MÉTODO DE PARTIÇÃO PRODUTO APLICADO A KRIGAGEM

Maria de Fátima Ferreira Almeida

Tese apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para fins de obtenção do título de Doutora em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro - 2019

MÉTODO DE PARTIÇÃO PRODUTO APLICADO A KRIGAGEM

Maria de Fátima Ferreira Almeida

Orientador: Prof. Dr. **José Sílvio Govone**

Coorientador: Prof. Dr. **Gérson Rodrigues dos Santos**

Tese apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para fins de Exame Geral de defesa de tese, como parte dos requisitos para obtenção do título de Doutora em Biometria.

BOTUCATU
São Paulo - Brasil
Fevereiro - 2019

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Almeida, Maria de Fátima Ferreira.
Método de partição produto aplicado a Krigagem / Maria de Fátima Ferreira Almeida. - Botucatu, 2019

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Sílvio Govone
Coorientador: Gérson Rodrigues dos Santos
Capes: 90000005

1. Krigagem. 2. Geologia - Métodos estatísticos. 3. Biometria - Estudos longitudinais. 4. Análise espacial (Estatística). 5. Teoria bayesiana de decisão estatística.

Palavras-chave: Modelo hierárquico Bayesiano; Acurácia; Krigagem; Ponto de mudança.

Dedicatória

Á Deus pela infinita bondade, pelas bênçãos recebidas, pela saúde e equilíbrio meu e de minha família, pelo dom da coragem e pela crença de que a distância física no trajeto para o doutorado era mais curta que nos mapas de voo e nas rotas dos ônibus e pela tranquilidade para a realização do meu trabalho.

Ao meu orientador José Sílvio Govone, por ter acreditado em mim e não medir esforços, orientando com paciência e sempre ponderando algumas limitações de distância, trabalho e familiares, que muitas vezes me impediram cumprir as tarefas em tempo hábil.

Aos meus pais, Xisto F. dos Santos e Percília B. dos Santos, que mesmo desconhecendo o teor do meu trabalho sempre valorizaram a escola e com simplicidade transmitiram mensagens de apoio que se tornaram a cada dia, mais importantes que os ensinamentos dos livros.

Às minhas filhas, Magaly Stefânia e Luma Fabiane e ao meu marido Geraldo A. Almeida, aos meus irmãos, sobrinhos e netos: Miguel e Maria Luiza, por ficarem do meu lado apoiando e tornando mais leves e alegres os meus dias.

Aos amigos que se mantiveram ao meu lado ao longo de todo o curso e aos que mesmo distantes compartilharam conhecimentos e apoio.

À minha querida professora primária, Dona Ailce Almeida, pela dedicação e a todos os demais professores que ao longo da minha vida estudantil, contríbuiram para que eu chegasse até aqui.

. . . um dia, conversando com Deus, eu disse que chegava,. . . e ele me trouxe até aqui, . . . ai eu cheguei. Obrigada meu Deus.

Palavras não são insuficientes para demonstrar a minha gratidão!

"*Bom mesmo é ir à luta com determinação, abraçar a vida com paixão, perder com classe e vencer com ousadia, porque o mundo pertence a quem se atreve e a vida é* ***muito*** *para ser insignificante*"

(Charlie Chaplin)

Agradecimentos

Aos professores, técnicos e auxiliares do Departamento de Bioestatística/ UNESP pelos conhecimentos e dedicação.

À Universidade Federal de Viçosa (UFV) pelo apoio à mobilidade acadêmica, em especial ao professor Dr. Gerson Rodrigues dos Santos (Programa de Pós Graduação em Estatística Aplicada e Biometria - PPESTBIO) pela coorientação, ao professor Fabyano Fonseca e Silva (Programa de Pós Graduação em Zootecnia) pelo apoio e discussões relevantes para esta tese e ao Grupo de Estudos e Pesquisas em Levantamentos Hidrográficos- GEPLH na pessoa do professor Dr. Ítalo Ferreira (Departamento de Engenharia Civil) pelo apoio e por ceder os dados para a pesquisa.

Ao Instituto Federal de Educação do Norte de Minas Gerais (IFNMG) inclusive colegas de trabalho e estudantes, pelo apoio.

Aos amigos que se mantiveram ao meu lado ao longo de todo o curso, e aos que mesmo distantes se fizeram presentes, em especial: Jacqueline Domingues, Vívian Brancaglioni, Felipe Teles, Alex Santos, Márcio Rodrigues, Egídio Martins e Clênia Toletino.

À CAPES, IFNMG e CNPQ, pelo apoio financeiro.

Sumário

Lista de Figuras

Página

Lista de Tabelas

Página

MÉTODO DE PARTIÇÃO PRODUTO APLICADO A KRIGAGEM

Maria de Fátima Ferreira Almeida: MARIA DE FÁTIMA FERREIRA ALMEIDA
Orientador: Prof. Dr. JOSÉ SÍLVIO GOVONE

RESUMO

As variáveis aleatórias no espaço estão definidas por *funções aleatórias* sujeitas à teoria das variáveis regionalizadas. Para assumir continuidade espacial com um número limitado de realizações da variável aleatória são necessárias as hipóteses de estacionariedade, as quais envolvem diferentes graus de homogeneidade espacial. Formalmente, uma variável regionalizada Z é estacionária se os momentos estatísticos de $Z(s+h)$ forem os mesmos para qualquer vetor h. A hipótese de estacionaridade de primeira ordem é definida como a hipótese de que o momento de primeira ordem da distribuição da função aleatória $Z(s)$ é constante em toda a área. A hipótese intrínseca é baseada no cálculo de médias globais das semivariâncias, com a pressuposição de estacionaridade de 1^a ordem e da estacionaridade da variância dos incrementos. Embora muitas variáveis sejam sucetíveis a dupla ou múltipla estacionaridade, estas estruturas espaciais não são levadas em consideração pelo semivariograma usual, e, consequentemente, causam sérios problemas de acurácia nos mapas

de Krigagem. Na perspectiva de solucionar o problema apontado, buscou-se identificar os locais dos pontos de mudança na média que definem mais de uma estrutura de semivariância, com o objetivo de melhorar a qualidade dos mapas de Krigagem Ordinária. Para isso, foi utilizado o Método de Partição Produto (MPP), com enfoque espacial, denominado Método de Partição Produto Espacial (MPPs). Para separar os grupos, foi criada uma função de busca de ponto de mudança na média utilizando o modelo hierárquico bayesiano, denominado Modelo de Partição Produto Espacial (MPPs). Utilizou-se dois bancos de dados para testar o potencial do modelo em separar grupos espacialmente dependentes, em que no primeiro havia suspeita de uma mudança na média, enquanto no segundo, "Dados2", havia suspeita de dois pontos de mudança na média. Na análise do primeiro banco de dados, o primeiro grupo, apesar de não obter um ajuste do semivariograma totalmente satisfatório, ainda assim obteve boa acurácia no mapa, enquanto que o segundo grupo, observou-se um ajuste satisfatório a um modelo diferente de semivariograma e obteve-se melhor acurácia, superando o primeiro grupo e o conjunto de amostra completa. No segundo banco de dados, "Dados 2", os três grupos se ajustaram a três semivariogramas distintos e geraram três mapas de Krigagem, nos quais, os mapas gerados para as três subáreas mostraram resultados satisfatórios, comprovando que a qualidade de ambos superaram a qualidade do mapa de krigagem feito para a amostra completa. Por meio dos resultados obtidos nos dois bancos de dados, concluiu-se que o método MPPs aplicado à Krigagem Ordinária garante mapas de melhor qualidade, por apresentar estimativas mais acuradas.

Palavras-chave: Ponto de Mudança, Método hierárquico Bayesiano espacial de partição produto - MPPs, Krigagem, Acurácia de mapas.

KRIGING INDUCED VIA METHOD OF SPATIAL PRODUCT PARTITION

Author: MARIA DE FÁTIMA FERREIRA ALMEIDA
Adviser: Prof. Dr. JOSÉ SÍLVIO GOVONE

SUMMARY

The random variables in space are defined by random functions subject to regionalized variable theory. To assume spatial continuity with a limited number of realization of the random variable, we need to assume stationarity hypotheses, which involve different degrees of spatial homogeneity. Formally, a regionalized variable Z is stationary if statistical moments of $Z(s+h)$ are the same for any vector h. The first order stationarity hypothesis is defined to be the hypothesis that first order moment of the distribution of the random function $Z(s)$ is constant throughout the area. The intrinsic hypothesis is based on the computation of global means of semivariate models, with the assumption of 1^{st} order stationarity and incremental variation stationarity. Although many variables are capable of double or multiple stationarity, these spatial structures are not taken into account by the usual semivariogram, and, consequently, cause acuracy problems in Kriging maps. In order

to solve the described problem, it was identify the points of change in the average with the objective of improving the quality and accuracy of the maps of Ordinary Kriging. To separate the groups, a mean change point search function was created using the Bayesian hierarchical model, called the Space Product Partition Model (MPPs). Two databases were used to test the model's potential to separate spatially dependent groups, in which the former suspected a change in mean while in the latter. " Data2 ", there were suspicion of two points of change in the average. n the analysis of the first database, the first group, although it did not obtain a totally satisfactory semivariogram adjustment, nevertheless obtained good accuracy in the map, whereas the second group, a satisfactory adjustment was observed to a different model of semivariogram and we obtained better accuracy, surpassing the first group and the complete sample set. In the second database, Data 2, the three groups conformed to three distinct semivariograms and generated three Kriging maps. In the three subareas, referring to the second database, "Data2", the results proved that the quality of the three maps exceeded the quality of the kriging map made for the complete sample. Based on the results obtained, it was concluded that the MPPs method applied to Ordinary Kriging guarantees maps of better quality, since they present more accurate estimates. **Keywords**: Point of Change, Bayesian Hierarchical Model MPPs, Ordinary Kriging, Accuracy.

1 INTRODUÇÃO

A representação quantitativa de fenômenos físicos, assim como a predição de valores desconhecidos, sempre foi preocupação da ciência. Para tais representações, predições são feitas por meio de métodos probabilísticos (que levam em consideração a incerteza através de modelos probabilísticos) ou mesmo determinísticos (que desconsideram a incerteza que possa circundar o fenômeno em estudo). A Geoestatística utiliza de modelos probabilísticos para descrever os processos estocásticos espaciais contínuos.

De acordo com Vieira (2000) a Geoestatística teve seu limiar entre os anos de 1957 a 1962, quando o engenheiro de minas, francês, *G. Matheron*, de posse das observações de D.G. Krige, engenheiro de minas Sul africano, desenvolveu a *Teoria das variáveis regionalizadas*, representadas na prática por certa quantidade de dados brutos disponíveis, a partir dos quais foram obtidas as informações sobre as características do fenômeno. A partir do limiar da Geoestatística, muitos métodos de krigagem foram propostos e formalizados obedecendo a subdivisão em Krigagens lineares e não lineares. As Krigagens Lineares são assim denominadas porque os estimadores representam combinações lineares de seus parâmetros.

Landim (2011) destaca que até 1968 a Geoestatística foi utilizada para obter estimativas de reservas de hidrocarbonetos, e entre 1968 a 1970 foi fundamentada a Teoria da Krigagem Universal cujo nome foi dado por G. Matheron, em homenagem a D.G. Krige, para aplicação à cartografia submarina com tendência sistemática. De acordo com o autor, partir de 1970 muito tem se desenvolvido na Geoestatística, destacando sua ampla utilização no campo das ciências agrárias, geologia aplicada à agricultura de precisão e à preservação ambiental, dentre outros

autores, cita-se (Ramirez & Colin, 1994; da Silva et al., 2012).

Huijbregts (1975) destaca que a Geoestatística explora a aparente aleatoriedade dos dados para avaliar as medidas de correlação espacial dos mesmos, considerando uma determinada vizinhança.

Dentre as técnicas de Krigagem linear as mais utilizadas são: a krigagem simples, a krigagem ordinária e a krigagem universal, enquanto que as krigagens não lineares, destacam-se a krigagem Disjuntiva, a krigagem Indicativa e a krigagem Lognormal (seção 2.3 do capítulo 2). Diversas aplicações do método de Krigagem Ordinária (KO) e Krigagem Indicativa (KI) são encontradas na agricultura, entre outros trabalhos, destacam-se (Imai et al., 2003; da Silva et al., 2012).

Os métodos geoestatísticos de Krigagem são baseados na função semivariograma a qual é calculada por meio de médias globais das semivariancias para cada distância entre os pontos. Estes métodos utilizam os valores estimados do semivariograma com a pressuposição de estacionaridade de 1^a ordem, por esse motivo a estacionaridade da média é condição essencial para se aplicar geoestatística. Atendida a pressuposição da estacionaridade da média, tem-se a garantia de variância mínima dos erros e média dos erros tendendo a zero.

Apesar da estacionaridade da média, denominada estacionaridade de primeira ordem, ser condição essencial para se aplicar geoestatística, muitos fatores tais como, custo de pesquisa, ausência de delineamento e a inexistência de método objetivo para medí-la, impedem ou dificultam comprovar a falta de estacionaridade para qualquer variável em diferentes tamanhos de malha amostral.

O método usual para medir a estacionaridade é o semivariograma, que é uma função baseada nas médias da dependência espacial em função das distâncias dentro de uma janela móvel. Por ser baseada em médias móveis, esta função não possibilita calcular os valores das dependências espaciais em todas as sub-regiões da malha amostral. A baixa dependência espacial pode ser devida a presença de dupla ou múltipla estacionaridade na estrutura de semivariancia, que exige ajuste de mais de um semivariograma, ficando um único modelo de semivariograma, inadequado

para cobrir todas as sub-regiões da malha amostral.

Devido a dificuldade em detectar a estacionaridade de primeira ordem, ou sua ausência em algumas partes da malha amostral, a geoestatística tem sido usada, mesmo quando esta estacionaridade nãoé totalmente atendida, resultando em mapas de Krigagem de pouca acurácia.

Na expectativa de encontrar maneiras de melhorar a acurácia dos mapas de krigagem, muitos pesquisadores têm se dedicado ao estudo da qualidade dos mapas de krigagem e as relações existentes com o tamanho e comportamento amostral, como por exemplo, o trabalho de (Uribe-Opazo et al., 2012) mostrarando que o tamanho amostral não é fator determinante, pois, outras condições precisam ser atendidas para garantir a qualidade dos mapas.

Uma suposição sobre os fatores que interferem na qualidade dos mapas de krigagem podem estar associados ao comportamento dos dados. Tentativas de otimização da malha utilizando a geoestatística foram também propostas por (Oliver & Webster, 2015), porém, limitando-se ao estudo de caso específico para delineamentos planejados.

Muitos dos estudos feitos até então não avançaram em relação ao problema da perda de qualidade dos mapas de Krigagem em geral, por não associarem o problema dos mapas gerados com a possibilidade da presença de duas ou mais médias estacionárias, que causam perda da dependência espacial, devido a presença de pontos de mudança na estrutura espacial.

Page & Quintana (2016) apontaram problemas de acurácia em mapas de krigagem Ordinária os quais atribuíram aos problemas de agrupamento implícitos na estrutura de dependência espacial, que frequentemente não são levados em consideração nos modelos de predição, ficando tais mapas restritos a estruturas de covariância globais que mascaram tendências nos dados e produzem mapas suavizados.

Problemas como os descritos no parágrafo anterior foram abordados recentemente por Page & Quintana (2016) os quais propuseram modelos bayesianos baseadas em Modelos de Partição Produto(MPP), que modelando diretamente o

agrupamento de locais em *clusters* espacialmente dependentes garantem que a configuração espacial em locais distantes tenham probabilidades pequenas, enquanto que em locais próximos tenham altas probabilidades. Para isto, propuseram modelos que geram distribuições a posteriori capazes de captar as características locais de cada cenário e estabelecer os locais de pontos de mudança estruturais na média, baseando-se nas distâncias entre as amostras.

A metodologia sugerida por Page & Quintana (2016) apresenta-se dificuldade computacional quando o tamanho da amostra a ser utilizada é grande, porque o artigo foi discutido por meio de exemplo utilizando um conjunto de poucos pontos amostrais fictícios fazendo-se uso de um método de partição aleatória. Porém uma partição aleatória requer testar todas as combinações possíveis das partes do conjunto de dados, o que torna inviável computacionalmente em bancos de dados grandes, como é o caso comumente utilizado em geoestatística.

Os Modelos de Partição Produto(MPP), foram utilizados até então para identificar pontos de mudança na média em conjuntos amostrais cujas variáveis se referiam a taxas de doença em dados de área (Barry & Hartigan, 1992) utilizando matriz de contiguidade de borda e em amostras ordenadas em função do tempo(Yao, 1984b; Loshi & Cruz, 2005), dentre outros casos envolvendo séries temporais.

Em modelos Bayesianos, θ é o parâmetro de interesse e o seu verdadeiro valor é desconhecido. Com o intuito de tentar reduzir o grau de incerteza que se tem sobre o verdadeiro parâmetro, modelos probabilísticos são propostos para descrever o comportamento de θ.

O método proposto nesta tese tem o objetivo de encontrar o local k do ponto de mudança na média μ utilizando MPPs, para isso, é apresentado o modelo

$$\begin{aligned} \theta &= \alpha + \beta_j X + \epsilon \\ X &= (x_i - x_k) \end{aligned}$$

em que α é a esperança de Y ser o ponto de mudança, β é o parâmetro que indexa a covariável X em que $X = (x_i - x_k)$ e ϵ é o erro aleatório.

No modelo, X, a variável regressora, refere-se a diferença entre as distâncias médias de cada ponto em relação aos seus vizinhos, colocada em ordem crescente, critério adotado de acordo com (Barry & Hartigan, 1992).

Para um valor fixo de Y, a função $l(\theta; y) = p(y|\theta)$ fornece a plausibilidade ou verossimilhança de cada um dos possíveis valores de θ, enquanto $p(\theta)$ representa as informações que se supõe conhecer a priori, chamada distribuição a priori de θ. Com estas duas fontes de informação, a informação a priori e a informação da verossimilhança são expressas como produto e a distribuição a posteriori de θ, $p(\theta|y)$ (Ehlers, 2011) é traduzida em

$$p(\theta|y) \propto p(Y = y|\theta)\, p(\theta)$$

em que $p(Y = y|\theta)$ representa a verossimilhança dos dados.

A distribuição a posteriori representa o produto da verossimilhança pela distribuição a priori, em que no modelo é descrita por:

No modelo com um ponto de mudança na média (corte),

$$p_i(\alpha, \beta_1, \beta_2, k|y) \propto \text{Verossimilhança} \times \text{prioris}$$

$$p_i(\alpha, \beta_1, \beta_2, k|y) \propto N(\mu_i, \sigma^2)\Gamma(0,001; 0,001)N(0,00; 0,001)N(0,1)cat(U\{1,2\})$$

No modelo com dois pontos de mudança na média (cortes),acrescenta-se β_3 e k_1, em que k e k_1 representam as duas posições dos pontos de mudança na média no espaço e β_3 refere-se ao coeficiente que indexa o terceiro grupo.

Na construção do modelo considera-se que $Y_i \sim Normal(\mu, \sigma^2)$, $i = \{1, \ldots, n\}$, a função de distribuição dos dados escrita por meio de um modelo hierárquico normal, μ representa uma regressão nos parâmetros da covariável X e J_i representa uma variável categórica ($cat\{1,2\}$) com $J_i = 1$, para o grupo 1 se $i \leq k$ e $J_i = 2$ para o grupo 2, se $i > k$, para o caso de suspeita de um ponto de mudança na média no espaço amostral, ou J_i representa uma variável categórica ($cat\{1,2,3\}$) que assume $J_i = 1$, para o grupo 1 se $i \leq k$, $J_i = 2$ para o grupo 2, se $i < k < k_1$ e $J_i = 3$ para o grupo 3, se $i \geq k_1$, no caso de suspeita de dois pontos de mudança na média amostral.

Para o modelo espacial hierárquico bayesiano de ponto de mudança (MPPs), θ é o parâmetro que representa a esperança do ponto de mudança da média amostral, regido pela covariável espacial X(distância média entre os vizinhos), com 95% de credibilidade, em que $X = x_1 - x_k$.

Fazendo a esperança $E[Y] = \alpha$ no ponto de mudança, os coeficientes β_1 e β_2, indexam os grupos antes do ponto de mudança e após o ponto de mudança, respectivamente, e obtém-se o modelo preditivo posteriori cuja fórmula é expressa por um modelo hierárquico que não tem forma analítica conhecida, apresentado em (1):

$$\begin{aligned} Y_i &\sim N(\mu, \sigma^2) \\ \mu &= \alpha + \beta_{J_i}(x_i - x_k) \\ J_i &= 1, \text{se} \qquad i \leq k \\ J_i &= 2, \text{se} \qquad i < k_1 \\ \sigma &= \tau^{-1} \end{aligned} \tag{1}$$

em que $y_i = \gamma_i$, vetor de resposta, é composto pelas semivariancias locais dos dados, a qual é assumida seguir distribuição normal. O modelo para dois pontos de mudança na média é similar ao modelo (1) feito para um ponto de mudança na média, acrescentando-se apenas os parâmetros que descrevem um terceiro grupo, como mostrado na expressão (64) do capítulo 4.

No modelo (1) são assumidas cinco priores, sendo duas delas não informativas que representam τ e α, uma priori informativa β que assume normal padrão, uma distribuição uniforme que descreve os valores de k e k_1 e uma uniforme categórica que descreve os grupos j e uma variável categorica, $cat(U\{1, 2\})$, em que foram assumidas duas categorias, $\{1, 2\}$. São utilizados também valores iniciais para k e k_1, valores iniciais para os parâmetros betas e um valor inicial para τ para gerar as distribuições. O local do ponto de mudança na média, k, é a variável aleatória de interesse estimada pelo modelo, uniformemente distribuída de $1, \ldots, n$.

Os modelos foram programados em linguagem R CORE Team (2015)

e WINBUGS Lunn et al. (2000), em que se verifica no modelo (1), após ser obtido k, assume-se que as observações seguem duas distribuições normais independentes $y_i|\phi \sim N(\mu_1, \tau), i = 1, \ldots, k$ e $y_i|\gamma \sim N(\mu_2, \tau), i = k+1, \ldots, n$, em que ϕ e γ representam as distribuições, uma para cada parte da partição dos dados, onde o parâmetro de média da distribuição normal de cada uma das partes é descrito por um modelo de regressão $\mu_i = \alpha + \beta_{J_i}(x_i - x_k)$ e $\tau \sim Gama(\alpha, \beta_j)$.

Na modelagem de dois pontos de mudança na média, a quantidade de parâmetros a ser estimados pelo modelo aumenta, como é o caso de k e β que passa a ter dois parâmetros (k e k_1) e três parâmetros, (β_1, β_2 e β_3) que definem três grupos. Consequentemente, pode se verificar que aumentando o número de pontos de mudanças na média, aumenta não somente o número de parâmetros do modelo, mas também o tempo de processamento.

Após ser definir os grupos (partes da partição) feito pelo MPPs, é aplicada a geoestatística com ajuste dos semivariogramas às partes e um mapa de krigagem para cada uma, quando possível a ambas. Se apenas algumas das partes apresentam dependência espacial suficientes, ás partes com dependências espaciais insuficientes são aplicadas outras metodologias ou a mesma metodologia com condições limitadas e assumindo que estes mapas, sob tais condições, são pouco confiáveis.

Para o modelo com dois pontos de mudança na média a quantidade de parâmetros a ser estimados pelo modelo aumenta, como é o caso de k e β que passa a ter dois e três parâmetros, respectivamente. De acordo com as características do modelo (1), pode se observar que aumentando o número de pontos de mudanças na média, o número de parâmetros do modelo e o tempo de processamento também aumentam.

O local do ponto de mudança na média, k, é a variável aleatória de interesse estimada pelo modelo, uniformemente distribuída de $1, \ldots, n$.

Para o modelo (1), dado k, assume-se que as observações seguem duas distribuições normais independentes $y_i|\phi \sim N(\mu_1, \tau), i = 1, \ldots, k$ e $y_i|\gamma \sim N(\mu_2, \tau), i = k+1, \ldots, n$, em que ϕ e γ representam as distribuições, uma para cada

parte da partição dos dados, onde o parâmetro de média da distribuição normal de cada uma das partes é descrito por um modelo de regressão $\mu_i = \alpha + \beta_{J_i}(x_i - x_k)$ e $\tau \sim Gama(\alpha, \beta_j)$.

Os fenômenos abrangidos pela Geoestatística são do tipo estruturados e se caracterizam por sua variável aleatória ser contínua no espaço e possuir dependência espacial, por isso a estrutura de semivariancia é considerada função do ajuste de parâmetros globais, assumindo assim a pressuposição de estacionaridade da média e em alguns casos, a finitude da variância e da covariância.

Os casos em que as pressuposições de estacionaridade de 1^a e 2^a ordens (apresentada com maiores detalhes no capítulo 2, na seção 2.2.1), quando válidas para todas as direções, são ditos fenômenos isotrópicos, raramente encontrados. Para muitos casos, estas pressuposições são assumidas, mas não são totalmente atendidas, sendo consideradas válidas por falta de um método objetivo de mensuração.

Os métodos de krigagem adotam o semivariograma ominidirecional que é uma média da estrutura global de semivariancia em todas as direções para cada distância entre pontos. No entanto, a semivariancia calculada nem sempre é a mesma em todas as direções e locais da área e quando a pressuposição de estacionaridade de primeira ordem não é totalmente atendida, os mapas gerados explicam muito pouco do fenômeno em estudo, embora feitos para cobrir toda a área porque quando a variável espacial tem aparentemente, mais de uma estrutura de covariância, ao se ajustar um modelo único de semivariograma estas estruturas não são levadas em consideração e mascaram a dependência espacial pela contaminação dos dados. Em malhas amostrais com estas características, ao se ajustar o semivariograma global ocorre uma suavização exagerada nos pontos estimados, provocada pela contaminação dos dados. Além disso, ao assumir um modelo único, impõe-se uma estrutura de covariância enganosa na estimação, levando a baixa acurácia nos mapas gerados por Krigagem.

A constatação apresentada no parágrafo anterior leva a suposição que o MPP possa ser adaptado a um método denominado Modelo de Partição Pro-

duto Espacial (MPPs), para identificar mudanças na estrutura espacial da média em dados com dependência espacial, ou seja com presença de duas ou mais médias estacionárias, que mascaram a estrutura de semivariancia espacial.

Na perspectiva de solucionar o problema apontado nos parágrafos anteriores, pretende-se identificar pontos de mudança na média que provocam a mudança da estrutura de semivariancia e estabelecer cortes na malha amostral, com o objetivo de melhorar a qualidade e acurácia dos mapas gerados por Krigagem Ordinária. Para isso, nesta tese é desenvolvido o MPP dando enfoque espacial, que denomina-se aqui, Modelo de Partição Produto espacial(MPPs), que é um método bayesiano, cuja teoria está fundamentada em (Smith, 1975) e primeiro uso prático foi proposto em sua forma geral por Hartigan (1990) para identificar pontos de mudanças na média.

O objetivo desta tese consiste em criar um novo Método de Partição Produto Espacial (MPPs) direcionado à aplicação de Krigagem Ordinária como uma metodologia para resolver o problema da dupla ou múltipla estacionaridade da média de uma variável espacialmente distribuída para garantir que os mapas de Krigagem tenham maior acurácia.

O MPP convencional garante que as médias de cada grupo sejam mais homogêneas, porém não garante que os grupos tenham dependência espacial, como proposta de garanti-la, sugere-se na tese, a transformação dos dados utilizando como variável resposta as semivariancias locais, cada uma correspondente a um ponto, e como covariável a distância média entre vizinhos, para que o ponto de mudança, sendo regido pela distancia média entre vizinhos, possa identificar o local da mudança da média em função da distancia entre as amostras. Os pontos de mudança na média serão os locais dos cortes na malha amostral que formam as partes da partição.

Esta tese foi pensada com o potencial de atingir várias áreas que utilizam-se da geoestatística e por isso, pretende-se propor uma metodologia para usuários de geoestatística e estatística espacial aplicada a diversos campos das ciências, sendo assim, a ordenação dos capítulos foi pensada com o propósito de apresentar a metodologia de forma simples, clara e prática.

A escolha de não adotar métodos clássicos de agrupamento, facilmente disponíveis em alguns trabalhos, deve-se ao fato que nestes métodos não se tem a garantia que os grupos terão dependência espacial e nem tão pouco que serão mais prováveis de obter melhores estimativas, por não se basearem em distribuições de probabilidades, enquanto, que o MPPs é um método bayesiano, por isso, baseia-se em distribuições de probabilidades por assumirem funções de probabilidade que geram os dados (de onde obtém-se a verossimilhança) e funções de probabilidade a priori para os parâmetros; podendo incorporar informações externas aos dados, chamadas informações a priori para se obter a distribuição a posteriori para as estimativas.

A aplicação da métodologia MPPs tem como primeiro passo a transformação da variável resposta em semivariancias locais e o uso da covariável espacial " Distancia Média entre Vizinhos". Essa transformação e a restrição do número de cortes à supostas quantidades de pontos de mudança na média, estabelecidas para permitir gerar agrupamentos com mais pontos amostrais localizados na mesma subregião do seu grupo e com dependência espacial, condição necessária para aplicação em geoestatística.

O conjunto completo da tese está dividido em seis capítulos: Introdução, Geoestatística, Método de Partição Produto, Material e métodos, Resultados e discussão e Conclusão, ficando no primeiro capítulo reservado à Introdução, o segundo capítulo incumbido de apresentar um estudo sobre Geoestatística iniciando-se com o tema "Processos Estocásticos Espaciais", seguindo por "Variáveis Regionalizadas", "Semivariograma" e finalizando com o estudo dos principais "Métodos de Krigagem".

No terceiro capítulo apresenta-se uma revisão sobre o Método de Partição Produto (MPP) em sua proposta original e espacial, com um tópico específico sobre pontos de mudança na média sob o enfoque bayesiano.

No quarto capítulo apresenta-se a Metodologia MPPs (Método de Partição Produto Spacial) aplicada ao método de Krigagem, que qualifica o método utlizado nesta tese.

O quinto capítulo apresenta os Resultados e Discussão do método aplicado a dados altimétricos e batimétricos, e a seguir, no sexto capítulo apresenta a Conclusão, em que são levantadas, também, algumas sugestões de trabalhos futuros.

Finalmente, apresenta-se todas as referências consultadas para construção da Tese.

2 GEOESTATÍSTICA

2.1 Caracterização da Geoestatística

Fenômenos naturais frequentemente são caracterizados por sua distribuição no espaço por meio de quantidades mensuráveis, chamadas variáveis regionalizadas. Utilizando a conotação etimológica das raízes da "Geociência", a geoestatística foi assim denominada por sua principal e marcante abrangência nas aplicações estatísticas relacionadas aos estudos de ciências da terra.

Embora derivada da geologia e áreas afins, a geoestatística pode ser definida de maneira formal como "a aplicação do formalismo de funções aleatórias ao reconhecimento e predição de fenômenos naturais" (Manzione, 2002), estando deste modo, envolvida em problemas de predição espacial .

O conceito de variáveis regionalizadas, fundamento da teoria geoestatística, parte de processos estocásticos gerais e redefine o espaço de estado como posição no espaço geográfico.

2.1.1 Processos estocásticos

Para a transposição do conceito geral de processos estocásticos para uma abordagem geoestatística, torna-se relevante considerar que os processos estocásticos temporais tem como suporte o tempo (contínuo ou discreto) e este representa uma sequência contínua não negativa do tempo, enquanto que os processos estocásticos espaciais tem como suporte o espaço, os de natureza contínua (designando uma região de abrangência) ou os de natureza discreta (pontos isolados). Para os dois casos, diferentemente do tempo, não existem uma sequência lógica das loca-

lizações, já pré definida. Em ambos os suportes, espaço e tempo, Fernandez (1973) apresentou uma definição clara de processos estocásticos, utilizando-se de um exemplo baseado no fenômeno observado por A. Einstein , N. Wiener, P. Levis dentre outros. Para compreensão da definição, suponha que para cada realização de um experimento aleatório o resultado seja uma função num intervalo real. Considere as duas possibilidades, indicadas na Figura 1.

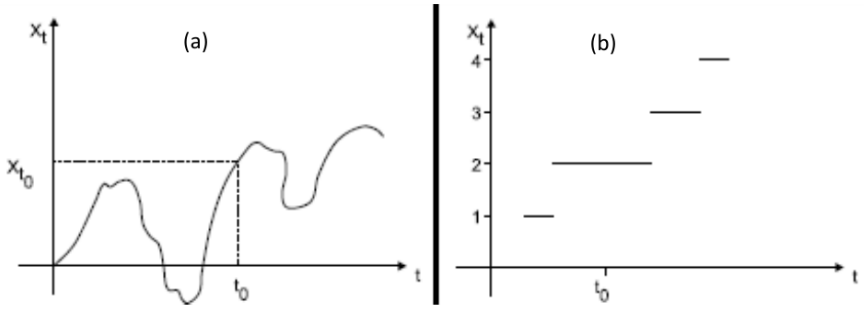

Figura 1: (a) Realização de um experimento aleatório de movimento de partículas de um gás (b) Número de partículas observadas por intervalo de tempo

Fonte: (Fernandez, 1973), adaptado.

Nesta representação, t é usualmente interpretado como tempo, mas poderia ser uma distância a um ponto fixo, um volume ou outra variável.

Suponha que ao tempo 0 tenha-se observado uma partícula de um gás, na origem. Para cada tempo $t \geq 0$ registrou-se o valor da 1^a das coordenadas da Figura $1(a)$ em que a curva da figura representa a trajetória das partículas de um gás em relação ao tempo. Considere que o fenômeno estudado refere-se aos eventos de contar o número de partículas emitidas por este gás num intervalo de tempo $[0, t] \geq 0$. Uma realização típica deste experimento está representada na Figura 1(b). Assim, x_{t_0}, na Figura 1(a) é o valor da 1^a coordenada no instante inicial, e na Figura 1(b), x_{t_0} indica o intervalo de tempo (t_0) indica o número de partículas emitidas no intervalo referente ao $[0, t_0]$. Um gráfico semelhante seria obtido para diferentes fenômenos observados, tais como, o número de mensagens que chegam por minuto no telefone de uma pessoa, o tempo que uma pessoa fica numa fila até ser

atendido, etc.

Pode se dizer que um processo estocástico é um modelo matemático utilizado para estudar fenômenos aleatórios que tem como resultado funções. Essas funções, didaticamente chamadas trajetórias, estão definidas em um espaço de parâmetros T, tomado usualmente como um intervalo da reta real.

Pelo fato de cada ponto w do espaço amostral ω associar uma trajetória, um processo estocástico também pode ser considerado uma função de duas variáveis $w \in \omega$ e $t \in T$, a valores em um conjunto E, chamado *espaço de estados*, usualmente, o conjunto dos números reais não negativos $[0, \infty)$ ou um conjunto finito enumerável.

Fixando t, tem-se uma variável aleatória sobre ω. Assim, todo processo estocástico é uma família de variáveis aleatórias (a valores em E) $X_t : t \in T$.

Atualmente abordagens de fenômenos aleatórios espacias tem sido tema frequente. Tais fenômenos são representados por uma função aleatória, ou seja, por um processo estocástico de índices de uma ou mais dimensões. Pode se observar que para um valor qualquer fixo de t, X_t é uma variável aleatória, que descreve o estado do processo no tempo t.

De acordo com Fernandez (1973) uma realização de um processo estocástico espacial é uma trajetória que a cada ponto do espaço faz corresponder uma variável aleatória.

De maneira formal, Clarke & Disney (1979) definem processo estocástico como um conjunto não vazio T que é chamado espaço paramétrico e na associação com cada $t \in T$ de uma variável aleatória $Z_t : \omega \to E$, todas elas definidas sobre o mesmo espaço de probabilidades.

Cressie (1993) destaca que quando os dados são representados por variáveis aleatórias regionalizadas, estes representam uma realização $z(x)$ de um processo estocástico que é caracterizado por suas distribuições finito-dimensionais, ou seja, de todas as possíveis combinações da distribuição conjunta das variáveis $Z(x_1), \ldots, Z(x_k), \forall k = 1, \ldots, n$.

Como se pode observar, de acordo com a definição de processos estocásticos, uma hipótese ideal para garantir a validade das estimativas da variável de interesse, seria de um processo estocástico gaussiano, ao qual Bhattacharya & Waymire (1990) definem como o conjunto de todos os vetores finito-dimensionais, que possuem distribuição normal multivariada. Porém, esta condição é difícil de ser verificada na prática, por ser impossível observar todas as realizações de um processo estocástico. Devido a este impedimento, é inserido um número mínimo de hipóteses necessárias, chamadas hipóteses de "estacionaridade", a fim de garantir a acurácia das estimativas.

Para compreensão dos processos estocásticos espaciais, o esquema da Figura 5, apresenta-se um organograma em que estão subdivididos os processos estocásticos.

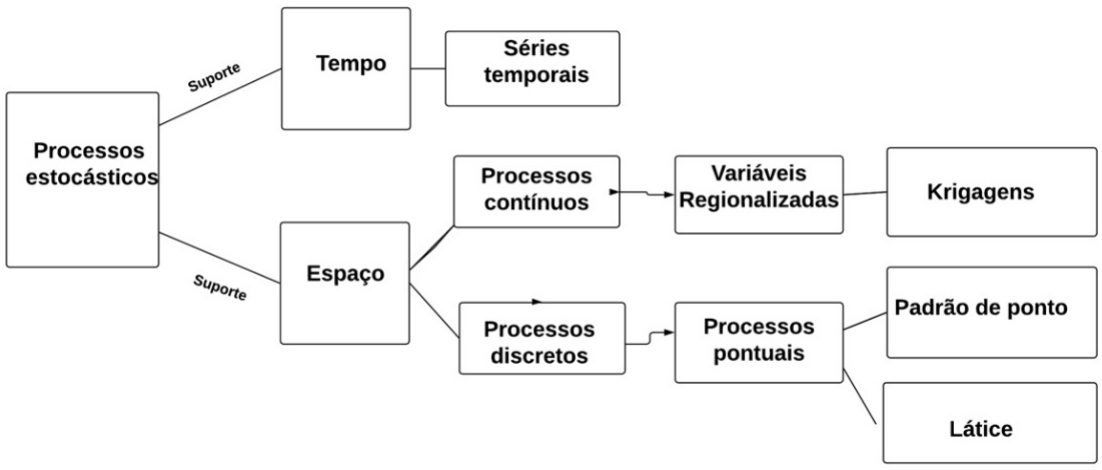

Figura 2: Subdivisões dos processos estocásticos

No organograma da Figura 2, estão representadas as subdivisões dos processos estocásticos os quais compreendem o estudo das séries temporais e dos processos espaciais, tais como látice, padrão de ponto e Krigagem. Destaca-se que, nos processos espaciais, o espaço de estado é definido em uma região geográfica ou em um conjunto de pontos, em posições distintas no espaço geográfico.

2.1.2 Variáveis regionalizadas

Yamamoto & Landim (2013) utilizaram a seguinte explicação para definir variável regionalizada: "ao se retirar uma amostra em um determinado ponto, o teor de uma substância é um valor único, fisicamente determinado". Retirando uma amostra em um ponto próximo, será possível obter um valor diferente do anterior, mesmo que muito próximo, dentro da precisão do método utilizado. Mesmo diferentes, os dois valores estarão correlacionados entre si, se o fenômeno apresentar alguma correlação espacial.

Olea (1975) apud Yamamoto & Landim (2013), define uma variável regionalizada como qualquer função numérica com uma distribuição e variação espacial, mostrando uma continuidade aparente, mas cujas variações não podem ser previstas por uma função determinística.

As variáveis aleatórias no espaço real podem ser unidimensionais, bidimensionais ou tridimensionais e estão definidas por *funções aleatórias* sujeitas à Teoria das Variáveis Regionalizadas. Com base nisso, o valor de uma propriedade do solo, como o pH, em qualquer posição s da área, é apenas uma, das infinitas que são possíveis para representá-lo. Associa-se a cada lugar s não apenas um valor, mas um conjunto completo de valores com uma média, uma variância e momentos de ordem superior de uma distribuição.

Segundo Cressie (1993), os processos estocásticos espaciais podem ser estudados em termos de 1^0 e 2^0 momentos (médias e covariâncias) para amostras finitas. Os modelos estatísticos espaciais decompõem a variação estatística de variáveis aleatórias em um termo de tendência determinística, $\mu(s)$, e um termo estocástico residual, $\epsilon(s)$, como segue ,

$$Y(s) = \mu(s) + \epsilon(s), s \in \Re \tag{2}$$

Nesta relação , $\mu(s)$ é considerado a média de $Y(s)$,

$$\begin{aligned} \epsilon(s) &= Y(s) - \mu(s) \\ E[\epsilon(s)] &= E[Y(s) - \mu(s)] \\ E[\epsilon(s)] &= 0 \end{aligned}$$

$\mu(s)$ é uma função determinística em $\Re$ que se situa como uma função de tendência espacial de valores típicos de um processo estocástico como estrutura global Y e por definição $\epsilon(.)$ é um processo estocástico em $\Re$ com média identicamente nula. Diante disso, $\epsilon(.)$ é entendido como um processo espacial residual e representa as variações locais em torno de $\mu(.)$, a estrutura local do processo de Y.

