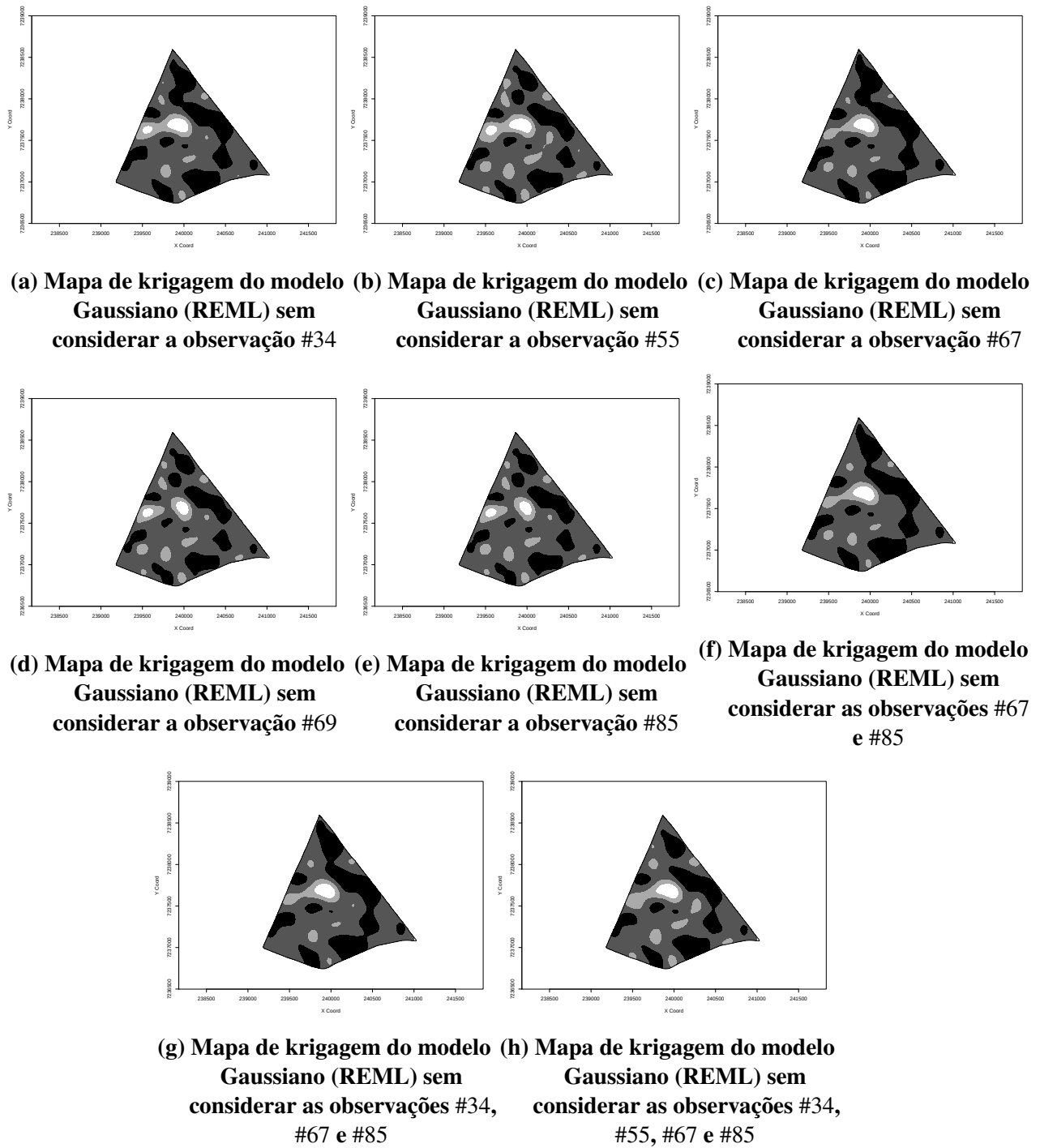


Figura 23 – Demais mapas de krigagem para o modelo Gaussiano utilizando o método da REML para os dados de produtividade de soja.

Na Tabela 21 é apresentada a matriz de erros que fornece a quantidade de área em hectares (*ha*) da produtividade de soja (*ton/ha*) de cada mapa temático apresentado na Figura 22, bem como as áreas que coincidiram quando o mapa foi construído considerando todos os pontos e sem considerar o grupo de observações. Pôde-se observar que a terceira classe sofreu a menor mudança entre os mapas, com uma área de aproximadamente 86 *ha*.

