



GEAN PEREIRA DAMACENO

**MODELOS DE REGRESSÃO PARA DADOS DE ÁREAS COM
VALORES EXTREMOS NO ESTUDO DA INCIDÊNCIA DE
DENGUE EM MINAS GERAIS**

LAVRAS – MG

2025

GEAN PEREIRA DAMACENO

**MODELOS DE REGRESSÃO PARA DADOS DE ÁREAS COM VALORES
EXTREMOS NO ESTUDO DA INCIDÊNCIA DE DENGUE EM MINAS GERAIS**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

Prof. DSc. João Domingos Scalon
Orientador

Prof. DSc. Douglas Mateus da Silva
Coorientador

**LAVRAS – MG
2025**

**Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos da Biblioteca
Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Damaceno, Gean Pereira

Modelos de regressão para dados de áreas com valores extremos no estudo da incidência de dengue em minas gerais / Gean Pereira Damaceno. 6^a ed. rev., atual. e ampl. – Lavras : UFLA, 2025.

116p. : il.

Dissertação(mestrado)–Universidade Federal de Lavras, 2025.

Orientador: Prof. DSc. João Domingos Scalon.

Coorientador: Prof. DSc. Douglas Mateus da Silva .

Bibliografia.

1. Modelos espaciais robustos. 2. Dados de áreas. 3. Dengue. 4. Saúde pública. 5. Dependência espacial. I. Universidade Federal de Lavras. II. Título.

CDD 808.066

GEAN PEREIRA DAMACENO

**MODELOS DE REGRESSÃO PARA DADOS DE ÁREAS COM VALORES
EXTREMOS NO ESTUDO DA INCIDÊNCIA DE DENGUE EM MINAS GERAIS**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

APROVADA em 30 de setembro de 2025.

Prof. DSc. RENATO RIBEIRO DE LIMA	Universidade Federal de Lavras (UFLA)
Prof. DSc. LOURENÇO MANUEL	Universidade Eduardo Modlane (UEM)
Prof. DSc. CARLOS PEREIRA DA SILVA	Universidade Federal de Lavras (UFLA)
Prof. DSc. DEIVE CIRO DE OLIVEIRA	Universidade Federal de Alfenas (UNIFAL-MG)

Documento assinado digitalmente
 **JOAO DOMINGOS SCALON**
Data: 05/01/2026 16:16:30-0300
Verifique em <https://validar.it.gov.br>

Prof. DSc. João Domingos Scalon
Orientador

Prof. DSc. Douglas Mateus da Silva
Co-Orientador

**LAVRAS – MG
2025**

Dedico este trabalho à minha mãe, Maria das Dores Pereira Damaceno, minha avó, Rita Pereira Damaceno, e minha filha, Sarah Silva Damaceno.

Aos meus irmãos, Fernando Pereira Damaceno, Reisiane Pereira Damaceno e Wilberson Pereira Damaceno, que estão sempre ao meu lado nos momentos mais difíceis e demonstrando apoio.

AGRADECIMENTOS

Agradeço a Deus por concluir mais uma etapa, com muita saúde e com o objetivo de sempre estar apto a desenvolver e aprender novas habilidades.

Agradeço à minha família por sempre estar me apoiando nas tomadas de decisão que faço.

Agradeço em especial ao meu ex-orientador e professor do Instituto Federal do Piauí, Dr. Ernandes Moura Guedes.

Agradeço ao meu ex-professor do Instituto Federal do Piauí, Gabriel Pinto.

Agradeço aos meus amigos que Deus os colocou em minha vida: Giovani Festa Paludo, Carlos Silva, Júlia Peixoto, Iulli Jahel, Diana Rebaza, Valdeline Mequelino, Alex Nhancololo, Rafael Genaro, Melissa Wendy, Louiziane Carvalho, Lucas Barbosa, Elias sabe, Osvaldo, Mateus Peixoto, Ramon Ferreira da Silva, Viviane Costa, Camila Giovana, Dheyson Sandes, Lorrann Carreiro, Yasmin Santana, João Vitor, João Bonifácio, Geovane Fernandes, Pedro Mambelli, Lara Gualberto, Raul Rezende, Tiago Brito, Ali William Canaza, Roxana Churata e Gabriel Edson.

Agradeço aos professores do Departamento de Estatística, principalmente, ao professor João Domingos Scalón e ao professor Douglas Mateus da Silva, que estão me orientando e me apoiando.

Agradeço a agência de fomento CAPES pelo financiamento da bolsa estudantil.