No contexto de modelagem espacial, levando em consideração que a maioria das variáveis tendem a exibir algum grau de continuidade sobre o espaço, espera-se que estas apresentem valores similares em locais próximos no espaço. Conseguinte a isto, os resíduos espaciais $\epsilon(s)$, por definição, são constituintes de todas as variáveis espaciais não observadas que interferem em Y e não são captadas pela tendência global $\mu(s)$.

Estatisticamente, o fato das amostras mais próximas serem similares, caracteriza-se por exibir dependência estatística positiva.

O modelo estatístico usual de "ruído" como efeitos aleatórios de uma coleção de variáveis aleatórias independentes é restritivo e não representa adequadamente problemas de modelagem espacial. A medida do grau de dependência espacial residual entre duas variáveis aleatórias X e Y é dada em termos da covariância entre elas.

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$
$$Cov(X,Y)) = E[(X - \mu_x)(y - \mu_y)]$$

Existem três configurações distintas para a covariância entre duas variáveis aleatórias, são elas: $Cov(x,y)$ positiva, $Cov(x,y)$ negativa e $Cov(x,y)$ nula (Fernandez, 1973).

Em se tratando de variáveis regionalizadas, as variáveis X e Y são descritas na literatura como $Z(s)$ e $Z(s+h)$, respectivamente, em que s e $s+h$

são locações distintas da variável Z a uma distância h para a qual as variáveis regionalizadas terão sempre covariâncias positivas ou nulas.

Guerra (1988) explica que as variáveis regionalizadas são funções que variam de um lugar para outro da região de estudo, exibindo uma aspecto de continuidade.

A regionalização é o caráter estruturado dos fenômenos e a linguagem que permite tratá-los como tal é a das funções aleatórias. Deste modo, a intuição de que as características em locais próximos são mais propensas de serem semelhantes é formalizada na teoria das funções aleatórias(Andriotti, 2003).

Na teoria das variáveis regionalizadas, $Z(s)$, é definida como uma variável aleatória que assume valores z em função da posição s dentro da área de estudo.

De acordo com Isaaks & Srivastava (1989) o conjunto formado pelas variáveis $Z(s)$ medidas em uma determinada área é considerado uma função aleatória, pois são variáveis aleatórias regionalizadas cuja dependência espacial é explicada por alguma função aleatória de probabilidade. Por serem funções aleatórias, a teoria das variáveis regionalizadas pressupõe que a medida de variação da Variável Aleatória Regionalizada, pode ser expressa pela soma de três componentes, a saber:

- (a) uma componente estrutural associada a um valor médio constante ou uma tendência constante;
- (b) uma componente aleatória espacialmente correlacionada;
- (c) um ruído ou erro aleatório residual.

Assim, fazendo s representar a posição no espaço, o valor predito de Z na posição s, é dado pela equação

$$Z(s) = \mu(s) + \epsilon^{'}(s) + \epsilon^{''} \tag{3}$$

A representação gráfica desta relação funcional é mostrada na Figura 3, por Burrough (1992) e nesta observa-se que diferente da representação feita em Cressie (1993), o

erro estruturado engloba o erro aleatório, fazendo a decomposição do componente de erro em estruturado $\epsilon^{'}(s)$ e aleatório $\epsilon^{''}$.

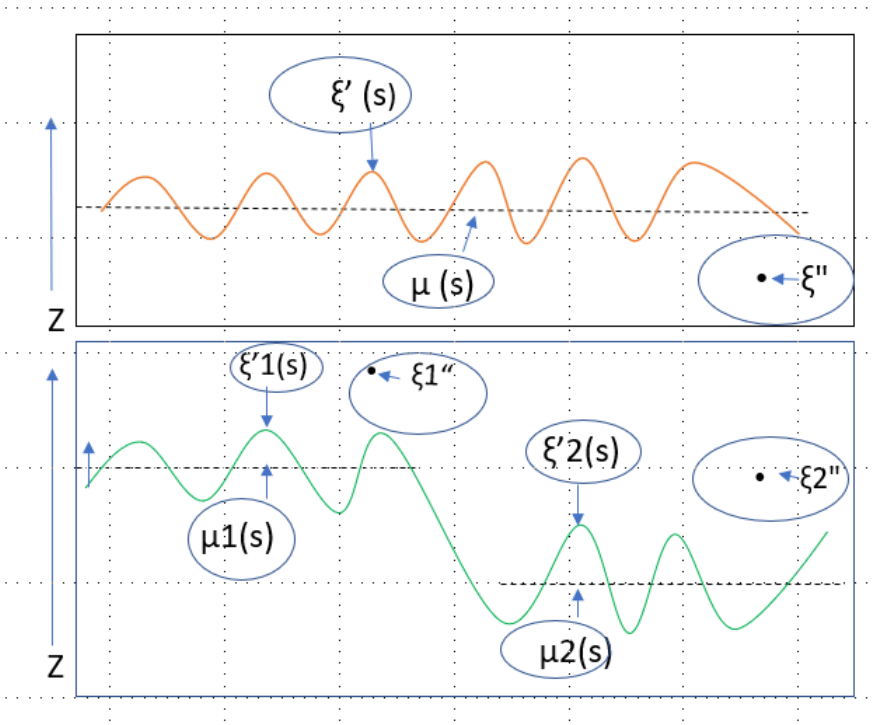

Figura 3: Principais componentes da variação espacial

Fonte: Camargo (1998), adaptado de Burrough (1992)

Burrough (1992), define o comportamento de uma variável aleatória no espaço como a soma de três componentes, a média global como $\mu(s)$, o erro estruturado espacialmente, $\epsilon^{'}(s)$, e o erro aleatório, $\epsilon^{''}$, ao qual estão associados a erros de medidas, erros de delineamento, dentre outros, cujas causas não são identificáveis.

Para estimar uma realização de uma variável aleatória no espaço é necessário estudar todo o conjunto que compõe sua estrutura. Para isto, a consideração da dependência espacial que define a porção do *erro estruturado* viola as pressuposições básicas para estimação pela estatística clássica convencional. Diferente dos métodos convencionais de estimação, a krigagem está fundamentada na teoria das variáveis regionalizadas. A hipótese mais comum é a chamada "estacionaridade de 2^a ordem" (Camargo, 1998):

- A componente determinística, $\mu(s)$, é constante (não há tendências na região)
- A variância das diferenças entre duas amostras depende somente da distância h entre elas.

isto é:

$$Var[Z(s) - Z(s+h)] = E[Z(s) - Z(s+h)]^2 = 2\gamma(h) \tag{4}$$

onde $\gamma(h)$ é chamado de semivariância. Assim, supondo $\mu(s)$ constante, a variação local das amostras (e seu relacionamento espacial) pode ser caracterizado pela semivariância $\gamma(h)$ que é uma função da diferença entre as amostras em função da distância entre elas, representada num gráfico chamado semivariograma.

2.1.3 Estacionaridade

Todos os conceitos da geoestatística têm suas bases em funções e variáveis aleatórias.

Andriotti (2003) define uma função aleatória como estacionária, aquela cuja distribuição de probabilidades é invariante por translação, ou seja, os fatores controladores do seu comportamento agem de forma similar em toda a área em estudo.

A interpretação probabilística de que a variável regionalizada $Z(s)$ é e uma realização particular de certa função aleatória $Z(s_i)$, é consistente quando se pode inferir toda, ou parte da lei de distribuição de probabilidade que define essa função aleatória, (Journel & Huijbregts, 1978).

Como mencionado em 2.1.2, cada ponto s_i, na prática, tem apenas uma realização $Z(s_i)$, assim, o número de pontos é sempre finito, tornando impossível inferir sobre $Z(s)$. Para assumir continuidade espacial, com um número limitado de realização da variável aleatória, faz-se necessário o uso de hipóteses de estacionariedade, as quais envolvem diferentes graus de homogeneidade espacial.

Vieira (2000) afirma que formalmente, uma variável regionalizada Z é estacionária, se os momentos estatísticos de $Z(s+h)$ forem os mesmos, para qualquer vetor h. De acordo com o número k de momentos estatísticos que são constantes, a variável é chamada de estacionária de ordem k.

A hipótese de estacionaridade de primeira ordem é definida como sendo a hipótese de que o momento de primeira ordem da distribuição da função aleatória $Z(s_i)$ é constante em toda a área, ou seja:

$$E[Z(s_i)] = E[Z(s_i + h)] = \mu \tag{5}$$

em que μ é a média dos valores amostrais para todo h, onde h é a distância que separa as amostras;
$E[Z(s_i)]$ = esperança matemática da função aleatória $Z(s_i)$;
$E[Z(s + h)]$ = esperança matemática da função aleatória $Z[(s_i + h)]$.
Segue que:

$$E[Z(s_i) - Z(s_i + h)] = 0. \tag{6}$$

Vieira (2000)define uma função aleatória $Z(s_i)$ estacionária de 2^a ordem se:

- 1^a - O valor esperado $E[Z(s_i)]$ existir e não depender da posição s, ou seja, para qualquer s_i dentro da área A, $E[Z(s_i)] = \mu$.

- 2^a - A cada par de variáveis aleatórias, $Z(s_i)$, $Z(s_i + h)$, a função covariância, $Cov(h)$, existir e for função somente de h.

  A expressão da covariância em função de h é expressa pela fórmula

$$Cov(h) = E[Z(s_i)Z(s_i + h)] - \mu^2 \tag{7}$$

  A hipótese de estacionaridade de segunda ordem 7, implica a existência de uma variância finita dos valores medidos, $Var[Z(si)] = Cov(0)$, mas,em alguns fenômenos, esta hipótese pode não ser satisfeita, devido ao fato que, alguns fenômenos físicos apresentam capacidade de dispersão infinita.

  No operador de primeira ordem ao se considerar a linearidade do valor esperado, E, supondo que a função aleatória $Z(s_i)$ tenha valores esperados $E[Z(s_i)] = \mu(s_i)$ e $E[Z(s_i + h)] = \mu(s_i + h)$, e variâncias $Var[Z(s_i)]$ e $Var[Z(s_i + h)]$, respectivamente, para locais s_i e s_i+h, resulta que a covariância, $Cov(s_i, s_i+h)$, entre $Z(s_i)$ e $Z(s_i + h)$ é dada por,

$$Cov(s_i, s_i + h) = E[Z(s_i)Z(s_i + h)] - \mu(s_i)\mu(s_i + h) \tag{8}$$

  e o variograma, muito importante neste contexto, é definido como $2\gamma(s_i, s_i+h)$ em que

$$2\gamma(s_i, s_i + h) = E[Z(s_i) - Z(s_i + h)]^2 \tag{9}$$

  logo o semivariograma é definido estatisticamente como,

$$\gamma(s_i, s_i + h) = \frac{1}{2}(E[Z(s_i) - Z(s_i + h)]^2) \tag{10}$$

  A variância de $Z(s_i)$ é definida para $h = 0$,

$$\begin{aligned} Var[Z(s_i)] = &\quad E[Z(s_i)Z(s_i + 0) - \mu(s_i)\mu(s_i + 0)] \\ Var[Z(s_i)] = &\quad E[Z^2(s_i) - \mu^2(s_i)] = Cov(s_i, s_i) \end{aligned} \tag{11}$$

  e a variância de $Z(s_i + h)$ é definida como

$$Var[Z(s_i + h)] = \quad E[Z^2(s_i + h) - \mu^2(s_i + h)] = Cov(s_i + h, s_i + h). \tag{12}$$

A justificativa para o uso da semivariância e não da função covariância se deve ao fato de que nos casos de fenômenos que apresentam dispersão infinita, a hipótese de estacionaridade de segunda ordem torna-se muito restritiva. Para esses fenômenos, uma hipótese menos restritiva, chamada hipótese intrínseca, pode ser aplicável, a qual exige apenas a estacionaridade do semivariograma, sem nenhuma restrição quanto a existência ou nã de variância finita.

Existem três hipóteses de estacionaridade de uma função aleatória, $Z(s_i)$, sendo necessário a satisfação da primeira e de uma das duas restantes, antes de se fazer qualquer aplicação geoestatística.

- 3^a Hipótese: Intrínseca

  Uma função é intrínseca, quando além de satisfazer a primeira hipótese, primeiro momento estatístico, satisfaz, também o incremento $[Z(s_i) - Z(s_i + h)]$, o qual, tem variância finita e não depende de s_i, para qualquer vetor h.

  As expressões da covariância, da semivariância e da variância foram definidas pelas equações (8), (9), (10) e (11).

$$Var[Z(s_i) - Z(s_i + h)] = E[Z(s_i) - Z(s_i + h)]^2. \tag{13}$$

  Como se observa, a função $\gamma(h)$ representa o semivariograma que é denominado semivariograma clássico de Matheron, expresso usualmente por:

$$\gamma[h] = \frac{1}{2}E[Z(s_i) - Z(s_i + h)]^2.$$

  O semivariograma é estimado assumindo a hipótese intrínseca, mais utilizada na geoestatística por ser menos restritiva.

2.2 Semivariograma

O objetivo da análise geoestatística é identificar a estrututa de variabilidade espacial entre as amostras e a partir disso estabelecer uma medida dessa dependência por meio de ajuste de uma função teórica. De antemão, a verificação

da dependência espacial é feita por meio de uma função emprírica, chamada função semivariograma.

Diante da grande importância do semivariograma nas estimativas por krigagem (abordada na próxima seção), muitos estudos têm sido feitos para avaliar a relação da qualidade deste estimador, tamanho amostral, tipo de desenho amostral, dentre outros fatores envolvidos na análise espacial.

Wang & Qi (1998) estudaram a influência do tamanho amostral com duas análises de dados, sendo um dos conjuntos de dados satisfazendo a suposição estacionária intrínseca de segunda ordem e o outro violando essa suposiçã. Os autores verificaram que na Krigagem da média, na Krigagem simples e na Krigagem Ordinária, quando satisfeitas as suposições de estacionaridade, mesmo diminuindo de 2500 amostras para 625 amostras, não há perda de qualidade nos resultados dos ajustes dos semivariogramas e nem nos mapas das krigagens. Diante disso, fatores como a malha amostral regular e o atendimento às pressuposições de estacionaridade são decisivos na qualidade dos mapas de krigagem.

O estimador do semivariograma da equação (12) é dado pela equação:

$$\hat{\gamma}(h) = \frac{1}{n_h} \sum_{i=1}^{n_h} [Z(s_i) - Z(s_i + h]^2 \tag{14}$$

em que n_h é o número de pares de valores medidos $Z(s_i)$ e $Z(s_i + h)$, separados por uma distância h.

Oliver & Webster (2015) apontaram que partir do ajuste do semivariograma é possível saber se a dependência espacial dos dados pode ser considerada ou não, baseando-se no tamanho do raio de estimação, quando este for menor que a menor distância entre as amostras, torna-se inviável fazer krigagem.

De acordo com Chung et al. (1995) e Basseto et al. (2016), muito têm se discutido acerca dos modelos de semivariogramas com o objetivo de encontrar um modelo que estime com mais precisão, sem interferência de *outliers*. Seguindo neste intuito, muitos modelos foram sugeridos, tais como, o estimador New_1 e New_2 propostos por Li & Lake (1994), sob a justificativa que na análise variográfica é

comum delimitar uma distância máxima chamada *cutoff*,, com base nestes modelos, as estimativas para as distâncias superiores ao *cutoff*, tendem a ser menos precisas, devido à menor quantidade de pares de pontos utilizados.

Um segundo estimador foi proposto concomitantemente por Cressie & Hawkins (1980) que substituia o estimador da média pela mediana, sendo assim, um estimador mais robusto, chamado de estimador da mediana.

A modificação da proposta inicial do estimador robusto, que consistiu em substituir a média pela mediana no estimador robusto de Cressie e Hawkins, teve a justificativa de que a presença de uma observação discrepante pode ainda distorcer a estimativa da média. A mediana, por outro lado, é mais tolerante a erros, pois na presença de valores discrepantes, ela se mantém pouco, ou totalmente inalterada.

Outros estimadores foram propostos, tais como o "Pairwise" de Isaaks & Srivastava (1989) e o "Estimador das Diferenças" de Haslett (1997), ambos tiveram propósitos similares aos demais citados, o primeiro, teve o objetivo de proporcionar melhor visualização da continuidade espacial e o segundo, foi criado baseado na função de variância, com algumas poucas modificações que não apresentaram melhorias significativas, se comparados aos modelos já existentes, pois, dependendo do comportamento dos dados, a falta de robustez em ambos, foram de alguma forma mais ou menos presente.

Journel & Huijbregts (1978) destacaram que, sob a hipótese de estacionaridade de 2^a ordem, a covariância e a variância são equivalentes para caracterizar a correlação entre duas variáveis em relação a uma distância h, sendo a variância um caso particular da covariância quando $h = 0$,

$$Cov[Z(s), Z(s)] = E[Z^2(s)] - \mu^2 = Var[Z(s)],$$

porém, a explicação para preferir utilizar o semivariograma na geoestatística se deve a maior abrangência do método, pois em muitos processos aleatórios não existem covariâncias.

Sob a hipótese intrínseca, as pressuposições básicas do semivariograma

são:

- a esperança matemática existe e não depende do refencial s

$$E[Z(s)] = \mu, \forall s \tag{15}$$

em que μ é uma constante.

- a variância da diferença $[Z(s) - Z(s+h)]$ existe para toda distância h e não depende a posição s

$$Var([Z(s+h) - Z(s)]) = E([Z(s+h) - Z(s)]^2) = 2\gamma(h), \tag{16}$$

em que $2\gamma(h)$ é o variograma, enquanto a estacionaridade de 2^a ordem é muito restritiva, pois, $E[Z(s)] = \mu =$ constante

e

$$Cov(h) = E[Z(s+h), Z(s)] - \mu^2 = K(h). \tag{17}$$

Logo, a $Cov(h)$ precisa existir e ser constante em função da distância h, para qualquer s_i dentro da área A.

da SILVA et al. (1989) apresentaram algumas características necessárias e os parâmetros de semivariograma típico, dentre elas, destaca-se uma função positiva crescente e dependente de h,ou seja, quando h aumenta, $\gamma(h)$ também aumenta até uma valor máximo, chamado *patamar*, que se estabiliza, a partir desta medida de h, denominada *alcance*.

Em alguns casos, partindo do *efeito pepita* a semivariância cresce além do patamar(*sill*) até determinada distância e depois decresce e apresenta flutuações abaixo do valor do patamar para grandes distâncias, estes casos caracterizam periodicidade nos dados, cujo tratamento específico é chamado de *densidade espectral*, esta discussão pode ser encontrada em McBratney et al. (1981). Outro comportamento irregular ocorre quando a função semivariograma se mantém positiva, porém constante para qualquer distância h, indicando ausência total de dependência espacial, esse comportamento é denominado "pepita puro". Para esses casos, os dados

apresentam-se com distribuição espacial completamente aleatória e a única estatística aplicável é a estatística clássica.

Ao se mencionar as características do semivariograma, associa-se a elas os parâmetros que as definem, são eles (Andriotti, 2003):

- O efeito pepita (C_0), é a descontinuidade do semivariograma, que representa a variância nâo explicada, ou ao acaso, frequentemente causada por erros de medições ou variações das propriedades que não podem ser detectadas na escala de amostragem para $h = 0$ (Vieira, 2000).

- O alcance (a), representa a distância até onde as amostras apresentam comportamento estruturado com correlação espacial, ou seja, a partir dele o comportamento das amostras tornam-se independentes.

- O patamar(C), determina a variabilidade máxima entre os pares de valores em função de h quando h chega a um o valor máximo, chamado alcance (a), ou seja, quando a ordenada do alcance encontra o seu máximo e a partir daí a covariância entre os dados tornam-se nula.

- A Contribuição (C_1), representa a diferença entre o patamar e o efeito pepita.

Uma representação gráfica do semivaiograma clássico é dada pela Figura 4 com os parâmetros de alcance(a), Contribuição(C_1), Patamar(C) e Efeito Pepita (C_0).

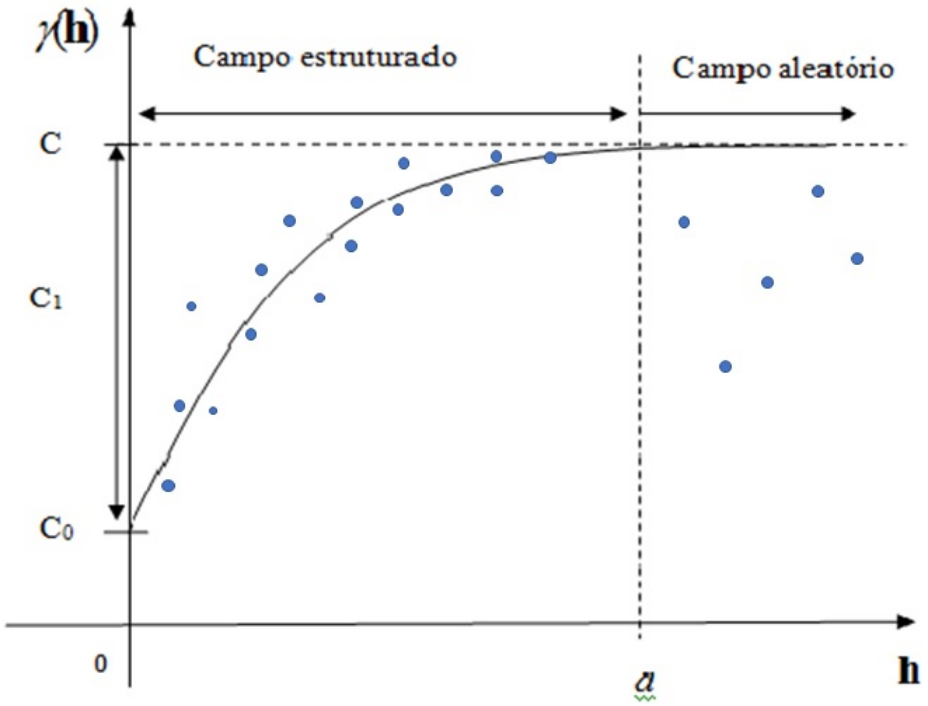

Figura 4: Representação do semivariograma com os parâmetros C_0, C_1, C e a.

O ajuste do semivariograma é feito em duas etapas, na primeira etapa é feito um ajuste direto dos dados ao modelo empírico, este gera diretamente o semivariograma empírico mas não permite fazer inferências da estrutura de variabilidade espacial. Um segundo ajuste é feito por meio de estimadores de semivariancias, utilizando algum dos métodos de ajuste de modelo teórico aos dados, tais como o "Mínimos Quadrados Ordinários"(OLS), "Mínimos Quadrados Ponderados"(WLS), "Máxima Verossimilhança"(ML) ou "Máxima Verossimilhança Restrita"(REML), que possibilita caracterizar a estrutura de dependência espacial ao se fazer um ajuste do modelo teórico á nuvem de pontos.

Após a adoção de um dos métodos de ajuste dos parâmetros é feita a validação cruzada com a qual é possível verificar a qualidade do ajuste através de uma regressão linear entre os valores reais e estimados os quais se estima os coeficientes β_0 e β_1.

A Interpretação dos parâmetros de qualidade, β_0 e β_1, é similar à

regressão linear simples, sendo β_0 o intercepto e β_1 o coeficiente angular. Quanto mais o coeficiente β_0 se aproximar de zero e o coeficiente β_1 se aproximar de 1 melhor será o modelo.

O ajuste de modelo de semivariograma por meio de modelos teóricos, torna-se necessários à aplicação da Krigagem.

Alguns autores, tais como, Webster & Oliver (1992) e Uribe-Opazo et al. (2012) analisaram a estrutura de dependência espacial no ajuste de semivariograma em diferentes dimensões de malha, para comparar os métodos de estimação de Mínimos Quadrados Ordinários (OLS), Mínimos Quadrados Ponderados (WLS), Máxima Verossimilhança (ML) e Máxima Verossimilhança Restrita (RML) e observaram que para dados normais, em todos os dados, ambos os métodos testados estimam os parâmetros do semivariograma de forma semelhante, porém outros fatores como espaçamento do *lag*, podem interferir no ajuste e estimação dos parâmetros de dependência espacial, tanto ou mais que o critério de ajuste e o tamanho amostral, ressalvada a condição de haver um número suficiente de pares de pontos no ajuste.

Os modelos teóricos de semivariogramas, se dividem em modelos com patamar e modelos sem patamar, sendo o segundo, usado apenas quando os dados apresentarem dispersão infinita. Os principais modelos com patamar são: Modelo Esférico, Modelo Exponencial e modelo Gaussiano (Andriotti, 2003).

O modelo Esférico é dado pela função:

$$\gamma(h) = C_0 + C_1 \left[\frac{3h}{2a} - \frac{1}{2}(\frac{h}{a})^2\right], 0 \leq h \leq a$$
$$\gamma(h) = C_0 + C_1, h > a \tag{18}$$

Alguns autores como Trangmar et al. (1987); Salviano (1996) destacaram que os modelos de semivariograma esféricos e exponencial são os mais adequados para descrever o comportamento de atributos de plantas e de solos.

No modelo Exponencial a semivariância aumenta mais lentamente partindo da origem em direção ao patamar, e não se pode dizer onde exatamente o modelo atinge o patamar, pois este é obtido assintoticamente (Almeida, 2013).

O modelo Exponencial é representado pela função (19):

$$\gamma(h) = C_0 + C_1 \left[1 - exp\left\{-3\frac{h}{a}\right\}\right], 0 \leq h \leq a$$
$$\gamma(h) = C_0 + C_1, h > a \tag{19}$$

Silva et al. (2011) sugerem que caso o efeito pepita seja muito pequeno e a estrutura de variabilidade cresça de maneira bastante suave, o variograma pode ser melhor ajustado pelo modelo Gaussiano. O modelo Gaussiano é altamente desejável por apresentar boas propriedades, como continuidade na variabilidade a medida que os pontos se afastam entre si. De modo similar ao modelo exponencial, o modelo Gaussiano atinge o patamar(C), assintoticamente, em 95% do alcance a, maiores detalhes sobre a demonstração deste percentual, podem ser obtidos em (Almeida, 2013).

A função que descreve o modelo Gaussiano é dada pela expressão (20).

$$\gamma(h) = \begin{cases} 0, & \text{se} \quad h = 0, \\ C_0 + C_1 \left[1 - exp\left\{-3\left(\frac{h}{a}\right)^2\right\}\right], & \text{se} \quad 0 \leq h \leq a, \\ C_0 + C_1, & \text{se} \quad h > a \end{cases}, \tag{20}$$

este modelo é, muitas vezes, usado para modelar fenômenos extremamente contínuos (Isaaks & Srivastava, 1989; Uribe-Opazo et al., 2012).

Outros modelos teóricos muito usados para ajuste de semivariograma são, o modelo Cúbico e o modelo Matérn.

O modelo Cúbico é descrito pela expressão (21):

$$\gamma(h) = \begin{cases} 0, & \text{se} \quad h = 0, \\ C_0 + C_1[7(\frac{h}{a})^2 - \frac{35}{4}(\frac{h}{a})^3 + \frac{7}{2}(\frac{h}{a})^5 + \frac{3}{4}(\frac{h}{a})^7], & \text{se} \quad 0 \leq h \leq a, \\ C_0 + C_1, & \text{se} \quad h > a \end{cases}, \tag{21}$$

e caracteriza-se por ser suave na origem (Goovaerts, 2000).

O modelo Matérn e sua função de correlação é dado pela expressão (22):

$$\gamma(h) = \begin{cases} 0, & \text{se} \quad h = 0, \\ C_0 + C_1[1 - \frac{2}{\Gamma(v)}(\frac{h\sqrt{v}}{a})^v \kappa v(\frac{2h\sqrt{v}}{a})], & \text{se} \quad h \geq 0; v \geq 0 \end{cases} \tag{22}$$

em que k é a função Bessel (Kreh, 2012), Γ e υ são, a função gama e o parâmetro de suavização, respectivamente.

O semivariograma Matérn é válido em R^d, d$\geq$ 1 e assume qualquer tipo de comportamento próximo à origem, e para sua aplicação, adota-se h^{2v} se υ(v) não for inteiro e $h^{2v}\log(h)$ quando υ(v) for inteiro.

Quando o modelo Matérn utiliza-se a constante $\kappa = 0,5$ é denominado modelo exponencial e quando assume $\kappa = 0,3$, denomina-se simplesmente, "Matérn".

O modelo Circular é representado pela expressão(23):

$$\gamma(h) = \begin{cases} 2\frac{C_0+C_1}{\pi} \times \frac{h}{a} \times \left[\sqrt{1-\left[\frac{h}{a}\right]^2} + arcsin(\frac{h}{a})^2\right], & \text{se} \quad 0 \leq h \leq a, \\ C_0 + C_1, \qquad \text{se} \quad h > 0 \end{cases} \tag{23}$$

O Modelo Circular é válido apenas nos planos unidimensionais e bidimensionais, para os planos tridimensionais são aplicados o modelo esférico.

2.2.1 Isotropia e anisotropia

A garantia de validade do semivariograma é assegurada pela hipótese intrínseca assumida no ajuste do semivariograma, porém, a propriedade de média estacionária requer assumir que os parâmetros do semivariograma sejam os mesmos em qualquer direção, propriedade denominada *isotropia*.

Isotropia, refere-se a qualidade segundo a qual uma característica de interesse (para a geoestatística, a característica é ser variável regionalizada) tem o mesmo valor ou intensidade, independente da direção que ocorre, ou seja, acontece de forma homogênea em diferentes direções. Em contrapartida, a *anisotropia* acontece quando a característica de interesse varia conforme se modifica a direção de ocorrência do fenômeno.

A Isotropia é a característica desejável para toda pesquisa que utiliza-se de dados espaciais para análise geoestatística, porém na maioria dos casos esta propriedade não é alcançada.

Durante o procedimento da análise estrutural existem situações em que obter um variograma comum para todas as direções (omnidirecional) parece tarefa impossível, ao passo que obter um variograma para cada direção impossibilita a geração do mapa temático. Quando as semivariancias dos valores observados sofrem forte influência da direção, esta interfere para que haja mais de um semivariograma para a mesma variável no mesmo experimento, fenômeno denominado *anisotropia*.

A presença da anisotropia em algum ângulo direcional acarreta distorção no mapa de Krigagem. Diante deste problema, muitos autores dedicaram a buscar metodologias capazes de resolver ou amenizar estas distorções, tais como Journel & Huijbregts (1978), Isaaks & Srivastava (1989), Vieira (2000), Olea (2006), dentre outros.

Existem várias formas de detectar a anisotropia, uma delas se dá pelo cálculo de semivariogramas experimentais direcionais (usualmente 0^0, 45^0, 90^0 e 135^0), em que é feita uma inspeção visual avaliando suas similaridades para as diferentes direções adotadas. Outra forma acontece por meio do esboço gráfico de uma elipse (conhecido também como diagrama da rosa), calculada através dos alcances obtidos em direções distintas.

A Figura 5 representa a ilustração da anisotropia.

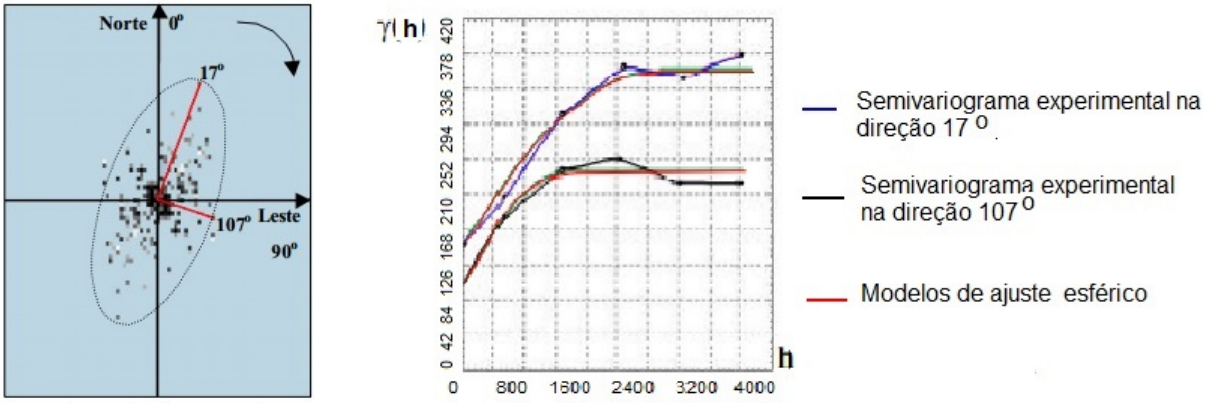

Figura 5: Representação do semivariograma com a presença de anisotropia.
Fonte:(Câmara et al., 2002)

Sobre o semivariograma é possível detectar rapidamente os eixos de anisotropia, isto é, as direções de maior e menor continuidade espacial da variável analisada.

Geralmente, ocorre que semivariogramas determinados ao longo de diferentes direções da área, podem indicar variações diferentes para a mesma variável, caso típico de anisotropia, que por sua vez, podem ser classificadas em, Anisotropia Geométrica, Anisotropia Zonal ou Anisotropia Combinada, maiores detalhes sobre o assunto, podem ser encontrados em Deutsch & Journel (1992) e outros.

.

2.3 Krigagem

O termo *krigagem* foi utilizado pela primeira vez em 1963 por G.P.M. Matheron para homenagear o engenheiro de minas, Daniel G.Krige e por volta de 1970 a técnica foi batizada com este nome. Esta técnica consiste na interpolação de valores de pontos para locais não amostrados o qual o modelo assegura não tenden-

ciosidade e variância mínima entre os valores amostrados e estimados. A partir deste período a técnica foi se consolidando, principalmente devido ao fato de ser capaz de tornar um campo de observações pontuais em um campo contínuo, cujos padrões espaciais são considerados e mantidos sob algumas hipóteses, necessárias para garantia de precisão nas estimativas.

Supondo que exista uma relação espacial entre os n valores amostrais, regularmente distribuídos ou não, o valor Z^* a ser interpolado para qualquer local s_0, é igual a

$$Z^*(s_0) = \sum_{i=1}^{n_h} \lambda_i * z_i(s_i) \tag{24}$$

em que Z^* é o valor interpolado, λ_i são os pesos atribuídos a cada ponto amostral por meio do Sistema de Krigagem e $z(s_i)$ é o valor amostral da variável Z no ponto s_i, $i = 1, 2, 3, \ldots, n$.

Yamamoto & Landim (2013) destacam que a diferença fundamental entre os diversos métodos estimadores existentes é a maneira como os Z_i são escolhidos e os respectivos pesos λ_i são calculados e aplicados durante o processo de estimativa.

Nos modelos determinísticos, utiliza-se critérios puramente geométricos como por exemplo, Distâncias Euclidianas ou Método do Inverso do Quadrado da Distância, os quais não fornecem medidas de incertezas. Contrários a estes métodos, nos modelos estocásticos, os valores são interpretados como provinientes de processos aleatórios, capazes de quantificar a incerteza associada ao estimador, a esta classe pertencem os modelos geoestatísticos chamados de "krigagem".

2.3.1 Estimador de Krigagem

O estimador de Krigagem estima valores Z^*, para qualquer local de uma região espacial, s_0, dentro do raio de estimação, onde não se tenha valores medidos, por meio de combinação linear dos valores amostrados. O estimador é

representado pela equação:

$$Z^*(s_0) = \sum_{i=1}^{n_h} \lambda_i z(s_i) \tag{25}$$

em que n_h é o o número de valores medidos a uma distância h e λ_i são os pesos associados aos valores Z nas posições s_i, envolvidos na estimativa.

As estimativas geoestatísticas podem ser feitas diretamente sobre os dados originais, por modelagem linear, ou sobre os dados transformados, por modelagem não linear. Os tipos de Krigagem linear são, Krigagem Ordinária, Krigagem da Média, Krigagem Simples e Co-Krigagem, enquanto que, os principais tipos de Krigagem não linear são, Krigagem Multigaussiana, Krigagem Lognormal e Krigagem Indicadora (Yamamoto & Landim, 2013; Vieira, 2000).

2.4 Sistema de equações de Krigagem Linear

A lógica envolvida na dedução do sistema relaciona-se com a aplicação das condições de estacionaridade de primeira ordem e estacionaridade intrínseca ao estimador, em que Vieira (2000) trata todo o processo matemático de manipulação das equações, para a dedução do sistema que origina os pesos da Krigagem. A partir daí, é calculada também, a variância de Krigagem.

Resumidamente, faz-se a esperança da diferença entre o estimador e o valor amostrado, inserindo a primeira condição de estacionaridade (25), que resultam em (26):

$$E(Z^*(s_0) - Z(s_0)) = E[\sum_{i=1}^{n_h} \lambda_i Z(s_i) - Z(s_0)] = 0. \tag{26}$$

Observa-se que, para um ponto amostrado numa posição qualquer, s_0, a esperan ca do valor estimado, coincide com o valor observado, e assim, a segunda hipótese, assumindo como necessária a primeira hipótese, resulta em (27):

$$Var[Z^*(s_0) - Z(s_0)] = E(Z^*(s_0) - Z(s_0))^2 = \text{mínima} \tag{27}$$

em que aplicando a linearidade do operador, obtém-se $\sum_{i=1}^{n_h} \lambda_i = 1$.

Este fato mostra que, para que a estimativa não tenha tendência, os pesos devem somar 1, para quaisquer que seja a distribuição de seus valores. Deste modo, desenvolvendo a equação (27), tem-se a variância de Krigagem como segue,

$$\sigma_k^2 Z^*(s_0) = E[Z^*(s_0) - Z(s_0)]^2 = E[Z^{*2}(s_0) + Z^2(s_0) - 2Z^*(s_0)Z(s_0)].$$

Fazendo as manipulações algébricas necessárias e simplificando-as, obtém-se

$$\sigma_k^2 Z^*(s_0) = \sum_{i=1}^{n_h}\sum_{j=1}^{n_h} \lambda_i \lambda_j Cov(s_i, s_j) + Cov(0) - 2\sum_{j=1}^{n_h} Cov(s_i, s_j). \tag{28}$$

Por meio da condição de restrição, dada pela soma dos pesos que se iguala a 1, a equação (28) pode ser minimizada pelas técnicas de Lagrange descritas em Almeida (2013), em que para satisfazer as condições expressas em (26) e (27) é preciso que as n derivadas parciais sejam iguais a zero.

Após organizadas as equações, tem-se o sistema de Krigagem:

$$\sum_{j=1}^{n_h} \gamma(s_i, s_j) + \delta = \gamma(s_i, s_0); i = 1 \ldots n_h, \tag{29}$$

com isto, pode-se observar que o sistema (29) contém $n + 1$ equações e $n + 1$ incógnitas, com $n = n_h$, em que, uma única solução, produz n pesos λ e um multiplicador de Lagrange, δ.

O sistema de Krigagem da equação (29), pode ser escrito em notação matricial como:

$$[C][\lambda] = [b],$$

em que $[C]$ é a matriz de covariância amostral e $[b]$ é o vetor de valores das covariancias relativas às distâncias dos pontos vizinhos ao ponto a ser estimado, e $[\lambda]$ é o vetor de pesos, ou, quando se usa o semivariograma,

$$[\gamma][\lambda] = [b], \tag{30}$$

em que $[\gamma]$ é a matriz de semivariancias.

A solução da equação (30), é dada por:

$$[\lambda] = [C]^{-1}[b].$$

A variância da estimativa, $\sigma_k^2 Z^*(s_0)$, fica

$$\sigma_k^2 Z^*(s_0) = \delta + \sum_{i=1}^{n_h} \lambda_i \gamma(s_i, s_j). \tag{31}$$

Na sequência, apresenta-se os diferentes tipos de krigagem linear.

2.4.1 Krigagem Simples ou Estacionária

Considerando um local s_0 e n valores obtidos em pontos vizinhos (com vizinhança definida pelo *alcance*, também conhecido como (*cuttoff*), do semivariograma ou por uma matriz de vizinhança de raio fixo, uma estimativa linear ponderada desse local pode ser escrita como (Journel & Journel, 1989)

$$Z_{ks}^*(s_0) = \mu_0 + \sum_{i=1}^{n} \lambda_i [z(s_i) - \mu_i]$$

em que $\mu_i = E[Z(s_i)]$ são as médias, as quais são assumidas como conhecidas, μ_0, a média assumida no ponto s_0 e $\lambda_i, i = 1, \ldots, n$, os pesos relacionados aos n dados.