Tabela 21 – Matriz de erros da quantidade de área em *ha* para os dados de produtividade de soja

Mapa utilizando todos os pontos	Mapa sem considerar as observações #34, #55, #67, #69 e #85				Total
	1.7750 – 2.2923	2.2924 – 2.8095	2.8096 – 3.3268	3.3269 – 3.8440	
1.7750 – 2.2923	1.7481	1.6045	0.0827	0	3.4353
2.2924 – 2.8095	0.1300	15.7594	2.7843	0	18.6737
2.8096 – 3.3268	0	6.4822	86.3639	0.7746	93.6207
3.3269 – 3.8440	0	0	6.2579	45.3624	51.6203
Total	1.8781	23.8461	95.4888	46.1370	167.35

Para medir a semelhança entre os mapas temáticos apresentados na Figura 22, foi considerado o cálculo do índice Kappa que apresentou o resultado de $K = 0.8132$, o que indica moderada exatidão entre os mapas. O intervalo de confiança de 95% obtido para essa medida foi de $[0.8118, 0.8147]$.

5.2.3 Considerações Gerais

De acordo com as análises realizadas utilizando métodos geoestatísticos foi possível novamente verificar a variabilidade espacial dos dados em estudo. Inicialmente consideramos os estimadores de máxima verossimilhança e máxima verossimilhança restrita para estimar os parâmetros dos modelos Gaussiano, exponencial, e da família Matérn para $\kappa = (0.7, 1.0, 1.5)$ considerando todas as observações. Observamos que os valores dos parâmetros estimados, em ambos os métodos de estimação, apresentaram valores semelhantes quando considerado o mesmo modelo, conforme pode ser verificado nas Tabelas 13 e 14. Ao realizar o método de validação cruzada, em ambos os métodos de estimação utilizados, o modelo Gaussiano foi o escolhido. Comparando os critérios de validação cruzada dos modelos escolhidos nos respectivos métodos de estimação, máxima verossimilhança e máxima verossimilhança restrita, obteve-se que o segundo apresentou melhores resultados.

Aplicando o método de influência local, observamos a presença de sete observações influentes no modelo, as observações #34, #55, #67, #69, #70, #71 e #85. Ao realizar o ajuste do

modelo sem considerar o grupo de observações (#34, #55, #67, #69 e #85), pôde-se verificar que este modelo apresentou o melhor ajuste segundo os critérios utilizados. A fim de comparação, apresentamos os mapas temáticos do modelo escolhido considerando todas as observações e do modelo retirando o conjunto de observações.

6 CONCLUSÕES

No presente trabalho, explorou-se uma nova abordagem para o método de influência local em modelos espaciais lineares gaussianos considerando o método de estimação de máxima verossimilhança restrita. O interesse em estudar este tópico é devido a importância de se modelar a estrutura de dependência espacial, característica presente em conjuntos de dados geoestatísticos. Deseja-se considerar um método que alcance melhores resultados nos ajustes dos modelos e que apresente uma melhor detecção de possíveis valores atípicos nos conjuntos de dados, que podem distorcer os resultados e nos levar a conclusões e interpretações erradas sobre a análise geoestatística.

Nos Capítulos 1 e 2, foram apresentadas uma introdução e o referencial teórico sobre a Geoestatística. No Capítulo 2 também foi apresentado de forma mais específica os modelos teóricos utilizados no decorrer do trabalho que usam das características espaciais para descrever os dados. Em seguida, foram apresentados os métodos de estimação dos parâmetros que compõem estes modelos. Nessa dissertação consideramos os métodos de máxima verossimilhança e máxima verossimilhança restrita, com a finalidade de comparar os resultados considerando os diferentes métodos. Também foram apresentados os métodos de interpolação, conhecidos como krigagem. Esta ferramenta é essencial para construção dos mapas que auxiliam na interpretação do ajuste do modelo. Em seguida apresentou-se os critérios que são utilizados na escolha do melhor modelo ajustado, os critérios de validação cruzada.

No Capítulo 3 são apresentados os métodos de diagnóstico de influência, em particular o método de influência local na variável resposta, perturbação de Zhu. Esta metodologia foi estudada na identificação de possíveis observações influentes que podem prejudicar o ajuste do modelo.

O Capítulo 4 apresenta um estudo de simulações para avaliar o desempenho da metodologia apresentada neste trabalho na detecção de observações influentes. Os resultados das simulações mostraram que o método utilizado apresenta um bom desempenho para diferentes valores de perturbação. Ou seja, pode ser coerente considerar este método para melhor identificação de observações influentes nos conjuntos de dados.

A aplicação a dois conjuntos distintos de dados reais, apresentados no Capítulo 5, evidenciou o que já havia sido verificado através do estudo de simulação, que ao estimar os parâmetros utilizando método de máxima verossimilhança restrita foi obtido um melhor ajuste do modelo. Além disso, os resultados das análises geoestatísticas mostraram que, quando foi

considerada a retirada dos pontos influentes detectados pelo modelo, os modelos ajustados apresentaram melhores resultados.