Agradeço à Universidade Federal de Lavras (UFLA) que me recebeu de portas abertas, ofertando e oferecendo apoio em momentos extras.

RESUMO

A dengue tem se tornado uma ameaça crescente nos últimos anos, apresentando surtos frequentes e sobrecarregando os sistemas de saúde. A propagação do vírus, transmitido pelo *Aedes aegypti*, está fortemente relacionada à fatores ambientais e urbanos, tornando essencial a utilização de métodos estatísticos para entender sua disseminação e implementar ações mais eficazes para combatê-lo. Modelos tradicionais de regressão linear, embora amplamente usados, assumem a independência das observações e não consideram a autocorrelação espacial que, quando presente, pode comprometer a precisão das estimativas em áreas geograficamente interdependentes. Para superar essa limitação, modelos espaciais como a Regressão Simultânea Autorregressiva (SAR) e a Regressão Condicional Autorregressiva (CAR) surgiram como soluções. Esses modelos são capazes de capturar os padrões espaciais que, normalmente, estão presentes em dados de áreas. Contudo, esses modelos são sensíveis à valores atípicos (ou *outliers*) que, em geral, prejudicam a qualidade das análises. A Regressão Geograficamente Ponderada (GWR), que incorpora a variação espacial por meio da estimação local dos coeficientes, foi proposta como alternativa mas, também, pode ser influenciada por valores atípicos. A Regressão Geograficamente Ponderada Robusta (RGWR) foi proposta como uma solução aprimorada que pode proporcionar maior robustez contra valores atípicos e, consequentemente, levar a estimativas mais eficazes. Este estudo aplica a RGWR para investigar a incidência de dengue no estado de Minas Gerais, no ano de 2023, com o objetivo de melhorar a precisão das estimativas frente a presença de *outliers* que é muito comum nesse tipo de dados. As variáveis utilizadas no estudo foram as seguintes: incidência de dengue como variável dependente e esgotamento sanitário adequado, área urbanizada, urbanização de vias públicas e arborização de vias públicas como variáveis independentes. Os resultados mostraram que a RGWR pode proporcionar uma abordagem mais precisa para entender a distribuição espacial da dengue em dados de áreas com a presença de *outliers* o que pode contribuir para uma gestão mais eficiente e eficaz das ações de saúde pública.

Palavras-chave: Dados de área. Dengue. Modelos Robustos. Outliers. Estatística espacial.

ABSTRACT

Dengue fever has become a growing threat in recent years, with frequent outbreaks and overwhelming healthcare systems. The spread of the virus, transmitted by *Aedes aegypti*, is strongly related to environmental and urban factors, making the use of statistical methods essential to understand its dissemination and implement more effective actions to combat it. Traditional linear regression models, although widely used, assume the independence of observations and do not consider spatial autocorrelation, which, when present, can compromise the accuracy of estimates in geographically interdependent areas. To overcome this limitation, spatial models such as Simultaneous Autoregressive Regression (SAR) and Conditional Autoregressive Regression (CAR) have emerged as solutions. These models are able to capture spatial patterns that are normally present in area data. However, these models are sensitive to outliers, which generally impair the quality of the analyses. Geographically Weighted Regression (GWR), which incorporates spatial variation through local estimation of coefficients, has been proposed as an alternative but can also be influenced by outliers. Robust Geographically Weighted Regression (RGWR) has been proposed as an improved solution that can provide greater robustness against outliers and, consequently, lead to more effective estimates. This study applies RGWR to investigate the incidence of dengue fever in the state of Minas Gerais, in the year of 2023, aiming to improve the accuracy of estimates in the face of outliers, which are very common in this type of data. The variables used in the study were: dengue fever incidence as the dependent variable and adequate sanitation, urbanized area, street urbanization, and street tree planting as independent variables. The results showed that RGWR can provide a more accurate approach to understanding the spatial distribution of dengue in data from areas with outliers which can contribute to more efficient and effective management of public health actions.

Keywords: Area data. Dengue. Robust models. Outliers. Spatial Statistics.

INDICADORES DE IMPACTO

O presente trabalho apresenta impactos de caráter tecnológico e social, com potencial aplicação direta na gestão de saúde pública.

Em relação aos aspectos tecnológicos (item a), o estudo contribui para o aprimoramento da análise de dados epidemiológicos ao avaliar modelos de regressão espacial robustos (GWR e RGWR) capazes de lidar com *outliers* valores extremos. Isso oferece aos pesquisadores e gestores ferramentas estatísticas mais confiáveis para mapear a incidência de doenças, evitando distorções causadas por surtos localizados.

