Universidade de Brasília
Departamento de Estatística

Comparação entre os métodos de Krigagem e Regressão Geograficamente Ponderada para Estimação de Superfícies a Partir de Fenômenos Não-Contínuos

Gabriela Carneiro de Almeida

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Gabriela Carneiro de Almeida

Comparação entre os métodos de Krigagem e Regressão Geograficamente Ponderada para Estimação de Superfícies a Partir de Fenômenos Não-Contínuos

Orientador: Prof. Dr. Alan Ricardo da Silva

Relatório apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Agradecimentos

Expresso meus sinceros agradecimentos à Universidade de Brasília pela oportunidade de realizar este trabalho e pelo apoio ao longo deste processo. Agradeço especialmente ao meu orientador, Prof. Dr. Alan Ricardo da Silva, pela paciência, apoio e incentivo constantes. Também sou grata à minha família pelo amor, compreensão e apoio incondicional ao longo de toda a minha jornada acadêmica e por sempre apoiarem meus sonhos.

Agradeço aos professores e colegas de curso pela troca de conhecimento, debates construtivos e colaboração, em particular ao meu colega de curso Cesar Augusto Galvão, que sempre me incentivou a seguir em frente mesmo nos momentos mais desafiadores. Também sou grata à professora Profa. Dra. Cira Souza Pitombo e à sua aluna Samille Santos Rocha por disponibilizarem dados e informações relevantes para a elaboração deste trabalho.

Não posso deixar de agradecer às minhas amigas, Talita Souza Carmo e Camila Gomes Cabral, pelo apoio incondicional sempre. E à amiga Larisse Lima, cujo apoio foi fundamental para que eu conseguisse concluir o curso.

Por fim, dedico este trabalho a todas as pessoas que, de alguma forma, contribuíram para o meu crescimento pessoal e acadêmico. Obrigado a todos que tornaram possível e mais leve a minha caminhada durante o curso.

Resumo

O estudo conduzido teve como objetivo comparar os métodos de Krigagem e Regressão Geograficamente Ponderada para a estimação de superfícies em fenômenos não-contínuos, usando dados relacionados à demanda por transportes em uma região central da cidade de São Paulo. Na metodologia, foram analisados os procedimentos de criação de superfícies por Krigagem e Regressão Geograficamente Ponderada, levando em consideração a natureza dos dados e a distribuição espacial das amostras. Foram determinados os parâmetros necessários para cada método, como o parâmetro de suavização na Regressão Geograficamente Ponderada e a análise dos semivariogramas na Krigagem. Os resultados da análise das superfícies de renda familiar e viagens por automóvel indicaram que as duas metodologias utilizadas para a modelagem das superfícies produzem resultados semelhantes, porém, a Regressão Geograficamente Ponderada mostrou-se vantajosa no sentido do ajuste ser mais fácil, não necessitar das suposições previstas para a Krigagem e por gerar estimativas com desvios padrão menores do que a Krigagem. Em conclusão, o estudo destacou a possibilidade de geração de superfícies a partir da técnica de Regressão Geograficamente Ponderada, que são comparáveis àquelas geradas pela técnica de krigagem, sem os pressupostos requeridos por essa última.

Palavras-chaves: Krigagem; Regressão Geograficamente Ponderada; Estimação de Superfícies; Fenômenos Não-Contínuos; Fenômenos Contínuos.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

A análise espacial de dados geográficos é importante para estudar fenômenos levando em consideração sua localização espacial (Câmara et al., 2004). A capacidade de apresentar fenômenos visualmente em um padrão espacial é uma caracteristica que pode ser explorada em diversas áreas de estudo, como saúde, agronomia, ecologia, urbanização, dentre outras (Câmara et al., 2004). É possível categorizar os fenômenos em análise espacial em três tipos: eventos ou padrões pontuais; áreas com contagens e taxas agregadas; e superfícies (Câmara et al., 2004). Entender o fenômeno que será analisado é extremamente importante para aplicar as técnicas disponíveis que são mais adequadas.

Especificamente no caso de superfícies, a técnica de Krigagem foi desenhada para modelar fenômenos contínuos como temperatura, depósitos minerais etc. No estudo feito por Wan et al. (2021), são comparados diversos métodos para prever a temperatura de uma área por meio de dados coletados em estações de medição. Apesar de as medições serem feitas de forma pontual, o fenômeno "temperatura" em si é contínuo, o que torna a utilização da metodologia de Krigagem adequada. Ainda nesse contexto de fenômenos contínuos, a Krigagem pode ser utilizada para estimar a concentração de metais pesados no solo a partir de dados pontuais de coleta (Fu et al., 2022), ou para estimar a taxa de precipitação anual (Bostan et al., 2012), bem como caracterizar solos de acordo com a sua composição (Wan et al., 2021; Nawar et al., 2017) e até para para análise espacial em menor escala, como por exemplo, para verificar a temperatura de um recife de corais (Gorospe e Karl, 2011) ou medir a atividade cortical do cérebro de mamíferos (Trumpis et al., 2021), dentre outras aplicações.

No entanto, a técnica de Krigagem ordinária é frequentemente aplicada na modelagem de fenômenos não-contínuos, o que pode resultar na violação dos pressupostos estabelecidos, quais sejam, a estacionariedade, fenômenos contínuos e isotropia. Exemplos

de uso equivocado da Krigagem ordinária incluem a previsão do tráfego em vias (Selby e Kockelman, 2013), a modelagem do número de viagens no transporte público (Rocha et al., 2019), a previsão de locais críticos de criminalidade (Deshmukh e Annappa, 2019) e a estimativa de preços de propriedades (Derdouri e Murayama, 2020). Todas essas variáveis estão associadas a fenômenos não-contínuos.

Face o que foi exposto, fica claro que uma confusão recorrente reside na equívoca associação entre o nível de mensuração de uma variável aleatória contínua e a, erroneamente presumida, natureza contínua do fenômeno em questão. Logicamente, todo fenômeno contínuo gera uma variável aleatória contínua, mas nem toda variável aleatória contínua vem de um fenômeno contínuo, conforme a Figura 1.1. Em geral, uma variável aleatória discreta não pode ser derivada de um fenômeno contínuo. No entanto, o estudo de De Oliveira (2014) apresenta metodologias capazes de discretizar fenômenos contínuos de forma eficaz. Um exemplo disso é $\Lambda(s)$, que denota o nível de emissão do radionuclídeo Césio-137 em uma localização s, enquanto Y_i representa o número de fótons emitidos e capturados na região de amostragem s_i durante um período t_i por um medidor de raios Gama (Diggle et al., 1998). No entanto, é importante ressaltar que essa abordagem é altamente específica e não pode ser generalizada.

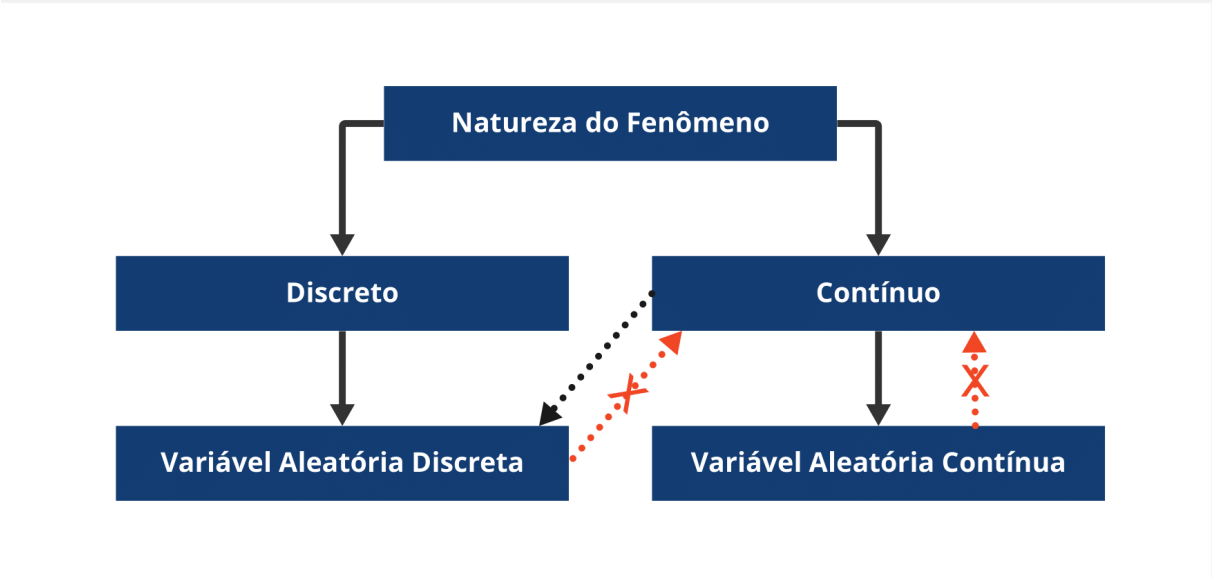


Figura 1.1: Natureza do fenômeno *versus* nível de mensuração da variável.

Já a temperatura, por exemplo, é um fenômeno contínuo medido em alguns locais discretos (devido à incapacidade humana de medí-la em todos os locais), e sua métrica é logicamente uma variável aleatória contínua. Já a quantidade de passageiros em alguns pontos de parada, é uma fenômeno não-contínuo, e sua métrica é uma variável aleatória

discreta. No caso, não é razoável supor que os pontos de parada não amostrados terão comportamentos similares aos vizinhos amostrados.

Numericamente, se a temperatura em um determinado local é de 28 graus Celsius e, a 100 metros após esse local, a temperatura medida é de 29 graus Celsius, então é razoável inferir que a temperatura entre esses pontos esteja situada em uma faixa entre 28 e 29 graus Celsius, por meio de um processo de interpolação. No que diz respeito à contagem de passageiros em pontos de parada próximos, essa lógica pode não ser aplicada.

A técnica de Regressão Geograficamente Ponderada (RGP) é utilizada em geral na modelagem de dados com dependência espacial, geralmente oriundos de fenômenos que envolvem contagens e taxas agregadas (Fotheringham et al., 2002). No entanto, a RGP também pode ser utilizada para estimar superfícies a partir de pontos não amostrados. A versão clássica da RGP é mais adequada para dados contínuos que seguem uma distribuição normal (Fotheringham et al., 2002). Entretanto, para analisar dados discretos, é mais recomendado que seja utilizada a Regressão Poisson Geograficamente Ponderada (RPGP) (Nakaya et al., 2005) ou a Regressão Binomial Negativa Geograficamente Ponderada (RBNGP) (Da Silva e Rodrigues, 2014).

Dessa forma, este estudo visa comparar os métodos de Krigagem e Regressão Geograficamente Ponderada para a estimação de superfícies a partir de fenômenos não-contínuos.

1.1 Objetivos

Esse trabalho tem como objetivo geral a comparação entre os métodos de Krigagem e Regressão Geograficamente Ponderada (RGP) para estimação de pontos não amostrados utilizando dados relativos à demanda por transportes em uma região central da cidade de São Paulo.

Os objetivos específicos são:

- Entender como é feita a criação de superfícies utilizando as metodologias de Krigagem e RGP;
- Reproduzir a análise dos dados feita no estudo de Rocha et al. (2019) utilizando as metodologias de modelagem de Krigagem e de RGP.

Capítulo 2

Análise de Superfícies

2.1 Introdução

A estimação de superfícies geralmente é realizada usando dados coletados de amostras pontuais. No entanto, a eficácia das metodologias empregadas para criar essas superfícies está diretamente ligada à natureza intrínseca dos dados. Quando os dados discretos são uma representação de um fenômeno contínuo, a estimação de superfícies via Krigagem pode ser aplicada com sucesso. Por outro lado, ao lidar com dados discretos provenientes de fenômenos não contínuos, é aconselhável recorrer a metodologias baseadas na RGP para a criação das superfícies. É importante observar que a RGP tradicional assume a normalidade dos dados, porém, quando essa suposição não é atendida, as metodologias RPGP e RBNGP surgem como alternativas mais adequadas para a geração de superfícies, mesmo na ausência de covariáveis. Neste Capítulo, serão abordadas técnicas de estimação de superfícies específicas para cada tipo de fenômeno subjacente aos dados coletados.

2.2 Estimação de superfícies

A fim de otimizar a aplicação de dados pontuais em contexto de geoprocessamento, é necessário empregar um método de interpolação (Câmara et al., 2004). O procedimento de interpolação desempenha um papel crucial na estimativa de valores em pontos não amostrados, resultando na criação de uma matriz de grade regular, onde cada elemento da matriz é associado a um valor numérico, geralmente denominado seu canto inferior direito, a partir do qual são definidos intervalos regulares nas direções horizontal e vertical (Câmara et al., 2004). Essa matriz gerada representa uma região da superfície

terrestre, começando a partir de uma coordenada inicial (Câmara et al., 2004).

Ainda, segundo Câmara et al. (2004), a construção de superfícies que sejam coerentes com a realidade do fenômeno em estudo requer a modelagem de sua variabilidade espacial. Em geral, os modelos destinados a gerar superfícies por meio da interpolação representam a variável em estudo como uma combinação da variabilidade global e local, considerando três principais tipos de modelos:

- Modelos determinísticos locais: nestes modelos, a estimativa de cada ponto na superfície é obtida através da interpolação de amostras vizinhas. A proximidade entre os pontos amostrados pode ser quantificada utilizando funções que consideram a distância entre eles. Ainda, nesse modelo, não são feitas suposições estatísticas quanto à variabilidade espacial, e a suposição predominante é de que os efeitos locais prevalecem;

- Modelos determinísticos globais: nesses modelos, a categoria de interpoladores utilizada para representar os fenômenos estudados considera a variabilidade em grande escala, enquanto a variação local é considerada irrelevante;

- Modelos estatísticos de efeitos locais e globais (Krigagem): os pontos da superfície são estimados com base em amostras próximas, usando técnicas de interpolação que incorporam estimadores estatísticos. Nessas circunstâncias, tanto a variabilidade local quanto a global também são modeladas por meio de métodos estatísticos.

No âmbito de modelos estatísticos que incorporam efeitos locais e globais, a Krigagem destaca-se como uma metodologia que emprega a técnica de interpolação em um conjunto limitado de dados amostrados. Seu objetivo é estimar os valores associados a um fenômeno contínuo em um espaço específico. As próximas seções fornecerão uma análise mais aprofundada da Krigagem, explorando suas aplicações e princípios fundamentais.

2.3 Semivariograma

A estimação de superfícies por Krigagem possibilita a quantificação dos erros associados às estimativas, o que confere uma avaliação da confiabilidade dos resultados obtidos. O processo de estimação implica na atribuição de pesos a cada unidade amostral, sendo essa ponderação determinada com base no semivariograma (Câmara et al., 2004). Nesse sentido, o semivariograma é uma ferramenta importante da geoestatística, pois permite a modelagem da estrutura de covariância espacial dos dados em análise.

Considerando Z como uma variável de interesse para um estudo, e u como uma coordenada no espaço, $Z(u)$ será a variável Z observada no ponto u. Dessa forma, o calculo do variograma entre dois pontos u e $u+h$, separadas por uma distância h é dado por (Câmara et al., 2004):

$$2\gamma(h) = E\left[Z(u) - Z(u+h)\right]^2 = Var\left[Z(u) - Z(u+h)\right]. \tag{2.3.0.1}$$

O estimador do variograma é dado por:

$$2\hat{\gamma}(h) = \frac{1}{N(h)} \sum_{i=1}^{N(h)} \left[Z(u_i) - Z(u_i+h)\right]^2, \tag{2.3.0.2}$$

sendo o semivariograma dado por:

$$\hat{\gamma}(h) = \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} \left[Z(u_i) - Z(u_i+h)\right]^2, \tag{2.3.0.3}$$

onde: $\hat{\gamma}(h)$ é o semivariograma estimado e $N(h)$ é o número de pares de valores mensurados para uma distância h, $Z(u)$ e $Z(u+h)$, separados pelo vetor distância h.

No entanto, essa fórmula não é robusta, podendo ser sensível a situações em que a variabilidade local não é constante (heteroscedasticidade), o que impede a estimação correta de seus parâmetros. Na prática, pode-se fazer a hipótese de que o fenômeno em estudo é isotrópico (mesmo comportamento em todas as direções). Desse modo, a determinação experimental do semivariograma depende apenas da distância entre amostras do fenômeno e não da direção entre elas (Câmara et al., 2004). Neste estudo não será dado enfoque a fenômenos anisotropicos, caracterizados pela variabilidade da sua distribuição espacial não depender somente da distância, mas também da direção.

Dessa forma, sob pressuposições de estacionariedade e média constante, pode-se obter a formulação de um padrão idealizado para o semivariograma experimental, conforme ilustrado na Figura 2.1. Antecipa-se que observações geograficamente próximas apresentem maior similaridade entre si do que aquelas separadas por distâncias mais extensas. Dessa forma, espera-se que o valor absoluto da diferença entre duas amostras, $Z(u)$ e $Z(u+h)$, aumente à medida que a distância entre elas cresce, atingindo um valor no qual os efeitos locais não teriam mais influência.

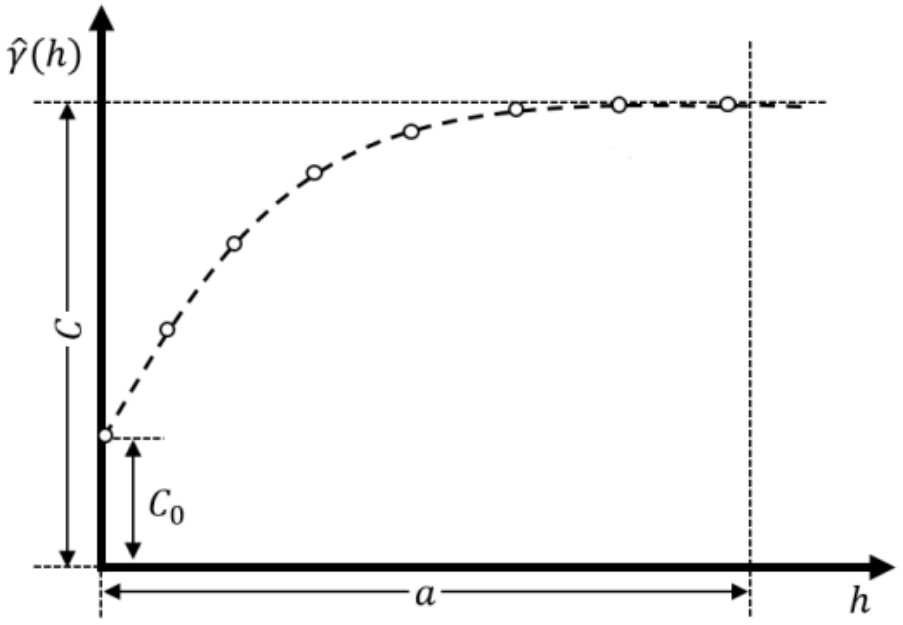


Figura 2.1: Parâmetros do semivariograma.
Fonte: Câmara et al. (2004)

onde:

- Alcance (a): A distância na qual as amostras demonstram correlação espacial;
- Patamar (C): representa o valor do semivariograma associado ao seu alcance (a). A partir desse ponto em diante, presume-se a inexistência de uma dependência espacial entre as amostras. Isso ocorre porque a variância da diferença entre pares de amostras ($Var[Z(u) - Z(u+h)]$) torna-se aproximadamente constante;
- Efeito Pepita (C_0): é o valor idealmente atribuído a $\gamma(0)$, que é igual a zero. No entanto, na prática, à medida que h se aproxima de zero, $\gamma(h)$ tende a um valor positivo denominado Efeito Pepita. Este efeito indica a descontinuidade do semivariograma para distâncias inferiores à menor distância entre as amostras. Esse efeito representa a semivariância para a distância zero e reflete a componente da variabilidade espacial que não pode ser atribuída a uma causa específica, representando, assim, a variabilidade ao acaso. A descontinuidade observada pode ser parcialmente devida a erros de medição, tornando difícil quantificar se a principal contribuição provém desses erros ou da variabilidade de pequena escala não captada pela amostragem.