Dessa forma, justifica-se a relevância da abordagem apresentada nesta dissertação, considerar um método de estimação que nos dê uma interpretação que traduza de forma mais real a informação e que explique mais fielmente o conjunto de dados considerado. Este ponto é uma tarefa primordial, uma vez que as análises geoestatísticas muitas vezes são utilizadas para a tomada de decisões de fundamental importância.

Para trabalhos futuros deseja-se:

- Verificar o comportamento que o parâmetro de perturbação pode causar nos dados considerando a abordagem utilizada neste trabalho.
- Estudar novos métodos para propor um valor ótimo de perturbação para aperfeiçoar a detecção de pontos influentes.

REFERÊNCIAS

- ANSELIN, L. Spatial effects in econometric practice in environmental and resource economics. **American Journal of Agricultural Economics**, JSTOR, v. 83, n. 3, p. 705–710, 2001.
- ARMSTRONG, M. Problems with universal kriging. **Mathematical Geology**, Springer, v. 16, n. 1, p. 101–108, 1984.
- ASSUMPÇÃO, R.; URIBE-OPAZO, M.; GALEA, M. Analysis of local influence in geostatistics using student's t-distribution. **Journal of Applied Statistics**, Taylor & Francis, v. 41, n. 11, p. 2323–2341, 2014.
- BÁRDOSSY, A. Copula-based geostatistical models for groundwater quality parameters. **Water Resources Research**, Wiley Online Library, v. 42, n. 11, 2006.
- BIVAND, R. S. *et al.* **Applied spatial data analysis with R**. [S.l.]: Springer, 2008. v. 747248717.
- BORSSOI, J.; URIBE-OPAZO, M.; GALEA, M. Técnicas de diagnóstico de influência local na análise espacial da produtividade da soja. **Engenharia Agrícola**, v. 31, n. 2, p. 376–387, 2011.
- BORSSOI, J. A. *et al.* Local influence of explanatory variables in gaussian spatial linear models. **The Chilean Journal of Statistics**, v. 2, n. 2, p. 29–38, 2011.
- BORSSOI, J. A.; URIBE-OPAZO, M. A.; ROJAS, M. G. Diagnostic techniques applied in geostatistics for agricultural data analysis. **Revista Brasileira de Ciência do Solo**, SciELO Brasil, v. 33, n. 6, p. 1561–1570, 2009.
- CAMBARDELLA, C. *et al.* Field-scale variability of soil properties in central iowa soils. **Soil Science Society of America Journal**, Soil Science Society of America, v. 58, n. 5, p. 1501–1511, 1994.
- CASELLA, G.; BERGER, R. L. **Statistical inference**. [S.l.]: Duxbury Pacific Grove, CA, 2002. v. 2.
- CASTRILLON-CANDÁS, J. E.; GENTON, M. G.; YOKOTA, R. Multi-level restricted maximum likelihood covariance estimation and kriging for large non-gridded spatial datasets. **Spatial Statistics**, Elsevier, v. 18, p. 105–124, 2016.
- CHICA-OLMO, M.; ABARCA-HERNANDEZ, F. Computing geostatistical image texture for remotely sensed data classification. **Computers & Geosciences**, Elsevier, v. 26, n. 4, p. 373–383, 2000.
- CHILÈS, J.; DELFINER, P. Geostatistics: Modeling spatial uncertainty, A John Wiley & Sons. **Inc., Publication, ISBN-13**, p. 978–0, 1999.
- CHRISTENSEN, R. The equivalence of predictions from universal kriging and intrinsic random-function kriging. **Mathematical Geology**, Springer, v. 22, n. 6, p. 655–664, 1990.
- CHRISTENSEN, R.; JOHNSON, W.; PEARSON, L. M. Prediction diagnostics for spatial linear models. **Biometrika**, JSTOR, p. 583–591, 1992.
- CHRISTENSEN, R.; JOHNSON, W.; PEARSON, L. M. Covariance function diagnostics for spatial linear models. **Mathematical Geology**, Springer, v. 25, n. 2, p. 145–160, 1993.