Sob a condição de estacionaridade de segunda ordem, a média e a variância de todos os locais são constantes, ou seja, dependem apenas das distâncias euclidianas que separam as amostras (Harlan, 2013).

O estimador da Krigagem Simples é calculado como

$$Z_{ks}^*(s_0) = \mu + \sum_{i=1}^{n} \lambda_i [Z(s_i) - \mu_i]. \tag{32}$$

Os pesos ótimos da Krigagem Simples são definidos por meio de uma nova função aleatória, que é a diferença entre a função aleatória $Z(s_i)$ e sua média:

$$Y(s_i) = Z(s_i) - E[Z(s_i)] \tag{33}$$

em que $E[Y(s_i)] = 0$, obtendo-se por meio da equação (33), a estimativa dos resíduos. Desta forma, a covariância de $Z(s_i)$ é igual a covariância de $Y(s_i)$, ou seja:

$$Cov(s_i, s_j) = Cov_Y(s_i, s_j) = E[Y(s_i)Y(s_j)] \tag{34}$$

com s_i e s_j, posições distintas das amostras.

A variância do erro e sua forma em termos de resíduos, podem ser escritas como

$$\sigma^2(s_0) = Var[Z^*_{KS}(s_0) - Z(s_0)]$$

$$\sigma^2(s_0) = Var[\sum_{i=1}^{n} \lambda_i Y(s_i) - Y(s_0)].$$

Para encontrar o ponto de mínimo da função da variância do erro, calcula-se as derivadas parciais em relação aos pesos, e as igualam a zero, resultando assim, no sistema de equações normais:

$$\sum_{j=1}^{n} \lambda_j Cov(s_i, s_j) = Cov_Y(s_i, s_0); i = 1, \ldots, n. \tag{35}$$

O Sistem de krigagem da expressão 30, escrito por meio do sistema de equações matriciais, que permite estimar os ponderadores da Krigagem Simples, é expresso da seguinte forma:

$$\begin{bmatrix} C(s_1,s_1) & C(s_1,s_2) & \ldots & C(s_1,s_n) \\ C(s_1,s_1) & C(s_2,s_2) & \ldots & C(s_2,s_n) \\ \cdots & \cdots & \ldots & \cdots \\ C(s_n,s_1) & C(s_n,s_2) & \ldots & C(s_n,s_n) \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} C(s_0,s_1) \\ C(s_0,s_2) \\ \cdots \\ C(s_0,s_n) \end{bmatrix}$$

em que $C(s_i, s_j)$ representa os valores de covariância entre dois pontos nas posições s_i, e s_j, ou quando utilizadas as semivariancias entre dois pontos, substituindo C por γ.

2.4.2 Krigagem da Média

Como as condições impostas na Krigagem Simples de média conhecida e constante em todo o domínio amostral nem sempre acontecem, é preciso estimar a média em uma região caracterizada por uma vizinhança com n pontos mais próximos $Z(s_i), i = 1, \ldots, n$. A média pode ser estimada para essa vizinhança, como mostrado em (36)(Wackernagel et al., 1997):

$$\mu^* = \sum_{i=1}^{n} \lambda_i^{KM} Z(s_i) \tag{36}$$

a qual λ_i^{KM} são os pesos da krigagem da média.

Assumindo a existência da média em toda a região, a esperança, $E[Z(s_i)] = \mu$ e o erro sistemático $(\mu^* - \mu) = 0$. Desenvolvendo estas duas pressuposições, obtém-se a condição de não viés, $\sum_{i=1}^{n} \lambda_i^{KM} = 1$.

A variância do erro de estimativa é obtida em relação a função covariância:

$$Var[\mu^* - \mu] = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i^{KM} \lambda_j^{KM} C(s_i - s_j). \tag{37}$$

Uma forma de encontrar os ponderadores de forma otimizada, é feita por meio da minimização da variância de erro de estimativa (37), resultando na função objetivo no qual contém o Multiplicador de Lagrange δ, que aplicando derivadas parciais e igualando-as a zero, dá origem ao sistema de Krigagem da média , com $n + 1$ equações, como segue:

$$\sum_{j=1}^{n} \lambda_j^{KM} C(s_i - s_j) - \mu_{KM} = 0; i = 1, \ldots, n$$

$$\sum_{j=1}^{n} \lambda_j^{KM} = 1.$$

No modelo de Krigagem da média, a variância de estimativa da Krigagem é igual ao multiplicador de Lagrange. Então, ao invés de utilizar a média conhecida e constante, pode-se substituir a equação (32) pela média estimada,

$$Z_{ksm}^*(s_0) = \sum_{i=1}^{n} \lambda_i^{KS} Z(s_0) + \sum_{i=1}^{n} \lambda_i^{KM} \left[Z(s_0) - \sum_{i=1}^{n} \lambda_{i=1}^{KM} Z(s_0) \right]$$

em que organizada, resulta em

$$Z_{ksm}^*(s_0) = \sum_{j=1}^{n} \lambda_j^{KS} \left[\lambda_i^{KS} + \left(1 + \sum_{j=1}^{n} \lambda_j^{KS} \right) \lambda_{i=1}^{KM} \right] Z(s_0). \tag{38}$$

Na equação (38), a expressão entre colchetes é o peso da Krigagem Ordinária, e o termo entre parênteses é o peso da média. A Krigagem Ordinária

é a soma da Krigagem simples com a média calculada para cada local através da Krigagem da média (Wackernagel et al., 1997).

A diferença entre o estimador de Krigagem Ordinária e o de Krigagem Simples é que no segundo não é incluído o multiplicador de Lagrange, por considerar a variável aleatória estacionária de 2^a ordem e a média conhecida e constante em todo o domínio amostral. Deste modo, assumindo covariâncias e variâncias finitas, a matriz de covariância do sistema é uma matriz triangular que assume valores nulos na diagonal principal.

2.4.3 Krigagem Ordinária

O método de Krigagem Ordinária é o mais utilizado de todas as krigagens por ser o método convencional de Krigagem, ser simples e permitir obter as estimativas dos valores da variável em locais não amostrados, por meio de combinação linear dos valores da amostra em locais próximos.

Por meio do estimador da Krigagem Ordinária, definido na equação (25), assumindo-se válidas as hipóteses de estacionaridade tem-se a garantia que, os pesos, λ_i, são não viesados e a variância das estimativas é mínima:

$$\sigma_E^2 = Var[Z(s_0) - Z^*_{KO}(s_o)]. \tag{39}$$

Como não se conhece $Z(s_0)$, a solução é um modelo probabilístico, em que seus valores são considerados realizações de um processo estocástico da variável aleatória, $Z(s_i), i = 1, \ldots, n$.

Para se obter o sistema de krigagem, é calculada a esperança da variância do erro de estimativa (27) que resulta na equação (40):

$$\sigma_E^2 = C(0) - 2\sum_i \lambda_i C(s_0, s_i) - \sum_{ij} \lambda_i \lambda_j C(s_i, s_j), \tag{40}$$

em que σ_E^2 é a expressão do erro de estimativa em termos de covariância. O ponto de mínimo define os pesos ótimos utilizando a técnica de multiplicador de Lagrange,

em que as restrições aplicadas são não viés e variância mínima para a lagrangiana (Yamamoto, 2001).

Aplicando cada uma das derivadas parciais da lagrangiana e igualando a zero, e aplicando a derivada em relação ao multiplicador de Lagrange δ, chega-se ao sistema de equações de Krigagem ordinária, descrito na forma matricial pelas equações (29) e (30) nas quais é mantida a restrição $\sum_{j=1}^{n} \lambda_j = 1$, e expresso por meio de matriz de covariâncias (ou semivariancias), vetor de pesos e Multiplicador de Lagrange, como segue:

$$\begin{bmatrix} C(s,s_1) & C(s_1,s_2) & \dots & C(s_1,s_n) & 1 \\ C(s_1,s_1) & C(s_2,s_2) & \dots & C(s_2,s_n) & 1 \\ \cdots & \cdots & \cdots & \dots & 1 \\ C(s_n,s_1) & C(s_n,s_2) & \cdots & C(s_n,s_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} * \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \cdots \\ \lambda_n \\ -\delta \end{bmatrix} = \begin{bmatrix} C(s_0,s_1) \\ C(s_0,s_2) \\ \cdots \\ C(s_0,s_n) \\ 1 \end{bmatrix} \tag{41}$$

ou em termos de função variograma, substituindo-se, no sistema(41) o C por γ.

A variância da Krigagem Ordinária é definida por:

$$\sigma^2_{KO} = \sum_{i=1}^{n} \lambda_i \gamma(s_0, s_i) + \delta$$

em que γ é substituído por C, caso se utilize a função covariância por meio do correlograma.

2.4.4 Cokrigagem

O método geoestatístico de Cokrigagem é utilizado quando duas variáveis amostradas juntas nos mesmos pontos são correlacionadas e ambas apresentam dependência espacial. O ajuste do semivariograma é feito por meio de semivariograma cruzado, utilizando as duas variáveis.

O estimador de Cokrigagem representa uma combinação linear de duas variáveis, Z_1 e Z_2. Desta forma, supondo que se queira estimar valores, Z_2^*, para

qualquer local, s_0, para que a estimativa seja uma combinação linear de ambas, Z_1 e Z_2, o estimador de Cokrigagem é descrito por:

$$z_2^*(s_0) \quad = \quad \sum_{i=1}^{n_1} \lambda_{1i} z_1(s_{1i}) + \sum_{j=1}^{n_2} \lambda_{2j} z_2(s_{2j}) \tag{42}$$

em que n_1 e n_2 representam os tamanhos amostrais das variáveis, Z_1 e Z_2.

A equação (42) indica que a estimativa de Z_2, é uma combinação linear de ambas as variáveis, Z_1 e Z_2, com pesos, λ_1 e λ_2, distribuídos de acordo com a dependência espacial de cada uma das variáveis entre si, e da correlação cruzada entre elas (Vieira, 2000).

As mesmas condições da Krigagem Ordinária, são consideradas para as duas variáveis, requeridas para garantir a otimalidade da Cokrigagem, e sob elas são deduzidas as equações, como segue:

$$\sum_{j=1}^{n_1} \lambda_{1j} \gamma_{11}(s_{1j}, s_{1k}) + \sum_{j=1}^{n_2} \lambda_{2j} \gamma_{22}(s_{1j}, s_{1k}) - \delta_1 = \gamma_{12}(s_{1k}, s_0); i = 1, \ldots, n_1$$

$$\sum_{j=1}^{n_1} \lambda_{1j} \gamma_{12}(s_{1j}, s_{2k}) + \sum_{j=1}^{n_2} \lambda_{2j} \gamma_{22}(s_{2j}, s_{2i}) - \delta_2 = \gamma_{22}(s_{2j}, s_0); i = 1, \ldots, n_2$$

pois $\sum_{i=1}^{n_1} \lambda_{1i} = 0$ e $\sum_{j=1}^{n_2} \lambda_{2j} = 1$.

A variância de estimativa é:

$$\sigma^2_{K2} Z_2^*(s_0) = \delta_1 + \delta_2 + \sum_{i=1}^{n_1} \lambda_{1i} \gamma_{12}(s_{1i}, s_0) + \sum_{j=1}^{n_2} \lambda_{2j} \gamma_{22}(s_{2j}, s_0).$$

A Cokrigagem é indicada quando duas variáveis são espacialmente correlacionadas, e por algum motivo prático, uma delas não pode ser amostrada numa quantidade grande de pontos. Assim, utiliza-se a outra variável, de conhecimento amostral mais amplo, para contribuir nas estimativas dos valores desta variável.

A Cokrigagem é menos utilizada pelo fato de ser mais exigente com a covariável envolvida na estimação. Detalhes sobre o método podem ser encontrado em (Vieira, 2000).

2.5 Krigagem não linear

Dentre os métodos de Krigagem existentes, tem-se aqueles utilizados quando a distribuição é assimétrica positiva, para esses casos, há necessidade de transformação dos dados, para evitar a influência dos poucos valores elevados na estimativa dos pontos, ou quando se deseja fazer inferência sobre a probabilidade de ocorrência de um limite mínimo, ou máximo, em relação aos valores da variável. Nestes casos, existe a necessidade de fazer uma transformação binária. Situações desta natureza, em que são necessárias transformações nos dados para se fazer Krigagem, são chamadas de Krigagem não lineares.

Dentre as krigagens não lineares destacam-se, a Krigagem Gaussiana, a krigagem Logarítmica e a Krigagem Indicadora, também chamada Indicatriz ou Indicativa.

2.5.1 Krigagem Indicativa

O método de Krigagem Indicativa, consiste na transformação de dados, numéricos ou temáticos, em indicadores $0, 1$, que possibilita predizer mapas de probabilidades.

Ao se transformar os dados numéricos do atributo em função indicadora, tem-se a vantagem de transformar em um novo conjunto, não paramétrico, em que nenhum tipo de distribuição para a variável aleatória, é considerada a priori.

Ao se fazer a conversão dos dados, em um indicador binário, $\{0, 1\}$, possibilita a estimativa da função de distribuição da variável aleatória, que por vezes, permite a determinação de incertezas e a inferência de valores do atributo, em locais não amostrados.

A principal vantagem da Krigagem Indicativa (não linear), em relação aos métodos de krigagem linear, é conseguir modelar atributos com alta variabilidade, por ser não paramétrico, e não sofrer interferência de dados discrepantes (Felgueiras, 1999).

2.5.2 Estimador de Krigagem Indicativa

O procedimento de Krigagem Indicativa, requer uma transformação não linear, chamada de Codificação Indicativa, que transforma cada valor do conjunto amostral, $z(s_i)$, em valores por indicação.

A Codificação Indicativa, sobre um conjunto de dados amostrais numéricos, $z(s = s_0)$, para um valor de corte z_k, gera um conjunto amostral por indicação, $I(s = s_0, z_k)$, do tipo:

$$I(s = s_0, z_k) = \begin{cases} 1, & \text{se} \quad z \leq z_k, \\ 0, & \text{se} \quad z > z_k. \end{cases}$$

Uma desvantagem do étodo, é que a escolha do ponto de corte exige muito conhecimento sobre a variável de interesse, por não existir um critério pré estabelecido, e desta forma, a escolha é feita de acordo com o interesse do estudo.

Landim (2006) sugire a adoção de ponto de corte na mediana, ou em mais de um ponto de corte, nos quantis da distribuição.

O estimador da Krigagem Indicativa pode ser escrito como (Journel & Huijbregts, 1978):

$$KI^*(s_0; z_k) = \sum_{i=1}^{n} \lambda_i I(s_i; z_k)$$

em que s_i, representa os locais dos pontos, e z_k, é a indicadora do ponto de corte.

A incerteza associada à estimativa da Krigagem Indicativa, pode ser medida por meio da variância de interpolação (Yamamoto et al., 2012):

$$S_0^2(s_0; z_k) = \sum_{i=1}^{n} \lambda_i [I(s_i; z_k) - KI^*(s_0; z_k)]^2$$

em que s_0 é a posição do ponto estimado, e S_0^2 é a medida de incerteza deste ponto.

A estimativa da função de distribuição acumulada condicional é dada por:

$$F^*_{I(s_0;z_k)} = KI^*(s_0; z_k) \tag{43}$$

Yamamoto & Landim (2013) apontam que o método de KI tem a

limitação da necessidade da existência de uma distribuição mínima de $0's$(zeros) e $1's$(uns) no conjunto amostral, para o ajuste de um modelo de variografia.

Na transformação dos dados em 0 e 1 tem-se a vantagem de mudar a variável para indicadora, sendo a partir daí, submetidas às facilidades e redução dos cálculos, porque a partir desta transformação tornam-se válidas, todas as propriedades operatórias de função indicadora, descrita em (Mood et al., 1974).

A definição, os axiomas e propriedades de função de probabilidade e função indicadora descritas em Mood et al. (1974), permitem estabelecer relação com a Krigagem Indicativa, porque ao interpolar um valor para um ponto qualquer não amostrado, estima-se um valor dentro do intervalo $[0, 1]$, ou seja, estima-se um valor que obedece aos três axiomas de probabilidade de Kolmogorov, permitindo com isso uma aproximação discretizada da *fdac* (função distribuição acumulada condicional) de $Z(s)$, que de acordo com Felgueiras (1999), são probabilidades discretizadas, que podem ser usados diretamente, para se estimar valores característicos da distribuição, tais como: valor médio, variância, mediana, quantis e outros, e obter uma estimativa, $\mu_{z(s)}$, do valor esperado $E[Z(s)]$:

$$E[Z(s)] = \int_{-\infty}^{\infty} z f[s; z|(n)]dz.$$

A partir da função densidade de probabilidade condicional às n amostras, a função de distribuição condicional, $f(s, z|(n))$, estima a média da variável nos $K + 1$ intervalos, usando os K valores de corte, z_K, pela aproximação da função acumulada crescente *fdac*:

$$\mu_{Z(s)} = \int_{-\infty}^{\infty} z dF(s, z|(n)) \approx \sum_{K=1}^{k+1} z_K [F^*(s, z_K|(n)) - F^*(s, z_{K-1}|(n))]$$

em que os valores das classes $k = 1, 2, \ldots, K$, são valores estimados das *fdac's*, para cada valor do atributo, $F^*(s, z_k|(n)); k = 1, 2, \ldots, K$.

$$\begin{aligned} z_0 = z_{min}; \qquad z_k = z_{max}; \\ z_k = \frac{z_0 + z_k}{2}; \quad F(s, z_0|(n)) = 0; \quad F(s, z_k|(n)) = 1. \end{aligned} \tag{44}$$

O valor da mediana $q_{0,5}(s) = F^{-1}[s, 0, 5|(n)]$ é inferido aplicando–se a função ajuste da distribuição sobre os valores de corte, com probabilidades acumuladas vizinhas ao valor $0, 5$.

• Dados temáticos

Felgueiras (1999) mostrou que é possível realizar Krigagem Indicativa em dados temáticos e que a diferença entre Krigagem Indicativa em relação a dados numéricos é apenas a codificação. Para dados temáticos esta codificação pode ser expressa como:

$$I(s = s_0, z_k) = \begin{cases} 1, & \text{se} \quad z = z_k, \\ 0, & \text{se} \quad z \neq z_k, \end{cases}$$

no qual os valores de corte $z_k, K = 1, 2, ...k$, são valores das K classes que pertencem ao domínio da função aleatória $Z(s)$ definida na região A.

A esperança condicional da variável aleatória temática por indicação $I_E(s; Z_K)$ é definida como:

$$I_E(s = s_0, z_K) = 1 \cdot prob\{I(s, z_K) = 1|(n)\} + 0 \cdot prob\{I(s, z_K) = 0|(n)\}$$

$$I_E(s = s_0, z_K) = 1 \cdot prob\{I(s, z_K) = 1|(n)\} = F^*(s, z_K|(n)).$$

Para dados temáticos, tanto a transformação por indicação aplicada ao método de Krigagem Simples, quanto aplicada a Krigagem Ordinária, fornecem, para cada classe, z_K, uma estimativa que é também, a melhor estimativa mínima quadrática da esperança condicional da variável aleatória $\{s; Z_K|(n)\}$.

Com esta propriedade, pode-se calcular estimativas de valores da *fdpc* (função de distribuição de probabilidade condicional), de $Z(s)$ para todas as classes, Z_K, no domínio de $Z(s)$. O conjunto desses valores, é uma estimativa da *fdpc* da variável aleatória temática na área, A.

A estimativa de incerteza para atributos temáticos, é definida pela moda da distribuição, determinada por um único valor de probabilidade, refente a classe z_K, associada a posição s:

$Inc(s) = 1 - P_K(s)$, em que $P_K(s) = Prob\{Z(s) = z_k\}$.

Para atributos numéricos, é comum expressar as incertezas em função de intervalos de confiança, por exemplo, para um nível de confiança de $0,95$, supondo normalidade, tem-se:

$$Prob\{Z(s) \in \mu_z(s) \pm 2\sigma(s)\} = 0,95$$

em que $\sigma^2 = E\{(Z(s)) - E\{Z(s)\}^2\}^2$, é a variância global da distribuição.

Deutsch & Journel (1992) criaram um método para eliminar pesos negativos, em distribuições de dados resultantes de Krigagem, para isso, fizeram um método para resolver o problema, a seguida, Deutsch et al. (1998), propuseram um algoritmo mais simples e funcional, para contornar o problema, descrito da seguinte forma:

Obtidos os resultados da solução do sistema de Krigagem Indicativa, os valores são ordenados em ordem crescente ou decrescente, e verificado o valor negativo de maior módulo, ou seja, o menor valor da distribuição dos pesos c, este valor torna-se considerado uma constante c:

$$c = -min(W_i, i = 1, 2, \ldots, n).$$

Esta constante, c, é adicionada a todos os pesos, que em seguida, são normalizados para retornar a soma dos pesos igual a 1:

$$W_i^c = \frac{W_i + c}{\sum_{j=1}^{n}(W_j + c)}; i = 1, \ldots, n.$$

Depois de feita a correção dos valores, a constante c é eliminada, e os demais valores preservados.

2.6 Krigagem Multigaussiana

Baseada na curva teórica de distribuição normal, a transformada Gaussiana é realizada quando os dados apresentam assimetria positiva forte provocando a influência dos poucos valores altos na estimativa de pontos da vizinhança, caracterizada por valores mais baixos. A solução apresentada por Yamamoto & Landim

(2013) é a obtenção de um único semivariograma, como se faz em Krigagem Indicativa da mediana, porém, no caso de variáveis discretas, tal transformação é impossível por estas estarem decompostas em k tipos.

A transformação neste caso, é feita por meio de uma função matemática, atribuindo a cada valor z, um novo valor $y = f(z)$.

A transformação multigaussiana é indicada para dados discretos com grande dispersão, ou seja, quando a variância dos dados excede o valor da média, e eleva o coeficiente de variação para maior que 1.

Os valores da variável de interesse, são classificados em ordem crescente, e obtém -se as classes, o primeiro ponto pertence a primeira classe $(r(z_1)) = 1$, e o n-ésimo ponto, pertence a n-ésima classe $(r(z_n)) = n$.

As proporções das classes são obtidas, dividindo-se cada classe, pelo número total de observações somado uma unidade (Journel & Huijbregts, 1978).

Calculando a inversa da função Gaussiana desses quantis, obtém-se os escores da distribuição normal padrão, como mostrado na equação (45),

$$Y(z(s_i)) = G^{-1}\left(\frac{r(Z(s_i))}{n+1}\right) \tag{45}$$

em que $G^{-1}(.)$ é a função gaussiana inversa que fornece o escore da distribuição normal padrão, para o quantil $\frac{r(z(s_i))}{n+1}$ (maiores detalhes e aplicação podem ser encontrados em (Yamamoto & Landim, 2013).

2.7 Krigagem Lognormal

Este tipo de Krigagem é feito por meio de transformação logaritmica, muito indicada para casos em que o coeficiente de variação é superior a $1,254$, ou a assimetria é positiva, sendo utilizada com o objetivo de aproximar de uma distribuição normal em termos de média e desvio padrão.

A transformção é feita aplicando logaritmo natural aos dados originais, como mostrado na equação (46),

$$y(s_i) = ln(z(s_i)) \tag{46}$$

em que $z(s_i)$ são os valores da variável de interesse, nos locais s_i.

Yamamoto & Furuie (2010) propuseram a transformação

$$y(s_i) = ln\left(\frac{r(z(s_i))}{n+1}\right)$$

que não altera a forma da distribuição, mas provoca translação nos valores em relação a mediana, melhorando a simetria.

;

3 MODELO DE PARTIÇÃO PRODUTO-MPP

A preocupação em encontrar estruturas mais homogêneas em banco de dados sempre foi de interesse de pesquisadores em análises estatísticas. Tal preocupação se se deve ao fato de que, a maioria dos métodos estatísticos têm como pressupostos, a homogeneidade de variância e normalidade dos dados, que em pesquisas de campo, não são atendidos, comprometendo a qualidade dos resultados. Além destas exigências, em dados reais, especialmente em conjuntos de dados espaciais, a estacionaridade da média é um fator determinante na qualidade dos mapas, porque é a primeira condição necessária, para se aplicar geoestatística. Tal condição, tem relação com a simetria e comportamento dos dados no espaço.

Em face da preocupação em garantir a homogeneidade, simetria e estacionaridade da média, muitos métodos de deteção de agrupamentos (*clusters*)[1], tem sido amplamente desenvolvidos com suporte no espaço, no tempo ou no espaço-tempo.

Na literatura são encontrados métodos clássicos, determinísticos e bayesianos de detecção de *clusters*, dentre eles, os Métodos Eurísticos, Modelagem de Amostragem de Espécie (SSM)(Beaumont et al., 2008) e Métodos de Partição Produto (MPP)(Smith, 1975).

Smith (1975) definiu teoricamente o método MPP, por meio de partição definida por um ponto de mudança da média, por considerar que o estudo de ponto de mudança na média, pode ser tratado sob abordagem Bayesiana, sendo possível fazer inferências sobre o mesmo, por considerá-lo na sequência de v.a. para a qual

[1] *clusters* são aglomerações de amostras observadas com um padrão característico em um determinado espço, tempo ou espaço-tempo

houve mudança de distribuição.

3.1 Partição definida por ponto de mudança na média: sob o enfoque bayesiano

Smith (1975) definiu que uma sequência de v.a. $y_1, \ldots, y_n$ é dita ter um ponto de mudança para k, $1 \leq k \leq n$, se

$$Y_i \sim F_1(y|\theta_1) \quad \text{e} \quad Y_i \sim F_2(y|\theta_2) \tag{47}$$

em que $F_1(y|\theta_1) \neq F_2(y|\theta_2)$, para as quais deve-se considerar, F_1 e F_2, formas funcionais conhecidas, mas o ponto de mudança k, desconhecido, para todo $i = 1, \ldots, k$, em F_1 e para todo $i = k+1, \ldots, n$, em F_2.

Dada uma sequência de observações, $y_1, \ldots, y_n$, θ_1 e θ_2, para vetores de paâmetros conhecidos ou desconhecidos, em que no último caso, além de fazer inferência sobre k, é interessante fazer inferências sobre θ_1 e θ_2, por meio do cálculo da probabilidade a posteriori do ponto de mudança ter ocorrido para vários pontos possíveis, $1 \leq k < n$. A partir destas probabilidades, são calculadas estimativas bayesianas e testes de hipóteses, usando probabilidade a posteriori.

Para obter a distribuição a posteriori de ponto de mudança, assume-se que a distribuição dos dados admite densidades $p_1(y|\theta_1)$ e $p_2(y|\theta_2)$.

A distribuição conjunta de $y_1, \ldots, y_n$, condicional a θ_1, θ_2 e ao ponto de mudança k, ocorrido no intervalo $1 \leq k < n$, é dada por:

$$p(y_1, \ldots, y_n|k, \theta_1, \theta_2) = \quad p_1(y_1, \ldots, y_k|\theta_1)p_2(y_{k+1}, \ldots, y_n|\theta_2). \quad = \prod_{i=1}^{k} p_1(y_i|\theta_1) \prod_{i=k+1}^{n} p_2(y_i|\theta_2)$$

ou seja, a verossimilhaça conjunta condicional aos parâmetros θ_1, θ_2 e k, $p(y_1, \ldots, y_n|k, \theta_1, \theta_2)$, é igual ao produto das probabilidades condicionais p_1, relativa a θ_1 e a p_2 relativa a θ_2.

Assume-se ainda, uma distribuição a priori especificada, sobre o conjunto de possíveis pontos de mudança, expressa por $p_0(k), 1 \leq k \leq n$, tal que

$$p_0(1) + p_0(2) + \ldots + p_0(n) = 1$$

De acordo com Smith (1975), a análise a posteriori do problema, depende criticamente das pressuposições feitas sobre θ_1 e θ_2, ao qual depende do conhecimento ou não do comportamento destes parâmetros.

- Caso com θ_1 e θ_2 conhecidos

Sabe-se via Teorema de Bayes que dados θ_1 e θ_2, as probabilidades a posteriori de possíveis pontos de mudança, tendo observado $y_1, \ldots, y_n$, são dadas por:

$$p_n(k|\theta_1, \theta_2) \propto p(y_1, \ldots, y_n|k, \theta_1, \theta_2)p_0(k), \qquad 1 \leq k \leq n \tag{48}$$

o qual nota-se que a equação(48) pode ser escrita na forma

$$p_n(k|\theta_1, \theta_2)/p_0(k) \propto p_2(y_{k+1}, \ldots, y_n|\theta_2)/p_1(y_{k+1}, \ldots, y_n|\theta_1) \tag{49}$$

Para um dado $p_0(k)$, segue que $p_n(k|\theta_1, \theta_2)$ é grande, quando a razão de probabilidade de "mudança" sobre "nenhuma mudança" é grande, quando baseada nas observações finais $n - k$.

Na equação (50),

$$p_0(k) = \frac{1}{n} \quad , 1 \leq k \leq n, \tag{50}$$

a moda a posteriori é também a estimativa da máxima verossimilhança, dada pelo valor de k que maximiza o lado direito da expressão (49).

- Caso com θ_1 conhecido e θ_2 desconhecido

Para a análise, independentemente da atribuição da priori $p_0(k), 1 < k < n$, atribui-se uma densidade a priori, $p_0(\theta_2|\theta_1)$, sobre Θ_2 (espaço paramétrico de p_2), que representa o conjunto de possíveis valores de θ_2 que poderia depender de θ_1. Neste caso, dado θ_1, tem-se

$$p_n(k|\theta_1) \propto \quad p(y_1, \ldots, y_n|k, \theta_1)p_0(k) \tag{51}$$

em que

$$p(y_1, \ldots, y_n|k, \theta_1) = \int_{\Theta_2} p(y_1, \ldots, y_n|k, \theta_1, \theta_2) p_0(\theta_2|\theta_1) d\theta_2$$

e desta forma, tem-se:

$$p_n(k|\theta_1)/p_0(k) \propto \int_{\Theta_1} p_2(y_{k+1}, \ldots, y_n|\theta_2)\, p_0(\theta_2|\theta_1) d\theta_2 / p_1(y_{k+1}, \ldots, y_n|\theta_1) \quad (52)$$

O taxa de verossimilhança esperada, $p_n(k|\theta_1)$, sendo agora determinado em relação a priori para θ_2, baseado nas observações finais, $n - k$.

A densidade marginal a posteriori para θ_2, é dada por:

$$p_n(\theta_2|\theta_1) = \sum_k p_n(\theta_2|k, \theta_1)\, p_n(k|\theta_1)$$

em que

$$p_n(\theta_2|k, \theta_1) \propto p(y_1, \ldots, y_n|k, \theta_1, \theta_2)\, p_0(\theta_2|\theta_1)$$

- Caso com θ_1 e θ_2 desconhecidos

Assumindo uma função de densidade a priori, $p_0(\theta_1, \theta_2)$, no espaço paramétrico $\Theta_{1,2}$, o intervalo de valores possíveis de (θ_1, θ_2), obtém -se

$$p_n(k) \propto p(y_1, \ldots, y_n|k)\, p_0(k)$$

em que

$$p(y_1, \ldots, y_n|k) = \int_{\Theta_{1,2}} p(y_1, \ldots, y_n|k, \theta_1, \theta_2)\, p_0(\theta_1, \theta_2) d\theta_1 d\theta_2.$$

Inferência sobre θ_1 e θ_2 podem ser baseadas em
$p_n(\theta_1, \theta_2) = \sum_k p_n(\theta_1, \theta_2|k) p_n(k)$, em que

$$p_n(\theta_1, \theta_2|k) \propto p(y_1, \ldots, y_n|k, \theta_1, \theta_2)\, p_0(\theta_1, \theta_2).$$

Assumindo uma distribuição a priori uniforme, o conjunto de moda a posteriori que dá a estimativa de máxima verossimilhança é em $\hat{k}, \hat{\theta}_{1,\hat{k}}, \hat{\theta}_{2,\hat{k}}$, em que $\hat{k}$ maximiza
$p_1(y_1, \ldots, y_k|\hat{\theta}_{1,k}) p_2(y_{K+1}, \ldots, y_n|\hat{\theta}_{2,k})$, e $\hat{\theta}_{1,k}$ maximiza $p(y_1, \ldots, y_k|\theta_1)$.

3.2 Modelos Espaciais de Partição Produto-MPPs

De acordo com Quintana (2006) sob permutabilidade, MPP e SSM induzem ao mesmo tipo de estrutura de partição, ambos são discutidos no contexto de detecção de agrupamentos, em modelos de regressão linear normal e na estimativa de densidade univariada.

O MPP teve sua primeira aplicação em dados de saúde introduzida por Barry & Hartigan (1992) e Barry & Hartigan (1993), nestes modelos as unidades amostrais de determinado grupo são consideradas amostradas de uma distribuição comum e, a priori, grupos são formados de acordo com o produto das distribuições, para as quais, Quintana & Iglesias (2003) propuseram um algoritmo de seleção de uma única partição, para resolução de um problema específico, usando critério de decisão.

Müller et al. (2008) discutiram a partir do MPP, a formulação preditiva de modelos de probabilidade para partições, focando em modelos especificados através de uma sequência de probabilidades condicionais, para juntar grupos já existentes ou iniciar um novo grupo.

Métodos para detecção de agrupamentos (*clusters*), têm sido cada vez mais utilizados nas diversas áreas da ciência, especialmente no campo da saúde, a pesquisa científica tem unido esforços para encontrar relação entre covariáveis e doenças e entre doenças e espaço físico (enfoque espacial).

3.2.1 Definição

O MPP permite identificar os pontos onde há mudança nos parâmetros de média e variância da distribuição dos dados, o qual é descrito pelo produto das probabilidades dos conjuntos das partições dos dados, que está contido o conjunto das partes, que são grupos (*clusters*), que compõem a partição verdadeira, o conjunto da variável resposta, o conjunto de variáveis preditoras, os locais destes objetos (*IDs*) e a cardinalidade do número de agrupamentos.

Para definir o método de Partição Produto MPP espacialmente

considera-se uma coleção de objetos $S = (1, \ldots, n)$ representados pelo conjunto de índices, $I_n = (1, \ldots, n)$, associados as unidades experimentais, um conjunto $S^* = \{S_1^*, \ldots S_k^*\}$ denotando uma partição[2], a qual sendo conhecida, assume-se que há independência entre os agrupamentos, S_h^*, ou seja, os parâmetros de cada grupo podem assumir distribuições diferentes, dada a partição a qual pertence estes grupos. O conjunto de índices $I_n = \{1, 2, \ldots, n\}$, representa os elementos no agrupamento h, para $h = 1, \ldots, k$, maiores detalhes podem ser obtidos em (Barry & Hartigan, 1992, 1993; Quintana & Iglesias, 2003; Quintana, 2006; Müller et al., 2008; Park & Dunson, 2010).

Cada conjunto S_i em S^* consiste de $n_i \geq 1$ elementos de S, que resultam em $\sum_{i=1}^{k} n_i = n$, ou seja, juntando-se os n_i elementos de cada agrupamento somam o total de elementos do espaço amostral n.

Um conjunto S de locais de agrupamentos na partição S_i^*, é definido pelos índices dos n elementos no agrupamento h , os quais assume-se que $S_1, \ldots, S_k$, são os conjuntos de agrupamentos classificados em ordem crescente, ou seja, $min\{S : S \in S_1\} < min\{S : S \in S_2\} < \ldots < min\{S : S \in S_k\}$.

O número k de agrupamentos vai de 1 a n e o modelo que define a partição em clusters é dado, inicialmente, pelo produto das probabilidades das partes de S^*, denominadas *coesões* [3] a priori $c(S_h^*)$ dado por:

$$\pi(S^*) = c_o \prod_{h=1}^{k} c(S_h^*)$$

em que $c(S_h^*)$ é definida por Hartigan (1990) como uma medida especificada para cada subconjunto S_h^* de acordo com o o grau de similaridade(*coesão*) dos componentes no subgrupo, e c_0 é escolhido de modo que a soma das probabilidades sobre todas as possíveis partições resultem em 1.

[2]Uma partição é uma família de subconjuntos $S_1, \ldots, S_j$ com a propriedade de que cada objeto contido em S^*, encontra-se em exatamente um dos S_j da partição (Hartigan, 1990)

[3]Coesões são graus de similaridades que se acredita existirem a priori entre as amostras em um mesmo subconjunto (Barry & Hartigan, 1993).

Condicionalmente a S^*, tem-se que $y_h = \{y_i : y_i \in S_h^*\}$ representa os dados referentes aos elementos no grupo h.

A função de densidade de y, em que $y = \{y_1, \ldots, y_n\}$(o vetor de resposta), dada a partição S^*, é expressa por:

$$f(y|S^*) = \prod_{h=1}^{k} f_h(y_h)$$

em que

$$f_h(y_h) = \int \prod_{i \in S_h^*} f(y_i|\theta_h) dG_0(\theta_h) \tag{53}$$

na equação (53), $f(\cdot|\theta_h)$ é a função densidade de y_h no agrupamento h, condicionada por θ, $G_0(\theta_h)$ representa a mistura das características de todos os agrupamentos e θ_h é a característica de cada agrupamento específico, ou seja, $G_0(\theta_h)$ são as misturas definidas pelas funções prioris e representa a mistura de processo de cada grupo.

O MPP, cuja teoria está fundamentada em Smith (1975), também pode ser implementado via modelo hierárquico, do seguinte modo:

Considere $S = (S_1, \ldots, S_j)$ o vetor indicador do grupo, com $j \in \{1, \ldots, n\}$, em que j indica a ordenação dos grupos. Dado que se conhece a partição, os grupos são independentes entre si, ou seja, existe um conjunto de parâmetros $\theta^* = \{\theta_1, \ldots, \theta_j\}$, um para cada grupo S_j, e o modelo é assim descrito:

$$y_i|\theta^*, S \sim^{ind} f(\theta^*_{S_j}) \tag{54}$$

Os pesos dos grupos são determinados por meio de um conjunto de funções, uma para cada grupo S_j, expressas como segue,

$$\begin{aligned} S_j &\sim^{iid} \textstyle\sum_{l=1}^{j} \Pi \delta_{\theta_l^*}; \\ \theta_j^* &\sim^{iid} G_o \end{aligned} \tag{55}$$

em que δ_{θ^*} é a soma das probabilidades de todos os elementos em θ_j^*.

Uma forma de representar o MPP através de modelo hierárquico é fazendo uma transformação de variáveis de modo que, fazendo Φ_i representar o

conjunto de parâmetros $\{\theta^*, S\}$, a função (56) torna-se equivalente as expressões (54) e (55), e se resume na seguinte forma hierárquica:

$$\begin{aligned} y_i &\sim f(\Phi_i) \\ \theta_j^* &\sim G \end{aligned} \tag{56}$$

em que G é a função que reune as probabilidades de todos os grupos, ou seja,

$$G = \sum_{l=1}^{j} \Pi_k \delta\theta_j^*$$

e δ_θ^* é o processo de mistura.

Como se observa, o modelo se resume em

$$G \sim D(\alpha G_0).$$

Os parâmetros em um mesmo grupo são originados de uma mesma distribuição G_0, ou seja, G_0 é a mistura de processos de cada grupo e, D representa as distribuições desses grupos.

Ao assumir distribuições específicas para os parâmetros, os métodos são ditos paramétricos (Reich et al., 2007).

As funções preditivas descrevem como os indivíduos são sequencialmente designados para os grupos já formados ou para iniciar um novo grupo. Como a escolha dessas funções a priori é arbitrária, o problema reside em saber quais distribuições devem ser utilizadas. Em modelagem Bayesiana geral (não somente espacial), uma escolha sugerida por Reich et al. (2007) é a priori *stick-breaking*, que oferece uma forma de modelagem de distribuição dos parâmetros, como uma quantidade desconhecida a ser estimada pelos dados.

A priori *stick-breaking* para distribuições F desconhecidas, é dada pela mistura

$$F \equiv^d \sum_{i=1}^{m} p_i \delta(\theta_i) \tag{57}$$

em que o número de componentes m de mistura pode ser finita, p_i são as probabilidades de mistura e $\delta(\theta_i)$ é a distribuição de Dirac com massa no ponto para θ_i.

No modelo (57) as de m partes somam 1, então, $\sum_{i=1}^{m} p_i = 1$.
A primeira probabilidade de mistura é modelada como $p_i = v_1, v_1 \sim Beta(a, b)$.
As locações $\theta_i \sim^{iid} F_0$, sendo F_0 uma distribuição conhecida.

Nestes modelos hierárquicos, a partição S^* é induzida pelo agrupamento dos elementos nos grupos S_j (vetor indicador do agrupamento) e θ corresponde ao conjunto de valores únicos de $\Phi_{i=1}^{n}$, um para cada agrupamento, os quais Hartigan (1990) denominou de coesões.