A estimação dos parâmetros do semivariograma é feita de forma iterativa, por meio de algum método de estimação em conjunto com um método de otimização, que serão

apresentados nas próximas seções. Após realizar o ajuste no semivariograma experimental, torna-se crucial a escolha do semivariograma teórico a partir de um conjunto de funções matemáticas que descrevem a relação espacial. A determinação do modelo apropriado ocorre quando a configuração da curva do variograma experimental se alinha com a curva associada à função matemática, representando, assim, a tendência do modelo inicial. Na próxima seção serão apresentados alguns dos modelos matemáticos frequentemente utilizados para essa finalidade.

2.4 Modelos teóricos de Semivariograma

Os valores obtidos no semivariograma experimental são ajustados por modelos teóricos. Os modelos aqui apresentados são modelos básicos, denomidados de modelos isotrópicos. Como já manifestado anteriormente, os modelos anisotrópicos não serão abordados neste trabalho, mas eles existem e podem ser explorados mais detalhadamente em Cressie (1993).

2.4.1 Modelo Exponencial

O modelo exponencial de semivariograma é valido em qualquer dimensão de $\mathbb{R}^d$, $d \geq 1$ e é descrito pela equação:

$$\gamma(h) = \begin{cases} 0 & \text{se } h = 0; \\ C_0 + C_1 \left[1 - \exp\left(\frac{-3h}{a}\right)\right] & \text{se } h \neq 0. \end{cases} \tag{2.4.1.1}$$

ou, equivalentemente, como:

$$\gamma(h) = \begin{cases} 0 & \text{se } h = 0; \\ C_0 + C_1 \left[1 - \exp\left(\frac{-h}{a}\right)\right] & \text{se } h \neq 0. \end{cases} \tag{2.4.1.2}$$

Teoricamente, os semivariogramas atingem seu patamar a uma distância h finita, denotada como alcance. Na prática, o semivariograma atinge o patamar apenas quando h tende para infinito, ou seja, assintoticamente. Então, para efeito de análise, adota-se como alcance prático a distância correspondente a 95% do patamar (C_1). Assim, o alcance

prático para o modelo é igual a $3a$, quando a Equação (2.4.1.2) é observada.

$$\begin{aligned} \gamma(h) &= \left[1 - \exp\left(\frac{-h}{a}\right)\right] \\ 0.95 &= \left[1 - \exp\left(\frac{-h}{a}\right)\right] \\ 0.05 &= \exp\left(\frac{-h}{a}\right) \\ -3 &\approx \frac{-h}{a} \\ h &\approx 3a \end{aligned} \tag{2.4.1.3}$$

A expressão (2.4.1.3) mostra a definição do alcance prático $A = 3a$. O alcance prático é atingido quando a distância assume o valor equivalete a $3a$ se o semivariograma estiver ajustado conforme a Equação (2.4.1.2). Já para a Equação (2.4.1.1), $A = a$, ou seja, o valor encontrado para a já corrresponde ao alcance prático (Conceição, 2013).

2.4.2 Modelo Gaussiano

O modelo Gaussiano de semivariograma também é valido em qualquer dimensão de $\mathbb{R}^d$, $d \geq 1$. Além disso, seu patamar é atingido assintoticamente, sendo seu alcance prático igual a $A = \sqrt{3}$.

O semivariograma Gaussiano é definido como:

$$\gamma(h) = \begin{cases} 0 & \text{se } h = 0; \\ C_0 + C_1\left[1 - \exp\left(\frac{-h}{a}\right)^2\right] & \text{se } h \neq 0. \end{cases} \tag{2.4.2.1}$$

sendo o alcance prático para este modelo encontrado de maneira análoga ao do semivariograma exponencial. Ao lidar com o semivariograma Gaussiano, é crucial estar atento ao efeito pepita. Quando esse efeito é nulo, o modelo pode apresentar problemas numéricos (Conceição, 2013).

2.4.3 Modelo Seno

Já o semivariograma seno demonstra variações periódicas que diminuem gradualmente à medida que a defasagem aumenta. Este modelo é considerado válido em $\mathbb{R}^d$,

onde $d \leq 3$ e sua equação é dada por (2.4.3.1).

$$\gamma(h) = \begin{cases} 0 & \text{se } h = 0; \\ C_0 + C_1 \left[1 - \frac{\sin\left(\frac{\pi h}{a}\right)}{\frac{\pi h}{a}}\right] & \text{se } h > 0. \end{cases} \tag{2.4.3.1}$$

A periodicidade inerente ao modelo seno pode levar a correlações negativas no processo. Dessa forma, é crucial observar que o valor mínimo que a função pode assumir não deve ser inferior a -0.218, um patamar atingido quando $h \approx 4.5a$.

2.4.4 Modelo Mátern

A partir da função Mátern é possível derivar outros modelos, por exemplo o modelo exponencial quando $\nu = 0.5$, com fórmula dada por:

$$\gamma(h) = \begin{cases} 0 & \text{se } h = 0; \\ C_0 + C_1 \left[1 - \frac{2}{\Gamma(\nu)} \left(\frac{h\sqrt{\nu}}{a}\right)^{\nu} K\nu \left(\frac{2h\sqrt{\nu}}{a}\right)\right] & \text{se } h > 0, \nu > 0. \end{cases} \tag{2.4.4.1}$$

em que K é a função Bessel, $\Gamma(\nu)$ é a função Gama e ν é o parâmetro de suavização.

O semivariograma Mátern é valido em $\mathbb{R}^d$, $d \geq 1$ e pode assumir qualquer tipo de comportamento próximo a origem, assumindo a forma $h^{2\nu}$ se ν não for inteiro e $h^{2\nu}\log(h)$ para ν inteiro.

2.5 Estimação dos parâmetros do semivariograma

2.5.1 Método dos mínimos quadrados ponderados

Uma abordagem para ajustar os parâmetros de um modelo de semivariograma é o método de mínimos quadrados ponderados. Esse método procura uma solução ótima para esses parâmetros, visando minimizar a discrepância entre os valores experimentais e os valores previstos pelo modelo teórico (Cressie, 1993), de modo que a equação a ser minimizada é:

$$\frac{1}{2}\sum_{i=1}^{k} N(h_i)\left[\frac{\hat{\gamma}(h_i)}{\gamma(h_i;\eta)} - 1\right], \tag{2.5.1.1}$$

onde i é o *lag* e η é o conjunto de parâmetros.

As estimativas são obtidas por meio de um processo de otimização, requerendo a

especificação de parâmetros iniciais. Quanto mais próximos esses parâmetros iniciais estiverem dos valores ótimos, mais eficiente será a convergência do método. Uma abordagem para determinar os parâmetros iniciais foi sugerida por Jian et al. (1996):

$$a_{inicial} = \frac{h_k}{2}, \tag{2.5.1.2}$$

$$C_{0inicial} = \max\left(0, \hat{\gamma}(h_1) - \frac{h_1}{h_2 - h_1} - (\hat{\gamma}(h_2) - \hat{\gamma}(h_1))\right), \tag{2.5.1.3}$$

$$C_{1inicial} = \frac{\hat{\gamma}(h_{k-2}) + \hat{\gamma}(h_{k-1}) + \hat{\gamma}(h_k)}{3} - C_{0inicial}, \tag{2.5.1.4}$$

onde h_k indica o k-ésimo *lag*.

Após a conclusão de qualquer procedimento de modelagem, é essencial avaliar a qualidade do ajuste do modelo teórico em comparação com o modelo derivado dos dados. O Critério de Informação de Akaike (AIC), conforme proposto por Cressie (1993) e expresso na Equação (2.5.1.5), é empregado para examinar tal ajuste:

$$AIC_{MQP} = k \log(QM_{Res}) + 2p, \tag{2.5.1.5}$$

em que p é o número de parâmetros, k é o número de classes do semivariograma e QM_{Res} é o quadrado médio dos resíduos, calculado por:

$$QM_{Res} = \sum_{i=1}^{k} \frac{1}{k} \left[\hat{\gamma}(h_i) - \gamma(h_i; \eta)\right]^2. \tag{2.5.1.6}$$

2.5.2 Máxima Verossimilhança

Antes de abordar o método da Máxima Verossimilhança para ajuste dos parâmetros em si, é pertinente estabelecer a relação existente entre a função de covariância e a função de correlação, conforme a seguinte expressão:

$$\rho(h) = \frac{C(h)}{C(0)}. \tag{2.5.2.1}$$

Para expressar a função de covariância por meio do covariograma e a função de correlação através do correlograma, é essencial que a covariância no ponto $h = 0$ seja positiva, como destacado por Cressie (1993). Uma relação adicional importante é aquela que se estabelece entre o covariograma e o variograma, conforme ilustrado na Figura 2.2,

como definido por:

$$\gamma(h) = C(0) - C(h). \tag{2.5.2.2}$$

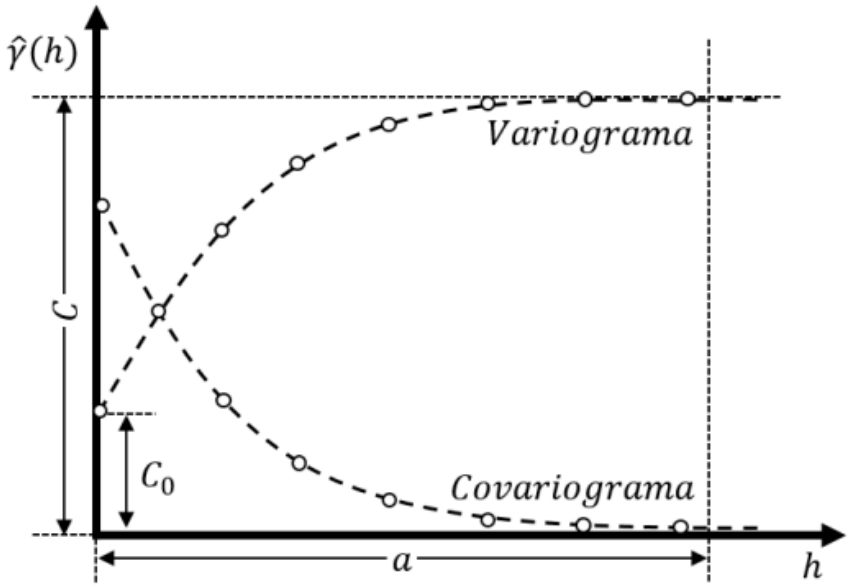


Figura 2.2: Variograma e Covariograma.
Fonte: Câmara et al. (2004)

Uma alternativa para expressar o variograma é através da função de correlação, utilizando as Equações (2.5.2.1) e (2.5.2.2) para obter:

$$\gamma(h) = C(0)[1 - \rho(h)]. \tag{2.5.2.3}$$

Diante disso, o método da máxima verossimilhança é uma alternativa ao método de mínimos quadrados e é amplamente empregado para estimar diversos parâmetros simultaneamente. A função de verossimilhança L é obtida por meio de:

$$L(\eta; z_1, ..., z_n) = \prod_{i=1}^{n} f_Z(z_i; \eta), \tag{2.5.2.4}$$

onde $f_Z(z; \eta)$ é a função de densidade conjunta da normal multivariada e η é o vetor de parâmetros a serem estimados. Ao aplicar o logaritmo e derivar esta função em relação a

η, obtem-se (2.5.2.5) a ser maximizada (Diggle e Ribeiro, 2007):

$$L(\boldsymbol{\beta}, \tau^2, \sigma^2, \phi) = -\frac{1}{2}\left[n\log(2\pi) + \log|\boldsymbol{\Sigma}| + (\boldsymbol{z} - \boldsymbol{\mu}(\boldsymbol{z}))^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}(\boldsymbol{z}))\right], \qquad (2.5.2.5)$$

onde, $\boldsymbol{\mu}(\boldsymbol{z}) = \boldsymbol{D}\boldsymbol{\beta}$ e $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{R}(\phi) + \tau^2\boldsymbol{I}$ com tamanho $n \times n$, obtida pela Equação (2.5.2.1).

Os parâmetros $\boldsymbol{\beta}, \tau^2, \sigma^2, \phi$ representam os coeficientes de regressão, o efeito pepita, a variância amostral e a amplitude das distâncias h, respectivamente. A estimação inicial desses parâmetros pela máxima verossimilhança não requer o uso do variograma, ao contrário do método dos mínimos quadrados, possibilitando seu cálculo direto a partir dos dados amostrados.

Seguindo, a matriz de covariáveis $\boldsymbol{D}$ tem tamanho $n \times (1 + p)$ quando há p variáveis explicativas. Quando não há variáveis explicativas, $\boldsymbol{D}$ é um vetor de números 1 com tamanho n. Por fim, $\boldsymbol{R}(\phi)$ é a função de correlação entre as distâncias e o parâmetro ϕ, definida nos modelos Gaussiano (2.5.2.6), Exponencial (2.5.2.7), Seno (2.5.2.8) e Mátern (2.5.2.9).

Modelo Gaussiano:

$$R(\phi) = \exp\left(-\frac{h^2}{\phi^2}\right); \qquad (2.5.2.6)$$

Modelo Exponencial:

$$R(\phi) = \exp\left(-\frac{h}{\phi}\right); \qquad (2.5.2.7)$$

Modelo Seno:

$$R(\phi) = \left(\frac{\phi}{\pi h}\right)\sin\left(\frac{\pi h}{\phi}\right); \qquad (2.5.2.8)$$

Modelo Mátern:

$$R(\phi, k) = (2^{k-1}\Gamma(k))^{-1}\left(\frac{h}{\phi}\right)^k K_k\left(\frac{h}{\phi}\right). \qquad (2.5.2.9)$$

Após realizar a estimativa inicial dos parâmetros, será empregado um método de otimização para maximizar a função de verossimilhança, com o objetivo de identificar o conjunto de parâmetros η que apresenta a maior probabilidade de gerar os dados observados.

Por fim, a avaliação da qualidade do ajuste é realizada por meio do Critério de

Informação descrito por Akaike (1998), expresso por:

$$AIC_{MV} = -2\log(L_\eta) + 2p, \tag{2.5.2.10}$$

no qual L_η é a função de verossimilhança maximizada e p é o número de parâmetros, incluindo a variância.

Quando o tamanho da amostra é pequeno em relação à quantidade de parâmetros estimados, isto é, $\frac{n}{p} < 40$, emprega-se a metodologia de correção de viés descrita por Hurvich e Tsai (1989):

$$AIC_{MVc} = AIC_{MV} + \frac{2(p+1)(p+2)}{n-p-2}. \tag{2.5.2.11}$$

Dessa forma, a fórmula ajustada leva em conta o tamanho da amostra n e o número de parâmetros p na avaliação do ajuste do modelo.

2.6 Krigagem

O termo "Krigagem" tem sua origem no nome de Daniel G. Krige, o qual foi o precursor na introdução do uso de médias móveis, visando mitigar a superestimação sistemática de reservas minerais (Câmara et al., 2004). Posteriormente, o método foi aprimorado pelo matemático francês Georges Matheron (Bailey e Gatrell, 1995). A aplicação dessa técnica requer os passos:

(a) Análise exploratória dos dados;

(b) Análise estrutural (modelagem da estrutura de correlação espacial);

(c) Interpolação estatística da superfície.

Conforme já mencionado, o diferencial fundamental da Krigagem reside na sua capacidade de realizar a estimação de uma matriz de covariância espacial, que desempenha um papel central na atribuição de pesos às diferentes amostras, ao tratamento da redundância dos dados, à vizinhança a ser considerada no procedimento inferencial e ao erro associado ao valor estimado (Câmara et al., 2004).

A estrutura teórica da técnica é fundamentada no conceito de variável regionalizada, que é uma variável distribuída no espaço ou no tempo, cujos valores são tratados como realizações de uma função aleatória. Esse conceito possibilita a incorporação de processos espaciais locais (Câmara et al., 2004). A variação espacial de uma variável

regionalizada pode ser descrita como a soma de três componentes distintas: (a) uma componente estrutural, relacionada a um valor médio constante ou a uma tendência constante; (b) uma componente aleatória espacialmente correlacionada; e (c) um resíduo aleatório ou erro residual. Considerando um vetor u representando uma posição, o valor da função aleatória Z em u é definido como segue:

$$Z(u) = \mu(u) + \varepsilon'(u) + \varepsilon''. \tag{2.6.0.1}$$

onde $\mu(u)$ é uma função deterministica que descreve a componente estrutural de Z, $\varepsilon'(u)$ é um termo de variação espacial, e ε'' é um resíduo aleatório ou erro residual com distribuição normal com média zero e variância σ^2 (Câmara et al., 2004).

Ainda, nesse caso, o conceito de contiguidade não está implicitamente definido, então, para caturar a componente de variação espacial, uma abordagem mais natural é a criação de uma matriz $\boldsymbol{D}$ de distâncias entre os pontos, sendo $d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ a distância entre os pontos i e j (Fotheringham et al., 2000). É importante destacar que a Krigagem disponibiliza estimadores que exibem as propriedades de não viés e eficiência, fornecendo estimativas não tendenciosas e com variância mínima (Câmara et al., 2004).