- CLARK, I. The semivariogram—part 1. **Engineering Mining Journal**, v. 180, n. 7, p. 90–94, 1979.
- COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**, Sage Publications Sage CA: Thousand Oaks, CA, v. 20, n. 1, p. 37–46, 1960.
- COOK, R. D. Assessment of local influence. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 48, n. 2, p. 133–155, 1986.
- COX, D. R.; HINKLEY, D. V. **Theoretical Statistics**. [S.l.]: Chapman and Hall/CRC, 1979.
- CRESSIE, N. Geostatistics. **The American Statistician**, Taylor & Francis, v. 43, n. 4, p. 197–202, 1989.
- CRESSIE, N. The origins of kriging. **Mathematical geology**, Springer, v. 22, n. 3, p. 239–252, 1990.
- CRESSIE, N. A. Statistics for spatial data. John Wiley and Sons. **Inc., New York**, 1993.
- DE BASTIANI, F. **Influência local em modelos espaciais lineares com distribuição da família de contornos elípticos**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2012.
- DE BASTIANI, F. *et al.* Influence diagnostics in elliptical spatial linear models. **Test**, Springer, v. 24, n. 2, p. 322–340, 2015a.
- DE BASTIANI, F. *et al.* Case-deletion diagnostics for spatial linear mixed models. **Spatial Statistics**, Elsevier, v. 28, p. 284–303, 2018.
- DE BASTIANI, F. *et al.* Global influence diagnostics in gaussian spatial linear model with multiple repetitions. **Procedia Environmental Sciences**, Elsevier, v. 26, p. 74–77, 2015b.
- DIGGLE, P.; RIBEIRO, J. Pj. model based geostatistics. **SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA**, v. 14, p. 192, 2007.
- DOWD, P.; PARDO-IGÚZQUIZA, E. The incorporation of model uncertainty in geostatistical simulation. **Geographical and Environmental Modelling**, Taylor & Francis, v. 6, n. 2, p. 147–169, 2002.
- DUTTA, S.; MONDAL, D. *et al.* Reaml estimation with intrinsic matérn dependence in the spatial linear mixed model. **Electronic Journal of Statistics**, The Institute of Mathematical Statistics and the Bernoulli Society, v. 10, n. 2, p. 2856–2893, 2016.
- FARACO, M. A. *et al.* Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. **Revista Brasileira de Ciência do Solo**, Sociedade Brasileira de Ciência do Solo, v. 32, n. 2, 2008.
- GALEA, M.; PAULA, G. A.; URIBE-OPAZO, M. On influence diagnostic in univariate elliptical linear regression models. **Statistical Papers**, Springer, v. 44, n. 1, p. 23–45, 2003.
- GARCIA-PAPANI, F. *et al.* Birnbaum-saunders spatial regression models: Diagnostics and application to chemical data. **Chemometrics and Intelligent Laboratory Systems**, Elsevier, v. 177, p. 114–128, 2018.

GNEITING, T. Nonseparable, stationary covariance functions for space–time data. **Journal of the American Statistical Association**, Taylor & Francis, v. 97, n. 458, p. 590–600, 2002.

GOOVAERTS, P. Geostatistical modelling of uncertainty in soil science. **Geoderma**, Elsevier, v. 103, n. 1-2, p. 3–26, 2001.

GOOVAERTS, P. *et al.* **Geostatistics for natural resources evaluation**. [S.l.]: Oxford University Press on Demand, 1997.

GUEDES, L. P. C. *et al.* Anisotropy to analyze spatial variability of some spatially referenced soil chemical variables. **Revista Brasileira de Ciência do Solo**, SciELO Brasil, v. 32, n. 6, p. 2217–2226, 2008.

ISAAKS, E. H.; SRIVASTAVA, M. R. **Applied Geostatistics**. [S.l.: s.n.], 1989.

JONES, R. H. Fitting a stochastic partial differential equation to aquifer data. **Stochastic Hydrology and Hydraulics**, Springer, v. 3, n. 2, p. 85–96, 1989.

JOURNEL, A. Geostatistics: models and tools for the earth sciences. **Mathematical Geology**, Springer, v. 18, n. 1, p. 119–140, 1986.

JOURNEL, A. G. **Fundamentals of geostatistics in five lessons**. [S.l.]: American Geophysical Union Washington, DC, 1989. v. 8.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining Geostatistics**. [S.l.]: Academic press London, 1978. v. 600.

KITANIDIS, P. K. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. **Water Resources Research**, Wiley Online Library, v. 19, n. 4, p. 909–921, 1983.

KITANIDIS, P. K. Parametric estimation of covariances of regionalized variables 1. **JAWRA Journal of the American Water Resources Association**, Wiley Online Library, v. 23, n. 4, p. 557–567, 1987.

KITANIDIS, P. K.; SHEN, K.-F. Geostatistical interpolation of chemical concentration. **Advances in Water Resources**, Elsevier, v. 19, n. 6, p. 369–378, 1996.

KITANIDIS, P. K.; VOMVORIS, E. G. A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations. **Water Resources Research**, Wiley Online Library, v. 19, n. 3, p. 677–690, 1983.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on the witwatersrand. **Journal of the Southern African Institute of Mining and Metallurgy**, Southern African Institute of Mining and Metallurgy, v. 52, n. 6, p. 119–139, 1951.

KRIPPENDORFF, K. **Content analysis: An introduction to its methodology**. [S.l.]: Sage publications, 2004.

LANDIM, P. M. B. Introdução aos métodos de estimação espacial para confecção de mapas. **DGA, IGCE, UNESP/Rio Claro, Lab. Geomatemática, Texto Didático**, v. 2, p. 20, 2000.