3.3 Representação de MPP paramétrico

De acordo com Hartigan (1990) seja $y_1, y_2, ..., y_n$ uma sequência de dados observados, considerada uma amostra de dados espacialmente georreferenciados, pode ser representada como $y_{s_1}, y_{s_2}, \ldots, y_{s_n}$, em que $y_1, y_2, \ldots, y_n$ representa o conjunto de observações nos locais $s_1, s_2, \ldots, s_n$, com o conjunto de índices, $I = 1, \ldots, n$ e $\rho = i_0, i_1, \ldots, i_b$, o conjunto de índices no agrupamento.

Assumindo que um valor particular $i_0, i_1, \ldots, i_b$ de ρ $0 = i_0 < i_1 < i_2 < \ldots < i_b = n$ divide a malha amostral(*grid*) em $B = b$ blocos (grupos) contíguos, com pontos finais $[i_j, i_{j-1}, \ldots, b-1]$, um bloco no espaço pode ser definido por um intervalo $[i_j, i_{j+1}] = [i_j + 1, \ldots, i_{j+1}]$ com $i_j \in I = 1, \ldots, n$, o conjunto de índices dos pontos amostrais e $j \in [i_j, i_{j-1}, \ldots, b-1]$, o conjunto de índices dos elementos no bloco, ou seja, $i_j + 1$ é o local onde o j-ésimo ponto de mudança ocorreu.

Uma partição em I indica o conjunto amostral Y, com os elementos representados pelos agrupamentos, como segue.

$$Y_{i_j, i_{j+1}} = Y_{i_j} + 1, \ldots Y_{i_{j+1}}.$$

Denota-se por B o número de blocos ou segmentos contíguos, deste modo, a variável aleatória B está associada com ρ de forma que se $\rho = i_0, i_1, \ldots, i_b$, B assume o valor b (definição apresentada em Ferreira et al. (2014) adaptada neste trabalho para abordagem espacial do modelo).

Uma escolha adequada de funções a priori para os grupos, quando a

quantidade de pesos dos agrupamentos for m, $m \to \infty$, é utilizar o Processo *Dirichlet*. No entanto, quando utilizado pode ocorrer formação de grande número de grupos muito pequenos, inapropriados para ser utilizados em análises geoestatísticas.

Quando se tem a convicção baseada em uma análise preliminar, ou uma vaga ideia da quantidade de grupos m, torna-se possível assumir que existem m distribuições e o problema se reduz ao caso de Modelos de Partição Produto Paramétrico (MPPP).

Para o caso paramétrico, assume-se m distribuições distintas, ou m distribuições iguais com parâmetros distintos, conforme for o caso. Para cada parte da partição assumida existe uma distribuição, e para cada parâmetro destas distribuições, é assumida uma distribuição a priori.

Como se pode observar do exposto no parágrafo anterior, quanto maior o valor de m, maior será o grau de complexidade do modelo, por isso, o conhecimento a priori do comportamento dos dados por meio de análise exploratória, é fundamental para reduzir o número de distribuições a uma quantidade satisfatória e exequível computacionalmente.

3.3.1 Algumas Propostas de Modelagem de Agrupamentos

Uma proposta de modelagem espacial bayesiana é apresentada em Gamerman & Lopes (2006) utilizando áreas ou *pixels* da região de estudo. De acordo com os autores, em cada área ou *pixels*, uma variável de interesse Y_i é observada, com média $h(\theta_i)$, para alguma função h, como por exemplo, o número de casos de doença numa determinada cidade, a quantidade de carbono no ar medida em vários pontos de uma região, etc. Para ambos os casos, nota-se que a informação da variável, seja discreta ou contínua, univariada ou multivariada, está associada a influência local.

Uma proposta de modelagem espacial, levando em consideração as influências locais, com o objetivo de reduzir o ruido em informações espaciais foi sugerida por Besag et al. (1991) ,o qual utilizou-se especificação da priori, com as informações dos pontos e suas respectivas localizações, as quais são descritas na forma

de Pares de Diferença(PD), como segue,

$$P(\theta) \propto \exp\{-\sum W_{ij} h(\theta_i - \theta_j)\}.$$

Os autores apresentaram algumas possibilidades e interpretações para os pesos W_{ij} e para a função h. Em particular, consideraram $h(y) = \frac{y^2}{2W}$, levando a uma distribuição diferente, denotada por $\theta \sim PD(w, \omega)$, em que uma constante de proporcionalidade, α_i, inclui o termo $W^{-d/2}$, que é uma matriz de distância elevada à metade do inverso da quantidade de vizinhos. A distribuição $\theta_i \sim \{\theta_j, W\}$, ou seja, θ_i é função de θ_j, na vizinhança de W, dada pela distribuição normal $N(\alpha_i, R)$, em que

$$\alpha_i = \frac{1}{n_i} \sum_{j-1}^{d} W_{ij}\theta_j$$

$$R_i = \frac{1}{n_i} W$$

$$n_i = \sum_{j-1}^{d} W_{ij}; i = 1, 2, \ldots d.$$

As formas PD são referidas na literatura como modelos autoregressivos condicionais(CAR), mas Gamerman & Lopes (2006) ressaltam que tais nomenclaturas são usadas indistintamente. Uma típica escolha de pesos é dada por $W_{ij} = I(S_j \in N_i)$,uma função indicadora de vizinhança, em que N_i define a vizinhança de S_i, $i = 1, \ldots, d$.

Os autores consideraram para definir o raio de cobertura da vizinhança, uma matriz $W_{ij} \sim {d_{ij}}^{-\tau}$, em que d_{ij} é alguma distância entre *pixel*, i e j, às posições dos pontos e τ é uma quantidade que mede a força da dependência espacial. Esta distribuição também pode ser identificada com os Campos Aleatórios Gaussianos, (*Gaussian Markov Random Fields* - GMRF).

Gamerman & Lopes (2006) afirmam que um processo de variação espacial θ' que toma valores em uma região S, segue um *Gaussian Random Field*, se $\theta = \theta(s_1), \ldots, \theta(s_d)$ possuir uma distribuição normal d-variável, para qualquer d e para todo conjunto de posições $s_1, s_2, \ldots, s_d \in S$. Esta priori é baseada na noção de vizinhança definida no espaço, e conduz ao modelo:

$$Y|\theta, \sigma^2 \sim N(\theta, \sigma^2 Id)$$

$$Y|W \sim CAR(w, \omega)$$

$$\omega \sim F_w.$$

Os autores ressaltam que um modelo conveniente, aplicado quando o risco de casos de doença em dada região precisa ser avaliado, é descrito pela expressão

$$Y_i \sim P_{0i}(e^{\theta i}), i = 1, 2, \ldots d.$$

Neste modelo, se nenhuma covariável estiver presente e os pesos forem dados pelos indicadores de vizinhos, obtém-se a distribuição a posteriori para θ e W_d , representada por:

$$\pi(\theta, W_d) \propto \prod_{i=1}^{d} \exp\{(\theta_i y_i - e^{\theta_i})\} W^{-d/2} \exp\{\frac{-1}{2W} \sum_{i<j} (\theta_i - \theta_j)^2\} p(W)$$

$$\Pi_i(\theta_i, W_d) \propto \exp\{\theta_i Y_i - e^{\theta_i}\} W^{-d/2} exp \frac{-1}{2W} \sum_{i<j} (\theta_i - \theta_j)^2 P(W)$$

em que $P(W)$ é a distribuição a priori de W.

A distribuição condicional completa de θ_i é:

$$\Pi_i(\Theta_i) \propto \exp\{\theta_i y_i - \exp \theta_i\} - \frac{n_i}{2W} (\theta_i - \overline{\theta}_i)^2,$$

em que n_i é o número de elementos na vizinhança $W_{i,j} = 1$, que define os valores de θ_j, e

$$\overline{\theta}_i = \frac{1}{n_i} \sum_{\{j, W_{i,j}=1\}} \theta_j.$$

Gamerman & Lopes (2006) apontaram que além desta opção de modelagem, a distância baseada no Processo Gaussiano GRF, sob a pressuposição de homocedasticidade (variância comum) e isotropia (dependência apenas da distância entre amostras) são frequentemente assumidas. Sob essas pressuposições, $\theta \sim N(\mu I_d, \tau^2 R\lambda)$ em que $R\lambda = (p_{ij})$, com $p_{ij} = \rho\lambda(|s_i - s_j|)$, para a mesma função de correlação definida por $\rho\lambda$, possivelmente dependendo do parâmetro λ, tipicamente um escalar ou de baixa dimensão.

Page & Quintana (2016) sugeriram quatro modelos de funções a priori, dentre eles, um modelo básico de MPP (referido como $sPPM$) que investiga a probabilidade de associação de agrupamentos, e mostraram que é possível combinar MPP espaciais com probabilidades, para produzir estruturas marginais espaciais com propriedades atraentes, (como por exemplo, não estacionaridade) que equilibram estrutura local e global. Para isso, propuseram algumas funções de coesão, que contrapõem ao modelo proposto por Yao (1984a) quanto a independência, capazes de garantir a dependência espacial. A partir destas sugestões, os autores apontaram a possibilidade de incorporar informações espaciais, através da função de coesão que é regida pela localização, descrita como segue,

$$P_r(\rho) \propto \prod_{h=1}^{k} C(S_h, S*_h)$$

em que $C(S_h, S*_h)$ representa a coesão de cada parte(grupo) na parti cão. Tal função coesão, é semelhante a abordagem de Park & Dunson (2010) que estenderam MPP para incorporar covariáveis, poém, neste modelo admitiram apenas partições conectadas espacialmente,como mostrado em (58):

$$C_{[i_j, i_{j+1}]} = \begin{cases} M\,\Gamma(|S|), \text{ se } \mathbf{S} \text{ é espacialmente conectado} \\ 0, \text{ outro caso} \end{cases} \tag{58}$$

em que, $M\Gamma(|S|)$ é usado para favorecer um pequeno número de grandes grupos, com o número de aglomerado regulado por M.

Para implementar a função de forma que atribua pequena probabilidade aos agrupamentos não conectados espacialmente, sugeriram quatro modelos

candidatos. Para o primeiro modelo, utilizaram idéias de tecelagem encontradas em Holmes et al. (2005), na medida em que as distâncias para um centróide de aglomerados, são usadas para penalizar partições com agrupamentos espacialmente dispersos.

Seja $\overline{S}_h$ o centróide do agrupamento S_h cujas coordenadas são calculadas usando:

$$\overline{S}_{h_k} = \frac{1}{n_h} \sum_{i \in s_h} S_{ik}; k = 1, 2$$

em que n_h =$|S_h|$ e $D_h = \sum_{i \in S_h} d(S_i, \overline{S}_h)$ e $|S_{i,k}|$ =soma das distâncias entre os pontos s_i e s_h e $D_h = \|.\|$.

No intuito de evitar pequenos agrupamentos locais, e pequenas áreas impostas pela função decrescente de D_h, que são obtidas por meio da expressão (58), os autores propuseram usar $\Gamma(D_h)$, para além de D_h, que requer o uso de $\Gamma(D_h)$ para garantir que $|S_h|$ sejam ponderados de forma semelhante. Na expressão (58), D_h é contínuo em $\Re^+$ e $\Gamma(.)$ não é monótona, aumentando de 0 a 1. Ao considerar $\Gamma(D_h)I[D_h \geq 1] + D_h I[D_h < 1]$ e um parâmetro α, para controle sobre a penalização das distâncias, obtiveram a expressão:

$$C_1(S_h, S_h^*) = \left\{ \frac{M\,\Gamma(|S_h|}{\Gamma(\alpha D_h)I[D_h \geq 1] + (D_h I[D_h < 1])} \text{se} |S_h| > 1 M, \text{se} |S_h| = 1, \right.$$

em que $C_1(S_h, S_h^*) = M$ para $|S_h| = 1$, para evitar problemas associados a $D_h = 0$, uma vez que todos $S_1, \ldots, S_n$ são distintos, e desta forma, $D_h = 0 \Leftrightarrow |S_h| = 1$. Assim, $\Gamma|S_i| = 1$ e $M\,\Gamma|S_h| = 1$.

Analisando os ajustes que compõem a função de coesão priori C_1, verifica-se as seguintes características:

- $\Gamma(\alpha(D_h))$ dificulta formar grupos com uma unidade;
- Dificulta formar grupos com região geográficas muito grandes;
- Facilita formar grupos de áreas pequenas (mais concentradas).

A segunda opção de função de coesão, proposta por Page & Quintana (2016), baseou-se num limite rígido de agrupamentos, em relação a algum α, ($\alpha > 0$) especificado, da forma:

$$C_2(S_h, S_h^*) = M\,\Gamma|S_h|\,\Pi_{i,j\in S_h} I[\|S_i - S_j\| \leq \alpha.]$$

Neste modelo de coesão C_2, pode se verificar as seguintes características:

- Limita o grupo a um raio máximo.
- Dificulta a permissão para formar grupos com uma unidade;
- Permite a formação de grupos com áreas de quaisquer tamanhos até o limite máximo;
- Não controla a formação de grupos de áreas pequenas (mais concentradas).

A terceira opção de função de coesão, é mencionada por Park & Dunson (2010) ao tratar covariáveis dependentes em um cenário espacial, por meio de modelo MPP espacial, em que se especifica, essencialmente, as covariáveis s_i, através da coesão, como segue,

$$C_3(S_h, S_h^*) = M\,\Gamma|S_h| \int \prod_{i\in S_h} q(s_i|\epsilon_h)q(\epsilon_h)d\epsilon_h,$$

ao qual, Müller et al. (2011) chamaram $q(s_i|\epsilon_h)q(\epsilon_h)$ de um modelo auxiliar e, apesar de que em teoria, $q(.|.)$ e $q(.)$ podem representar alguma combinação de funções, Page & Quintana (2016) a trataram como um par de densidades conjugadas, para os quais, um modelo conjugado Gaussiano ou modelo Inverso Gaussiano de Wishart pode ser apropriado. Neste caso, $\epsilon_h = (m, V)$ pode denotar a média e covariância, $q(s|\epsilon) = N(s|m, V)$, uma densidade Gaussiana bivariada e $q(\epsilon) = NIW(m, v|\mu_0, k_0, \nu_0, \wedge_0)$, uma densidade Normal Inversa de Wishart. Neste caso, C_3 pode ser referido como um modelo de coesão auxiliar.

Uma caracteriza que diferencia o modelo C_3 dos demais, é devido a cada ponto representar uma variável aleatória, o qual associa-se um agrupamento.

Como todo grupo, há independência entre eles, ou seja, ϵ_h é uma variável aleatória latente, que pode ser a distância máxima em cada grupo, ou de maneira independente, similar ao modelo C_1. Algumas propriedades importantes que marcam este modelo são assim expressas:

- Permite Juntar grupos grandes.
- A metodologia de agrupamento é feita por meio de uma distribuição e esta não depende do grupo em si, diferente de C_2 que é fixo;
- Cada grupo tem a sua variável aleatória, que representa o agrupamento.

O útimo modelo, proposto por Page & Quintana (2016) tem a mesma forma de C_3, porém, em vez de empregar um modelo conjugado preditivo a priori, um modelo preditivo a posteriori é utilizado. A formulação de C_4 é descrita pela seguinte expressão:

$$C_4(S_h, S_h^*) = M\,\Gamma|S_h| \int \prod_{i \in S_h} q(s_i|\epsilon_h) q(\epsilon_h|s_{h^*}) d\epsilon_h,$$

uma vez que o preditor a posteriori é, tipicamente mais alto que o preditor a priori, C_4, e com isso, coloca mais peso nas partições que são locais. Este modelo pode ser usado para dados de área, ou pontos georreferenciados, em que para o último, os autores sugerem o uso de um modelo conjugado $N_2(s_i|m_h, v_h)NIW(m_h, V_h|S_h^*)$.

Uma propriedade importante que diferencia o modelo C_4 dos demais, é que ele impõe características do grupo na variável aleatória, como por exemplo, a inclusão de centróide de vizinhança.

3.3.2 Estrutura espacial via verossimilhança e priori

De acordo com Page & Quintana (2016), uma estratégia de modelagem completamente válida, é assumir observações independentes condicionalmente ao agrupamento ρ. Neste caso, toda dependência espacial terá origem no agrupamento espacial produzido pelo modelo de partição produto espacial, e alternativamente, poderá ser incluído no modelo de verossimilhança, uma estrutura global ou

de agrupamento espacial específica, para produzir maior riqueza de estrutura espacial. Para este caso, os autores consideraram a correlação entre duas observações como a distância entre elas de forma crescente até ∞, ou decrescente até a distância zero, calculadas sob modelos de pouca probabilidade, para cada uma das coesões. Desconsiderando a dependência espacial na verossimilhança, o modelo básico é dado por:

$$f(y|\pi) = \prod_{i=1}^{k_n} f_h(y_h^*)$$

$$Pr(\pi) \propto \prod_{h=1}^{k_n} C(S_h, s_h^*)$$

em que $\mathbf{y} = (y(s_1), \ldots, y(s_n))$ e $f_h(y_h^*) = \int \prod_{i \in s_h} f(y(s_i)|\theta) dG_0(\theta)$ com $f(.|\theta)$ denotando a verossimilhança, e G_0 a função a priori em θ. Este modelo pode ser escrito hierarquicamente, usando rótulos $c_1, \ldots, c_n$ na seguinte forma:

$$\begin{aligned} y(s_i)|\theta, c_i &\sim^{ind} f(\theta_{c_i}^*), \text{para } i = 1, \ldots, n \\ \theta_\ell^* &\sim^{iid} G_0, \text{para } \ell = 1, ..., k_n \end{aligned} \tag{59}$$

com $\theta_1^*, \ldots, \theta_{k_n}^*$ denotando os parâmetros específicos de agrupamentos, tal que $\theta_\ell = \theta_{c_i}^*$, e k_n é o número de grupos,sobre os quais observa-se que todos os elementos pertencentes ao mesmo grupo, pertencem a mesma distribuição.

No cenário espacial, $c_1, \ldots, c_n$, são variáveis latentes dependentes e multinomial, cujas componentes de probabilidades são oriundas de um modelo espacial. A estrutura espacial poderá ser incluída na verossimilhança, hierarquicamente, como efeito espacial aleatório, e desta forma, o modelo precisa ser ajustado de acordo com tais efeitos, sendo o efeito espacial aleatório, definido por um agrupamento específico ou global. Quando estão disponíveis covariáveis, a relaç ao entre elas com a resposta, também pode ser modelada como agrupamento específico ou não.

Para definir um modelo de verossimilhança com as covariáveis disponíveis, considera-se $x(s_i) = x_i$ e $y(s_i) = y_i$, que denota vetores de covariável e variável resposta na locação s_i.

Fazendo $\beta_1^*, \ldots, \beta_{k_n}^*$ denotar os parâmetros de agrupamento específico, tal que, $\beta_h^* \sim^{iid} N(\mu, T)$ e assumindo que π e $\{\beta_h^*\}_{h=1}^{k_n}$, são mutuamente independentes, o modelo, sob verossimilhança, é dado por:

$$y(s_i)|x_i, c_i, \beta^*, \phi^2 \sim N(x_i^{'}\beta_{c_i}^*, \phi^2),$$

em que, x_i poderá representar algum comportamento espacial, ou alguma forma de correlação com o espaço. Nesta modelagem, os grupos são definidos por $c_1, c_2, c_3, \ldots, c_n$, cuja formulção é expressa da seguinte forma:

$$\begin{aligned} y(s_i)|\theta, c_i &\sim^{ind} f(\theta_{c_i}^*) \\ \theta_l^* &\sim G(\theta_l) \end{aligned} \tag{60}$$

para $l = 1, \ldots, k_n$. Neste caso, a expressão (60) completa a hierarquia.

4 MATERIAL E MÉTODOS

O estudo se aplica a dados cuja variável resposta é paramétrica e contínua, por exemplo, taxas de incidências e proporções e suas relações com o espaço, quantidades de substâncias no solo ou no ar,altimetrias de terrenos, profundidades (batimetrias) de rios e lagos, entre outros casos com características de continuidade. A especificação do espaço paramétrico da variável resposta com suporte em R, foi adotada nesta tese, devido às características dos conjuntos de dados utilizados para testar o método exigirem. A metodologia proposta nesta tese, nomeadamente MPPs, se justifica por ser um modelo destinado a aplicação a dados contínuos em que a Krigagem Ordinária ou outros tipos de Krigagem são comumente usadas.

A escolha dos dados para aplicaçõa da metodologia proposta, MPPs, o qual foi utilizado um banco de dados de altimetria de terreno e um segundo banco de dados de profundidade de um lago, foi motivada pela caracterização da presença do problema de dupla e tripla estacionaridade, respectivamente, que a metodologia MPPs se propõe corrigir. A metodologia busca responder a seguinte questão: É posível obter amostras mais homogêneas ao se estabelecer grupos, de modo a garantir a estacionaridade da média espacialmente, e melhorar a qualidade e acurácia dos mapas de Krigagem?

Em algumas situações, a garantia de estacionaridade da média, é comprometida devido a presença de mais de uma estrutura de semivariância, e este fator impede a obtenção de um bom ajuste do semivariograma, e como consequência, um modelo ruim de semivariograma provoca perda de acurácia no mapa de krigagem.

A proposta desta tese é utilizar uma metodologia que possa fazer um

ou mais particionamentos da malha amostral, baseando-se na probabilidade de obter a melhor partição, ou seja, a mais provável de obter dados mais homogêneos nos grupos formados, baseando-se em distribuições de probabilidade para localizar os pontos de mudança na média, espacialmente. Esta proposta, tem como alternativa mais viável, a construção de um método Bayesiano de MPP (Método de Partição Produto), denominado nesta tese, MPPs (Método de Partição Produto Espacial) por incorporar características espaciais ao modelo.

Para aplicar a metodologia proposta foram utilizados dois conjuntos de dados, um conjunto de dados (Dados1) com 479 amostras georreferenciadas, extraídas em uma fazenda no município de Canaã, Zona da Mata Mineira, no ano de 2011, cuja variável de interesse é a altitude do terreno, medida ao nível do mar.

O segundo banco de dados refere-se a dados batimétricos (Dados2) com 1411 amostras georreferenciadas, extraídas da lago São Bartolomeu, que é a lagoa na proximidade da Universidade Federal de Viçosa, (UFV), os quais obtiveram-se amostras da profundidade em cada ponto, medidas em valores negativos a partir do ponto zero (z=0), determinado pelo limite da superfície da água. Estas amostras foram coletadas e organizadas em arquivos com extensões *.txt*, *.csv* e *.dat* pelo Grupo de Estudos e Pesquisas em Levantamentos Hidrográficos da UFV (GEPLH-UFV), cedidas para aplicação deste trabalho.

Na Tabela 1 são apresentadas algumas medidas descritivas e espaciais, referentes aos conjuntos de dados completos que justificaram a necessidade de aplicar um ponto de corte no conjunto de dados altimétricos (Dados1), e dois pontos de corte no conjunto de dados batimétricos (Dados2).

Tabela 1: Medidas descritivas referentes aos dados de altmetria(Dados1) e de batimetria(Dados2)

	Dados altimétricos coordenadas			Dados batimétricos coordenadas		
	X	Y	Distância	X	Y	Distância
Mínima	739475,1	7702222	86,81	72164,9	7702159	1,0017
Máxima	755361,0	7723030	21.161,08	721756,8	7702306	154,55

O método MPPs oferece uma forma prática para solucionar o problema da estacionaridade de 1^a ordem, o qual frequentemente, pesquisadores tem tentado resolver para conseguir ajustar semivariogramas. Problemas de estacionaridade permeiam as discussões em busca do que fazer quando uma amostra de tamanho considerável, apresenta-se na variável de interesse, natureza e características de falta de dependência espacial devido à amostra apresentar algumas características que mascaram a dependência espacial, tornando-a insatisfatória para um bom ajuste do semivariograma.

Dentre as suspeitas levantadas sobre o problema da falta de estacionaridade da média é que os dados tenham mais de uma média estacionárias predominantes, que em muitos casos, se devem a própria natureza dos dados, como por exemplo, na saúde, a existência de aglomerados de focos de doenças, a presença de focos de incêndios ou derramamento de substâncias no solo que contamina apenas uma parte da área, as características topográficas de um terreno,tais como, desmatamentos, erosão, e outros fatores que não se pode controlar, dependendo do tipo de variável envolvida.

A proposta desta tese é utilizar uma metodologia que possa fazer um particionamento do *grid* em duas ou mais partes, procurando encontrar a melhor partição, ou seja, a mais provável de obter dados mais homogêneos nos grupos do conjunto amostral, baseando em distribuições de probabilidade para localizar os pontos de mudança na média espacialmente. Uma solução para se encontrar os locais

das mudanças na média no espaço é utilizando o modelo Bayesiano de MPP(Método de Partição Produto), aqui tratado como MPPs(Método de Partição Produto Espacial), incorporando características espaciais ao modelo.

Para corrigir o problema da falta de estacionaridade de 1^a ordem, foi desenvolvido um método espacial (MPPs) o qual foi aplicado a dados altimétricos e batimétricos, assumindo-se um número conhecido de pontos de mudança na estrutura de semivariância, com o propósito de encontrar os locais mais prováveis que mudanças estruturais na média tenham ocorrido.

A abordagem da metodologia MPPs está baseada nas discussões feitas nos capítulos (1), (2) e (3), utilizadas nesta tese como suporte a metodologia construída.

Após aplicar a metodologia de (MPPs) para definição dos grupos, é ajustado um semivariograma para cada grupo da partição, ficando cada grupo sujeito a estrutura de semivariância que melhor se adapta aos dados, para se aplicar Krigagem Ordinária.

A metodologia Bayesiana, denominada MPPs é proposta para encontrar a partição espacial mais provável de conter dependência espacial, de forma a garantir maior estacionaridade da média. Para isso, nos dois bancos de dados escolhidos, é considerads conhecido o número de cortes em cada um deles, em que para o primeiro é assumido um corte, e para o segundo, dois cortes, utilizando para isso, a variável resposta transformada em semivariancias locais em cada um deles, e como covariável a Soma das Distâncias entre Vizinhos (SUMD), que tiveram os mesmos resultados, quando foram comparados com a utilização da covariável Soma das Distâncias Médias entre Vizinhos (SumMediaDvi), ambas feitas por meio de programação específica em linguagem R.

Para aplicar o método é assumido para o total de dados n, uma quantidade k de grupos, $k \in \mathbb{N}$, $k < n$, ou seja, o número de grupos é um parâmetro conhecido mas os elementos que vão compor cada grupo constituem uma variável aleatória que segue distribuição Normal.

4.1 Dados altimétricos (Dados1)

Os dados foram obtidos da Zona da Mata Mineira, região de Canaã-MG, utilizando dados amostrais dos pontos altimétricos obtidos a partir de dados orbitais, via projeto SRTM, coordenadas UTM da folha $23S$ (IBGE, 2006), composto por 479 amostras da variável, medida em relação ao nível do mar em maio de 2011 para estudo de solo pela Universidade Federal de Viçosa, os quais se mostraram com forte suspeita de mudança na média em alguma subregião da área e grande assimetria, dificultando o ajuste de um único modelo de semivariograma.

A Figura 5 mostra possíveis mudanças na estrutura da dependência espacial, configurada pelas altitudes locais em função da covariável Somas das Distâncias entre os pontos Vizinhos-SUMD (Figura 6, superior direita) e das semivariancias em função da da covariável Distância Média entre Vizinhos e da covariável SUMD, respectivamente, (Figura 6, inferior esquerda e direita).

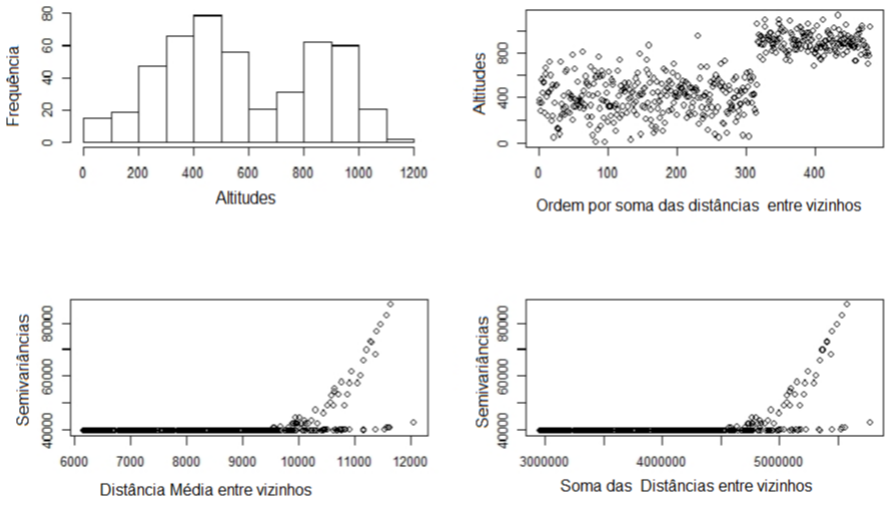

Figura 6: Inspeção de ponto de mudança na média dos dados altimétricos

4.2 Dados Batimétricos- Dados 2

Ferreira et al. (2012) relatam passo a passo a metodologia de coleta dos dados batimétricos que serviram de base para este estudo. Este dados foram cedidos pelo grupo de pesquisa (GEPLH) do Departamento de Engenharia Civil da UFV, os quais, de acordo com os autores, foram coletados em dezembro de 2010, através de um levantamento batimétrico realizado em um dos principais represamentos do Ribeirão São Bartolomeu, no campus da Universidade Federal de Viçosa(UFV) no município de Viçosa, Zona da Mata Mineira, utilizando-se de um ecobatímetro monofeixe equipado com um transdutor de dupla frequência (33/210 kHz) e seguindo as recomendações da *International Hydrographic Organization (IHO)*, que permitiram gerar um arquivo com pontos contendo as coordenadas planimétricas e as respectivas profundidades (XYZ_{bat}), exportado em formatos compatíveis aos softwares *ArcGIS* 10.2.2 e R 3.2.2 visando efetuar análises estatísticas e geoestatísticas.

Os dados batimétricos, apesar de precisos, necessitam de metodologias geoestatísticas para gerar mapas capazes de cobrir toda a área, porém, comumente as estimativas feitas por Krigagem, resultantes de dados desta natureza, apresentam-se com pouca precisão, devido à dificuldade de ajuste do semivariograma. Esta dificuldade de ajuste, na maioria das vezes, são provocadas por flutuações da média de uma subregião para outra, não satisfazendo, completamente, à hipótese de estacionaridade de 1^a ordem.

A perda da estacionaridade de 1^a ordem pode ser provocada por efeitos perturbadores do próprio meio ambiente, tais como, enchentes no centro do lago, queda de encontas nas extremidades do lago, intervenções nocivas do homem e outras circunstâncias ambientais que não podem ser controladas e que contribuem para a falta de normalidade e presença de assimetria acentuada nos dados.

A Figura 7 mostra possíveis mudanças na estrutura da dependência espacial, configurada pelas semivariancias locais, em função da covariável " Médias das Distâncias entre os pontos Vizinhos"(SumMediaDvi).

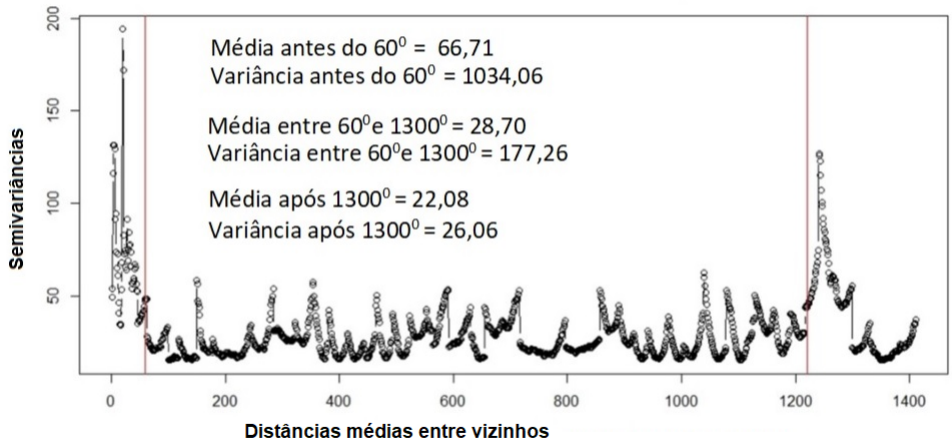

Figura 7: Inspeção de pontos de mudança na média global dos dados batimétricos

4.3 Metodologia proposta

A metodologia proposta para definir a partição nos conjuntos de dados é baseada no Modelo de Partição Produto, com adaptação para modelos espaciais com depenência espacial, nomeadamente considerada Método de Partição Produto Espacial (MPPs), que em inglês é denominada *Spatial Partition Product Method* (sPPM). Estas adaptações para captar a dependência espacial é estabelecida pela inserção da covariável espacial (SUMD ou SumMediaDvi), em que a escolha é feita de acordo com cada banco de dados, aquela que melhor definir os grupos de médias estacionárias.

4.4 Metodologia MPPs

Para garantir que os grupos obtidos pelos pontos de corte tenham dependência espacial, é estabelecida uma vizinhança dentro de um raio, r. Este raio é obtido de uma proporção da distancia mínima entre os vizinhos de acordo com a expressão proposta,

$$r = L * \sqrt{2} * min(distanciadosvizinhos), \tag{61}$$

A expressão (61) é criada para se obter uma matriz de vizinhança local. Para isso, L é considerado um número racional positivo capaz de garantir que nenhum ponto seja solitário, e também, que nenhum ponto tenha vizinhança completa, ou seja, todos os pontos sejam vizinhos de todos os demais pontos. Para garantir estas duas condições, a escolha do valor de L é feita por meio de análise exploratória dos dados. Após assumir um determinado valor de L e aplicar a função que calcula as semivariancias locais para cada ponto, são analisadas as medidas descritivas da variável transformada em semivariancias locais, observando-se os quartis, decis, percentis e também, algumas das medidas de tendência central e dispersão, a fim de verificar se em algum dos quantís analisados há vizinhança nula (ou seja, pontos solitários), caso haja, um novo valor de L é testado e repetido todo o processo, até encontrar o valor ideal para o raio r.

Assumido o raio r, é calculado a matriz de vizinhança e em seguida encontrado a distância mínima entre vizinhos $min(d1)$ em que $d1$ representa o menor valor de uma submatriz das distâncias simétricas diferentes de zero. Este valor é utilizado na expressão, seguindo os passos descritos em 4.4.1.

4.4.1 Etapas do método MPPs proposto

- Para construir a função semivariância local utilizando o *software* R, são utilizados os pacotes *spBayes* e *sp* do R Core **?**.

- Obtida as distâncias entre todos os pontos, obtém-se a distância mínima entre vizinhos para obter o raio r, de acordo com a expressão 61.

- Para construir a matriz de vizinhança, é imposta a restrição de um raio máximo r e aplicada a função *ifelse* do R que transforma as distâncias em indicadores de vizinhança. Com a matriz de vizinhança já especificada é feita a seguinte manipulação matricial:

  As colunas de coordenadas geográficas *UTM*, $dados[, 1]$ e $dados[, 2]$ e da variável de interesse (Altimetria ou Batimetria, conforme for o banco de dados utili-

zado), cada coluna da matriz de dados é multiplicada pela matriz de vizinhança (Vizinhos), por meio de produto direto $\otimes$, como segue:

$$\begin{aligned} M1 &= dados[,1] \otimes vizinhos \\ M2 &= dados[,2] \otimes vizinhos \\ M3 &= dados[,3] \otimes vizinhos. \end{aligned} \tag{62}$$

em que $dados[,1]$ e $dados[,2]$ representam os vetores referentes as coordenadas geográficas X_{UTM}, Y_{UTM}, e $dados[,3]$ representa a variável de interesse Z(altimetria ou batimetria), e *vizinhos* representa a matriz de vizinhança, quadrada de ordem n.

- Obtida a distância mínima e tendo calculado a matriz de vizinhança, após observarem que as medidas separatrizes e de tendências central e dispersão mostraram que todos os pontos possuem uma vizinhança restrita e maior que 1, aplica-se a função (refmetod2) nas três colunas do banco de dados georreferenciado.

  Observação: ao efetuar os produtos descritos na expressão (62), $M1$, $M2$ e $M3$ resultam em matrizes cujos valores são diferentes de zero para os vizinhos existentes e zero para os não vizinhos.

- Obtidas as matrizes $M1$, $M2$ e $M3$ é aplicada uma função para calcular a semivariância local de cada ponto em função de sua vizinhança, que resultará numa nova variável "Semivariância local" , função programada em linguagem R, que guardará as informações relativas a semvariância de cada ponto em função da distância em relação aos seus vizinhos.

Os valores das semivariancias locais representam a variável transformada a qual, ordenada em ordem crescente em relação à covariável " Soma das Distancia entre Vizinhos"(SUMD), ou "Média das Distâncias entre Vizinhos"(sumMediaDvi), é aplicado o modelo bayesiano hierárquico de MPPs.

O modelo hierárquico bayesiano escrito como Modelo de Partição Produto Espacial, nomeadamente MPPs, é utilizado para encontrar o ponto de mudança na média que definirá a divisão dos grupos. Para isso, com a variável de interesse Y, refere-se às "semivariancias locais", ordenada em função da covariável espacial X adotada (SUMD ou sumMediaDvi). No modelo é assumida como conhecida a quantidade de pontos de mudança na média, e desta forma, os elementos de cada grupo, bem como os locais de ocorrências destes pontos são variáveis aleatórias a serem estimadas via MPPs. Além disso, assume-se que os dados seguem distribuição normal (condição justificada no texto).

O MPPs é descrito para um ponto de mudança na média e para dois pontos de mudança na média, descritos como segue:

Modelo MPPs de um corte,

$$\begin{aligned} Yi &\sim Normal(\mu, \sigma^2) \\ \mu &= \alpha + \beta_{J_i}(x_i - x_k) \\ J_i &= 1; i \leq k; \\ J_i &= 2; i > k \\ \sigma &= \tau^{-1} \end{aligned} \tag{63}$$

Modelo MPPs de dois cortes,

$$\begin{aligned} Yi &\sim Normal(\mu_i, \sigma^2) \\ \mu_i &= \alpha_m + \beta_{J_i}(x_i - x_k) \\ J_i &= 1; i \leq k; \\ J_i &= 2; k \leq i \leq k1 \\ J_i &= 3; i \geq k1 \\ m &= \{1, \ldots, J-1\} \\ \sigma &= \tau^{-1} \end{aligned} \tag{64}$$

A esperança de Y, $E(Y) = \alpha$ ou $E(Y) = \alpha_m$, são o ponto ou pontos de mudança na média, para os modelos com um ou dois cortes, respectivamente. Os

coeficientes β_1 e β_2 são utilizados no modelo com suspeita de um ponto de mudança na média (63) e β_1, β_2 e β_3, no modelo com suspeita de dois pontos de mudança na média (64).

A distribuição normal foi escolhida por ser a mais indicada para ajustar o modelo com o objetivo de garantir que os grupos sejam mais homogêneos e simétricos em torno da média no espaço e para evitar a restrição de que os parâmetros α e β_{J_i} possam ser apenas positivos (restrição imposta quando é utilizada a função Gama), ou inteiros (restrição imposta quando é utilizada a distribuição Binomial).

Ressalta-se que as estimativas de pontos a ser interpolados por krigagem ordinária são feitas considerando um intervalos de confiança de 95% de uma distribuição normal. Diante disso, utilizar a distribuição normal parece uma escolha mais vantajosa.

Para implementação do modelo hierárquico bayesiano de MPPs, em suma são utilizados os seguintes passos:

- Calcular a matriz de distancias simétrica entre as coordenadas dos pontos.
- Encontrar a distância mínima diferente de zero e a distância máxima entre os pontos.
- Estabelecer uma matriz de vizinhança para todos os pontos baseada num fator relacionado a distância mínima, utiliza-se a expressão:

$$r = L\sqrt{2}min(distância)$$

  em que L é uma constante a qual se assume um valor capaz de garantir que todos os pontos tenham vizinhos.

- Multiplicar cada coluna do banco de dados por produto direto (produto de ronecker) pela matriz de vizinhança, gerando assim n bancos de dados.
- Aplicar a fórmula do semivariograma empírico às en-ésimas colunas resultantes da multiplicação da matriz de vizinhança pela variável resposta.

- Calcular a semivariância para cada ponto (γ_i)usando o raio r e a quantidade de elementos em cada vizinhança dentro do raio r estabelecido de acordo com o item 3.

- Calcular a média de cada coluna de distañcias D que se refere a vizinhança de cada ponto.

- Criar um arquivo com extensões de texto ou em bloco de notas das coordenadas geográficas Longitude (X_{UTM}) e Latitude (Y_{UTM}), variável resposta (semivariancias locais) e variável regressora (distância média entre vizinhos ou soma das distâncias entre vizinhos) com a variável regressora na forma crescente de ordenação utilizando o *software* R core (Team, 2015).

