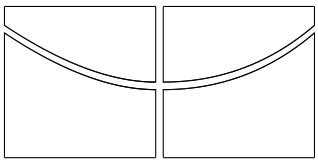


Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Dissertação de Mestrado

Análise do Problema da Unidade de Área Modificável Pela Regressão Geograficamente Ponderada

por

Alisson Carlos da Costa Silva

Orientador: Prof. Dr. Alan Ricardo da Silva

Janeiro de 2019

Alisson Carlos da Costa Silva

Análise do Problema da Unidade de Área Modificável Pela Regressão Geograficamente Ponderada

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Universidade de Brasília

Brasília, Janeiro de 2019

TERMO DE APROVAÇÃO

Alisson Carlos da Costa Silva

Análise do Problema da Unidade de Área Modificável Pela Regressão Geograficamente Ponderada

Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial à obtenção do título de Mestre em Estatística.

Prof. Dr. Alan Ricardo da Silva
Departamento de Estatística - EST/UnB
Orientador

Prof. Dr. André Luiz Fernandes Cançado
Departamento de Estatística - EST/UnB

Prof. Dr. Pedro Henrique Melo Albuquerque
Departamento de Administração - ADM/UnB

Brasília, 19 de Janeiro de 2019

Ficha Catalográfica

Silva, Alisson Carlos da Costa

Análise do Problema da Unidade de Área Modificável
pela Regressão Geograficamente Ponderada,
(UnB - IE, Mestre em Estatística, 2019)

Dissertação de Mestrado - Universidade de Brasília.
Departamento de Estatística - Instituto de Ciências Exatas.
Orientação: Alan Ricardo da Silva.

1. MAUP 2. Regressão Geograficamente Ponderada
3. Análise espacial 4. Agregação espacial

“Sê valente..."(Js 1:9)

Agradecimentos

Agradeço ao Criador de todas as coisas por sua presença contínua e inegável em todos os momentos de minha vida.

À minha Família, pelo apoio, cuidado e paciência nessa jornada.

Ao meu orientador, Professor Alan, pela disposição, paciência e inspiração.

Aos amigos do mestrado que caminharam comigo me apoiando e trazendo alegria nesses dois anos. Em especial aos do grupo de estudos: Adolfo, Alessandra, Erique, Geiziane e Márcia.

Aos professores do Departamento de Estatística, em especial aos professores Raul, Gladston e Ana Maria Nogalles, pelo apoio e encorajamento nesse período.

Aos amigos da Codeplan, em especial ao Bruno Cruz, Miriam, Iraci e Mirna. Vocês me inspiram a caminhar e acreditar que vale a pena investir em uma melhor qualidade de serviço público oferecido. Ao Arthur e Otávio, que fizeram essa caminhada um tanto mais leve, dividindo conosco a rotina do Núcleo de Estatística.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

O Problema da Unidade de Área Modificável, ou do inglês MAUP (*Modifiable Areal Unit Problem*) é caracterizado por situações em que a agregação espacial de unidades de dados influencia os resultados finais. Os estudos mais antigos sobre o MAUP datam de 1934 com Gehlke e Biehl (1934), porém ainda não há solução definitiva. Com a introdução da Regressão Geograficamente Ponderada (RGP) criada por Brunsdon et al. (1996), uma nova abordagem foi dada ao problema. Como uma fonte do MAUP está relacionada a heterogeneidade espacial, e a RGP pode modelar a variabilidade local, acredita-se que ela seja menos sensível aos efeitos do MAUP. No entanto, a RGP apresenta a limitação de não incorporar mecanismo de agregação de dados em sua estrutura. Murakami e Tsutsumi (2015) propuseram uma adaptação da RGP, a Regressão Geograficamente Ponderada Área Para Ponto (RGP-APP) que incorpora em sua estrutura mecanismos de agregação que permitem a estimação de parâmetros a nível dos dados desagregados, a partir de dados agregados. Neste trabalho, a RGP-APP foi aplicada a dados simulados e reais e os resultados mostraram uma capacidade limitada para a eliminação dos efeitos do MAUP. No entanto, a RGP-APP apresenta capacidade de mitigar os efeitos da agregação, e tem resultados satisfatórios quando comparados com resultados produzidos por modelos OLS e RGP aplicados a dados agregados.

Palavras-chave: MAUP; regressão geograficamente ponderada; análise espacial; agregação espacial.

Abstract

The Modifiable Areal Unit Problem(MAUP) is characterized by situations in which the spatial aggregation of data units influences the final results. Older studies on Maup date from 1934 with Gehlke e Biehl (1934), but there is still no definitive solution. With the introduction of the Geographically Weighted Regression(GWR), developed by Brunsdon et al. (1996), a new approach was given to the problem. As a source of MAUP is related to the spacial heterogeneity, and GWR can model local variability, it is believed that this method is less sensible to the effects of MAUP. However, GWR introduces the limitation of not incorporating the data aggregation mechanism into its structure. Murakami e Tsutsumi (2015) proposed an adaptation of GWR, called Area-to-Point Geographically Weighted Regression(ATP-GWR), that incorporates in its structure aggregation mechanisms that allow the estimation of parameters at disaggregated level of the data, from aggregated data. In this work, ATP-GWR was applied to simulated and real data. The results showed a limited capacity to eliminate the effects of MAUP. However, ATP-GWR has the ability to mitigate the effects of aggregation, and has satisfactory results when compared to the results produced by OLS and GWR models applied to aggregated data.

Keywords: MAUP; geographically weighted regression; spacial analysis; spatial aggregation.

Introdução

O Problema da Unidade de Área Modificável, ou do inglês MAUP (*Modifiable Areal Unit Problem*) é um problema caracterizado por situações em que a agregação espacial de unidades de dados influencia os resultados numa análise espacial para o mesmo conjunto de dados. O termo MAUP foi utilizado pela primeira vez por Openshaw e Taylor (1979) quando eles avaliaram a variação dos valores do coeficiente de correlação quando pequenas unidades de áreas eram agregadas para formar grandes unidades de áreas de forma hierárquica ou não. Eles chegaram a conclusão de que o coeficiente de correlação poderia assumir uma gama de valores distintos, dependendo do nível de agregação espacial. Openshaw e Taylor (1979) afirmam que a origem do problema está no fato de que os limites das unidades de áreas foram criados artificialmente e poderiam ser alterados conforme a conveniência. Quando esses limites são alterados, análises de dados tabulados a partir de diferentes limites produzirão resultados diferentes.

Para Openshaw e Taylor (1979), os efeitos do MAUP podem ser decompostos em dois componentes: efeitos de zoneamento e efeitos de escala. Enquanto o efeito de zoneamento refere-se a variabilidade introduzida por diferentes configurações de zoneamento no mesmo nível de agregação, o efeito da escala descreve a ocorrência de variações de resultados estatísticos usando dados agregados em diferentes níveis. Existem várias maneiras de particionar uma região, mesmo que o número de unidades de área seja mantido constante.

Quando o número de unidades de área é fixado ou relativamente estável, mas os limites são redesenhados para acomodar mudanças, basicamente ocorre o processo de zoneamento. Os dados coletados de acordo com diferentes sistemas de zoneamento da mesma região produzirão diferentes representações da região e diferentes resultados

analíticos quando os dados forem analisados. A inconsistência dos resultados baseados em dados de diferentes sistemas de zoneamento são os mais recorrentes quando se trabalha com dados espacialmente agregados. Outro processo através do qual o espaço é particionado está associado a dimensão da escala. Dada uma região de estudo, podemos particionar a região a diferentes níveis de detalhe ou resolução espacial.

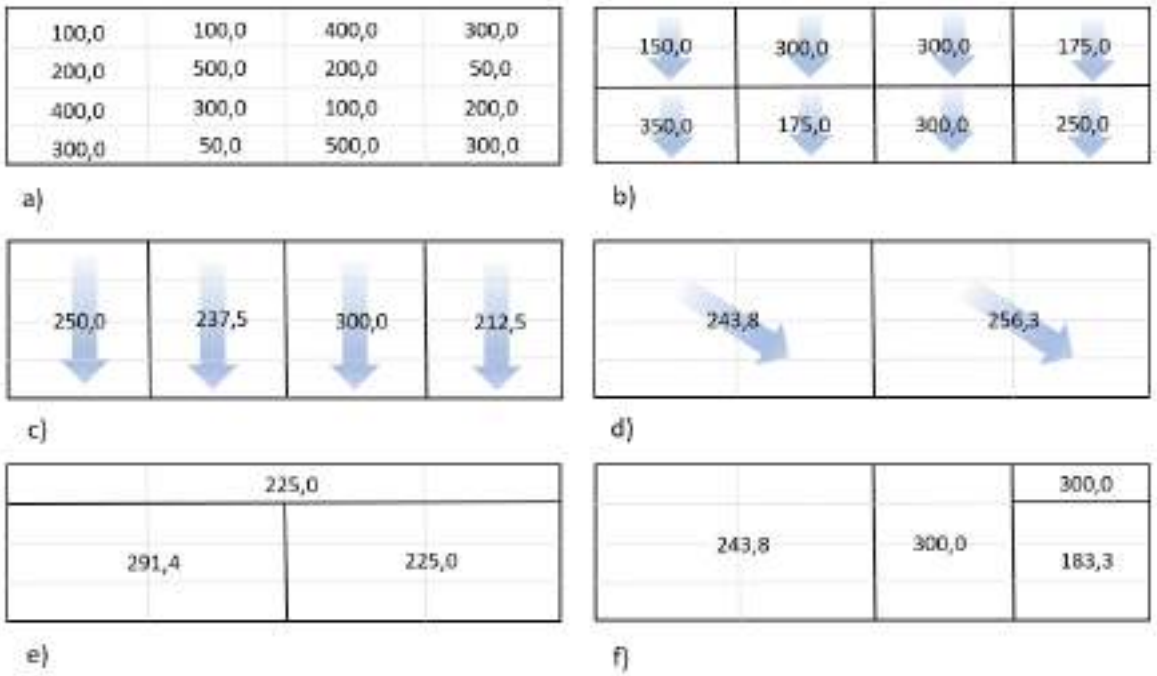


Figura 1: Nível desagregado: **(a)**; Efeito Escala em relação a **(a)**: **(b)**,**(c)**,**(d)** e **(f)**; Efeito Zoneamento em relação a **(e)**: **(f)**.

A Figura 1 ilustra um exemplo de agregação de áreas, onde a Figura 1(a) apresenta as unidades básicas de área. A Figura 1 (b), (c) e (d), mostra agregações das unidades da Figura 1(a), mantendo a mesma quantidade de áreas por agregação. Por exemplo, na Figura 1(b), cada nova área possui duas unidades de área da Figura 1(a); na Figura 1(c), as novas agregações possuem 4 unidades cada; já as novas agregações da Figura 1(d) possuem 8 unidades cada. Os efeitos dessas de agregações podem ser vistos na Tabela 1, onde os dados da Figura1(a) foram agregados e representados para as demais figuras pelas médias obtidas a partir de suas respectivas agregações. Note que apesar das diferentes quantidades de unidades agregadas na Figura 1(b), (c) e (d), as médias calculadas a partir dessas agregações são iguais à média calculada no nível desagregado (Figura 1(a)). No entanto, a variância dos dados tende a diminuir quanto maior for o nível de agregação, ou seja, quanto maior a quantidade de unidades

agregadas. Nesses casos, temos o processo de agregação em escala, onde as menores unidades de áreas da Figura 1(a) estão agregadas na Figura 1(b), (c) e (d), formando novos blocos. Ainda, na Figura 1(e) e (f), as unidades da Figura 1(a) são agregadas, mas agora formando novas áreas de tamanhos distintos. Os efeitos disso podem ser observados na Tabela 1, onde a média obtida a partir das áreas agregadas não são iguais à média a nível desagregado (Figura 1(a)) e nem iguais às médias das agregações representadas na Figura 1(b), (c) e (d). Em relação à Figura 1(a), essas duas novas configurações, também, produzem o efeito de escala.

Tabela 1: Média e desvio padrão por tipo de agregação

Agregação	Média	Desvio
A	250,0	149,4
B	250,0	74,4
C	250,0	27,0
D	250,0	36,8
E	247,2	38,5
F	256,8	55,7

No efeito escala, à medida que unidades de área menores são agregadas para formar unidades maiores, de forma hierárquica, valores originais das unidades menores com algum nível de variabilidade são resumidos ou substituídos por um valor representativo, que, na maioria dos casos, é uma medida de tendência central, como a média ou a mediana. Os valores extremos entre as unidades menores são agora removidos e, portanto, os dados mais agregados estão se tornando menos variados ou mais semelhantes. Assim, as correlações entre as variáveis tendem a ser maiores com níveis mais altos de agregação espacial. Esta natureza do efeito da escala foi melhor exemplificada pela trabalho de Openshaw e Taylor (1979), que mostra que o coeficiente de correlação poderia ter uma ampla gama de valores a um nível moderado para dados relativamente desagregados, até um nível de correlação muito alto para dados altamente agregados.

A Figura 1(e) e (f) ilustra o processo de zoneamento entre elas, onde os limites foram redesenhados de forma que as novas áreas não guardam relação de hierarquia, ou seja, as unidades da Figura 1(f), não contém as unidades da Figura 1(e). Na Figura 1(e) temos três novas áreas sendo uma formada por três unidades de área e as outras duas por 6 unidades, cada uma. Na Figura 1(f) as novas áreas são formadas por 4

polígonos irregulares, sendo que uma contém 1 unidade de área, outra 3 unidades, outra 4 e a última 8 unidades de área. Note ainda que a Figura 1(f) representa uma agregação que provocará o efeito de escala em relação a Figura 1(d) e que a agregação representada em 1(e) produzirá o efeito de zoneamento em relação a 1(d), uma vez que na primeira situação existe uma relação de hierarquia e na última não.

Como exemplos reais de agregações que podem caracterizar o efeito zoneamento, pode-se citar as zonas criadas pelas concessionárias de serviços públicos para definição de suas áreas de planejamento. Normalmente essas zonas têm como base ou a estrutura viária, ou a estrutura de rede de abastecimento de água ou energia. Muitas vezes essas zonas não apresentam relação entre si e nem com a divisão político-administrativa de uma cidade. Já para o efeito escala, a estrutura censitária é um excelente exemplo, onde domicílios são agregados em setores censitários, que por sua vez são agregados em subdistritos, que são agregados em distritos que compõem os municípios. Como será visto no Capítulo 1, a maior parte dos esforços produzidos pelos pesquisadores esteve voltada para solucionar o efeito escala.

Para avaliar os efeitos do MAUP, Wong (2009) utilizou os dados dos distritos congressionais da Geórgia e os dados censitários de Washington. No exemplo da Geórgia/EUA, os limites dos distritos congressionais (DCs) mudaram significativamente entre os 107º e 109º Congressos. A Figura 2(a) e (b) mostra dois mapas dos 107º e 109º distritos congressionais (DCs) na Geórgia. Os dois sistemas de particionamento tem padrões espaciais muito diferentes, embora apenas dois distritos tenham sido adicionados no 109º Congresso. Nenhum distrito antigo no 107º Congresso manteve seu território no 109º. Outra mudança é que a área em torno da área metropolitana de Atlanta tornou-se muito mais fragmentada espacialmente para acomodar mais DCs. A Figura 2 mostra apenas as mudanças de fronteira sem demonstrar os impactos potenciais na análise devido ao efeito de zoneamento. Os dados do Censo de 2000 foram tabulados de acordo com os limites do 107º Congresso, a fim de avaliar como a redefinição afetou as características dos DCs. A partir daí foram verificados os percentuais de negros e brancos nos Distritos no 107º e 109º Congressos. De forma geral, as estatísticas apresentaram importantes alterações entre o 107º e o 109º Congressos.

Para o efeito de zoneamento, segundo Wong (2009), o processo e os impactos gerais parecem ser mais difíceis de avaliar e compreender. Existem várias variáveis que

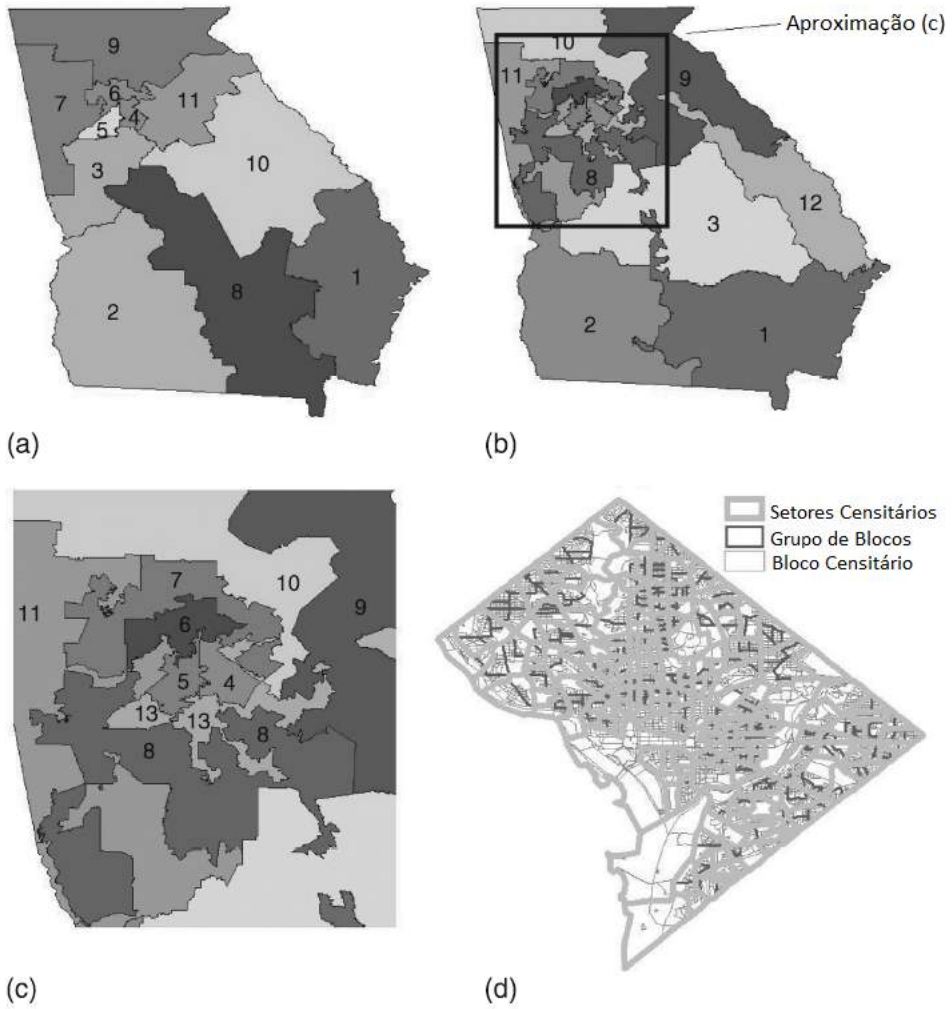


Figura 2: (a)Distritos congressionais - 107º Congresso; (b) Distritos congressionais - 109º Congresso A; (c) Distritos congressionais (zoom) - 109º Congresso; (d) Estrutura censitária - Censo 2000.

Fonte: Wong (2009)

atuam de forma independente e conjunta para determinar os impactos do efeito de zoneamento. Para ilustrar os papéis de algumas dessas variáveis, Wong (2009) fez uso de um estudo de simulação. Em seus resultados chegou à conclusão de que o efeito de zoneamento será mínimo se o fenômeno exibir um padrão um pouco aleatório. Mas se o fenômeno exibir uma forte autocorrelação espacial positiva, deve-se esperar alguns impactos significativos devido ao efeito de zoneamento. Além da distribuição espacial dos dados, outro fator importante na determinação dos impactos do MAUP é o mecanismo de agregação espacial, ou o processo usado para derivar um valor representativo para o unidades agregadas. Nas simulações realizadas por Wong (2009), foi utilizada

a média como processo, isto é, o valor médio dos dados originais foi usado para a unidade agregada. Mas há outras opções possíveis para valores representativos, como mediana, mínimo, máximo e outros.

Um outro fenômeno causado pela agregação de áreas é o que se chama de "falácia ecológica" ou "paradoxo de Simpson", caracterizado por situações onde os coeficientes de correlação podem ser inteiramente diferentes no indivíduo e nas áreas, levando a conclusões impróprias de relacionamentos a nível individual a partir de resultados agregados ao nível de unidade de área (Câmara et al., 2002). Por exemplo, para uma amostra de indivíduos onde são medidas duas características, é ajustado um modelo de regressão. Na Figura 3(a) está representada a reta ajustada para todo o conjunto de dados sem considerar a classificação por uma terceira variável. Nesse caso, percebe-se uma relação negativa entre as duas características observadas. Na Figura 3(b), os dados estão classificados por uma terceira característica que está discriminada pelas cores e ajusta-se uma reta para cada grupo. Os resultados nessa última situação mostram um comportamento diferente do obtido na situação anterior, pois agora a relação entre as duas variáveis é positiva, quando avaliada em cada um dos grupos.

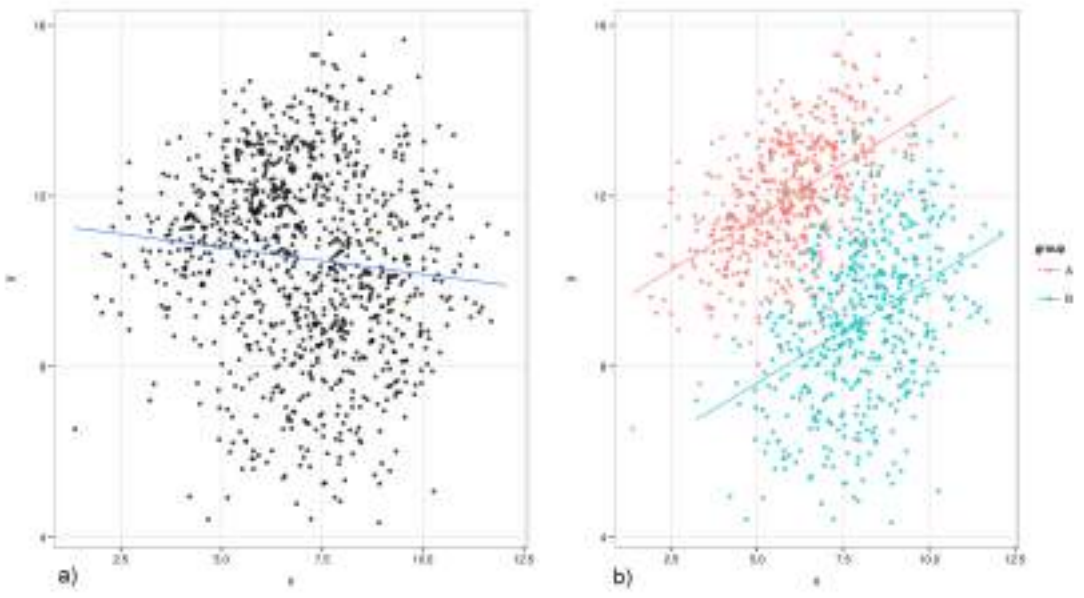


Figura 3: a)Reta ajusta sem classificação dos indivíduos; b) Retas ajustadas para cada grupo de indivíduos.

Fonte: Câmara et al. (2002)

Devido à necessidade de se preservar a confidenciabilidade de registros e impedir

que ocorra a identificação de indivíduos, grande parte dos bancos de dados produzidos em pesquisas domiciliares são disponibilizados de forma desagregada, porém sem a identificação de onde foram coletados. Essa é uma prática comum no caso da divulgação dos dados da amostra do Censo Demográfico Brasileiro ou da Pesquisa Nacional por Amostra de Domicílios - PNAD, onde os dados são disponibilizados a nível de pessoas e/ou domicílios, mas não se tem qualquer informação sobre em qual setor censitário estão localizados. Isso acaba por dificultar análises que utilizam referências geográficas, como por exemplo a utilização de modelos espaciais.

É possível que em alguns estudos deseje-se identificar relações causa-efeito entre diferentes medidas, utilizando modelos com dependência espacial. Um exemplo clássico é correlacionar anos de estudo do chefe de família e sua renda, que usualmente apresenta forte correlação. Uma vez que localidades de alta renda estão próximas entre si e o mesmo ocorre com localidades de baixa renda, torna-se mais adequado o uso de modelos com dependência espacial. A natureza desses dados já conduz a uma situação em que se pode observar o problema do MAUP, pois a definição espacial das fronteiras dos setores censitários pode afetar os resultados obtidos.

A aplicabilidade do estudo do MAUP se dá em várias áreas do conhecimento. Parenteau e Sawada (2011) investigaram os efeitos do uso de diferentes agregações espaciais nas conclusões sobre a relação entre NO2 e a morbidade respiratória. Ávila e Monasterio (2008) investigaram os efeitos do MAUP no estudo de associações espaciais globais da variável renda per capita para o Estado do Rio Grande do Sul. Xu et al. (2015) abordaram o problema do MAUP na análise de segurança de trânsito realizando um estudo de caso para demonstrar a existência do problema na modelagem de colisões de nível macro e, em seguida, foram apresentadas quatro estratégias potenciais para lidar com os efeitos do MAUP: O uso de dados desagregados quando possível, a captura da não estacionaridade espacial, a projeção de sistemas de zoneamento ótimos e a realização de análise de sensibilidade para registrar o escopo e a magnitude do MAUP. Murakami e Tsutsumi (2015) propuseram um tipo de Regressão Geograficamente Ponderada (RGP) com mecanismos de agregação, chamado Área Para Ponto (APP), ou do inglês *area-to-point* (ATP), com a intenção de atenuar o MAUP por meio de um estudo de simulação e um estudo empírico. A RGP-APP, que está intimamente relacionada com as abordagens geoestatísticas, estima os parâmetros de

tendência num nível de desagregação local utilizando variáveis agregadas. Apesar de todos os esforços, desde Openshaw e Taylor (1979), ainda não foi identificada uma solução definitiva para o MAUP.

Neste trabalho, pretende-se analisar os efeitos do MAUP por meio do método RGP-APP em dados simulados e em um estudo de caso real. A aplicação aos dados da Pesquisa Distrital por Amostra de Domicílio - PDAD 2018, conduzida pela Companhia de Planejamento do Distrito Federal - CODEPLAN torna-se interessante, pois o acesso a dados individuais dos domicílios pesquisados traz a vantagem da comparação dos resultados obtidos por meio de agregações com os resultados obtidos a nível de indivíduos.

O trabalho está organizado da seguinte forma: No Capítulo 1 são introduzidos os conceitos do MAUP e seus efeitos. No Capítulo 2 a Regressão Geograficamente Ponderada é abordada, tratando também das adaptações propostas por Murakami e Tsutsumi (2015). No Capítulo 3 são apresentados os materiais e métodos empregados e no Capítulo 4 os resultados dos estudos de simulação e aplicação aos dados da PDAD 2018. Por fim, as conclusões, limitações do trabalho e recomendações para trabalhos futuros estão no Capítulo 5.

Capítulo 1

O MAUP

1.1 Descoberta e Avaliação de Impacto

Os impactos do MAUP foram documentados minuciosamente nos últimos 80 anos. Considerando que alterações nos padrões das correlações constituem um impacto típico e fundamental do MAUP, espera-se que uma parte importante das análises estatísticas estejam sujeitas ao MAUP.

Embora Openshaw e Taylor (1979) tenham criado o termo, muitos pesquisadores antes deles haviam documentado alguns aspectos do MAUP. O mais antigo parece ser o trabalho de Gehlke e Biehl (1934), que constataram em seus estudos alguns padrões de mudanças no coeficiente de correlação quando os setores censitários foram agregados de diferentes formas. O trabalho de Robinson (1956) também deu um passo importante ao argumentar que era necessário um esquema de ponderação para corrigir o coeficiente de correlação para suportar os diferentes números de observações entre as unidades de área. Embora não tenha sido direcionado especificamente ao MAUP, Moellering e Tobler (1972) ofereceram uma melhor compreensão do processo de suavização do efeito de escala, explicando como a variância se altera conforme o nível de escala utilizado.

O trabalho de Sawicki (1973), e mais tarde o de Clark e Avery (1976), estiveram entre as primeiras tentativas de avaliar os efeitos de MAUP nas análises estatísticas gerais. Perle (1977) liga de forma direta o MAUP à questão da falácia ecológica, embora os potenciais problemas de utilização da correlação ecológica para inferir com-

portamento individual tenham sido bem documentados por Robinson (1956). Paralelamente, alguns geógrafos britânicos se concentraram em uma questão relacionada ao desenvolvimento de sistemas zonais ideais, em parte para fins de regionalização e em parte para lidar com o problema MAUP. A criação de zonas ou regiões é muitas vezes necessária na análise regional, e essas zonas ou regiões fornecem a base para os modelos de locação alocação. Goodchild (1979) foi o primeiro a reconhecer o efeito MAUP nos modelos de locação alocação. Os pesquisadores que se dedicavam à modelagem matemática ocasionalmente trabalharam nesse tópico (Fotheringham et al. (1995),Horner e Murray (2002)), mas esses estudos limitaram-se a avaliar os impactos do MAUP.

Depois que Openshaw e Taylor (1979) inventaram o termo MAUP em 1979, o próximo grande esforço conjunto para abordar o MAUP teve início em 1989, em parte devido à iniciativa de pesquisa do Centro Nacional de Análise de Informação Geográfica (NCGIA) sobre a precisão dos dados. Em meio a esses trabalhos, surgiram produções intermitentes na identificação de diferentes aspectos do MAUP. Batty e Sikdar (1982) desenvolveram a abordagem baseada em entropia para lidar com o problema da agregação no contexto do desenvolvimento de modelos gravitacionais. Putman e Chung (1989) também se juntaram aos geógrafos britânicos para abordar problemas de delineamento de zonas para modelos de interação espacial. Blair e Miller (1983) demonstraram os impactos do MAUP nos modelos de insumo-saída.

A formação do NCGIA e o lançamento da iniciativa de pesquisa de precisão de dados espaciais criaram um impulso para a pesquisa do MAUP desde 1989. Fotheringham (1989) reivindicou o reconhecimento de problemas de sensibilidade à escala na análise espacial, bem como a necessidade de realizar operações para a análise multiescala. No mesmo volume, Tobler (1989) argumentou que o MAUP é um problema espacial e, portanto, a solução deveria ser de natureza espacial. Posteriormente, ele propos uma estrutura de modelagem de migração que não era sensível às mudanças de escala, provavelmente a primeira técnica analítica espacial independente da escala a ser introduzida. Sem relação com o desenvolvimento do NCGIA, Arbia (1989) publicou uma monografia altamente detalhada sobre o MAUP.

1.2 Conceituação

Com a iniciativa de pesquisa do NCGIA, uma nova onda de atividades de pesquisa relacionadas ao MAUP surgiu no início da década de 1990, começando pelo artigo de Fotheringham e Wong (1991), um artigo frequentemente citado, abordando sistematicamente os impactos do MAUP em modelos de análise de correlação e regressão. Enquanto os pesquisadores ainda estavam interessados nos impactos do MAUP, a comunidade gradualmente se moveu para encontrar as soluções para o problema. Esta busca de soluções surgiu paralelamente ao esforço de vários pesquisadores que haviam fornecido evidência de que os efeitos de MAUP poderiam não ser tão intensos quanto alguns outros reivindicaram. Amrhein e Flowerdew (1992) mostraram que o MAUP tem um impacto limitado nas regressões de Poisson. Tentando identificar quando o MAUP seria significativo, Amrhein (1995) e Amrhein e Reynolds (1996) realizaram uma série de simulações, controlando várias propriedades estatísticas dos dados, incluindo vários níveis de autocorrelação espacial. Eles concluíram que os efeitos de MAUP podem não ser significativos, dados certos níveis de correlação espacial entre variáveis, mas suas relações são extremamente complexas.