LANDIM, P. M. B. Sobre geoestatística e mapas. **Terrae Didatica**, v. 2, n. 1, p. 19–33, 2006.

LUTKEPOHL, H. Handbook of matrices. **Computational Statistics and Data analysis**, v. 2, n. 25, p. 243, 1997.

MARCHANT, B.; LARK, R. Robust estimation of the variogram by residual maximum likelihood. **Geoderma**, Elsevier, v. 140, n. 1-2, p. 62–72, 2007.

MARDIA, K. V.; MARSHALL, R. J. Maximum likelihood estimation of models for residual covariance in spatial regression. **Biometrika**, Oxford University Press, v. 71, n. 1, p. 135–146, 1984.

MATÉRN, B. **Spatial Variation, volume 36 of Lecture Notes in Statistics**. [S.l.]: Springer-Verlag, Berlin,, 1986.

MATHERON, G. **Traité de géostatistique appliquée. 1 (1962)**. [S.l.]: Editions Technip, 1962. v. 1.

MATHERON, G. Principles of geostatistics. **Economic Geology**, Society of Economic Geologists, v. 58, n. 8, p. 1246–1266, 1963.

MATHERON, G. **Le krigeage universel**. [S.l.]: École nationale supérieure des mines de Paris Paris, 1969. v. 1.

MCBRATNEY, A.; WEBSTER, R. Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. **Journal of Soil Science**, Wiley Online Library, v. 37, n. 4, p. 617–639, 1986.

MELLO, J. d. *et al.* Ajuste e seleção de modelos espaciais de semivariograma visando à estimativa volumétrica de eucalyptus grandis. **Scientia Forestalis**, v. 69, n. 1, 2005.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2012. v. 821.

NEL, D. On matrix differentiation in statistics. **South African Statistical Journal**, South African Statistical Association (SASA), v. 14, n. 2, p. 137–193, 1980.

NETO, P. L. d. O. C. Estatística, edgard blucher. **São Paulo**, 2002.

OLEA, R. A. **Geostatistics for engineers and earth scientists**. [S.l.]: Springer Science & Business Media, 2012.

OLIVER, M. A.; WEBSTER, R. **Basic steps in geostatistics: the variogram and kriging**. [S.l.]: Springer, 2015. v. 106.

ORTEGA, E. M.; BOLFARINE, H.; PAULA, G. A. Influence diagnostics in generalized log-gamma regression models. **Computational Statistics & Data Analysis**, Elsevier, v. 42, n. 1-2, p. 165–186, 2003.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, Oxford University Press, v. 58, n. 3, p. 545–554, 1971.

PETERSEN, K. B.; PEDERSEN, M. S. The matrix cookbook, nov 2012. URL <http://www2.imm.dtu.dk/pubdb/p.php>, v. 3274, 2012.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2018. Disponível em: <<https://www.R-project.org/>>.

Ribeiro Jr, P. J.; DIGGLE, P. J. **geoR: Analysis of Geostatistical Data**. [S.l.], 2016. R package version 1.7-5.2. Disponível em: <<https://CRAN.R-project.org/package=geoR>>.

ROCHA, M. M. da; YAMAMOTO, J. K. Comparison between kriging variance and interpolation variance as uncertainty measurements in the capanema iron mine, state of minas gerais—brazil. **Natural Resources Research**, Springer, v. 9, n. 3, p. 223–235, 2000.

ROUSSEEUW, P. J.; ZOMEREN, B. C. V. Unmasking multivariate outliers and leverage points. **Journal of the American Statistical Association**, Taylor & Francis, v. 85, n. 411, p. 633–639, 1990.

SCHABENBERGER, O.; PIERCE, F. J. **Contemporary statistical models for the plant and soil sciences**. [S.l.]: CRC press, 2001.

SEBER, G. A.; LEE, A. J. **Linear regression analysis**. [S.l.]: John Wiley & Sons, 2012. v. 329.

SHERMAN, M. **Spatial statistics and spatio-temporal data: covariance functions and directional properties**. [S.l.]: John Wiley & Sons, 2011.

TOBLER, W. R. A computer movie simulating urban growth in the detroit region. **Economic Geography**, Taylor & Francis, v. 46, n. sup1, p. 234–240, 1970.

TRANGMAR, B. B.; YOST, R. S.; UEHARA, G. Application of geostatistics to spatial studies of soil properties. In: **Advances in Agronomy**. [S.l.]: Elsevier, 1986. v. 38, p. 45–94.

URIBE-OPAZO, M. A.; BORSSOI, J. A.; GALEA, M. Influence diagnostics in gaussian spatial linear models. **Journal of Applied Statistics**, Taylor & Francis, v. 39, n. 3, p. 615–630, 2012.

VERBEKE, G.; MOLENBERGHS, G. **Linear mixed models for longitudinal data**. [S.l.]: Springer Science & Business Media, 2009.