- Aplicar o método $MCMC$ por meio de Metropolis Hasting para gerar as distribuições de probabilidade dos locais dos pontos de corte k e $k1$, e então identificar os pontos de mudanças na média.

- Caso as distribuições condicionais não tenham forma analítica e as distribuições a posteriori não sejam conhecidas, usar o *software* WINBUGS (Lunn et al., 2000) ou bibliotecas do R, tais como *ggmcmc, ggplot2, mcmc*, dentre outras.

- Delimitar os dois agrupamentos e separar em dois bancos de dados distintos, ajustar um semivariograma para cada um e aplicar Krigagem em cada grupo de amostras, utilizando programação no *software R* (Team, 2015).

- Avaliar a qualidade dos mapas por meio de validação cruzada, variância de Krigagem e demais medidas como AIC e tirar conclusões sobre a metodologia proposta.

No modelo (1) são assumidas cinco priores, como segue em 65:

$$\begin{aligned} k &\sim U\{1,2,\ldots,n\} = \frac{1}{n} \\ \tau &\sim Gama(0,001;0,001) \\ \alpha &\sim N(0,00;0,000001) \\ \beta_j &\sim N(0;1) \\ J_i &\sim cat(U\{1,2\}) \end{aligned} \tag{65}$$

Em que $cat(U\{1,2\})$ representa uma variável categórica que são assumidas duas categorias, $\{1,2\}$. São utilizados também um valor inicial para k, dois valores iniciais para os parâmetros betas e um valor inicial para τ para que o WINBUGS gere as distribuições.

4.5 Análise exploratória

Para aplicar a metodologia proposta foram utilizados dois conjuntos de dados, um conjunto de dados (Dados 1) com 479 amostras georreferenciadas, extraídas em uma fazenda no município de Canaã, Zona da Mata Mineira, no ano de 2011, cuja variável de interesse é a altitude do terreno medida ao nível do mar.

O segundo banco de dados é um conjunto de dados batimétricos (Dados2) com 1411 amostras georreferenciadas, extraídas da lagoa da UFV em um estudo batimétrico, os quais obtiveram-se amostras da profundidade em cada ponto, medidas em valores negativos a partir de um ponto zero (z=0), determinado pelo limite da superfície da água; estas amostras foram coletadas e organizadas em arquivo com extensões *.txt*, *.csv* e *.dat* pelo Grupo de Estudos e Pesquisas em Levantamentos Hidrográficos da UFV (GEPLH-UFV) e cedidas para aplicação deste trabalho.

5 RESULTADOS E DISCUSSÃO

Na Tabela 2 são apresentadas algumas medidas descritivas, referentes aos dois conjuntos de dados completos que justificaram a necessidade de aplicar um ponto de corte no conjunto de dados altimétricos (Dados1), e dois pontos de corte no conjunto de dados batimétricos (Dados2) como o uso da metodologia proposta, MPPs.

Tabela 2: Medidas descritivas referentes aos bancos de dados completos de altimetria (Dados1) e de batimetria (Dados2)

Dados1	Mín	1º quartil	Mediana	Média	3ºquartil
	101,33	709,65	855,73	807,89	925,52
	Máx.	Var.	Ass.	curtose	C.V%
	1287,27	34.826,15	-0,76	3,59	25,6
Dados2	Mín	1º quartil	Mediana	Média	3ºquartil
	-5,54	-4,83	-4,46	-4,18	-3,81
	Máx.	Var.	Ass.	curtose	C.V%
	0	1,1209	2,09	8,57	25,3

Por meio do teste de normalidade *Shapiro-Wilk* verificou-se a não normalidade dos dois conjuntos de dados, Dados1 e Dados2, apresentaram (valor-p: $1,64 \times 10^{-11}$ e $2,2 \times 10^{-16}$, respectivamente), Assimetria ($-0,76$ e $2,09$), Curtose ($3,59$ e $8,57$) e Coeficiente de Variação ($25,6\%$ e $25,3\%$), na ordem respectiva aos conjuntos de dados, mostrando que os dois bancos de dados não são totalmente estacionários em termos de média, porque os coeficientes de Variação foram maiores que

20%, as Assimetrias e Curtoses foram elevadas, indicando alto grau de achatamento da curva de distribuição, principalmente no segundo banco de dados (Dados2), sinalizando que a média não é representativa de toda a área, ou seja, existe a possibilidade de duas ou mais médias estacionárias nos dois conjuntos de dados.

O comportamento espacial da média, apresentado na Figura 8, referente ao banco de dados altimétricos (Dados1), mostraram que em uma parte da área encontra-se com bastante variabilidade e na outra parte, com distribuição mais homogênea, indicando boa dependência espacial. De acordo com o gráfico mostrado à direita superior da mesma figura, a média anterior á amostra de posição 320^0 é inferior à média dos valores posteriores à posição 320^0. O gráfico superior esquerdo mostra por meio do histograma que aparentemente existem duas distribuições distintas, enquanto que na parte inferior da figura apresentam dois gráficos similares, diferenciando-se apenas pela covariável espacial adotada. Aparentemente, a maior parte da área tem pouca dependência espacial, e isto dificulta o ajuste de um único modelo de semivariograma, porque a parte com maior possibilidade de dependência espacial é contaminada pela influência das duas estruturas de médias.

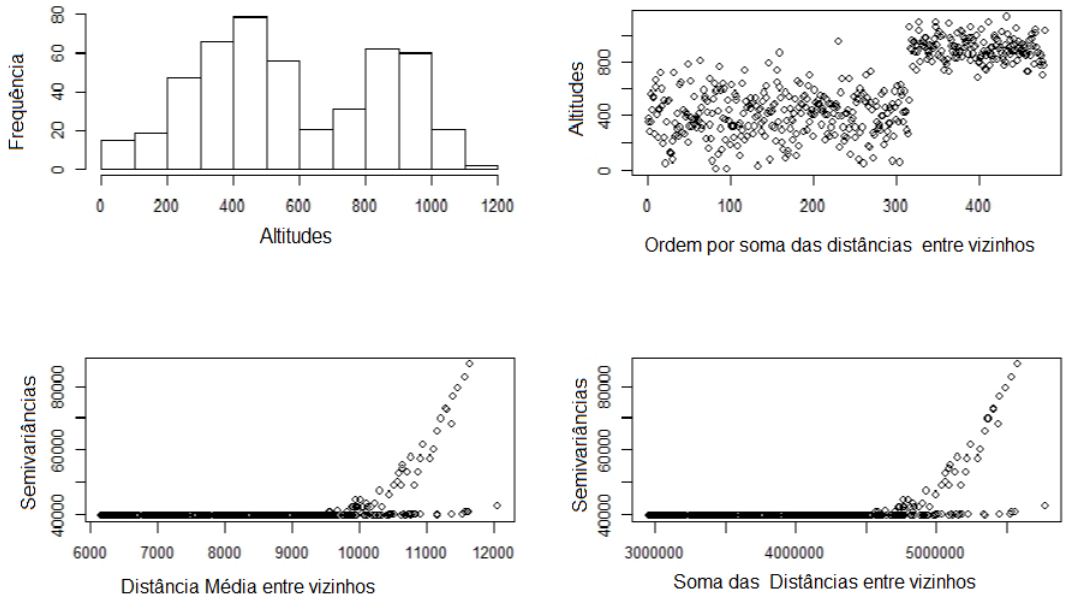

Figura 8: Inspeção de ponto de mudança na média dos dados altimétricos (Dados1)

A Figura 9 referente ao banco de dados batimétricos(Dados2), está mostrando o comportamento das semivariancias locais em função das Médias das Distancias entre Vizinhos, em que se tem forte suspeita de dois pontos de mudança na média, ou seja, a possibilidade de haver três grupos amostrais distintos. Como se pode observar na figura, até a amostra 60^0, a média das semivariancias locais ficou em torno de 66, a média entre as amostras de posições 60^0 e 1300^0 ficou em torno de 29 e após a amostra de ordem 1300^0 ficou em torno de 22, indicando menor variabilidade comparada às médias dos demais grupos. Aparentemente, cada uma das três partes da área tem dependência espacial diferentes, dificultando o ajuste de um único modelo de semivariograma porque o mesmo é arruinado pela influência das três médias.

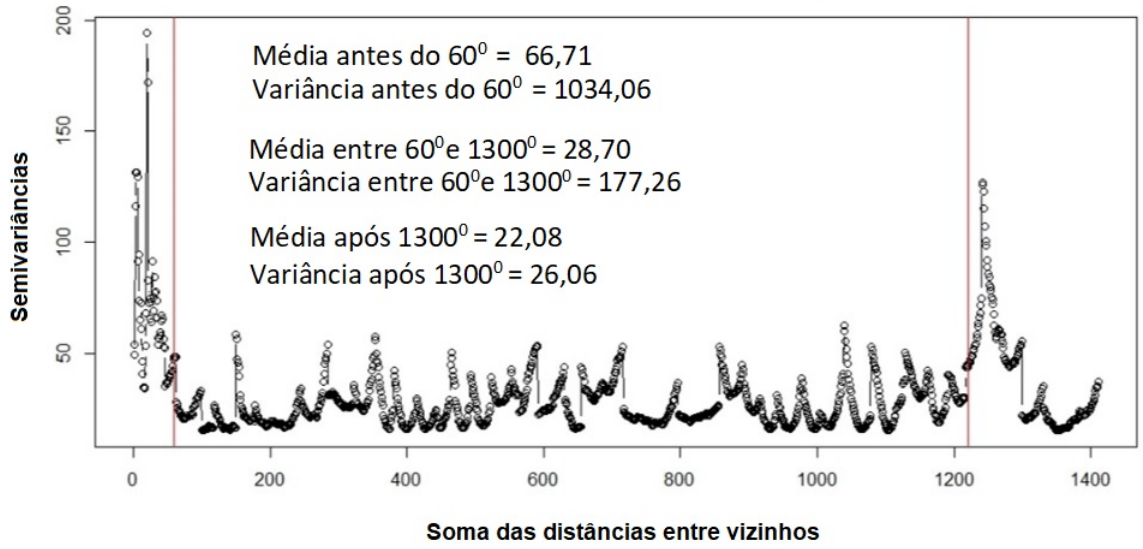

Figura 9: Inspeção de pontos de mudança na média dos dados batimétricos

A análise gráfica feita por meio das figuras 8 e 9, foi importante na tomada de decisão quanto a necessidade de se utilizar o método MPPs. Como em todo modelo espacial, os pontos não estão ordenados numa sequência natural como ocorre nos casos de séries temporais; para contornar o problema da falta de ordenação foi utilizada a estratégia de ordenar os dados em função da covariável espacial (SUMmediaDvi) e (SumD), nos respectivos bancos de dados. Em ambos os conjuntos de dados, utilizando qualquer das duas covariável, apontaram locais similares de possíveis mudanças na média.

Para a obtencão da distribuição a posteriori por meio do modelo bayesiano de MPPs, utilizou-se o *software* livre WINBUGS, Lunn et al. (2000), para detecção do ponto de mudança na média porque o modelo não tem distribuição a posteriori conhecida e as condicionais completas não podem ser obtidas analiticamente, sendo encontradas por aproximações numéricas, via *Metrópolis-Hasting*.

A seguir são apresentados os resultados da aplicação do método MPPs aos dois bancos de dados, Dados1 e Dados2.

5.1 Resultados para um ponto de Corte K - Dados 1

Utilizando o banco de dados de altimetrias (Dados1), foi como causas da dificuldade de se obter dependência espacial, a contaminação dos dados provocada pela duplicidade ou multiplicidade de estruturas de médias que dificultam obter a dependência espacial devido a contaminação dos dados pela interferência das várias médias, estas características estão retratadas no banco de dados de altimetria, cujas discussões do problema foram apresentadas e a aplicação do MPPs para resolver solucionar tal problema estão a seguir.

A Figura 8, apresentada anteriormente, mostrou claramente o problema de duplicidade da média, presente no cenário referente aos Dados1, em que para solucionar o problema, é assumida a existência de uma partição com duas partes, ou seja, a existência de um possível ponto de mudança K, na média das semivariancias locais em função da covariável SumD, em que foi aplicada a expressão do

modelo, apresentada no capítulo (4), nas seções (4.3, 4.4 e 4.4.1) e na expressão(61).

Para obter os resultados desejados, o banco de dados (Dados1) foi ordenado em relação á covariável SumD, com o objetivo de garantir que os agrupamentos se formem com grupos de dados contíguos espacialmente.

Com o uso do aplicativo WINBUGS, foram gerados os resultados apresentados na Tabela 3.

Tabela 3: Medidas descritivas referentes ponto de corte, k, estimado pelo modelo descrito no capítulo (4) e subseção (4.4.1)

Média	Desv.pad.	MC erro	LI(2,5%)	Mediana	LS(97,5%)	Burn-in	Ream.
314,2	5,901	0,1196	288,0	315,0	318,0	1001	4000

[1]

Como mostrado na Tabela 3, a escolha do ponto de corte ficou bem definido pela média e pela mediana, ambas muito próximas e dentro do intervalo de credibilidade.

A Figura 10 mostra o intervalo de credibilidade bem definido com limite inferior em 285 e superior em 320, indicando que o ponto de corte pode ser estabelecido dentro deste intervalo.

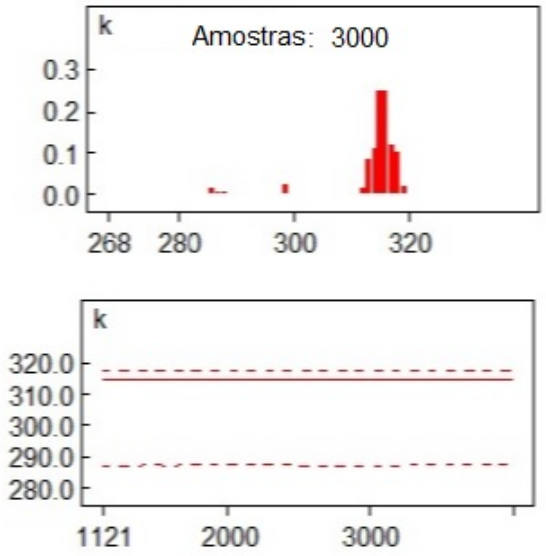

Figura 10: Densidade de K (acima); Intervalo de credibilidade do ponto de mudança na média (abaixo).

A Figura 11 mostra boa convergência e autocorrelação aproximadamente nula em toda a cadeia.

O ponto de corte K, determinado pela expressão do modelo (63) e mostrado na Tabela 3, foi detectado na posição do 315° elemento, ou seja, considerando as semivariancias locais em função da covariável espacial SUMD, é mais provável que o conjunto de dados possa ser dividido em dois grupos, sendo o grupo 1 (G1) formado de 1 até o 315° elemento ($n = 315$ pontos amostrais) e o grupo 2 (G2), iniciando no 316° elemento até o 479° elemento, ficando o grupo G2 com 164 pontos amostrais.

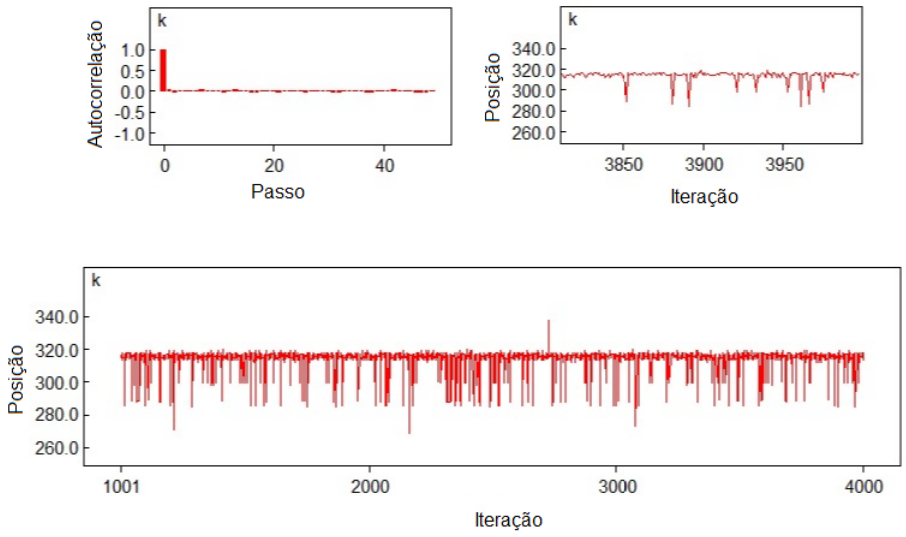

Figura 11: Convergência e autocorrelação da cadeia

Após definido o ponto de corte para o conjunto de dados altimétricos (Dados1), na posição 315°, foi feita a análise espacial dos três grupos amostrais, como mostrado a seguir.

5.1.1 Análise exploratória e espacial do Grupo 1 - Dados 1

Os resultados da análise exploratória do Grupo 1, estão apresentados na Tabela 4.

Tabela 4: Medidas referentes as coordenadas referentes ao grupo 1

	Coordenadas		Distancias
	X	Y	
Mínima	739486,9	7710262	86,85
Máxima	755361,0	7723030	18817,97

De acordo com a Tabela 4, a distância mínima se manteve a mesma, enquanto a distância máxima do Grupo 1 ficou menor que a distância máxima entre as amostras do grupo amostral total.

Tabela 5: Medidas descritivas referentes ao grupo1

Mínimo	1º quartil	Mediana	Média	3ºQuartil
101,33	637,00	692,705	756,75	764,00
Máximo	Variância	Assimetria	Curtose	C.V.%
1287,27	41725,85	0,338	-0,106	27

Como pode ser verificado na Tabela 5, ao ser aplicada a metodologia proposta, MPPs, e efetuado o corte, o Grupo 1 reduziu a Assimetria em relação a Tabela completa de $-0,76$ para $0,338$ e o Coeficiente de Variação de 26% manteve-se em torno de 27%.

O teste de normalidade Shapiro-Wilk, via estatística $W = 0,99627$ e $p - valor = 0,6637$, comprovou normalidade nos dados do Grupo 1.

A distribuição dos dados está representada na Figura 12, a qual

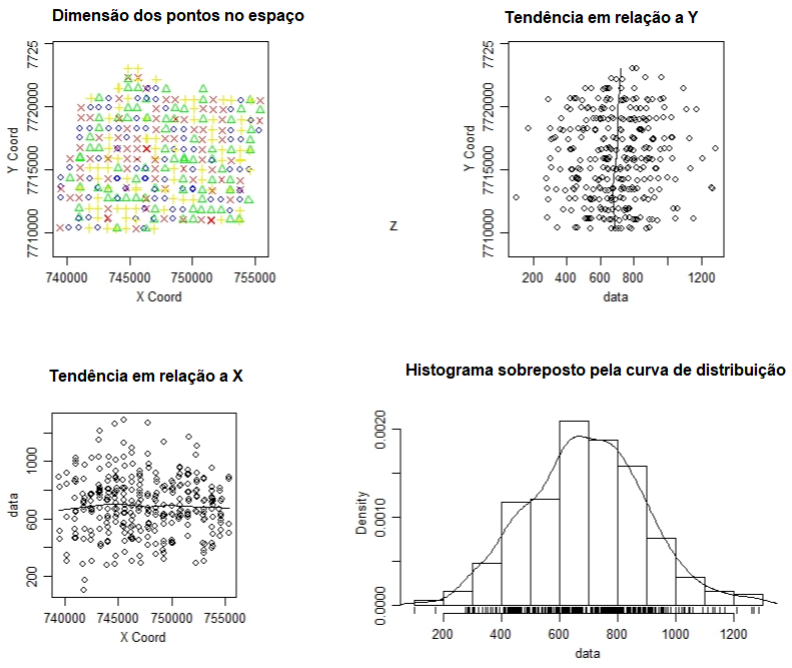

Figura 12: Representação exploratória referentes ao grupo 1

verifica-se que os dados estão bem distribuídos em torno da média.

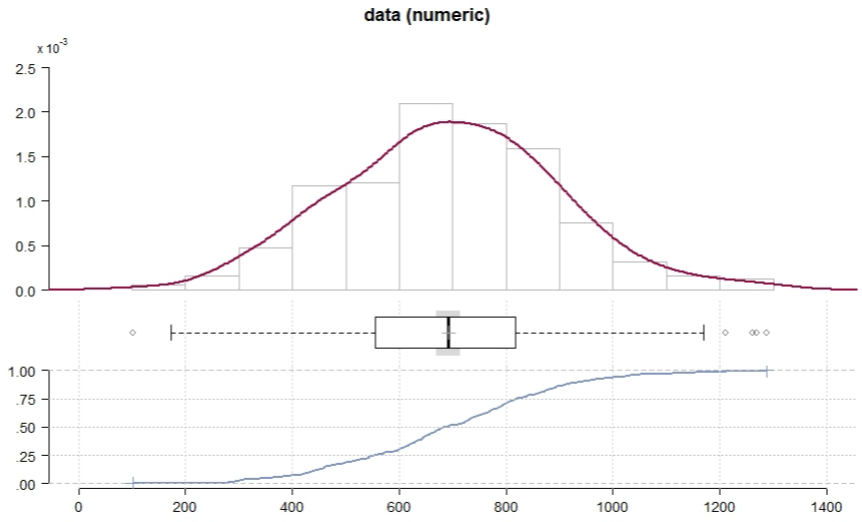

Figura 13: Curva de densidade e função acumulada referentes ao grupo 1

De acordo com a Figura 13, observa-se que o grupo amostral 1 segue distribuição normal e na Figura 14 está mostrado que embora o Grupo 1, na inspeção preliminar, apresentava-se com aparente falta de dependência espacial, foi possível o ajuste do modelo de semivariograma.

Como mostrado na Figura 14, o semivariograma empírico ajustou ao modelo teórico "Cúbico" por meio dos métodos OLS (Mínimos Quadrados Ordinários)[2] e WLS (Mínimos Quadrados Ponderados) [3], nos dois métodos obteve-se elevado Efeito Pepita.

O elevado valor do Efeito Pepita era esperado porque a Figura 8 já antecipava a ideia de possível pouca dependência espacial em parte dos dados. As representações à esquerda, estão mostrando que o comportamento das semivariancias são muito parecidos para os diferentes raios de estimação.

[2]é uma técnica para estimar parâmetros desconhecidos num modelo de regressão linear nos parâmetros que gera estimativas de coeficientes imparciais, relativamente próximas dos valores reais da população (variação mínima), garantido pelo teorema de Gauss-Markov como BLUE(*best linear unbiased estimator*).

[3]é um estimador semelhante ao OLS, com a diferença de ser ponderado pelas variâncias quando as observações apresentam variância não constante ao longo das observações (heterocedasticidade).

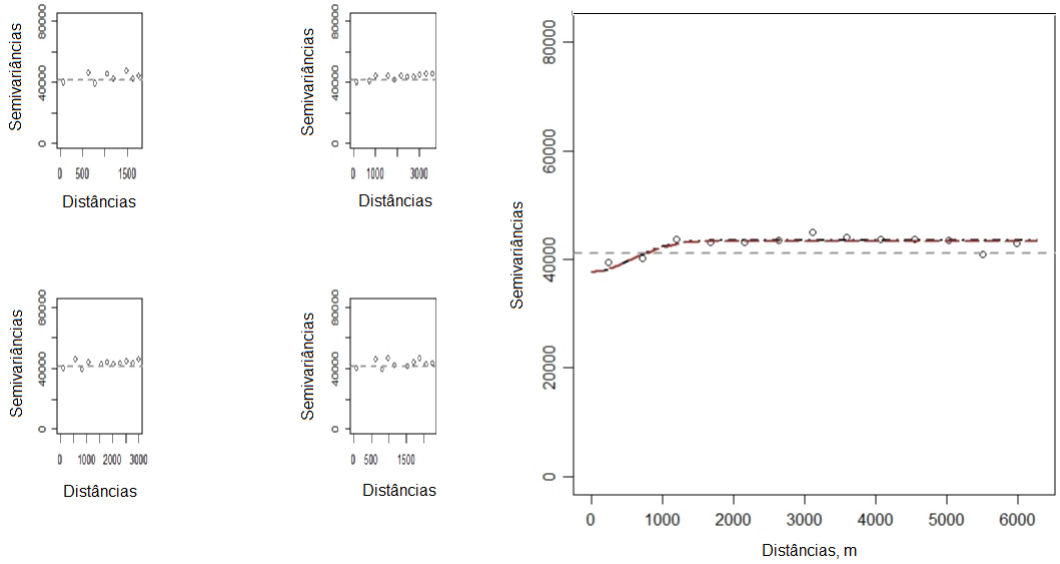

Figura 14: À esquerda: Semivariogramas empíricos do grupo 1 para diferentes raios de estimação; à direta: Semivariograma Cúbico com ajuste por OLS (linha preta) e WLS (linha marron).

Na Tabela 6 são mostrados os resultados do ajuste e validação do semivariograma.

Tabela 6: Parâmetros do Semivariograma e da qualidade do ajuste

GRUPO 1

Modelo	Método	Contribuição	Alcance prático	Efeito pepita	RSE	
Cúbico	WLS	5616,98	1874,21	37707,41	203,043	
SSE	SSE.aj.	ME	MEP	DesvpEP	Coef.reg.	
					Beta0	Beta1
0,00031	-0,0029	0,027	0,000039	0,985	698,90	-0,009

WLS: Mínimos quadrados ponderados; RSE: Erro Padrão residual; SSE: Erro quadrático padronizado; SSE.aj.: Erro quadrátrico padronizado ajustado; ME: Média dos erros; MEP: Média dos erros padronizados; DesvpEP: Desvio padrão dos erros padronizados; Coef.reg.: Coeficientes da Regressão de validação.

Nos resultados mostrados na Tabela 6, os parâmetros de qualidade do ajuste do Semivariograma Cúbico, confirmaram por meio de valida cão cruzada, que a média dos erros padronizados (MEP) ficou próxima de zero ($0,000039$) e o desvio padrão dos erros padronizados ficou próximo de um ($0,985$), indicando boa qualidade do ajuste. Embora o alcance prático de $1.874,21$ metros não seja um valor elevado comparado à distância mínima entre dois pontos ($(86,81)$ metros), pode-se considerar que, em média, cada ponto foi estimado por 21 pontos vizinhos.

Na Figura 15 é apresentado o mapa de Krigagem Ordinária e a representação gráfica dos pontos amostrais do Grupo 1 e na Figura 16, a sobreposição destes pontos sobre o mapa de krigagem Ordinária.

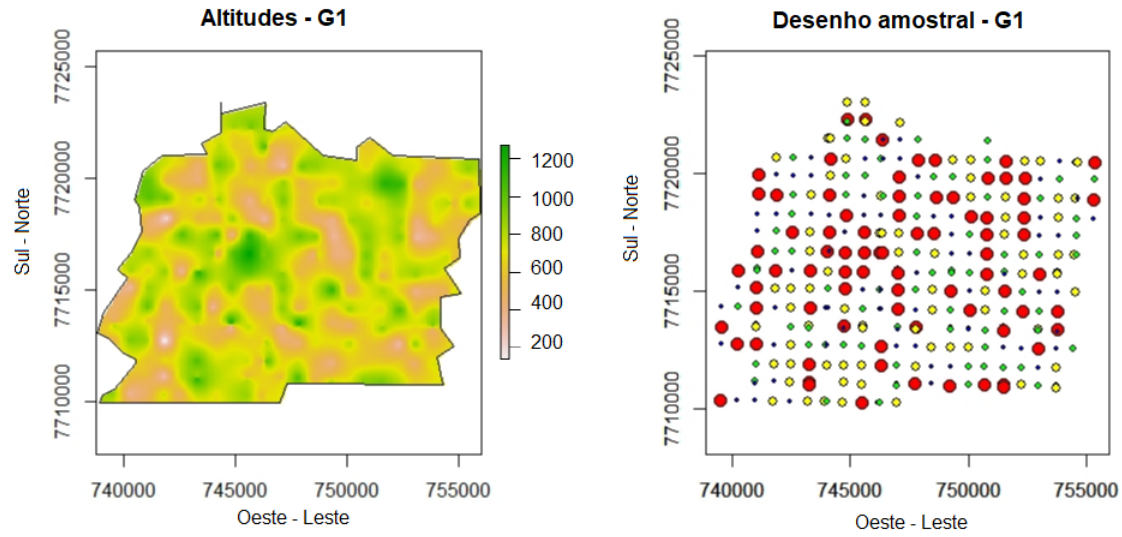

Figura 15: À esquerda: Mapa de Krigagem Ordinária; à direita: Mapa dos pontos amostrais do grupo 1.

Os pontos apresentados no desenho da Figura 15 à direita, são proporcionais as suas dimensões e cores, sendo os pontos azuis, muito pequenos,representando os menores valores, os pontos verdes, de tamanho pequeno, representando os valores maiores que aqueles representados com pontos azuis, os pontos de cor amarela, representando os valores intermediários e os pontos vermelhos,

representando os maiores valores do conjunto amostral referente ao Grupo 1.

Pode se constatar visualmente no mapa da Figura 16 que o método de Krigagem Ordinária estimou valores muito próximos dos valores reais.

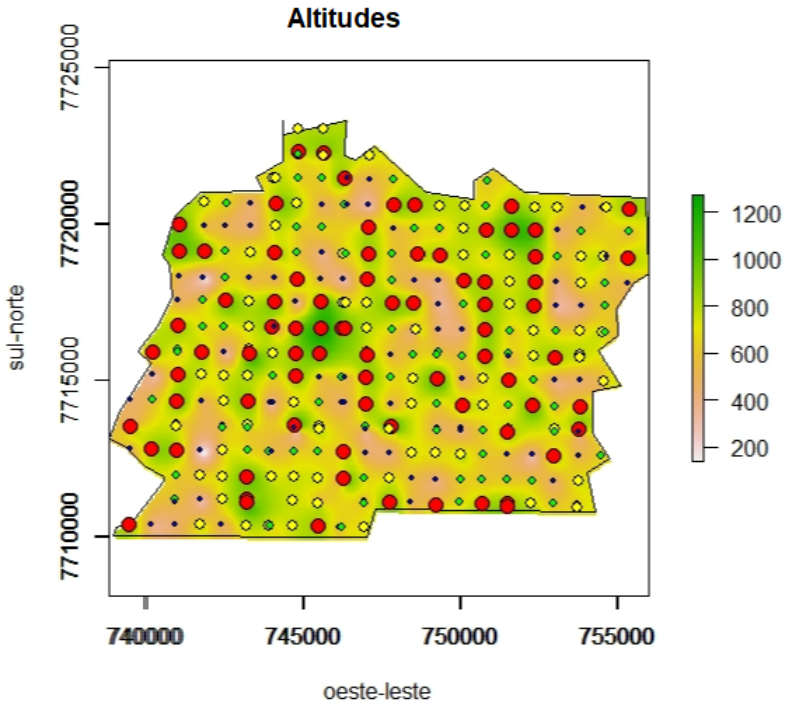

Figura 16: Mapa de Krigagem Ordinária do Grupo 1 com sobreposição dos pontos amostrais.

O teste Shapiro-Wilk aplicado à Krigagem Ordinária do Grupo 1, apresentaram estatísticas W= 0.9972 e $p - valor = 0.8747$, indicando que os valores preditos seguem distribuição normal.

A Figura 17 está mostrando a regressão feita entre os valores observados e preditos.

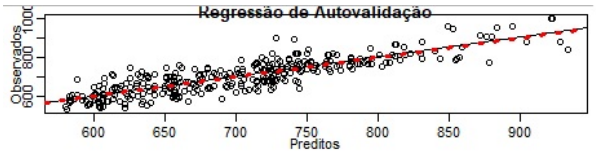

Figura 17: Valores preditos versus valores obervados na autovalidação .

De acordo com a Figura 17, a validação cruzada os valores preditos em função dos valores amostrais ajustaram bem á reta indicando que os valores preditos tem boa precisão.

Algumas medidas referentes a qualidade da Krigagem Ordinária serão apresentadas no final da seção 5.1.4, em uma tabela comparativa dos dois grupos em relação ao conjunto Amostral Total.

5.1.2 Análise exploratória espacial do Grupo 2 - Dados 1

Os resultados da análise exploratória referente ao Grupo 2 são apresentados na Tabela 7.

Tabela 7: Medidas referentes as coordenadas do grupo 2

	Coordenadas		Distancias
	X	Y	
Mínima	739475,1	7702222	86,81
Máxima	754423,5	7710251	14.960,89

De acordo com os resultados da Tabela 7, verificou-se que houve redução na distância m axima entre dois pontos comparada à distância máxima obtida no banco de dados completo.

Na Tabela 8 estão mostradas as medidas descritivas do Grupo 2, em que se verifica a diminuição do C.V. e da Assimetria comparada às mesmas medidas obtidas no banco de dados completo, indicando que o Grupo 2 apresenta maior concentração em torno da média.

Outro fator importante foi a garantia de normalidade dos dados, confirmada no teste Shapiro-Wilk de normalidade, cuja estatística $W = 0,98796$ e $p - valor = 0,1732$, garantindo com 95% de confiança, que não se rejeita a hipótese nula (de normalidade).

Tabela 8: Medidas descritivas referentes ao grupo 2

Mínimo	1º Q.	Mediana	Média	3º Q.
705,98	855,73	899,895	906,115	951,43
Máx.	Variância	Assim.	curt.	C.V.
1167,23	6987,16	0,298	3.39	0,092

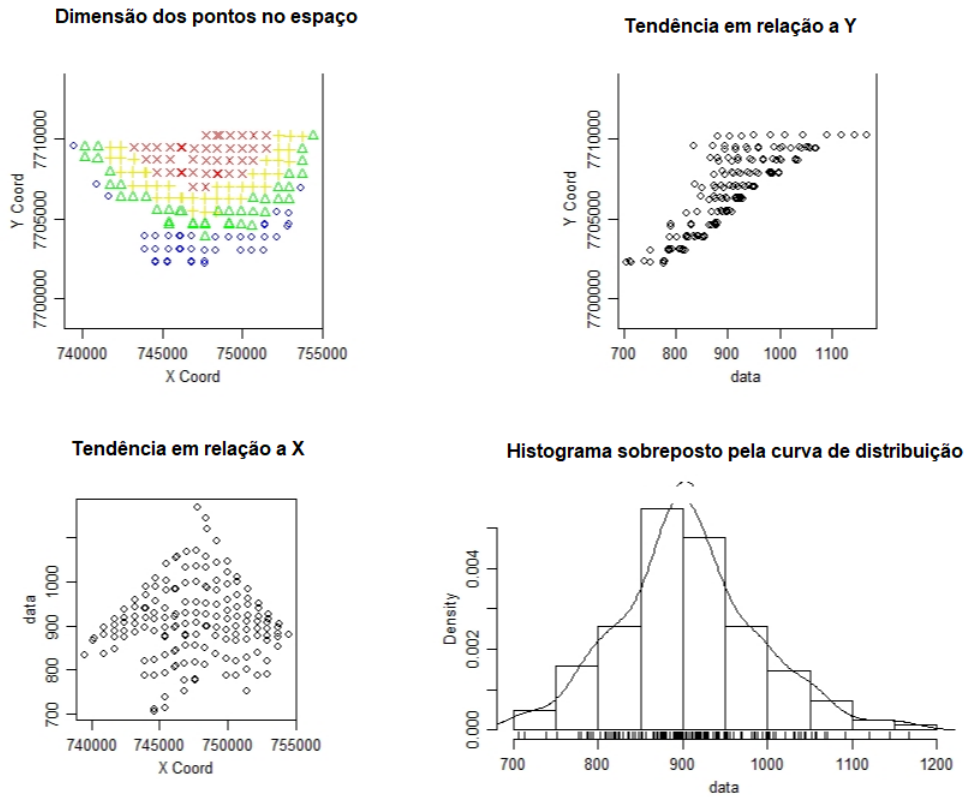

Figura 18: Representação gráfica dos dados referentes ao Grupo 2

Na Figura 18, no canto esquerdo superior, é observado que a distribuição de valores na área apresenta-se com continuidade e não aleatoriedade na distribuição, indicando comportamento com boa estrutura de covariância. No canto esquerdo (inferior) e no canto direito (superior), mostram tendências nas coordenadas X e Y, respectivamente e no canto direito (inferior), a figura mostra o comportamento dos dados aproximando-se da normalidade, que é um dos sinalizadores de

boa estrutura de covariância.

A Figura 19 mostra o semivariograma Gaussiano ajustado aos dados do Grupo 2, em que se pode observar que o semivariograma empírico ajustou perfeitamente ao modelo e obteve-se efeito pepita nulo.

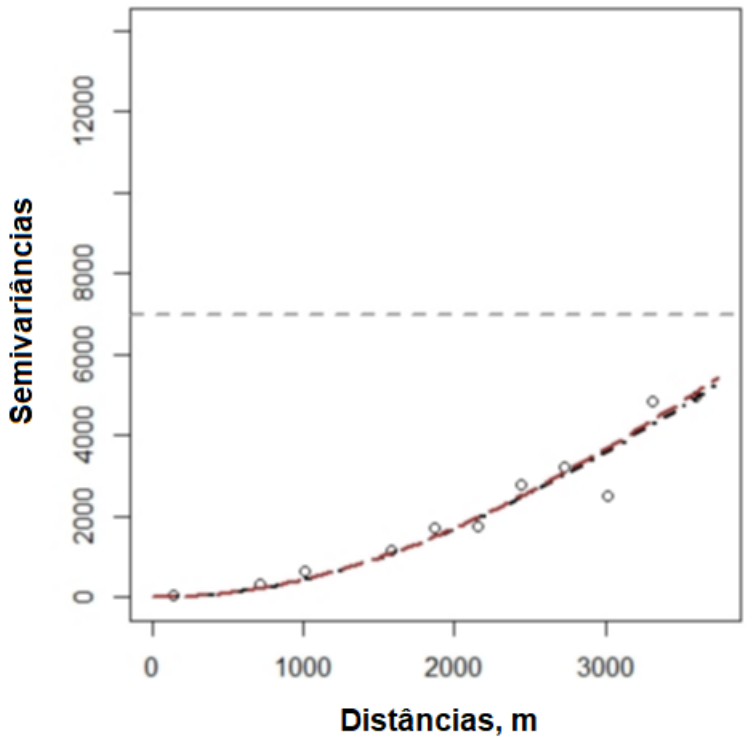

Figura 19: Semivariograma Gaussiano ajustado por WLS (linha marrom) e OLS (linha vermelha) ao grupo 2.

Os resultados referentes aos parâmetros do semivariograma ajustado podem ser observados através da Tabela 9.

Tabela 9: Qualidade do ajuste referente ao grupo 2

Método de ajuste	WLS
Modelo ajustado	Gaussiano
Alcance prático	10.795,54
Contribuição	17.465,762
Patamar	17.465,762
Efeito-pepita	0,0

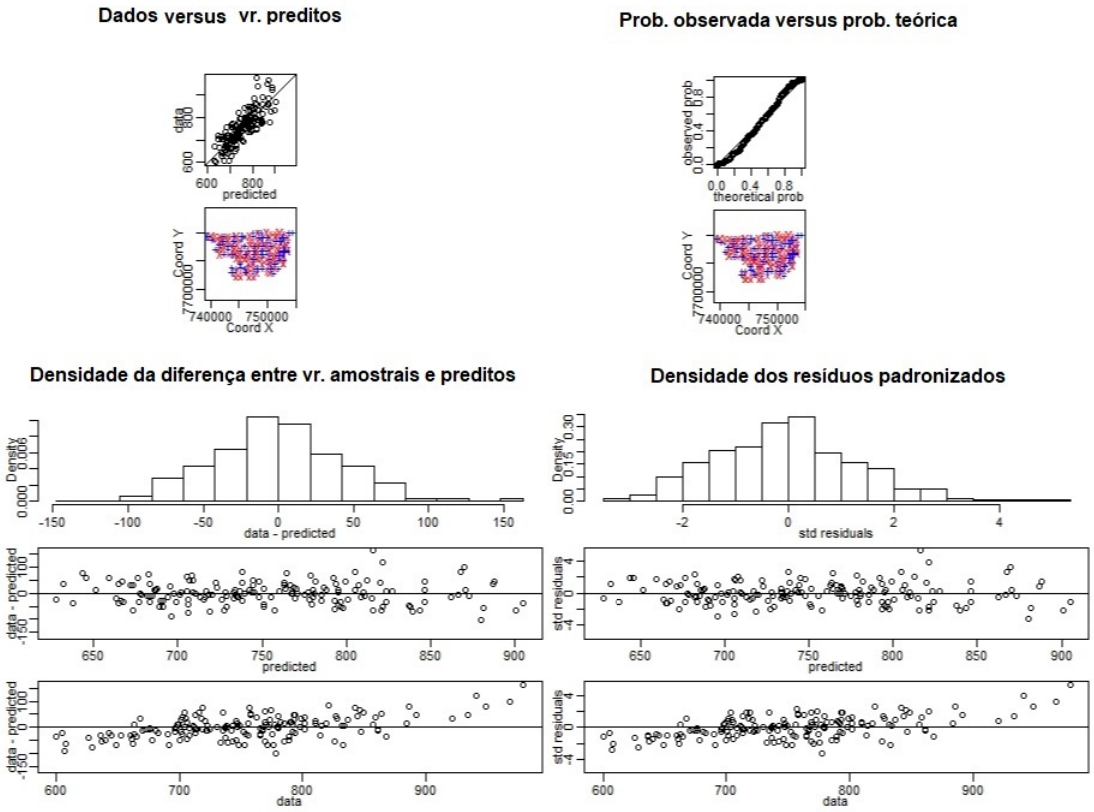

Figura 20: Probabilidades referentes aos valores preditos do grupo 2.