Embora a maioria das análises de impacto do MAUP tenha se concentrado na modelagem estatística ou matemática, algumas análises foram mais estreitamente focadas nas formulações de índices, particularmente usando índices para mensuração da segregação (Wong, 1996). Além de conceituar o efeito da escala na mensuração da segregação, essa linha de pesquisa também mostra que as medidas espaciais são provavelmente mais sensíveis à mudanças de escala do que medidas não espaciais (Wong, 2004). Um esforço coordenado durante esta fase da pesquisa MAUP foi a publicação de uma edição especial de *Geographical Systems* (Wong e Amrhein, 1996). Nesta edição especial, alguns pesquisadores ainda se concentraram nos efeitos de MAUP (por exemplo, Okabe (1996)), mas outros aprofundaram seus estudos sobre as fontes do MAUP, incluindo o conceito de mudança de suporte em geoestatística (Cressie et al., 1998).

Uma direção evidente foi desenvolver soluções. Holt et al. (1996) argumentaram que a origem do efeito da escala estava nas mudanças da correlação entre as variáveis e, portanto, propuseram uma estrutura para modelar as mudanças de correlação, le-

vando em consideração a autocorrelação espacial de forma implícita. Infelizmente, a complexidade do método computacional foi além de uma solução prática para o problema. A criação de um zoneamento ideal foi considerada como uma solução potencial para o MAUP no passado (Openshaw, 1977), e essa direção ainda era interessante nesta fase da pesquisa (Openshaw e Schmidt, 1996).

A maior parte da pesquisa sobre o MAUP mencionado acima manteve seu foco na agregação de dados de características de polígonos, uma operação comum na manipulação de dados de formato vetorial em SIG e frequentemente usada no tratamento de fenômenos socioeconômicos. No entanto, os impactos do MAUP também estão presentes na geografia física, modelagem ambiental e, em geral, na análise de dados de formato *raster*. Fora da geografia humana, alguns ecologistas e geógrafos começaram a desenvolver uma apreciação dos problemas de MAUP (Jelinski e Wu, 1996), e uma série de pesquisas seguiram nessa direção. Espa et al. (1996) podem ter sido os primeiros a vincular de forma explicita o efeito de escala na análise de dados de sensoriamento remoto ao MAUP, mas o efeito de escala ou dependência de escala não era algo novo para cientistas de sensoriamento remoto (Bian e Walsh, 1993)), já que os dados de detecção estão frequentemente disponíveis e podem ser facilmente tabulados em múltiplos níveis de escala (Bian, 1997).

Parte da questão, que historicamente tem sido um problema na análise de detecção remota, é selecionar a resolução apropriada para a análise (por exemplo, Townshend e Justice (1988)). Lam e Quattrochi (1992) analisaram vários conceitos relacionados à escala e resolução, tentando abordar a questão da escolha da escala ideal ou da resolução para análise de um fenômeno particular. Alguns pesquisadores também reconheceram que o efeito da escala é essencialmente um problema de mudança de suporte na geoestatística (Atkinson e Curran, 1995). O volume editado de Quattrochi e Goodchild (1997) coletou documentos que abordavam os impactos do MAUP em sensoriamento remoto, na modelagem do efeito de escala e no desenvolvimento de soluções (por exemplo, Bian (1997)). Ainda assim, nenhuma solução clara foi identificada.

Fora da literatura geográfica, o MAUP atraiu atenção adicional após a publicação da monografia de King (1997), que se concentrou em questões de inferência em disciplinas de ciências sociais, abordando também o MAUP. Sua afirmação de que uma

abordagem vinculada ao erro poderia resolver o efeito da escala, que estaria relacionado conceitualmente ao problema da falácia ecológica, provocou reações no meio geográfico. As respostas de Fotheringham et al. (2000) e Anselin (2000) não eram tão otimistas quanto ao fato de as soluções de King (1997) resolverem o problema da falácia ecológica e especificamente o MAUP.

1.3 Soluções Potenciais

Mesmo que na fase inicial da pesquisa os pesquisadores estivessem focados na generalização dos efeitos de MAUP e nos estudos de "análise de impacto", eles nunca pararam de buscar soluções. Robinson (1956) sugeriu métodos de ponderação simplistas para superar alguns dos efeitos de MAUP na análise de correlação.

Tobler (1989) argumentou que, pelo fato do MAUP ser um problema espacial, as soluções deveriam ser de natureza espacial. Assim, ele solicitou o desenvolvimento de técnicas analíticas espaciais insensíveis à escala para lidar com o MAUP e empregou um modelo de migração populacional que era relativamente insensível às mudanças de escala. O modelo de migração de Tobler (1989) é uma das poucas ferramentas analíticas que são relativamente insensíveis à escala. Outro que demonstrou algum nível de estabilidade na correlação em diferentes níveis de escala foi Wong (2001) que propôs a análise de correlação específica da localização. Mas todas essas ferramentas tem aplicações limitadas.

Uma "solução" espacial popular para o MAUP, mesmo antes de Openshaw e Taylor (1979) inventarem o termo, era criar sistemas de zoneamento ótimos (Openshaw, 1977). Dado que a maioria dos problemas de agregação envolvem múltiplas variáveis, as derivações dos sistemas zonais devem basear-se em variáveis múltiplas e múltiplos objetivos. Em geral, o princípio é criar sistemas zonais para minimizar as variações intrazonais e maximizar as variações interzonais. Mas muitas vezes não existe uma solução única e, portanto, os processos heurísticos parecem ser bem promissores (Wei e Chai, 2004).

Tate e Atkinson (2001) sugeriram a análise geoestatística como uma solução potencial para o problema da escala. As ferramentas geoestatísticas, especialmente os variogramas, podem identificar o alcance geográfico da autocorrelação espacial. Esta

é uma informação importante para entender e modelar o efeito escala. Atkinson e Tate (2000) alegaram que as ferramentas geoestatísticas não são usadas para redimensionar os dados, mas para redimensionar as estatísticas que descrevem os dados. Esta é uma ideia interessante, mas não foi totalmente validada ou operacionalizada. Mais recentemente, após a introdução da Regressão Geograficamente Ponderada (RGP), o uso da RGP para descrever a heterogeneidade espacial relacionada ao MAUP foi mais abordado (Fotheringham et al., 2000). Como uma fonte importante do efeito escala é a heterogeneidade espacial e a RGP pode modelar a variabilidade local, acredita-se que RGP pode ser mais robusta do que outros modelos globais e menos sensível ao efeito da escala (Fotheringham et al., 2002). Apesar disso, a RGP ainda não pode ser considerada como uma solução para o MAUP.

Um grupo de estatísticos da área social (Holt et al., 1996; Steel e Holt, 1996; Tranmer e Steel, 1998) tomou uma direção semelhante à abordagem geoestatística para redimensionar estatísticas em vários níveis de escala. Eles perceberam que o efeito escala pode ser reduzido ao mínimo quando as áreas agregadas possuem alto grau de homogeneidade interna (baixa variação) e a magnitude do efeito da escala será em parte uma função da homogeneidade interna. Como resultado, pode-se modelar o efeito de escala ou as estatísticas que descrevem os dados em diferentes níveis de escala, desde que possamos estabelecer as regras de agregação e como o efeito da escala está relacionado ao nível de homogeneidade interna. Uma vez que a base da maioria das estatísticas clássicas é a matriz variância-covariância, este grupo de pesquisadores propos usar a correlação no nível individual para estimar a correlação no nível agregado e então poder estimar a matriz de variância-covariância do nível agregado. As derivações estatísticas envolvidas foram muito sofisticadas e demandaram um grande esforço matemático e computacional. Como resultado, esta não foi uma solução prática para o MAUP.

Embora esforços enormes tenham sido empregados para lidar com o problema de escala, para alguns pesquisadores o problema de zoneamento parece ser mais fácil de lidar. Flowerdew e Green (1989) trataram do problema do zoneamento da mesma forma que resolver sistemas zonais incompatíveis. A abordagem geral consiste em usar métodos de interpolação espacial para transformar dados coletados de acordo com um padrão zonal para outro padrão. Fisher e Langford (1995) avaliaram a con-

fiabilidade dessa técnica no tratamento do problema de zoneamento. Uma técnica relacionada, o mapeamento dasimétrico, também mostrou ser eficaz para lidar com padrões zonais incompatíveis de uma perspectiva cartográfica (Fisher e Langford, 1996). Uma técnica mais antiga de suavização ou interpolação, a interpolação picnofilática suave introduzida por Tobler (1979), também foi revisitada e acredita-se ser uma solução candidata para o MAUP, especificamente em abordar o problema a partir da perspectiva da mudança de suporte (Gotway e Young, 2002).

Em síntese, os efeitos do MAUP podem ser abordados por modelos sofisticados e técnicas computacionalmente intensivas. Técnicas relativamente simples podem lidar com o problema de zoneamento, mas não com o problema de escala. Assim, sem métodos geralmente viáveis para lidar com o MAUP, o antigo apelo ao reconhecimento do MAUP ainda é a abordagem mais acessível para lidar com esse problema persistente a longo prazo (Fotheringham, 1989). Dado os avanços na tecnologia de Sistemas de Informação Geográfica (SIG) e ferramentas computacionais, e a disponibilidade de dados digitais em várias escalas, repetir a mesma análise, mas usando diferentes escalas ou esquemas de particionamento está ao alcance da maioria dos pesquisadores.

Estudos mais recentes tem focado seus esforços numa lista mais sintetizada de possíveis soluções para o problema MAUP. Xu et al. (2015) apresentaram quatro estratégias potenciais para lidar com os efeitos do MAUP: (1)O uso de dados desagregados quando possível, (2) a projeção de sistemas de zoneamento ótimos, (3) a realização de análise de sensibilidade para registrar o escopo e a magnitude do MAUP e (4) a captura da não estacionaridade espacial.

1.3.1 Uso de Dados Desagregados Sempre que Possível

Como o MAUP resulta da agregação de dados, uma solução óbvia seria evitar o uso de dados agregados e trabalhar com os dados individuais. Davis (2004) defendeu uma análise de dados de acidentes de trânsito a nível individual usando uma abordagem de mecanismo para desenvolver uma relação causal. Alternativamente, os modelos estatísticos poderiam ser aplicados empregando os dados da amostragem relativos a casos de colisão e não colisões (Abdel-Aty et al., 2004). No entanto, esta solução não

pode ser aplicada às situações em que o acesso a dados desagregados não é possível devido à natureza das medidas de dados (por exemplo, densidade populacional) ou a razões legais (por exemplo, proteção da privacidade). Além disso, é evidente que as variáveis de nível macro poderiam fornecer informações que não são capturados por dados de nível individual (Diez-Roux, 1998), e uma análise de nível desagregado pode ignorar em grande parte os efeitos do sistema (Kennedy et al., 2007).

Na verdade, o argumento acima poderia ser estendido à utilização de métodos de análise agregados ou desagregados. Os teóricos comportamentais argumentam que é impossível aprender algo sobre o comportamento individual se modelos agregados são utilizados, enquanto pesquisadores de fenômenos agregados argumentam que os modelos comportamentais não conseguem fornecer informações suficientes sobre os efeitos do sistema (Kennedy et al., 2007). Na verdade, há uma compreensão mais profunda do problema. Os fenômenos agregados são influenciados pela forma como os indivíduos se comportam e pela forma como indivíduos interagem uns com os outros. Assim, a reação de um sistema é mais do que a soma direta de como cada um dos componentes do sistema reagiram. Também é dependente das interações complexas entre indivíduos. Da mesma forma, as ações e interações individuais são impactadas pelo contexto do nível macro.

1.3.2 Criação de sistemas de zoneamento ótimos

Ao contrário da perspectiva estatística convencional acima, Openshaw (1984) sugeriu que o MAUP é um problema geográfico que requer uma solução geográfica e não estatística. Com base em reconhecimento de que as atividades de modelagem e calibração não eram independentes do delineamento de zona, Openshaw (1984) defendeu a criação deliberada de sistemas de zoneamento ótimos para a calibração específica do modelo. Ao gerar uma amostra aleatória de 261 sistemas alternativos de 22 zonas e 87 sistemas de 42 zonas do conjunto de 72 unidades espaciais básicas, foram criados sistemas de zoneamento ótimos para um modelo de interação espacial que asseguraram que os mínimos quadrados se ajustassem aos parâmetros fixados em valores arbitrários (Openshaw e Taylor, 1979). Essa ideia foi posteriormente expandida para um Procedimento de Zoneamento Automatizado (PZA) (Martin, 2003).

O PZA é uma abordagem de otimização baseada em heurística que começa com um sistema zonal inicial e refina iterativamente a solução reatribuindo objetos a regiões vizinhas. A vantagem do PZA é que permite a interação entre exercícios de modelagem e as gerações de zoneamento desde que o desempenho do modelo possa ser selecionado como uma das metas de otimização. No entanto, esta metodologia está sujeita a críticas como: (1) os pesquisadores estão projetando unidades espaciais para alcançar um resultado estatístico satisfatório (Swift et al., 2008); (2) o que constitui o ótimo em termos de análise multivariada provavelmente será bastante subjetivo como um sistema de zoneamento, pois o que é ótimo para uma variável pode não ser ótimo para outra (Fotheringham e Wong, 1991); (3) requer alto custo computacional, portanto, incapacidade de processar grandes conjuntos de dados (Guo, 2008).

1.3.3 Realização de análise de sensibilidade para mensurar o alcance e a magnitude do MAUP

Outra solução seria realizar uma análise de sensibilidade que examina o alcance e a magnitude do MAUP pela avaliação de resultados em vários níveis de agregação de dados e com diferentes definições de unidades de área (Fotheringham e Wong, 1991). Através desta análise de sensibilidade, é possível explorar o escopo dos resultados estatísticos que podem ser produzidos para vários sistemas de zoneamento, para determinar quais os fatores que são sensíveis à variação na escala ou na definição da unidade, para qual extensão, e para identificar qual nível de agregação de dados minimiza o impacto do MAUP. A chave da análise de sensibilidade está na forma como os esquemas zonais são gerados. Podem ser produto de uma agregação aleatória com uma restrição de contiguidade (Fotheringham e Wong, 1991), agregação de estruturas de grade regulares (Viegas, 2009), esquemas de agrupamento matemático (Yannis et al., 2007) ou resultado de um método automatizado de regionalização, como por exemplo o denominado regionalização com agrupamento dinâmico limitado e particionado (Xu et al., 2015).

Fotheringham e Wong (1991) realizaram uma análise de sensibilidade para explorar o impacto do MAUP na regressão linear. O procedimento aleatório com uma restrição de contiguidade foi utilizado para agregar 871 grupos de blocos para 800,

400, 200, 100, 50 e 25 unidades de área. Em cada nível de agregação, 20 sistemas de zonas diferentes foram criados. Eles descobriram que os resultados eram totalmente imprevisíveis. Por exemplo, um aumento de 0,1 na proporção de idosos criaria uma diminuição na renda familiar média prevista de apenas $308 com base em 800 zonas. Quando os dados foram agregados em apenas 25 zonas, os resultados sugeriram que o mesmo aumento resultaria em uma redução de $2.654. Além disso, no mesmo nível de agregação alguns dos arranjos de zoneamento relataram uma associação positivamente significativa, enquanto alguns arranjos de zoneamento relataram uma relação negativamente significativa entre as duas variáveis acima. Os resultados não são surpreendentes, pois a produção aleatória não reflete o processo agregado subjacente.

1.3.4 Captura da não estacionariedade espacial

Ao investigar a causa do MAUP, Fotheringham et al. (2002) afirmaram que ele pode ser causado devido à não estacionariedade espacial. Várias variáveis geralmente não variam de forma idêntica no espaço e a relação entre uma variável resposta e variáveis explicativas pode não ser necessariamente constante ou fixa em uma área de estudo. A incapacidade de capturar esta variação espacial local na estimação dos parâmetros pode resultar na questão do MAUP. É precisamente a presença dessa não-estacionaridade espacial que levou ao desenvolvimento de modelos de regressão linear que permitem que os coeficientes variem espacialmente. Dois potenciais modelos concorrentes desse tipo são : a Regressão Geograficamente Ponderada (RGP) ou do inglês *Geographically Weighted Regression* (GWR) (Fotheringham et al., 2002) e os modelos de Regressão Bayesiana com Coeficientes de Variação Espacial (RBCVE) ou do inglês *Spatially Varying Coefficient Process* (SVCP) (Gelfand et al., 2003). A RGP é semelhante em metodologia aos modelos de regressão linear local (Loader, 1999), exceto em que os pesos em RGP são determinados por uma função de kernel espacial ao invés de uma função de kernel na variável espaço. Com relação à RBCVE, os coeficientes de variação espacial são modelados como um processo espacial multivariado. A RBCVE difere da RGP na medida em que é um modelo estatístico único especificado de uma maneira hierárquica, enquanto a RGP é um conjunto de modelos de regressão espacial local, onde cada um se ajusta separadamente.

Recentemente, pesquisas revelaram que a regressão geograficamente ponderada de Poisson superou o tradicional modelo linear generalizado (isto é, modelo de Poisson) na captura de relacionamentos espacialmente não estacionários entre colisões e fatores de previsão (Hadayeghi et al., 2010).

Além dos esforços de Fotheringham et al. (2002) em utilizar a RGP como uma possível solução para o MAUP, mais recentemente outros pesquisadores tem trabalhado na mesma direção, propondo soluções a partir de adaptações realizadas no modelo RGP. Murakami e Tsutsumi (2015) propõem uma adaptação do modelo RGP Clássico que será abordado no Capítulo 2, juntamente com a abordagem clássica da RGP.

Capítulo 2

A Regressão Geograficamente Ponderada - RGP

2.1 Indicadores de Autocorrelação Espacial

A Primeira Lei da Geografia enunciada por Tobler (1979) afirma que "Tudo está relacionado a tudo, mas as coisas mais próximas estão mais relacionadas que as coisas mais distantes". Esse fato é a base para a definição da dependência espacial ou autocorrelação espacial. Câmara et al. (2002) comentam que essa dependência é uma característica inerente à representação dos dados através de subdivisões territoriais, ou seja, os dados de uma determinada área tendem a ser mais parecidos com os de seus vizinhos do que com os de áreas mais distantes. Vale ressaltar que o termo "vizinho" está baseado no padrão espacial adotado: geográfico ou conectividade.

A heterogeneidade espacial diz respeito a aspectos da estrutura socioeconômica do espaço geográfico, ou seja, é o processo em que as respostas variam de lugar para lugar (Anselin, 1988). Essa característica pode ocasionar a instabilidade estrutural (coeficientes variáveis) e variância não constante (heterocedasticidade) que distorcem os resultados do modelo.

Os indicadores de autocorrelação espacial caracterizam a dependência espacial dos dados. Os indicadores globais resumem esta caracterização em um único índice para toda a região de estudo, equanto que os indicadores locais lidam com desagregações dentro dessa região. Na subseção seguinte é inserido o conceito de matriz de

proximidade espacial que é necessário para a descrição desses índices.

2.1.1 Matriz de Proximidades

A matriz de proximidade espacial ou "matriz W", representa o peso ou o grau de conectividade ou de proximidade espacial entre as áreas i e j. Ela representa quantitativamente a estrutura espacial entre as áreas da região de estudo. Assim, dado um conjunto de n áreas, $A_1, \ldots, A_n$, os elementos w_{ij} da matriz $\mathbf{W}$, cuja dimensão é $n \times n$, representam alguma medida de proximidade entre as áreas A_i e A_j (Assunção, 2004). Por definição, a diagonal dessa matriz é nula, isto é, $w_{ii} = 0$ para todo $i = 1, 2, \ldots, n$. Considerando que a área i não sofre influência dela mesma. A definição dos elementos de w_{ij} é subjetiva e depende do fenômeno em estudo. Assunção (2004) apresenta algumas possibilidades descritas a seguir:

1. $w_{ij} = 1$, se A_i faz fronteira com A_j, e $w_{ij} = 0$ caso contrário;

2. $w_{ij} = 1$, se o centróide (ou centro político) de A_i está a uma distância menor do que k quilômetros de A_j, e $w_{ij} = 0$ caso contrário;

3. $w_{ij} = 1/(1 + d_{ij})$, onde $d_{i,j}$ é a distância entre os centróides das áreas A_i e A_j;

4. $w_{ij} = 1/(1 + t_{ij})$, onde $t_{i,j}$ é o tempo necessário para ir de A_i para A_j pela malha rodoviária (Silva, 2006);

É comum padronizar as linhas da matriz $\mathbf{W}$, criando uma nova matriz assimétrica, a fim de facilitar a derivação de fórmulas e as propriedades estatísticas envolvidas. A padronização consiste em fazer com que a soma da linha i seja igual a 1. Isso pode ser feito como:

$$w_{ij}^* = w_{ij}/w_{i.} \tag{2.1}$$

onde $w_{i.} = \sum_{j=1}^{n} w_{ij}$.

Como dito anteriormente, a matriz de proximidade apenas indica a estrutura espacial existente e não fornece informações sobre a existência de dependência espacial. Alguns indicadores específicos, como o I de Moran podem identificar e quantificar tal dependência. A Figura 2.1 ilustra a elaboração de uma matriz de proximidade espacial. Na próxima Seção é apresentado o conceito de indicadores de autocorrelação

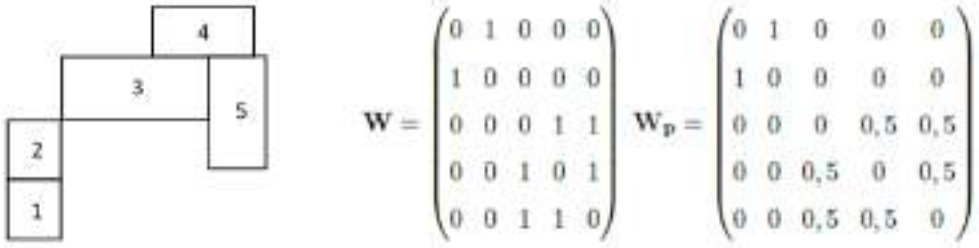


Figura 2.1: Exemplo de elaboração de matriz de proximidade espacial.
Fonte: Adaptado de LeSage (1999)

espacial, com enfoque no I de Moran.

2.1.1.1 Indicadores globais e locais

Os indicadores globais de autocorrelação são úteis na análise exploratória dos dados. O índice I de Moran é um indicador global, utilizado para verificar a existência de correlação espacial no conjunto de dados analisado. Ele apresenta uma única medida para a totalidade da região estudada permitindo testar a hipótese de existência de correlação espacial entre as regiões de acordo com a variável de interesse. O índice I de Moran é dado por (Câmara et al., 2002):

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{2.2}$$

onde n é o número de regiões a serem estudadas, x_i e x_j são os valores da variável de interesse nas regiões i e j e w_{ij} são os elementos da matriz de proximidade. O índice I de Moran é restrito ao intervalo $[-1, 1]$, onde valores próximos de -1 indicam correlação espacial negativa, valores próximos de 1 indicam correlação espacial positiva e um valor igual a 0 indica ausência de correlação espacial em relação à variável testada.

Desenvolvidos por Anselin (1995), os indicadores locais de associação espacial (LISA), são utilizados para verificar a existência de correlação dentro das unidades geográficas estudadas, identificando peculiaridades regionais. A existência de áreas com índices locais significativos, indica o aspecto não estacionário do dado estudado. O índice I_i de Moran local é dado por:

$$I_i = \frac{n\left(x_i - \bar{x}\right)\sum_{j=1}^{n} w_{ij}\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{2.3}$$

2.1.1.2 Diagrama de espalhamento de Moran

Anselin (1995) propôs uma forma gráfica para o índice global de Moran, denominada *Moran Scatterplot* ou "Diagrama de Espalhamento de Moran". O objetivo é comparar o valor do atributo na área A_i com a média dos valores dos atributos nas áreas próximas a A_i. Na forma matricial, o índice I de Moran é dado por:

$$I = \frac{\mathbf{z}'\mathbf{W}\mathbf{z}}{\mathbf{z}'\mathbf{z}} \text{ ou } I = (\mathbf{z}'\mathbf{z})^{-1}\,\mathbf{z}'\mathbf{W}\mathbf{z} \tag{2.4}$$

em que, $\mathbf{z}' =$ vetor $1 \times n$ dos desvios de y; $\mathbf{z}$ vetor $n \times 1$ dos desvios de y; $\mathbf{W} =$ matriz $n \times n$ de proximidade espacial.

Nota-se, a partir da Equação 2.4, que o índice I de Moran é o coeficiente angular da regressão linear de $\mathbf{Wz}$ em $\mathbf{z}$, ou seja, da reta de regressão do diagrama de dispersão de Moran (Camargo et al., 2004). Como o parâmetro β indica a inclinação da reta de regressão, é possível analisar a associação de $\mathbf{z}$ com a média dos seus vizinhos $\mathbf{Wz}$ através da disposição desses pontos ao redor da reta. O Diagrama de Espalhamento

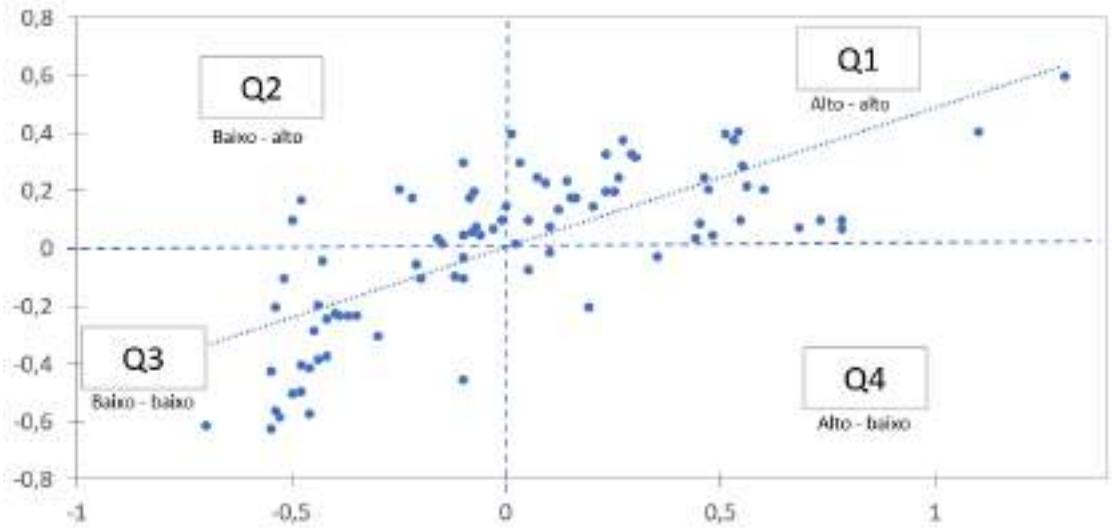


Figura 2.2: Exemplo de um diagrama de espalhamento de Moran.
Fonte: Adaptado de Assunção (2004)

de Moran é dividido em quatro quadrantes (Q1, Q2, Q3 e Q4), conforme a Figura 2.2. Os pontos que estão em Q1 são chamados Alto-Alto (ou *High-High*) por indicarem que para altos valores de $\mathbf{z}$, na média existe altos valores de $\mathbf{Wz}$; os pontos que estão em Q3 são chamados Baixo-Baixo (ou *Low-Low*) por indicarem que para baixos valores

de $\mathbf{z}$, na média existem baixos valores de $\mathbf{Wz}$; os pontos em Q2 e Q4 são chamados de Baixo-Alto (ou *Low-High*) e Alto-Baixo (ou *High-Low*) respectivamente, indicando que para baixos (ou altos) valores de $\mathbf{z}$, na média existem altos (ou baixos) valores de $\mathbf{Wz}$.

A aglomeração no primeiro e no terceiro quadrantes caracteriza uma dependência espacial com maior intensidade, enquanto que a aglomeração de pontos no segundo e quarto quadrante descaracteriza esse fato. Esses últimos sugerem que o fator "espaço" não influencia diretamente na valoração da variável y, caracterizando-se assim, como eventos aleatórios.

Outra forma de visualização do Diagrama de Espalhamento de Moran é dado pelo mapa de espalhamento de Moran (ou do inglês *Box Map*). O *Box Map* é a visualização georreferenciada do diagrama de dispersão de Moran. As áreas da região de estudo são pintadas de quatro cores, representando os quatro quadrantes. A combinação do mapa de espalhamento de Moran com o mapa de indicadores locais dá origem ao mapa de Moran. Seu intuito é indicar quais classificações do mapa de espalhamento de Moran são significativas de acordo com a significância dos índices locais. Portanto, assim como o mapa de indicadores locais, áreas significativas no mapa de espalhamento de Moran também são indícios de não-estacionariedade nos dados.

2.2 Modelos de Regressão

2.2.1 Modelo de Regressão Clássica

Usualmente recorre-se a regressão linear com o objetivo de melhor compreender a dependência funcional de uma variável em relação a uma outra ou em relação a um conjunto de outras variáveis. Em particular, na regressão linear temos uma relação da forma:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \tag{2.5}$$

em que y_i é a i-ésima observação da variável resposta, $i = 1, \ldots, n$, x_{ij} é o valor da j-ésima variável preditora para a i-ésima observação, $j = 1, \ldots, k$, e ε_i é o erro rela-

cionado à i-ésima observação. Supõe-se nesse modelo que os erros são independentes e normalmente distribuídos, com média 0, variância σ^2 e $\varepsilon \sim N(0, \sigma^2)$. Os coeficientes β_j são chamados de parâmetros do modelo e representam a taxa de mudança que uma alteração na variável explicativa acarreta na média condicional da variável resposta. Para manter coerência com o fato de que as inferências sobre a relação entre y_i e x_i assumem o conhecimento de x_i, pode-se escrever (2.5) como:

$$E\left(Y_i|x_i\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \tag{2.6}$$

Cada β_j é determinado por meio de uma amostra de $i = 1, \ldots, n$ observações, geralmente utilizando a técnica de Mínimos Quadrados Ordinários - MQO (ou do inglês, *Ordinary Least Squares* - OLS), cujas estimativas que minimizam a soma dos erros ao quadrado são obtidas matricialmente da seguinte forma:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \tag{2.7}$$

em que $\mathbf{X}$ é uma matriz $n \times (k+1)$ com o valor das variáveis explicativas para cada observação e o intercepto, $\mathbf{y}$ é um vetor $n \times 1$ com o valor da variável resposta para cada observação e $\hat{\boldsymbol{\beta}}$ é um vetor $(k+1) \times 1$ que contém as estimativas para os parâmetros. O valor esperado de $\hat{\boldsymbol{\beta}}$ é dado por:

$$E[\hat{\boldsymbol{\beta}}] = E[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}] = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}] = \boldsymbol{\beta} \tag{2.8}$$

Então $\hat{\boldsymbol{\beta}}$ é não viesado e sua variância é dada por:

$$\begin{aligned} Var(\hat{\boldsymbol{\beta}}) &= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \tag{2.9}$$

Os elementos da diagonal desta matriz são as variâncias dos estimadores dos parâmetros individuais e os elementos fora da diagonal são as covariâncias entre estes

estimadores.

A utilização dos modelos de regressão clássica pressupõe as seguintes características para os dados analisados:

1. Observações não correlacionadas;
2. Erros independentes e identicamente distribuídos;
3. Erros com distribuição normal com média zero e variância constante.

No entanto, Anselin (1988) afirma que dados espaciais agregados são caracterizados pela autocorrelação espacial e pela heterogeneidade ou estrutura espacial. Câmara et al. (2002) comentam que essa dependência é uma característica inerente à representação dos dados através de subdivisões territoriais, ou seja, os dados de uma determinada área tendem a ser mais parecidos com os de seus vizinhos do que com os de áreas mais distantes. Dessa forma, a avaliação dos efeitos espaciais é importante, pois a sua presença invalida os resultados dos modelos tradicionais de regressão, por violarem alguns dos pressupostos considerados.

2.2.2 Modelos de Regressão Espacial

Os modelos de regressão espacial foram desenvolvidos com a finalidade de se incorporar a autocorrelação espacial à estrutura de um modelo. Os modelos de regressão espacial também necessitam dos três principais pressupostos do modelo de regressão convencional, porém ao incorporar em sua estrutura o fator "espaço", eliminam na maioria das vezes, os problemas de autocorrelação e heterocedasticidade (Anselin, 1988).

Os modelos espaciais globais partem do pressuposto que o processo espacial analisado é estacionário. Segundo Anselin (1988), os modelos globais compõem a classe mais simples dentre os modelos de regressão espacial, pois eles captam a estrutura de correlação espacial dos dados em um ou, no máximo, dois parâmetros.