VIEIRA, S. R. *et al.* Geoestatística em estudos de variabilidade espacial do solo. **Tópicos em ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo**, v. 1, p. 1–53, 2000.

WACKERNAGEL, H. **Multivariate geostatistics: an introduction with applications**. [S.l.]: Springer Science & Business Media, 2013.

WARNES, J. J. A sensitivity analysis for universal kriging. **Mathematical Geology**, Springer, v. 18, n. 7, p. 653–676, 1986.

YAMAMOTO, J. Curso de geoestatística aplicada. **Instituto de Geociências–USP. São Paulo**, 2001.

ZHU, H. *et al.* Perturbation selection and influence measures in local influence analysis. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 35, n. 6, p. 2565–2588, 2007.

ZHU, H.-T.; LEE, S.-Y. Local influence for incomplete data models. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, Wiley Online Library, v. 63, n. 1, p. 111–126, 2001.

ZIMMERMAN, D. L. Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. **Mathematical Geology**, Springer, v. 21, n. 7, p. 655–672, 1989.

APÊNDICE A – CONTAS

Seja o logaritmo da máxima verossimilhança restrita:

$$l_R(\phi) = -\frac{n-p}{2} \log(2\pi) + \frac{1}{2} \log(|X^\top X|) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} \log |X^\top \Sigma^{-1} X| - \frac{1}{2} Z^\top P Z,$$

em que:

$$P = \Sigma^{-1}(I - A); \quad A = X(X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1}$$

Observe que:

$$P = \Sigma^{-1} - \Sigma^{-1} A$$

$$P = \Sigma^{-1} - \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1}.$$

Assim, temos que:

$$\begin{aligned} \frac{\partial l_R(\phi)}{\partial \phi_i} &= -\frac{1}{2} \left[\frac{\partial (\log |\Sigma| + \log |X^\top \Sigma^{-1} X| + Z^\top P Z)}{\partial \phi_i} \right] \\ &= -\frac{1}{2} \left[\frac{\partial (\log |\Sigma| + \log |X^\top \Sigma^{-1} X| + Z^\top P Z)}{\partial \phi_i} \right]. \end{aligned}$$

Agora, sejam respectivamente 1 e 2:

$$-\frac{1}{2} \frac{\partial \log |\Sigma|}{\partial \phi_i} - \frac{1}{2} \frac{\partial \log |X^\top \Sigma^{-1} X|}{\partial \phi_i} \quad (1);$$

$$-\frac{1}{2} Z^\top \frac{\partial P}{\partial \phi_i} Z \quad (2).$$

Primeiro, vamos resolver a primeira parte (1):

$$\begin{aligned}
& -\frac{1}{2} \frac{\partial \log |\Sigma|}{\partial \phi_i} - \frac{1}{2} \frac{\partial \log |X^\top \Sigma^{-1} X|}{\partial \phi_i} \\
&= -\frac{1}{2} \text{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \right) - \frac{1}{2} \text{tr} \left((X^\top \Sigma^{-1} X)^{-1} X^\top \frac{\partial \Sigma^{-1}}{\partial \phi_i} X \right) \\
&= -\frac{1}{2} \text{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \right) + \frac{1}{2} \text{tr} \left(X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} \right) \\
&= -\frac{1}{2} \text{tr} \left(\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \right) + \frac{1}{2} \text{tr} \left(\Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \right) \\
&= -\frac{1}{2} \text{tr} \left[\left(\Sigma^{-1} - \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \right) \frac{\partial \Sigma}{\partial \phi_i} \right] \\
&= -\frac{1}{2} \text{tr} \left[\Sigma^{-1} (I - X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1}) \frac{\partial \Sigma}{\partial \phi_i} \right] \\
&= -\frac{1}{2} \text{tr} \left[\Sigma^{-1} (I - A) \frac{\partial \Sigma}{\partial \phi_i} \right] \\
&= -\frac{1}{2} \text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_i} \right).
\end{aligned}$$

Agora, vamos desenvolver apenas a parte da derivada de (2), $\frac{\partial P}{\partial \phi_i}$:

$$\begin{aligned}
\frac{\partial P}{\partial \phi_i} &= \frac{\partial \Sigma^{-1} (I - A)}{\partial \phi_i} \\
&= \frac{\partial \Sigma^{-1}}{\partial \phi_i} (I - A) + \Sigma^{-1} \frac{\partial (I - A)}{\partial \phi_i} \\
&= -\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} (I - A) - \Sigma^{-1} \frac{\partial A}{\partial \phi_i}.
\end{aligned}$$