No âmbito social e de extensão (itens b e c), o trabalho possui potencial extensionista ao fornecer subsídios analíticos para a vigilância epidemiológica no território de Minas Gerais. Além disso, a metodologia proposta é replicável e pode ser estendida para todo o território nacional, permitindo a comparação de cenários regionais e o fortalecimento das políticas de combate a arboviroses em escala federal. A identificação de padrões espaciais da dengue auxilia na tomada de decisão e na alocação eficiente de recursos públicos para o combate ao vetor, beneficiando a população do estado, especialmente em áreas urbanas densas e regiões historicamente afetadas pela doença.

Quanto à classificação na Política Nacional de Extensão (item e), este trabalho se enquadra principalmente na área temática 6 - Saúde, com interface na área 7 - Tecnologia e Produção (desenvolvimento de métodos de análise).

Por fim, em alinhamento com a agenda global (item f), os resultados desta pesquisa corroboram os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especificamente o ODS 3 (Saúde e Bem-Estar), meta 3.3, que visa acabar com as epidemias de doenças tropicais negligenciadas e outras doenças transmissíveis, e o ODS 11 (Cidades e Comunidades Sustentáveis), ao promover o planejamento urbano baseado em evidências para o controle de endemias.

INDICADORES DE IMPACTO

This work presents technological and social impacts, with potential direct application in public health management.

Regarding technological aspects (item a), the study contributes to the improvement of epidemiological data analysis by evaluating robust spatial regression models (GWR and RGWR) capable of handling *outliers* (extreme values). This offers researchers and managers more reliable statistical tools to map disease incidence, avoiding distortions caused by localized outbreaks.

In the social and extension scope (items b and c), the work possesses extension potential by providing analytical support for epidemiological surveillance in the territory of Minas Gerais. Furthermore, the proposed methodology is replicable and can be extended to the entire national territory, allowing for the comparison of regional scenarios and the strengthening of policies to combat arboviruses on a federal scale. The identification of spatial patterns of dengue assists in decision-making and the efficient allocation of public resources for vector control, benefiting the state's population, especially in dense urban areas and regions historically affected by the disease.

Regarding the classification within the National Extension Policy (item e), this work falls mainly under thematic area 6 - Health, with an interface in area 7 - Technology and Production (development of analysis methods).

Finally, in alignment with the global agenda (item f), the results of this research support the UN Sustainable Development Goals (SDGs), specifically SDG 3 (Good Health and Well-being), target 3.3, which aims to end epidemics of neglected tropical diseases and other communicable diseases, and SDG 11 (Sustainable Cities and Communities), by promoting evidence-based urban planning for the control of endemic diseases.

LISTA DE FIGURAS

Figura 1.1 – ODS 3 – Saúde e Bem-Estar	21
Figura 2.1 – Mosquito da Dengue	23
Figura 2.2 – Diferenças morfológicas entre o mosquito da dengue macho e fêmea.	23
Figura 2.3 – Mosquitos <i>Aedes aegypti</i> capturados durante estudo em ambiente acadêmico.	26
Figura 2.4 – Casos de dengue no Brasil	27
Figura 2.5 – Tipos de vizinhança espacial.	30
Figura 2.6 – Representação genérica de uma matriz de pesos espaciais C de ordem $n \times n$	31
Figura 2.7 – Representação de uma área irregular (A) e sua respectiva matriz de adjacência binária W (B).	32
Figura 2.8 – Exemplo de diagrama de espalhamento de Moran	36
Figura 3.1 – Área de estudo sobre a incidência de dengue em Minas Gerais	61
Figura 3.2 – Matriz de vizinhança do tipo rainha (Queen) para os municípios de Minas Gerais.	62
Figura 4.1 – Boxplot da incidência de dengue para o ano de 2023	66
Figura 4.2 – Distribuição da variável dependente em função das covariáveis para o ano de 2023	67
Figura 4.3 – Gráfico e Histograma dos resíduos do modelo	70
Figura 4.4 – Mapa da incidência de dengue para o ano de 2023	73
Figura 4.5 – Diagrama de dispersão	74
Figura 4.6 – Índice de Moran Local LISA	75
Figura 4.7 – Municípios que são considerados outliers	76
Figura 4.8 – Gráficos dos resíduos dos modelos SAR e CAR: QQ-plots e histogramas.	78
Figura 4.9 – Resultados do procedimento de seleção de variáveis para o modelo GWR.	81
Figura 4.10 – Coeficientes de determinação locais para os modelos GWR (a) e RGWR(b).	84
Figura 4.11 – Comparação da distribuição espacial dos resíduos para os seis modelos ajustados.	91
Figura 4.12 – Coeficientes locais para Arborização (a) e Esgotamento Sanitário (b).	94
Figura 4.13 – Coeficientes locais para Áreas Urbanizadas (C) e Urbanização de Vias (D).	96
Figura 4.14 – Mapas dos erros-padrão dos coeficientes das covariáveis do estudo.	97