A forma geral de um modelo espacial autoregressivo global (Anselin, 1988) é dado

por:

$$\begin{aligned} \mathbf{y} &= \rho \mathbf{W_1 y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W_2 u} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N\left(0, \sigma^2 \mathbf{I}_n\right) \end{aligned} \tag{2.10}$$

em que ρ e λ são os parâmetros que captam a dependência espacial na variável dependente $\mathbf{y}$ e no erro aleatório $\mathbf{u}$, respectivamente, e $\mathbf{W_1}$ e $\mathbf{W_2}$ são as matrizes de proximidade espacial de $\mathbf{y}$ e $\mathbf{u}$, respectivamente, sendo possível adotar $\mathbf{W_1} = \mathbf{W_2}$.

Quando o padrão de autocorrelação espacial variar na região de estudo, o modelo de regressão espacial global não será capaz de representar adequadamente a dependência espacial dos dados. Nestas situações, modelos de regressão espacial local, que permitem que os parâmentros variem no espaço, são mais recomendados.

Os modelos com efeitos espaciais são utilizados quando o processo é não-estacionário, e por isso, há a necessidade de que os coeficientes da regressão reflitam essa heterogeneidade espacial (Câmara et al., 2002).

A Tabela 2.1, adaptada de Fotheringham et al. (2002), resume as características que diferem esses dois tipos de modelo, em se tratando de dados espaciais.

Tabela 2.1: Características dos modelos globais e locais

Modelos Globais	**Modelos Locais**
Resumem os dados para toda a região em estudo	-Desagregam localmente as estatísticas globais
-Geram uma única estatística	-Geram uma estatística para cada local
-Estatísticas não podem ser mapeadas ou analisadas por um Sistema de Informações Geográficas	-Estatísticas podem e devem ser mapeadas ou analisadas por um Sistema de Informações Geográficas
-Ênfase na similaridade no espaço	-Ênfase nas diferenças no espaço
-Procuram por regularidades ou padrões	-Procuram por exceções ou regiões em destaque

Fonte:Adaptada de Fotheringham et al. (2002)

A RGP é uma técnica utilizada para modelar a tendência espacial de forma contínua, com parâmetros variando no espaço. A ideia é ajustar um modelo de regressão a cada ponto observado, ponderando todas as demais observações como função da distância a esse ponto. Esse modelo foi desenvolvido por Brunsdon et al. (1996).

2.2.2.1 O modelo de Regressão Geograficamente Ponderada

A RGP estende a estrutura da Regressão Clássica, dada em (2.5), permitindo que os parâmetros locais sejam melhor estimados do que os parâmetros globais. Assim, o modelo RGP pode ser escrito como:

$$y_i = \beta_0 (u_i, v_i) + \sum_{k=1}^{p} \beta_k (u_i, v_i) x_{ik} + \varepsilon_i \tag{2.11}$$

em que (u_i, v_i) corresponde às coordenadas do i-ésimo ponto no espaço e $\beta_k(u_i, v_i)$ é o resultado da função contínua no ponto i, possibilitando dessa forma, definir uma superfície contínua dos valores dos parâmetros que refletem a variabilidade espacial. Em caso de dependência espacial, espera-se que os parâmetros estimados de localidades próximas tenham magnitudes e sinais relativamente similares.

Note que a Regressão Clássica (2.5) é um caso especial da Regressão Geograficamente Ponderada (2.11). Esta simplificação ocorre quando não há variação espacial nos parâmetros.

A forma matricial de (2.11) é dada por

$$\mathbf{y} = (\boldsymbol{\beta} \otimes \mathbf{X})\mathbf{1} + \boldsymbol{\varepsilon} \tag{2.12}$$

em que $\otimes$ é o operador que denota a multiplicação elemento por elemento e $\boldsymbol{\beta}$ é da forma:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \dots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \dots & \beta_k(u_2, v_2) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \dots & \beta_k(u_n, v_n) \end{bmatrix} \tag{2.13}$$

Considerando que o tamanho da amostra observada é n e o número de variáveis explicativas é k, tem-se que $\mathbf{X}$ é a matriz do modelo com dimensão $(n \times (k+1))$, cuja linha j contém a estimativa dos $(k+1)$ parâmetros para a amostra j.

Pode-se estimar os parâmetros da RGP matricialmente como:

$$\hat{\boldsymbol{\beta}}\left(u_i, v_i\right) = (\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{y} \tag{2.14}$$

na qual

$$\mathbf{W}(u_i, v_i) = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix} \tag{2.15}$$

$\mathbf{W}(u_i, v_i)$ é uma matriz diagonal e diferente para cada ponto i, contendo os pesos w_{ij} em sua diagonal principal, obtidos através das funções de ponderação ou *kernel*. Na RGP, uma observação j é ponderada de acordo com sua proximidade à localidade i, sendo que esse peso não se mantém constante para determinada observação, e sim depende da localidade i com que está sendo ponderada. Um ponto-chave dessa técnica diz respeito à noção do "círculo de inclusão" de observações ao redor do ponto i. A questão está relacionada a encontrar a melhor dimensão do raio a ser considerada. Outra questão importante está relacionada ao peso de cada localidade na estimação no ponto i.

O parâmetro de suavização (ou do inglês *bandwidth*) é um parâmetro que controla a variância da função de ponderação e determina a velocidade de decaimento do peso com a distância. O parâmetro de suavização pode ser fixo ou variar espacialmente de acordo com a disposição dos dados observados. Fotheringham et al. (2002) comentam que os resultados da RGP são relativamente insensíveis à escolha da função de ponderação, no entanto, são muito sensíveis à escolha do parâmetro de suavização. As duas principais funções de ponderação encontradas na literatura são a função normal ou Gaussiana e a função Bisquare. As fórmulas para ambas as funções são apresentadas a seguir:

Gaussiana:

$$w(i,j) = exp\left\{-\frac{1}{2}\left(d_{ij}/b\right)^2\right\} \tag{2.16}$$

Bisquare:

$$w(i,j) = \left[1 - \left(d_{ij}/b_{i(k)}\right)^2\right]^2 \tag{2.17}$$

A notação d_{ij} representa a distância do ponto i para a observação j, d é uma distância pré-determinada e b é o parâmetro de suavização.

A primeira função de ponderação (2.16), chamada de *Kernel* Gaussiana, decresce continuamente a medida que os pontos se distanciam e o seu parâmetro de suavização é fixo. Funções desse tipo são chamadas de Kernel espacial fixas, uma vez que assumem apenas um valor para o limite b (Figura 2.3). Em algumas situações, como quando os dados não estão igualmente espaçados, é recomendável que o parâmetro de suavização da função de ponderação varie espacialmente de acordo com a disposição dos dados observados. Assim, as áreas com menor densidade de pontos utilizam uma função *kernel* com maior parâmetro de suavização, enquanto que as áreas mais adensadas empregam um parâmetro de suavização menor (Figura 2.4). A função Bisquare (2.17) é um exemplo de função de ponderação que varia espacialmente.

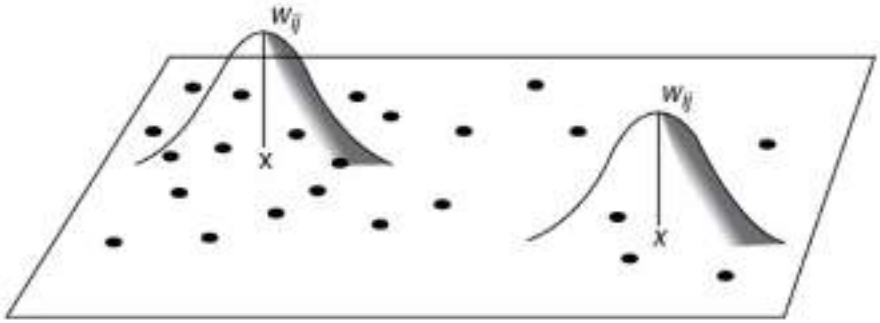


Figura 2.3: Função de ponderação espacial - Kernel fixo
Fonte: Fotheringham et al. (2002)

Para a definição do valor ótimo para o parâmetro de suavização, uma solução seria encontrar o valor b que minimize a soma $z = \sum_{i=1}^{n}(y_i - \hat{y}_i^*(b))^2$, em que $\hat{y}_i^*(b)$ seria o valor estimado pelo modelo (2.11) com o valor b indicando a suavização da função de ponderação. Porém, existe uma limitação, pois para essa forma o valor b ótimo é o valor tal que apenas o próprio elemento i permaneça na estimativa dos parâmetros no local i, ou seja, a soma se iguala a zero, gerando um modelo com um parâmetro para cada observação. Cleveland (1979) propôs uma solução que consiste no uso da técnica

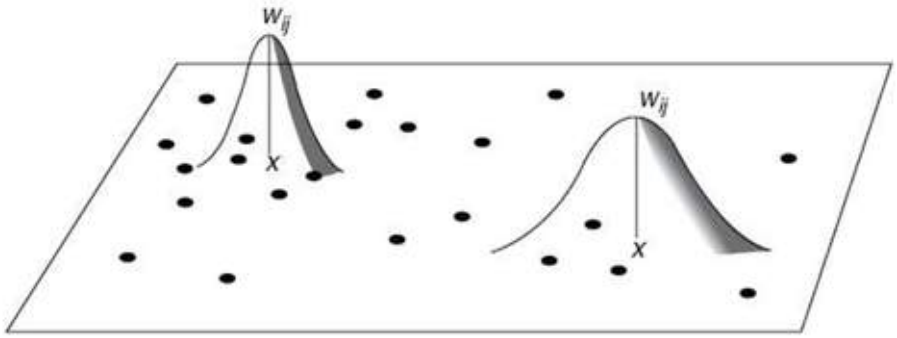


Figura 2.4: Função de ponderação espacial - Kernel variável
Fonte: Fotheringham et al. (2002)

de validação cruzada, ou do inglês *Cross-Validation (CV)*, em que o valor procurado é o que minimiza a soma

$$CV = \sum_{i=1}^{n} (y_i - \hat{y}^*_{\neq i}(b))^2 \tag{2.18}$$

em que $\hat{y}^*_{\neq i}(b)$ é o valor ajustado para o local i, desconsiderando o ponto i na estimação. Assim, somente os pontos próximos de i são selecionados. Note que $w_{i,j} = 1$ se $i = j$, porém, para fins computacionais, quando i e j são coincidentes, deve-se assumir que $w_{i,j} = 0$ na validação cruzada.

Charlton et al. (2009) sugerem ainda o uso do valor do critério de Akaike corrigido, no lugar de (2.18), que pode ser utilizado também como uma medida de ajuste para a comparação de modelos, cuja a forma, de acordo com Burnham e Anderson (2003), é dada por:

$$AIC_C = 2v_1 - 2l + 2\frac{v_1(v_1+1)}{n - v_1 - 1} \tag{2.19}$$

sendo l a log-verossimilhança do modelo dada por:

$$l = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \sum_i \frac{(y_i - \hat{y}_i)^2}{\hat{\sigma}^2} \tag{2.20}$$

em que n é o número de observações, $\hat{\sigma}$ é o valor estimado do desvio padrão dos resíduos via máxima verossimilhança e v_1 representa o número de parâmetros estimados pelo modelo.

Como em um modelo de regressão global, é importante calcular os erros padrões das estimativas locais, a fim de verificar a variabilidade e a validade estatística de tais estimativas. Considere que o estimador para as estimativas locais possa ser reescrito como:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \mathbf{CY} \tag{2.21}$$

em que $\mathbf{C} = (\mathbf{X}'\mathbf{W}(u_i, v_i)\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}(u_i, v_i)$. A variância das estimativas dos parâmetros é dada por:

$$\hat{Var}[\hat{\boldsymbol{\beta}}(u_i, v_i)] = \mathbf{CC}'\hat{\sigma}^2 \tag{2.22}$$

em que σ^2 é a soma dos quadrados dos resíduos normalizados da regressão local,

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{n - 2v_1 + v_2} \tag{2.23}$$

na qual $v_1 = tr(R)$, $v_2 = tr(R'R)$ e $tr()$ denota o traço da matriz.

As linhas $\mathbf{r}_j$ da matriz $\mathbf{R}$ são dadas por:

$$\mathbf{r}_j = \mathbf{X}_j[\mathbf{X}'\mathbf{W}(j)\mathbf{X}]^{-1}\mathbf{X}'\mathbf{W}(j) \tag{2.24}$$

em que $\mathbf{X}_j$ é a j-ésima linha da matriz do modelo $\mathbf{X}$. O traço da matriz $\mathbf{R}$ da RGP é igual ao traço da matriz de projeção (ou do inglês, *hat matrix*).

2.3 A Regressão Geograficamente Ponderada Área para Ponto - RGP-APP

Murakami e Tsutsumi (2015), propuseram um tipo de RGP com mecanismos de agregação, chamado Regressão Geograficamente Ponderada - Área para Ponto (RGP-APP) ou do inglês *area-to-point* (ATP-GWR). A RGP-APP, que está intimamente relacionada com as abordagens geoestatísticas, estima os parâmetros de tendência num nível de desagregação local utilizando variáveis agregadas.

Fotheringham et al. (2002) propuseram a aplicação da Regressão Geograficamente

Ponderada como uma abordagem para o MAUP. Como a RGP modela os padrões espaciais nos dados, que são uma fonte de MAUP, acredita-se que a RGP seja robusta ao MAUP. Entretanto, como a RGP não considera mecanismos de agregação, ela não pode ser considerada como uma medida que atenua o MAUP (Fotheringham et al., 2002; Wong, 2009). Como uma exceção, Young e Gotway (2007) mostram que a consideração de um mecanismo de agregação em uma análise baseada em RGP pode mudar os resultados da análise. No entanto, eles consideram o mecanismo de agregação para prever as variáveis explicativas no seu modelo RGP, e seu modelo RGP em si não o considera.

A aproximação mencionada para o MAUP tem algumas desvantagens, e medidas teoricamente suficientes para mitigar o MAUP ainda não existem (Siffel et al., 2006). Na geoestatística, o MAUP é considerado como um Problema de Mudança de Suporte (PMS) (ou do inglês *Change of Suport Problem* (COSP)) (Gotway e Young, 2002). A estrutura geral do PMS discute a mudança de suporte espacial (por exemplo, unidades, locais de dados). No entanto, enquanto a maioria dos estudos do PMS se concentra na mudança de suporte (por exemplo, interpolações), alguns discutem os vieses nas estimativas de parâmetros ao fazê-lo, incluindo o MAUP (Gotway e Young, 2002).

Atualmente, o PMS é um tópico de interesse em geoestatística, e a eficácia de abordagens geoestatísticas ao PMS foi demonstrada em vários estudos. No entanto, os estudos geográficos de MAUP e os estudos do PMS geostatísticos foram discutidos praticamente de forma independente (ver Haining et al. (2002)). A fim de combinar as discussões do MAUP em geografia e em geoestatística, o trabalho de Murakami e Tsutsumi (2015) estende a RGP, que tem sido discutida nos estudos geográficos de MAUP, como Fotheringham et al. (2002) e Wong (2009), e propõe um método chamado RGP-APP, com uma abordagem voltada para o problema de escala.

2.3.1 RGP para o MAUP

Existem dois tipos de variáveis: variáveis extensivas (ou volume) e variáveis intensivas (incluindo densidade, taxas e médias). O primeiro cresce em proporção ao tamanho da unidade espacial, enquanto o último independente do tamanho. Por

exemplo, população e área são variáveis extensivas, enquanto densidade populacional e taxa de zonas residenciais são variáveis intensivas. Assume-se que variáveis intensivas são variáveis explicativas para as quais a RGP tem sido aplicada geralmente, no entanto a discussão apresentada é facilmente estendida para variáveis extensivas. A seguir é apresentada uma abordagem baseada na RGP que mitiga o MAUP pela estimação dos parâmetros num nível desagregado, que são essencialmente livres de mecanismos de agregação, usando variáveis agregadas.

2.3.2 Modelo

Assume-se um modelo composto de dois submodelos. O primeiro é a RGP a nível desagregado, que é dado pela RGP padrão na Equação 2.11 como segue:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta_d} + \boldsymbol{\varepsilon} \\ E\left[\boldsymbol{\varepsilon}\right] &= 0 \\ Var\left[\bar{\boldsymbol{\varepsilon}}\right] &= \sigma^2\mathbf{I} \end{aligned} \tag{2.25}$$

em que d é um indexador da unidade desagregada e $\boldsymbol{\beta_d}$ é um vetor do parâmetro espacialmente variante na d-ésima unidade desagregada. O segundo modelo é a RGP a nível agregado, que é automaticamente obtido pela multiplicação da matriz de agregação $\mathbf{A}$ pelo lado esquerdo de (2.25):

$$\begin{aligned} \bar{\mathbf{y}} &= \bar{\mathbf{X}}\boldsymbol{\beta_d} + \bar{\boldsymbol{\varepsilon}} \\ E\left[\bar{\boldsymbol{\varepsilon}}\right] &= \bar{\mathbf{0}} \\ Var\left[\bar{\boldsymbol{\varepsilon}}\right] &= \sigma^2\mathbf{A}\mathbf{A}' \end{aligned} \tag{2.26}$$

em que $\bar{\mathbf{y}} = \mathbf{A}\mathbf{y}$ é um vetor de variável dependente de nível agregado, $\bar{\mathbf{X}} = \mathbf{A}\mathbf{X}$, $\bar{\boldsymbol{\varepsilon}} = \mathbf{A}\boldsymbol{\varepsilon}$ e $\bar{\mathbf{0}} = \mathbf{A}0$. Aqui, assume-se que $\mathbf{y}$ é desconhecido, enquanto $\bar{\mathbf{y}}$ e $\mathbf{X}$ (e $\bar{\mathbf{X}}$) são conhecidos.

A matriz $\mathbf{A}$ necessita ser determinada pela consideração da propriedade de preservação do volume (Lam, 1983) para a qual a agregação das variáveis desagregadas, $\mathbf{A}\mathbf{y}$, necessita ser igual aos valores atuais das variáveis agregadas, $\bar{\mathbf{y}}$. Em outras palavras, $\bar{\mathbf{y}} = \mathbf{A}\mathbf{y}$ necessita ser constante. Por exemplo, suponha que a densidade populacional em uma unidade agregada a é $\bar{y}_a$ e a unidade a compreende duas unidades

desagregadas d e d'; então, $\bar{\mathbf{y}} = \mathbf{A}\mathbf{y}$ é expresso como:

$$\bar{y_a} = \begin{bmatrix} A_{a,d} & A_{a,d'} \end{bmatrix} \begin{bmatrix} Y_d/S_d \\ Y_{d'}/S_{d'} \end{bmatrix}, \tag{2.27}$$

$$\frac{Y_d + Y_{d'}}{S_d + S_{d'}} = A_{a,d}\frac{Y_d}{S_d} + A_{a,d'}\frac{Y_{d'}}{S_{d'}}, \tag{2.28}$$

em que $A_{a,d}$ é o (a,d)-ésimo elemento de $\mathbf{A}$, Y_d e $Y_{d'}$ são as populações nas unidades desagregadas, e S_d e $S_{d'}$ são as áreas dessas unidades. A Equação 2.28 apenas estabelece que $\bar{y}_a$ necessita ser igual a média ponderada das densidades populacionais nas unidades desagregadas; consequentemente, são satisfeitas pela definição de A_d e $A_{d'}$ com $\frac{S_d}{S_d+S_{d'}}$ e $\frac{S_{d'}}{S_d+S_{d'}}$, respectivamente.

Em geral, $\mathbf{A}_{a,d}$ é definindo como se segue:

$$A_{a,d} \begin{cases} S_d / \sum_{d \subseteq a} S_d, & se \;\; d \subseteq a \\ 0, & c.c \end{cases} \tag{2.29}$$

em que $d \subseteq a$ denota as unidades desagregadas na a-ésima unidade agregada, $A_{a,d}$ assume um valor entre 0 e 1, e $\sum_{d \subseteq a} A_{a,d} = 1$.

Por definição, variáveis intensivas são sempre expressas como Y_d/S_d, como assumido em (2.27), e essa estrutura é aplicável a quaisquer variáveis intensivas desde que os dados em S_d estejam disponíveis. Por exemplo, quando $\bar{\mathbf{y}}$ consiste de variáveis divididas pela área (ex: densidade populacional), $A_{a,d}$ necessita ser dada por (2.29), em que S_d é a área da d-ésima unidade (ver (2.27)). Quando $\bar{\mathbf{y}}$ consiste de variáveis divididas pela população (por exemplo, rendimento per capita), $A_{a,d}$ necessita ser dado por (2.29) na qual S_d é a população da d-ésima unidade. No caso em que $\bar{\mathbf{y}}$ consiste dos preços médios da habitação, $A_{a,d}$ podem ser dados por (2.29) com seus S_d sendo definidos por variáveis que são proporcionais ao número de habitações, como o número de domicílios. Além disso, a estrutura é modificável para variáveis extensivas (ou volumes). Por substituição em (2.29) para $\mathbf{A}\mathbf{A}'$, esta matriz torna-se uma matriz diagonal com seu a-ésimo elemento $\sum_{d \subseteq a} A_{a,d}^2$, que indica um valor entre 0 e 1; então,

o a-ésimo elemento de $\bar{\boldsymbol{\varepsilon}} \sim N(\bar{\mathbf{0}}, \sigma^2 \mathbf{A}\mathbf{A}')$ em (2.26), $\bar{\varepsilon}_a$ é expresso como:

$$\begin{aligned} E\left[\bar{\varepsilon}_a\right] &= 0 \\ Var\left[\bar{\varepsilon}_a\right] &= \sigma^2 \textstyle\sum_{d \subseteq a} A_{a,d}^2 \end{aligned} \tag{2.30}$$

A Equação 2.30 indica que a RGP de nível agregado avalia as variações em cada unidade agregada por deflação da variação do nível desagregado, σ^2, com $\sum_{d \subseteq a} A_{a,d}^2$. A deflação é grande (isto é, $\sum_{d \subseteq a} A_{a,d}^2$ é pequeno) para as unidades agregadas incluindo muitas unidades desagregadas. Em adição, como ilustrado na Figura 2.5, as deflações são uniformes (não uniformes) quando as unidades agregadas são regularmente (irregularmente) formadas. Portanto, o modelo considera explicitamente a deflação da variância causada pela agregação, que é a principal fonte do MAUP.

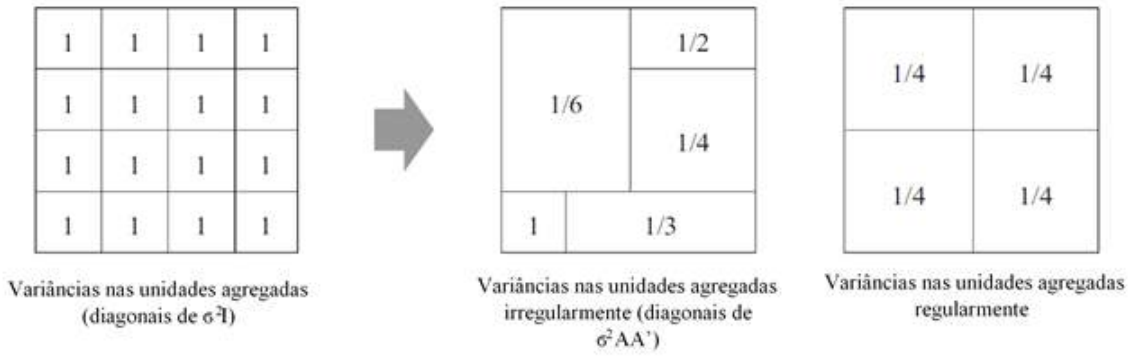


Figura 2.5: Deflação na variância devida a agregação
Fonte:Murakami e Tsutsumi (2015)

Além disso, o modelo RGP na Equação 2.25 pode mitigar a influência da dependência espacial, que é outra fonte principal do MAUP, especialmente o problema de escala. A dependência de escala é introduzida pelo delineamento do processo espacialmente dependente, consequentemente, a otimização do zoneamento (ou delimitação) é uma abordagem potencial para mitigar a dependência da escala (por exemplo, Openshaw (1984)). A ponderação local baseada em Kernel na RGP pode ser interpretada como zoneamento difuso ou janelas móveis (Páez et al., 2008). Especificamente, a RGP estima os β_s usando variáveis geograficamente ponderadas em uma zona difusa, cujo tamanho de zona é calibrado pelo parâmetro de suavização b. Como sugerido por Fotheringham et al. (2002), o zoneamento difuso calibrado retrata os padrões de dependência espacial relacionados ao MAUP e mitiga a dependência

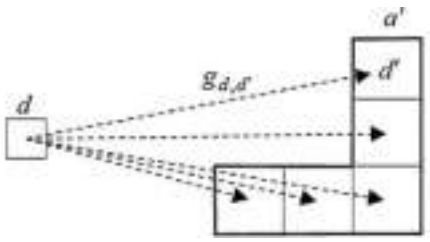
da escala.

2.3.2.1 Estimação dos parâmetros

O modelo de nível agregado (2.26) é idêntico ao modelo padrão RGP cujas variâncias são dimensionadas pelas diagonais de $\mathbf{AA'}$,$\sum_{d\subseteq a} A_{a,d}^2$. Assim, o estimador de $\boldsymbol{\beta_d}$ é dado como na RGP padrão:

$$\hat{\boldsymbol{\beta}}_d = \left(\bar{\mathbf{X}}'\bar{\mathbf{G}}_d^{1/2}\left(\mathbf{AA}'\right)^{-1}\bar{\mathbf{G}}_d^{1/2}\bar{\mathbf{X}}\right)^{-1}\bar{\mathbf{X}}'\bar{\mathbf{G}}_d^{1/2}\left(\mathbf{AA}'\right)^{-1}\bar{\mathbf{G}}_d^{1/2}\bar{\mathbf{y}}. \tag{2.31}$$

Suponha que $\mathbf{G_d}$ é uma matriz diagonal cujo d'-ésimo elemento é $g_{d,d'}$ que é definido pela função de distância entre as unidades desagregadas d e d', enquanto $\bar{\mathbf{G}}_d = \mathbf{AG}_d\mathbf{A}'$ é uma matriz diagonal cuja a'-ésima diagonal $\bar{g}_{d,a'} = S^2_{d'\subseteq a'} g_{d,d'\subseteq a'} / \sum_{d'\subseteq d} S^2_{d'\subseteq d}$ é a média ponderada da conectividade espacial entre a d-ésima unidade desagregada e as unidades desagregadas na a-ésima unidade agregada, $g_{d,d'\subseteq a'}$, e $\bar{\mathbf{G}}_d^{1/2}$ é uma matriz diagonal cuja diagonais são as raízes quadradas das diagonais de $\bar{\mathbf{G}}_d$. Conforme mostrado na Figura 2.6, esta especificação nos permite considerar as formas das unidades agregadas. Em suma, a Equação 2.31 é um estimador ponderado de mínimos quadrados que considera as deflações de variância por agregações com $\mathbf{AA'}$ e as formas das unidades agregadas com $\bar{\mathbf{G}}_d$.



$$\bar{g}_{d,a'} = \frac{\sum_{d'\subseteq a'} S^2_{d'} g_{d,d'}}{\sum_{d'\subseteq a'} S^2_{d'}}$$

Figura 2.6: Conectividade espacial entre d e a':$\bar{g}_{d,a'}$. $\bar{g}_{d,a'}$ considera a forma de unidades agregadas utilizando $g_{d,a'}$, que são descritas pelas setas, em que d' é um indexador de unidades desagregadas na a'-ésima unidade agregada.
Fonte:Murakami e Tsutsumi (2015)

O modelo de nível agregado, (2.26), é idêntico ao padrão RGP. Portanto, a matriz de variância-covariância de $\hat{\boldsymbol{\beta}}_d$ é dada como (Fotheringham et al., 2002)

$$\begin{aligned} Cov[\hat{\boldsymbol{\beta}}_d] &= \hat{\sigma}^2\mathbf{V}_d\mathbf{V}_d' \\ \mathbf{V_d} &= \left(\bar{\mathbf{X}}'\bar{\mathbf{G}}_d^{1/2}\left(\mathbf{AA}'\right)^{-1}\bar{\mathbf{G}}_d^{1/2}\bar{\mathbf{X}}\right)^{-1}\bar{\mathbf{X}}'\bar{\mathbf{G}}_d^{1/2}\left(\mathbf{AA}'\right)^{-1}\bar{\mathbf{G}}_d^{1/2}, \end{aligned} \tag{2.32}$$

em que $\hat{\sigma}^2$ denota a estimativa de σ^2. Ao substituir (2.31) em (2.26), os valores ajustados de $\bar{\mathbf{y}}$ são dados por $\hat{\bar{\mathbf{y}}} = \mathbf{ALy}$, em que $\mathbf{L}$ é uma matriz cuja d-ésima linha é $\mathbf{x_d}'\mathbf{V_d}$ e $\mathbf{x_d}$ é um vetor das variáveis explicativas observadas na d-ésima unidade desagregada. $\hat{\sigma}^2$ é dado como (Cressie et al., 1998):

$$\hat{\sigma}^2 = \frac{(\bar{\mathbf{y}} - \mathbf{A}\hat{\mathbf{y}})'(\bar{\mathbf{y}} - \mathbf{A}\hat{\mathbf{y}})}{tr\left\{(\mathbf{I} - \mathbf{AL})(\mathbf{I} - \mathbf{AL})'\right\}}, \tag{2.33}$$

em que $tr\{.\}$ é o operador traço e $\hat{\mathbf{y}}$ é um vetor cujo d-ésimo elemento é $\mathbf{x}_d'\boldsymbol{\beta}^d$. A significância de $\boldsymbol{\beta_d}$ pode ser testado usando os elementos diagonais de (2.32).

A estimativa de $\hat{\boldsymbol{\beta}}_{\boldsymbol{d}}$ é dada pela calibração do parâmetro de suavização b em $\bar{\boldsymbol{G}}_d$ e substituindo o parâmetro de suavização estimado em (2.31). A calibração acima pode ser realizada pela aplicação do método "$m - fold$" de validação cruzada nas cinco etapas seguintes: (a) os elementos em $\bar{\mathbf{y}}$ são divididos aleatoriamente em m subconjuntos; (b) sob um dado b, $1/m$ subamostras de $\bar{\mathbf{y}}$ são preditas usando as $(m-1)/m$ sub-amostras restantes; (c) o passo (b) é realizado para todos os m casos; (d) o erro quadrático de nível agregado ($\hat{\bar{\boldsymbol{\varepsilon}}}'\hat{\bar{\boldsymbol{\varepsilon}}}$ em (2.34)) é calculado; e (e) as etapas (a) a (d) são iteradas variando b, e o b ótimo, $\hat{b}$, satisfazendo (2.34) é definida como:

$$\hat{b} = arcmin_b \left[\hat{\bar{\boldsymbol{\varepsilon}}}'\hat{\bar{\boldsymbol{\varepsilon}}}\right] \quad \hat{\bar{\boldsymbol{\varepsilon}}} = \left(\mathbf{AA}'\right)^{-1/2}\left(\bar{\mathbf{y}} - \hat{\bar{\mathbf{y}}}\right) \tag{2.34}$$

$\hat{\bar{\mathbf{y}}} = \mathbf{A}\hat{\mathbf{y}}$, em que $\hat{\mathbf{y}}$ é um vetor cujo o d-ésimo elemento é $\mathbf{x}_d'\hat{\boldsymbol{\beta}}_{d(4/5)}$. Além disso, $\mathbf{x}_d$ é um vetor de variáveis explicativas na d-ésima unidade desagregada e $\hat{\boldsymbol{\beta}}_{d(4/5)}$ é dado com base em (2.31) como,

$$\hat{\boldsymbol{\beta}}_{d(4/5)} = \left(\bar{\mathbf{X}}'_{(4/5)}\bar{\mathbf{G}}^{-1/2}_{d(4/5)}\left(\mathbf{AA}'\right)^{-1}_{(4/5)}\bar{\mathbf{G}}^{-1/2}_{d(4/5)}\bar{\mathbf{X}}'_{(4/5)}\right)^{-1}\left(\bar{\mathbf{X}}'_{(4/5)}\bar{\mathbf{G}}^{-1/2}_{d(4/5)}\left(\mathbf{AA}'\right)^{-1}_{(4/5)}\bar{\mathbf{G}}^{-1/2}_{d(4/5)}\bar{\mathbf{y}}_{(4/5)}\right) \tag{2.35}$$

em que $\bar{\mathbf{X}}_{(4/5)}$ é $\bar{\mathbf{X}}$ cujo $1/5$ das linhas são descartadas, $\bar{\mathbf{y}}_{4/5}$ é $\bar{\mathbf{y}}$ cujo $1/5$ dos elementos são descartados, e $\bar{\mathbf{G}}_{d(4/5)}$ são $\bar{\mathbf{G}}_d$ e $\left(\mathbf{AA}'\right)_{(4/5)}$ cujo $1/5$ das linhas e colunas são descartadas. Os $1/5$ de elementos descartados necessitam incluir o d-ésimo elemento. Ao contrário da RGP padrão, que não considera os mecanismos de agregação, os elementos de $\hat{\boldsymbol{\beta}}_{\boldsymbol{d}}$ indicam os parâmetros de nível desagregado que variam espacialmente (note que os parâmetros $\boldsymbol{\beta_d}$ nas Equações 2.11 e 2.26 são idênticos). Essa abordagem

foi denominada RGP-APP, após a *ATP Kriging* (Kyriakidis, 2004), uma abordagem geoestatística intimamente relacionado ao PMS.