Nesse sentido, a estimação e predição de superfícies pela Krigagem engloba uma variedade de técnicas. Essas técnicas incluem a krigagem ordinária, amplamente adotada na prática, bem como a krigagem simples, krigagem lognormal, krigagem universal, krigagem fatorial, cokrigagem ordinária (krigagem ordinária para duas ou mais variáveis), krigagem indicatriz, krigagem disjuntiva e krigagem probabilística (Cressie, 1993), dentre outras. Este estudo concentra-se na aplicação da Krigagem ordinária, deixando de abordar outras técnicas correlatas.

2.6.1 Krigagem ordinária

A hipótese mais simples na análise de variáveis regionais é que a média, representada por $\mu(u)$, permanece constante na área de estudo, o que implica em uma falta de variação significativa em larga escala (Câmara et al., 2004). Essa premissa forma a base para a Krigagem ordinária, uma técnica de interpolação que assume que a média é constante. Isso significa que a diferença esperada entre valores observados em duas posições, u e $u+h$, separadas por uma distância h, é nula. A Krigagem ordinária é útil quando não há uma tendência espacial substancial na variável regionalizada (Câmara et al., 2004).

Seguindo a hipotese da Krigagem ordinaria, $\mu(u)$ é constante e denominada m.

Então, o valor esperado de Z entre as posições u e $u+h$ é:

$$E\left[Z(u) - Z(u+h)\right] = 0. \tag{2.6.1.1}$$

Ainda, como o fonômeno é estacionário de segunda ordem, a covariância entre os valores $Z(u)$ e $Z(u+h)$ permanece constante (Equação 2.6.1.2). Como resultado direto da estacionariedade da covariância, a variância também se mantém constante em toda a região de estudo (Equação 2.6.1.3).

$$C(h) = Cov\left[Z(u), Z(u+h)\right] = E\left[Z(u)Z(u+h)\right] - m^2. \tag{2.6.1.2}$$

$$\begin{aligned} Var\left(Z(u)\right) &= E\left[Z^2(u)\right] - 2m^2 + m^2. \\ &= E\left[Z^2(u)\right] - m^2 = C(0). \end{aligned} \tag{2.6.1.3}$$

Dessa maneira, sob as premissas de média constante e estacionariedade da covariância, a caracterização da variável regionalizada pode ser plenamente alcançada mediante a determinação da função $C(h)$, que é obtida por meio da Equação (2.3.0.1).

Desenvolvendo a soma de quadrados de (2.3.0.1) , pode-se chegar em:

$$\begin{aligned} &2\gamma(h) = E\left[Z(u) - Z(u+h)\right]^2. \\ &2\gamma(h) = 2C(0) - 2C(h) \quad \text{ou} \quad \gamma(h) = C(0) - C(h). \end{aligned} \tag{2.6.1.4}$$

A Equação (2.6.1.4) mostra que a covariância e o semivariograma representam formas alternativas de descrever a autocorrelação entre os pares $Z(u)$ e $Z(u+h)$, separados pelo vetor de distâncias h (Câmara et al., 2004). Nesse contexto, o semivariograma possibilita uma representação quantitativa da variação de um fenômeno regionalizado no espaço.

Agora, retomando o conceito da superfície na qual a variável Z representa um atributo observado em n pontos distintos, cujas coordenadas são expressas pelo vetor u, o estimador geral para o valor de Z em um ponto u_0 desconhecido é dado por:

$$\hat{Z}(u_0) = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(u_i), \tag{2.6.1.5}$$

onde λ_i são os pesos determinados para minimizar o erro das estimativas. Dessa forma, considerando um processo estocástico E como um processo estacionário com função de

semivariograma $\hat{\gamma}$, então qualquer valor de Z em um ponto u_0 pode ser predito por meio de uma combinação linear dos n valores observados.

Na Krigagem ordinária, para que a média seja constante, $\lambda_0 = 0$ e $\sum_{i=1}^{n} \lambda_i = 1$. Então, o estimador será:

$$\hat{Z}(u_0) = \sum_{i=1}^{n} \lambda_i Z(u_i). \tag{2.6.1.6}$$

Sendo $\hat{Z}(u_0)$ um estimador não tendencioso de $Z(u_0)$, então:

$$E\left[\hat{Z}(u_0) - Z(u_0)\right] = 0. \tag{2.6.1.7}$$

Para minimizar a variância do erro $Var(\hat{Z}(u_0) - Z(u_0)$ sob $\sum_{i=1}^{n} \lambda_i = 1$ é necessário minimizar:

$$E\left(\sum_{i=1}^{n} \lambda_i Z(u_i) - Z(u_0)\right)^2 - 2\alpha\left(\sum_{i=1}^{n} \lambda_i - 1\right), \tag{2.6.1.8}$$

utilizando o semivariograma definido em (2.3.0.1). Nesse caso, α é o multiplicador de Lagrange que garante $\sum_{i=1}^{n} \lambda_i = 1$ e desta condição obtem-se:

$$\left(\sum_{i=1}^{n} \lambda_i Z(u_i) - Z(u_0)\right)^2 = -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \left[Z(u_i) - Z(u_j)\right]^2 + 2\sum_{i=1}^{n} \lambda_i \left[Z(u_0) - Z(u_i)\right]^2. \tag{2.6.1.9}$$

Substituindo (2.6.1.9) em (2.6.1.8), tem-se:

$$-\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(u_i, u_j) + \sum_{i=1}^{n} \lambda_i \gamma(u_i, u_0) - 2\alpha(\sum_{i=1}^{n} \lambda_i - 1). \tag{2.6.1.10}$$

Derivando (2.6.1.10) em relação a $\lambda_i, ..., \lambda_n$ e igualando a zero, obtém-se:

$$-\sum_{j=1}^{n} \lambda_j \gamma(u_i, u_j) + \gamma(u_i, u_0) - \alpha = 0. \tag{2.6.1.11}$$

Os pesos γ_i são obtidos por meio de um sistema de equações, denominado Sistema

de Krigagem Ordinária:

$$\begin{cases} -\sum_{j=1}^{n} \lambda_j \gamma(u_i, u_j) + \gamma(u_i, u_0) - \alpha = 0 \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{2.6.1.12}$$

onde $\gamma(u_i, u_j)$ é a semivariância entre os pontos u_i e u_j e $\gamma(u_i, u_0)$ é a semivariância entre o i-ésimo ponto e o ponto u_0. Já o α é o multiplicador de Lagrange.

A variância da Krigagem ordinária é estimada por:

$$Var[\hat{Z}(u_0)] = E\left[\hat{Z}(u_0) - Z(u_0)^2\right] = 2\sum_{i=1}^{n} \lambda_i \gamma(u_i, u_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(u_i, u_j). \tag{2.6.1.13}$$

Cada estimativa possui uma variância associada de krigagem, que pode ser representada por $\sigma^2(u_0)$ e é definida pela Equação (2.6.1.13). Sendo assim, os pesos λ_i obtidos pelo sistema de krigagem ordinária são substituídos em (2.6.1.6), por meio da qual a estimativa da variância é obtida:

$$\sigma^2(u_0) = \sum_{i=1}^{n} \lambda_i \gamma(u_i, u_0) + \alpha(u_0). \tag{2.6.1.14}$$

Matricialmente, pode-se definir as matrizes $\mathbf{C}_+$, $\boldsymbol{\lambda}_+(u)$ e $\mathbf{C}_+(u)$ como:

$$\mathbf{C}_+ = \begin{bmatrix} C(u_1, u_1) & \dots & C(u_1, u_n) & 1 \\ \vdots & \vdots & \ddots & \vdots \\ C(u_n, u_1) & \dots & C(u_n, u_n) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \quad \boldsymbol{\lambda}_+(u) = \begin{bmatrix} \lambda_1(u) \\ \vdots \\ \lambda_n(u) \\ \alpha(u) \end{bmatrix} \quad \mathbf{C}_+(u) = \begin{bmatrix} C(u, u_1) \\ \vdots \\ C(u, u_n) \\ 1 \end{bmatrix}$$

Tendo como solução:

$$\boldsymbol{\lambda}_+(u) = \mathbf{C}_+^{-1}\mathbf{C}_+(u). \tag{2.6.1.15}$$

e variância:

$$\sigma_k^2(u) = \sigma^2 - \mathbf{C}_+{}^{\top}(u)\mathbf{C}_+{}^{-1}\mathbf{C}_+(u). \tag{2.6.1.16}$$

onde $\mathbf{C}_+^{-1}$ é a matriz de covariâncias e $\mathbf{C}_+(u)$ é a matriz de covariâncias entre os pontos u, ponto a ser estimado e u_i, ponto amostrado.

A krigagem ordinária representa um interpolador exato, significando que, ao empregar as equações mencionadas, os valores interpolados coincidirão exatamente com os valores dos pontos amostrais (Câmara et al., 2004).

2.6.2 Demonstração (2×2) do estimador exato da Krigagem - Sem perda de generalidade

Ao utilizar o modelo exponencial de semivariograma, por exemplo, observa-se que:

$$\mathbf{C} = \begin{bmatrix} \sigma^2 & \sigma^2 e^{\frac{-3|h|}{a}} \\ \sigma^2 e^{\frac{-3|h|}{a}} & \sigma^2 \end{bmatrix}$$

para determinar a matriz inversa $\boldsymbol{C}^{-1}$, é necessário calcular o determinate da matriz $\boldsymbol{C}$ conforme:

$$\boldsymbol{D} = (\sigma^2)^2 - (\sigma^2)^2 e^{-2\frac{-3|h|}{a}} = (\sigma^2)^2(1 - e^{-2\frac{-3|h|}{a}}).$$

Dessa forma, a matriz $\boldsymbol{C}^{-1}$ é dada por:

$$\mathbf{C}^{-1} = \begin{bmatrix} \frac{\sigma^2}{(\sigma^2)^2(1-e^{-2\frac{3|h|}{a}})} & \frac{-\sigma^2 e^{\frac{-3|h|}{a}}}{(\sigma^2)^2(1-e^{-2\frac{3|h|}{a}})} \\ \frac{-\sigma^2 e^{\frac{-3|h|}{a}}}{(\sigma^2)^2(1-e^{-2\frac{3|h|}{a}})} & \frac{\sigma^2}{(\sigma^2)^2(1-e^{-2\frac{3|h|}{a}})} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2(1-e^{-2\frac{3|h|}{a}})} & -\frac{e^{\frac{-3|h|}{a}}}{\sigma^2(1-e^{-2\frac{3|h|}{a}})} \\ -\frac{e^{\frac{-3|h|}{a}}}{\sigma^2(1-e^{-2\frac{3|h|}{a}})} & \frac{1}{\sigma^2(1-e^{-2\frac{3|h|}{a}})} \end{bmatrix}$$

como $\mathbf{c} = \begin{bmatrix} \sigma^2 \\ \sigma^2 e^{\frac{-3|h|}{a}} \end{bmatrix}$, então:

$$\lambda = \mathbf{C^{-1}c} = \begin{bmatrix} \\ \frac{\sigma^2-(\sigma^2 e^{-2\frac{3|h|}{a}})}{\sigma^2(1-e^{-2\frac{3|h|}{a}})} \\ \\ \frac{-\sigma^2 e^{\frac{-3|h|}{a}}+\sigma^2 e^{\frac{-3|h|}{a}}}{\sigma^2(1-e^{-2\frac{3|h|}{a}})} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \frac{1-e^{-2\frac{3|h|}{a}}}{1-e^{-2\frac{3|h|}{a}}} \\ \\ \frac{-e^{\frac{-3|h|}{a}}+e^{\frac{-3|h|}{a}}}{1-e^{-2\frac{3|h|}{a}}} \\ \\ \end{bmatrix} = \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix}$$

Seguindo para o calculo da variância é possível observar que:

$$\sigma_k^2(u) = \sigma^2 - \mathbf{c}'\lambda = \sigma^2 - \begin{bmatrix} \sigma^2 & \sigma^2 e^{\frac{-3|h|}{a}} \end{bmatrix} \begin{bmatrix} 1 \\ \\ 0 \end{bmatrix} \tag{2.6.2.1}$$

$$= \sigma^2 - \sigma^2 = 0.$$

Desse modo, observa-se que quando um ponto amostrado coincide exatamente com um ponto da grade regular, o valor estimado é exatamente igual ao valor amostrado e a variância desse valor estimado é igual a zero.

2.6.3 Limitações das Funções de Semivariograma

A escolha da função para o covariograma ou semivariograma não pode ser feita de maneira aleatória. Essas medidas de variabilidade influenciam as equações do sistema de Krigagem ordinária, que determinam os pesos de cada observação na interpolação. Portanto, é essencial garantir que a matriz de covariâncias seja definida como positiva para garantir uma solução única no sistema de krigagem ordinária (Isaaks e Srivastava, 1989).

Então, admitindo a hipótese de estacionaridade de segunda ordem, a covariância $C(0)$ deve satizfazer (2.6.3.1), cuja expressão representa a variância dos erros de predição (Conceição, 2013).

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(0) \geq 0, \tag{2.6.3.1}$$

onde λ_i e λ_i representam os pesos. Efetivamente, ao analisar as variâncias das combinações

lineares $\sum_{i=1}^{n} Z(u_i)$, resulta na obtenção da (2.6.3.1), onde $\lambda \in \mathbb{R}$ e $u_i \in \mathbb{R}$ para todo $n \in \mathbb{N}$ (Schlather et al., 2012). Então, a condição de ser definida positiva imposta à função covariância é tanto necessária quanto suficiente (Schlather, 1999).

Na presença de um processo intrinsecamente estacionário, a dependência espacial não pode ser mensurada utilizando a função covariância. Como alternativa, a variabilidade espacial é avaliada pela função semivariograma, que deve atender à condição expressa em (2.6.3.2). Nesse contexto, $-\lambda(h)$ é condicionalmente não negativo quando a soma dos pesos é nula:

$$-\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma(h) \geq 0. \tag{2.6.3.2}$$

Os incrementos $Z(u) - Z(0)$ de um processo intrinsecamente estacionário apresentam esperança nula apenas quando a condição (2.6.3.3) é atendida.

$$\lim_{r \to \infty} \frac{\gamma(h)}{h^2} = 0. \tag{2.6.3.3}$$

No caso de um processo totalmente aleatório, é necessário que h^2 cresça a uma taxa superior à do semivariograma $\gamma(h)$.

Então, a construção de modelos de semivariograma necessita de alguns cuidados importantes. Entendendo as funções densidade de probabilidades (fdp) como medidas positivas, Christakos (1984) utilizou este fato como argumento para construir famílias de modelos isotrópicos de semivariogramas, se amparando nas propriedades que definem uma fdp. Mas nem toda fdp fundamenta a construção de um semvariograma adequado, conforme abordado por Conceição (2013), de modo que a construção dos modelos partindo de premissas equivocadas, resulta em funções inadequadas. A Figura 2.3 apresenta as curvas de alguns modelos de semivariograma.

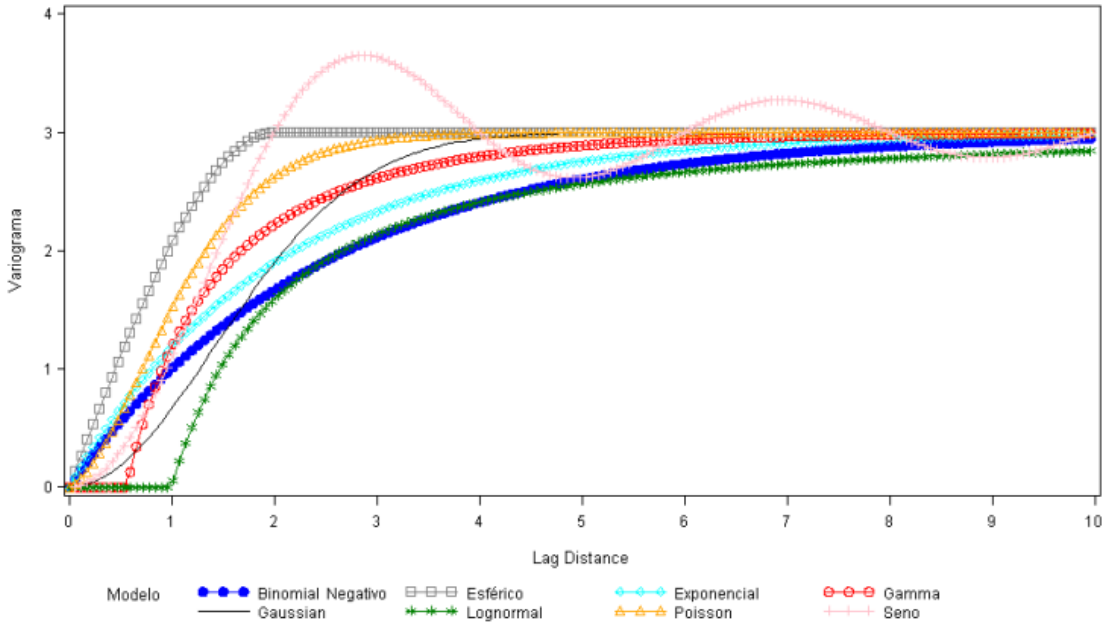


Figura 2.3: Curvas dos modelos clássicos de semivariograma e dos propostos por Conceição (2013).
Fonte: Conceição (2013)

2.7 Regressão Geograficamente Ponderada

A Regressão Geograficamente Ponderada (RGP) representa uma abordagem especializada para a análise espacial local de dados multivariados. Uma de suas vantagens notáveis é sua fundamentação na estrutura da regressão tradicional. Além disso, destaca-se por incorporar de maneira intuitiva e explícita as dependências espaciais locais à estrutura de regressão, proporcionando uma análise mais refinada e contextualizada (Fotheringham et al., 2002). É importante notar que a regressão clássica é um caso especial da RGP, sendo que a RGP retorna para uma regressão clássica quando não há dependência espacial nos dados.

Nesse sentido, os pressupostos do modelo da regressão clássica ainda precisam ser atendidos para a RGP sendo, erros normais, homocedásticos e independentes. Logo, esse modelo ainda tem limitações para tratar dados especiais de contagem. Para casos de contagem, os modelos mais adequados são Regressão de Poisson Geograficamente Ponderada (RPGP) desenvolvida por Nakaya et al. (2005) ou Regressão Binomial Negativa Geograficamente Ponderada (RBNGP) desenvolvida por Da Silva e Rodrigues (2014).

O modelo RGP é representado como:

$$y_j = \beta_0(u_j, v_j) + \sum_k \beta_k(u_j, v_j) x_{jk} + \varepsilon_j, \quad \varepsilon_j \sim N(0, \sigma^2). \tag{2.7.0.1}$$

Ao realizar a regressão para estimar o parâmetro $\boldsymbol{\beta}$ no ponto i, utiliza-se as informações dos pontos $j\{x_{ik}, y_i\}$ sem a necessidade de ter os dados do ponto i. Levando em consideração que pontos mais próximos são mais similares, pondera-se as informações dos pontos observados j com pesos que diminuem à medida que eles se afastam de i.