Desenvolvendo $\frac{\partial A}{\partial \phi_i}$:

$$\begin{aligned}
\frac{\partial A}{\partial \phi_i} &= \frac{\partial X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1}}{\partial \phi_i} \\
&= X \frac{\partial (X^\top \Sigma^{-1} X)^{-1}}{\partial \phi_i} X^\top \Sigma^{-1} + X (X^\top \Sigma^{-1} X)^{-1} \frac{\partial \Sigma^{-1}}{\partial \phi_i} \\
&= -X (X^\top \Sigma^{-1} X)^{-1} \left(X^\top \frac{\partial \Sigma^{-1}}{\partial \phi_i} X \right) (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} - X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma^{-1}}{\partial \phi_i} \Sigma^{-1} \\
&= X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma^{-1}}{\partial \phi_i} \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} - X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1}.
\end{aligned}$$

Substituindo na equação anterior:

$$\begin{aligned}
\frac{\partial P}{\partial \phi_i} &= -\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} (I - A) - \Sigma^{-1} \frac{\partial A}{\partial \phi_i} \\
&= -\Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} (I - A) - \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \\
&\quad + \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma^{-1}}{\partial \phi_i} \Sigma^{-1} \\
&= -\Sigma^{-1} \frac{\Sigma}{\partial \phi_i} \Sigma^{-1} (I - A) - \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\Sigma}{\partial \phi_i} \Sigma^{-1} (I - A) \\
&= \left(-\Sigma \frac{\partial \Sigma}{\partial \phi_i} - \Sigma^{-1} X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_i} \right) \Sigma^{-1} (I - A) \\
&= -\Sigma^{-1} (I - X (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1}) \frac{\partial \Sigma}{\partial \phi_i} \Sigma^{-1} (I - A) \\
&= -\Sigma^{-1} (I - A) \frac{\partial \Sigma}{\partial \phi_i} P \\
&= -P \frac{\partial \Sigma}{\partial \phi_i} P.
\end{aligned}$$

Substituindo na equação anterior, temos:

$$-\frac{1}{2} Z^\top \frac{\partial P}{\partial \phi_i} Z = -\frac{1}{2} Z^\top \left(-P \frac{\partial \Sigma}{\partial \phi_i} P \right) Z = \frac{1}{2} Z^\top P \frac{\partial \Sigma}{\partial \phi_i} P Z.$$

Agora, substituindo na primeira equação:

$$\begin{aligned}
\frac{\partial l_R(\phi)}{\partial \phi_i} &= -\frac{1}{2} \text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_i} \right) + \frac{1}{2} Z^\top P \frac{\partial \Sigma}{\partial \phi_i} Z \\
&= -\frac{1}{2} \left[\text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_i} \right) - Z^\top P \frac{\partial \Sigma}{\partial \phi_i} P Z \right].
\end{aligned}$$

Agora, sejam respectivamente 3 e 4:

$$-\frac{1}{2} \left[\text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_i} \right) \right] \quad (3);$$

$$\frac{1}{2} \left[Z^\top P \frac{\partial \Sigma}{\partial \phi_i} P Z \right] \quad (4).$$

Desenvolvendo 3, temos que:

$$\begin{aligned}
-\frac{1}{2} \left\{ \frac{\partial \left[\text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_i} \right) \right]}{\partial \phi_j} \right\} &= -\frac{1}{2} \left[\text{tr} \left(\frac{\partial P}{\partial \phi_j} \frac{\partial \Sigma}{\partial \phi_i} + P \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} \right) \right] \\
&= \frac{1}{2} \left[\text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} - P \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} \right) \right]
\end{aligned}$$

Agora, desenvolvendo 4, temos que:

$$\begin{aligned}
\frac{1}{2} \left[\frac{\partial \left(Z^\top P \frac{\partial \Sigma}{\partial \phi_i} P Z \right)}{\partial \phi_j} \right] &= \frac{1}{2} \left[Z^\top \frac{\partial P}{\partial \phi_j} \frac{\partial \Sigma}{\partial \phi_i} P Z + Z^\top P \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} P Z + Z^\top P \frac{\partial \Sigma}{\partial \phi_i} \frac{\partial P}{\partial \phi_j} Z \right] \\
&= \frac{1}{2} \left[-Z^\top P \frac{\partial \Sigma}{\partial \phi_j} P \frac{\Sigma}{\partial \phi_i} P Z + Z^\top P \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} P Z - Z^\top P \frac{\partial \Sigma}{\partial \phi_i} P \frac{\partial \Sigma}{\partial \phi_j} P Z \right] \\
&= \frac{1}{2} Z^\top P \left[\frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} - \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} - \frac{\partial \Sigma}{\partial \phi_i} P \frac{\partial \Sigma}{\partial \phi_j} \right] P Z.
\end{aligned}$$