Murakami e Tsutsumi (2015) examinaram a eficácia da RGP-APP para atenuar o MAUP por meio de um estudo de simulação. Segundo os autores, existem pelo menos duas abordagens de simulação para a RGP. A primeira utiliza os autovetores de uma matriz de proximidade de duplo centro e a segunda abordagem pressupõe a presença de parâmetros que variam no espaço e obedecem a processos espaciais, $\beta_d \sim N(0, \tau^2\mathbf{C})$ (por exemplo, Finley (2011)) em que $\mathbf{C}$ é uma matriz de covariância cujos elementos são parametrizados por uma função de decaimento-distância. Murakami e Tsutsumi (2015) chegaram à conclusão de que do ponto de vista do MAUP, a RGP-APP é superior à RGP padrão, pois por ela estima-se explicitamente parâmetros de nível desagregado, que são essencialmente livres de mecanismos de agregação. No seu estudo de simulação, confirmaram que o método é eficaz em estimar parâmetros em face do MAUP.

Murakami e Tsutsumi (2015) afirmam que o método apresenta três limitações principais. Primeiro, o estudo de simulação indica a ineficácia do método quando os parâmetros espacialmente variantes têm padrões espaciais locais. Fisher e Langford (1996) descobriram que os padrões espaciais locais de nível desagregado podem ser efetivamente capturados considerando-se dados auxiliares espaciais em mecanismos de agregação. Portanto, o método precisa ser estendido para considerar dados auxiliares espacialmente finos em $\mathbf{A}$. A segunda limitação é a multicolinearidade. Tal como na RGP padrão, o RGP-APP parece sofrer com a questão da multicolinearidade, particularmente quando o número de variáveis explicativas é grande. A aplicação de uma forma penalizada de RGP, como uma Regressão *Ridge* geograficamente ponderada ou um modelo de regressão LASSO geograficamente ponderada (Wheeler, 2007), pode ser útil para resolver esse problema. O terceiro problema é o desconhecimento da dependência espacial. A integração da RGP-APP e Krigagem APP pode ser útil para superar esse problema. Em termos de futuras pesquisa, os autores sugerem a integração dos estudos geográficos do MAUP como os estudos do PMS. A integração destas áreas de estudo seria um passo importante para o desenvolvimento de medidas mais sofisticadas para o enfrentamento do MAUP.

Capítulo 3

Materiais e Métodos

Este Capítulo apresenta os materias e métodos utilizados no trabalho. A primeira seção descreve os materiais a serem utilizados nos estudos de simulação e estudo com dados reais. A segunda seção descreve os métodos e procedimentos aplicados para a avaliação da RGP-APP. Toda a análise será feita no *software* SAS 9.4.

3.1 Materiais

Para a realização dos estudos de simulação serão utilizados polígonos regulares com diferentes dimensões e configurações para a agregação de dados. Para a análise de dados reais, será utilizada a estrutura censitária planejada para o Distrito Federal - DF, a divisão administrativa proposta pela Companhia de Planejamento do Distrito Federal - CODEPLAN e a setorização interna das Regiões Administrativas.

3.1.1 Estudo com Dados Simulados

Para a análise dos efeitos do MAUP e eficácia do método RGP-APP, serão utilizados grades regulares formadas por quadrados que representem as menores unidades espaciais para a simulação. Serão consideradas grades com dimensões 40×40 ($n = 1.600$).

Os dados simulados para esse ensaio, serão gerados utilizando dois modelos: A Equação 3.1 representa o Modelo I que considera valores fixos para os parâmetros que definem como as covariáveis z_1 e z_2 se relacionam com a variável dependente z.

O Modelo I é representado por:

$$z_i = 3 + 0.8z_{1i} - 0.1z_{2i} + \varepsilon_i \qquad i = 1, \ldots, n \tag{3.1}$$

em que z_i é a variável dependente, z_{1i} e z_{2i} são variáveis contínuas com distribuições $N(5, 10)$ e $Poi(2)$, respectivamente. ε_i é o erro aleatório $N(0, 1)/f$, onde f é um fator que permite controlar a variância do modelo. Foram utilizados dois valores para esse fator: 3 e 10. A Figura 3.1 ilustra o efeito desse fator nos dados.

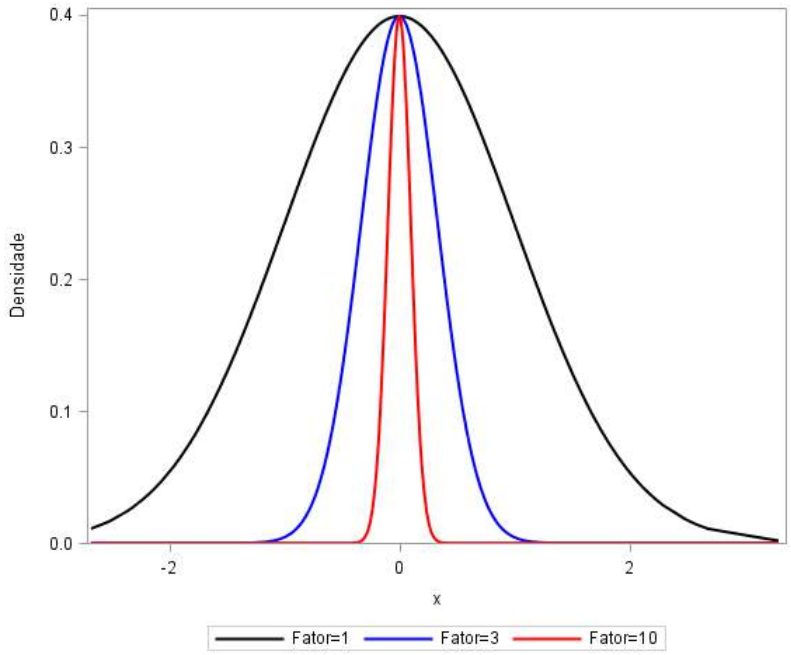


Figura 3.1: Efeito do fator de controle da variância

Os valores atribuídos aos parâmetros β_0, β_1 e β_2 são 3,0, 0,8 e -0,1, respectivamente. Esses valores são considerados os parâmetros verdadeiros para a simulação.

A Equação 3.2 representa o Modelo II, onde os parâmetros variam espacialmente e é dado por:

$$z_i = \beta_{0i} + \beta_{1i}z_{1i} + \beta_{2i}z_{2i} + \varepsilon_i \qquad i = 1, \ldots, n \tag{3.2}$$

em que z_{1i} e z_{2i} são variáveis contínuas com distribuições $U(0, 10)$ e $U(0, 1)$, respectivamente, β_{0i}, β_{1i} e β_{2i} se relacionam com as coordenadas x e y.

A Figura 3.2 representa a distribuição de z, β_0, β_1 e β_2 em função das coordenadas.

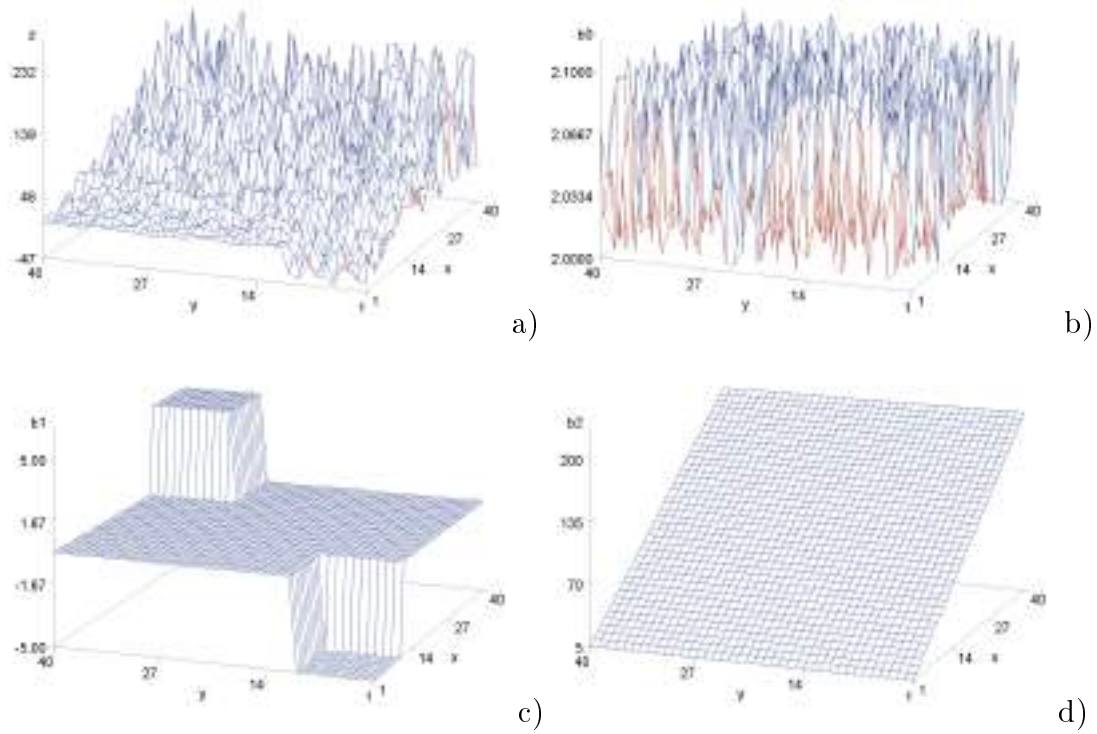


Figura 3.2: Exemplo de uma grade 40×40: a) Distribuição espacial de z; b)Distribuição espacial de β_0; c) Distribuição espacial de β_1; d) Distribuição espacial de β_2.

A Figura 3.3 ilustra a distribuição espacial da variável dependente z em uma grade regular de dimensão 40×40 com um fator para controle da variância fixado em 3.

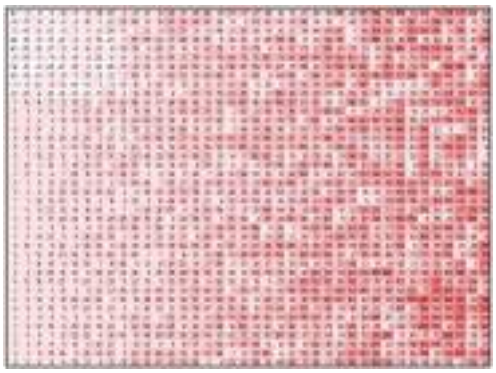


Figura 3.3: Grade 40 x 40 - Distribuição espacial de z

Para a avaliação dos efeitos do MAUP serão considerados quatro tipos de agregações das unidades básicas: horizontal, vertical, desigual e desigual 2. Os dois primeiros tipos, são representações de agregações em que o número de unidades agregadas para formar novas unidades é constante, formando sempre unidades de tamanhos iguais. A duas últimas são agregações que consideram unidades agregadas de tamanhos di-

ferentes. Cada um dos tipos de agregação será executada em dois níveis. As Figuras 3.4, 3.5, 3.6 e 3.7 mostram o formato das agregações propostas.

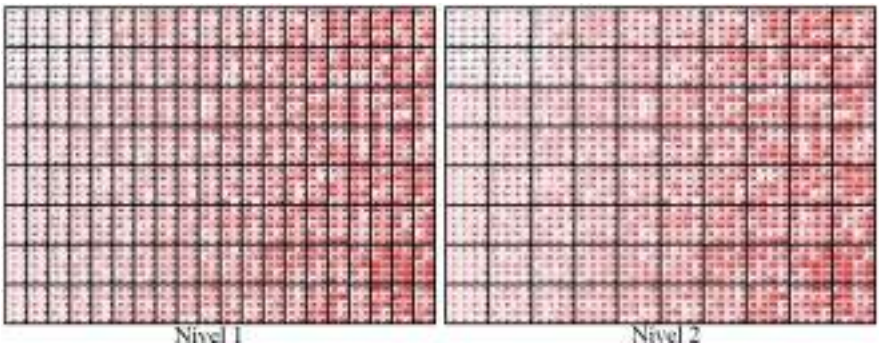


Figura 3.4: Grade 40 x 40 - Agregação vertical

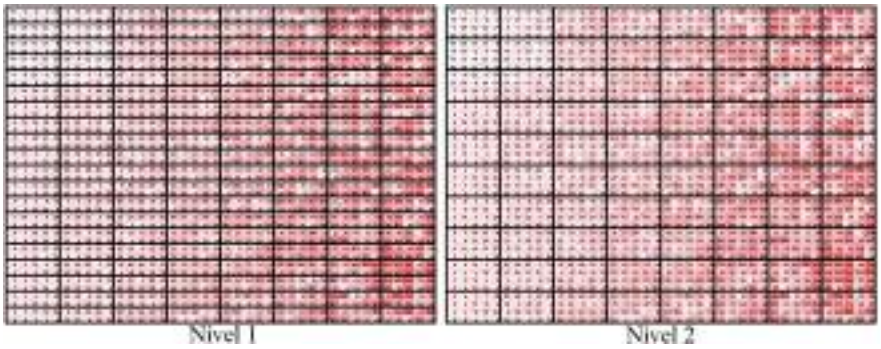


Figura 3.5: Grade 40 x 40 - Agregação horizontal

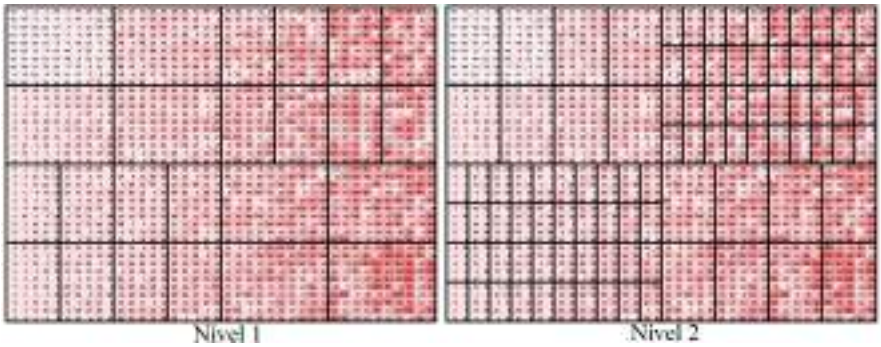


Figura 3.6: Grade 40 x 40 - Agregação desigual

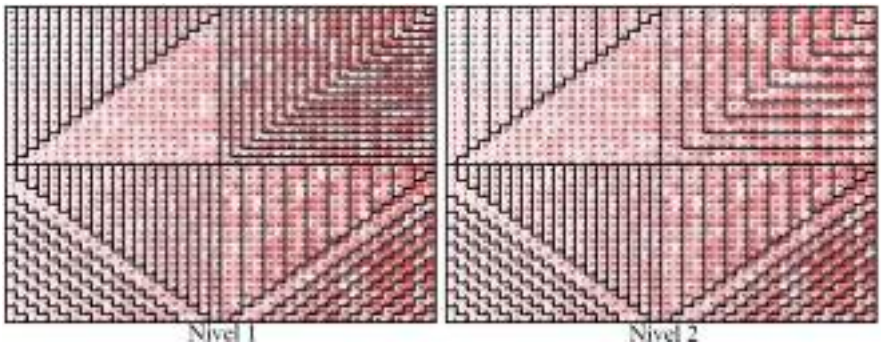


Figura 3.7: Grade 40 x 40 - Agregação desigual 2

3.1.2 Estudo com dados reais

Para a análise de dados reais, serão utilizados os microdados da PDAD de 2018, a delimitação elaborada pela CODEPLAN para a realização da Pesquisa Distrital por Amostra de Domicílios - PDAD e a setorização de Regiões Administrativas utilizada para fins de planejamento. Desta forma, serão utilizadas as seguintes estruturas espaciais:

- Domicílios visitados pela PDAD - CODEPLAN;
- Malha censitária - 2010 - IBGE;
- Delimitação das 31 Regiões Administrativas do DF - 2011 - CODEPLAN;
- Delimitação de setores internos das Regiões Administrativas.

Para simplificação da aplicação dos procedimentos descritos nesta Seção, serão considerados os dados dos responsáveis pelos domicílios residentes em duas Regiões Administrativas (RAs), Taguatinga e Ceilândia, que juntas apresentam uma amostra superior a 1.800 domicílios. A Figura 3.8 apresenta a delimitação das 31 Regiões Administrativas do DF, destacando a posição das duas RAs que serão estudadas.

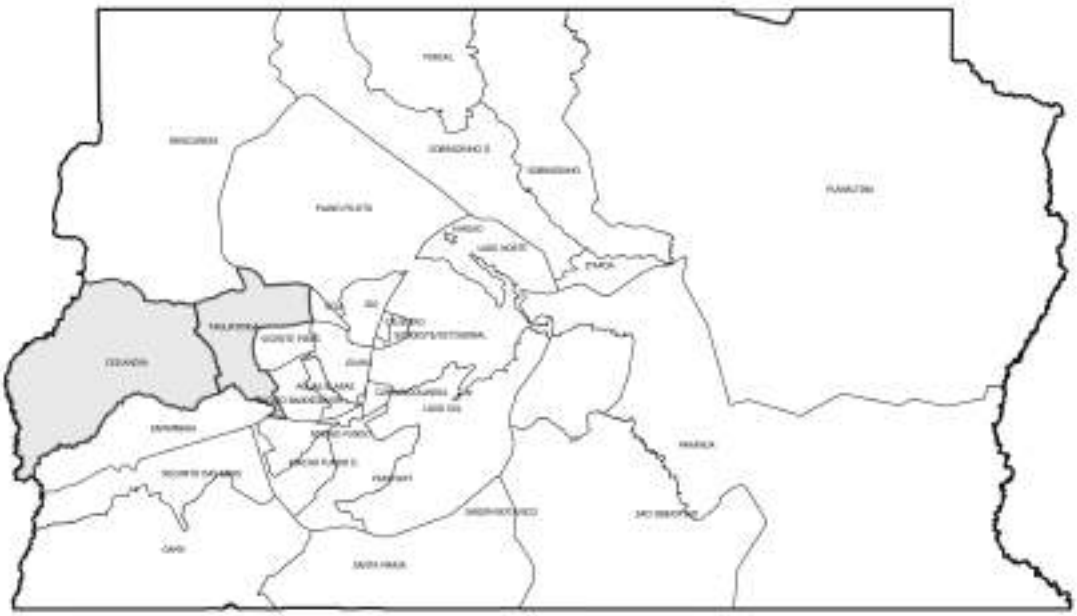


Figura 3.8: Distrito Federal e Regiões Administrativas

Para definição da forma de agregação dos dados foram utilizadas duas estruturas de setorização: Malha Censitária do DF e divisão setorial das RAs. Na Figura 3.9 estão representados os mais de 790 setores censitários e os 25 setores alcançados pela PDAD 2018.

a) b)

Figura 3.9: Agregações: a) Setores Censitários e b) Setores ("bairros")

A Pesquisa Distrital por Amostra de Domicílios - PDAD é realizada a cada dois anos pela Companhia de Planejamento do Distrito Federal, e tem como objetivo traçar o perfil sócio-econômico das famílias residentes em áreas urbanas do Distrito Federal.

Em sua última edição (PDAD-2018), foram visitados mais de 24.000 domicílios, onde foram registradas as coordenadas geográficas dos mesmos, e dos quais foram levantas informações sobre o perfil demográfico da população urbana, nível de instrução, trabalho e rendimento, infraestrutura domiciliar, características dos domicílios e posse de bens, dentre outras. A CODEPLAN concedeu, para a realização desse estudo, o acesso a dados individuais dos domicílios pesquisados, bem como as suas coordendas, o que traz a vantagem da comparação dos resultados obtidos por meio de agregações com os resultados obtidos a nível de indivíduos.

Para a avaliação do nível de desigualdade socioeconômica, será utilizado o modelo proposto por Mincer (1975), que tem sido base para uma extensa lista de estudos em economia. Esse modelo foi concebido para estimar retornos a educação, qualidade da educação e experiência, dentre outros fatores. Nesse modelo, os rendimentos dependem de fatores explicativos associados à escolaridade e à experiência. Pela identificação dos custos de educação e rendimentos do trabalho, é possível calcular a taxa interna de retorno da educação, que é a taxa de desconto que equaliza o custo e o ganho esperado dos investimentos em educação.

A equação Minceriana incorpora dois conceitos econômicos distintos em uma só equação:

a)Uma equação de preço revelando quanto o mercado de trabalho se dispõe a pagar por atributos produtivos como educação e experiência e;

b) a taxa de retorno da educação, que deve ser comparada com a taxa de juros de mercado para determinar a quantidade ótima de investimento em capital humano.

O modelo de regressão decorrente da equação de Mincer é dado por:

$$\ln(w) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 exp^2 + \varepsilon \tag{3.3}$$

em que w é o salário recebido pelo indivíduo, $educ$ é a sua escolaridade, geralmente medida por anos de estudo, exp é sua experiência, geralmente aproximada por $exp = idade - anos\ de\ estudo - 6$. β_1 é o retorno associado à escolaridade, β_2 e β_3 são os retornos associados à experiência e experiência ao quadrado e $\boldsymbol{\varepsilon}$ é um erro estocástico.

3.2 Métodos

A seguir serão descritos os procedimentos a serem adotados para a avaliação da eficiência do método RGP-APP.

3.2.1 Análise com Dados Simulados Utilizando a Extrapolação do Parâmetro de Suavização

Como mensionado no Capítulo 2, a RGP pode ser considerada uma extensão do modelo de regressão clássica (OLS) e os resultados obtidos pelos modelos se equiparam quando não há dependência espacial nos dados. Com isso, o parâmetro de suavização seria representado por um raio onde estariam incluídos todos os pontos do conjunto de dados. Numa primeira etapa, utilizando os dados simulados, afim de atestar a capacidade do RGP-APP em se estimar os parâmetros a nível desagregado, a RGP-APP será avaliada considerando um valor para o parâmetro de suavização que extrapole a maior distância observada entre os pontos do conjunto de dados e parâmetros fixos. Dessa forma espera-se que os resultados obtidos se aproximem dos valores estimados pelo modelo OLS obtidos com dados a nível desagregado.

Nessa etapa, serão considerados dois tipos de modelos: um sem covariáveis e o outro com covariáveis em sua estrutura, que serão aplicados aos dados simulados com parâmetros fixos, gerados pelo Modelo I representado pela Equação 3.1.

3.2.1.1 Modelo sem covariáveis

A concepção da RGP-APP parte de uma abordagem inspirada na krigagem APP (Kyriakidis, 2004), que é uma técnica da geoestatística utilizada para mitigar os efeitos do Problema de Mudança de Suporte (PMS). A Krigagem é utilizada para propósitos de interpolação e se apresenta como uma forma generalizada de modelos de regressão linear simples (modelo somente com intercepto), para estimação em um ponto sobre uma área ou dentro de um volume. É um método de média ponderada linearmente, onde seus pesos dependem não apenas da distância, mas também da direção e orientação dos dados vizinhos para o local sem amostragem.

Com o propósito de avaliar a resistência ao MAUP da RGP-APP como um modelo

de regressão linear simples, será conduzido um primeiro estudo onde se verificará a capacidade de estimação dos valores da média geral em diferentes níveis de agregação. Para tal, serão realizadas 100 repetições do ensaio, considerando grades de dimensão 40×40 nas configurações de agregação apresentadas nas Figuras 3.4, 3.5, 3.6 e 3.7 e fator de controle da variação fixado em 3. Para cada repetição serão calculadas as médias gerais dos dados desagregados, as médias ponderadas dos dados agregados nos 4 tipos de agregações (horizontal, vertical, desigual e desigual 2), as estimativas para o intercepto por meio do modelo de Regressão Linear Clássica, sem covariáveis, a nível desagregado e agregado, além das estimativas para o intercepto por meio do modelo RGP-APP.

Para que a comparação com os resultados globais da regressão OLS sejam válidas, serão considerados para o parâmetro de suavização valores que sejam maiores que a maior distância observada entre duas unidades desagregadas no conjunto de dados. Por meio das comparações, pretende-se identificar se a RGP-APP é eficiente em estimar a média geral da variável z a partir dos dados agregados.

3.2.1.2 Modelo com covariáveis

Nesta etapa, pretende-se avaliar um modelo RGP-APP que possua as duas covariáveis utilizadas para a geração dos dados em (3.1). Como no ensaio anterior, serão geradas 100 repetições, considerando grades de dimensão 40×40 e o nível 2 de agregação apresentado nas Figuras 3.4, 3.5, 3.6 e 3.7. O fator de variação utilizado será fixado em 3.

Para cada repetição, os valores estimados para β_0, β_1 e β_2 pela RGP-APP serão comparados aos valores estimados pela regressão OLS utilizando dados a nível desagregado e agregado. As comparações dos resultados obtidos pelos diferentes métodos permitirão avaliar se a RGP-APP tem o comportamento esperado quando utilizados dados com as características apresentadas, ou seja, se as estimativas produzidas pela RGP-APP se aproximam das estimativas produzidas pela regressão OLS a nível desagregado. Para fim de controle, serão calculadas as estimativas considerando um modelo RGP que servirá como uma referência do comportamento esperado pela RGP-APP.

3.2.2 Análise utilizando o parâmetro de suavização ótimo

Na segunda etapa, os modelos serão utilizados considerando o valor ótimo para o parâmetro de suavização, que será determinado pelo emprego do algoritmo *Golden Section Search*. Os procedimentos descritos nesta seção serão aplicados ao conjunto de dados reais (PDAD 2018) e aos dados gerados pelo Modelo II representado pela Equação 3.2 com parâmetros variando espacialmente em grades de dimensões 40×40. Serão utilizados os níveis 3 e 10 para o fator de controle da variação.

Para os dados simulados, o método RGP-APP será empregado a fim de se estimar os parâmetros β_0, β_1 e β_2 nas unidades básicas utilizando dados agregados para a variável dependente. As estimativas geradas pela RGP-APP em um modelo com duas covariáveis serão comparadas aos valores verdadeiros dos parâmetros e às estimativas da RGP clássica obtidas a nível desagregado e agregado.

Para o estudo de caso, após a verificação de dependência espacial utilizando o índice I de Moran, os coeficientes das variáveis do modelo de Mincer nas unidades básicas (domicílios), serão estimados a partir dos dados de rendimento agregados a nível de setores censitários e setores de Regiões Administrativas. Nessa etapa, os valores estimados pelo modelo RGP aplicado a dados desagregados serão considerados os valores verdadeiros dos parâmetros, servindo como referência para os demais modelos.

Pretende-se aqui avaliar se a RGP-APP apresenta maior resistência ao MAUP do que a RGP agregada e do que o modelo OLS, buscando definir alguma relação entre os resultados, a fim de minimizar a influência do MAUP.

No próximo Capítulo, serão apresentados os resultados dos estudos de simulação e do estudo de caso.

Capítulo 4

Análise dos Resultados

Neste Capítulo, serão apresentados os resultados obtidos nos estudos com dados simulados e logo em seguida os resultados do estudo com dados reais.

4.1 Estudos com Dados Simulados

Conforme descrito no Capítulo 3, nesta Seção serão apresentados os resultados da primeira etapa de análise realizada com dados simulados. Essa parte tem por objetivo verificar a capacidade do método RGP-APP em atenuar os efeitos do MAUP em dados simulados considerando as configurações descritas no Capítulo 3.

4.1.1 Análise Resultados Gerados com Dados Simulados Utilizando a Extrapolação do Parâmetro de Suavização

Os resultados apresentados a seguir foram gerados utilizando um valor para o parâmetro de suavização que extrapole a maior distância observada entre os pontos do conjunto de dados. Dessa forma, o raio de inclusão, considerará todos os pontos do conjunto de dados, fazendo com que os resultados dos modelos RGP e RGP-APP sejam os mesmos do modelo OLS. Isso será feito primeiramente para que a comparação seja feita apenas considerando a média dos parâmetros estimados, e porque nessa primeira simulação os parâmetros são fixos, conforme visto no Capítulo 3.

4.1.1.1 Modelo sem Covariáveis

No primeiro ensaio realizado, foi utilizado um modelo sem a presença de covariáveis (nesse caso, basta considerar que a covariável z_1 possui todos os valores iguais a 1 e o modelo seja estimado sem o intercepto), aplicando-o a dados gerados em grades de dimensão 40 x 40, no nível de agregação 2 e com fator de controle da variância fixado em 3. Para cada uma das 100 repetições, foram calculadas as médias gerais para os dados desagregados e para os dados agregados ponderados. Além das médias, foram obtidas as estimativas para a média utilizando o modelo OLS e o modelo RGP-APP. A Figura 4.1 apresenta o *Box Plot* da distribuição dos valores estimados para média geral por tipo de modelo para a agregação horizontal nas 100 repetições do ensaio. Cabe ressaltar que, para a estimação realizada pela RGP-APP, não foram utilizados os dados desagregados para a variável dependente z.

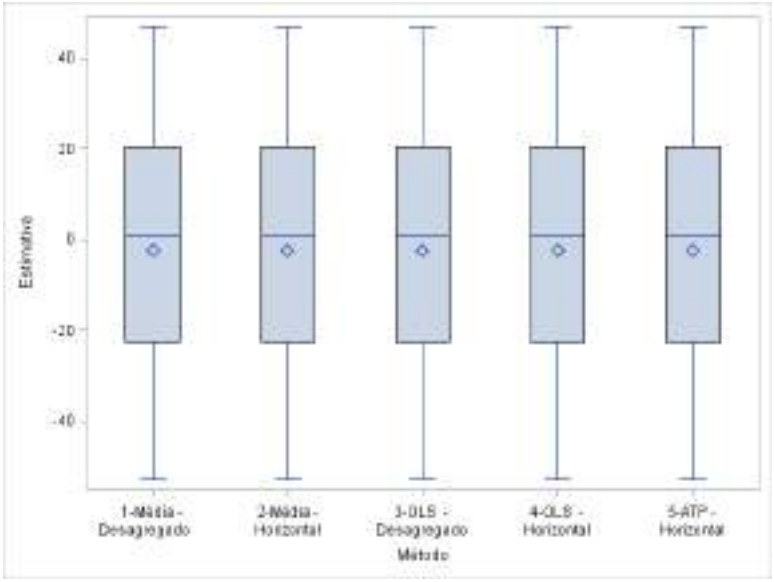


Figura 4.1: *Box Plot* da distribuição dos valores estimados para média geral

Os resultados mostram que as distribuições das estimativas produzidas nas 100 repetições pelos diversos modelos se apresentam iguais à distribuição da média geral calculada a nível desagregado. A Tabela 4.1 apresenta o percentual de casos em que a média geral agregada ou as estimativas para a média obtidas pelos modelos foram iguais a média obtida a nível desagregado.

Tabela 4.1: Percentual de casos em que a estimativa do intercepto foi igual a média geral a nível desagregado

Agregação	**Média Ponderada**	**OLS Agregada Ponderada**	**RGP-APP**
Horizontal	100%	100%	100%
Vertical	100%	100%	100%
Desigual	100%	100%	100%
Desigual 2	100%	100%	100%

Os dados mostram que em todas as repetições as estimativas obtidas pela RGP-APP é igual à média geral dos dados desagregados, atestando que a RGP-APP é resistente ao MAUP quando se considera um modelo de regressão simples. Na próxima Seção será avaliada a resistência da RGP-APP quando considerado um modelo com covariáveis.