Em sua forma matricial, (2.7.0.1) pode ser expressa como:

$$\mathbf{y} = (\boldsymbol{\beta}_k(u_j, v_j) \otimes \mathbf{X})\mathbf{1} + \boldsymbol{\varepsilon}, \tag{2.7.0.2}$$

onde $\otimes$ é o operador de Kronecker, que denota a multiplicação de Kronecker. Considerando uma amostra de tamanho n e o número de variáveis explicativas igual a k, então $\mathbf{X}$ é a matriz do modelo com dimensão $(n \times k+1)$, $\mathbf{1}$ é um vetor de 1's de dimensão $(k+1)$, e $\boldsymbol{\beta}_k(u_j, v_j)$ é um vetor de parâmetros com dimensão $(n \times k+1)$, cuja linha j é formada pela estimativa dos $(k+1)$ parâmetros para o ponto j da amostra, ou seja,

$$\begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \dots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \dots & \beta_k(u_2, v_2) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \dots & \beta_k(u_n, v_n) \end{bmatrix} = \boldsymbol{\beta}_k(u_j, v_j)$$

sendo,

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = [\mathbf{X}^\top \mathbf{W}(u_i, v_i)\mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{W}(u_i, v_i)\mathbf{y}, \tag{2.7.0.3}$$

onde $\boldsymbol{\beta}(u_i, v_i)$ representa a estimativa do vetor de parâmetros $\boldsymbol{\beta}$ no ponto de regressão (u_i, v_i), e $\boldsymbol{W}(u_i, v_i)$ é uma matriz $n \times n$. Os elementos fora da diagonal são iguais a zero, e os elementos na diagonal, denotados por w_{ij}, onde $j = 1, \dots, n$, indicam o peso da j-ésima observação no ponto de regressão i. Com N sendo o número total de pontos de regressão, a variável i varia de 1 até N. Em resumo, a técnica RGP realiza um conjunto de regressões equivalente ao número de pontos desejados, N, para a estimativa.

Denotando $\hat{\boldsymbol{\beta}}(u_i, v_i)$ por $\hat{\boldsymbol{\beta}}(i)$, e $\mathbf{W}(u_i, v_i)$ por $\mathbf{W}(i)$, a Equação (2.7.0.3) pode ser

reescrita como:

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}^\top \mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}^\top \mathbf{W}(i)\mathbf{y}, \tag{2.7.0.4}$$

onde:

$$\begin{bmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{in} \end{bmatrix} = \mathbf{W}(i)$$

A matriz de pesos $\mathbf{W}(i)$ deve ser calculada para cada ponto i. Nessa notação, considera-se (u_i, v_i) um ponto arbitrário no espaço, para o qual será feita a estimativa dos parâmetros. Por ser um ponto arbitrário, (u_i, v_i) não é necessariamente equivalente a um ponto onde há o dado observado, denotado como j. Assim, considera-se que o parâmetro no ponto i é igual aos parâmetros dos pontos j próximos de i.

A Equação (2.7.0.4) pode ser utilizada para estimar o erro padrão das estimativas locais no modelo RGP. Definindo a matriz C como:

$$\mathbf{C} = \left[\mathbf{X}^\top \mathbf{W}(i)\mathbf{X}\right]^{-1}\mathbf{X}^\top \mathbf{W}(i)\big], \tag{2.7.0.5}$$

tem-se:

$$\begin{aligned} \hat{\boldsymbol{\beta}}(i) &= \mathbf{C}y, \\ \widehat{Var}\left[\hat{\boldsymbol{\beta}}(i)\right] &= \mathbf{C}\mathbf{C}^\top \hat{\sigma}^2, \end{aligned} \tag{2.7.0.6}$$

onde $\hat{\sigma}^2$ é a soma dos quadrados dos resíduos normalizados da regressão local, sendo:

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n - 2v_1 + v_2}, \tag{2.7.0.7}$$

em que $v_1 = tr(\mathbf{R})$, $v_2 = tr(\mathbf{R}^\top \mathbf{R})$ e $tr()$ é o traço da matriz.

A matriz $\mathbf{R}$ relaciona as matrizes $\hat{\boldsymbol{\mu}}$ e $\mathbf{y}$, cujas linhas $\mathbf{r}_j$ são definidas como:

$$\mathbf{r}_j = \mathbf{X}_j[\mathbf{X}^\top \mathbf{W}(j)\mathbf{X}]^{-1}\mathbf{X}^\top \mathbf{W}(j), \tag{2.7.0.8}$$

onde $\mathbf{X}_j$ é a j-ésima linha da matriz do modelo $\mathbf{X}$.

2.7.1 Função de ponderação Espacial

Os pesos W_{ij} da matriz $\mathbf{W}(i)$ são determinados de acordo com a função de ponderação espacial, conforme a Figura (2.4).

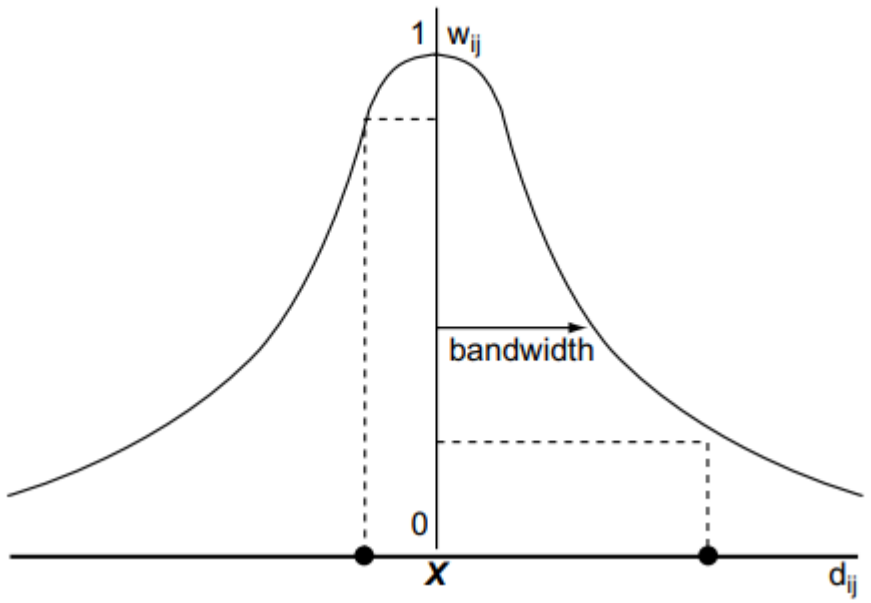


Figura 2.4: Função de ponderação espacial, onde **X** é ponto de regressão e ● são os pontos amostrais. Fonte: Fotheringham et al. (2002)

Algumas funções de ponderação descritas por Fotheringham et al. (2002) são:

- $w_{ij} = 1$ se $d_{ij} < d$, e $w_{ij} = 0$ caso contrário;
- $w_{ij} = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right]$;
- $w_{ij} = \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2$, se $d_{ij} < b$, e $w_{ij} = 0$ caso contrário.

onde d_{ij} é a distância do ponto i para a observação j, d é uma distância pré-determinada e b é o parâmetro de suavização, ou *bandwidth*, que controla a variância da função de ponderação e determina a velocidade de decaimento do peso com relação a distância. Esses exemplos representam apenas algumas das muitas funções de ponderação disponíveis, e há diversos outros tipos que devem ser adaptados conforme a natureza dos dados em análise. É importante destacar que os resultados da RGP são bastante sensíveis a escolha do parâmetro de suavização.

2.7.2 Parâmetro de suavização - Bandwidth

Um dos métodos para a determinação do parâmetro de suavização é o método de validação cruzada, conforme descrito por Cleveland (1979) na regressão local, sendo este representado pela seguinte equação:

$$CV = \sum_{j=1}^{n} \left[y_j - \hat{y}_{\neq j}(b) \right]^2 . \tag{2.7.2.1}$$

Na Equação (2.7.2.1), $\hat{y}_{\neq j}(b)$ representa o valor ajustado para o ponto j, sem considerar a própria observação j no ajuste, sendo b o valor do parâmetro de suavização ótimo que minimiza (2.7.2.1).

2.7.3 Criação de superfícies com a RGP

Agora, relembrando o modelo clássico de regressão linear simples, tem-se que:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{2.7.3.1}$$

onde, p é o número de covariáveis adicionadas ao modelo.

A sua forma matricial é escrita como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{2.7.3.2}$$

Quando se busca estimar o atributo em pontos não amostrados, desconsiderando a influência de covariáveis, o objetivo consiste em trabalhar com a média do modelo de regressão, representada, neste caso, pelo intercepto β_0. Para tal, a matriz de covariáveis $\mathbf{X}$ pode ser substituida por um vetor de 1's e, dessa forma:

$$\hat{\mathbf{Y}} = E(\mathbf{Y}|\mathbf{X}) = \hat{\boldsymbol{\beta}}_0. \tag{2.7.3.3}$$

Essa abordagem pode ser estendida para a Regressão Geograficamente Ponderada (RGP), de modo que, com base na Equação (2.7.0.3), a média de cada ponto de regressão (u_i, v_i), representada pelo valor de $\beta_0(u_i, v_i)$, dependerá exclusivamente da matriz de pesos $\mathbf{W}(i)$, que indica a ponderação dos pontos vizinhos amostrados. Desta forma, serão con-

duzidas tantas regressões quantos pontos de regressão forem definidos. Assim, a geração de superfícies torna-se possível por meio da modelagem utilizando RGP.

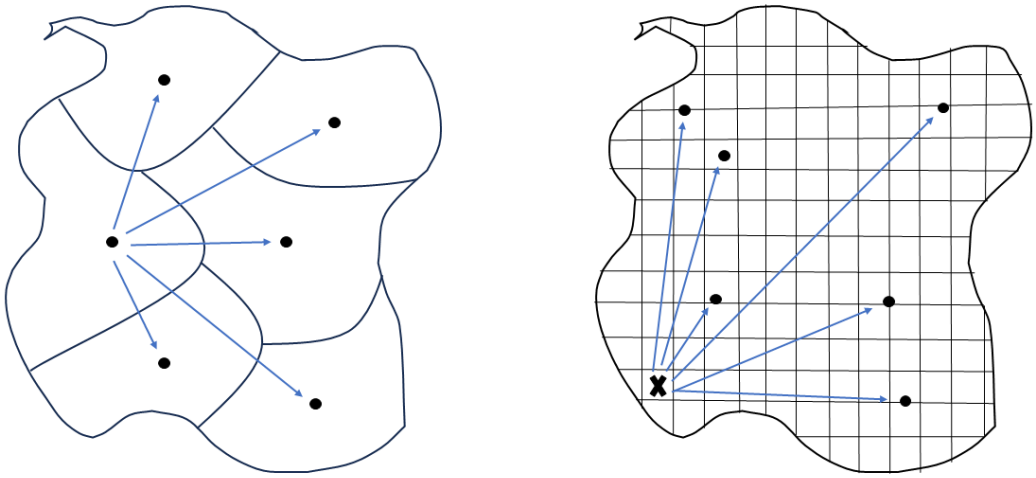


Figura 2.5: Determinação da matriz de pesos $W_{(i)}$ para áreas (à esquerda) e para a criação de superfícies (à direita).

No caso da RGP, a matriz $\mathbf{W}(i)$ para gerar as superfícies segue o mesmo princípio que a análise de áreas, conforme evidenciado na Figura 2.5. Na análise de áreas, à esquerda da figura, a matriz de pesos $\mathbf{W}(i)$ é determinada pela distância, indicada pelas setas azuis, do centróide de cada um dos polígonos em relação aos centróides dos demais polígonos que compõem o mapa, representados pelos círculos pretos. No contexto das superfícies, à direita da figura, o princípio é análogo, porém cada ponto central de cada quadrado da grade regular atua como um ponto de regressão, denotado por $\mathbf{X}$, e a matriz de pesos $\mathbf{W}(i)$ é definida pela distância entre cada um desses pontos de regressão e os pontos amostrados, também representados pelos círculos pretos.

Ainda, considerando que a variância é calculada por meio da Equação (2.7.0.7) e que os pontos y_j não são conhecidos, a variância poderá ser estimada por meio dos pontos amostrados.

Além disso, é plausível considerar que a superfície de parâmetros $\boldsymbol{\beta}(u, v)$ é aproximadamente plana nos locais próximos ao ponto de regressão (Figura 2.6).

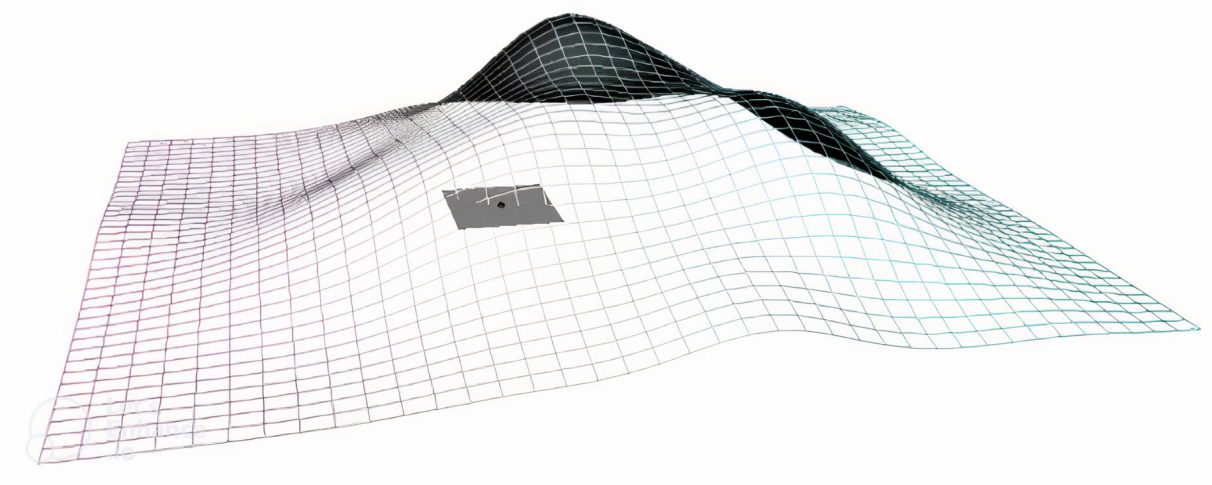


Figura 2.6: A superfície de $\beta_0(u_i, v_i)$ é assumida como plana no ponto de regressão i.
Fonte: Da Silva e Rodrigues (2014)

Considerando a substituição das covariáveis por um vetor de 1's, nota-se que a estrutura da Krigagem apresenta semelhanças com a da RGP. Para melhor visualização dessa analogia, retoma-se o que foi apresentado na seção 2.6.1. Nessa seção, é apresentada a Equação (2.6.1.15), onde $\boldsymbol{\lambda}(u)$ representa um vetor $(n \times 1)$, $\mathbf{C}_+^{-1}$ corresponde a uma matriz $(n \times n)$, e $\mathbf{C}_+(u)$ é um vetor $(n \times 1)$. Nesse contexto, a matriz $\mathbf{C}_+^{-1}$ na Krigagem desempenha um papel análogo à matriz $(\mathbf{X}^\top \mathbf{W}(j)\mathbf{X})^{-1}$ na RGP, enquanto a matriz $\mathbf{C}_+(u)$ corresponde à $\mathbf{X}^\top \boldsymbol{W}(j)\mathbf{y}$ na RGP.

Diante dessa consideração, a aplicação de RGP surge como uma alternativa para a geração de superfícies quando os requisitos essenciais para a utilização do semivariograma não são satisfeitos. Dessa maneira, a RGP tem a capacidade de produzir superfícies que se assemelham às superfícies produzidas pela Krigagem, oferecendo uma abordagem menos restritiva em termos de critérios metodológicos.

Capítulo 3

Materiais e Métodos

3.1 Introdução

Neste Capítulo, serão delineados os materiais e procedimentos metodológicos utilizados no presente estudo. Os dados empregados têm origem na Pesquisa Origem-Destino (OD) da Região Metropolitana de São Paulo, realizada em 2007 pela Companhia do Metropolitano de São Paulo – Metrô/SP (Metrô - São Paulo, 2007). As análises serão conduzidas por meio de algoritmos implementados no *software* SAS 9.4.

3.2 Materiais

A Pesquisa Origem e Destino, conhecida como Pesquisa O/D, tem sido conduzida na Região Metropolitana de São Paulo (RMSP) desde 1967, com uma periodicidade de dez anos. Seu propósito central é coletar informações atualizadas sobre as viagens realizadas pela população metropolitana em dias úteis típicos. Essa pesquisa desempenha um papel fundamental como o principal meio de obtenção de dados sobre deslocamentos, constituindo-se como a base essencial para estudos de planejamento de transporte na região.

A pesquisa de 2007 abrangeu 469 zonas de tráfego, com uma amostra composta por aproximadamente 30.000 domicílios, da região da grande São Paulo, com um banco de dados composto por 124 variáveis.

3.2.1 Tratamento das variáveis

A partir da região da grande São Paulo, foi selecionada uma área de estudo referente a região central do município de São Paulo, de acordo com o estudo desenvolvido por Rocha et al. (2019). As variáveis selecionadas para a investigação de dados relacionados à demanda por transporte compreenderam o percentual de viagens domiciliares realizadas por automóvel (calculado como a razão entre o número de viagens por automóvel e o total de viagens realizadas por domicílio) e a renda média familiar (expressa em salários mínimos).

3.2.1.1 Renda Familiar

A variável "Renda familiar" foi processada de modo que, nos casos em que dois ou mais domicílios compartilhavam as mesmas coordenadas geográficas, calculou-se a média aritmética dos valores de renda correspondentes. Dessa forma, foi atribuída uma consideração exclusiva a uma única ocorrência de $Z(u)$ em uma determinada localização espacial. Posteriormente, esse processo resultou em um banco de dados abrangendo a área em análise, contendo um total de 8.498 domicílios.

3.2.1.2 Viagens por automóvel

Antes de iniciar o tratamento dessa variável, em concordância com o procedimento delineado por Rocha et al. (2019), foi estabelecida uma grade regular de 85 metros. Ou seja, para a variável "Viagens por automóvel", foi adotado um método de *Buffer* como parte do processo metodológico. O cálculo do índice de viagens residenciais executadas por automóvel foi obtido através da divisão do total de viagens automotivas pelo número total de deslocamentos feitos por domicílio. Em seguida, a média desses percentuais foi calculada para todos os lares dentro de uma célula da grade regular e o resultado obtido foi então associado às coordenadas do centróide de cada uma das quadrículas. Consequentemente, essa etapa resultou na construção de um banco de dados referente à área em análise, contendo um total de 3.859 pontos. A Figura 3.1 ilustra o processo de agregação previamente descrito.