Dessa forma, temos que:

$$\frac{\partial^2 l_R(\phi)}{\partial \phi_i \partial \phi_j} = \frac{1}{2} \left[\text{tr} \left(P \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} - P \frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} \right) \right] + \frac{1}{2} Z^\top P \left[\frac{\partial^2 \Sigma}{\partial \phi_i \partial \phi_j} - \frac{\partial \Sigma}{\partial \phi_j} P \frac{\partial \Sigma}{\partial \phi_i} - \frac{\partial \Sigma}{\partial \phi_i} P \frac{\partial \Sigma}{\partial \phi_j} \right] P Z.$$

Influência Local - Pertubação de Zhu

$$l(\theta/\omega) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (Z\omega - X\beta)^\top \Sigma^{-1} (Z\omega - X\beta)$$

em que:

$$Z\omega = Z - \Sigma^{\frac{1}{2}} \omega; \quad \Sigma = \phi_1 I + \phi_2 R; \quad \beta = (\beta_1, \dots, \beta_p); \quad Z = X\beta + \varepsilon; \quad \omega = (\omega_1, \dots, \omega_n)^\top$$

$$\Delta_{R\phi} = \frac{\partial^2 l(\theta/\omega)}{\partial \phi_j \partial \omega^\top}$$

$$\begin{aligned}
\frac{\partial l(\theta/\omega)}{\partial \omega^\top} &= -\frac{1}{2} \left[\frac{\partial (Z\omega - X\beta)^\top \Sigma^{-1} (Z\omega - X\beta)}{\partial \omega^\top} \right] \\
&= -\frac{1}{2} \left[(Z\omega - X\beta) \Sigma^{-1} \frac{\partial (Z\omega - X\beta)}{\partial \omega^\top} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial (Z\omega - X\beta)}{\partial \omega^\top} \right] \\
&= -\frac{1}{2} \left[(Z\omega - X\beta) \Sigma^{-1} \frac{\partial Z\omega}{\partial \omega^\top} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial Z\omega}{\partial \omega^\top} \right] \\
&= -\frac{1}{2} \left[(Z\omega - X\beta) \Sigma^{-1} \frac{\partial (Z - \Sigma^{1/2} \omega)}{\partial \omega^\top} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial (Z - \Sigma^{1/2} \omega)}{\partial \omega^\top} \right] \\
&= \frac{1}{2} \left[(Z\omega - X\beta) \Sigma^{-1} \frac{\partial (\Sigma^{1/2} \omega)}{\partial \omega^\top} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial (\Sigma^{1/2} \omega)}{\partial \omega^\top} \right] \\
&= \frac{1}{2} \left[(Z\omega - X\beta) \Sigma^{-1} \Sigma^{1/2} + (Z\omega - X\beta)^\top \Sigma^{-1} \Sigma^{1/2} \right] \\
&= \frac{1}{2} \left[2(Z\omega - X\beta) \Sigma^{-1} \Sigma^{1/2} \right] \\
&= (Z\omega - X\beta)^\top \Sigma^{-1} \Sigma^{1/2}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l(\theta/\omega)}{\partial \phi_j \partial \omega^\top} &= \frac{\partial (Z\omega - X\beta)^\top \Sigma^{-1} \Sigma^{1/2}}{\partial \phi_j} \\
&= \left[\frac{\partial (Z\omega - X\beta)^\top}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} + (Z\omega - X\beta)^\top \frac{\partial \Sigma^{-1}}{\partial \phi_j} \Sigma^{1/2} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma^{1/2}}{\partial \phi_j} \right] \\
&= \left[\frac{\partial (Z\omega)^\top}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} - (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma^{1/2}}{\partial \phi_j} \right] \\
&= \left[\frac{\partial (Z - \Sigma^{1/2} \omega)^\top}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} - (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma^{1/2}}{\partial \phi_j} \right] \\
&= \left[-\frac{\partial (\Sigma^{1/2} \omega)^\top}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} - (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma^{1/2}}{\partial \phi_j} \right] \\
&= \left[-\omega^\top \frac{\partial \Sigma^{1/2}}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} - (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2} + (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma^{1/2}}{\partial \phi_j} \right] \\
&= (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma^{1/2}}{\partial \phi_j} - (Z\omega - X\beta)^\top \Sigma^{-1} \frac{\partial \Sigma}{\partial \phi_j} \Sigma^{-1} \Sigma^{1/2}
\end{aligned}$$

APÊNDICE B – PROVA

MATRIZ CHAPÉU/MATRIZ DE PROJEÇÃO (H)

Seja $H = X(X^T X)^{-1} X^T$, para provar que a matriz H é simétrica, devemos mostrar que a matriz $A = A^T$. Dessa forma, para provar que a matriz H é simétrica, vamos mostrar que $H = H^T$:

$$H^T = [X(X^T X)^{-1} X^T]^T$$

$$H^T = X [(X^T X)^{-1}]^T X^T$$

$$H^T = X [(X^T X)^T]^{-1} X^T$$

$$H^T = X(X^T X)^{-1} X^T$$

$$H^T = H \quad \blacksquare$$