4.1.2 Modelo com covariáveis

Para esta etapa do estudo com dados simulados, foram realizadas 100 repetições do ensaio, considerando grades de dimensão 40×40 no nível de agregação 2 e fator de controle da variância fixado em 3. As estimativas geradas pelos modelos RGP clássico e RGP-APP foram obtidas considerando o parâmetro de suavização fixado em um valor que extrapole a maior distância entre dois pontos observada no conjunto de dados. Nesta situação, o comportamento dos resultados obtidos pelo modelo OLS e RGP servem como referência do que se espera para o modelo RGP-APP.

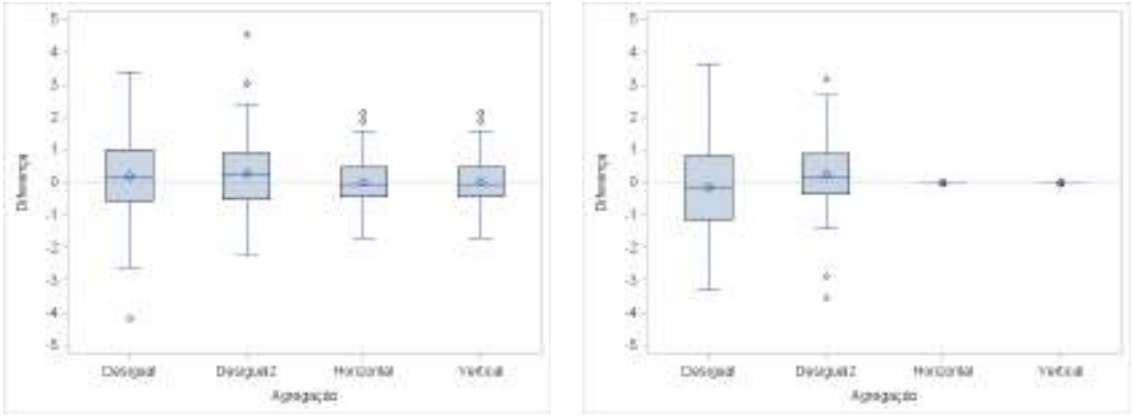
A Tabela 4.2 apresenta as estimativas globais para os parâmetros utilizando os dados a nível desagregados para os modelos OLS e RGP e agregados para a RGP-APP. Nota-se que, em média, não há diferenças entre as estimativas globais produzidas para β_0, β_1 e β_2 pelos modelos RGP e OLS, quando considerado o nível desagregado, o que confirma a capacidade da RGP de estimar os parâmetros nesse nível. Já para a RGP-APP, as estimativas dos parâmetros são iguais, em média, às estimativas do modelo OLS aplicados a dados agregados quando utilizadas as agregações horizontais

ou verticais. Quando utilizada as agregações desiguais, não é possível identificar um padrão de comportamento, conforme os dados da Tabela 4.2. Para melhor avaliação deste resultado, foram calculadas as diferenças entre as estimativas obtidas pela regressão RGP-APP e as estimativas OLS a nível agregado e desagregado em cada uma das repetições.

Tabela 4.2: Estimativas globais

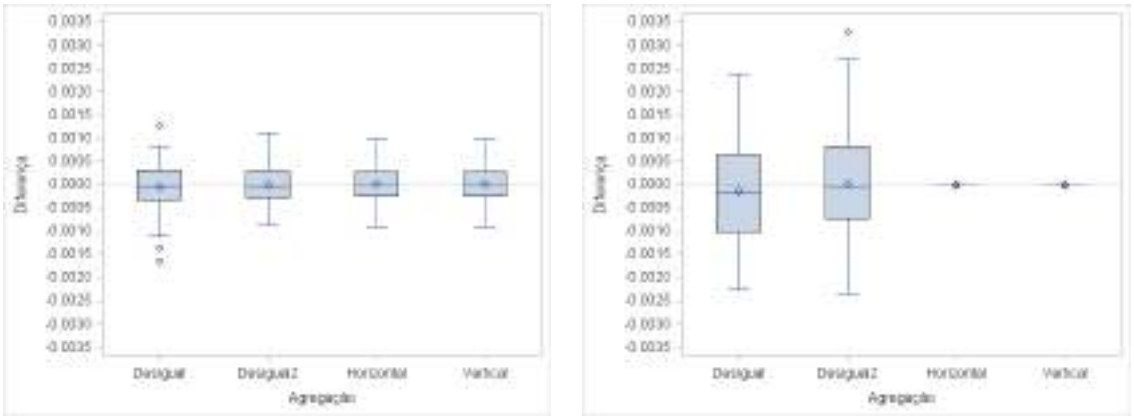
Par.	Modelo	Agregação	#Zonas	Média	Desvio	Min	Max
β_0	OLS	**Desagregado**	1600	3,02319	0,033806	2,53943	3,42433
		Horizontal	100	3,02582	0,631268	0,93267	5,06004
		Vertical	100	3,02582	0,631268	0,93267	5,06004
		Desigual	40	3,34849	4,307909	-4,24170	8,77439
		Desigual 2	56	3,04318	2,924211	-1,29057	7,77061
	RGP	**Desagregado**	1600	3,02319	0,033806	2,53943	3,42433
	RGP-APP	**Horizontal**	100	3,02583	0,631276	0,93262	5,05990
		Vertical	100	3,02583	0,631276	0,93262	5,05990
		Desigual	40	3,21308	1,909951	-1,25722	6,09854
		Desigual 2	56	3,29647	1,623549	0,77294	7,94862
β_1	OLS	**Desagregado**	1600	0,79998	0,000001	0,79708	0,80283
		Horizontal	100	0,79999	0,000001	0,79710	0,80292
		Vertical	100	0,79999	0,000001	0,79710	0,80292
		Desigual	40	0,80008	0,000002	0,79601	0,80320
		Desigual 2	56	0,79996	0,000002	0,79542	0,80363
	RGP	**Desagregado**	1600	0,79998	0,000001	0,79708	0,80283
	RGP-APP	**Horizontal**	100	0,79999	0,000001	0,79710	0,80292
		Vertical	100	0,79999	0,000001	0,79710	0,80292
		Desigual	40	0,79994	0,000001	0,79728	0,80313
		Desigual 2	56	0,79998	0,000001	0,79734	0,80290
β_2	OLS	**Desagregado**	1600	-0,10007	0,000000	-0,10184	-0,09810
		Horizontal	100	-0,10031	0,000004	-0,10907	-0,09561
		Vertical	100	-0,10031	0,000004	-0,10907	-0,09561
		Desigual	40	-0,10098	0,000024	-0,12373	-0,08864
		Desigual 2	56	-0,10004	0,000012	-0,11837	-0,09108
	RGP	**Desagregado**	1600	-0,10007	0,000000	-0,10184	-0,09810
	RGP-APP	**Horizontal**	100	-0,10031	0,000004	-0,10907	-0,09561
		Vertical	100	-0,10031	0,000004	-0,10907	-0,09561
		Desigual	40	-0,10061	0,000012	-0,11175	-0,09277
		Desigual 2	56	-0,10063	0,000012	-0,12374	-0,09171

As Figuras 4.2 , 4.3 e 4.4 apresentam as distribuições das diferenças entre as estimativas obtidas pela regressão OLS a nível desagregado.



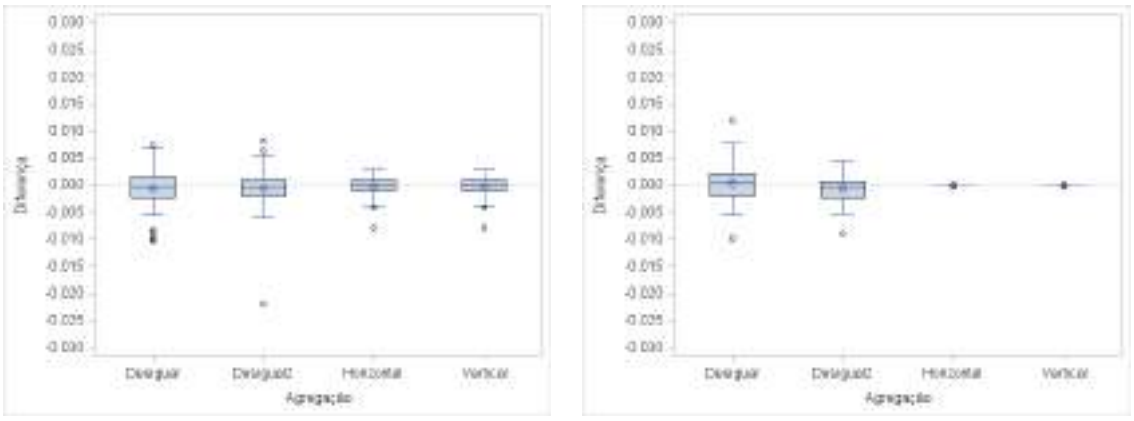
a) b)

Figura 4.2: *Box Plot* das diferenças entre estimativas para β_0: a) Referência OLS desagregado e b) Referência OLS agregado



a) b)

Figura 4.3: *Box Plot* das diferenças entre estimativas para β_1: a) Referência OLS desagregado e b) Referência OLS agregado



a) b)

Figura 4.4: *Box Plot* das diferenças entre estimativas para β_2: a) Referência OLS desagregado e b) Referência OLS agregado

As Figuras 4.2 b) , 4.3 b) e 4.4 b), mostram que quando consideradas as agregações horizontal e vertical, todas as diferenças foram iguais a zero. Isso confirma que, no cenário estabelecido para esse ensaio, a RGP-APP produz estimativas para os três parâmetros iguais às estimativas produzidas pelo modelo OLS, quando utilizados dados agregados de forma horizontal ou vertical. Para as agregações desiguais, as diferenças parecem estar mais próximas de zero quando a referência é a OLS desagregada, indicando que, nesta situação, a RGP-APP parece produzir estimativas mais próximas das estimativas OLS a nível desagregado.

A Tabela 4.3 apresenta o percentual de casos em que as estimativas produzidas pela RGP-APP se aproxima mais da OLS com dados desagregados, por parâmetro e tipo de agregação. Os resultados evidenciam a tendência das estimativas obtidas pela RGP-APP em se aproximarem das estimativas produzidas pela regressão OLS quando utilizados dados agregados de forma horizontal ou vertical. Quando utilizados dados com a agregação desigual, a RGP-APP produziu estimativas mais próximas dos parâmetros de nível desagregado, principalmente para o parâmetro β_1. Quando utilizada a agregação desigual 2, em 77% dos casos foram produzidas estimativas que se aproximam mais dos parâmetros a nível desagregado para β_1. Para os outros parâmetros o percentual de casos é menor que 50%. Destaca-se que apesar da utilização da agregação com unidades de tamanhos iguais, em situações práticas o que ocorre é a agregação com unidades de tamanhos desiguais.

Tabela 4.3: Percentual de casos em que as estimativas da RGP-APP se aproxima mais da OLS desagregada

Parâmetro	**Agregação**	**Frequência**
β_0	Horizontal	0%
	Vertical	0%
	Desigual	55%
	Desigual 2	46%
β_1	Horizontal	0%
	Vertical	0%
	Desigual	75%
	Desigual 2	77%
β_2	Horizontal	0%
	Vertical	0%
	Desigual	59%
	Desigual 2	45%

Dessa forma, torna-se evidente que quando utilizado o modelo com covariáveis, a RGP-APP não apresenta o mesmo desempenho de quando utilizado o modelo sem covariáveis. Na próxima subseção será demostrado matematicamente o efeito da inclusão de covariáveis no modelo de regressão quando comparados os valores entre diferentes níveis de agregação.

4.1.3 Análise do efeito da inclusão de covariáveis

Como visto nas Seções 4.1.1 e 4.1.2 o modelo RGP-APP estima com precisão o valor de β_0 para o modelo sem covariáveis em 100% dos casos analisados para qualquer tipo de agregação utilizada, produzindo valores iguais aos parâmetros obtidos a nível desagregado. Já quando são incluídas as covariáveis z_1 e z_2, o modelo produz estimativas para β_1 e β_2 que tendem aos valores obtidos pela regressão OLS com dados agregados em unidades que tenham o mesmo tamanho. Quando as unidades são desiguais, não há uma tendência evidente. Para melhor compreensão desse resultado, será apresentado nesta Seção a forma de estimação dos parâmetros associados

às covariáveis.

De forma geral, para um modelo de regressão linear com uma covariável (x), tendo como variável dependente (y), pelo método de Mínimos Quadrados Ordinários (MQO), as estimativas para β_0 e β_1 são obtidos pela minimização da soma dos quadrados dos desvios L.

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]^2 \tag{4.1}$$

Para minimizar 4.1, deve-se derivar L em relação aos parâmetros β_0 e β_1 e depois igualar a zero. Assim,

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \quad e \tag{4.2}$$

$$\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i \tag{4.3}$$

Daí,

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad e \tag{4.4}$$

$$-2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0. \tag{4.5}$$

Assim, as Equações de MQO serão dadas por:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i \end{cases} . \tag{4.6}$$

Resolvendo o sistema temos que,

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4.7}$$

em que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ são as médias de x e de y, respectivamente. Substituindo (4.7) em (4.6) tem-se que,

$$\begin{aligned} \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + n\hat{\beta}_1 \bar{x}^2. \end{aligned} \tag{4.8}$$

Então,

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \tag{4.9}$$

Assim,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{4.10}$$

Seja β_{1d} o parâmetro associado à covariável x quando se consideram dados desagregados e $\hat{\beta}_{1d}$ a sua estimativa, e β_{1d} o parâmetros associado à covariável x quando se consideram dados agregados e $\hat{\beta}_{1a}$ a sua estimativa. Poderíamos expressá-los como

$$\hat{\beta}_{1d} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)} \tag{4.11}$$

e

$$\hat{\beta}_{1a} = \frac{\sum_{i=1}^{k} f_i \bar{x}_i \bar{y}_i - \left(\sum_{i=1}^{k} f_i\right) \left(\frac{\sum_{i=1}^{k} f_i \bar{x}_i}{\sum_{i=1}^{k} f_i}\right) \left(\frac{\sum_{i=1}^{k} f_i \bar{y}_i}{\sum_{i=1}^{k} f_i}\right)}{\sum_{i=1}^{k} f_i \bar{x}_i^2 - \left(\sum_{i=1}^{k} f_i\right) \left(\frac{\sum_{i=1}^{k} f_i \bar{x}_i}{\sum_{i=1}^{k} f_i}\right)^2} = \frac{Cov(\bar{x}, \bar{y})}{Var(\bar{x})} \tag{4.12}$$

Dos resultados acima, temos que em um modelo de regressão linear simples, $\hat{\beta}_0$ (4.7) dependerá somente da média de y, que pode ser calculada de forma precisa com dados agregados quando existe a informação para a ponderação das novas áreas. Desta forma, os resultados para $\hat{\beta}_0$ independem do nível de agregação. A RGP-APP, por considerar a ponderação dos dados pela área das novas unidades de agregação, é capaz de estimar com precisão a média geral dos dados desagregados a partir dos dados agregados em um modelo que considere apenas o intercepto.

Como foi notado por Murakami e Tsutsumi (2015) e ilustrado na Figura 2.5,

as variâncias tendem a ser deflacionadas a medida em que se aumenta o nível de agregação das variáveis. Como (4.11) e (4.12) dependem também da covariância entre x e y e da variância de x, as estimativas de β_{1d} e β_{1a} sofrerão também os efeitos da agregação. Desta forma, pode-se concluir que a igualdade de (2.25) em relação a (2.26), só é verdadeira para o modelo sem covariáveis.

Se considerarmos um modelo com duas covariáveis x_1 e x_2, os produtos entre variância e covariância poderia causar um efeito mais intenso, pois seguindo o mesmo raciocínio teríamos:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2, \tag{4.13}$$

$$\hat{\beta}_{1d} = \frac{Cov(x_1,y)Var(x_2)-Cov(x_2,y)Cov(x_1,x_2)}{Var(x_1)Var(x_2)-[Cov(x_1,x_2)]^2} \tag{4.14}$$

e

$$\hat{\beta}_{2d} = \frac{Cov(x_2,y)Var(x_1)-Cov(x_1,y)Cov(x_1,x_2)}{Var(x_1)Var(x_2)-[Cov(x_1,x_2)]^2} \tag{4.15}$$

O exemplo a seguir ilustra com um caso prático a situação para um modelo com uma covariável. Considere o seguinte conjunto de dados desagregados apresentados na Tabela 4.4:

Tabela 4.4: Exemplo: dados desagregados para o cálculo dos parâmetros

Id	**y**	**x**	**xy**	x^2
1	3	5	15	25
1	2	4	8	16
1	4	1	4	1
2	2	0	0	0
2	2	2	4	4
3	1	1	1	1
3	0	3	0	9
Média	**2**	**16/7**	32/7	8
Soma	14	16	**32**	**56**

Temos que β_{1d} seria dado por:

$$\begin{aligned} \beta_{1d} &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{32 - 7 \times \frac{16}{7} \times 2}{56 - 7 \times (\frac{16}{7})^2} \\ &= 0 \end{aligned} \tag{4.16}$$

e β_{0d} seria dado por:

$$\beta_{0d} = 2 - 0 \times \frac{16}{7} = 2 \tag{4.17}$$

Para os dados agregados teríamos o seguinte conjunto de dados:

Tabela 4.5: Exemplo: dados agregados para o cálculo dos parâmetros

id	$\bar{y}$	$\bar{x}$	f_i	$f_i\bar{y}$	$f_i\bar{x}$	$f_i\bar{y}\bar{x}$	$f_i\bar{x}^2$
1	3	$\frac{10}{3}$	3	9	10	30	$\frac{100}{3}$
2	2	1	2	4	2	4	2
3	$\frac{1}{2}$	2	2	1	4	2	8
Média	$\frac{11}{6}$	$\frac{19}{9}$	1	2	$\frac{16}{7}$	$\frac{36}{7}$	$\frac{130}{9}$
Soma	$\frac{11}{2}$	$\frac{19}{3}$	7	14	16	36	$\frac{130}{3}$

e β_{1a} pode ser dado por:

$$\beta_{1a} = \beta_{1a} = \frac{\sum_{i=1}^{k} f_i \bar{x}_i \bar{y}_i - \left(\sum_{i=1}^{k} f_i\right)\left(\frac{\sum_{i=1}^{k} f_i \bar{x}_i}{\sum_{i=1}^{k} f_i}\right)\left(\frac{\sum_{i=1}^{k} f_i \bar{y}_i}{\sum_{i=1}^{k} f_i}\right)}{\sum_{i=1}^{k} f_i \bar{x}_i^2 - \left(\sum_{i=1}^{k} f_i\right)\left(\frac{\sum_{i=1}^{k} f_i \bar{x}_i}{\sum_{i=1}^{k} f_i}\right)^2} = \frac{36 - 7 \times \frac{16}{7} \times \frac{14}{7}}{\frac{130}{3} - 7 \times (\frac{16}{7})^2} = 0,5915 \tag{4.18}$$

e β_{0a} seria dado por:

$$\beta_{0d} = 2 - 0,5915 \times \frac{16}{7} = 0,648 \tag{4.19}$$

No caso de um modelo sem covariáveis teríamos que β_{0a} seria dado por:

$$\beta_{0d} = 2 - 0 \times \frac{16}{7} = 2 \tag{4.20}$$

sendo exatamente igual ao resultado obtido em (4.17).

Note que para o cálculo de β_{1d} as médias representadas pelas parcelas $\bar{x}$ e $\bar{y}$ são equivalentes às médias ponderadas de $\bar{x}f_i$ e $\bar{y}f_i$ no cálculo de β_{1a}. Todos os demais termos contribuem para que as estimativas obtidas em diferentes níveis sejam diferentes. Como o intercepto é dado em função dos demais parâmetros, é esperado que ele sofra os maiores impactos causados pela agregação dos dados. Espera-se também que quanto maior seja o número de parâmetros, maior seja o impacto causado. Na próxima seção os resultados da RGP-APP aplicada a dados simulados com a utilização do parâmetro de suavização ótimo serão comparados aos resultados da RGP clássica. Por não se tratar de um modelo que considere a estrutura espacial dos dados, espera-se que o modelo de regressão OLS sofra com mais intensidade os efeitos do MAUP.

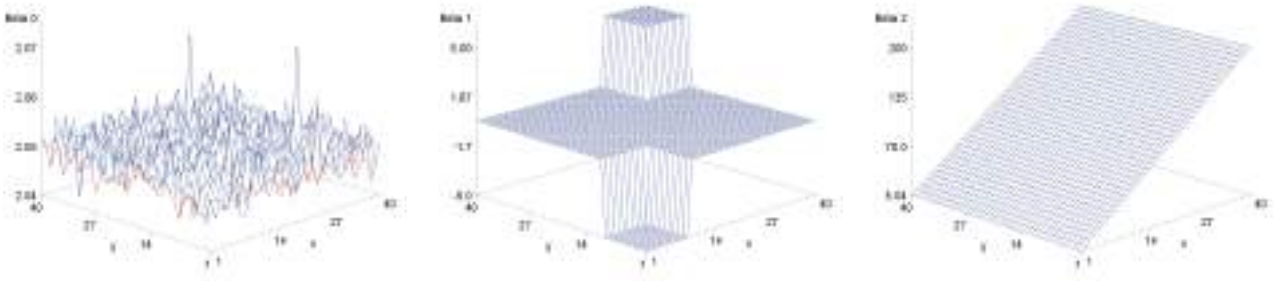
4.1.4 Simulação com parâmetro de suavização ótimo

Na Seções anteriores, considerou-se um cenário em que os parâmetros apresentavam distribuição espacial uniforme. Foi por conta desta característica que o modelo de regressão OLS foi tomado como referência, quando utilizados valores grandes para o parâmetro de suavização. Nesta Seção, consideram-se dados onde distribuição espacial dos parâmetros se relaciona com as coordenadas x e y, ou seja, a forma como a variável dependente se relaciona com as covariáveis se altera conforme a localização das observações. Pretende-se avaliar neste cenário a capacidade do modelo RGP-APP em mitigar os efeitos do MAUP em relação aos modelos OLS e RGP. Por não se tratar de um modelo de regressão espacial, espera-se que o modelo OLS seja o que sofra os maiores impactos do MAUP neste ensaio.

4.1.4.1 Comparação das estimativas produzidas pelos modelos OLS, RGP (aplicados a dados desagregados) e RGP-APP com os valores reais dos parâmetros

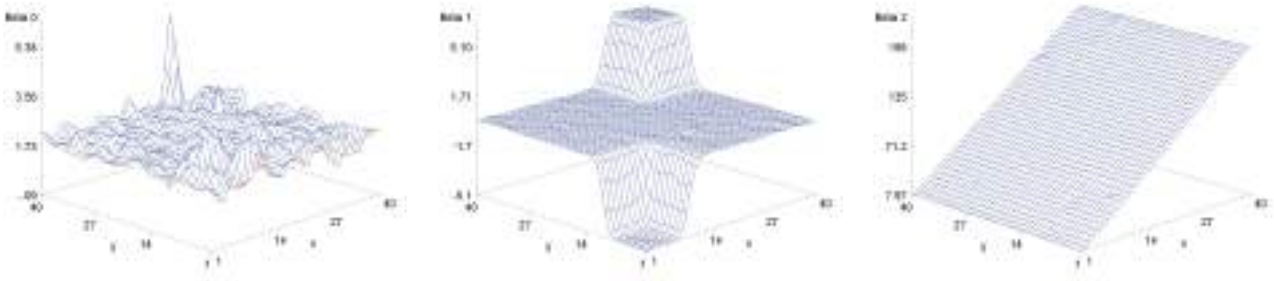
Conforme apresentado no Capítulo 2, a RGP-APP tem como proposta a estimação dos parâmetros a nível desagregado utilizando dados agregados para a variável dependente. Dessa forma, o bom funcionamento da RGP-APP produzirá dados próximos aos valores reais dos parâmetros e próximos aos valores estimados pelo modelo RGP aplicado a dados desagregados. Neste ensaio, foram utilizados dados gerados em uma grade de dimensão 40×40, utilizando os dois fatores de variação, como descrito no Capítulo 3.

As Figuras 4.5, 4.6 e 4.7 ilustram a distribuição espacial da média dos parâmetros reais, da média das estimativas produzidas pela RGP aplicada a dados desagregados e da média das estimativas RGP-APP utilizando dados agregados de forma vertical. Os resultados apresentam uma capacidade muito superior da RGP em relação a RGP-APP em reproduzir a distribuição espacial dos parâmetros. No entanto, é importante ressaltar que para obtenção desse resultado, foram utilizados dados agregados no modelo RGP-APP e que mesmo nessa condição a RGP-APP apresenta boa capacidade de estimação da tendência dos parâmetros. As médias estimadas pelo modelo OLS foram iguais a $2,53$, -0,04 e 102,69 para β_0, β_1 e β_2, respectivamente. Note que esses valores são aproximadamente os pontos médios das superfícies.



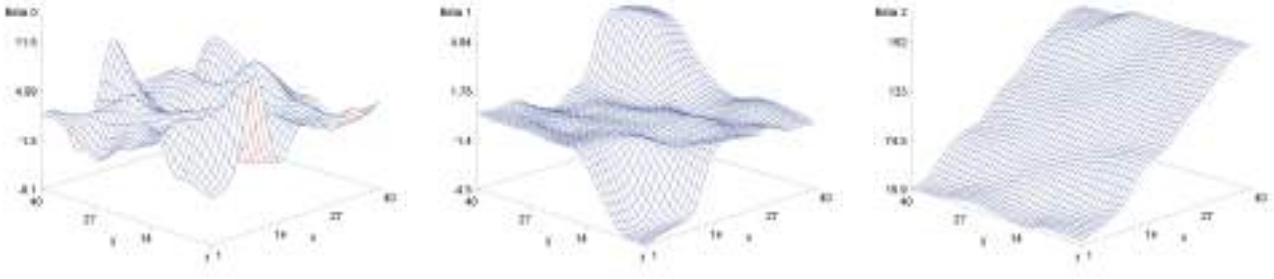
a) b) c)

Figura 4.5: Distribuição espacial dos parâmetros reais: a) β_0; b) β_1; c) β_2



a) b) c)

Figura 4.6: Distribuição espacial das estimativas - RGP aplicada a dados desagregados: a) β_0; b) β_1; c) β_2



a) b) c)

Figura 4.7: Distribuição espacial das estimativas - RGP-APP aplicada a dados agregados de forma vertical no nível 1: a) β_0; b) β_1; c) β_2

A Tabela 4.6 apresenta algumas estatísticas descritivas dos parâmetros gerados para os dados simulados, bem como para as estimativas geradas pelos modelos OLS, RGP e RGP-APP.

Tabela 4.6: Média, mínimo, máximo e desvio padrão das estimativas geradas pelos modelos RGP, RGP-APP e valores reais dos parâmetros

Nível de Agregação	Parâmetro	Modelo	Fator de Variação							
			3				10			
			Média	Min	Max	Desv	Média	Min	Max	Desv
1	β_0	Real	2,05	2,00	2,10	0,03	2,05	2,00	2,10	0,03
		RGP	2,04	-42,35	52,96	2,69	2,04	-41,56	52,96	2,57
		APP desigual	1,96	-244,36	162,57	38,69	3,37	-165,80	205,54	37,67
		APP desigual 2	2,66	-1407,98	7254,91	111,00	2,39	-1655,25	1468,62	80,15
		APP horizontal	1,82	-128,30	148,34	17,55	2,28	-200,15	127,81	17,61
		APP vertical	2,00	-128,30	171,69	17,32	2,30	-114,48	123,58	16,96
	β_1	Real	0,07	-5,00	5,00	1,86	0,07	-5,00	5,00	1,86
		RGP	0,07	-8,72	8,24	1,79	0,06	-8,49	7,61	1,80
		APP desigual	-0,10	-27,95	25,10	5,56	-0,33	-23,11	21,22	5,52
		APP desigual 2	-0,07	-485,32	195,96	11,79	-0,13	-127,78	251,63	10,74
		APP horizontal	0,09	-21,25	16,13	2,88	0,02	-22,38	26,99	2,91
		APP vertical	0,09	-31,12	23,05	2,92	0,06	-20,66	22,23	2,93
	β_2	Real	102,55	5,00	200,10	57,72	102,55	5,00	200,10	57,72
		RGP	102,56	-35,77	238,02	57,51	102,57	-40,89	248,01	57,54
		APP desigual	104,27	-204,24	408,89	77,10	103,77	-204,24	338,24	76,42
		APP desigual 2	102,90	-9053,39	1563,11	155,18	103,92	-2471,55	1992,98	121,94
		APP horizontal	102,78	-109,53	379,64	60,09	102,47	-125,04	305,01	59,95
		APP vertical	102,39	-140,02	379,64	59,96	102,11	-140,02	330,49	60,86
2	β_0	Real	2,05	2,00	2,10	0,03	2,05	2,00	2,10	0,03
		RGP	2,04	-38,28	49,16	2,65	2,04	-54,58	47,33	2,67
		APP desigual	8,37	-661,83	927,33	101,61	-0,71	-1650,23	867,71	97,30
		APP desigual 2	5,39	-1454,77	1059,90	85,83	2,06	-1358,03	802,44	79,64
		APP horizontal	2,73	-335,60	252,57	36,59	2,88	-307,27	252,57	38,00
		APP vertical	2,73	-335,60	252,57	36,59	2,88	-307,27	252,57	38,00
	β_1	Real	0,07	-5,00	5,00	1,86	0,07	-5,00	5,00	1,86
		RGP	0,07	-9,13	7,81	1,80	0,07	-8,08	7,81	1,79
		APP desigual	-0,77	-237,76	144,76	14,77	0,35	-73,87	172,68	13,26
		APP desigual 2	-0,91	-245,86	102,93	12,65	-0,17	-95,20	99,66	10,70
		APP horizontal	-0,18	-43,07	37,78	5,38	-0,07	-39,67	41,59	5,44
		APP vertical	-0,18	-43,07	37,78	5,38	-0,07	-39,67	41,59	5,44
	β_2	Real	102,55	5,00	200,10	57,72	102,55	5,00	200,10	57,72
		RGP	102,56	-30,59	250,54	57,52	102,54	-28,86	242,64	57,51
		APP desigual	98,41	-1139,76	3236,22	158,09	105,18	-987,65	1475,49	149,87
		APP desigual 2	105,45	-1105,43	1891,66	127,36	105,01	-771,75	2647,94	122,26
		APP horizontal	103,68	-217,95	573,06	72,73	102,31	-329,89	528,19	74,01
		APP vertical	103,68	-217,95	573,06	72,73	102,31	-329,89	528,19	74,01

De forma geral, as médias das estimativas produzidas pelo modelo RGP apresentam maior proximidade às médias dos valores reais dos parâmetros verdadeiros

do que as médias das estimativas produzidas pela RGP-APP. Nota-se também que a variabilidade para o conjunto de estimativas geradas pela RGP é bem menor do que a variabilidade do conjunto de estimativas geradas pela RGP-APP. Mas há de se destacar que, nesta etapa do ensaio, a RGP-APP foi aplicada a dados agregados da variável dependente, enquanto a RGP foi aplicada a dados desagregados.

As estimativas produzidas pela RGP-APP, quando utilizadas as configurações de agregação com unidades desiguais se apresentaram mais distantes dos valores verdadeiros do que quando utilizadas configurações que considerem unidades iguais. Quando considerada a agregação desigual 2, o conjunto de estimativas gerado pela RGP-APP apresenta uma variabilidade muito grande, apesar de produzir, em média, estimativas próximas às médias dos valores reais dos parâmetros. Por exemplo, as estimativas para β_2 quando considerada a agregação desigual 2 variam entre $-9053,39$ e $1563,11$. Esse resultado pode estar associada à forma complexa com que as unidades foram formadas nessa agregação.