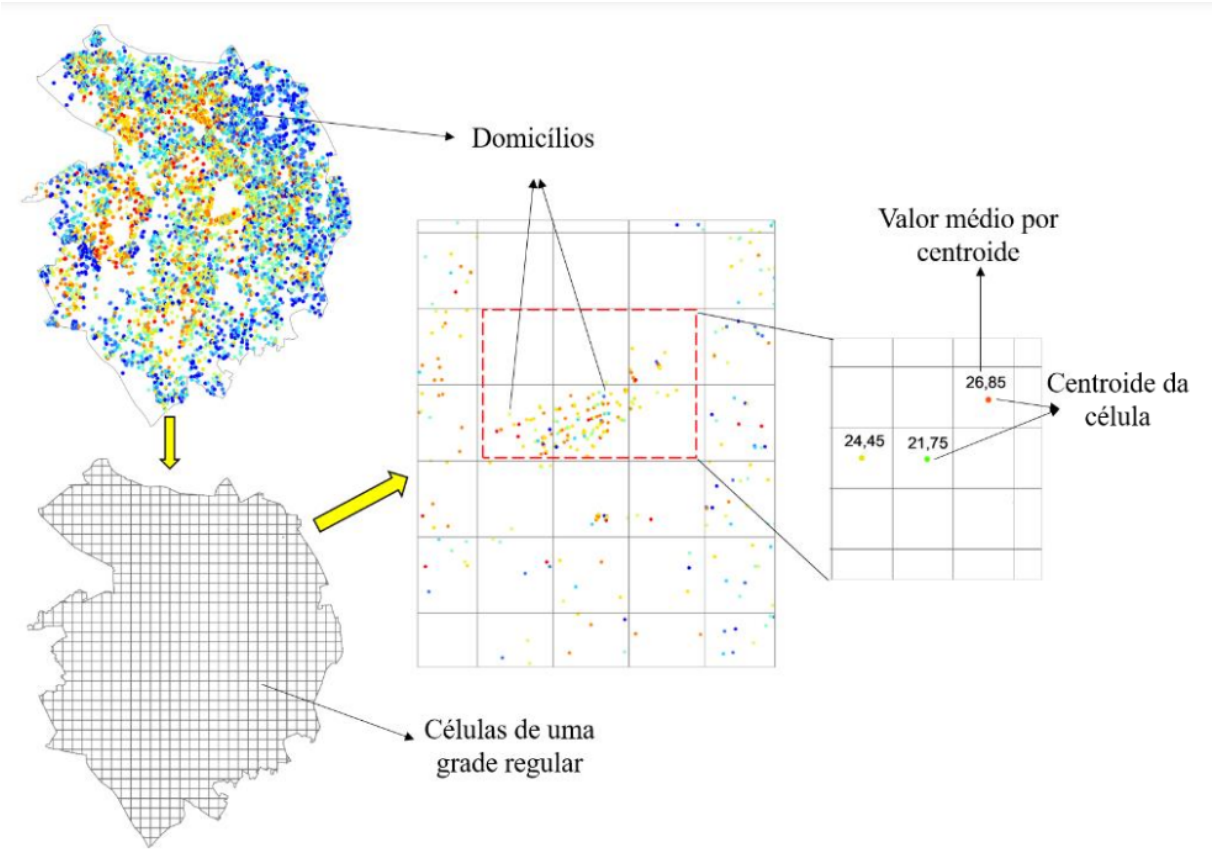


Figura 3.1: Exemplo da agregação de dados em células de uma grade regular.
Fonte: Rocha et al. (2019)

Ainda, no contexto da variável "Viagens por automóvel", dois tipos de superfícies foram geradas com base na proporção de viagens por automóvel e no número de viagens feitas por automóvel.

3.2.2 Métodos para criação de superfícies por Krigagem

Após a definição da grade regular e tratamento das variáveis, procedeu-se à aplicação da geoestatística, incluindo o cálculo e ajuste de semivariogramas com base no modelo Gaussiano. Nesse sentido, a estimação das superfícies para a variável "Renda Familiar" foi realizada por meio de amostras de tamanho 100, 500, 1.000 e 5.000 domicílios, a fim de verificar a influência do tamanho da amostra nas superfícies estimadas. Já a estimação das superfícies para a variável "Viagens por automóvel" foi realizada utilizando todos os dados disponíveis.

3.2.3 Métodos para criação de superfícies pela RGP

Para a aplicação da RGP, a representação visual da demanda por transporte público na região central de São Paulo foi feita com os mesmos dados utilizados na construção das superfícies pela Krigagem, para fins de comparação.

Na RGP as superfícies foram geradas conforme detalhado na Seção 2.7.3, com a substituição da matriz de covariáveis por um vetor de 1's e conforme descrito por Da Silva (2016). Além disso, os parâmetros de suavização foram ajustados utilizando o modelo gaussiano para os dados em formato de porcentagem e os modelos binomial negativo e gaussiano para os dados de contagem, para fins de comparação.

Capítulo 4

Análise dos Resultados

4.1 Criação das superfícies - renda familiar

O banco de dados referente a área total de São Paulo foi delimitado para a área central, como pode ser visto na Figura 4.1. Para a variável "Renda familiar", observa-se a ausência de um padrão definido de distribuição, uma vez que existem áreas onde a renda é menor misturadas entre áreas onde a renda é maior, embora pareça que as maiores rendas estejam concentradas na região mais central. Ou seja, estamos claramente diante de um fenômeno não-contínuo (mesmo que a variável seja contínua), onde deseja-se estimar a renda familiar em toda a região de estudo.

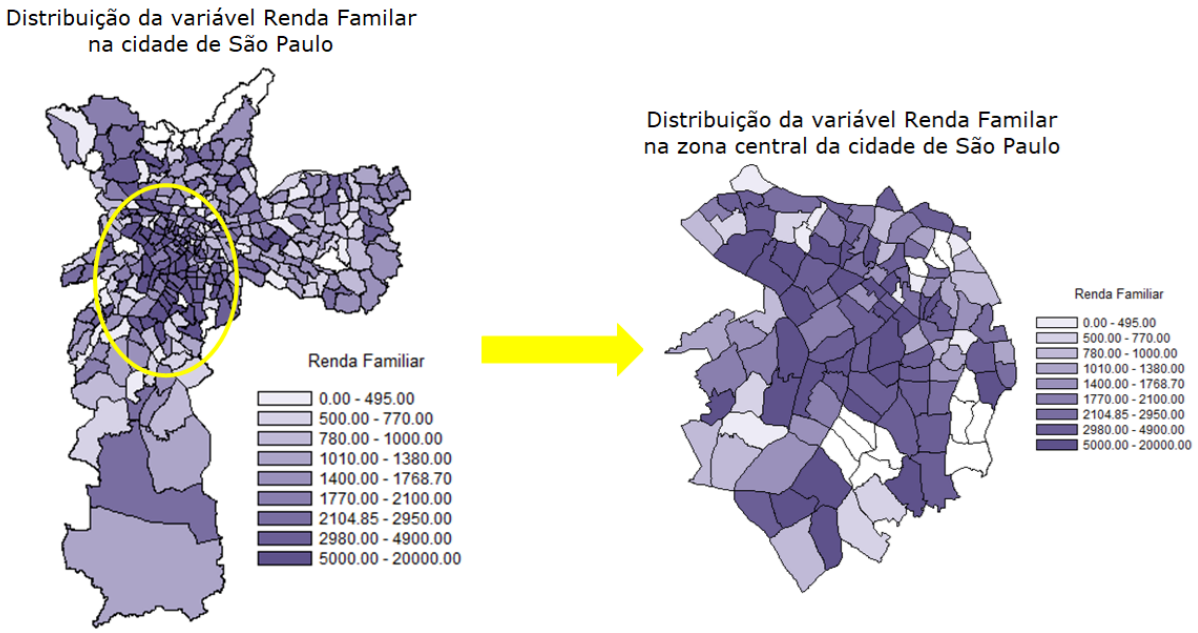


Figura 4.1: Distribuição da variável renda familar na área de estudo.

A redução da área foi necessária devido à grande quantidade de dados contidos no banco de dados original. Na Figura 4.2, apresenta-se a distribuição dos 8.498 domicílios resultantes do tratamento dos dados, oferecendo uma representação visual da área selecionada para análise.

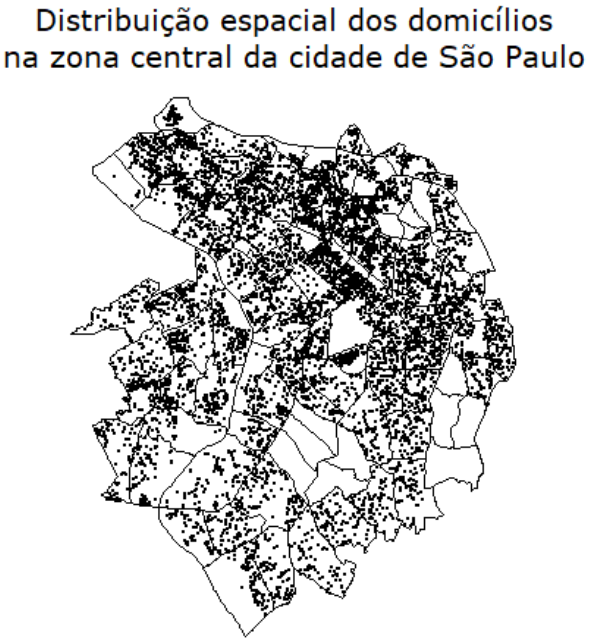


Figura 4.2: Distribuição dos 8.498 domicílios da região central de São Paulo.

Mesmo após a redução da área de estudo, ainda persiste um grande número de domicílios na região central, o que implica em uma demanda computacional considerável para gerar as superfícies desejadas. É crucial destacar a significativa carga computacional associada a essas técnicas, especialmente à Krigagem.

Desde 2011, as previsões climáticas do Instituto Nacional de Pesquisas Espaciais (INPE) têm sido conduzidas por um supercomputador capaz de realizar 258 trilhões de cálculos por segundo, permitindo um detalhamento de até 5 quilômetros (AEB, 2010). Embora essa escala tenha alcançado uma precisão notável, é relevante observar que o sistema computacional atual do INPE, denominado Tupã, está desatualizado e está prestes a ser substituído por uma nova supermáquina capaz de previsões com resolução em nível de metros, oferecendo um alto grau de precisão nas estimativas de superfícies (Correia, 2023).

Assim, antes de proceder à análise das superfícies geradas pela metodologia de RGP e pela Krigagem, foi necessário realizar uma amostragem dos domicílios, também com o objetivo de verificar a influência da quantidade de pontos nas superfícies estimadas.

Nesse contexto, o propósito é avaliar a eficácia das técnicas mencionadas, por meio da análise das diferentes amostras de 100, 500, 1.000 e 5000 domicílios (Figura 4.3).

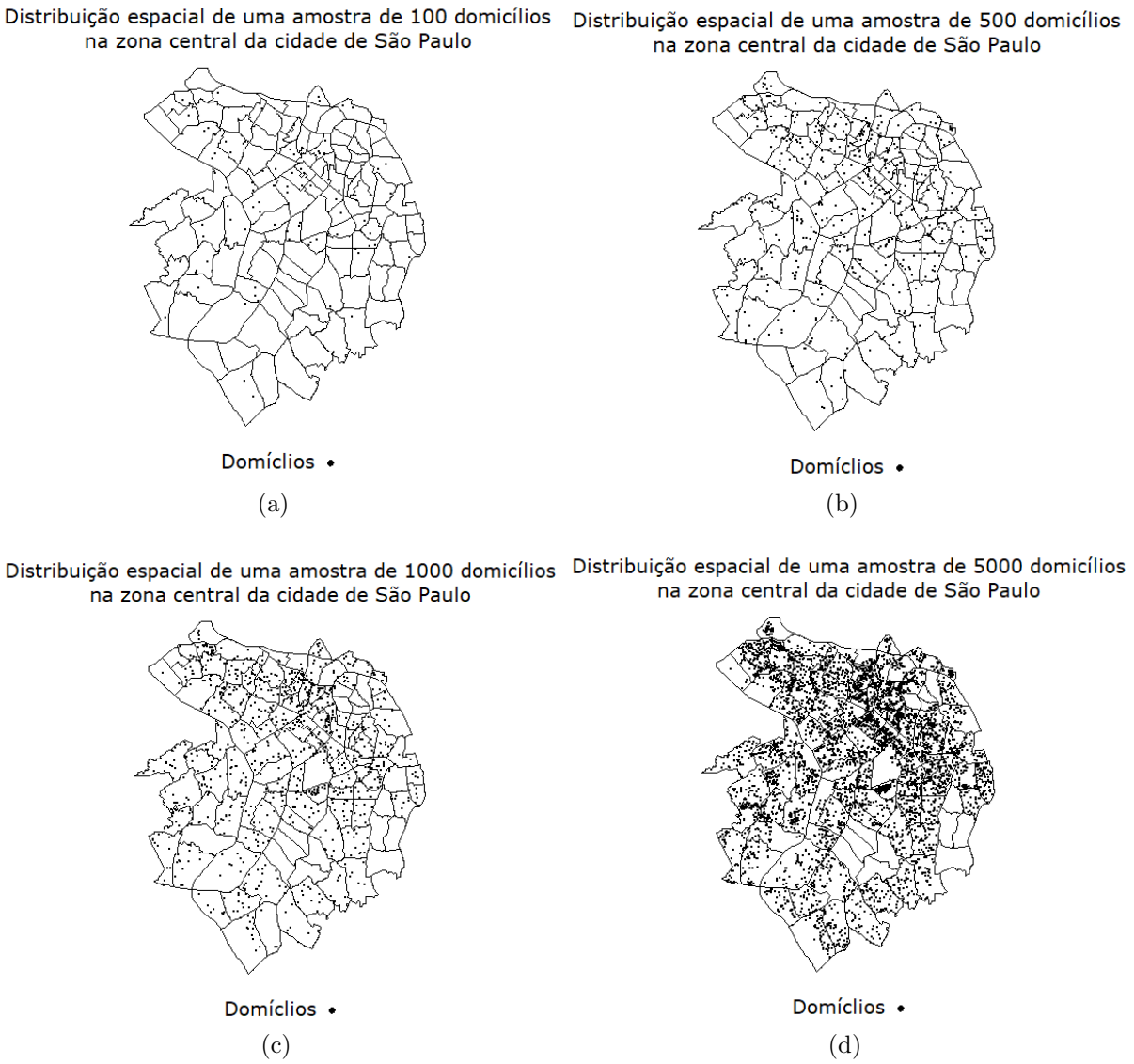


Figura 4.3: Distribuição espacial das amostras de (a) 100, (b) 500, (c) 1.000 e (d) 5.000 domicílios na zona central da cidade de São Paulo.

A aplicação da técnica de Regressão Geograficamente Ponderada inicia-se com a otimização do parâmetro de suavização para cada uma das amostras, resultando nos valores apresentados na Tabela 4.1. É possível observar na Figura 4.4 que os valores atribuídos aos parâmetros de suavização estão, de fato, otimizados, uma vez que são mínimos globais das funções, e que em geral, ele tende a diminuir à medida que a amostra aumenta.

Os parâmetros de suavização obtidos para as amostras de 100, 500 e 1.000 domicílios foram divididos, respectivamente, por um fator de 2.5, 1.5 e 1.5, para acomodar

Tabela 4.1: Valores dos parâmetros de suavização para cada amostra selecionada.

Amostra	100	500	1.000	5.000
Parâmetro de Suavização	1.668,652	787,5056	819,80685	382,60341

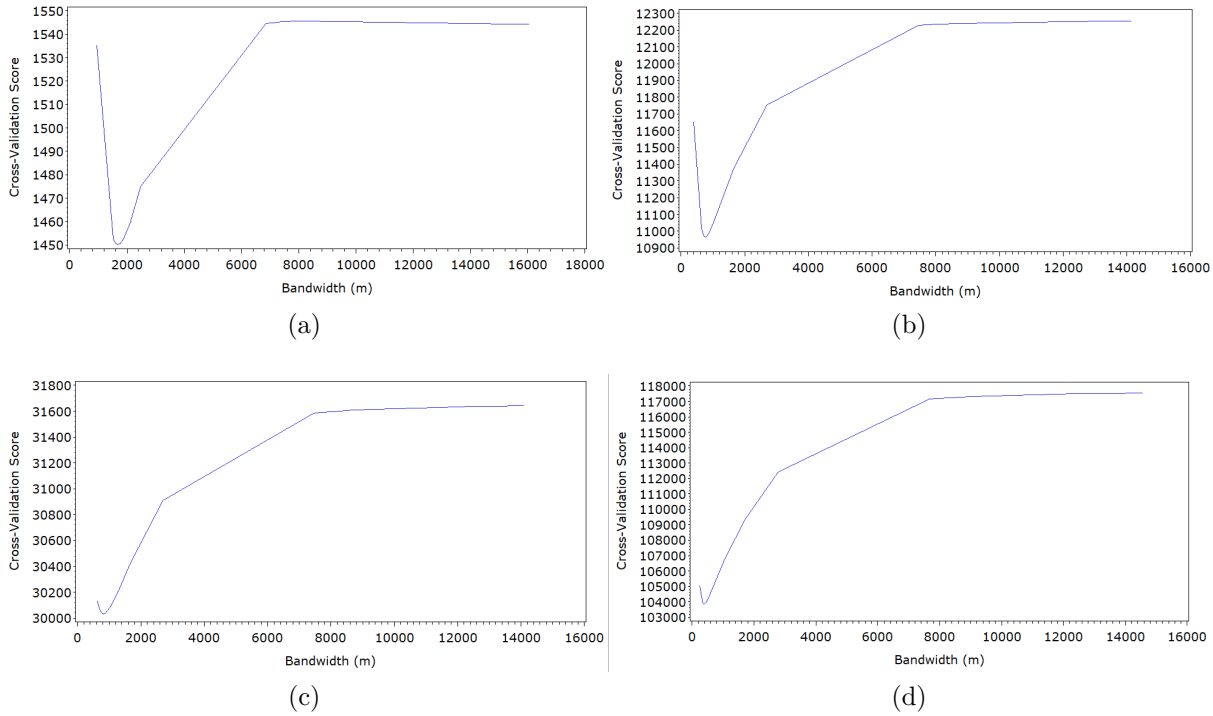


Figura 4.4: Valor do Bandwidth calculado para as amostras de (a) 100, (b) 500, (c) 1000 e (d) 5000 domicílios.

os percentis 10 e 90, visto que as estimativas do modelo não foram capazes de capturar os valores mais extremos da variável com o parâmetro de suavização encontrado. Essa adaptação visa assegurar a obtenção de superfícies que representem com maior precisão a variação da renda na região analisada.