Como se pode observar na Figura 20, os resíduos e os resíduos padronizados, comparados aos dados e aos valores preditos, aproximaram de zero. A densidade dos resíduos e dos resíduos padronizados mostraram-se bastante simétricos e a regressão das probabilidades teóricas e preditas aproximaram-se de uma reta com inclinação de 45°, indicando uma função identidade, assim como a regressão dos valores preditos e observados.

O teste de normalidade, Shapiro-Wilk dos valores preditos para o Grupo 2, apresentou normalidade, comprovada pelas estatísticas $W = 0,99624$ e $p - valor = 0,173$, em que, com 95% de confiança, não se rejeita a hipótese nula (de normalidade).

A Figura 21 mostra os mapas de Krigagem Ordinária do Grupo 2 com sobreposição dos pontos amostrais.

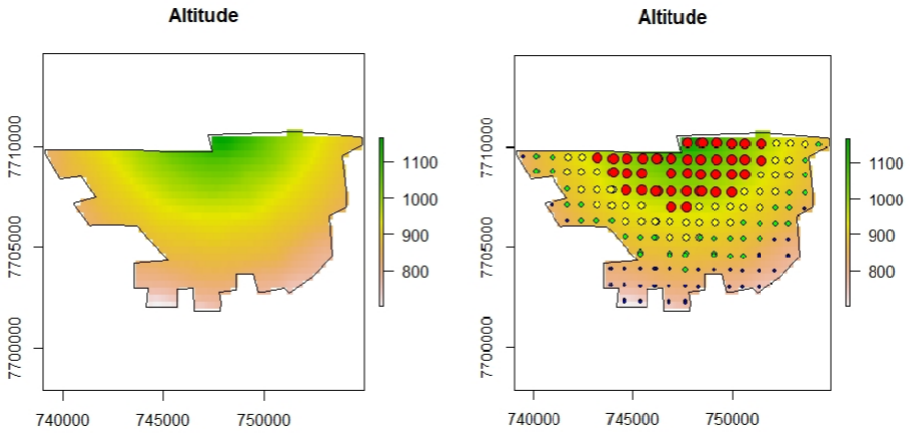

Figura 21: À esquerda: Mapa de Krigagem Ordinária; à direita: sobreposição dos pontos amostrais do Grupo 2.

O Mapa de Krigagem Ordinária do Grupo 2, mostrado na Figura 21, apresentou excelente acurácia, comprovadamente observada com sobreposição dos pontos amostrais na região do mapa de Krigagem Ordinária, em que se nota correspondência à dimensão da cor.

Visualmente, os valores de cada faixa de dados amostrais do Grupo 2 foram preditos corretamente, além das medidas de Assimetria, Curtose e Coeficiente de Variação apresentarem inferiores ou muito próximas dos valores amostrais do próprio grupo.

5.1.3 Medidas da Krigagem Ordinária da "Amostra Total" - Dados 1

Algumas medidas descritivas da variável de interesse (altitudes), referentes ao banco de dados completo (Dados1), foram apresentadas anteriormente nas Tabelas 1 e 2, bem como o resultado do teste de normalidade Shapiro-Wilk, que comprovou a não normalidade dos dados.

Com a tentativa de verificar o comportamento dos dados do banco de dados completo de Altimetria, foram realizadas análises da densidade de distribuição e distribuição acumulada dos dados, os quais observou-se que o valor de assimetria

$-0,78$ e curtose $0,58$ caracteriza-se por uma curva afilada à direita.

A Figura 22 mostra a densidade da distribuição e a função acumulada da amostra completa. Apesar da constatação de dificuldades de ajuste do semiva-

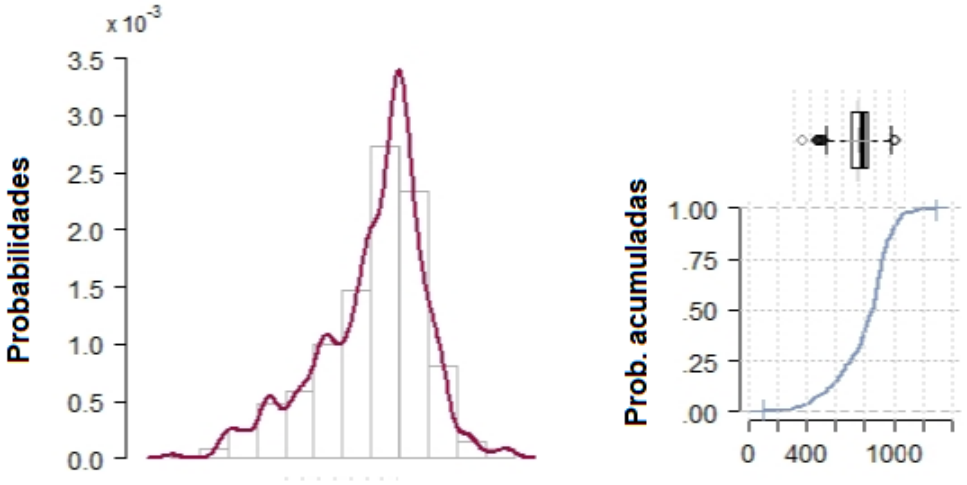

Figura 22: À esquerda: Função de Densidade de Probabilidades e à direita: Distribuição Acumulada de Probabilidades referentes a amostra completa.

riograma, o mesmo ajustou-se ao modelo "Matérn", utilizando o método de ajuste OLS e WLS, como mostrado na Figura 23.

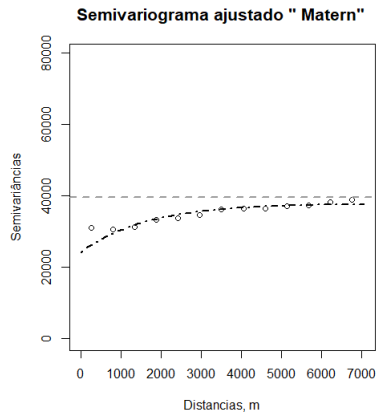

Figura 23: Semivariograma Matérn ajustado via WLS (linha pontilhada) e OLS(linha tracejada).

Pode se observar na Figura 23 que, apesar que o semivariograma ajustou-se ao modelo "Matérn", o elevado efeito pepita e o pequeno alcance prático comparado à distância máxima entre os pontos amostrais, de 21.161, 08 metros, indicam que a qualidade do mapa gerado não é boa.

Por meio da Tabela 10 são apresentados os valores referentes aos parâmetros de ajuste do semivariograma para a amostra completa.

Tabela 10: Parâmetros ajustados referentes a amostra completa

Método de ajuste	OLS(Mínimos Quadrados Ordinários)
Modelo ajustado	Matérn
Alcance prático	2.841,47
Contribuição	16.065,81
Patamar	36.166,46
Efeito-pepita	20.100,85

A Tabela 10 está mostrando que os dados ajustaram ao modelo Matérn por meio do método de ajuste OLS com o melhor alcance prático conseguido, de 2.841, 47 metros e a Contribuição de 16.065, 81, enquanto que o efeito pepita foi

de 20.100, 85. Apesar de elevado efeito pepita, pode-se assumir que os dados tem dependência espacial pois o modelo de semivariograma ajustou com um alcance bem maior que a distância mínima entre as amostras que é de 86, 81 metros. Entretanto, o efeito pepita elevado é o indicativo de que a Krigagem Ordinária não terá boa acurácia.

A Figura 24 mostra o estudo probabilístico dos valores preditos no semivariograma, comparado aos valores reais e as respectivas probabilidades observadas e teóricas da distribuição normal.

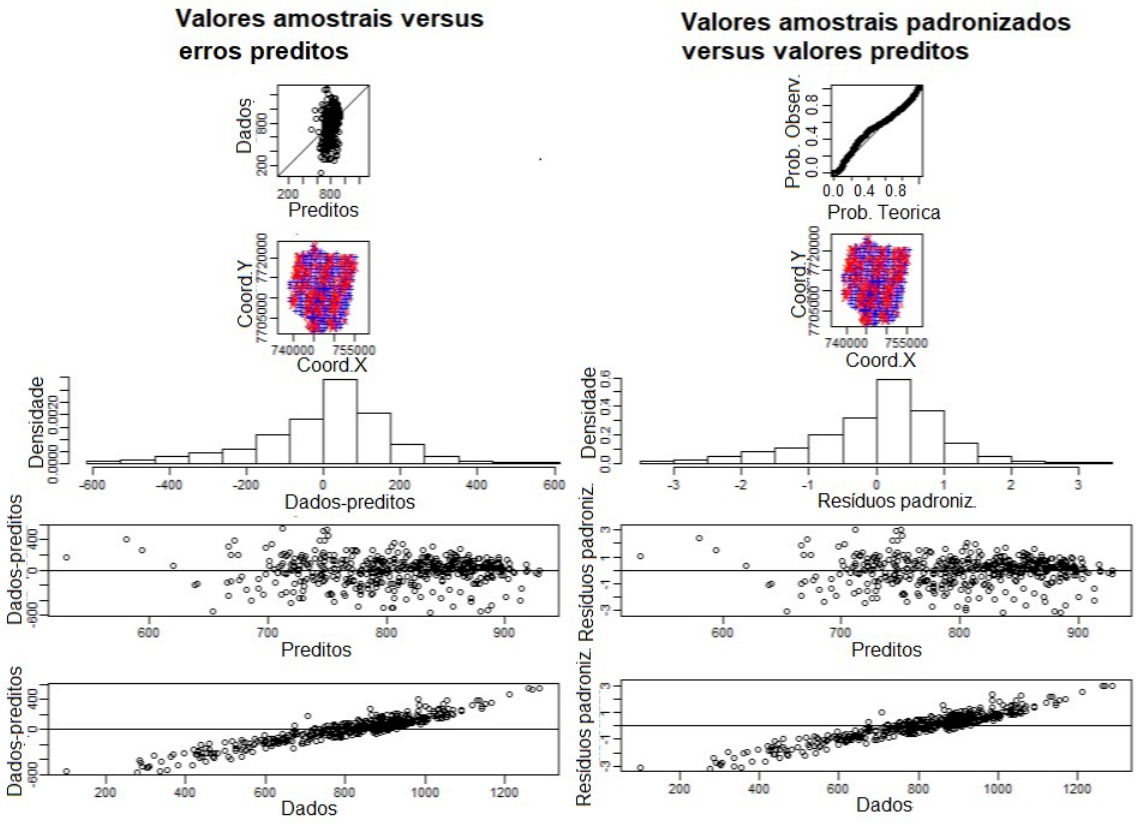

Figura 24: Probabilidades dos valores preditos para a amostra completa.

De acordo com a Figura 24, pode se observar que a reta de ajuste dos dados aos valores preditos não se ajustou bem e a densidade das diferenças entre os valores reais e os valores preditos, embora mais acentuada em torno de zero,

são mais frequentes à direita de zero, e os valores dos resíduos padronizados, que deveriam estar em torno da linha horizontal, apresentam-se com tendência negativa mais acentuada para os valores inferiores a 800, e com tendência positiva acentuada para valores superiores a 1000, indicando superestimação dos valores inferiores a 800 e subestimação dos valores superiores a 1000.

A Tabela 11 apresenta alguns parâmetros referentes ao semivariograma ajustado.

Tabela 11: Qualidade do ajuste da amostra completa

RSE	G.l.	R^2	R^2_{ajust}
166,4282	2 e 477	0,2063	0,2047

RSE=Erro Padrão Residual, G.l.=Grau de liberdade,
R^2_{ajust}= Percentual explicado pelo modelo ajustado.

A Tabela 11 mostra através do $R^2_{ajust.}$=0,204, que o modelo de semivariograma ajustado explica menos que 50% dos dados, não representando bem os dados.

O teste de normalidade Shapiro-Wilk, apresenta as estatísticas $W = 0,96754$ e $p-valor = 8.358 \times 10^{-9}$, não aceitando a hipótese nula (de normalidade), confirmando que os valores dos erros estimados pelo modelo de semivariograma não são normais.

A Figura 25 apresenta os mapas de Krigagem Ordinária e malha amostral referente a amostra completa.

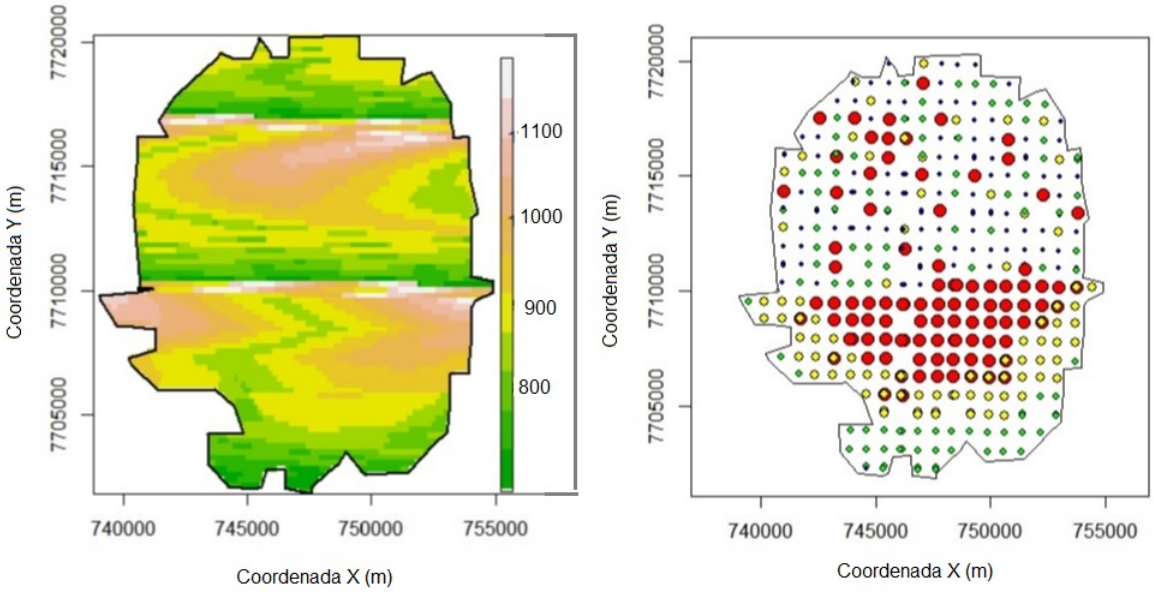

Figura 25: À esquerda: Krigagem Ordinária da amostra completa; à direita: Malha amostral completa.

Como se pode observar na Figura 25, o mapa de Krigagem Ordinária da amostra completa não apresenta boas estimativas, principalmente na subárea superior, onde as amostras apresentaram-se mais aleatórias.

Como se pode observar na Figura 25, a imagem referente a Krigagem Ordinária da amostra completa (à esquerda), confrontada com os valores amostrais (à direita), mostra que as cores no mapa de predições n ao são condizentes com as escalas de cores dos pontos amostrais. As dispersões entre os valores reais e preditos são calculados por meio de mapas de Variâncias de Krigagem. Os mapas das Variâncias de Krigagem dos três grupos estão representados na Figura 28.

Como se pode observar na Figura 26, o Grupo amostra total teve maior variabilidade nas estimativas por Krigagem Ordinária que os outros dois grupos resultante do corte da malha amostral.

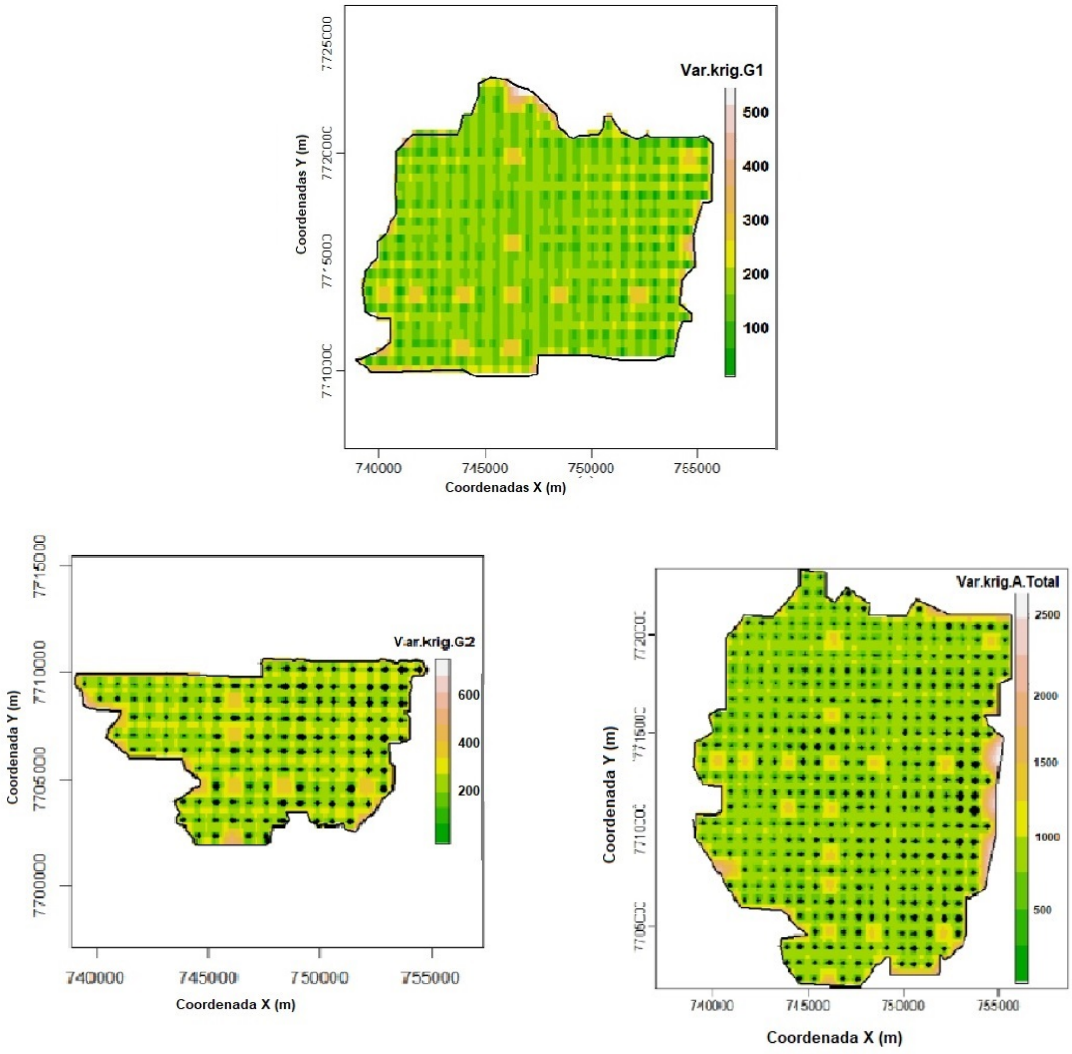

Figura 26: Variâncias de Krigagem do Grupo 1 (centro superior); Grupo 2 (à esquerda) e Amostra Total (à direita).

5.1.4 Comparação dos resultados dos grupos - Dados 1

As tabelas 12, 13 ,14 e 15 apresentam os resultados comparativos entre os ajustes realizados no Grupo 1, Grupo 2 e Amostra Total.

Tabela 12: Resultados comparativos dos parâmetros dos semivariogramas

Nome	Grupo1	Grupo2	Amostra total
Modelo	Cúbico	Gaussiano	Matérn
Alcance Prático	1.874,21	10.795,54	2.841,47
Contribuição	5.616,98	17.465,76	16.065,81
Efeito pepita	37.707,41	0	20.100,85

Tabela 13: Resultados comparativos da validação dos valores preditos versus valores reais

Nome	Grupo1	Grupo2	Amostra total
coef. linear	-3,99	- -	228,18
coef. angular	1,005	-	0,7102
$R^2_{ajust.}$	≈ 0	0,674	0,205
RSE	203,1	42,73	116,46
Norm.p-valor	0,6147 (0,874)	0,1732(0,17324)	NA($\approx$ 0,0)
Média.erro(WLS/OLS)	0,0162	0,0157055	1,9260
Desv.pad.(sd.error)	0,9844	0,9843	NaN ($\approx$ Inf.)
Média(erro.pad.)	0,000039	0,00009	NaN ($\approx$ Inf.)
AIC	4.245	4.245,5	6.322
Deviance Residual	12.910.000	12.908.136	14.930.000
Deviance nula	12.910.000	12.908.156	18.960.000

Deviance Residual: G1(gl.=313); G2(gl.=163); Amostra Total(gl.=477)

Deviance Nula: G1(gl.=314); G2(gl.=164); Amostra Total(gl.=478)

Os valores dentro dos parênteses() referem aos dados amostrais

Como se pode notar, nas Figuras 16, 21 e 25, os mapas de Krigagem Ordinária das Figuras 16 (Grupo 1) e 21 (Grupo 2) mostraram com clareza, por meio da sobreposição dos pontos amostrais, que as predições estão muito próximas dos verdadeiros valores. O mapa de Krigagem Ordinária da Figura 25, referente ao

Tabela 14: Resultados comparativos dos valores preditos por Krigagem Ordinária versus(valores reais) dos três grupos

Medida	Grupo 1	Grupo 2	Amostra total
Média	803,84(756,75)	911,52(906,12)	766,92(807,89)
Mediana	801,98(692,70)	903,1(899,89)	803,84(855,73)
Variância	2.814,9(41.725,85)	1.218,22(6.987,16)	24.282,53(34.826,174)
Desv.Pad.	53,05(204,27)	34,90 (83,59)	155,8(186,62)
Min.	592,35(101,33)	701,8(705,98)	592,36(101,33)
Max.	1164,38(1287,27)	1.166,95 (1.167,23)	1.164,38(1287,27)
Curt.	0,284(-0,106)	3,18 (3,39)	1,33(3,59)
Assim.	0,0034(0,338)	0,03(0,298)	0.23(-0,76)
C.V. %	6,6 (27)	0,15(9,2)	20 (23,1)

Os valores dentro dos parênteses() referem aos dados amostrais

Tabela 15: Resultados comparativos da qualidade de predição por Krigagem Ordinária nos três grupos

Nome	Grupo 1	Grupo 2	Amostra total
Variância de Krigagem	2.814,9	1.218,22	16.960,62
Erro padrão de estimação	2,99	2,7	7,1
Média dos erros de estimação	213,88	386,936	398,17
C.V. dos erros preditos%	24	9,9	32,71

grupo “Amostra Total”, mostrou que na maior parte da área, as predições não são condizentes com os valores amostrais, principalmente quando os valores amostrais são próximos às medidas de tendência central.

Comparando os mapas de Krigagem Ordinária das figuras 16, 21 e 25, pode se verificar que o mapa gerado na figura 25 (referente à Amostra Total), apresentou maior quantidade de erros na mudança de escala de cores, enquanto que os mapas 16 e 21 mostraram mais precisos, em relação à escala de valores da amostra

comparados à escala de cores do mapa de Krigagem Ordinária.

Os resultados da Krigagem Ordinária dos três mapas estão medidos em números que definem a qualidade, apresentados de forma comparativa nas Tabelas 12,13,14 e 15.

De acordo com tais resultados, pode-se verificar que os mapas feitos por parte da malha via método MPPs, tiveram melhores medidas de qualidade.

Na Tabela 12 está mostrando que os dois grupos ajustaram-se a dois modelos distintos de semivariograma. O Grupo 1 ajustou ao modelo de semivariograma Cúbico e o Grupo 2 ajustou-se ao modelo de semivariograma Gaussiano, enquanto o banco de dados completo, "Amostra total " , ajustou-se ao modelo Matérn.

Os grupos 1 e 2, apresentaram maior alcance prático que a Amostral Total. O Grupo 2 obteve maior contribuição e não apresentou efeito pepita. Os grupos 1 e 2 apresentaram normalidade de resíduos, enquanto o grupo de Amostra Total não apresentou normalidade.

De acordo com as Tabelas 13 e 14, os valores de Variância de Krigagem foram maiores no grupo "Amostra Total", bem como a média dos erros de estimação e C.V. dos erros preditos, mostrando que a Krigagem Ordinária do Grupo 2 seguida do Grupo 1, foram mais precisas.

A Tabela 13 mostra, por meio de indicadores de ajuste da regressão dos valores estimados em relação aos valores reais, que o Grupo 2, apesar de ter menor quantidade amostral,(163 amostras), estimou melhor que o Grupo 1(316 amostras) e também melhor que o grupo de Amostra Total (479 amostras).

Os Grupos 1 e 2, como era esperado pelo modelo de $MPPS$, apresentaram normalidade nos dados, enquanto o grupo de Amostra Total não apresentou normalidade.

Os resultados comparativos que mais chamam atenção são os indicadores de qualidade de predição da Krigagem Ordinária, pois estão diretamente ligados a qualidade e acurácia dos mapas gerados. Dentre estes resultados, verifica-se que o banco de dados da Amostra total teve, em média, maior erro padrão de estimação

$(7,1)$, enquanto os grupos 1 e 2 tiveram $2,99$ e $2,70$, respectivamente. Os Coeficientes de Variação dos valores preditos também foram inferiores nos grupos 2 e 1, respectivamente, em relação ao grupo "Amostra Total", mostrando que as estimativas foram mais concentradas em torno das médias estimadas. A variância de krigagem foi menor no grupo 2, seguida do grupo 1, enquanto que o grupo "Amostra Total" teve valores superiores a ambos os grupos. Outro ponto de destaque foi o Coeficiente de Variação dos erros preditos, em que o grupo "Amostra Total" foi mais elevado, $32,71\%$ enquanto os grupos 1 e 2 tiveram $9,9\%$ e 24%, respectivamente.

Os valores de Varância de Krigagem, Erro Padrão de estimação, Média dos erros de estimação e C.V. dos erros preditos, Tabela 13, que foram maiores no grupo " Amostra Total" comparados ao Grupo 1 e ao Grupo 2, mostraram que a Krigagem Ordinária no Grupo 1 e no Grupo 2 foram mais precisas, além de outras medidas importantes como a Deviance Residual e o Critério de Informação Akaike (AIC), os quais o grupo "Amostra Total" obteve maiores valores, indicando pior qualidade.

A tabela 14 mostra, por meio de valores comparativos da predição por Krigagem Ordinária, em relação aos valores amostrais, que os resultados dos valores preditos para o Grupo 2 e para o Grupo 1, ficaram mais próximos dos valores amostrais que o grupo "Amostra Total". O Grupo 2, superou a qualidade do Grupo 1, mas apesar disso, o Grupo 1 mostrou-se melhor que o grupo "Amostra Total", ficando mais próximo das estatśisticas amostrais em relação ao valor máximo e à assimetria.

Nos resultados que avaliaram a qualidade de predição por Krigagem Ordinária nos três grupos, Tabela 15, verificou-se que o Grupo 2, superou em quase todos os quesitos avaliados, ficando inferior ao Grupo 1 apenas na Média dos Erros de Estimação, enquanto o grupo "Amostra Total" ficou pior em todos os quesitos avaliados.

5.2 Resultados obtidos pelo modelo de dois pontos de Corte, k e k_1 - Dados 2

Utilizando o banco de dados de batimetria (Dados2), após observar possível presença de tripla estacionaridade da média, mostrado na Figura 11, e assumida a possibilidade da existência de uma partição com três partes, ou seja, a possível existência de dois pontos de mudanças na média das semivariancias locais, K e k_1, em função da Médias das Distâncias entre Vizinhos (SumMediaDvi), foi aplicada o modelo, apresentado no capítulo 4, nas seções 4.3,4.4 e 4.4.1, descrito pela expressão (64).

Para obter os resultados, o banco de dados(Dados2) foi ordenado em relação a covariável SumMediaDvi, a fim de que os agrupamentos mantivessem em grupos contíguos espacialmente.

Com o uso do software WINBUGS, assumindo como valores iniciais, $k = 20$ e $k1 = 600$, foram gerados os seguintes resultados para os pontos de mudança:

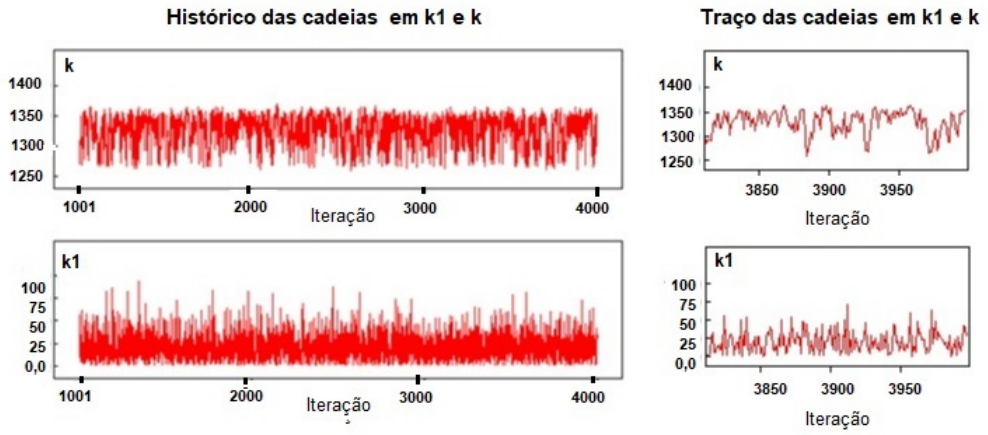

Figura 27: Histórico da cadeia para os locais dos pontos de mudança k e k1 (à esquerda); Traço da cadeia para os locais dos pontos de mudança k e k1 (à direita).

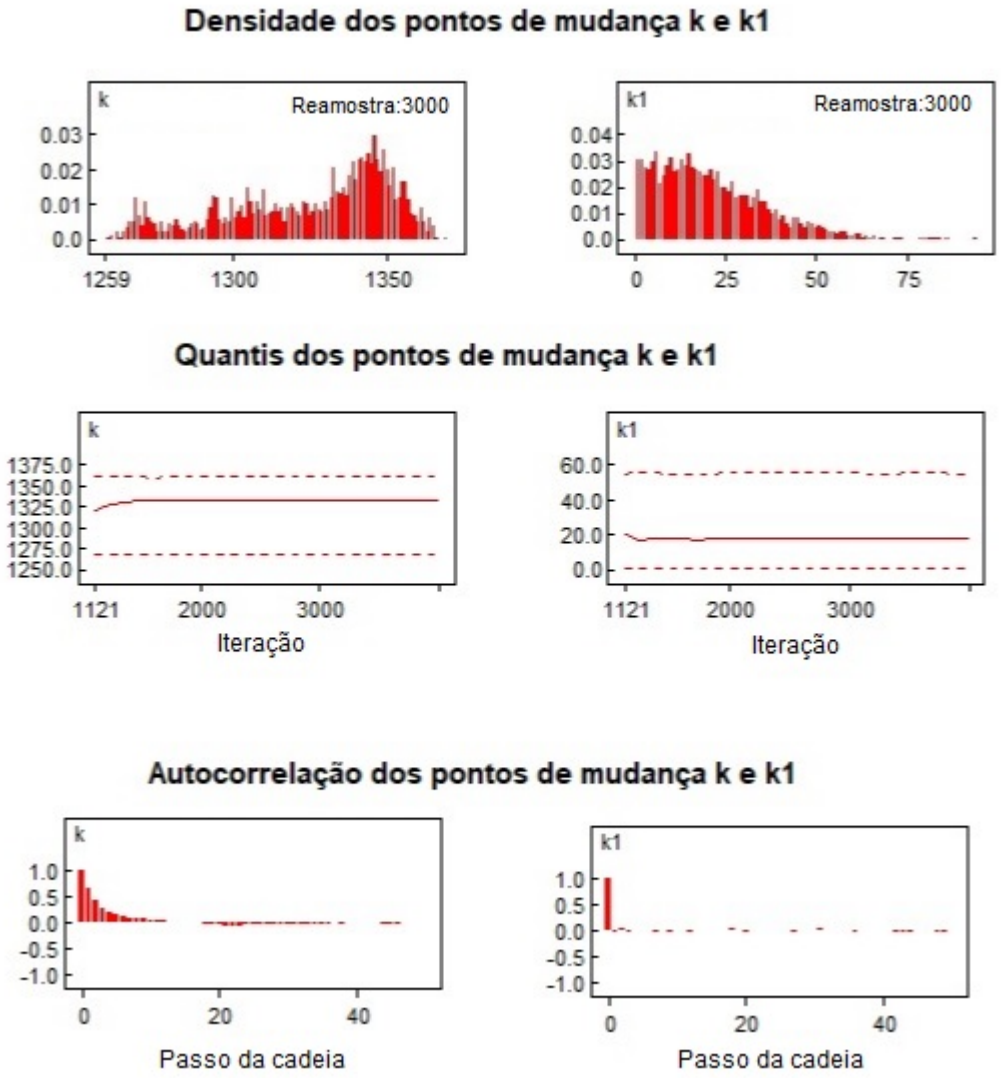

Figura 28: Densidade, Intevalo de Credibilidade e Autocorrelação das cadeias dos pontos de mudança k e k1.

Os valores das estatísticas referentes à análise do método bayesiano MPPs, apresentadas nas Figuras 27 e 28 estão descritos na Tabela 16.

Tabela 16: Estatísticas dos locais dos pontos de mudança na média

Nó	Média	Desv.Pad.	Erro MC	2,5%	Mediana	97,5%	Burn-in	Ream.
k	1.325,00	26,13	1,086	1.269	1333	1361	1001	3000
k1	20,99	14,86	0,271	1,0	18,0	56,0	1001	3000

$Thin^*$: 50

*$Thin$ é o espaçamento entre as amostras simuladas a fim de evitar a autocorrelação da cadeia.

De acordo com a Tabela 16, o Intervalo de Credibilidade do ponto de mudança na média em $k1$ está entre 1 e 56 com 95% de probabilidade. Logo, embora a média seja a posição 21°, foi tomado o 56° elemento para aplicar o corte, por ser o ponto final da distribuição do ponto de mudança, a fim de garantir uma quantidade mínima de amostras, suficientes para se aplicar geoestatística.

Para o segundo corte foi adotada a média da distribuição do ponto de mudança porque a distribuição de densidade neste ponto, ficou bem definida no 1.325° elemento, próximo ao limite superior do Intervalo de Credibilidade.

Os grupos ficaram assim divididos:

- Grupo 1: do 1° ao 56° elemento, $n = 56$ amostras.

- Grupo 2: do 57° a 1.325° elemento, $n = 1269$ amostras.

- Grupo 3: do 1.326° a 1.411° elemento, $n = 86$ amostras.

Em relação aos tamanhos amostrais mínimos, tem-se satisfeita uma das condições para se aplicar geoestatística, que é garantir uma quantidade suficiente de pares de pontos.

A Tabela 17 mostra os resultados exploratórios dos três grupos estabelecidos a partir dos pontos de corte adotados.

Tabela 17: Estatísticas dos grupos G1, G2 e G3, definidos a partir dos pontos de mudança na média.

Medidas	G1	G2	G3
Distância mínima	1,01 m	1,002 m	1,00245 m
Distância máxima	15,82 m	120,96 m	154,55 m
Média	-4,78 m	-4,27 m	-3,324 m
Valor mínimo	-5,37 m	-5,54 m	-4,75 m
Valor máximo	-4,53 m	0 m	-1,77 m
Desvio padrão	0,2 m	0,92 m	0,73 m
Variância	0,04 m	0,85 m	0,53 m
Coef. Variação%	4,2	22	21,8
Assimetria	-1,21	2,01	-0,09
Curtose	4,09	9,34	1,71

m = metros.

5.2.1 Análise espacial do Grupo G1 - Dados 2

Os resultados da análise exploratória espacial do grupo G1, referente ao banco de dados batimétricos, estão apresentados na figura 29:

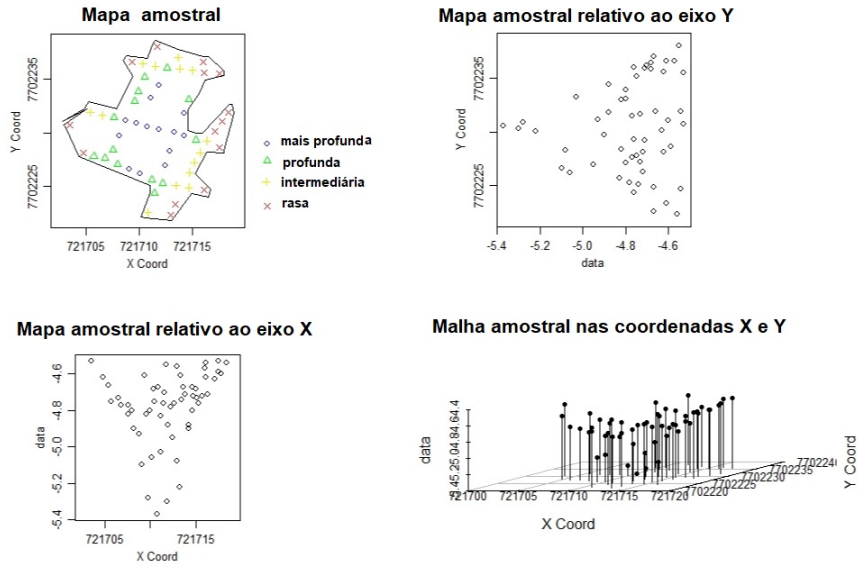

Figura 29: Representação gráfica dos pontos amostrais relativos às coordenadas X e Y.

Os parâmetros do semivariograma ajustado ao grupo G1 estão apresentados na Tabela 18 e na Figura 30, o qual foi ajustado o modelo Cúbico.

Tabela 18: Parâmetros do semivariograma ajustado ao grupo G1.

Medidas	Valores
Método de ajuste	OLS(ordinary least squares)
Alcance prático	9,69 m
Contribuição	0,059
Efeito-pepita	0,004
Modelo ajustado	cúbico

m = metros

Os resultados apresentados na Tabela 18 e na Figura 32 mostraram que embora o alcance prático não tenha obtido um valor elevado (9, 69 metros), comparado a distância mínima (1, 006 metros) e a distância máxima (15, 824 metros) entre dois pontos, é possível considerar que, em média, cada ponto foi estimado por mais da metade do total de pontos do grupo.

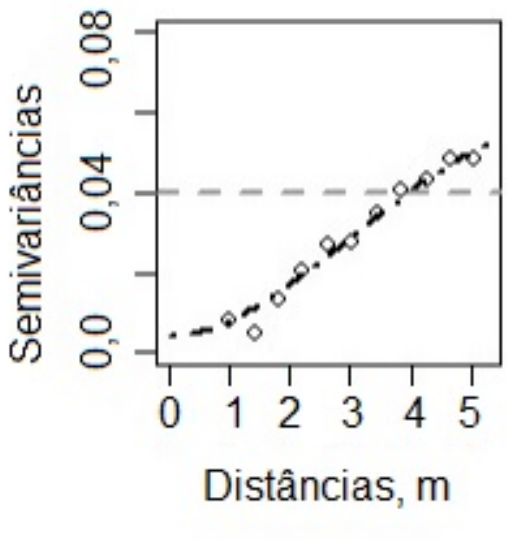

Figura 30: Semivariograma ajustado por OLS e WLS referente a G1 - dados batimétricos.

A Figura 31 está mostrando algumas medidas importantes dos dados amostrais comparados aos valores preditos, tais como, densidade dos erros e densidade dos erros padronizados, regressão dos valores amostrais e preditos, probabilidades teórica e observada, medidas do erros e dos erros padronizados em função dos valores preditos e em função dos valores amostrais.