Outro resultado importante é que a RGP-APP produz estimativas mais próximas dos valores reais quando se trabalha com o nível 1 de agregação. Por exemplo, o modelo RGP-APP gerou uma estimativa de 2,66 para β_0 quando utilizou-se dados agregados no nível 1 na configuração desigual. Já quando se considerou o nível 2 com a mesma configuração de agregação, a estimativa saltou para 8,77, enquanto o valor verdadeiro do parâmetro é de 2,05. As Figuras 4.8 e 4.9 apresentam as distribuições das diferenças entres os valores estimados pelos modelos e os valores reais dos parâmetros. Além de confirmar os resultados citados acima, a análise dessas figuras também permite a visualização do aumento da variabilidade dos valores estimados com a mudança do nível de agregação.

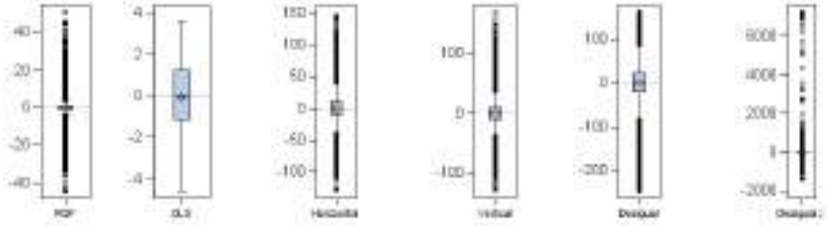
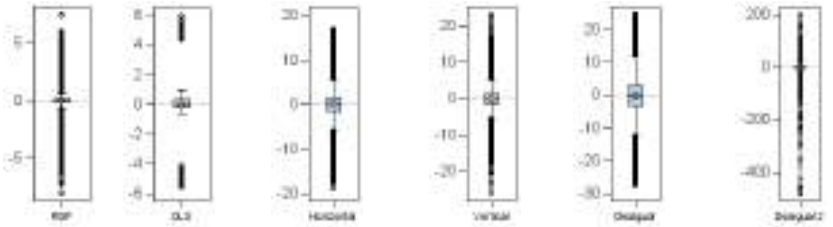
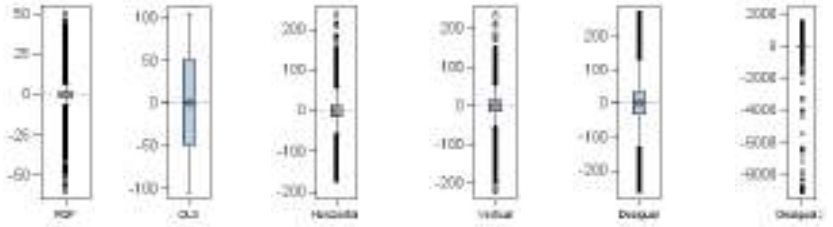
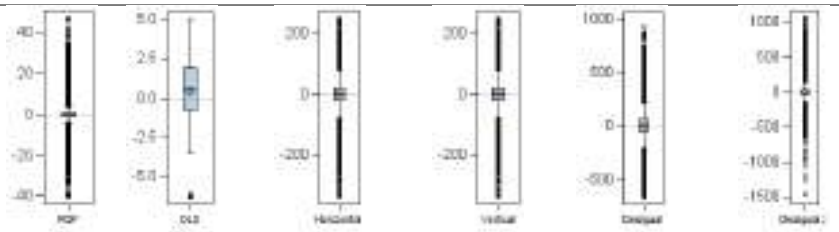
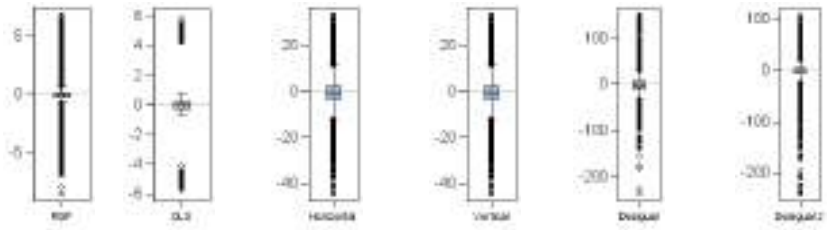
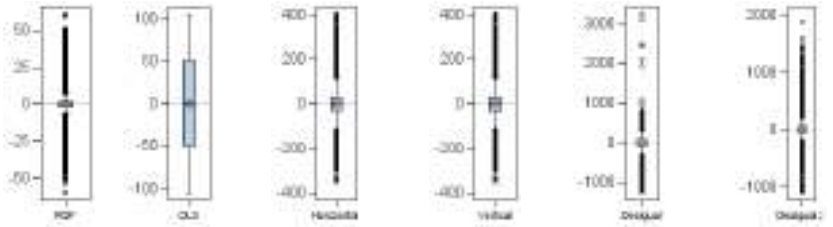
Nível Agr.	Par.	RGP	OLS	RGP-APP			
				Horizontal	Vertical	Desigual	Desigual 2
1	β_0						
	β_1						
	β_2						
2	β_0						
	β_1						
	β_2						

Figura 4.8: Distribuição das diferenças entre as estimativas produzidas pela RGP e RGP-APP e os valores verdadeiros dos parâmetros - Fator de controle da variância = 3

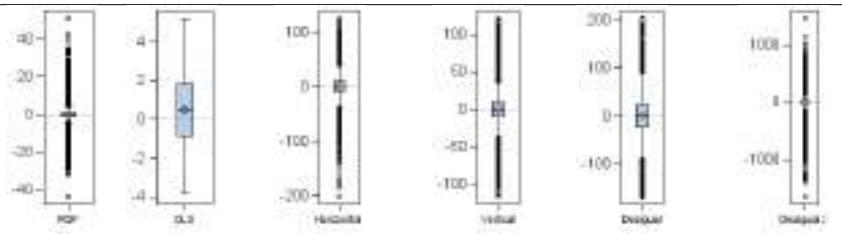
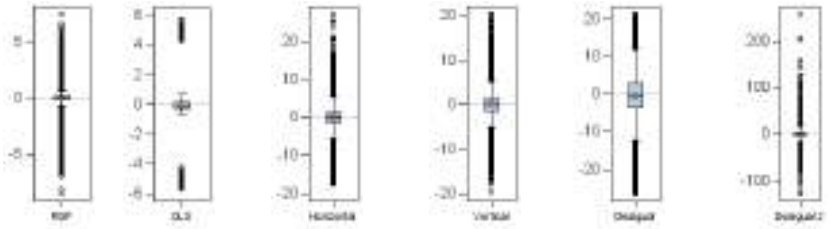
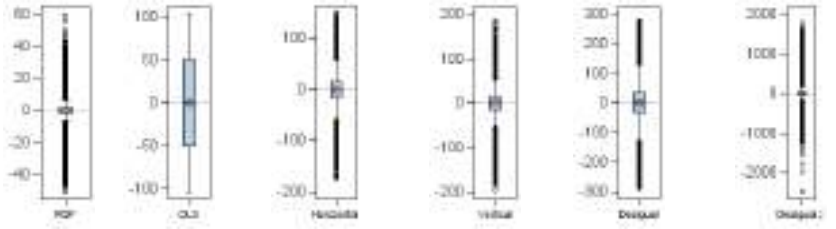
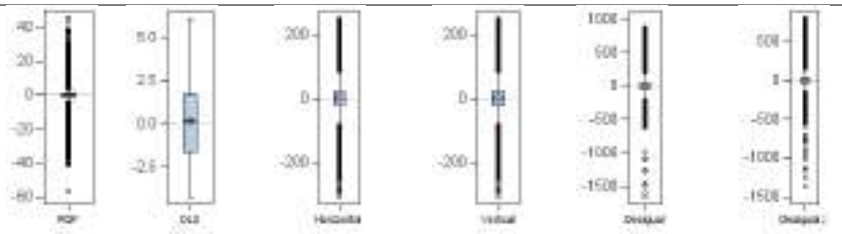
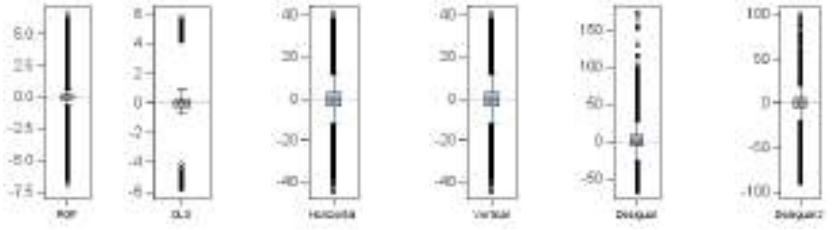
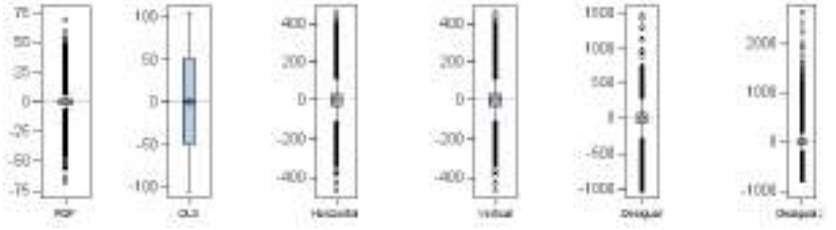
Nível Agr.	Par.	RGP	OLS	RGP-APP Horizontal	RGP-APP Vertical	RGP-APP Desigual	RGP-APP Desigual 2
1	β_0						
	β_1						
	β_2						
2	β_0						
	β_1						
	β_2						

Figura 4.9: Distribuição das diferenças entre as estimativas produzidas pela RGP e RGP-APP e os valores verdadeiros dos parâmetros - Fator de controle da variância = 10

O uso de diferentes níveis do fator de controle da variação produz impacto bem

mais discretos nos valores das estimativas que a alteração do nível de agregação. Não é possível identificar um padrão de melhora ou piora, pois em alguns casos a média das estimativas se aproxima dos valores reais, enquanto que para outros casos ela se distancia. No entanto, a análise dos dados apresentados na Tabela 4.7, que apresenta o erro quadrático médio das estimativas geradas pelos modelos, confirma que o impacto provocado pela alterção do nível de agregação é muito maior que o impacto provocado pelo fator de controle da variância nos dados gerados. Os resultados confirmam ainda que o modelo RGP-APP tende a produzir estimativas mais distantes dos valores reais em situações em que são empregadas as configurações desiguais, em especial a agregação desigual 2.

Tabela 4.7: Erro Quadrático Médio das estimativas geradas pelos modelos OLS, RGP e RGP-APP

Parâmetro	Modelo	Erro Quadrático Médio			
		Fator = 3		Fator = 10	
		Nível agreg. =1	Nível agreg. =2	Nível agreg. =1	Nível agreg. =2
β_0	**RGP**	7,22	7,00	6,62	7,14
	OLS	3,57	4,07	3,45	4,39
	APP horizontal	308,05	1339,31	310,02	1445,00
	APP vertical	299,84	1339,31	287,83	1445,00
	APP desigual	1497,15	10363,55	1420,81	9475,68
	APP desigual 2	12321,97	7377,70	6423,65	6342,40
β_1	**RGP**	0,32	0,32	0,33	0,33
	OLS	3,56	3,55	3,54	3,56
	APP horizontal	6,51	26,87	6,53	28,72
	APP vertical	6,48	26,87	6,17	28,72
	APP desigual	27,92	214,83	27,35	174,25
	APP desigual 2	136,26	157,22	113,43	110,83
β_2	**RGP**	17,35	16,93	16,23	17,86
	OLS	3340,62	3342,24	3339,61	3342,88
	APP horizontal	651,31	2727,16	631,54	2775,06
	APP vertical	623,38	2727,16	617,95	2775,06
	APP desigual	3217,94	22425,03	2824,97	20354,96
	APP desigual 2	21414,75	13658,96	11893,52	13037,57

Nesta Seção considerou-se uma situação em que os modelos RGP e OLS foram aplicados a dados desagregados e a RGP-APP foi aplicada considerando dados agregados para a variável dependente. Porém, a RGP-APP surge como uma proposta para situações em que não existe o acesso a nível de indivíduos para algumas variá-

veis. Dessa forma, na próxima Seção, os modelos RGP e OLS serão aplicados a dados agregados, tanto a variável dependente quanto as covariáveis.

4.1.4.2 Comparação das estimativas produzidas pelos modelos OLS, RGP (aplicados a dados agregados) e RGP-APP com os valores reais dos parâmetros

Nesta Seção, a RGP e a OLS serão aplicadas ao conjunto de dados agregados nos diferentes cenários apresentados no Capítulo 3. Para comparação dos resultados produzidos, serão consideradas as médias dos parâmetros verdadeiros e as médias das estimativas produzidas pela RGP-APP em cada uma das 100 repetições realizadas.

A Tabela 4.8 apresenta algumas estatísticas descritivas das estimativas produzidas pelos modelos OLS, RGP e RGP-APP, além dos valores reais dos parâmetros quando considerado o nível 1 de agregação. Como esperado, o modelo OLS se apresentou mais sensível aos impactos devidos a agregação produzindo, em média, as estimativas mais distantes em quase todos os tipos de agregação para os três parâmetros considerados neste ensaio. Como exemplo, a média das estimativas produzidas pelo modelo OLS para o parâmetro β_1, quando considerada a agregação desigual 2, foi de $0,87$, enquanto o valor real da média de β_1 foi de $0,07$. Além disso, os resultados mostram que a RGP-APP gera, em média, estimativas semelhantes às geradas pela RGP aplicada a dados agregados.

De forma geral, os modelos tendem a ser menos precisos quando consideradas agregações com quantidade desigual de unidades agregadas, especialmente quando considerada a agregação desigual 2. No entanto, a RGP-APP se apresenta mais resistente que os outros modelos aos efeitos desse tipo de agregação. Por exemplo, as estimativas produzidas para β_0, pelos modelos OLS, RGP e RGP-APP quando considerada da agregação desigual 2, foram $-2,07$, $-1,88$ e $1,49$, respectivamente, enquanto o valor real da média de β_0 foi de $2,05$.

Tabela 4.8: Média, mínimo, máximo e desvio padrão das estimativas geradas pelos modelos OLS, RGP, RGP-APP e valores reais dos parâmetros - Dados agregados no nível 1

Parâmetro	Método	Fator							
		3				10			
		Média	Min	Max	Desv	Média	Min	Max	Desv
β_0	Real	2,05	2,05	2,05	0,00	2,05	2,05	2,05	0,00
	OLS vertical	2,31	-37,49	36,83	16,05	5,39	-30,37	29,89	13,98
	OLS horizontal	2,32	-43,29	39,03	16,50	5,70	-27,89	35,61	13,55
	OLS desigual	-4,17	-128,71	111,27	57,32	8,43	-157,81	138,83	55,35
	OLS desigual 2	-2,07	-99,99	101,32	44,30	6,43	-153,54	101,24	47,32
	RGP vertical	1,99	-8,67	6,77	2,63	2,30	-8,67	8,90	3,09
	RGP horizontal	1,81	-7,27	9,21	2,94	2,19	-7,27	8,45	2,72
	RGP desigual	1,53	-18,19	20,60	8,65	3,11	-15,09	32,15	8,33
	RGP desigual 2	-1,88	-56,28	30,06	15,98	1,59	-41,26	72,01	18,56
	APP vertical	2,00	-8,09	6,63	2,55	2,30	-8,09	8,93	3,01
	APP horizontal	1,82	-7,85	9,80	3,09	2,28	-7,85	9,91	2,92
	APP desigual	1,73	-37,81	23,23	11,19	3,60	-23,67	27,06	10,34
	APP desigual 2	1,49	-57,06	101,81	18,94	1,68	-37,66	45,72	15,34
β_1	Real	0,07	0,07	0,07	0,00	0,07	0,07	0,07	0,00
	OLS vertical	-0,14	-7,01	7,65	2,61	-0,61	-7,01	3,85	2,05
	OLS horizontal	-0,12	-6,05	8,19	2,56	-0,67	-6,05	3,70	1,97
	OLS desigual	-0,05	-18,94	18,22	7,07	-1,83	-16,96	18,28	6,88
	OLS desigual 2	0,87	-10,93	14,10	5,80	0,02	-14,84	19,53	6,72
	RGP vertical	0,09	-0,84	1,90	0,38	0,06	-1,05	1,90	0,41
	RGP horizontal	0,09	-0,93	1,63	0,43	0,04	-0,83	1,63	0,37
	RGP desigual	-0,02	-2,93	3,33	1,23	-0,10	-4,18	2,96	1,26
	RGP desigual 2	0,70	-4,25	7,44	2,17	0,52	-5,59	4,97	2,27
	APP vertical	0,09	-0,53	1,81	0,35	0,06	-0,97	1,81	0,40
	APP horizontal	0,09	-0,97	1,69	0,45	0,02	-1,02	1,69	0,40
	APP desigual	-0,03	-4,14	4,23	1,68	-0,26	-3,50	2,85	1,35
	APP desigual 2	0,38	-6,07	5,24	2,21	0,38	-5,18	6,04	2,03
β_2	Real	102,55	102,55	102,55	0,00	102,55	102,55	102,55	0,00
	OLS vertical	104,08	48,26	167,83	23,46	102,62	61,16	156,96	21,45
	OLS horizontal	103,80	53,33	165,40	23,83	102,54	58,06	160,34	20,91
	OLS desigual	115,78	-18,52	349,55	78,60	108,53	-73,95	287,95	78,62
	OLS desigual 2	111,52	-27,62	271,32	68,19	102,79	-21,24	268,69	64,01
	RGP vertical	102,47	95,55	111,01	3,74	102,08	93,02	111,01	3,71
	RGP horizontal	102,83	95,37	111,11	3,77	102,52	92,88	109,14	3,27
	RGP desigual	104,39	78,66	129,61	11,76	102,06	75,88	128,13	10,38
	RGP desigual 2	110,05	54,90	161,62	22,30	104,89	10,05	176,27	23,58
	APP vertical	102,39	96,15	112,36	3,67	102,11	93,15	110,63	3,61
	APP horizontal	102,78	94,67	112,25	4,02	102,47	92,78	109,69	3,48
	APP desigual	104,20	73,10	145,01	15,63	102,81	72,29	134,96	13,81
	APP desigual 2	106,97	-18,96	174,26	26,97	106,44	46,65	157,53	20,89

Tabela 4.9: Média, mínimo, máximo e desvio padrão das estimativas geradas pelos modelos OLS, RGP, RGP-APP e valores reais dos parâmetros - Dados agregados no nível 2

Parâmetro	Método	Fator							
		3				10			
		Média	Min	Max	Desv	Média	Min	Max	Desv
β_0	**Real**	2,05	2,05	2,05	0,00	2,05	2,05	2,05	0,00
	OLS vertical	8,29	-62,75	85,02	30,71	1,42	-78,18	55,41	31,16
	OLS horizontal	8,29	-62,75	85,02	30,71	1,42	-78,18	55,41	31,16
	OLS desigual	11,68	-257,80	275,79	111,25	7,41	-225,42	253,01	110,29
	OLS desigual 2	9,66	-112,88	106,28	48,17	8,01	-133,03	107,27	48,47
	RGP vertical	2,64	-16,63	25,61	8,02	2,69	-14,70	25,90	7,00
	RGP horizontal	2,64	-16,63	25,61	8,02	2,69	-14,70	25,90	7,00
	RGP desigual	3,73	-93,94	113,90	34,54	1,08	-65,64	61,35	25,26
	RGP desigual 2	2,91	-66,39	62,39	23,84	3,71	-55,53	86,56	21,00
	APP vertical	2,73	-17,53	26,47	8,22	2,88	-15,02	23,78	7,10
	APP horizontal	2,73	-17,53	26,47	8,22	2,88	-15,02	23,78	7,10
	APP desigual	6,88	-89,85	96,56	29,77	0,95	-114,21	62,99	27,69
	APP desigual 2	4,33	-55,57	63,27	22,02	2,96	-50,75	61,49	18,92
β_1	**Real**	0,07	0,07	0,07	0,00	0,07	0,07	0,07	0,00
	OLS vertical	-1,41	-16,14	11,26	4,36	-0,69	-12,09	10,94	4,47
	OLS horizontal	-1,41	-16,14	11,26	4,36	-0,69	-12,09	10,94	4,47
	OLS desigual	-3,73	-32,34	29,74	14,42	-4,01	-46,47	30,68	14,84
	OLS desigual 2	-0,78	-15,41	11,98	5,98	0,03	-15,84	15,36	6,16
	RGP vertical	-0,17	-3,44	2,02	1,02	-0,05	-2,89	1,96	0,98
	RGP horizontal	-0,17	-3,44	2,02	1,02	-0,05	-2,89	1,96	0,98
	RGP desigual	-0,31	-12,85	16,83	4,99	0,48	-8,60	11,78	4,08
	RGP desigual 2	-0,17	-8,42	5,38	2,88	-0,37	-8,79	7,19	2,65
	APP vertical	-0,18	-3,58	2,49	1,07	-0,07	-2,70	1,98	1,01
	APP horizontal	-0,18	-3,58	2,49	1,07	-0,07	-2,70	1,98	1,01
	APP desigual	-0,46	-11,71	15,16	4,27	0,24	-9,45	8,57	3,77
	APP desigual 2	-0,55	-9,34	7,07	2,68	-0,27	-7,35	8,22	2,55
β_2	**Real**	102,55	102,55	102,55	0,00	102,55	102,55	102,55	0,00
	OLS vertical	104,66	-20,29	241,46	48,98	111,72	-28,77	210,01	50,13
	OLS horizontal	104,66	-20,29	241,46	48,98	111,72	-28,77	210,01	50,13
	OLS desigual	121,10	-268,80	454,82	175,61	132,82	-334,05	642,28	173,07
	OLS desigual 2	92,75	-51,02	267,70	68,62	87,94	-61,19	255,14	72,33
	RGP vertical	103,70	77,73	136,46	12,39	102,45	82,80	129,44	9,54
	RGP horizontal	103,70	77,73	136,46	12,39	102,45	82,80	129,44	9,54
	RGP desigual	103,29	-40,50	213,19	47,01	100,28	-43,01	189,94	39,82
	RGP desigual 2	97,88	3,38	204,07	33,47	98,36	-0,72	204,25	34,31
	APP vertical	103,68	75,93	136,87	12,72	102,31	82,15	129,70	9,83
	APP horizontal	103,68	75,93	136,87	12,72	102,31	82,15	129,70	9,83
	APP desigual	98,39	-17,70	195,44	40,59	102,98	-30,89	353,92	44,56
	APP desigual 2	99,47	3,83	198,97	32,23	99,70	6,12	198,97	30,01

A comparação dos dados da Tabela 4.8 com os dados da Tabela 4.9 evidencia o forte impacto do MAUP quando considerado o nível de agregação 2. Mais uma vez, o modelo OLS sofreu os maiores impactos dessa nova agregação. Por exemplo, a média das estimativas produzidas para β_0 no nível de agregação 1 foi de $2,31$, enquanto que, no novo nível, essa estimativa saltou para $8,29$, se distanciando ainda mais da média do valor real do parâmetro igual a 2,05. No caso das estimativas para β_1, percebe-se que houve inversão de sinal em relação ao valor real, para todas as estimativas produzidas pelos modelos considerados.

As Figuras 4.10 e 4.11 apresentam as distribuições das diferenças entre as médias das estimativas produzidas pelos modelos e as médias dos valores verdadeiros dos parâmetros por repetição. Na Figura 4.10, o fator de controle da variância foi fixado em 3 e os resultados mostraram que a RGP-APP gera, em média, distribuições semelhantes às geradas pela RGP aplicada a dados agregados, com algumas diferenças em relação a variabilidade dessas estimativas. Nota-se, ainda, que a regressão OLS não captura a tendência dos parâmetros, fazendo com que seus resultados sofram mais ainda com os efeitos da agregação.

Percebe-se, também, que a variabilidade é maior quando são utilizadas as agregações desiguais e níveis maiores de agregação. Quando utilizada a agregação desigual no primeiro nível de agregação, as estimativas geradas pela RGP apresentam menor variabilidade do que aquelas geradas pela RGP-APP. No entanto, quando considerado o nível de agregação 2, a RGP parece sofrer mais com os efeitos da agregação do que RGP-APP que produziu estimativas com menor variabilidade.

Analisando a Figura 4.11, onde o fator de controle da variância foi fixado em 10, percebe-se que as diferenças possuem comportamento semelhante às produzidas com o fator fixado em 3. No entanto, nota-se uma redução na variabilidade das estimativas produzidas, resultado ratificado pelos dados da Tabela 4.10 que apresenta os erros quadráticos médios das estimativas produzidas pelos modelos.

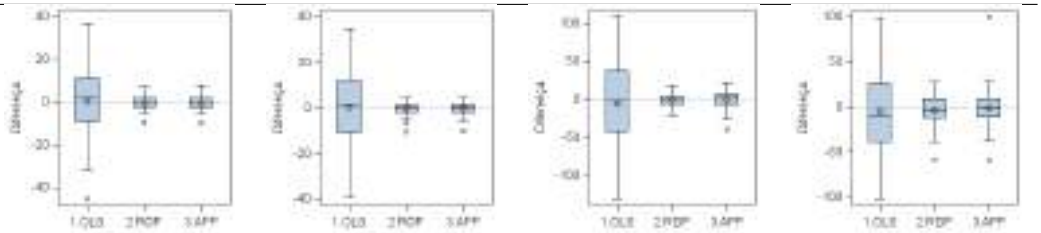
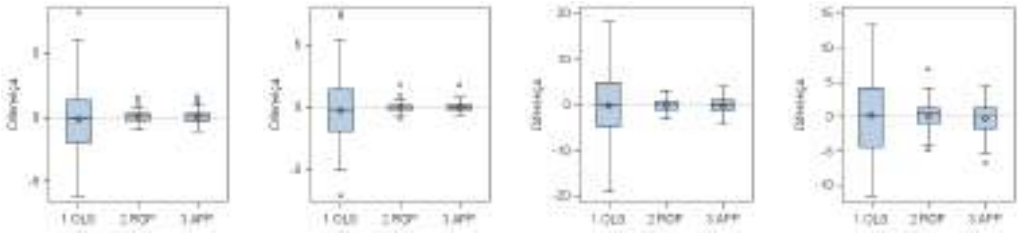
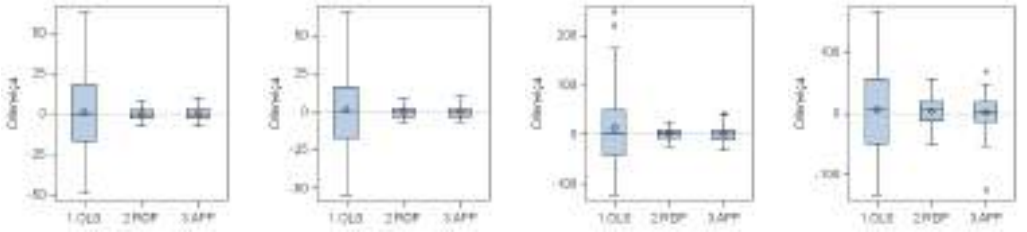
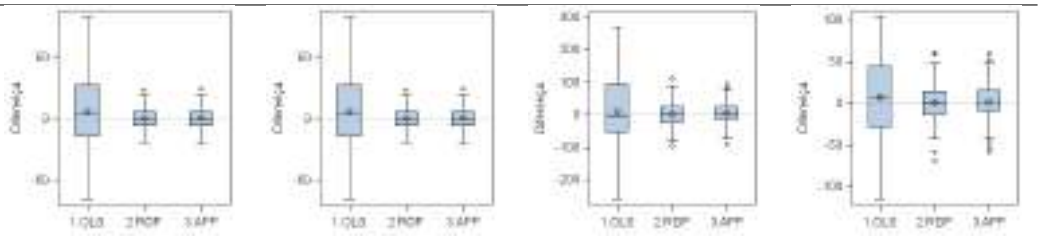
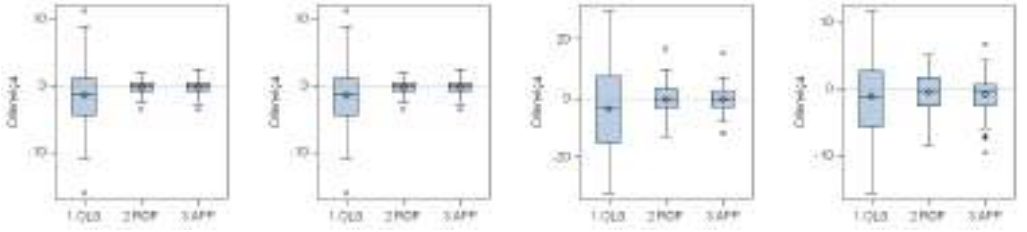
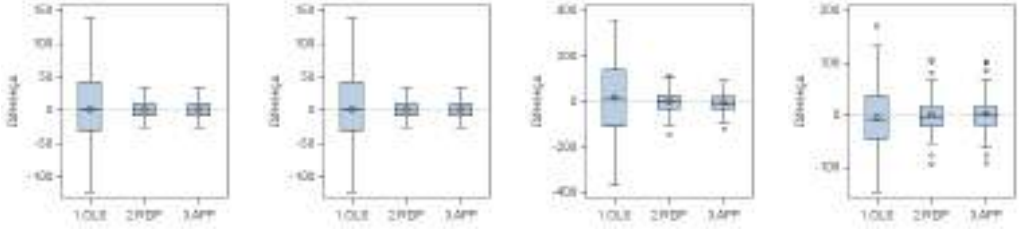
Nível Agr.	Par.	Tipo de Agregação			
		Horizontal	Vertical	Desigual	Desigual 2
1	β_0				
	β_1				
	β_2				
2	β_0				
	β_1				
	β_2				

Figura 4.10: Distribuição das diferenças entre as estimativas produzidas pela RGP e RGP-APP e os valores verdadeiros dos parâmetros - Fator de controle da variância = 3

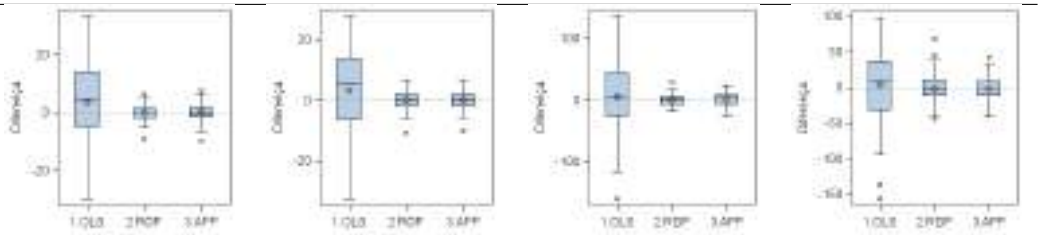
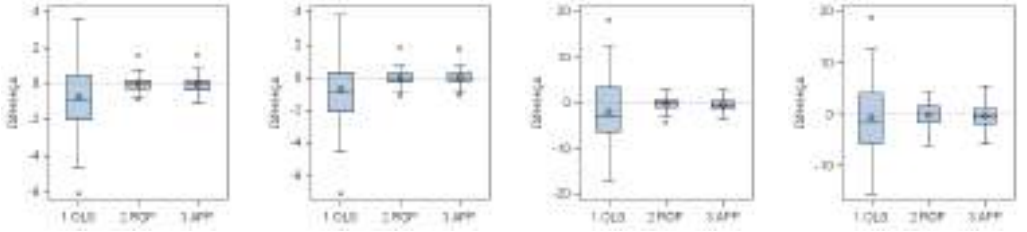
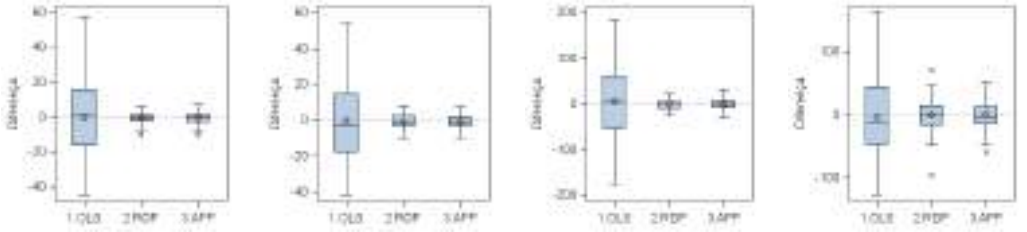
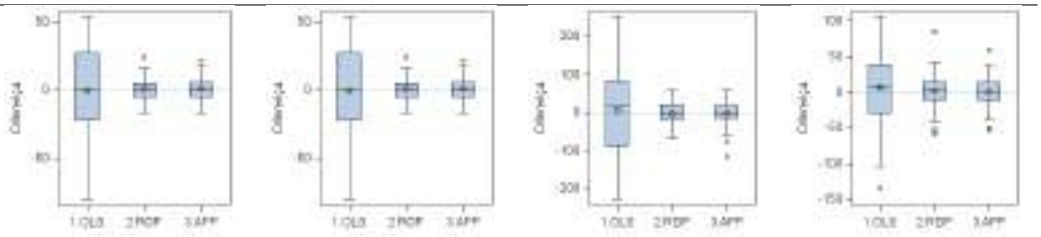
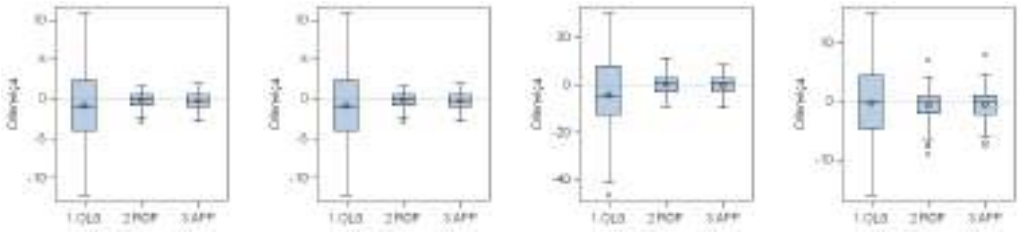
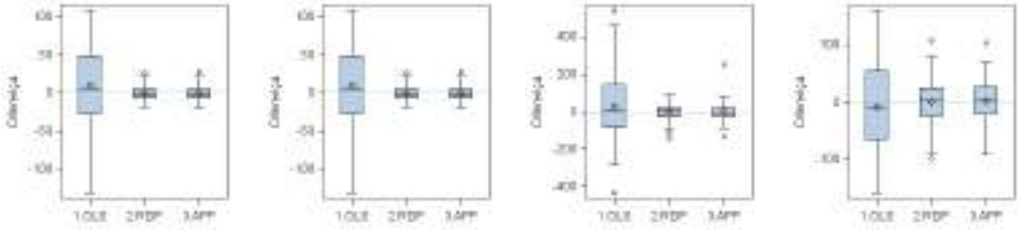
Nível Agr.	Par.	Tipo de Agregação			
		Horizontal	Vertical	Desigual	Desigual 2
1	β_0				
	β_1				
	β_2				
2	β_0				
	β_1				
	β_2				