Ainda, antes de avançar para a criação das superfícies utilizando a técnica de Krigagem, foi necessário ajustar os parâmetros dos semivariogramas. Inicialmente, a estimação desses parâmetros foi realizada utilizando o método da máxima verossimilhança, conforme detalhado na seção 2.5.2. No entanto, a estimação não convergiu, indicando que o tipo de fenômeno subjacente aos dados não é adequado para a aplicação da técnica de Krigagem. Diante da falta de convergência do modelo pela máxima verossimilhança, os parâmetros iniciais dos semivariogramas foram estimados por meio do método dos mínimos quadrados, seguindo o modelo teórico Gaussiano de semivariograma, como evidenciado nos resultados apresentados na Tabela 4.2.

Após definir o modelo teórico a ser ajustado, o processo de estimação dos semi-

variogramas foi conduzido de forma individual para cada uma das amostras de 100, 500, 1.000 e 5.000 domicílios, as quais já haviam sido utilizadas anteriormente para o ajuste do parâmetro de suavização.

Tabela 4.2: Valores ajustados do semivariograma para as amostras de 100, 500, 1.000 e 5.000 domicílios.

Amostra	**Patamar (C)**	**Alcance (a)**	**Efeito pepita (C_0)**	**lagd**
100	39,146155	135,45043	6,1169459	56,886428
500	38,3448	110,20975	16,904166	29,957011
1.000	37,18567	343,89	24,935093	100
5.000	25,1119	606,43	20,2064	100

Em geral, o semivariograma é definido por três parâmetros: alcance (a), patamar (C) e efeito pepita (C_0). No entanto, esses não são os únicos componentes que formam a curva do semivariograma. Também é importante definir o número de classes (lag) e a amplitude dessas classes (lagd), já que esses parâmetros também influenciam no formato do semivariograma. Dessa forma, os parâmetros lag e lagd foram determinados automaticamente conforme descrito por Da Silva et al. (2016) (Tabela 4.2). Mesmo com a definição automática de lag e lagd, para as amostras de 1.000 e 5.000, esses valores foram ajustados manualmente, devido a problemas na estimação das superfícies. Observou-se a ocorrência de estimativas negativas (devido à pesos negativos), o que não é desejável na Krigagem, mas pode acontecer conforme Deutsch (1996).

Conforme esperado, e ilustrado na Figura 4.5, observa-se que o ajuste do semivariograma se torna mais apropriado à medida que o tamanho da amostra aumenta. No entanto, é notável que a função acumulada dos pontos, à medida que a distância aumenta, não apresenta um padrão suave (linha azul). Em vez disso, ela exibe variações, algumas vezes ascendentes e outras vezes descendentes, indicando que o modelo ajustado não é ideal em nenhum dos casos, ja que esse não é cum comportamento esperado para uma função de distribuição acumulada.

Ajustando todos os parâmetros necessários para a estimação das superfícies por RGP e Krigagem, prosseguiu-se com a construção das superfícies utilizando ambas as técnicas. A Figura 4.6 apresenta as estimativas da distribuição de renda em salários mínimos, onde as áreas com cores mais frias indicam menor renda, enquanto as áreas mais quentes representam maior renda. A análise das figuras revela que as estimativas

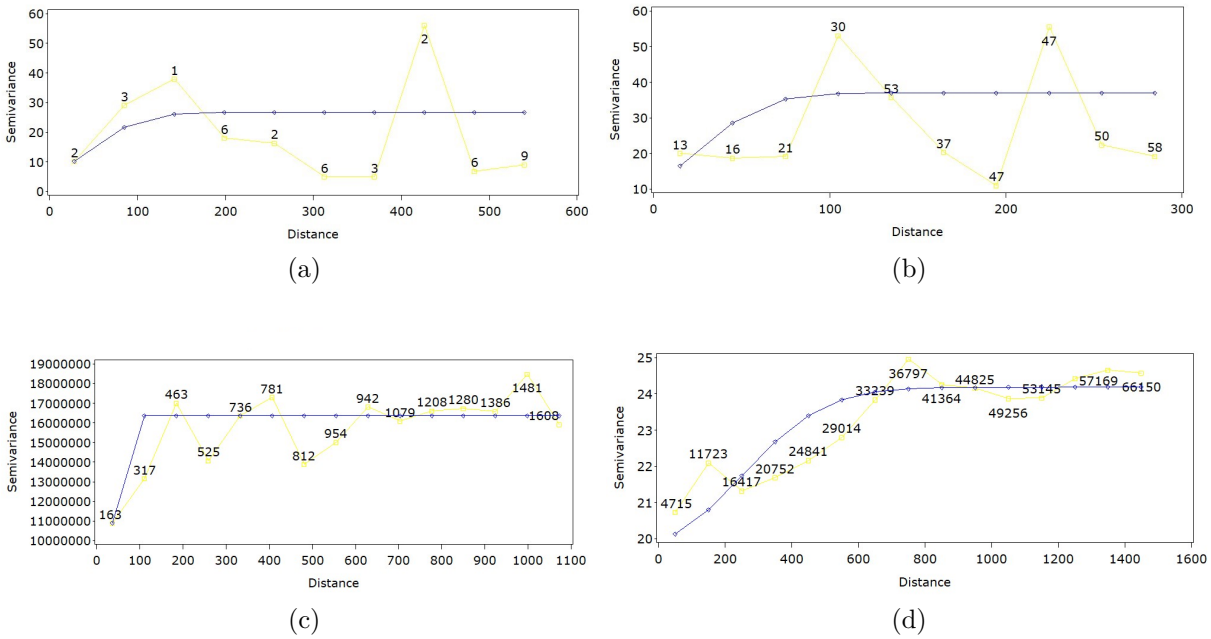


Figura 4.5: Semivariograma ajustado para as amostras de (a) 100, (b) 500, (c) 1.000 e (d) 5.000 domicílios.

obtidas por meio da técnica de RGP produzem superfícies mais precisas, mesmo com uma amostra menor de domicílios, capturando de forma mais detalhada as nuances da distribuição dos pontos amostrados. Por outro lado, as estimativas geradas pela técnica de Krigagem são restritas às áreas próximas aos domicílios amostrados, melhorando à medida que o tamanho da amostra aumenta.

É importante destacar que as estimativas da Krigagem podem incluir intervalos de renda familiar com valores negativos. De acordo com Deutsch (1996), pesos negativos na krigagem ordinária podem surgir quando os dados próximos ao local de estimativa excluem valores atípicos. Dependendo do variograma ajustado e da quantidade de exclusões, esses pesos negativos podem ter um impacto significativo nas estimativas. Além disso, a aplicação de pesos negativos a valores extremos pode resultar em estimativas que extrapolam o intervalo dos dados observados. Esse resultado pode ser atribuído à natureza não contínua do fenômeno da renda familiar. Não é esperado que a renda apresente uma distribuição suave em uma determinada área, pois é possível encontrar domicílios com renda mais elevada em proximidade com domicílios de renda mais baixa.

Ainda, conforme esperado, observa-se que quanto maior for a quantidade de pontos amostrados, maior será a concordância da superfície estimada com a distribuição real da renda na área tanto para as superfícies estimadas por RGP quando para as superfícies estimadas por Krigagem.

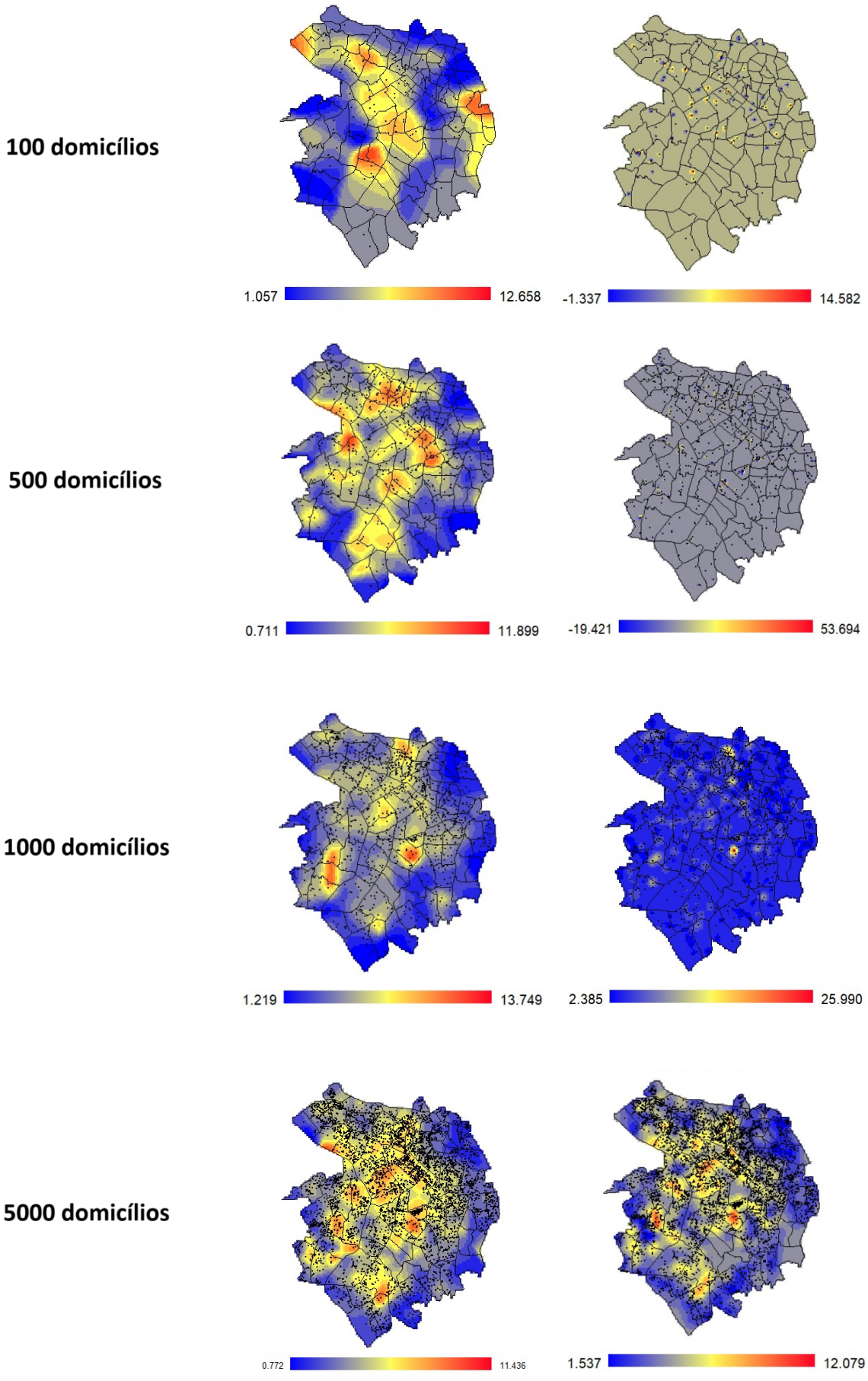


Figura 4.6: Superfícies estimadas por meio das técnicas de RGP e Krigagem, respectivamente, para as amostras de 100, 500, 1.000 e 5.000 domicílios.

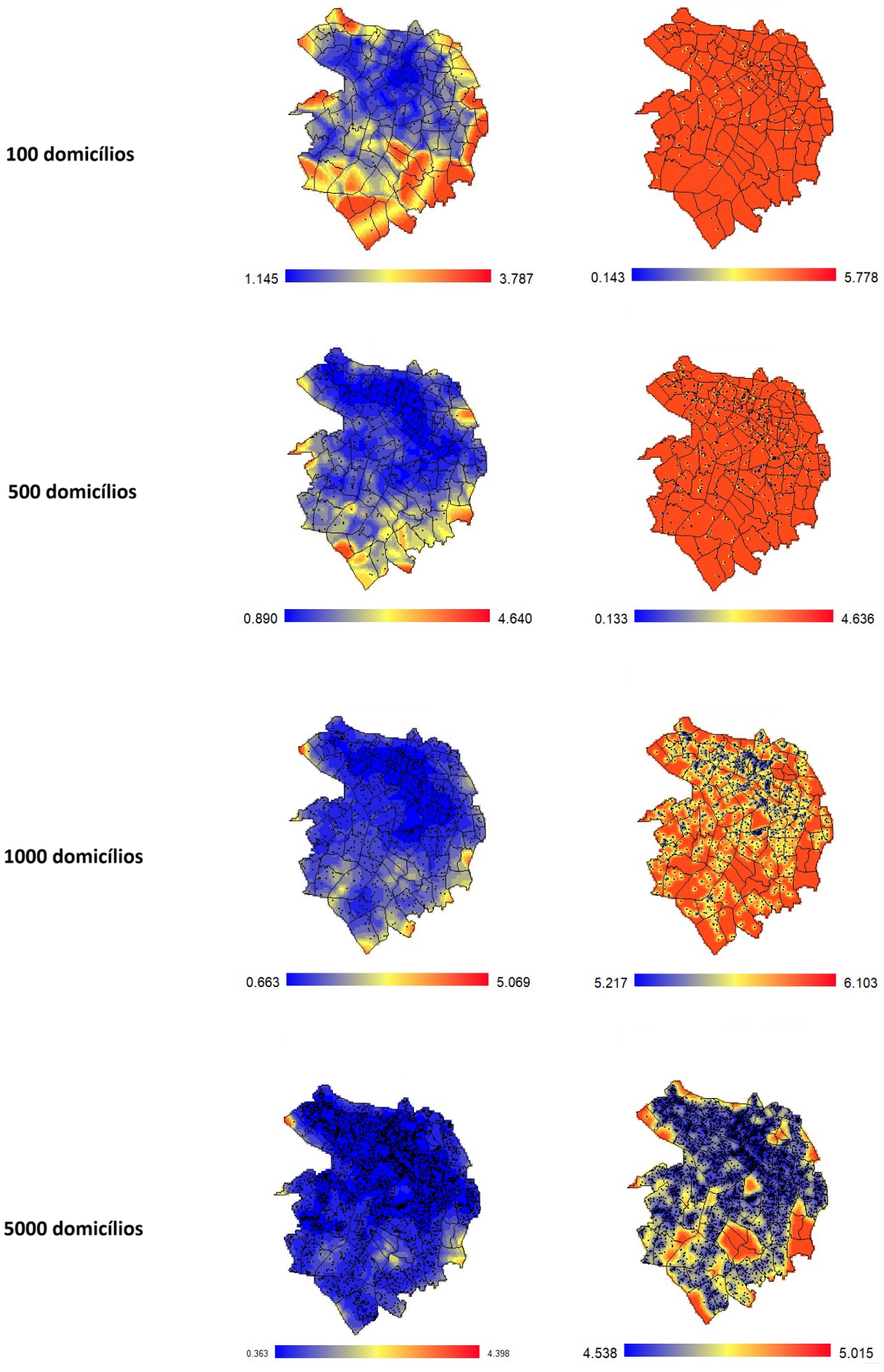


Figura 4.7: Desvio padrão das superfícies estimadas por meio das técnicas de RGP e Krigagem, respectivamente, para as amostras de 100, 500, 1.000 e 5.000 domicílios

A Figura 4.7 apresenta o desvio padrão das estimativas mostradas na Figura 4.6, onde as cores mais frias indicam desvios de menor magnitude e as cores mais quentes indicam desvios de maior magnitude. Observa-se que, devido à sua natureza como estimador exato, a Krigagem produz estimativas com erros menores, principalmente nas proximidades dos pontos amostrados. No entanto, fora dessas áreas, onde não há pontos amostrados, o desvio padrão em torno da média aumenta consideravelmente. Por outro lado, a RGP não limita os menores desvios apenas às proximidades dos domicílios amostrados, e esses desvios são de menor magnitude em comparação com os obtidos pelas estimativas da Krigagem.

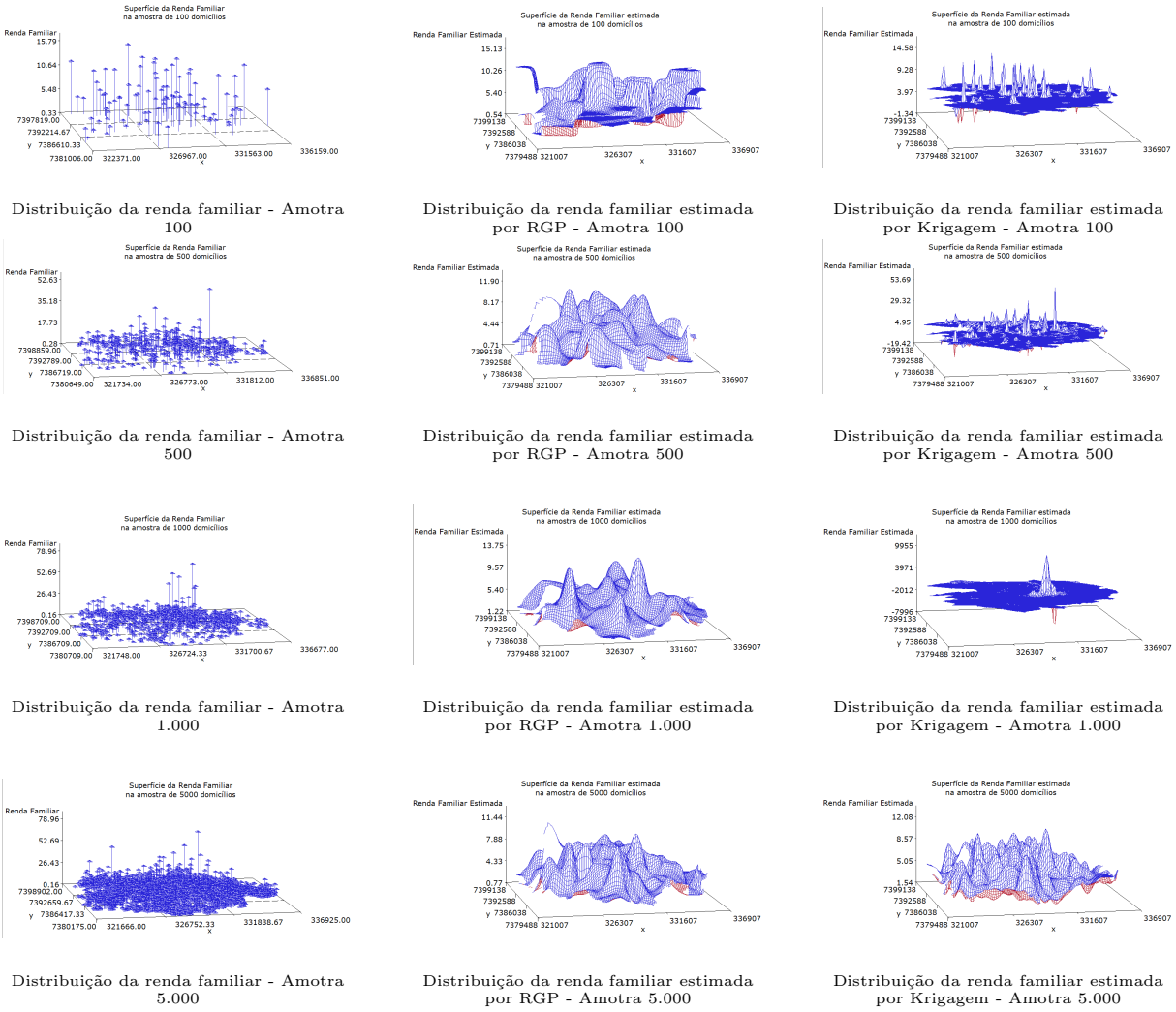


Figura 4.8: Superfícies de distribuição de renda familiar e suas estimativas utilizando RGP e Krigagem para as amostras de 100, 500, 1.000 e 5.000 domicílios.