Figura 4.15 – Mapas de t-valor para Arborização (a), Esgotamento Sanitário (b), Área Urbanizada (c) e Urbanização de Vias (d).	99
Figura 4.16 – Mapas de p-valor para as covariáveis ambientais selecionadas.	101
Figura 4.17 – Mapas dos p-valores locais dos coeficientes estimados pelo modelo RGWR robusto para as variáveis explicativas: Arborização (e), Esgotamento Sanitário (f), Área Urbanizada (g) e Urbanização de Vias (h).	102

LISTA DE TABELAS

Tabela 4.1 – Estatísticas descritivas das variáveis	65
Tabela 4.2 – Resultados da Regressão Linear Múltipla (OLS)	68
Tabela 4.3 – Resultado do teste de normalidade nos resíduos (Shapiro-Wilk)	69
Tabela 4.4 – Assimetria e Curtose dos Resíduos do Modelo de Regressão Múltipla	69
Tabela 4.5 – Resultado do teste de homocedasticidade (Breusch-Pagan)	70
Tabela 4.6 – Resultado do teste de independência (Durbin-Watson)	71
Tabela 4.7 – Resultados do modelo robusto (MM-type)	72
Tabela 4.8 – Resultados do índice de Moran Global para a incidência de dengue para o ano de 2023	73
Tabela 4.9 – Comparação dos Modelos de Regressão Espacial: SAR (lag) e CAR (erro) .	77
Tabela 4.10 – Teste de Normalidade Shapiro-Wilk para os Resíduos dos Modelos SAR e CAR	79
Tabela 4.11 – Resumo dos resíduos dos modelos ajustados	79
Tabela 4.12 – Modelos GWR e RGWR (robusto) com AICc e AIC	82
Tabela 4.13 – Comparação dos Testes F1 de Leung et al. (2000) para GWR e RGWR . .	85
Tabela 4.14 – Teste F3 de LEUNG, MEI & ZHANG (2000) para os Modelos GWR e RGWR	85
Tabela 4.15 – Estatísticas descritivas dos resíduos dos modelos RGWR e GWR.	86
Tabela 4.16 – Resumo do VIF local para variáveis explicativas no modelo GWR	87
Tabela 4.17 – Medidas resumo dos resíduos dos modelos ajustados	88
Tabela 4.18 – Dispersão dos Resíduos (Intervalo Interquartil) por Modelo.	90
Tabela 4.19 – Comparação das métricas dos modelos ajustados	93
Tabela 4.20 – Tempos de execução e consumo de memória dos modelos estatísticos . . .	93
Tabela 4.21 – Resumo de valores significativos e não significativos por variável	103

SUMÁRIO

	GLOSSÁRIO	17
1	INTRODUÇÃO	19
1.1	Objetivos específicos	20
1.2	Objetivos de Desenvolvimento Sustentável - ODS	20
2	REFERENCIAL TEÓRICO	22
2.1	Vetor <i>Aedes aegypti</i>	22
2.2	Casos de Dengue	24
2.3	Estatística Espacial	27
2.4	Dados de área	28
2.4.1	Estrutura da dependência espacial	29
2.4.2	Matriz de contiguidade espacial	29
2.4.3	Autocorrelação espacial	32
2.5	Modelos de Regressão na análise de incidência da dengue: Regressão li- near e Modelos com Dependência Espacial	36
2.5.1	Modelos de Regressão Linear Múltipla	37
2.5.1.1	Análise das pressuposições sobre a variável aleatória erro	39
2.5.2	O modelo espacial autoregressivo (Spatial Autoregressive Model - SAR) .	40
2.5.3	Modelo de Erro Espacial (Conditional Autoregressive Model - CAR) . . .	40
2.5.4	Modelo de Regressão Múltipla Robusta	42
2.6	Regressão Geograficamente Ponderada (GWR)	49
2.6.1	O modelo GWR	50
2.6.2	Calibração do Modelo e Seleção do <i>Bandwidth</i>	50
2.6.3	Teste de Não-Estacionariedade Espacial	51
2.6.4	Regressão Geograficamente Ponderada Robusta (RGWR)	51
2.6.5	Regressão Geograficamente Ponderada (GWR)	52
2.6.6	Regressão Geograficamente Ponderada Robusta (RGWR)	54
2.7	Matriz de Pesos do Modelo RGWR	56
2.7.1	Estimação dos Parâmetros	58
2.8	Critérios para avaliação dos modelos	59
3	MATERIAL E MÉTODOS	61
3.1	Área de estudo	61