Figura 4.11: Distribuição das diferenças entre as estimativas produzidas pela RGP e RGP-APP e os valores verdadeiros dos parâmetros - Fator de controle da variância = 10

Tabela 4.10: Erro Quadrático Médio das estimativas geradas pelos modelos OLS, RGP e RGP-APP - Dados agregados

Parâmetro	Modelo	Agregação	Erro Quadrático Médio			
			Fator = 3		Fator = 10	
			Nível agreg.=1	Nível agreg.=2	Nível agreg.=1	Nível agreg.=2
β_0	OLS	Horizontal	268,82	972,70	194,99	961,74
		Vertical	254,41	972,70	204,73	961,74
		Desigual	3282,64	12344,48	3073,50	12071,76
		Desigual 2	1954,72	2354,82	2236,29	2361,42
	RGP	Horizontal	8,59	63,98	7,33	48,86
		Vertical	6,83	63,98	9,53	48,86
		Desigual	74,24	1183,69	69,83	632,47
		Desigual 2	267,75	563,31	341,20	439,42
	APP	Horizontal	9,51	67,42	8,51	50,65
		Vertical	6,40	67,42	9,01	50,65
		Desigual	123,77	900,45	108,31	760,14
		Desigual 2	354,59	485,27	233,07	355,03
β_1	OLS	Horizontal	6,51	21,01	4,37	20,37
		Vertical	6,78	21,01	4,64	20,37
		Desigual	49,39	220,11	50,38	234,70
		Desigual 2	33,33	36,56	45,03	37,63
	RGP	Horizontal	0,18	1,09	0,13	0,97
		Vertical	0,14	1,09	0,17	0,97
		Desigual	1,49	24,78	1,60	16,67
		Desigual 2	4,68	8,39	5,10	7,41
	APP	Horizontal	0,20	1,20	0,16	1,02
		Vertical	0,12	1,20	0,16	1,02
		Desigual	2,80	18,30	1,92	14,11
		Desigual 2	4,87	7,81	4,13	6,73
β_2	OLS	Horizontal	562,47	2379,92	432,75	2572,51
		Vertical	545,90	2379,92	455,34	2572,51
		Desigual	6276,14	30873,91	6154,61	30568,71
		Desigual 2	4623,91	4679,28	4065,24	5259,39
	RGP	Horizontal	14,14	153,26	10,59	90,06
		Vertical	13,80	153,26	13,87	90,06
		Desigual	139,91	2188,30	107,00	1574,86
		Desigual 2	508,89	1109,72	551,49	1167,82
	APP	Horizontal	15,98	161,51	11,97	95,69
		Vertical	13,34	161,51	13,12	95,69
		Desigual	243,88	1648,12	188,94	1966,34
		Desigual 2	719,58	1035,21	432,54	899,51

Os resultados apresentados na Tabela 4.10 mostram ainda que a RGP-APP se mostrou mais resistente aos efeitos da agregação do que o modelo RGP. Por exemplo,

enquanto o erro quadrático médio da estimativa produzida pela RGP para β_1 cresceu mais de 15 vezes quando se alterou o nível de agregação, esse crescimento foi menor do que 7 vezes para o erro quadrático médio produzido pela RGP-APP.

A Tabela 4.11 apresenta o percentual de casos em que a média das estimativas por repetição do ensaio para a RGP-APP se aproximou mais das médias dos valores reais do que as estimativas produzidas pela RGP ou OLS. Os resultados mostram que a RGP-APP apresenta uma superioridade em relação a OLS em quase todos os casos avaliados. Nota-se que essa superioridade é ainda maior quando considerados o nível de agregação 2 e o fator de controle da variância fixado em 10. A única situação em que o desempenho não melhora com a alteração do fator de variação é para a agregação desigual no nível 2.

Já em relação a RGP, a RGP-APP apresentou melhores resultados para quase todas as situações em que foram utilizadas as agregações horizontais ou verticais. A única exceção foi quando se considerou a agregação horizontal com fator de controle da variância fixado em 10. Quando consideradas as agregações desiguais no nível de agregação 1, a RGP apresentou melhor desempenho do que a RGP-APP. No entanto, quando considerada a agregação 2, a RGP-APP apresenta um melhor desempenho do que a RGP em boa parte dos casos. Esse resultado indica que a RGP-APP apresentou maior resistência aos efeitos da agregação que a RGP aplicada a dados agregados.

Tabela 4.11: Percentual de vezes em que a RGP-APP se aproximou mais dos parâmetros reais que a OLS ou RGP aplicadas a dados agregados

Modelo	Nível agregação	Fator de controle da variância	Tipo de agregação	β_0	β_1	β_2
OLS	**1**	**3**	Vertical	71%	73%	77%
			Horizontal	71%	70%	76%
			Desigual	66%	70%	69%
			Desigual 2	68%	66%	69%
		10	Vertical	92%	86%	95%
			Horizontal	87%	90%	93%
			Desigual	86%	91%	92%
			Desigual 2	83%	86%	84%
	2	**3**	Vertical	88%	87%	89%
			Horizontal	88%	87%	89%
			Desigual	85%	89%	87%
			Desigual 2	74%	67%	80%
		10	Vertical	90%	91%	89%
			Horizontal	90%	91%	89%
			Desigual	97%	90%	94%
			Desigual 2	80%	78%	80%
RGP	**1**	**3**	Vertical	71%	70%	71%
			Horizontal	54%	55%	53%
			Desigual	43%	44%	42%
			Desigual 2	50%	49%	49%
		10	Vertical	51%	51%	51%
			Horizontal	44%	44%	42%
			Desigual	34%	35%	33%
			Desigual 2	38%	38%	39%
	2	**3**	Vertical	66%	66%	65%
			Horizontal	66%	66%	65%
			Desigual	55%	57%	59%
			Desigual 2	49%	50%	50%
		10	Vertical	67%	67%	66%
			Horizontal	67%	67%	66%
			Desigual	57%	58%	57%
			Desigual 2	48%	49%	49%

Na próxima Seção será analisada a resistência da RGP-APP ao MAUP quando aplicada a um conjunto de dados reais provenientes da PDAD 2018. Será empregada uma versão espacial de um modelo que associa os rendimentos dos responsáveis pelos domicílios ao nível de escolaridade e experiência profissional.

4.2 Estudo com Dados Reais

Nesta Seção é realizada a avaliação da RGP-APP em uma aplicação aos dados da PDAD 2018, conforme descrito no Capítulo 3.

4.2.1 Análise da Autocorrelação Espacial

Para a avaliação da presença de autocorrelação espacial na variável rendimento do responsável pelo domicílio, foi calculado o Índice I de Moran para o conjunto de dados agregados a nível de setor censitário, considerando a média dos rendimentos como a medida de agregação. Foi obtido o valor de 0,11, indicando a presença de baixa autocorrelação espacial. A Figura 4.12 apresenta o Mapa de Moran para a região estudada. A partir deste ponto, afim de se obter melhor visualização dos resultados, as figuras representarão apenas os setores visitados pela PDAD 2018.

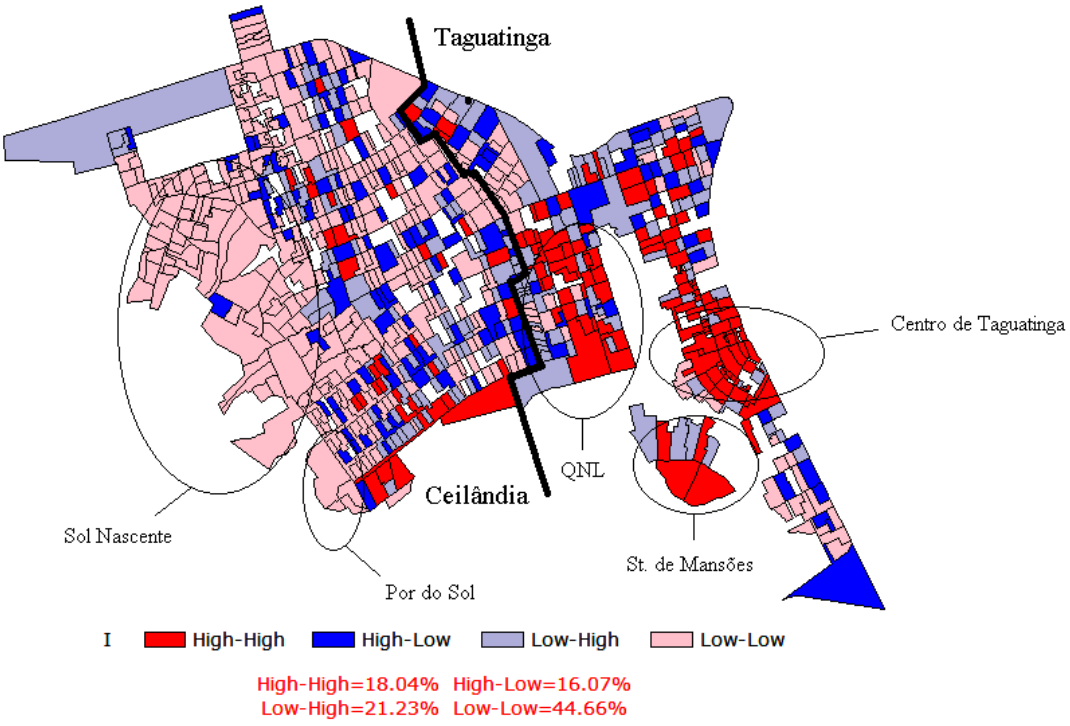


Figura 4.12: Mapa de Moran - Rendimento do responsável

Nota-se que em áreas próximas ao centro de Taguatinga, no Setores de Mansões de Taguatinga e parte da quadra QNL há concentração de setores classificados como Alto-Alto, ou seja, setores que apresentaram altos rendimentos, cercados por outros

setores que também apresentaram altos rendimentos. À medida em que se afasta dessas regiões, nota-se uma mudança de padrão. Em Ceilândia, quase a totalidade do setores Por do Sol e Sol Nascente foi classificada como Baixo-Baixo.

4.2.2 Análise dos dados a nível desagregado

Nesta primeira etapa, foram considerados os modelos OLS e RGP a fim de se estimar os parâmetros no nível mais desagregado dos dados, ou seja, no nível de domicílio. É importante destacar que os parâmetros estimados pela RGP no nível de domicílio serão considerados como os verdadeiros valores dos parâmetros para a análise nesta Seção. A realização da análise nesse nível, só foi possível pela autorização da Companhia de Planejamento do Distrito Federal - CODEPLAN, que concedeu acesso aos microdados da PDAD 2018 com identificação das coordenadas dos domicílios.

A Figura 4.13 apresenta a distribuição espacial dos domicílios visitados. Para todas as etapas do estudo com dados reais, foi definido o uso do parâmetro de suavização do tipo fixo.

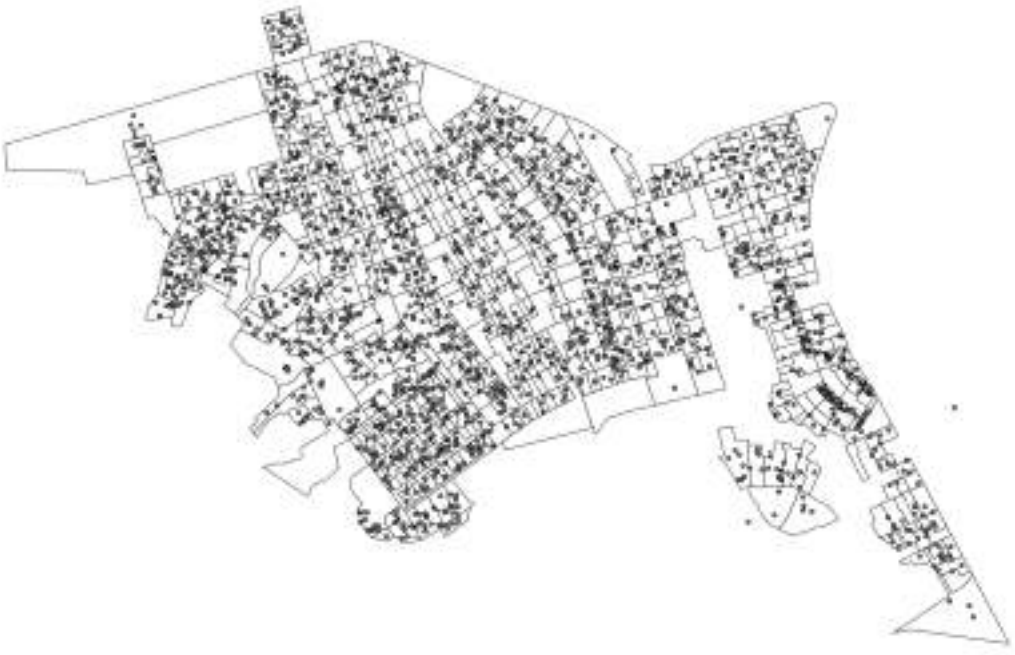


Figura 4.13: Distribuição espacial dos domicílios

A Figura 4.14 ilustra a busca do parâmetro de suavização ótimo, definido em

$b = 2,11\,km$. Ou seja, nesta etapa serão considerados no ajuste do modelo para um determinado domicílio, todos os vizinhos situados até 2,11km de distância.

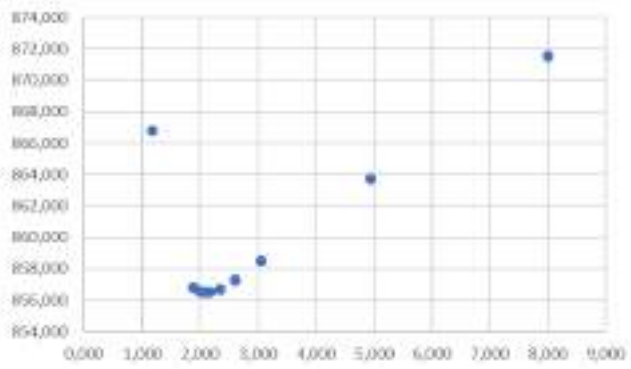


Figura 4.14: Parâmetro de suavização da RGP que minimiza o CV

A Tabela 4.12 apresenta as estimativas globais obtidas pelo modelo OLS e as médias das estimativas produzidas pelo modelo RGP.

Tabela 4.12: Estimativas dos parâmetros pelos modelos OLS e RGP - Dados desagregados

Método	Estatística	Intercepto	Escolaridade	Experiência	Experiência^2
OLS	Estimativa	5,74	0,13	0,0167	0,000003
	Erro Padrão	0,10	0,01	0,0044	0,000058
	Valor t	56,68	23,89	3,7591	0,052179
	Pr>|t|	0,00	0,00	0,0002	0,958393
RGP DESAGREGADA	Média	5,82	0,12	0,0191	-0,000053
	Mínimo	5,45	0,04	-0,0032	-0,000463
	P25	5,56	0,11	0,0087	-0,000173
	P50	5,77	0,13	0,0212	-0,000089
	P75	6,02	0,13	0,0294	0,000055
	Máximo	6,81	0,15	0,0450	0,000255

Os resultados mostram que as médias das estimativas produzidas pelo modelo RGP se aproximam das estimativas do modelo OLS, principalmente para os coeficientes das variáveis Escolaridade e Experiência. No entanto, o modelo RGP é um modelo de regressão local, que tem o potencial de identificar aspectos locais ignorados pelo modelo OLS. Os dados mostram que pelo modelo de regressão OLS a cada ano de estudo acrescentado ao indivíduo, considerados constantes todos os demais fatores, haverá uma retorno de aproximadamente 0,13% em seu rendimento individual. Já para o modelo RGP esse retorno varia entre 0,04% e 0,15%, dependendo de sua

localização espacial. Percebe-se que para a variável Experiencia^2 o efeito estimado é muito próximo de zero para o modelo RGP, e que para o modelo OLS esta não se mostrou significativa.

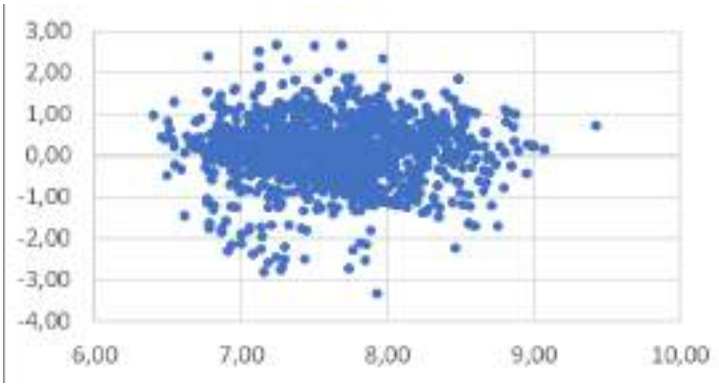


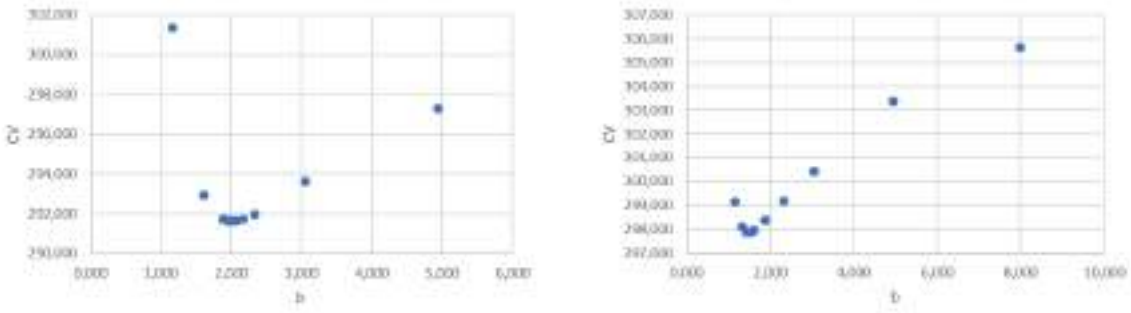
Figura 4.15: Resíduos do modelo RGP

A Figura 4.15 mostra que os resíduos produzidos pelo modelo RGP estão distribuídos de forma aleatória em torno de zero. O Coeficiente de Determinação do modelo RGP foi de $R^2 = 0,31$, enquanto que o do modelo OLS foi de $R^2 = 0,27$.

4.2.3 Análise dos dados agregados a nível de setor censitário

Para avaliação dos efeitos do MAUP e do desempenho do modelo RGP-APP em atenuar os seus efeitos, os dados foram agregados por setores censitários, considerando como medida de agregação a média dos rendimentos dos responsáveis, a média dos anos de estudo e a média dos anos de experiência. Nesta etapa, para os modelos OLS e RGP, foram utilizadas tanto as covariáveis quanto a variável dependente de forma agregada. Já para a RGP-APP, somente a variável dependente teve seus valores agregados por setores censitários, enquanto as covariáveis foram utilizadas em seu formato original, ou seja, a nível de domicílios.

Os parâmetros de suavização determinados para a RGP e RGP-APP foram de $b = 2,02\,km$ e $b = 1,50\,km$, respectivamente. A Figura 4.16 ilustra o processo de busca do valor ótimo para os dois modelos.



a) b)

Figura 4.16: Parâmetro de suavização que minimiza o CV: a) RGP b) RGP-APP

Comparando os resultados apresentados nas Tabelas 4.12 e 4.13, vemos que neste primeiro nível de agregação, as estimativas já sofreram os efeitos da agregação, pois foram alteradas em seus valores. Por exemplo, as estimativas produzidas para o intercepto estimado pelo modelo OLS para dados desagregados e agregados foram de aproximadamente 5,74 e 5,60, respectivamente. Já para o modelo RGP os valores foram de 5,82 para dados desagregados para 5,75 para dados agregados. Essas variações caracterizam o efeito do MAUP.

As estimativas obtidas pelo modelo RGP-APP apresentaram certa robustez, produzindo estimativas mais próximas aos valores reais dos parâmetros do que os modelos OLS e RGP. Os resultados mostram, por exemplo, que a taxa de retorno para cada ano de experiência adicionado ao responsável pelo domicílio estimada pelo modelo RGP-APP é em média de 0,02% sobre seu rendimento individual, podendo variar entre -0,04% a 0,05%, dependendo de sua localização.

Tabela 4.13: Estimativas dos parâmetros pelos modelos OLS e RGP - Dados agregados por setores censitários

Método	Estatística	Intercepto	Escolaridade	Experiência	Experiência^2
OLS	**Estimativa**	5,60	0,14	0,0269	-0,000114
	Erro Padrão	0,15	0,01	0,0067	0,000089
	Valor t	36,38	16,99	4,0365	-1,279150
	Pr>|t|	0,00	0,00	0,0001	0,201245
RGP AGREGADA	**Média**	5,75	0,12	0,0297	-0,000183
	Mínimo	5,09	-0,02	-0,0498	-0,001009
	P25	5,34	0,11	0,0116	-0,000416
	P50	5,64	0,13	0,0352	-0,000275
	P75	6,12	0,14	0,0479	0,000069
	Máximo	8,08	0,15	0,0765	0,000820
RGP-APP	**Média**	5,84	0,13	0,0184	0,000028
	Mínimo	5,35	0,02	-0,0364	-0,000391
	P25	5,48	0,12	0,0008	-0,000244
	P50	5,74	0,13	0,0220	-0,000081
	P75	6,04	0,15	0,0363	0,000196
	Máximo	7,63	0,19	0,0486	0,000577
RGP DESAGREGADA	**Média**	5,82	0,12	0,0191	-0,000053
	Mínimo	5,45	0,04	-0,0032	-0,000463
	P25	5,56	0,11	0,0087	-0,000173
	P50	5,77	0,13	0,0212	-0,000089
	P75	6,02	0,13	0,0294	0,000055
	Máximo	6,81	0,15	0,0450	0,000255

As Figuras 4.17, 4.18 e 4.19 ilustram a distribuição espacial das estimativas dos parâmetros β_0, β_1 e β_2 produzidas pelo modelo RGP com dados desagregados, agregados e RGP-APP. Como citado anteriormente, os resultados gerados pela RGP aplicada a dados desagregados, são a referência para esta Seção. Desta forma, o comportamento ideal para a RGP-APP seria uma aproximação a RGP desagregada. É importante ressaltar que os parâmetros foram estimados pela utilização de uma grade de 5.308 pontos, proporcionando uma boa resolução às figuras. No entanto os ajustes foram realizados com uma amostra de 1.741 domicílios.

a) b) c)

Figura 4.17: Distribuição espacial do Intercepto: a) RGP desagregada; b) RGP agregada por setores censitários; c) RGP-APP

a) b) c)

Figura 4.18: Distribuição espacial da Escolaridade: a) RGP desagregada; b) RGP agregada por setores censitários; c) RGP-APP

a) b) c)

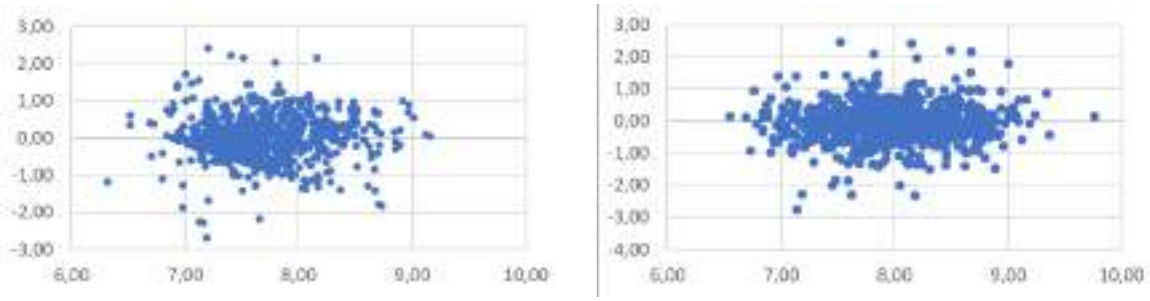
Figura 4.19: Distribuição espacial da Experiência: a) RGP desagregada; b) RGP agregada por setores censitários; c) RGP-APP

a) b) c)

Figura 4.20: Distribuição espacial da Experiência^2: a) RGP desagregada; b) RGP agregada por setores censitários; c) RGP-APP

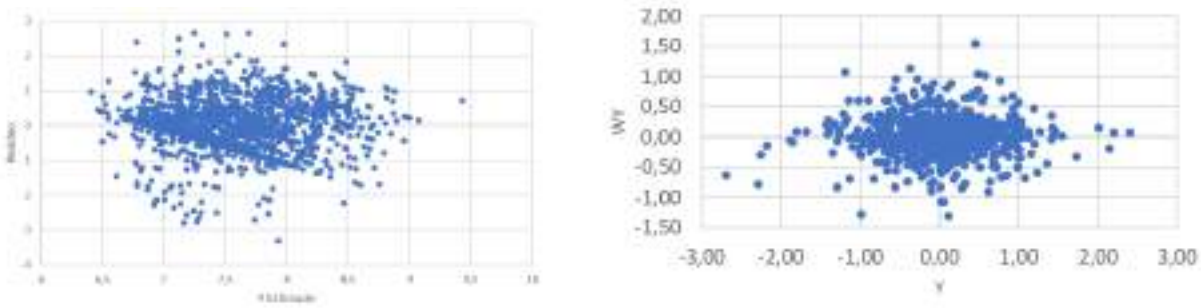
A análise das Figuras 4.17, 4.18 e 4.19 sugere que, tanto a RGP aplicada a dados agregados, quanto a RGP-APP apresentam uma boa aproximação da distribuição gerada pelo modelo RGP aplicado a dados desagregados, com uma ligeira vantagem da RGP-APP. Além disso, os resultados mostram, por exemplo, que a variável escolaridade trará maiores retornos a indivíduos que residem em áreas com alta concentração de renda, como o centro de Taguatinga, o Setor de Mansões e parte do setor QNL.

Os resíduos produzidos pelos modelos RGP aplicado a dados agregados e RGP-APP apresentam distribuição em torno de zero, como ilustra a Figura 4.21. O coeficiente de determinação para os dois modelos foi de $0,33$ e $0,32$, respectivamente. A Figura 4.22 apresenta os resíduos produzidos pela regressão OLS e o diagrama de espalhamento de Moran aplicado aos resíduos. A Figura 4.22 b) apresenta uma melhor distribuição dos resíduos, indicando que não há dependência espacial no resíduos produzidos pela RGP. Esse resultado é confirmado pelo valor de $I = 0$ para os resíduos.



a) b)

Figura 4.21: Distribuição dos resíduos produzidos: a) RGP e b) RGP-APP



a) b)

Figura 4.22: a) Distribuição dos resíduos - OLS e b) Diagrama de espalhamento de Moran aplicado aos resíduos da RGP

Como apresentado no Capítulo 2, o objetivo da RGP-APP é produzir estimativas desagregadas a partir de dados agregados. Isso é possível pois $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, e como $\hat{\boldsymbol{\beta}}$, teoricamente é o mesmo valor quando os dados estão desagregados, então basta multiplicá-lo pela matriz $\mathbf{X}$ de dados desagregados, gerando por sua vez o vetor de valores estimados $\hat{\mathbf{y}}$ de mesma dimensão. Note que para gerar o vetor de valores estimados $\hat{\mathbf{y}}$ para dados agregados, basta multiplicar $\hat{\boldsymbol{\beta}}$ pela matriz $\mathbf{X}$ de dados agregados.

Desta forma, é possível realizar a comparação ponto a ponto das estimativas geradas pela RGP-APP com a as estimativas geradas pela RGP aplicada a dados desagregados. As Figuras 4.23 e 4.24 apresentam as distribuições dos parâmetros estimados pela RGP-APP para dados agregados por setores censitários e pela RGP desagregada, respectivamente.

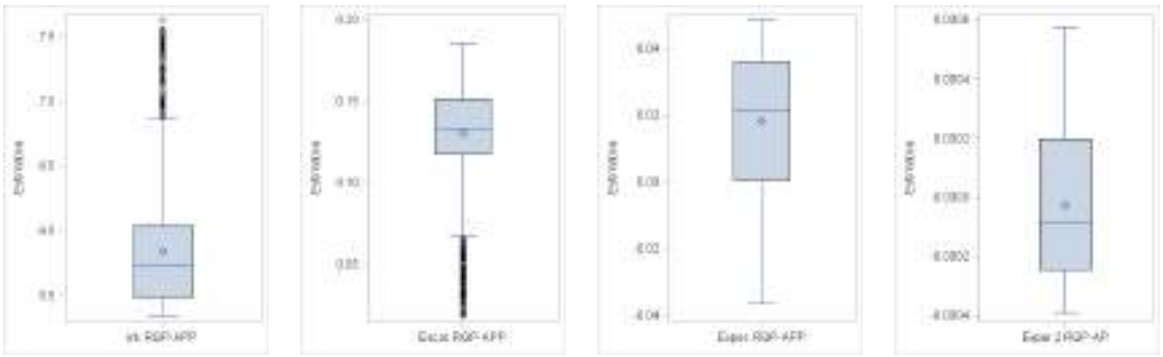


Figura 4.23: Distribuição das estimativas - RGP-APP agregada por setores censitários

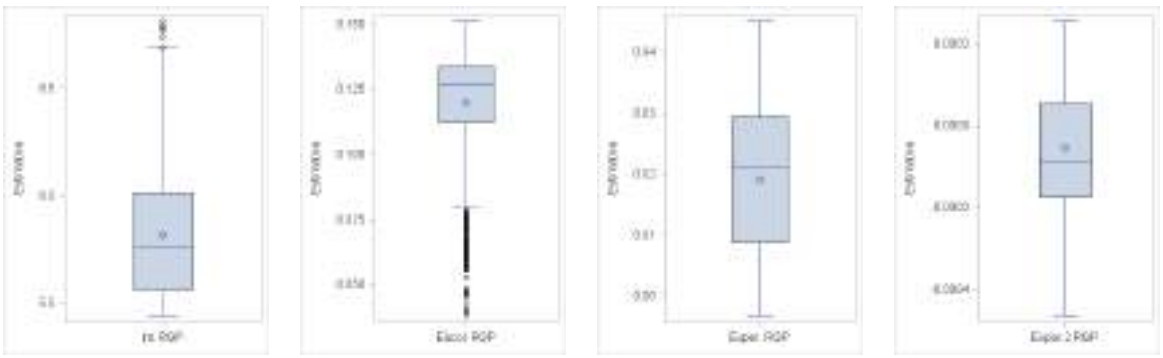


Figura 4.24: Distribuição das estimativas - RGP dados desagregados

Verifica-se que as estimativas produzidas pela RGP-APP apresentam distribuição semelhante às das estimativas produzidas pela RGP desagregada. Este é um resultado importante da RGP-APP, dado que somente ela produz estimativas a nível desagregado a partir de dados agregados. Para uma melhor avaliação desse resultado, a Figura 4.25 apresenta a distribuição das diferenças ponto a ponto entre as estimativas produzidas pelos modelos RGP desagregado e RGP-APP.