Ao destacar de maneira mais acentuada as vantagens derivadas da aplicação da RGP na elaboração de superfícies em situações onde o fenômeno em estudo é não-contínuo, a Figura 4.8 evidencia a capacidade da RGP em capturar de forma mais precisa as ca-

racterísticas da distribuição da variável renda familiar, mesmo a partir de uma amostra de 100 domicílios. Essa melhoria torna-se ainda mais evidente à medida que o tamanho da amostra aumenta. Notavelmente, com exceção da amostra de 5.000 domicílios, que mostra superfícies bastante semelhantes independentemente da técnica utilizada.

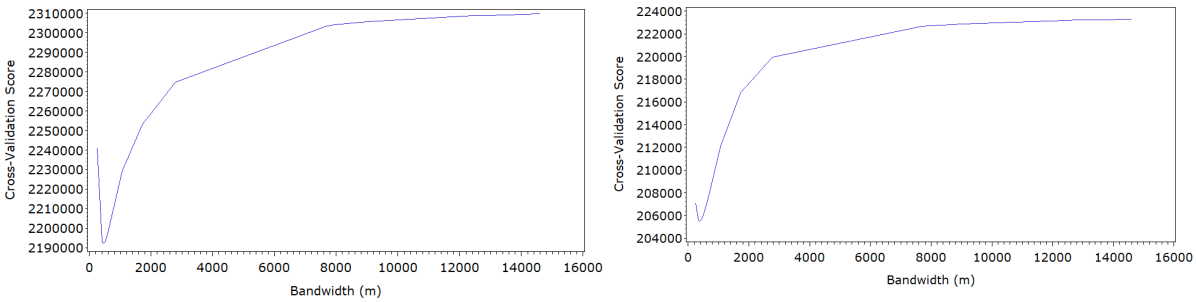
4.2 Criação das superfícies - Viagens por automóvel

Seguindo com a criação das superfícies para a variável "Viagens por automóvel", a aplicação do *Buffer* resultou em um total de 3.859 domicílios, não sendo necessário o uso de amostras para obter as representações visuais das superfícies de distribuição estimadas.

Antes de criar as representações, foram determinados os parâmetros de suavização ótimos para aplicar a RGP, considerando tanto a porcentagem de viagens por automóvel quanto o número de viagens por automóvel (Figura 4.9). Os parâmetros ajustados estão na Tabela 4.3.

Tabela 4.3: Valores dos parâmetros de suavização para a variável"Viagens por automóvel".

Variável	**Parâmetro de suavização**
Porcentagem de viagens	443,018
Número de viagens (Gaussiano)	383,729
Número de viagens (binomial negativo)	383,729



(a) (b)

Figura 4.9: Parâmetro de suavização ótimo para (a) porcentagem de viagens realizadas por automóvel e (b) número de viagens realizadas por automóvel.

O parâmetro de suavização para a porcentagem de viagens por automóvel foi ajustado dividindo-o por 1.38, considerando os percentis 10 e 90 da distribuição, resultando em um valor de $h = 322$. Observou-se que esse valor de parâmetro permaneceu constante para os dados de contagem, independentemente do modelo utilizado (Gaussiano

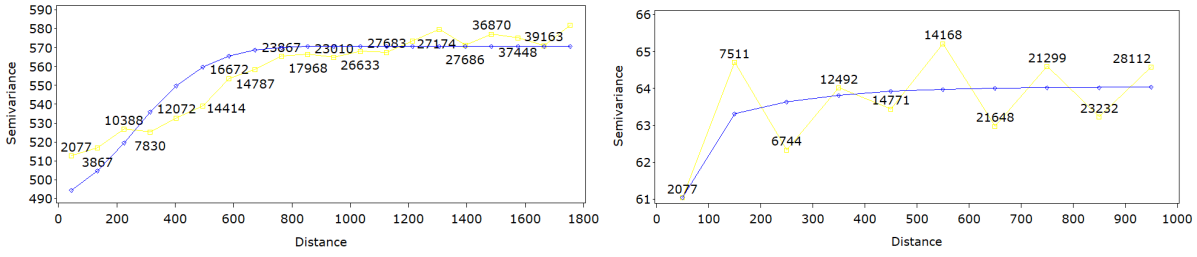
ou binomial negativo). Portanto, para estimar as superfícies para os dados de contagem, utilizou-se o parâmetro $h = 383,729$. Novamente, é possível observar na Figura 4.9 que os valores atribuídos aos parâmetros de suavização estão, de fato, otimizados, uma vez que representam os mínimos globais das funções.

Partindo para o ajuste do semivariograma para aplicação da técnica de Krigagem tanto na porcentagem de viagens por automóvel quanto na contagem de viagens por automóvel, a Tabela 4.4 apresenta os parâmetros estimados.

Tabela 4.4: Parâmetros estimados do semivariograma para os dados de porcentagem de viagens por automóvel e para a contagem de viagens por automóvel.

Dados veículos	**Patamar** (C)	**Alcance** (a)	**Efeito pepita** (C_0)	**lagd**
Porcentagem de viagens	575,0622	617,03	512,94	90
Contagem de viagens	64,0672	498,43	63,3860	100

Assim como o resultado obtido para a variável renda, observa-se que o ajuste do semivariograma não é ideal para a variável que representa o número de viagens por automóvel (Figura 4.10). Já para a proporção de viagens por automóvel, o semivariograma parece ter um bom ajuste. Como mencionado anteriormente, as variações na função acumulada dos pontos não correspondem a uma função de distribuição acumulada pela distância. Esse fato indica que a modelagem pela Krigagem não é a mais adequada para estimar superfície utilizando esse tipo de dado, no caso a variável discreta número de viagens por automóvel.



(a) (b)

Figura 4.10: Semivariograma ajustado para (a) porcentagem de viagens realizadas por automóvel e (b) número de viagens realizadas por automóvel.

Partindo para a criação das superfícies para a variável “Viagens por automóvel”, a Figura 4.11 apresenta as superfícies estimadas.

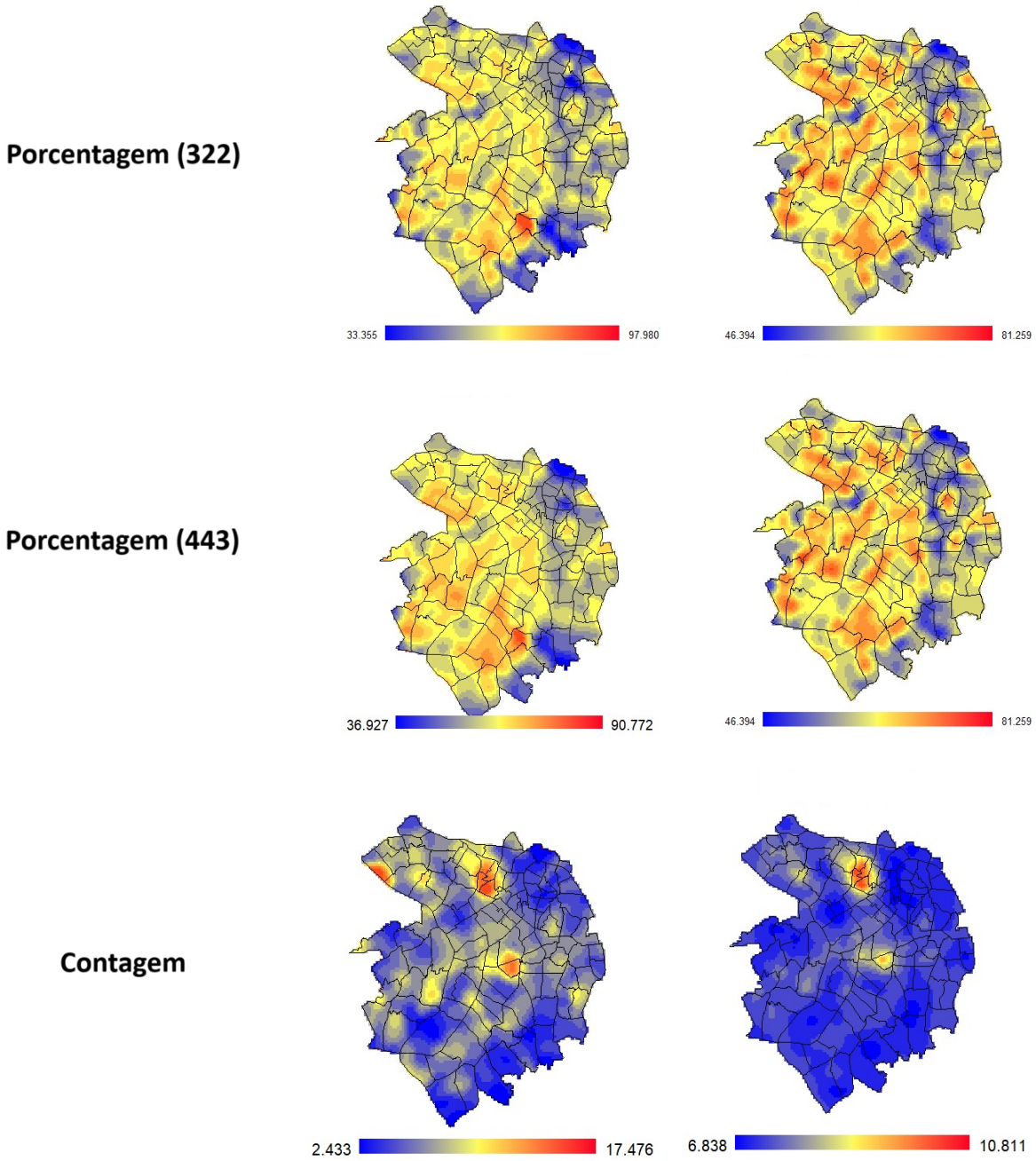


Figura 4.11: Superfícies estimadas por meio das técnicas RGP (esquerda) e Krigagem (direita), para a porcentagem e número de viagens por automóvel.

Ao analisar as superfícies resultantes, nota-se uma significativa semelhança nas estimativas geradas pela RGP, quer seja utilizando o parâmetro de suavização igual a 443 ou o parâmetro corrigido pelo fator para abranger os percentis 10 e 90, especialmente para os dados de porcentagem. Além disso, as superfícies estimadas tanto pela Krigagem quanto pela RGP para esses dados apresentam uma consistência notável, refletindo os resultados previamente observados para a variável de renda familiar em cenários com

amostras de maior número de domicílios. Isso resulta em superfícies geradas pela RGP e Krigagem muito similares.

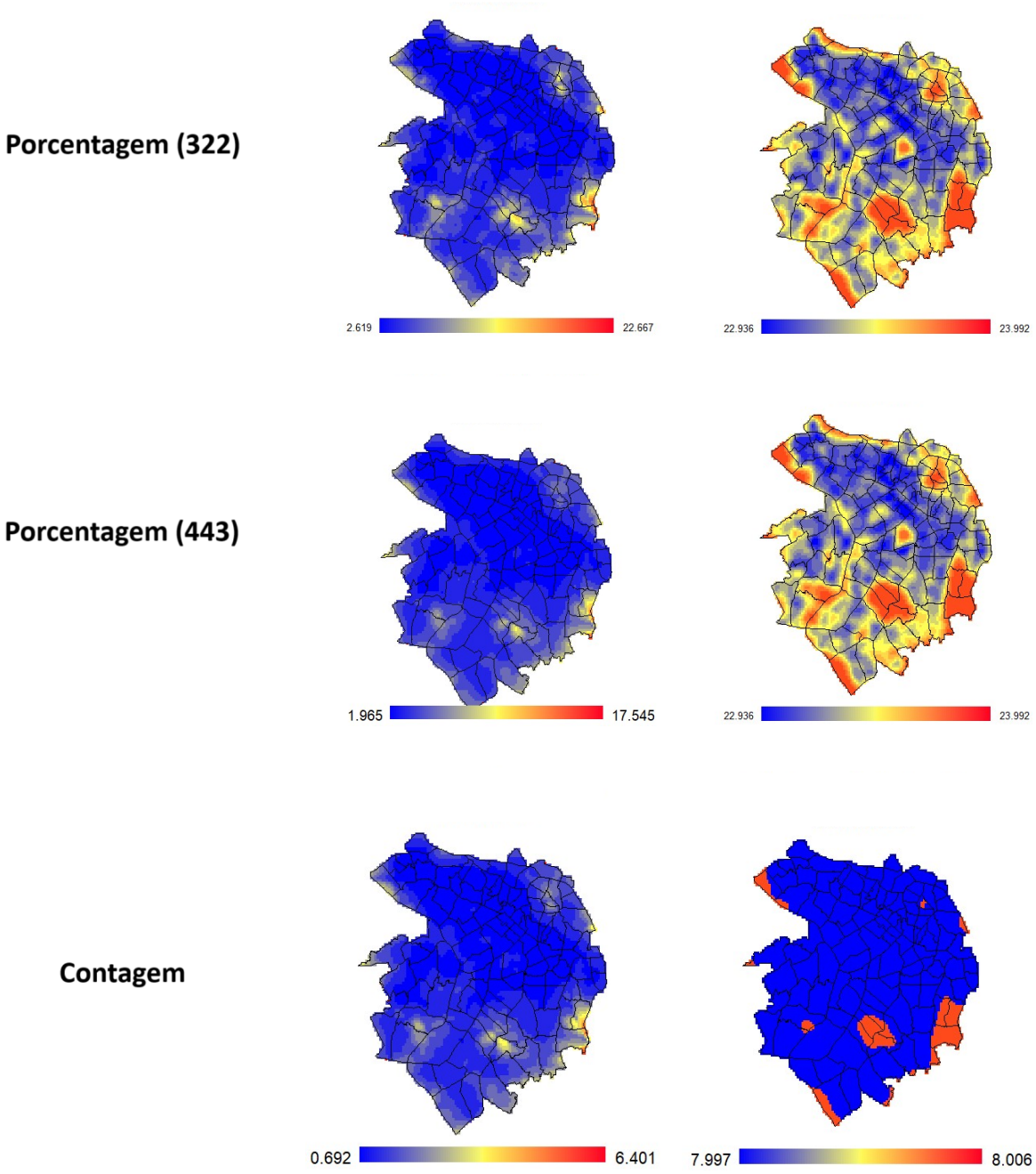


Figura 4.12: Desvio padrão das superfícies estimadas por RGP e Krigagem, respectivamente, para porcentagem de viagens por automóvel e contagem de viagens por automóvel.

No entanto, para os dados de contagem, a superfície estimada pela RGP demonstra uma capacidade superior em capturar as nuances da distribuição em comparação com a abordagem da Krigagem. Vale ressaltar que a superfície estimada para o número de viagens por veículo pela RGP foi feita utilizando a distribuição Gaussiana, visto que a

superfície gerada pela distribuição binomial negativa gerou estimativas muito próximas à média. Esse fato não era esperado e merece ser estudado com mais profundidade.

A Figura 4.12 apresenta o desvio padrão das estimativas mostradas na Figura 4.11. Considerando os erros das estimativas, fica claro que, em geral, eles são menores ao utilizar a RGP, confirmando resultados anteriores relacionados à variável renda familiar. Essa diferença se destaca ainda mais ao analisar os dados de contagem, o que reforça a conclusão de que a Krigagem não é eficiente para dados não-contínuos. De forma consistente, a estimativa de Krigagem, para esse caso, é mais precisa em áreas densamente amostradas, porém, essa precisão diminui significativamente em regiões com poucos ou nenhum ponto amostrado.

4.3 Comparação entre RGP e Krigagem

Uma consideração importante a ser feita é que não é possível comparar as estimativas geradas por ambas as técnicas de forma justa. A Krigagem, em primeiro lugar, possui a característica de ser um estimador exato, ou seja, quando um ponto amostrado coincidir exatamente com a grade regular, o valor estimado corresponde precisamente ao valor medido e com variância zero, conforme visto na seção 2.6.2. Além disso, a precisão das estimativas está relacionada ao tamanho da grade regular estabelecida, ou seja, ela é significativa melhor nas proximidades dos pontos amostrados, diminuindo à medida que a distância em relação aos pontos aumenta.

	x	y	estimate	STDERR
1	332519	7397108	1.6447368421	0
2	333057	7397047	2.7616578947	0
3	334441	7393592	0.4961677632	0
4	334909	7393886	1.8289473684	0
5	335144	7392263	11.842105263	0
6	333212	7392152	9.8684210526	0
7	333006	7392912	1.2236842105	0
8	331821	7393835	2.9736842105	0
9	332268	7392432	3.1801469298	0
10	331679	7393012	1.5514912281	0
11	331157	7394593	3.2865263158	0
12	331206	7394491	2.8607105263	0
13	331507	7394469	4.8684210526	0
14	331562	7395167	9.2105263158	0
15	329610	7394712	3.9223157895	0
16	329787	7394422	6.8088070175	0
17	330233	7394805	4.6677763158	0
18	330609	7397057	1	0
19	331154	7397594	1.6842105263	0
20	331200	7397424	1.3190789474	0

Estimativas dos pontos na amostra de 100 domicílios.

	x	y	estimate	STDERR
1	331994	7395544	15	0
2	332558	7395876	10.526315789	0
3	332882	7396098	2.8947368421	0
4	333172	7395763	1.1950877193	0
5	332501	7397121	3.3684210526	0
6	332950	7397131	2.7774539474	0
7	333553	7396410	1.3037105263	0
8	333831	7396364	2.6315789474	0
9	333920	7396106	1.7105263158	0
10	333947	7396708	2.397	0
11	334590	7396225	1.9555657895	0
12	334598	7396129	0.6315789474	0
13	334598	7396186	3.1118508772	0
14	334609	7396050	0.547754386	0
15	334720	7396076	2.3947368421	0
16	334731	7396007	1.0529239766	0
17	333948	7395696	1.2529678363	0
18	334139	7395432	1.4434210526	0
19	334302	7395173	0.9777828947	0
20	334563	7395355	2.0141929825	0

Estimativas dos pontos na amostra de 500 domicílios.

	x	y	estimate	STDERR
1	331912	7395542	1.052666325	0.0002277349
2	331982	7395441	2.6934557967	0.0002262192
3	331988	7395827	12.119692605	0
4	331994	7395544	15.000034744	0.0002565515
5	332236	7395703	4.2845314202	0
6	332488	7395828	2.6316136559	0
7	332496	7396076	1.0526662881	0
8	332583	7395892	2.6316136573	0
9	332658	7396060	1.9079294458	0
10	332702	7395843	1.4955487433	0
11	332968	7395601	2.8947715515	0
12	333086	7395592	0.5395083939	0
13	332501	7397121	3.3684557614	0
14	332620	7397229	3.9223504981	0
15	332739	7396869	7.1910347091	0
16	332831	7396812	3.9223504978	0
17	332989	7396703	3.0916867557	0
18	333022	7396814	1.5789820772	0
19	333124	7396946	1.2658110255	0
20	333133	7396978	3.2959426032	0

Estimativas dos pontos na amostra de 1.000 domicílios.