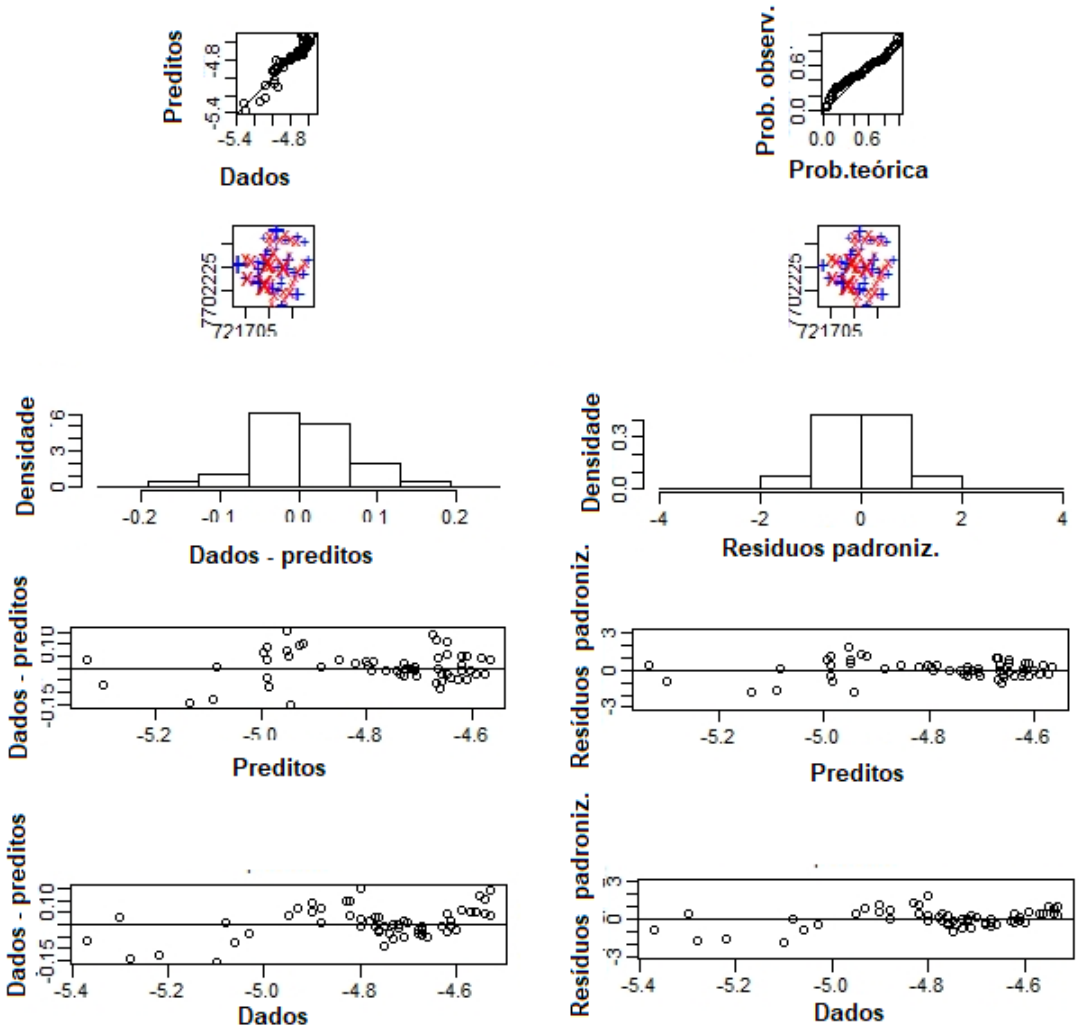

Figura 31: Representação gráfica dos valores preditos × valores amostrais do grupo G1 - Dados batimétricos

Na Figura 31, o gráfico da densidade da diferença entre valores amostrais e preditos, observa-se que as maiores densidades de probabilidades são referentes às menores diferenças. Os resíduos e os residuos padronizados em função dos dados e dos valores preditos mostram os pontos concentrados em torno do zero quando os valores preditos referem-se às maiores profundidades, em módulo. A reta de ajuste entre os dados e os valores estimados, também mostrou maior concentrados em torno

da linha diagonal, indicando muita proximidade entre os valores estimados e os valores amostrais.

A Figura 32 mostra o mapa de Krigagem Ordinária e a Figura 33 mostra o mapa de Variância de krigagem do Grupo G1, referente aos dados batimétricos.

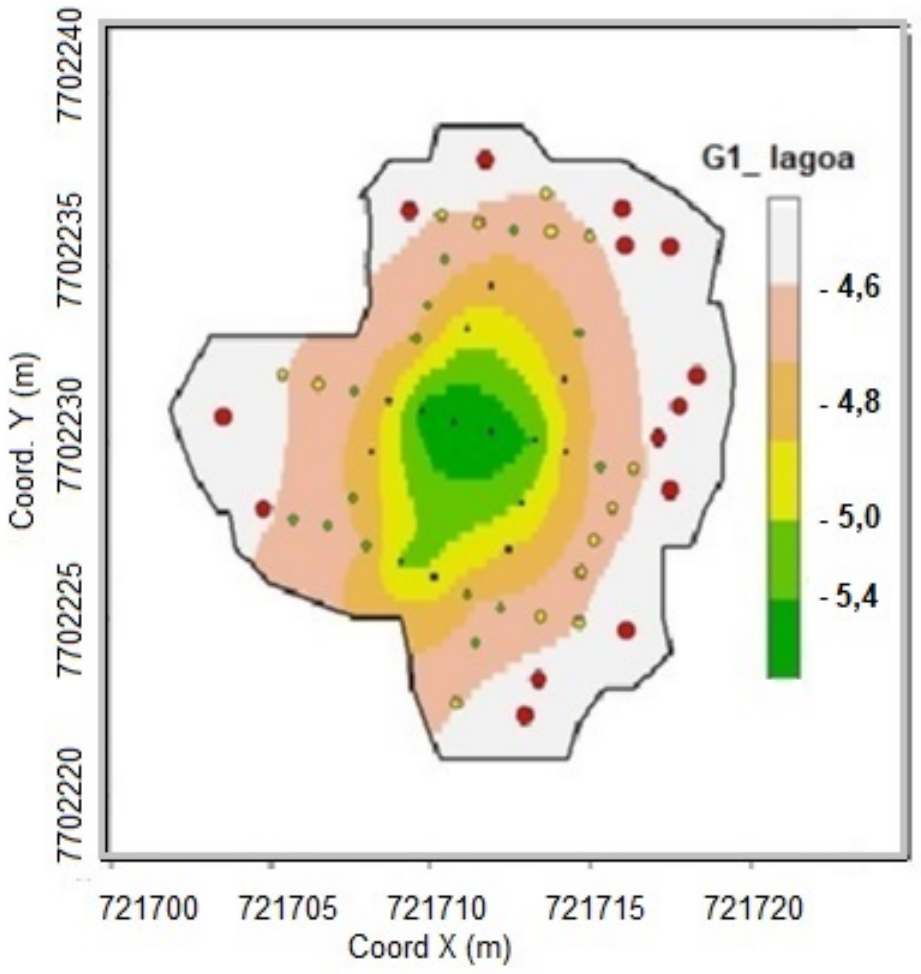

Figura 32: Mapa de Krigagem Ordinária do Grupo G1 - Dados batimétricos.

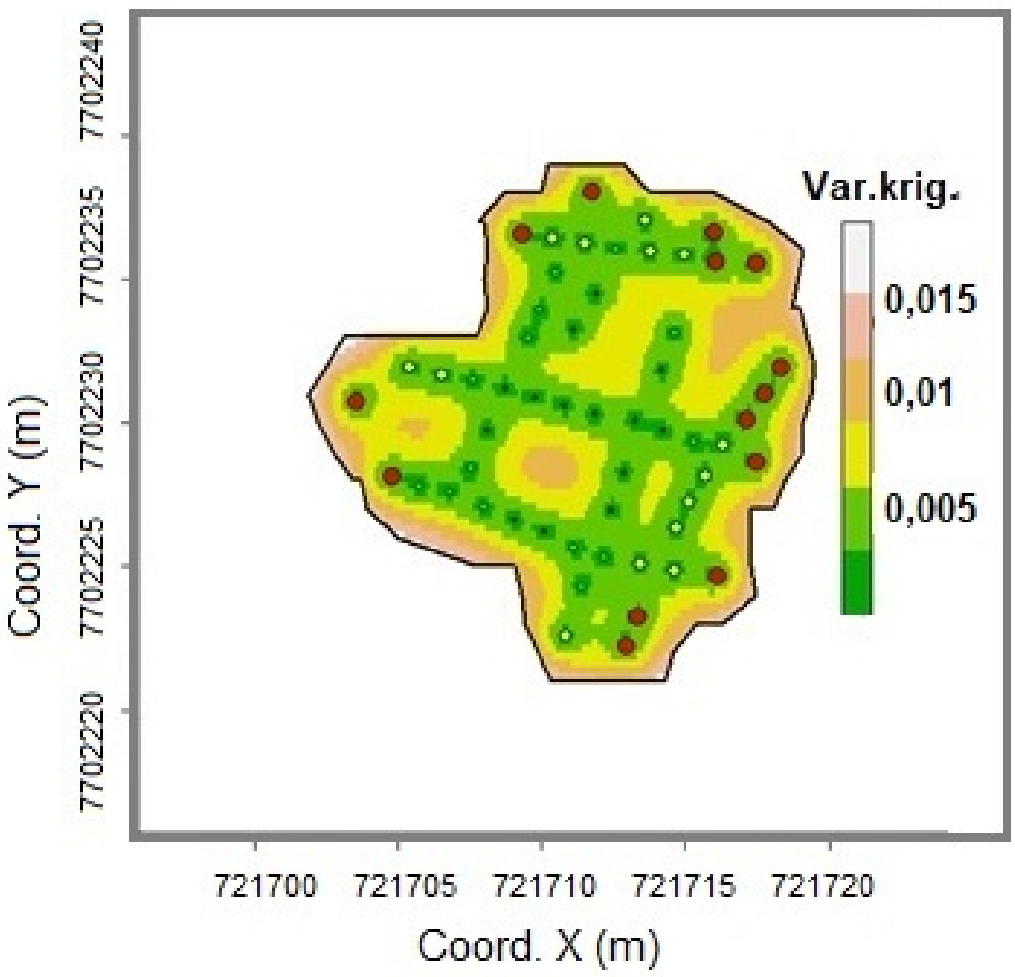

Figura 33: Mapa de Variância de Krigagem do Grupo G1- Dados batimétricos.

Como se pode notar no mapa da Figuras 32, as estimativas das maiores profundidades variaram de $-5,4$ metros a $-4,6$ metros, muito próximas dos valores amostrais. Os pontos vermelho são, em módulo, representativos das menores profundidades e os pontos verde escuros são em, módulo, representativos das maiores profundidades encontrados na amostra do grupo G1 e pode-se notar que os valores ão condizentes com as estimativas por Krigagem Ordinária.

O mapa de variâncias de krigagem (Figura 33), mostra que as predições nos locais mais próximos aos pontos amostrais, apresentam as menores variâncias

de krigagem (inferiores a 0,005) e nos locais próximos às extremidades da malha, apresentam as maiores variâncias de krigagem (em torno de 0,015). Ou seja, as extremidades da malha são os locais onde as predições por Krigagem Ordinária mais erraram, isto se justifica porque as estimativas nestes locais, são feitas com menor quantidade de vizinhos.

5.2.2 Análise espacial do grupo G2 - Dados batimétricos

Os resultados da análise exploratória espacial do Grupo 2 referente ao banco de dados batimétrico, estão apresentados na figura 36.

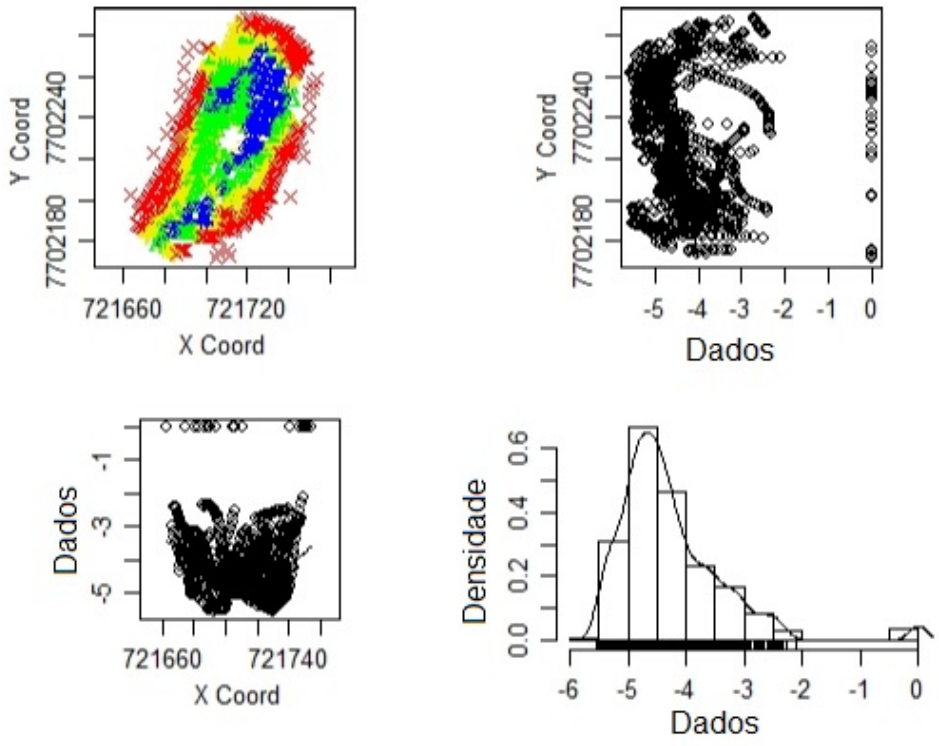

Figura 34: Malha de pontos (canto esquerdo superior); projeção de pontos sobre o eixo X(canto esquerdo inferior); projeção de pontos sobre o eixo Y (canto direito superior) e densidade amostral (canto direito inferior).

Como mostrado na Figura 34, a amostra apresenta uma grande concentração de dados de profundidades elevadas, em módulo, ao longo do eixo Y e com

alguns valores de profundidade nula em torno das margens da lagoa, como mostra o gráfico do no canto superior esquerdo (em formato de X vermelho maiores). Na separação do grupo G2, prevaleceu uma certa continuidade dos valores, que pode ser percebida pela distribuição das cores da Figura 34 (canto esquerdo superior), em que nota-se um "buraco"no centro da figura, que é o local onde foram retiradas as amostras do grupo G2.

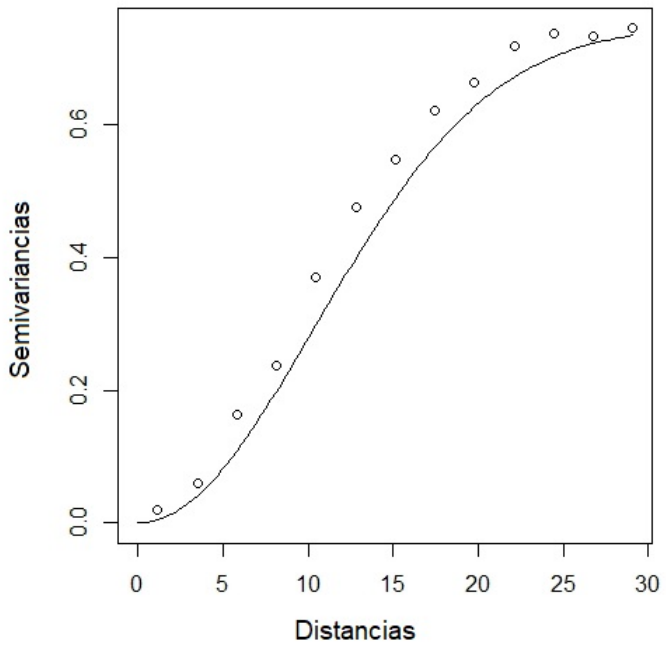

Figura 35: Semivariograma Gaussiano ajustado ao grupo G2.

Os parâmetros do semivariograma ajustado ao grupo G2, estão apresentados na Figura 35 e na Tabela 19, mostrando um excelente ajuste ao modelo de semivariograma Gaussiano. O alcance prático mostrou ser razoável, $22,24$, a contribuição foi de $0,75$, próximo ao valor da variância dos dados amostrais $0,85$ e não houve efeito pepita, indicando que o modelo é adequado aos dados.

Como se pode notar, na Tabela 19, o grupo G2 ajustou bem ao modelo de semivariograma Gaussiano, sem efeito pepita e com alcance prático igual a $22,24$ metros, indicando que cada ponto, tem em média, 20 vizinhos de estimação.

Tabela 19: Parâmetros do semivariograma ajustado ao grupo G2.

Parâmetros	Valores
Método de ajuste	OLS
Alcance prático	22,24 m
Contribuição	0,75
Efeito-pepita	0,0
Modelo ajustado	Gaussiano

m= metros.

A Figura 36 apresenta as análises de densidade e regressão entre os valores amostrais e preditos, bem como os erros de predições e resídulos padronizados.

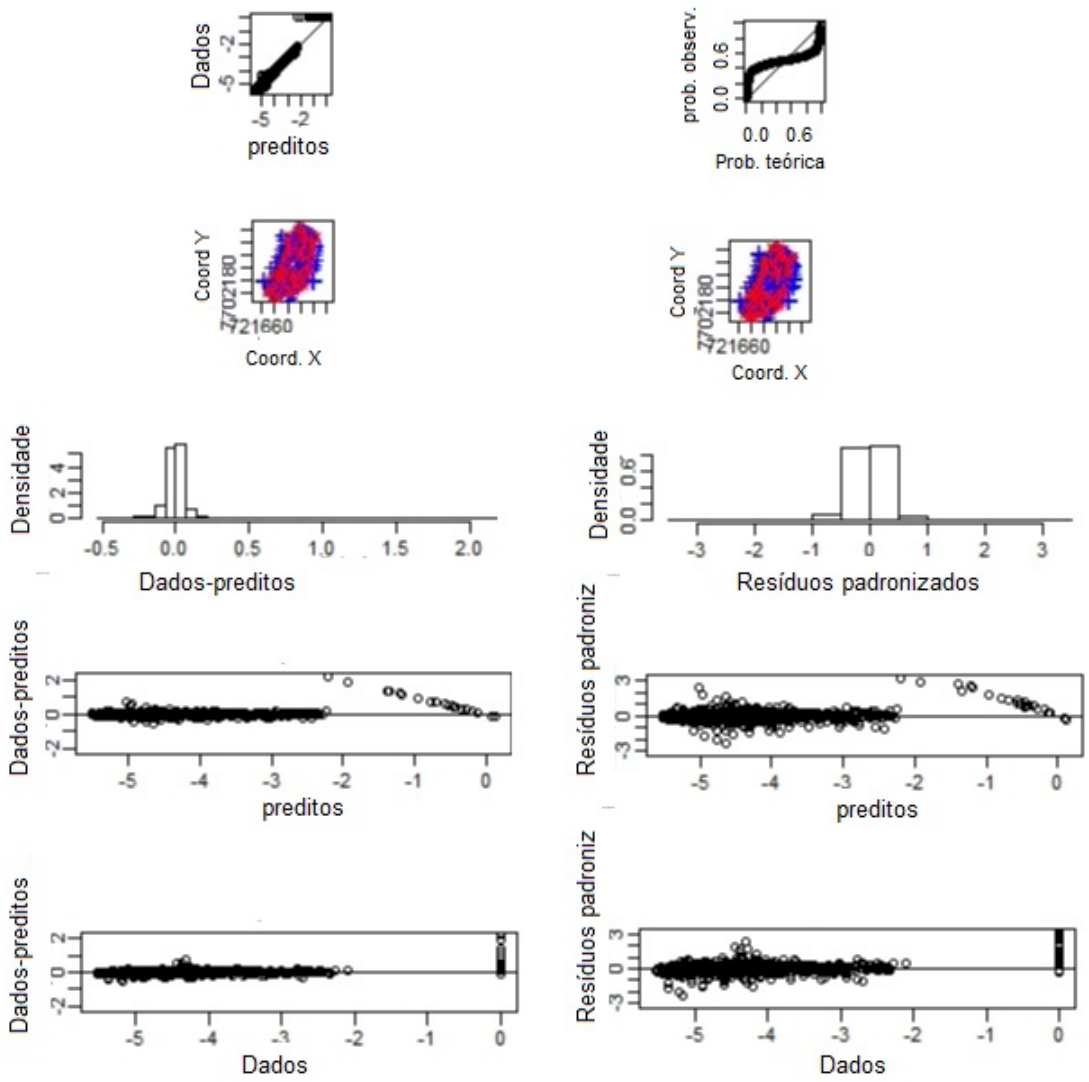

Figura 36: Gráficos de densidade e probabilidades dos valores preditos e dos erros de predições do grupo G2 - Dados batimétricos.

De acordo com a Figura 36, as densidades dos erros entre os valores preditos e os valores reais e as densidades dos resíduos padronizados concentraram-se em torno de 0. A regressão linear entre os dados e os valores preditos, mostraram que os valores preditos e os resíduos padronizados ajustaram bem a reta de regressão dos dados amostrais com alguns desvios notáveis nas predições das menores profun-

didades (em módulo).

A Figura 37 apresenta os mapas de Krigagem Ordinária (à esquerda) e mostra por meio de figura comparativa, o mapa de Krigagem Ordinŕia com sobreposição dos pontos amostrais (à direita).

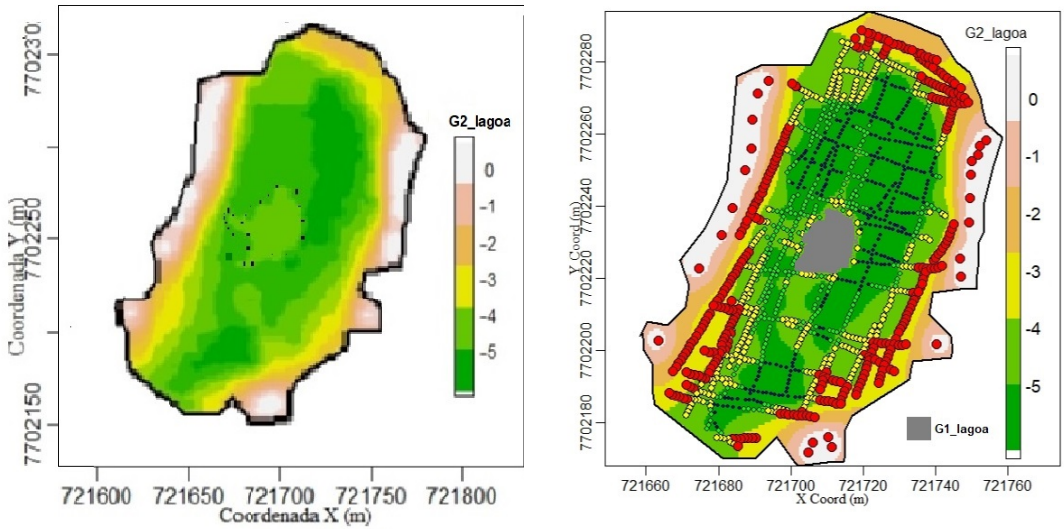

Figura 37: Krigagem Ordinária do grupo G2 (à esquerda); Krigagem Ordinária do grupo G2 com sobreposição dos pontos amostrais (à direita).

A Figura 37 (à direita) mostra que a região em cinza, pertencente ao grupo G1 e nas demais partes da figura, as predições feitas por Krigagem Ordinária, divididas em 6 cores, as quais, as maiores profundidades (em módulo) estão representadas em tom verde escuro a verde claro, as profundidades (em módulo) intermediárias estão representadas em amarelo e laranja e as menores profundidades (em módulo), estão representadas em salmon e branco. De acordo com estas escalas de cores, pode ser observado que os pontos amostrais (grandes, em vermelho) indicaram as menores profundidades (em módulo) correspondentes às predições feitas por Krigagem Ordinária, similarmente as demais escalas também apresentaram as escalas de cores estimadas de acordo com os pontos amostrais. Percebe-se pelo mapa de Krigagem Ordinária, que as estimativas são correspondentes, adequadamente, aos valores amostrais, indicando boa qualidade do mapa.

A Figura 38 apresenta os mapas das estimativas de Variâncias da krigagem distribuídas na região amostral.

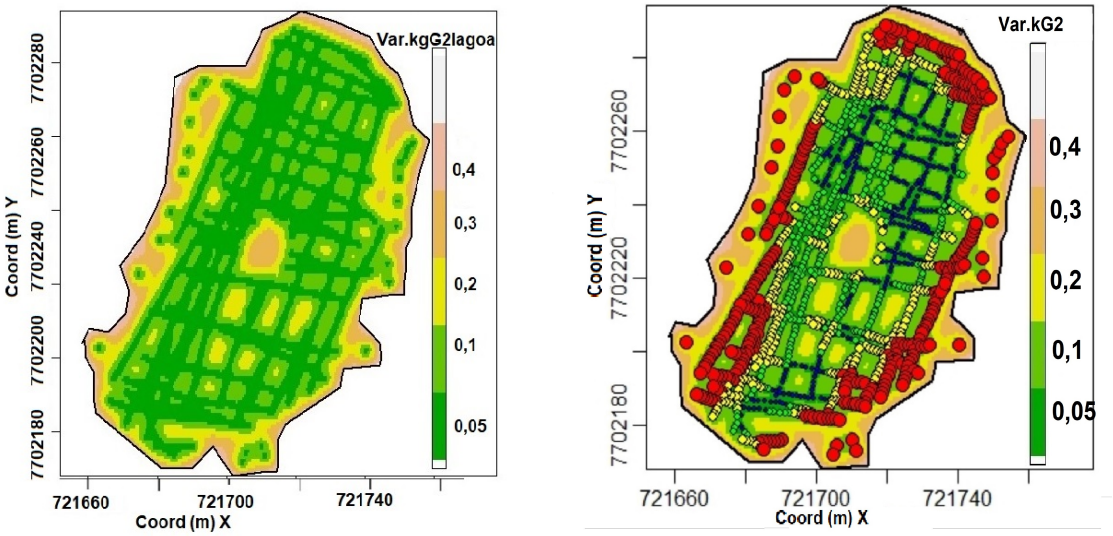

Figura 38: Variância de krigagem do grupo G2 (à esquerda); Variância de krigagem com sobreposição dos pontos amostrais do grupo G2 (à direita).

De acordo com a Figura 40 (à esquerda) é apresentado o mapa das Variâncias de krigagem, em que se observa que as regiões nas quais existem menos pontos amostrais e nas extremidades da malha, são as regiões da malha que possuem as maiores estimativas de variância e também mais ao centro (devido a falta de amostra). Os locais onde existem mais pontos amostrais, possuem as menores estimativas de variância e estas coincidem com os locais de maiores profundidades (em módulo). Em suma, em todas as regiões do *grid* as variâncias de Krigagem tiveram as menores estimativas, variando de $0,05$ a $0,4$.

5.2.3 Análise espacial do grupo G3 - Dados batimétricos

Os resultados da análise exploratória espacial do Grupo 3 referentes ao banco de dados batimétricos, estão apresentados na Figura 39.

Como mostrado na Figura 39, a amostra apresenta uma grande con-

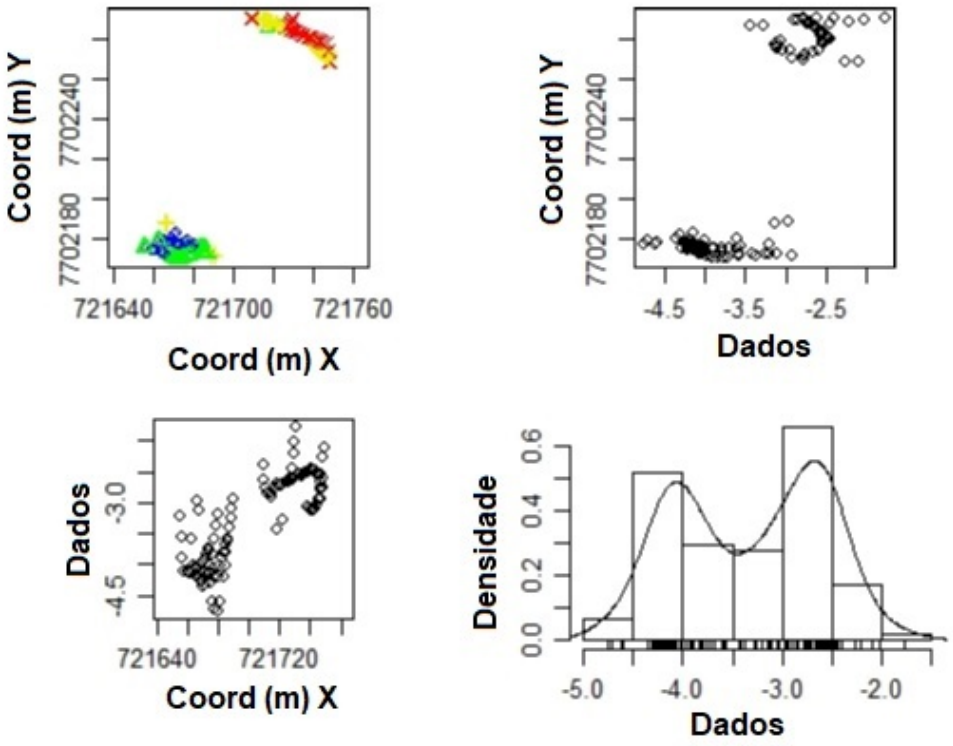

Figura 39: Malha de pontos (canto esquerdo superior); projeção de pontos sobre o eixo X(canto esquerdo inferior); projeção de pontos sobre o eixo Y (canto direito superior) e densidade amostral (canto direito inferior).

centração de dados de profundidades nulas, baixas e intermediárias, a partir da coordenada X (721700) deslocando para a direita e eixo Y, em torno da margem esquerda inferior da lagoa estão os maiores valores de profundidades, em torno de $3,0$ e $4,7$ metros (em módulo) e no canto direito superior estão as menores profundidades, em módulo. Na separação do grupo G2 com o grupo G3, ao se retirar o grupo G2, prevaleceu uma certa descontinuidade no grupo G3 em relação ao espaço amostral. Apesar disso, os valores foram agrupados por possuírem características similares de semivariancias e portanto, ajustaram a um mesmo semivariograma.

A distribuição das cores da Figura 40 (canto esquerdo inferior e canto direito superior) estão separadas por um grande "intervalo"vazio entre as amostras, apesar disso, foi possível um bom ajuste do semivariograma.

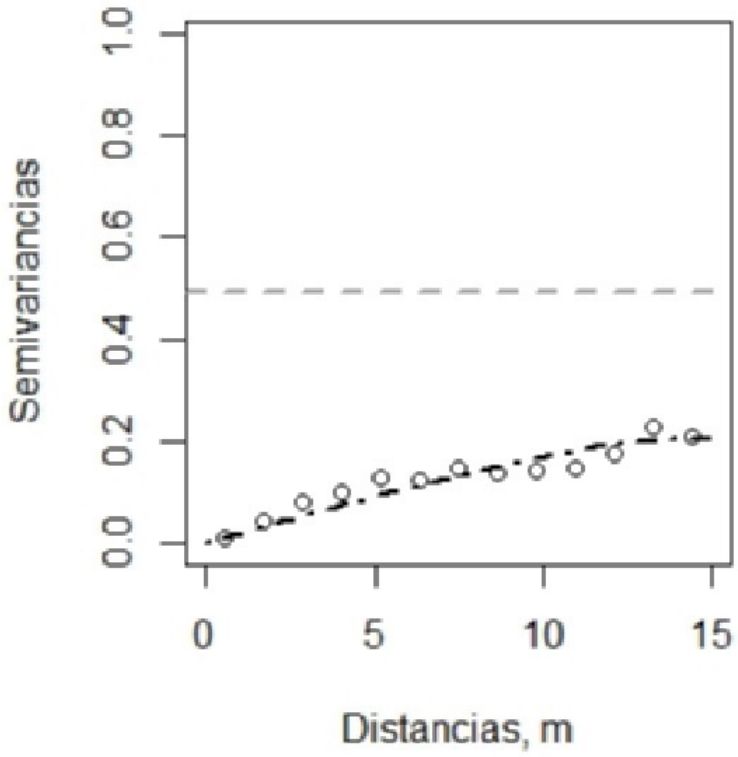

Figura 40: Semivariograma Circular ajustado ao grupo G3

Os parâmetros do semivariograma ajustado ao grupo G3 estão apresentados na Figura 40 e na Tabela 20, mostrando um excelente ajuste ao modelo Circular, com alcance prático de $23,58$ metros e sem efeito pepita, indicando que o modelo é adequado aos dados.

Tabela 20: Parâmetros do semivariograma Circular ajustado ao grupo G3

Parâmetros	Valores
Método de ajuste	OLS
Alcance prático	23,58 m
Contribuição	0,255
Efeito-pepita	0,0
Modelo ajustado	Circular

m= metros

A Figura 41 apresenta as análises de densidade e regressão entre valores amostrais e preditos, bem como erros de predições e resíduos padronizados. De acordo com a Figura 41, as densidades dos erros entre os valores preditos e os valores amostrais e as densidades dos resíduos padronizados concentraram-se em torno de 0. A regressão linear entre os dados observados e os valores preditos mostraram que os valores preditos e os valores amostrais ajustaram bem a reta, além disso, a distribuição de probabilidade teórica aproximou muito de uma reta, em relação à probabilidade observada.

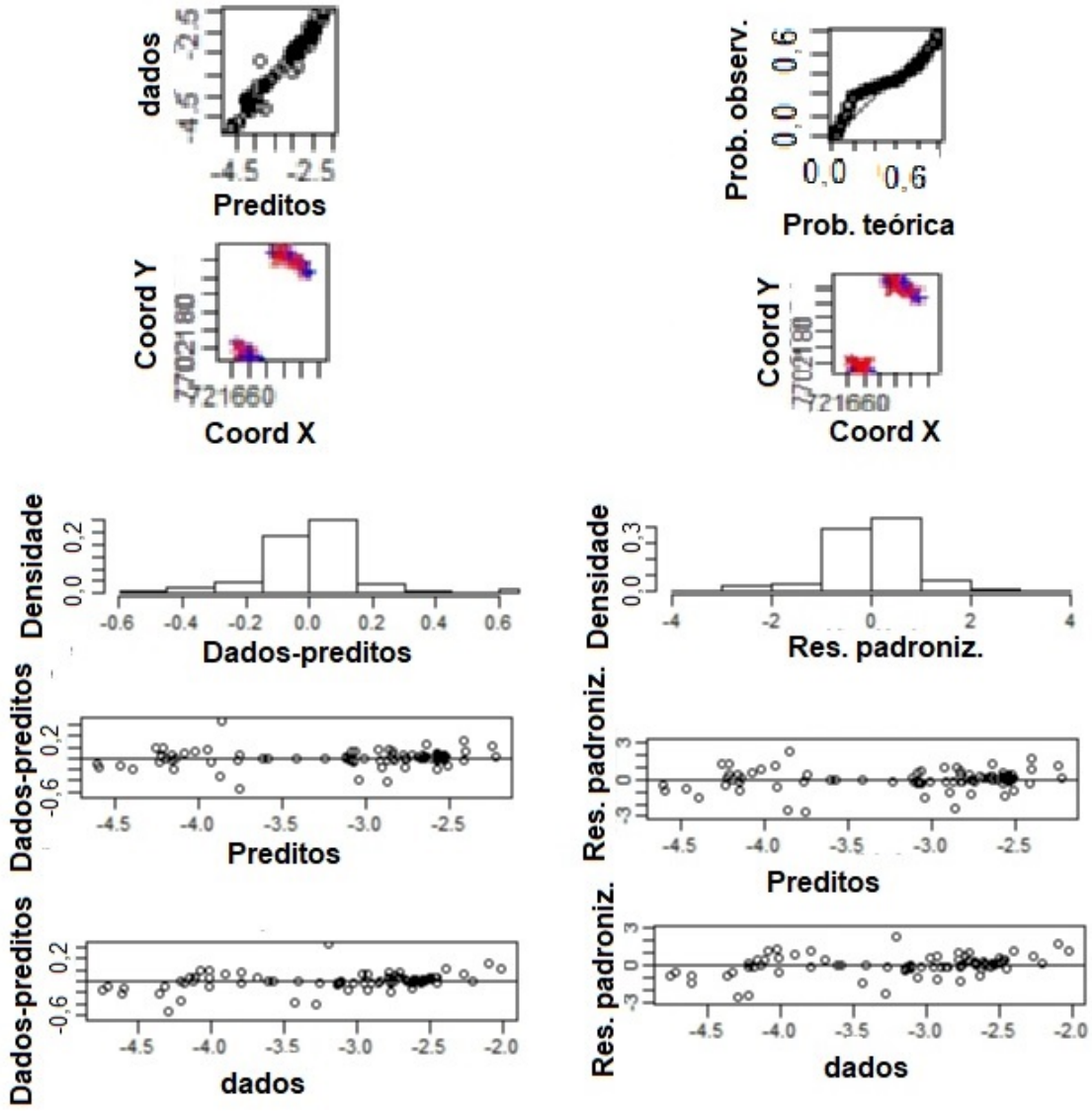

Figura 41: Gráficos de densidade e probabilidades dos valores preditos e dos erros do grupo G3 - dados batimétricos

A Figura 42 apresenta os mapas de Krigagem Ordinária do grupo G3 e a sobreposição da amostra.

Na Figura 42 à esquerda, mostra o mapa de Krigagem Ordinária do grupo G3, em que pode se observar, no canto esquerdo inferior a concentração da região dos maiores valores (em módulo), referentes às maiores profundidades do grupo. No canto direito superior, estão as estimativas dos pontos de menores pro-

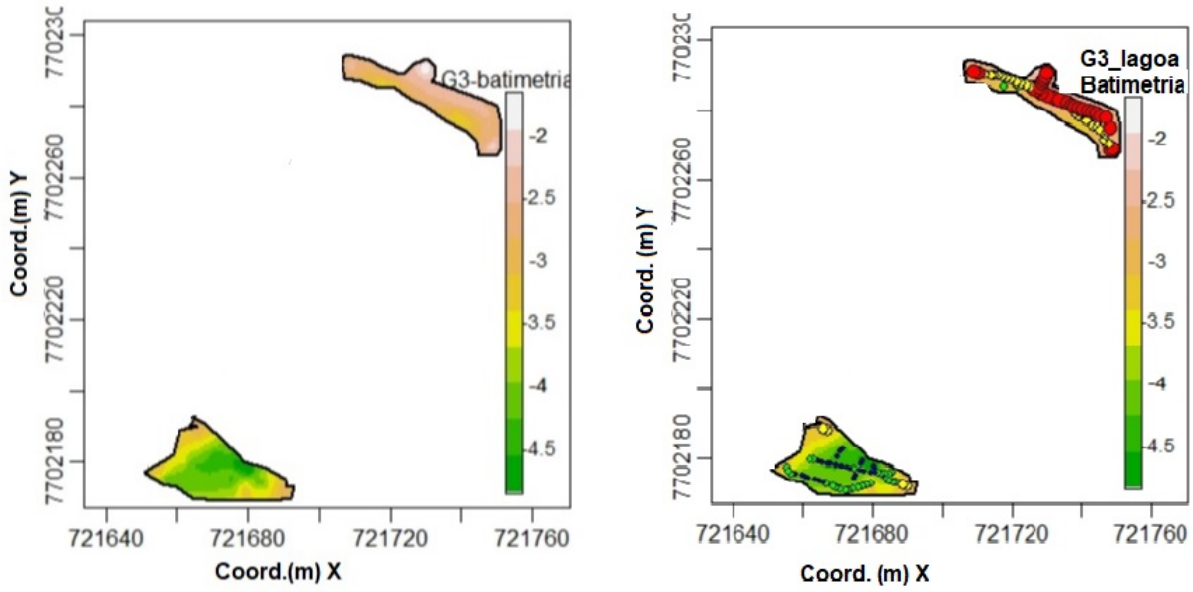

Figura 42: Krigagem Ordinária do grupo G3 (à esquerda); Krigagem Ordinária do grupo G3 com sobreposição dos pontos amostrais(à direita).

fundidades, e o intervalo em branco entre as duas regiões, refere-se a à sub-área onde ficam localizados os grupos G1 e G2. Apesar de haver um intervalo sem amostras no espaço, este fato não afetou o ajuste do semivariograma, ficando bem ajustada as duas partes do grupo G3, ao mesmo modelo de semivariograma e parâmetros. As duas partes da figura estão mostrando que as predições por Krigagem Ordinária se dividiram em 6 cores, em que as maiores profundidades (em módulo) estão em tom verde escuro, verde claro e amarelo e as menores profundidades (em módulo) estão representadas em laranja, salmon e branco. De acordo com esta escala de cores, pode-se verificar que os pontos amostrais (grandes e em tons vermelho), indicaram as menores profundidades e estão correspondendo às predições feitas por Krigagem Ordinária, similarmente às demais escalas, que também apresentaram os pontos amostrais de acordo com as escalas de cores estimadas. Percebe-se pelo mapa de Krigagem Ordinária, que as estimativas estão correspondendo adequadamente aos valores da amostra, indicando boa qualidade do mapa.