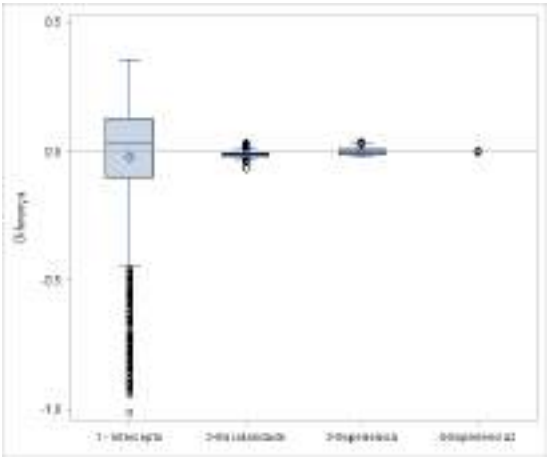


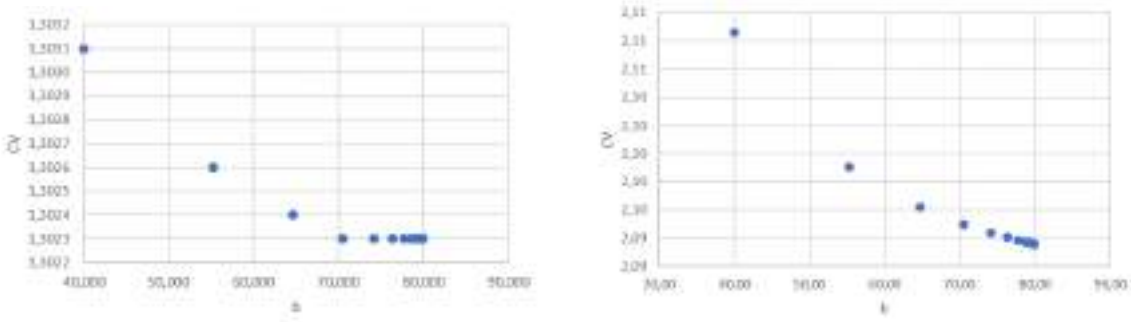
Figura 4.25: Distribuição das diferenças ponto a ponto entre as estimativas geradas pela RGP desagregada e RGP-APP

Nota-se que as diferenças estão distribuídas próximo de zero, sendo o intercepto o que apresenta maior variação, como já era esperado, conforme o que foi apontado na Seção 4.1.3. Esse resultado indica a boa capacidade da RGP-APP em estimar os parâmetros a nível desagregado para esse conjunto de dados. Na próxima Seção, será realizado mais um nível de agregação, onde a RGP-APP será mais uma vez avaliada em sua capacidade de atenuar os efeitos do MAUP.

4.2.4 Análise dos dados agregados a nível de setores de Regiões Administrativas

A setorização interna das Regiões Administrativa é uma das subdivisões territoriais utilizadas no Distrito Federal para fins de planejamento e formulação de políticas públicas. Para a realização da análise, os setores de RAs foram definidos como agregações de setores censitários. Nesta Seção, repetem-se os procedimentos da Seção anterior, agregando os dados a nível da setorização das RAs.

A Figura 4.26 ilustra o processo de obtenção dos parâmetros de suavização para o modelo RGP agregado em setores de RA e para o modelo RGP-APP. A princípio, buscou-se utilizar o mesmo procedimento empregado para a determinação do parâmetro de suavização ilustrado pela Figura 4.14. No entanto, mesmo utilizando um valor superior à distância máxima observada no conjunto de dados, não foi definida uma forma convexa da curva de forma que se pudesse identificar o mínimo da função. Assim, foram testados outros valores acima da distância máxima, adicionando intervalos de 20 quilômetros, até que se encontrasse um ponto em que a função estabilizasse. Na Figura 4.26 a), o valor para o CV estabilizou a partir de 70 km. A Figura 4.26 b) indica que, mesmo com um valor alto, a função não estabilizou. Dessa forma, optou-se por definir os parâmetros de suavização em $b = 80\,km$ para a RGP agregada e $b = 80\,km$ para a RGP-APP. Com isso, o raio de inclusão passará a considerar todas as observações do conjunto de dados para o cálculo das estimativas dos parâmetros, fazendo com que todas as estimativas tenham valores muito próximos às estimativas globais geradas pelo modelo OLS. Essa situação indica a perda de dependência espacial nesse nível de agregação.



a) b)

Figura 4.26: Parâmetro de suavização que minimiza o CV: a) RGP-APP b) RGP

A Tabela 4.14 apresenta as estimativas obtidas pelos modelos OLS e RGP, obtidas com dados agregados por setores de RAs. A comparação desses resultados com os da Tabela 4.12 torna evidente uma situação caracterizada pelo MAUP, onde as estimativas produzidas pelo modelo OLS e RGP sofrem o impacto da agregação dos dados. O intercepto foi o parâmetro que apresentou maior sensibilidade à agregação, variando de $5,82$, para $0,58$, quando utilizado o modelo RGP e de $5,74$ para $0,58$ quando o

modelo OLS foi empregado. Os demais parâmetros também apresentaram variações, porém um pouco mais discretas. Já as estimativas obtidas pelo modelo RGP-APP para o Intercepto e para o coeficiente da variável Experiência se apresentaram mais próximas dos valores reais que as produzidas pelos outros modelos. No entanto, o coeficiente da variável Escolaridade produzido pela RGP-APP foi o que ficou mais distante do valor real.

Tabela 4.14: Estimativas dos parâmetros pelos modelos OLS, RGP e RGP-APP - dados agregados por setores de regiões administrativas

Método	Estatística	Intercepto	Escolaridade	Experiência	Experiência^2
OLS	Estimativa	0,58	0,17	0,3227	-0,004536
	Erro Padrão	3,05	0,03	0,1988	0,003181
	Valor t	0,19	6,61	1,6232	-1,425838
	Pr>|t|	0,85	0,00	0,1195	0,168612
RGP AGREGADA	Média	0,58	0,17	0,3227	-0,004536
	Mínimo	0,58	0,17	0,3227	-0,004536
	P25	0,58	0,17	0,3227	-0,004536
	P50	0,58	0,17	0,3227	-0,004536
	P75	0,58	0,17	0,3227	-0,004536
	Máximo	0,58	0,17	0,3227	-0,004536
RGP-APP	Média	5,92	0,21	-0,0335	0,0007431
	Mínimo	5,92	0,21	-0,0335	0,0007431
	P25	5,92	0,21	-0,0335	0,0007431
	P50	5,92	0,21	-0,0335	0,0007431
	P75	5,92	0,21	-0,0335	0,0007431
	Máximo	5,92	0,21	-0,0335	0,0007431
RGP DESAGREGADA	Média	5,82	0,12	0,0191	-0,000053
	Mínimo	5,45	0,04	-0,0032	-0,000463
	P25	5,56	0,11	0,0087	-0,000173
	P50	5,77	0,13	0,0212	-0,000089
	P75	6,02	0,13	0,0294	0,000055
	Máximo	6,81	0,15	0,0450	0,000255

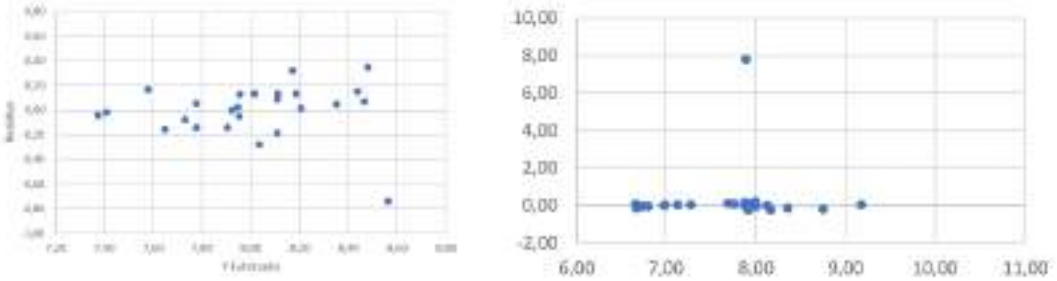
Enquanto a variação estimada pelo modelo RGP aplicado aos dados desagregados estimou retorno entre 0,04% e 0,15% a cada ano de escolaridade adicionado ao responsável pelo domicílio, o modelo RGP-APP estimou esse retorno em 0,21% a cada ano de estudo adicionado, sem variação espacial devido a perda de dependência espacial mencionada anteriormente.

A Figura 4.27 mostra a distribuição espacial do intercepto produzida pelo modelo RGP-APP. Pela falta de dependência espacial nos resultados produzidos nesse nível

de agregação, os resultados apresentam distribuição espacial uniforme, dessa forma a apresentação espacial dos resultados ficará restrita a essa figura.

Figura 4.27: Distribuição espacial das estimativas para o intercepto - RGP-APP

A Figura 4.28 a) apresenta a distribuição dos resíduos produzidos pelos modelos RGP. Nota-se que os resíduos produzidos por esse modelo apresentam uma distribuição aleatória dos valores dos resíduos em torno de zero. Para o modelo RGP-APP, com exceção do resíduo gerado por um *outlier*, os demais apresentam distribuição em torno de zero.



a) b)

Figura 4.28: Distribuição dos resíduos: a) RGP e b) RGP-APP

Nota-se que para este nível de agregação da variável dependente, a RGP-APP sofre um efeito mais intenso do MAUP do que na agregação a nível de setor censitário.

A Figura 4.29 apresenta a distribuição das diferenças calculadas ponto a ponto entre as estimativas produzidas pela RGP-APP e a RGP com dados desagregados. Os resultados mostram que as estimativas se distanciaram mais da RGP desagregada do que no nível de agregação por setores censitários. Mais uma vez as estimativas para o intercepto foram as que sofreram o maior impacto com a agregação dos dados.

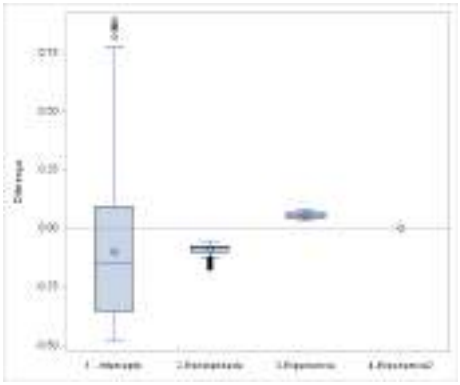


Figura 4.29: Distribuição das Estimativas - RGP dados desagregados

De forma geral, nos dois níveis de agregação avaliados, o modelo RGP-APP se mostrou mais robusto que os modelos OLS e RGP para dados agregados. Porém seu desempenho foi melhor em um menor nível de agregação, o de setores censitários. O modelo RGP-APP apresenta, ainda, a vantagem de produzir dados ao nível de domicílios, o que possibilita uma análise mais qualificada dos dados, dando maior destaque às peculiaridades locais.

Capítulo 5

Conclusão

Nesta Dissertação foram analisados a origem e os efeitos do MAUP e buscou-se a implementação de um método derivado da Regressão Geograficamente Ponderada (RGP), conhecido como Regressão Geograficamente Ponderada Área para Ponto - RGP-APP, para a mitigação de seus efeitos. O MAUP é caracterizado por situações em que a agregação espacial de unidades de dados influencia os resultados finais, o que pode dar margem a análises incorretas sobre determinado fenômeno. Os seus efeitos podem ser decompostos em dois componentes: efeitos de zoneamento e efeitos de escala. Enquanto o efeito de zoneamento refere-se a variabilidade introduzida por diferentes configurações de zoneamento no mesmo nível de agregação, o efeito da escala descreve a ocorrência de variações de resultados estatísticos usando dados agregados em diferentes níveis.

Os estudos mais antigos sobre o MAUP datam de 1934 com Gehlke e Biehl (1934) e várias propostas para a mitigação de seus efeitos foram apresentadas desde então, porém sem nenhuma solução definitiva. Mais recentemente com a introdução da Regressão Geograficamente Ponderada (Brunsdon et al., 1996), uma nova abordagem foi dada ao problema. Como uma fonte importante do efeito escala é a heterogeneidade espacial e a RGP pode modelar a variabilidade local, acredita-se que ela seja menos sensível aos efeitos do MAUP que outros modelos de regressão espacial globais. No entanto, a RGP apresenta a limitação de não incorporar mecanismo de agregação de dados em sua estrutura.

Neste trabalho foi implementado o algoritmo do modelo RGP-APP, proposto por

Murakami e Tsutsumi (2015), no *software* SAS 9.4 e realizada a avaliação desse modelo quanto à sua resistência ao MAUP. A RGP-APP surge de uma abordagem inspirada nos modelos de Geoestatística, em especial a Krigagem Área Para Ponto incorporando em sua estrutura mecanismos de agregação que permitem a estimação de parâmetros a nível dos dados desagregados, a partir de dados agregados.

Por meio de análise de dados simulados, análise de dados reais e demonstrações matemáticas, a RGP-APP foi avaliada em sua capacidade de resistência ao MAUP. Dado que surge de uma abordagem inspirada na Krigagem, que é uma forma generalizada de regressão linear simples, em um primeiro momento a RGP-APP foi avaliada como um modelo sem a presença de covariáveis. Como resultado, a RGP-APP mostrou-se eficiente em estimar a média geral da variável dependente a partir dos dados agregados em qualquer nível e configuração de agregação. Ou seja, para modelos que contenham somente o intercepto como parâmetro, a RGP-APP elimina todos os efeitos do MAUP.

Numa segunda etapa, avaliou-se o modelo RGP-APP considerando a presença de covariáveis. Os resultados mostraram que a RGP-APP não apresenta a mesma resistência demonstrada para o modelo sem covariáveis. No Capítulo 4 foi demonstrado que a igualdade de (2.25) em relação a (2.26), citada por Murakami e Tsutsumi (2015), é verdadeira somente para o modelo sem covariáveis. Nesse ensaio verificou-se um melhor desempenho da RGP-APP quando lida com agregações em unidades de diferentes tamanhos.

No terceiro ensaio, a RGP-APP foi avaliada considerando um modelo com duas covariáveis e a utilização do parâmetro de suavização ótimo, sendo o mesmo otimizado pelo algoritmo *Golden Section Search*. Os resultados mostraram que a RGP-APP apresenta melhor desempenho em situações em que são utilizados níveis menores de agregação (ou seja, agregações que consideram uma quantidade menor de unidades agregadas) e quando são utilizadas agregações que considerem a mesma quantidade de unidades agregadas. Verificou-se também que a RGP-APP é mais resistente aos efeitos do MAUP que os modelos OLS e RGP aplicados a dados agregados.

No estudo de caso, as estimativas produzidas pela RGP-APP em diferentes níveis de agregação foram comparadas com as estimativas produzidas pela RGP aplicada a dados desagregados. Nesse estudo a RGP-APP se mostrou mais resistente aos efeitos

do MAUP do que a regressão OLS e a RGP aplicadas a dados agregados. Nessa etapa, a RGP confirmou sua capacidade de mitigar os efeitos do MAUP frente a aplicação de outros modelos, com a vantagem de produzir estimativas a nível desagregado.

Dessa forma, concluí-se que a RGP-APP tem capacidade limitada para a eliminação dos efeitos do MAUP, sendo isso possível apenas em modelos sem covariáveis. Mesmo não eliminando os efeitos do MAUP em modelos com covariáveis, a RGP-APP apresenta capacidade de mitigar os efeitos da agregação, e tem resultados satisfatórios quando comparados com resultados produzidos por modelos OLS e RGP aplicados a dados agregados.

Ainda, como demonstrado no Capítulo 2, a RGP-APP apresenta a vantagem de produzir estimativas para os parâmetros a nível desagregado utilizando dados agregados. Dessa forma, a RGP-APP torna viável a estimação da variável dependente a nível desagregado, a partir do uso de covariáveis desagregadas e de uma matriz de agregação. No entanto, recomenda-se que o pesquisador que desejar aplicar a RGP-APP esteja atento aos seguintes aspectos:

1. A obtenção do parâmetro de suavização ótimo envolve um processo iterativo que pode demandar um tempo considerável de processamento, fazendo com que a estimação dos parâmetros da RGP-APP seja mais demorada do que a do modelo OLS ou de outros modelos de regressão espacial. O uso de computadores com alta capacidade de processamento ou de outras tecnologias, como computação em nuvem, pode minimizar a diferença de tempo demandada pelos métodos;

2. Em situações em que existe dependência espacial, a utilização de modelos de regressão clássica (OLS) ou de outros modelos de regressão espacial, em detrimento da RGP-APP, pode produzir resultados viesados e que não destacam as peculiaridades locais. Porém, em situações onde não é constatada a dependência espacial e não há o interesse por inferências a nível de indivíduo, o uso desses tipos de modelos pode trazer economia de tempo e resultados que atendem à demanda do pesquisador.

5.1 Limitações do Trabalho

Como foi visto, o trabalho buscou analisar os efeitos do MAUP pela Regressão Geograficamente Ponderada, e pela sua versão mais atual conhecida como RGP-APP. Apesar da metodologia proposta ter se mostrado adequada, o trabalho ainda conta com algumas limitações:

- Buscou-se inicialmente trabalhar com variáveis categóricas, mas não foi decidido ao certo qual estatística utilizar para a agregação: média ou soma, uma vez que para ambas os resultados foram muito distintos;
- Poderia ser utilizado também *shapes* irregulares para as simulações, a fim de se utilizar uma estrutura de dados mais realística;
- A adição de mais um nível de agregação com uma quantidade maior de unidades seria importante para a avaliação da RGP-APP, no entanto, devido ao tempo de processamento (em média 50 min para cada repetição), optou-se por trabalhar com apenas dois níveis.

5.2 Recomendações para Trabalhos Futuros

A RGP-APP foi desenvolvida recentemente, no ano de 2015, e naturalmente ainda há muito o que explorar sobre suas características e propriedades. A seguir são listadas alguns tópicos para estudos futuros:

- Avaliar a incorporação de mecanismos de agregação para variáveis binárias ou categóricas. Neste trabalho, as variáveis utilizadas eram quantitativas, discretas ou contínuas e o estudo do desenvolvimento de novos mecanismos de agregação para variáveis categóricas é de grande importância para a utilização do método RGP-APP;
- Avaliar a significância dos parâmetros estimados;
- Conduzir mais estudos de simulação controlando a correlação entre as covariáveis, a fim de verificação a robustez do modelo RGP-APP à presença de multicolinearidade;

- Incorporar algum fator de ponderação ao modelo para que o produto das covariáveis e a variável dependente, e o quadrado das covariáveis para os dados agregados sejam os mesmos dos dados desagregados, eliminando dessa forma o MAUP.

Referências Bibliográficas

Abdel-Aty, M., Uddin, e N., P. (2004). Predicting freeway crashes based on loop detector data using matched case-control logistic regression. *Transportation Research*, 1897(12):88–95.

Amrhein, C. e Reynolds, H. (1996). Using spatial statistics to assess aggregation effects. *Journal of Geographical Systems*, 2:143–158.

Amrhein, C. G. (1995). Searching for the elusive aggregation effect: evidence from statistical simulations. *Environment and planning A*, 27(1):105–119.

Amrhein, C. G. e Flowerdew, R. (1992). The effect of data aggregation on a poisson regression model of canadian migration. *Environment and Planning A*, 24(10):1381–1391.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

Anselin, L. (1995). Hierarchical bayesian space-time models. *Kluwer Academic Publishers*, 27(2):1538–4632.

Anselin, L. (2000). The alchemy of statistics, or creating data where no data exist. *Annals of the Association of American Geographers*, 90(3):586–592.

Arbia, G. (1989). *Statistical effects of spatial data transformations: a proposed general framework*. Taylor and Francis.

Assunção, R. M. (2004). Índices de auto-correlação espacial. *Belo Horizonte: UFMG, Departamento de estatística*.

Atkinson, P. M. e Curran, P. J. (1995). Defining an optimal size of support for remote sensing investigations. *IEEE Transactions on Geoscience and Remote Sensing*, 33(3):768–776.

Atkinson, P. M. e Tate, N. J. (2000). Spatial scale problems and geostatistical solutions: a review. *The Professional Geographer*, 52(4):607–623.

Batty, M. e Sikdar, P. (1982). Spatial aggregation in gravity models. 1. an information-theoretic framework. *Environment and Planning A*, 14(3):377–405.

Bian, L. (1997). Multiscale nature of spatial data in scaling up environmental models. In: *Scale in remote sensing and GIS*, D. A. Quattrochi e M. F. Goodchild, ed., pages 13–26. Lewis Publishers: Boca Raton, FL, USA.

Bian, L. e Walsh, S. J. (1993). Scale dependencies of vegetation and topography in a mountainous environment of montana. *The Professional Geographer*, 45(1):1–11.

Blair, P. e Miller, R. E. (1983). Spatial aggregation in multiregional input-output models. *Environment and Planning A*, 15(2):187–206.

Brunsdon, C., Fotheringham, A. S., e Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, 28(1538-4632):281–298.

Burnham, K. P. e Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Camargo, E. C. G., Fucks, S. D., e Câmara, G. (2004). Análise espacial de superfícies. *Análise espacial de dados geográficos. Planaltina: Embrapa Cerrados*, pages 79–122.

Charlton, M., Fotheringham, S., e Brunsdon, C. (2009). Geographically weighted regression. *White paper. National Centre for Geocomputation. National University of Ireland Maynooth*.

Clark, W. A. e Avery, K. L. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4):428–438.

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

Câmara, G., Monteiro, A. M., Carvalho, M. S., e Druck, S. (2002). *Análise Espacial de dados Geográficos*. DPI - INPE.

Cressie, N., Wikle, C. K., e Berliner, L. M. (1998). Local indicators of spatial association—lisa. *Geographical Analysis*, 5:117–154.

Davis, G. A. (2004). Possible aggregation biases in road safety research and a mechanism approach to accident modeling. *Accident Analysis & Prevention*, 36(6):1119–1127.

Diez-Roux, A. (1998). Bring context back into epidemiology: variables and fallacies in multilevel analysis. *American Journal of Public Health*, 88(2):216–222.

Espa, G., Arbia, G., e Benedetti, R. (1996). Effects of the maup on image classification. *Geographical Systems*, 3(2-3):123–141.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2(2):143–154.

Fisher, P. F. e Langford, M. (1995). Modelling the errors in areal interpolation between zonal systems by monte carlo simulation. *Environment and planning A*, 27(2):211–224.

Fisher, P. F. e Langford, M. (1996). Modeling sensitivity to accuracy in classified imagery: A study of areal interpolation by dasymetric mapping. *The Professional Geographer*, 48(3):299–309.

Flowerdew, R. e Green, M. (1989). Statistical methods for inference between incompatible zonal systems. *Accuracy of spatial databases*, pages 239–247.

Fotheringham, A., Brunsdon, C., e Charlton, M. (2002). Geographically weighted regression: the analysis of spatially varying relationship. *Chichester*.

Fotheringham, A. e Wong, D. (1991). The modifiable areal unit problem in multivariate statistical analysis. In: *Environment and Planning A*, S. A. Fotheringham e P. A. Rogerson, ed., pages 1025–1044. SAGE.

Fotheringham, A. S. (1989). Scale-independent spatial analysis. In: *Accuracy of spatial databases*. Taylor and Francis London.

Fotheringham, A. S., Brunsdon, C., e Charlton, M. (2000). *Quantitative geography: perspectives on spatial data analysis*. Sage.

Fotheringham, A. S., Densham, P. J., e Curtis, A. (1995). The zone definition problem in location-allocation modeling. *Geographical Analysis*, 27(1):60–77.

Gehlke, C. E. e Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29(185A):169–170.

Gelfand, A., Kim, H., e Sirmans, C. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of American Statistical Association*, 98(462):387–396.

Goodchild, M. F. (1979). The aggregation problem in location-allocation. *Geographical Analysis*, 11(3):240–255.

Gotway, C. A. e Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partition (redcap). *International Journal of Geographical Information Science*, 22(7):801–823.

Hadayeghi, A., Shalaby, A., e Persaud, B. (2010). Development of planning level transportation safety tools using geographically weighted poisson regression. *Accident Analysis and Prevention*, 42(2):676–688.

Haining, R. P., Kerry, R., e Oliver, M. A. (2002). Geography, spatial data analysis, and geostatistics: An overview. *Geographical Analysis*, 42:7–31.

Holt, D., Steel, D., e Tranmer, M. (1996). Area homogeneity and the modifiable areal unit problem. *Geographical Systems*, 3(2/3):181–200.

Horner, M. W. e Murray, A. T. (2002). Excess commuting and the modifiable areal unit problem. *Urban Studies*, 39(1):131–139.

Jelinski, D. E. e Wu, J. (1996). The modifiable areal unit problem and implications for landscape ecology. *Landscape ecology*, 11(3):129–140.

Kennedy, V., Jones, A., e Haynes, R. (2007). District variations in road curvature in england and wales and their association with road-traffic crashes. *Environment and Planning*, 26(1):1222–1237.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton, NJ: Princeton University Press.

Kyriakidis, P. C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, 36(3):259–289.

Lam, N. (1983). Spatial interpolation methods: A review. *Cartography and Geographic Information Science*, 10(2):129–150.

Lam, N. S.-N. e Quattrochi, D. A. (1992). On the issues of scale, resolution, and fractal analysis in the mapping sciences. *The Professional Geographer*, 44(1):88–98.

LeSage, J. P. (1999). The theory and practice of spatial econometrics. *University of Toledo. Toledo, Ohio*, 28:33.

Loader, C. (1999). *MLocal Regression and Likelihood*. Springer.

Martin, D. (2003). Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science*, 17(2):181–196.

Mincer, J. (1975). Education, experience, and the distribution of earnings and employment: an overview. In: *Education, income, and human behavior*, pages 71–94. NBER.

Moellering, H. e Tobler, W. (1972). Geographical variances. *Geographical Analysis*, 4(1):34–50.

Murakami, D. e Tsutsumi, M. (2015). Area-to-point parameter estimation with geographically weighted regression. *Journal of Geographical Systems*, 17(1):207–225.

Okabe, A. (1996). Spatial aggregation bias in a regression model containing a distance variable. *Geographical Systems*, 3(2):77–99.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. *Transactions of the institute of british geographers*, 2(4):459–472.

Openshaw, S. (1984). *The Modifiable Areal Unit Problem*. Geo Books, Norwich.

Openshaw, S. e Schmidt, J. (1996). Parallel simulated annealing and genetic algorithms for re-engineering zoning systems. *Geographical Systems*, 3(4):201–220.

Openshaw, S. e Taylor, P. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. *Statistical Applications in the Spatial Sciences*, 26(1):127–144.

Parenteau, M. e Sawada, M. C., C. (2011). The modifiable areal unit problem (maup) in the relationship between exposure to no2 and respiratory health. *International Journal of Health Geographics*, 10(58):1–15.

Perle, E. D. (1977). Scale changes and impacts on factorial ecology structures. *Environment and Planning A*, 9(5):549–558.

Páez, A., Long, F., e Farber, S. (2008). Moving window approaches for hedonic price estimation: an empirical comparison of modelling techniques. *Urban Studies*, 45(8):1565–1581.

Putman, S. e Chung, S. (1989). Effects of spatial system design on spatial interaction models. 1: The spatial system definition problem. *Environment and Planning A*, 21(1):27–46.

Quattrochi, D. A. e Goodchild, M. F. (1997). *Scale in remote sensing and GIS*. CRC press.

Robinson, A. H. (1956). The necessity of weighting values in correlation analysis of areal data. *Annals of the Association of American Geographers*, 46(2):233–236.

Sawicki, D. S. (1973). Studies of aggregated areal data: problems of statistical inference. *Land Economics*, 49(1):109–114.

Siffel, C., Strickland, M. J., Gardner, B. R., Kirby, R. S., e Correa, A. (2006). Role of geographic information systems in birth defects surveillance and research. *Birth Defects Research Part A: Clinical and Molecular Teratology*, 76(1):825–833.

Silva, A. d. (2006). *Avaliação de modelos de regressão espacial para análise de cenários do transporte rodoviário de carga*. PhD thesis, Dissertação de Mestrado. Faculdade de Tecnologia, Universidade de Brasília, DF, 122p.

Steel, D. e Holt, D. (1996). Rules for random aggregation. *Environment and Planning A*, 28(6):957–978.

Swift, A., Liu, L., e Uber, J. (2008). Reducing maup bias of correlation statistics between water quality and gi illness. *Computers, Environment and Urban System*, pages 134–148.

Tate, N. e Atkinson, P. M. (2001). *Modelling scale in geographical information science*. John Wiley & Sons.

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74(367):519–530.

Tobler, W. R. (1989). Frame independent spatial analysis. In: *Accuracy of spatial databases*, M. Goodchild e G. S., ed. Taylor and Francis London.

Townshend, J. R. e Justice, C. O. (1988). Selecting the spatial resolution of satellite sensors required for global monitoring of land transformations. *International Journal of Remote Sensing*, 9(2):187–236.

Tranmer, M. e Steel, D. G. (1998). Using census data to investigate the causes of the ecological fallacy. *SAGE Journals*, 30:817–831.

Viegas, J. (2009). Effects of the modifiable areal unit problem on the delineation of traffic analysis zones. *Environment and Planning B: Planning and Design*, 36(4):625–643.

Ávila, R. e Monasterio, L. (2008). o maup e a análise espacial: um estudo de caso para o rio grande do sul (1991-2000). *Revista Análise Econômica*, 26(49):233–259.

Wei, B. C. e Chai, W. Y. (2004). A multiobjective hybrid metaheuristic approach for gis-based spatial zoning model. *Journal of Mathematical Modelling and Algorithms*, 3(3):245–261.

Wheeler, D. C. (2007). Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A*, 39(10):2464–2481.

Wong, D. W. (1996). Aggregation effects in georeferenced data. In: *Practical handbook of spatial statistics*, S. Arlinghaus, ed. CRC Press.

Wong, D. W. (2001). Location-specific cumulative distribution function (lscdf): An alternative to spatial correlation analysis. *Geographical Analysis*, 33(1):76–93.

Wong, D. W. (2004). Comparing traditional and spatial segregation measures: A spatial scale perspective1. *Urban Geography*, 25(1):66–82.

Wong, D. W. S. (2009). The modifiable areal unit problem (maup). In: *The SAGE Handbook of Spatial Analysis*, S. A. Fotheringham e P. A. Rogerson, ed., pages 105–124. SAGE.

Wong, W. e Amrhein, C. (1996). Research on the maup: Old wine in a new bottle or real breakthrough? *Journal of Geographical Systems*, 3:73–76.

Xu, P., Huang, H., e Dong, N. (2015). The modifiable areal unit problem in traffic safety: Basic issue, potential solutions and future research. *Journal of Traffic and Transportation Engineering*, 26(1):127–144.

Yannis, G., Papadimitriou, E., e Aotoniou, C. (2007). Multilevel modeling for the regional effect of enforcement on road accidents. *Accident Analysis and Prevention*, 39(4):818–825.

Young, L. J. e Gotway, C. A. (2007). Linking spatial data from different sources: the effects of change of support. *Stochastic Environmental Research and Risk Assessment*, 21(5):589–600.