Figura 4.13: Estimativas dos valores da variável renda familiar para os vinte primeiros domicílios amostrados em três conjuntos de amostras distintos

Utilizando a variável renda familiar como exemplo, na Figura 4.13 é observada a estimativa resultante da aplicação da técnica de Krigagem em pontos exatos. Notavelmente, essa estimativa é idêntica ao valor real no ponto amostrado, refletindo um desvio

padrão nulo. Essa constatação ganha maior clareza ao se gerarem as superfícies dos pontos exatos (Figura 4.14).

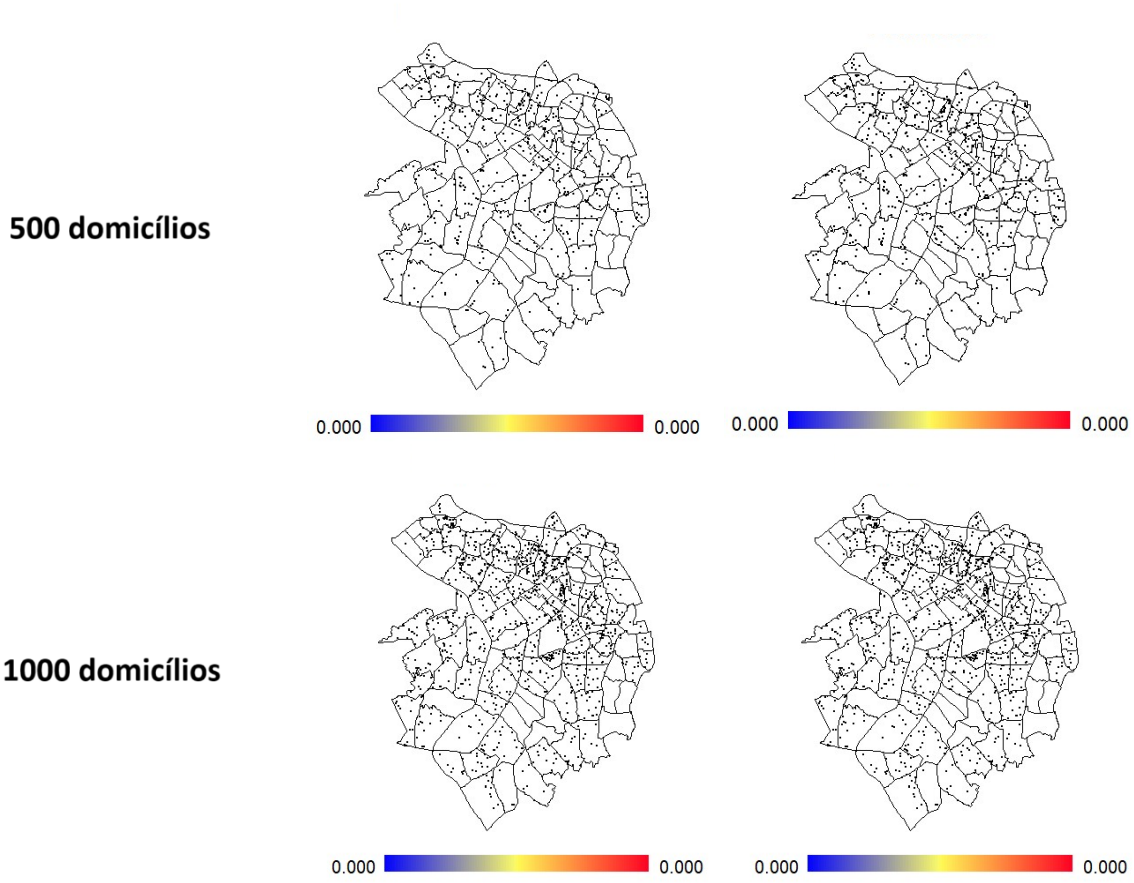


Figura 4.14: Superfícies estimadas pela Krigagem no ponto exato para as amostras de 500 e 1000 domicílios e seus respectivos desvios padrão.

Nesse sentido, as análises apresentadas em Wang et al. (2013) e Ku Wang e Li (2012), podem ser limitadas devido à natureza não comparável dos métodos empregados, os quais se baseiam em abordagens de modelagens distintas. Além disso, nos estudos mencionados, são retiradas amostras dos dados para a estimação das superfícies por meio da Krigagem. No entanto, como foi visto na análise da variável renda familar, as superfícies estimadas pela Krigagem que utilizaram apenas uma amostra dos dados não foi capaz de reproduzir uma superfície comparável à quase totalidade dos dados, ao contrário das superfícies geradas pela RGP. Em outras palavras, quando todos os dados estão disponíveis, é fundamental considerar todos eles na estimação, como discutido na seção 4.1.

Ainda com base nas análises deste estudo e nos resultados obtidos, é evidente que a RGP apresenta uma aplicação consideravelmente mais simples em comparação com a Krigagem. Isso é notável já na etapa de estimação dos parâmetros, pois a RGP

requer apenas a definição do parâmetro de suavização, enquanto a Krigagem demanda a estimação de quatro parâmetros. Além disso, a Krigagem ordinária possui pressupostos mais rigorosos em comparação com a RGP como a suposição de estacionariedade, isotropia e fenômenos contínuos.

Ainda, é importante destacar que a Krigagem ordinária realiza basicamente uma análise de médias, enquanto que a RGP foi capaz de gerar a mesma superfície que a Krigagem, conforme demonstrado nas análises anteriores, utilizando apenas um vetor de 1's (ou o intercepto). Isso levanta a questão legítima de por que a RGP não é mais amplamente adotada na geração de superfícies, especialmente em fenômenos não-contínuos.

A explicação para essa questão pode estar na forma da implementação computacional da técnica RGP nos *softwares* disponíveis. Por exemplo, em um modelo de regressão com apenas o intercepto (ou seja, sem covariáveis), espera-se que ele produza o mesmo resultado que um teste t quando um vetor de 1's é inserido no lugar das covariáveis. No *software* SAS, essa análise pode ser realizada inserindo um vetor de 1's no lugar das covariáveis $\mathbf{X}$ ou executando o modelo sem especificar as covariáveis, como mostrado na parte inferior à esquerda da Figura 4.15, pois a sintaxe do *software* SAS aceita as duas configurações. Em ambos os casos, a estimativa do intercepto é a mesma, representando a média da variável de interesse em estudo. O resultado do teste t também confirma que o resultado é igual nos dois casos. Na RGP, a lógica é semelhante, mas há uma grade regular de pontos (semelhante à Krigagem), e as médias são estimadas com base na distância desses pontos da grade regular para os dados amostrados.

Já no *software* MGWR, utilizado para a RGP, não é possível executar o modelo sem covariáveis, como indicado pela mensagem de erro (parte superior da Figura 4.15), embora seja possível executar a análise sem o intercepto. No entanto, pode ser desafiador para o usuário criar um vetor de 1's no arquivo CSV. Da mesma maneira, o *software* R não permite a execução da análise sem as covariáveis (parte inferior à direita da Figura 4.15). Essas limitações na manipulação dos dados nos *softwares* existentes podem estar contribuindo para a menor adoção da RGP na geração de superfícies, já que a maioria dos *softwares* não oferece a manipulação necessária dos dados nos pacotes implementados, e nem todos os usuários possuem o conhecimento para implementar a RGP por conta própria. Em contraste, a geração da superfície pela Krigagem é mais acessível.

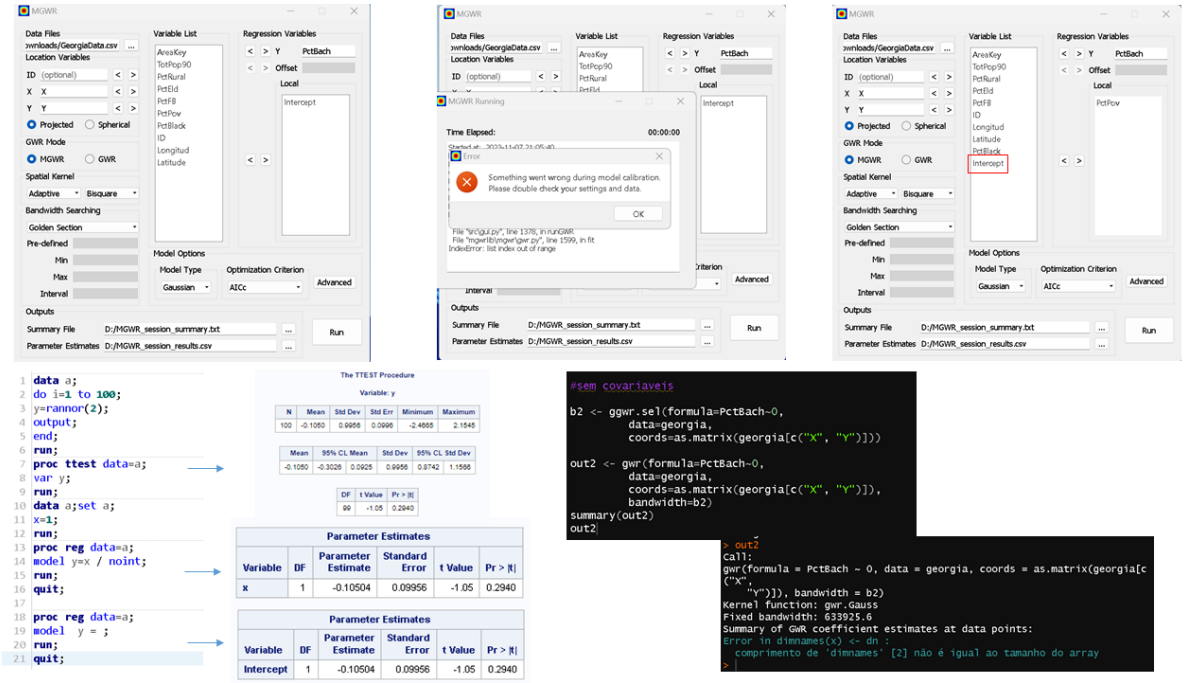


Figura 4.15: Comparação da implementação de funções de regressão em diferentes *softwares* disponíveis.

Capítulo 5

Conclusões

Com base na análise comparativa entre os métodos de Krigagem e Regressão Geograficamente Ponderada para estimação de superfícies a partir de fenômenos não-contínuos, é possível inferir que a Krigagem destaca-se como uma técnica adequada para a estimativa de superfícies em fenômenos contínuos, como temperatura, concentração de metais pesados no solo, taxa de precipitação e caracterização de solos.

Por outro lado, a Regressão Geograficamente Ponderada pode ser aplicada para a geração de superfícies a partir de fenômenos contínuos e não-contínuos, como os elecandos acima e também para a previsão do tráfego de veículos em vias e a modelagem do número de viagens no transporte público, pois leva em consideração a variação espacial desses fenômenos, resultando em superfícies com qualidade similar ou até superior às obtidas pela Krigagem.

Dessa forma, a partir das recomendações propostas de como gerar superfícies nos *softwares* disponíveis, tendo apenas a variável observada, a RGP se mostra uma ferramente robusta para a estimação de superfícies sem que o analista precisa se preocupar com a natureza da variável.

Referências Bibliográficas

AEB (2010). Agência espacial brasileira. inpe inaugurará supercomputador para previsões de tempo e estudos de mudanças climáticas. Disponível em: https://www.gov.br/aeb/pt-br/assuntos/noticias/inpe-inaugurara-supercomputador-para-previsao-de-tempo-e-estudos-de-mudancas-climatic~:text=O%20novo%20supercomputador%20permitir%C3%A1%20ao%20INPE%20gerar%20previs%C3%B5es,Sul%20e%2020%20km%20para%20todo%20o%20globo. Acesso em Jan. 2024.

Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York.

Bailey, T. & Gatrell, A. (1995). *Interactive Spatial Data Analysis*. Longman Scientific & Technical.

Bostan, P., Heuvelink, G., & Akyurek, S. (2012). Comparison of regression and kriging techniques for mapping the average annual precipitation of turkey. *International Journal of Applied Earth Observation and Geoinformation*, 19:115–126.

Christakos, G. (1984). On the problem of permissible covariance and variogram models. *Water Resources Research*, 20:251–265.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Conceição, S. F. (2013). *Discussão sobre a obtenção de funções semivariograma a partir de distribuições de probabilidade*. Dissertação de Mestrado. Programa de Pós-Graduação em Estatística. Universidade de Brasília.

Correia, F. (2023). Novo supercomputador promete revolucionar a previsão de tempo e clima no brasil. Disponível em: https://https://olhardigital.com.br/2023/02/25/ciencia-e-espaco/

`novo-supercomputador-vai-revolucionar-a-previsao-do-tempo-no-brasil/`. Acesso em Fev. 2024.

Cressie, N. A. C. (1993). *Statistics for Spacial Data*. John Wiley and Sons.

Câmara, G., Monteiro, A. M., Fucks, S. D., & Carvalho, M. S. (2004). *Análise Espacial de Dados Geográficos*. EMBRAPA.

Da Silva, A. R. (2016). Working with proc spp, proc gmap and proc ginside to produce nice maps. Disponível em: `https://support.sas.com/resources/papers/proceedings16/8540-2016.pdf`. Accesso em Jan. 2024.

Da Silva, A. R., de Oliveira Sousa, P. H. T., & Conceição, S. F. (2016). Another approach to fit variogram models. Não publicado.

Da Silva, A. R. & Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression - incorporating overdispersion. *Statistics and Computing*, 24:769–783.

De Oliveira, V. (2014). Poisson kriging: A closer investigation. *Spatial Statistics*, 7:1–20.

Derdouri, A. & Murayama, Y. (2020). A comparative study of land price estimation and mapping using regression kriging and machine learning algorithms across fukushima prefecture, japan. *Journal of Geographical Sciences*, 30:794–822.

Deshmukh, S. S. & Annappa, B. (2019). *Prediction of Crime Hot Spots Using Spatiotemporal Ordinary Kriging*. Springer Singapore.

Deutsch, C. V. (1996). Correcting for negative weights in ordinary kriging. *Computers and Geosciences*, 22(7):765–773.

Diggle, P. J. & Ribeiro, P. J. (2007). *Model-Based Geostatistics*. Springer.

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-Based Geostatistics. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 47(3):299–350.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis*. SAGE.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: the analysis of spatially varying relationships*. Wiley.

Fu, P., Yang, Y., & Zou, Y. (2022). Prediction of soil heavy metal distribution using geographically weighted regression kriging. *Bulletin of Environmental Contamination and Toxicology*, 108:344–350.

Gorospe, K. D. & Karl, S. A. (2011). Small-scale spatial analysis of in situ sea temperature throughout a single coral patch reef. *Journal of Marine Sciences*, 2011:1–12.

Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Isaaks, E. H. & Srivastava, R. M. (1989). *Applied Geostatistics*. Oxford University Press.

Jian, X., Olea, R. A., & Yu, Y.-S. (1996). Semivariogram modeling by weighted least squares. *Computers & Geosciences*, 22:387–397.

Ku Wang, C. Z. & Li, W. (2012). Comparison of geographically weighted regression and regression kriging for estimating the spatial distribution of soil organic matter. *GIScience and Remote Sensing*, 49(6):915–932.

Metrô - São Paulo (2007). Pesquisa origem e destino. Disponível em: `https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino`. Acesso em Set. 2023.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted poisson regression for disease association mapping. *Statistics in medicine*, 24:2695–717.

Nawar, S., Corstanje, R., Halcro, G., Mulla, D., & Mouazen, A. M. (2017). *Chapter Four - Delineation of Soil Management Zones for Variable-Rate Fertilization: A Review*, volume 143 of *Advances in Agronomy*. Academic Press.

Rocha, S. S., Pitombo, C. S., & Costa, L. H. M. (2019). Escolha da escala para modelagem geoestatística de variáveis de demanda por trasnportes. In: *33°Congresso de Pesquisa e Ensino em Transportes da AMPET*, pages 2644–2655.

Schlather, M. (1999). Introduction to positive definite functions and to unconditional simulation of random fields. Technical report, Department of Mathematics and Statistics, Faculty of Applied Sciences, Lancaster University, UK.

Schlather, M., Porcu, E., & Montero, J. M. (2012). *Advances and Challenges in Space-time Modelling of Natural Events*, (1st ed.). Springer.

Selby, B. & Kockelman, K. M. (2013). Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. *Journal of Transport Geography*, 29:24–32.

Trumpis, M., Chiang, C.-H., Orsborn, A. L., Bent, B., Li, J., Rogers, J. A., Pesaran, B., Cogan, G., & Viventi, J. (2021). Sufficient sampling for kriging prediction of cortical potential in rat, monkey, and human µecog. *Journal of Neural Engineering*, 18:036011.

Wan, H., Li, J., Shang, S., & Rahman, K. U. (2021). Exploratory factor analysis-based co-kriging method for spatial interpolation of multi-layered soil particle-size fractions and texture. *Jornal of soils and sediments*, 21:3868–3887.

Wang, K., Zhang, C., & Li, W. (2013). Predictive mapping of soil total nitrogen at a regional scale: A comparison between geographically weighted regression and cokriging. *Applied Geography*, 42:73–85.