A Figura 43 apresenta os mapas das estimativas de Variâncias da krigagem na malha amostral. Na Figura 43 estão representados dois mapas de Variância

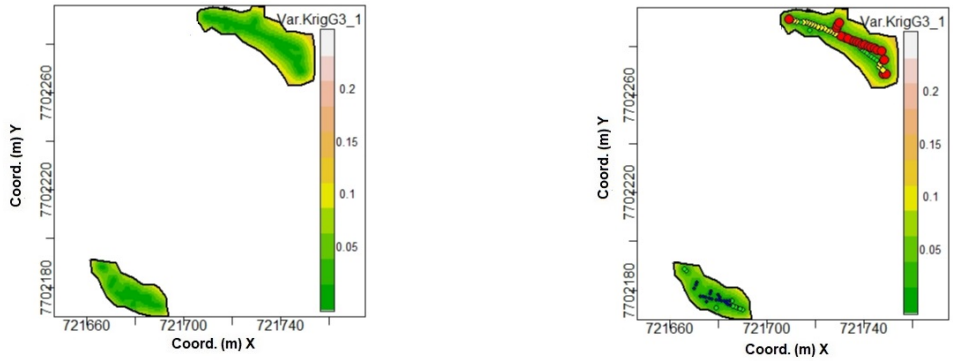

Figura 43: Variância de krigagem do grupo G3 (à esquerda); Variância de krigagem com sobreposição dos pontos amostrais do grupo G3 (à direita).

da krigagem do grupo G3, em que o mapa à esquerda apresenta apenas as estimativas de variâncias de Krigagem e o mapa à direita as variâncias de krigagem com sobreposição dos valores observados. Pode se perceber que nos locais onde foram contém as amostras, as estimativas das variâncias de krigagem foram muito baixas, inferiores a $0,05$, enquanto nos demais locais, especialmente em algumas partes das extremidades da figura, os valores de variâncias de krigagem foram maiores. Este comportamento era esperado porque esta região da malha é penalizada pela redução da vizinhança de estimação, porém em todas as duas partes da malha amostral as variâncias de krigagem foram pequenas indicando boa qualidade do mapa.

5.2.4 Análise espacial da Amostra Completa - Dados Batimétricos

Os resultados da análise exploratória espacial do Conjunto amostral "Amostra Completa"referente aos dados batimétricos, estão apresentados na Figura 44:

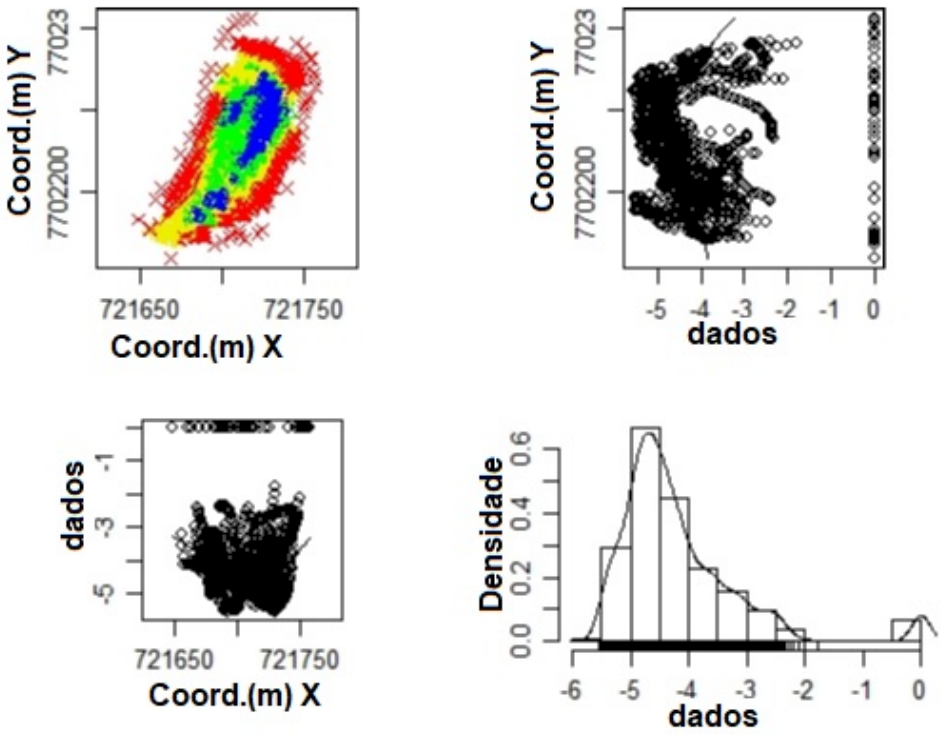

Figura 44: Malha de pontos (canto esquerdo superior); projeção de pontos sobre o eixo X (canto esquerdo inferior); projeção de pontos sobre o eixo Y (canto direito superior) e densidade amostral (canto direito inferior).

A Figura 44, está mostrando uma grande concentração de dados de profundidades elevadas (em módulo), ao longo do eixo X e do eixo Y e alguns alguns valores de profundidade nula em torno das margens da lagoa (X em vermelho). Observa-se a continuidade dos valores que podem ser notado pela distribuição das cores (canto esquerdo superior) em que as maiores profundidades estão concentradas no centro ao longo do eixo Y.

A Figura 45, apresenta o modelo de semivariograma ajustado aos dados, pelos métodos OLS e WLS.

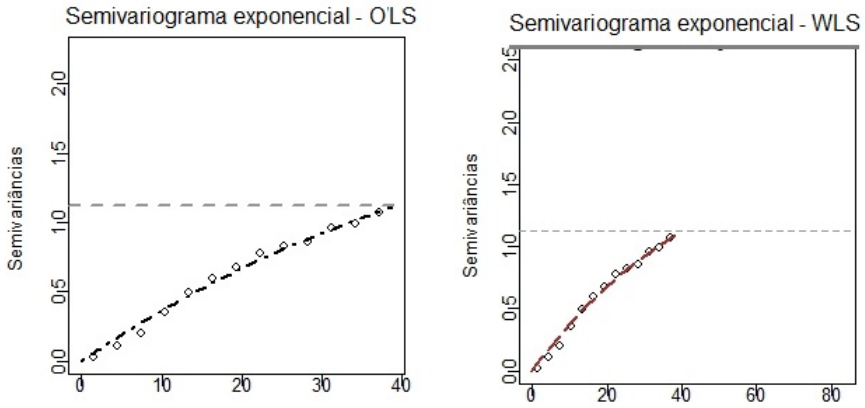

Figura 45: Semivariograma Exponencial ajustado ao grupo "Amostra Total"

A Tabela 21 mostra os parâmetros que foram ajustado o semivariograma Exponencial ao grupo " Amostra total".

Tabela 21: Parâmetros do semivariograma Exponencial ajustado ao grupo "Amostra total"

Parâmetros	Valores
Método de ajuste	OLS
Alcance prático	165,79 m
Contribuição	2,205
Efeito-pepita	0,0
Modelo ajustado	Exponencial

m= metros

Os parâmetros do semivariograma ajustado, estão apresentados na Figura 45 e na Tabela 21, mostrando um ajuste razoável ao modelo Exponencial, com alcance prático de $165,8$ metros e sem efeito pepita, indicando que o modelo é adequado aos dados.

A Figura 46 apresenta alguns resultados importantes referentes aos dados amostrais e preditos, tais como, densidade dos erros e dos erros padronizados, regressão dos valores preditos em função dos dados amostrais, probabilidade teórica

em função da probabilidade observada em que mostram ajuste ruim, erros e erros padronizados em função dos valores preditos e em função dos dados amostrais.

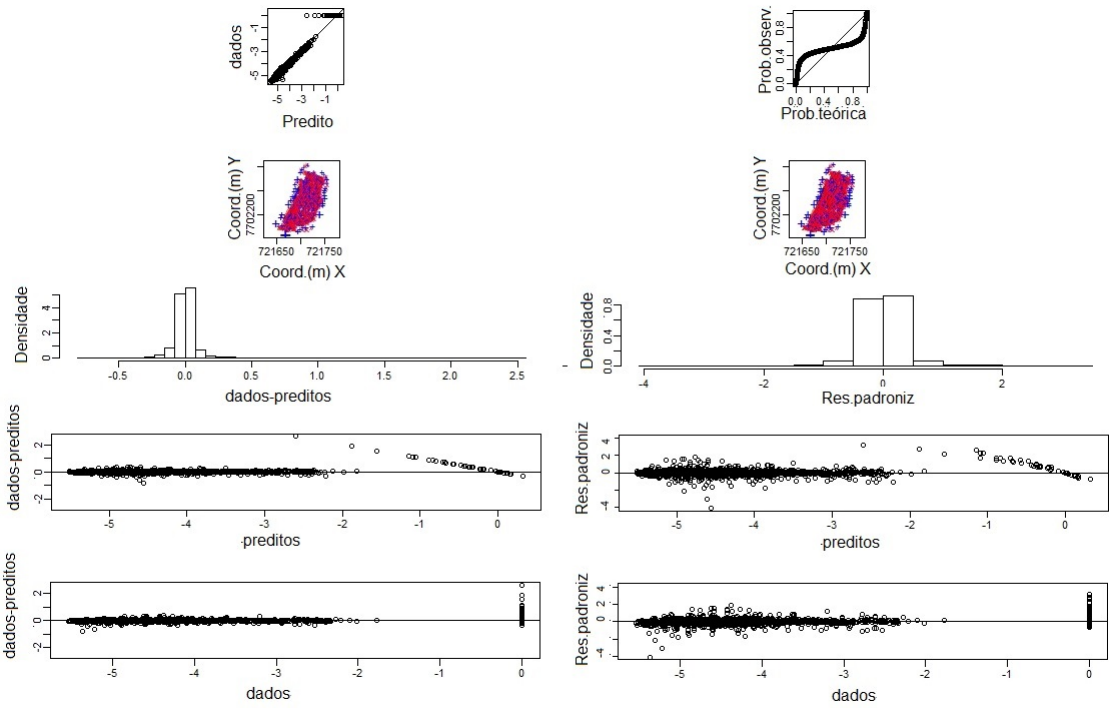

Figura 46: Representação gráfica dos valores preditos versus valores da Amostra Completa - Dados batimétricos

Na Figura 46, o gráfico da densidade das diferenças entre os valores amostrais e os valores preditos está mostrando que as maiores densidades estão nas diferenças nulas. Outros aspectos a ser considerados, são a relação entre essas diferenças e os dados e entre essas diferenças e os valores preditos, bem como, entre os resíduos e resíduos padronizados e os dados, em que a núvem de pontos uma mostra concentração em torno da linha horizontal, quando os valores preditos são maiores que 2 (em módulo) e um deslocamento bastante acentuado da linha horizontal para os valores preditos entre 2 e zero (em módulo). Os pontos de ajuste entre os dados e os erros, também mostraram-se concentrados em torno da linha horizontal, a regressão entre os dados e os valores preditos mostraram boa aproximação à reta, porém a regressão entre probabilidade observada na amostra e probabilidade teórica

não apresentou boa aproximação.

A Figura 47 mostra o mapa de Krigagem Ordinária da Amostra Completa referente aos dados batimétricos.

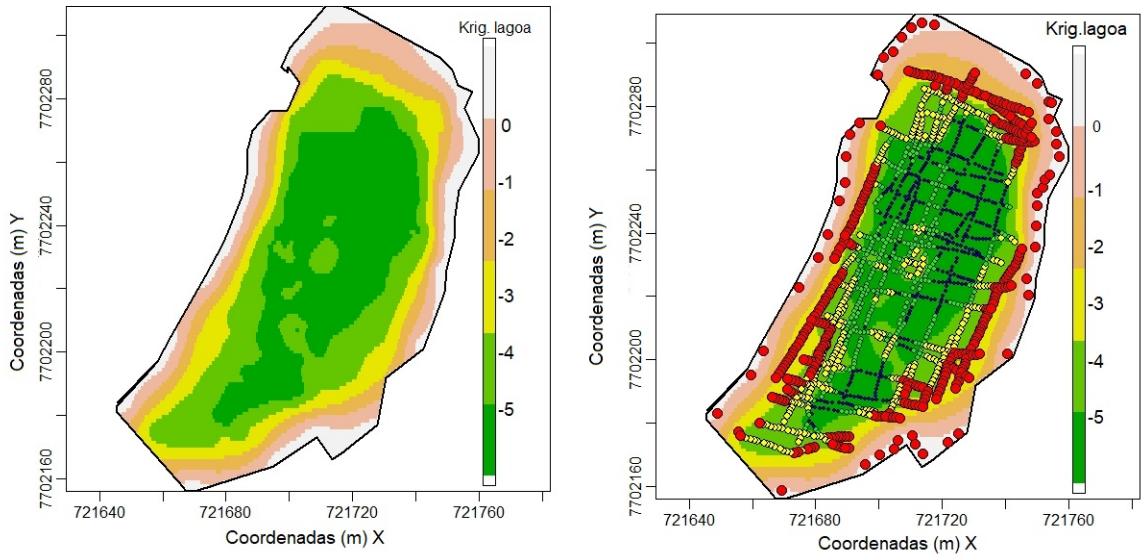

Figura 47: Mapa de Krigagem Ordinária da Amostra Completa - Dados batimétricos

Pode-se notar no mapa da Figuras 47, que as estimativas das maiores profundidades variaram para mais de $-5,0$ metros a $-4,5$ metros, nas partes mais centrais da lagoa ao longo do eixo Y, porém, nota-se pela sobreposição de pontos, que estas profundidades deveriam concentrar apenas no centro do mapa, porque apenas neste local da área foram encontradas as maiores profundidades. Os pontos em vermelho nas extremidades da lagoa têm dois formatos, que representam profundidades inferiores a 2 (em módulo) e profundidades nulas. Nestes pontos, as predições por Krigagem Ordinária estão aparentemente condizentes com os dados amostrais, mas as estimativas das profundidades de $3,5$ metros a $4,5$ metros (em módulo), não são aparentemente condizentes com a amostra.

A Figura 50 apresenta o mapa da Variância de Krigagem do grupo "Amostra completa".

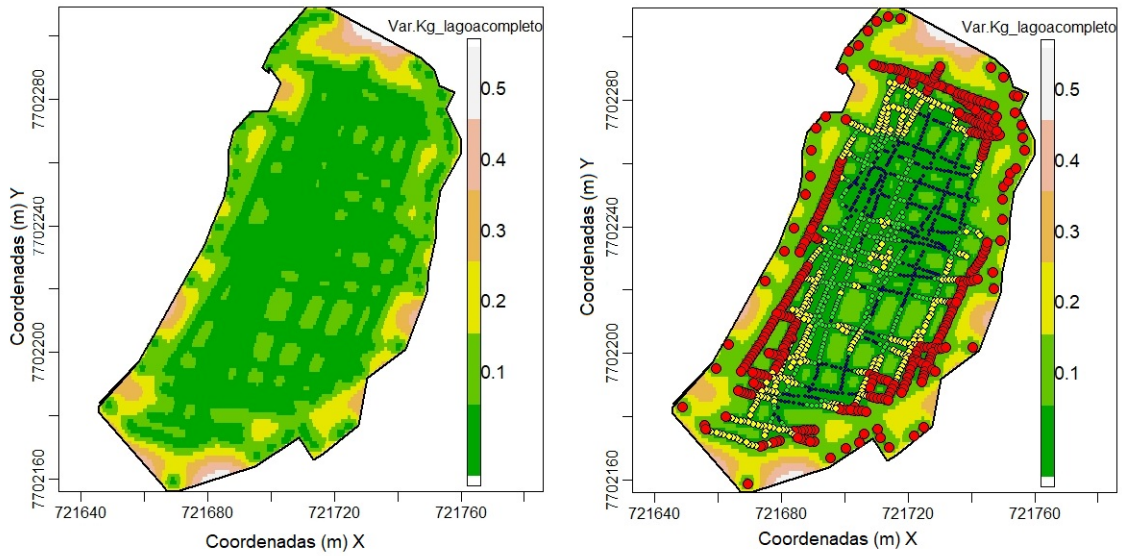

Figura 48: Mapa de Variância de Krigagem da Amostra Completa - Dados batimétricos

Pode se observar no mapa de variâncias de krigagem (Figura 50), que os locais com maior concentração de pontos amostrais apresentaram as menores variâncias de krigagem (menores ou iguais a 0,1) e os locais onde as predições por Krigagem Ordinária apresentaram maiores variâncias foram nas regiões das extremidades da malha, que era esperado, devido a menor quantidade de vizinhos nestes pontos.

5.2.5 Comparação dos resultados dos grupos - Dados batimétricos

Nas tabelas 22, 23, 24 e 25 estão apresentados os resultados comparativos entre os ajustes realizados no grupo G1, G2, G3 e Amostra Total.

Tabela 22: Resultados comparativos dos parâmetros dos semivariogramas dos três grupos e do grupo "Amostra Total"

Medidas	G1	G2	G3	Amostra Total
Modelo	Cúbico	Gaussiano	Circular	Exponencial
Alcance Prático	9,69	22,24	23,58	165,79
Contribuição	0,059	0,75	0,254	2,205
Efeito pepita	0,005	0	0	0

Tabela 23: Resultados comparativos da validação dos valores preditos × valores reais dos três grupos e da "Amostra Total"

Medidas	G1	G2	G3	Amostra Total
coef. linear	0,162	0,19	0,0853	0,1875
coef. angular	1,033	1,04	1,027	1,0422
$R^2_{ajust.}$	0,901	0,982	0,964	0,981
RSE	0,06445	0,123	0,00016	0,0122
Média.erro(WLS/OLS)	0,006	0,0073	0,011	0,011
Desv.pad.(sd.error)	0,712	0,41	0,80	0,418
Média(erro.pad.)	0,029	0,0091	0,00016	0,012
AIC	-144,185	-1446	-84,571	-1442
Deviance residual	0,224	23,66	0,9446	29,61
Devíance nula	2,192	1.080,00	30,601	1580,48
gl. dev. residual	54	1267	63	1409
gl. dev. nula	55	1268	64	1410

Tamanhos amostrais: G1(56); G2(1269); G3(86); Amostra Total(1411)

Obs:. *gl. refere-se ao grau de liberdade.

As discussão dos resultados dos três grupos estão na seção 5.2.6.

Tabela 24: Resultados comparativos dos valores preditos por Krigagem Ordinária × (valores reais) dos três grupos e do grupo Amostra Total

Medidas	G1	G2	G3	Amostra Total
Média	-4,542(-4,78)	-2,43(-4,27)	-2,843(-3,54)	1,097 (-4,18)
Mediana	-4,522	-2,35	-2,832	1,11
Variância	0,008 (0,04)	1,154 (0,85)	0,0467(0,48)	0,453(1,12)
Desv.Pad.	0,08 (0,2)	1,074 (0,92)	0,216 (0,69)	0,673 (1,06)
Min.	-5,369 (-5,37)	-5,57 (-5,54)	-4,357 (-4,36)	-5,57 (-5,54)
Max.	-4,48 (-4,53)	-0,285 (0)	-1,783 (-1,77)	2,353 (0)
Curt.	32,132 (4,09)	4,05 (9,34)	14,99 (2,25)	57,59 (8,57)
Assim.	4,985 (1,21)	0,692 (2,01)	2,19 (0,71)	6,96 (2,09)
C.V. %	1,97 (4,2)	44,3 (22,0)	7,6 (19,5)	61,3 (25,32)

∗Os valores dentro dos parênteses() referem aos dados amostrais

Tabela 25: Resultados comparativos da qualidade de predição por Krigagem Ordinária nos três grupos e no grupo Amostra Total

Medidas	G1	G2	G3	Amostra Total
Variância de Krigagem	0,0066	0,127	0,083	0,102
Erro padrão de estimação	0,00041	0,0025	0,0076	0,00223
Média dos erros de estimação	0,006	0,0091	0,011	0,0105

5.2.6 Discussão dos resultados dos três grupos e Amostra Total - Dados batimétricos

A segunda aplicação da metodologia MPPs, referente aos dados batimétricos (Dados2) com 1411 amostras, apontou dois pontos de mudanças e com isso, foram obtidos os cortes na 56^a posição e na 1325^a posição do banco de dados, ordenado pela covariável espacial 'Distância Média entreVizinhos", formando três grupos. Pode se observar pelo mapa de Krigagem Ordinária (Figura 47) que o grupo de "Amostra Total", ao se assumir um modelo único de semivariograma, obteve-se

um mapa de Krigagem Ordinária com um prolongamento da área de maior profundidade porque a estacionaridade de 1^a ordem não era totalmente satisfeita e devido a contamina cão dos dados pela mudança de estrutura da média, provocou maior suavização das faixas de profundidades, estendendo a região da faixa de maior fluxo, com as altas profundidades, para uma área maior que a área existente na amostra.

Observa-se também que o grupo G3 foi formado pela junção das duas "cabeceiras" da malha amostral, na qual em uma das extremidades estão os maiores valores amostrais, e em outra, os menores valores, seguidos de muitos valores nulos que foram retirados da amostra por constituir os pontos mais extremos das margens que impediam um bom ajuste do semivariograma.

Considerando que o valor do patamar, por definição do método de Krigagem Ordinária, precisa ser o mais próximo possível da variância dos dados, aceitável até o dobro da variância dos dados, de acordo com as Tabelas 22 e 24, ogrupo G1 1 teve variância $0,04$ metros e patamar $0,064$ metros, o grupo G2 teve variância $0,85$ metros e patamar $0,75$ metros e o grupo G3 teve variância $0,48$ metros e patamar $0,254$ metros, em comparação com o grupo "Amostra Total" que teve variância $1,12$ metros e patamar $2,205$ metros. Diante disso, pode-se observar que os grupos G1, G2 e G3 obtiveram os melhores valores de semivariancias, aproximando, sem extrapolar a variância dos dados.

Os valores de $R^2_{ajust.}$ da Tabela 23 mostram que em todos os grupos e no grupo " Amostra Total" tiveram acima de 90% de qualidade, indicando que os valores estimados pelo variograma ajustaram bem à reta de regressão com os valores observados e o RSE foi menor no Grupo 3, enquanto a média dos erros do método OLS/WLS foi menor no grupo G1 e no grupo G2.

A média e o desvio dos erros padronizados no grupo G3, seguido do grupo G1, ficaram mais próximos de 0 e 1, respectivamente, mostrando que os grupos G3 e G1 foram os que menos erraram no processo de estimação.

A Deviance Residual foi menor no grupo G1 e no grupo G3, seguida do grupo G2, indicando que os semivarigramas que melhor se ajustaram foram nos

grupos G1, G2 e G3, respectivamente. O mesmo foi observado nos valores de deviance residual e nula, relativas aos graus de liberdade de cada grupo, mostrando que o pior ajuste foi obtido no grupo "Amostra Total".

A Tabela 24 mostra que a Krigagem Ordinária do grupo " Amostra Total" obteve valor máximo predito superior ao valor amostral(nível da água), predizendo um ponto a $2,35$ metros acima do nível da água. Em relação à média, ao valor máximo e ao valor mínimo, o grupo G1, teve os melhores valores preditos comparados aos demais grupos. O grupo G3 aproximou mais os valores preditos dos valores amostrais mínimos e máximos, o grupo G2 teve a melhor assimetria, próxima de zero e o grupo G3 teve o menor C.V..

Quanto a Variância de Krigagem, a Tabela 25 mostra que o grupo G1 e o grupo G3 tiveram os menores valores, enquanto o grupo G1 teve também, o menor Erro Padrão de Estimação.

Diante dos resultados apresentados para os dados batimétricos (Dados 2), em todos os aspectos da qualidade da Krigagem Ordinária, a partição da malha amostral por meio do MPPs mostrou-se satisfatório porque ao separar os grupos amostrais, no grupo G2 notou-se que a estacionaridade de 1^a ordem foi mantida adequadamente, sem contaminação dos dados menos estacionários, e no grupo G1, por ser a região de maior flutuação da média, conseguiu-se ainda assim, obter um mapa de Krigagem Ordinária melhor que o mapa gerado para o grupo da "Amostra Total".

Ao aplicar o MPPs, pode-se verificar na Figura 37, que no grupo G2, a Krigagem Ordinária fez predições adequadas à escala de valores dos pontos amostrais e a Krigagem Ordinária do grupo G1 (Figura 32) fez predições condizentes à escala de cores, mostrando maior continuidade do grupo e também menor suavização do mapa.

Baseando-se na discussão apresentada, pode-se observar em termos da qualidade da Krigagem Ordinária que o grupo G1, o grupo G3 e o grupo G2, nesta ordem, tiveram os melhores mapas de Krigagem Ordinária, enquanto o grupo "Amostra

Total" mostrou-se com baixa qualidade, superestimando os valores elevados devido à influência das flutuações dos valores de semivariancias locais que provocam a falta de estacionaridade de 1^a ordem e dificultam um bom ajuste do semivariograma, tendo como consequências as más predições.

5.3 Algumas possibilidades e restrições a aplicação do método de Krigagem via MPPs

Diante da capacidade da metodologia, pode parecer que esta irá resolver todos os problemas de falta de estacionaridade da média, mas algumas restrições devem ser consideradas, pois, o método é aplicado para solucionar o problema de dupla ou múltipla estacionaridade da média, ou seja, quando há uma suspeita de aparente mudança abrupta na média e esta mudança se estabiliza somente após um espaço contínuo considerável envolvido em um número significativo de amostras, caracterizando assim a mudança na estrutura de semivariância.

Outro ponto a ser considerado é a necessidade de obtenção de um número significativo de amostras em cada grupo, tais que permitam aplicar Krigagem em cada grupo, com a vantagem de que o método permite a versatilidade de aplicar apenas aos grupos que atendam às pressuposições necessárias. Quando, impondo um número pequeno de pontos de corte, um dos grupos não atender a esta condição, uma possibilidade é aplicar métodos determinísticos no grupo não atendido, tais como o Inverso da Distância.

Para evitar a restrição da estacionaridade de primeira ordem, o método pode parecer permitir a aplica cão indiscriminada de Krigagem, porém deve se considerar que o corte é efetuado utilizando valores das semivariancias locais, em função das Distâncias Médias entre Vizinhas e portanto, caso não exista a dependência espacial, o método não é indicado e poderá não convergir.

A Krigagem Ordinária via Partição escolhida por MPPs, apresenta muitas vantagens em relação a Krigagem Ordinária da área global, dentre elas, per-

mite, em caso específico de possibilidade de duas ou mais estruturas de dependência espacial, diminuir o tamanho amostral, quando o tamanho amostral total for custoso computacionalmente, garantindo melhor qualidade dos mapas.

Utilizando a Krigagem Ordinária transformada em imagem *raster* e salvando em *shape* é possível juntar os dois grupos de mapas em um só mapa.

O método poderá ser estendido a quase todos os tipos de Krigagem, inclusive a Krigagem Indicativa, para esta, poderá ser adaptada a função priori a dados discretos, Poisson ou Binomial e a função de verossimilhança a uma distribuição Binomial Negativa.

6 CONCLUSÃO

Conclui-se que o método MPPs aplicado à krigagem é capaz de identificar os pontos de mudança da média e garantir grupos com médias mais estacionárias, assim, pode se afirmar que o método MPPs é mais viável de ser utilizado para se aplicar Krigagem Ordinária em comparação ao método de krigagem Ordinária clássico, quando a distribuição amostral dos dados apresentar características de se ter dependência espacial mas a estacionaridade de primeira ordem não for totalmente satisfeita e mostrar haver possibilidade de mais de uma estacionaridade da média, sendo o método indicado por permitir encontrar os pontos de mudança na média e identificar os locais mais prováveis de haver estas mudanças, formando grupos de médias mais estacionárias no espaço, ideais para produzir mapas de krigagem mais precisos e acurados.

6.1 Sugestões de trabalhos futuros

Diante da nova concepção que se forma ao inserir um método Bayesiano de partição para buscar grupos mais estacionários em termos de média e variância, abre-se o leque de possibilidades e uma das formas de ampliação é o seu uso em dados de saúde, em estudos epidemiológicos, relacionando covariáveis espaciais juntamente com covariáveis de saúde para se criar mapas regionalizados de melhor qualidade e favorecer políticas de prevenção e tratamento de epidemias e pandemias de forma otimizada (um trabalho já iniciado).

Outra possibilidade é testar o método para outros tipos de krigagem, inclusive a krigagem indicativa, utilizando outras funções a priori e de verossimi-

lhança.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, M. F. F. Uso da Krigagem Indicativa na seleção de áreas propícias ao cultivo de café em consorciação ou rotação com outras culturas. Viçosa-MG, 2013. 137p. Dissertação (Mestrado) - Universidade Federal de Viçosa.

ANDRIOTTI, J. L. S. **Fundamentos de Estatística e Geoestatística**. São Leopoldo: UNISINOS:Coleção Manual Universitário, 2003. 165p.

BARRY, D.; HARTIGAN, J. A. Product partition models for change point problems. **The Annals of Statistics**, v.20, n.1, p.260–279, 1992.

BARRY, D.; HARTIGAN, J. A. A Bayesian analysis for change point problem. **Journal of the American Statistical Association**, v.88, n.421, p.309–319, 1993.

BASSETO, V. F.; GONZATTO, O. A. J.; ROSSONI, D. F.; JARDEL, M. H. Estimadores de Semivariância: Uma Revisão. **Ciência e Natura**, v.38, n.3, 2016.

BEAUMONT, L. J.; HUGHES, L.; PITMAN, A. Why is the choice of future climate scenarios for species distribution modelling important? **Ecology letters**, v.11, n.11, p.1135–1146, 2008.

BESAG, J.; YORK, J.; MOLLIÉ, A. Bayesianimage restoration, with two applications in spatial statistics. **Annals of the Institute of Statistical Mathematics**, v.43, n.1–20, p.21–59, 1991.

BHATTACHARYA, R. N.; WAYMIRE, E. C. **Stochastic processes with applications**. Philadélphia: Society for Industrial and Applied Mathematics, 1990. 691p.

BURROUGH, P. A. Development of intelligent geographical information systems. **International journal of geographical information systems**, v.6, n.1, p.1–11, 1992.

CÂMARA, G.; MONTEIRO, A. M.; FUCKS, S. D.; CARVALHO, M. S. Análise espacial e geoprocessamento. **Análise espacial de dados geográficos**, v.2, 2002.

CAMARGO, E. C. G. Geoestatística:fundamentos e aplicações. **Geoprocessamento para projetos ambientais. São José dos Campos:INPE**, 1998.

CHUNG, C. K.; CHONG, S.; VARSA, E. C. Sampling strategies for fertility on a stoy silt loam soil. **Communications in Soil Science and Plant Analysis**, v.26, n.5-6, p.741–763, 1995.

CLARKE, A. B.; DISNEY, R. L. **Probabilidade e Processos Estocásticos**. LTC, 1979.

CRESSIE, N.; HAWKINS, D. M. Robust estimation of the variogram: I. **Journal of the International Association for Mathematical Geology**, v.12, n.2, p.115–125, 1980.

CRESSIE, N. A. C. Statistics for spatial data: Wiley series in probability and mathematical statistics. **online**, 1993.

DEUTSCH, C. V.; JOURNEL, A. G. **GSLIB: Geostatistical Software Library and User's Guide. Hauptbd**. Oxford university press, 1992.

DEUTSCH, C. V.; JOURNEL, A. G.; ET AL. Geostatistical software library and users guide. **Oxford University Press, New York**, 1998.

EHLERS, R. S. Inferência bayesiana. **Departamento de Matemática Aplicada e Estatística, ICMC-USP**, p.64, 2011.

FELGUEIRAS, C. A. Modelagem ambiental com tratamento de incertezas em sistemas de informação geográfica: o paradigma geoestatístico por indicação. São José dos Campos, 1999. Tese (Doutorado) - INPE-Instituto Nacional de Pesquisas Espaciais.

FERNANDEZ, P. J. **Introdução a Teoria das Probabilidades**. Rio de Janeiro-RJ: Livros Técnicos e Científicos, 1973. 174p.

FERREIRA, I. O.; RODRIGUES, D. D.; DE P SANTOS, A. Levantamento batimétrico automatizado aplicado à gestão de recursos hídricos. Estudo de caso: Represamento do ribeirão São Bartolomeu, Viçosa -Mg. **Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação**, v.4, p.1–8, 2012.

FERREIRA, J. A.; LOSCHI, R. H.; COSTA, M. A. Detecting changes in time series: A product partition model with across-cluster correlation. **Signal Processing**, v.96, p.212–227, 2014.

GAMERMAN, D.; LOPES, H. F. **Markov Chain Monte Carlo: stochastic simulation for Bayesian inference**. Chapman and Hall/CRC, 2006.

GOOVAERTS, P. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. **Journal of hydrology**, v.228, n.1-2, p.113–129, 2000.

GUERRA, P. A. **Geoestatística Operacional**. Brasília: Ministério das Minas e Energia, 1988. 145p.

HARLAN, W. S. A quick derivation of geostatistical Kriging, 2013.

HARTIGAN, J. A. Partition models. **Communications in statistics-Theory and methods**, v.19, n.8, p.2745–2756, 1990.

HASLETT, J. On the sample variogram and the sample autocovariance for non-stationary time series. **Journal of the Royal Statistical Society: Series D (The Statistician)**, v.46, n.4, p.475–484, 1997.

HOLMES, C. C.; DENISON, D. G. T.; RAY, S.; MALLICK, B. K. Bayesian Prediction via Partitioning. **J. Comput. Graph. Statist**, v.14, n.1, p.811–830, 2005.

HUIJBREGTS, C. J. **Regionalized variables and quantitative analysis of spatial data. In: Davis, J.C. and McCullagh, M.J. (ed) Display and analysis of spatial data**. New York: John Wiley, 1975. 38-53p.

IBGE, A. M. S. Malha municipal digital do Brasil: situação em 2005. **Rio de Janeiro: IBGE**, 2006.

IMAI, N. N.; VICENTE, J.; LIMA, D. L. T.; VILMA, M.; SILVA, E. A.; VOLL, E.; OLIVEIRA Análise comparativa da interpolação por krigagem ordinária e krigagem por indicação no caso de ervas daninhas em cultura de soja. In: PROJETO MUDANÇA DO REFERENCIAL GEODÉSICO, 2003. ; resumos. Belo Horizonte: XXI Congresso Brasileiro de Cartografia. Publicação em CD-Rom sem paginação, 2003.

ISAAKS, E. H.; SRIVASTAVA, R. M. An introduction to applied geostatistics. Rel. téc., Oxford university press, 1989.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining geostatistics**. Academic press, 1978.

JOURNEL, A. G.; HUIJBREGTS, C. J. **Mining Geostatistics**. Academic press, 1978.

JOURNEL, A. G.; JOURNEL, A. G. **Fundamentals of Geostatistics in Five Lessons**. American Geophysical Union Washington, DC, 1989. 8v.

KREH, M. Bessel functions. **Lecture Notes, Penn State-Göttingen Summer School on Number Theory**, v.82, 2012.

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

LANDIM, P. M. B. **Análise estatística de dados geológicos multivariados**. Oficina de Textos, 2011.

LOSHI, R. H.; CRUZ, F. R. B. Extension to the Product Partition Model Computing the Probability of a change. **Computational Statistics and data analisis**, v.48, n.2, p.255–268, 2005.

LUNN, D. J.; THOMAS, A.; BEST, N.; SPIEGELHALTER, D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. **Statistics and computing**, v.10, n.4, p.325–337, 2000.

MANZIONE, R. L. Variabilidade espacial de atributos químicos do solo em Araguari-MG. Botucatu, 2002. 141p. Dissertação (Mestrado) - Universidade Estadual Paulista Júlio de Mesquita Filho- UNESP.

MCBRATNEY, A. B.; WEBSTER, R.; BURGESS, T. M. The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I: Theory and method. **Computers & Geosciences**, v.7, n.4, p.331–334, 1981.

MOOD, A.; GRAYBILL, F. A.; DBOES, C. **Introduction to the Theory of Statistics**. United State of America: McGraw-Hill, 1974. 480p.

MÜLLER, P.; QUINTANA, F.; ROSNER, G. Bayesian clustering with regression. Rel. téc., Working paper, 2008.

MÜLLER, P.; QUINTANA, F.; ROSNER, G. L. A product partition model with regression on covariates. **Journal of Computational and Graphical Statistics**, v.20, n.1, p.260–278, 2011.

OLEA, R. A. **Optimum mapping techniques using regionalized variable theory**. Kansas Geological Survey, 1975.

OLEA, R. A. A six-step practical approach to semivariogram modeling. **Stochastic Environmental Research and Risk Assessment**, v.20, n.5, p.307–318, 2006.

OLIVER, M. A.; WEBSTER, R. **Basic Steps in Geoestatistics: The variogram and Kriging**. University Reading, United Kigdom of Great Britain and North Ireland: Springer, 2015. 100p.

PAGE, G. L.; QUINTANA, F. A. Spatial Product Partition Models. **Bayesian Analysis**, v.11, n.1, p.265–298, 2016.

PARK, J.; DUNSON, D. B. Bayesian Generalized Product Partition Model. **National Cancer Institute and Duke University**, v.20, n.1, p.1203–1226, 12010.

QUINTANA, F. A. A predictive view of Bayesian clustering. **Journal of Statistical Planning and Inference**, v.136, n.8, p.2407–2429, 2006.

QUINTANA, F. A.; IGLESIAS, P. L. Bayesian Clustering and product partition models. **Journal of the Royal Statistical Society**, v.65, n.2, p.557–574, 2003.

RAMIREZ, J. J.; COLIN, R. T. Especie nueva del género Jatropha(Euphorbiaceae) de la sección Mozinna. In: ANALES DEL INSTITUTO DE BIOLOGÍA. SERIE BOTÁNICA, 65, Instituto de Biología, Universidad Nacional Autónoma de México, 1994. ; resumos. México: Anales del Instituto de Biología, Universidad Nacional Autónoma de México:Serie Botánica, 1994. 1.

REICH, B. J.; FUENTES, M.; ET AL. A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. **The Annals of Applied Statistics**, v.1, n.1, p.249–264, 2007.

SALVIANO, A. A. C. **Variabilidade de atributos de solo e de Crotalaria juncea em solo degradado do municipio de Piracicaba-SP.**. ESALQ, 1996.

SILVA, A. F.; QUARTEZANI, W. Z.; ZIMBACK, C. R. L.; LANDIM, P. M. B. Aplicação da geoestatística em ciências agrárias. **Botucatu, SP: FEPAF**, 2011.

DA SILVA, A. F.; ZIMBACK, C. R. L.; LANDIM, P. M. B. Classificação de imagens em áreas cultivadas com citros por técnicas de sensoriamento remoto e geoestatística. **Energia na Agricultura**, v.27, n.3, p.01–15, 2012.

DA SILVA, A. P.; LIBARDI, P. L.; VIEIRA, S. R. Variabilidade espacial da resistência à penetração de um Latossolo Vermelho-Escuro ao longo de uma transeção. **Revista Brasileira de Ciência do solo**, v.13, n.1, p.1–5, 1989.

SMITH, A. A Bayesian approach to inference about a change-point in a sequence of random variables. **Biometrika**, v.62, n.2, p.407–416, 1975.

TEAM, R. C. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing, 2015.

TRANGMAR, B. B.; YOST, R. S.; WADE, M. K.; UEHARA, G.; SUDJADI, M. Spatial Variation of Soil Properties and Rice Yield on Recently Cleared Land 1. **Soil Science Society of America Journal**, v.51, n.3, p.668–674, 1987.

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

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo. **Tópicos em ciência do solo. In: Sociedade Brasileira de Ciência do Solo**, v.1, n.1, p.1–54, 2000.

WACKERNAGEL, H.; OLIVEIRA, V.; KEDEM, B. Multivariate geostatistics. **SIAM Review**, v.39, n.2, p.340–340, 1997.

WANG, X. J.; QI, F. The effects of sampling design on spatial structure analysis of contaminated soil. **Science of the total environment**, v.224, n.1-3, p.29–41, 1998.

WEBSTER, R.; OLIVER, M. A. Sample adequately to estimate variograms of soil properties. **Journal of soil science**, v.43, n.1, p.177–192, 1992.

YAMAMOTO, J. K. **Avaliação e Classificação de Reservas Minerais**. Edusp, 2001. 38v.

YAMAMOTO, J. K.; FURUIE, R. A. Um estudo sobre estimativa de dados lognormais. **Geociências (São Paulo)**, v.29, n.1, p.5–19, 2010.

YAMAMOTO, J. K.; LANDIM, P. M. **Geoestatstica: Conceitos e Aplicações**. Oficina de Textos, 2013.

YAMAMOTO, J. K.; MAO, X. M.; KOIKE, K.; CROSTA, A. P.; LANDIM, P. M. B.; HU, H. Z.; WANG, C. Y.; YAO, L. Q. Mapping an uncertainty zone between interpolated types of a categorical variable. **Computers & Geosciences**, v.40, p.146–152, 2012.

YAO, Y. Estimation of a noise discrete-time step function: Bayes and empirical. **Bayes approaches**, v.12, n.4, p.1434–1447, 1984a.

YAO, Y. Estimation of a noise discrete-time step function:Bayes and empirical Bayes approaches. **The annals of statistics**, v.12, n.4, p.1434–1447, 1984b.