3.2	Matriz de vizinhança espacial	62
3.3	Descrição dos dados e ferramentas computacionais	63
4	RESULTADOS e DISCUSSÃO	65
4.1	Análise exploratória usual	65
4.2	Modelo de regressão linear múltipla	68
4.2.1	Ajuste do modelo	68
4.2.2	Análise da adequabilidade do ajuste	69
4.3	Modelo de regressão linear múltipla robusto	71
4.4	Análise exploratória espacial	72
4.4.1	Mapa da incidência de dengue	73
4.4.2	Diagrama de dispersão de Moran	74
4.5	Modelos autorregressivo simultâneo (SAR) e Autoregressivo Condicional (CAR)	76
4.5.1	Ajustes dos modelos SAR e CAR	77
4.5.2	Análise de resíduos dos modelos SAR e CAR	78
4.6	Comparação dos quatro modelos ajustados	79
4.7	Seleção de Variáveis para os Modelos GWR RGWR	80
4.8	Modelos GWR e RGWR	81
4.8.1	Ajuste dos Modelos GWR RGWR	82
4.8.2	Coeficiente de determinação local	83
4.8.3	Teste F1	84
4.8.4	Teste F3	85
4.8.5	Análise comparativa dos resíduos dos modelos GWR e RGWR	86
4.8.6	Verificação da Multicolinearidade	87
4.9	Comparação de todos os modelos ajustados	88
4.9.1	Estatística descritiva dos resíduos	88
4.9.2	Intervalo interquartilico dos resíduos	89
4.9.3	Mapas dos Resíduos	90
4.9.4	Métricas para seleção do modelo	92
4.10	Análise visual do modelo RGWR por meio de mapas	94
4.10.1	Mapas dos coeficientes: Arborização e esgotamento sanitário	94
4.10.2	Mapa dos coeficientes: Área urbanizada e urbanização de vias	95

4.10.3	Mapa dos erros-padrão	97
4.10.4	Mapa dos valores-t	98
4.10.5	Mapas dos valores-p	100
4.10.6	Mapas dos valores- p estatisticamente significativos	102
5	CONCLUSÃO	105
	REFERÊNCIAS	106
	APÊNDICES	111
.1	Códigos para Modelagem	111
.1.1	Script em R	111

LISTA DE ABREVIATURAS

art. Artigo

cap. Capítulo

Eq. Equação

Fig. Figura

Tab. Tabela

LISTA DE SIGLAS

AIC Critério de Informação de Akaike (*Akaike Information Criterion*)

AICc Critério de Informação de Akaike Corrigido

BIC Critério de Informação Bayesiano (*Bayesian Information Criterion*)

CAR Modelo de Erro Espacial (*Spatial error model*)

CV Validação Cruzada (*Cross-Validation*)

DATASUS Departamento de Informática do SUS

GWR Regressão Geograficamente Ponderada (*Geographically Weighted Regression*)

IBGE Instituto Brasileiro de Geografia e Estatística

IWLS Mínimos Quadrados Re-ponderados Iterativamente (*Iteratively Reweighted Least Squares*)

MG Minas Gerais

MQO Mínimos Quadrados Ordinários (*Ordinary Least Squares*)

MS Ministério da Saúde

ODS Objetivos de Desenvolvimento Sustentável

OMS Organização Mundial da Saúde

RGWR Regressão Geograficamente Ponderada Robusta (*Robust Geographically Weighted Regression*)

SAR Modelo Autorregressivo Simultâneo (*Simultaneous Autoregressive Model*)

SINAN Sistema de Informação de Agravos de Notificação

SUS Sistema Único de Saúde

UFLA Universidade Federal de Lavras

VIF Fator de Inflação de Variância (*Variance Inflation Factor*)

WLS Mínimos Quadrados Ponderados (*Weighted Least Squares*)

LISTA DE SÍMBOLOS

n Número total de áreas/municípios observados

y_i Variável resposta (taxa de incidência) na área i

β Vetor de coeficientes de regressão

ε Vetor de erros aleatórios (ruído branco)

σ^2 Variância do erro

ρ Coeficiente de autocorrelação espacial (modelo SAR)

λ Parâmetro de dependência espacial no erro (modelo CAR)

W Matriz de pesos espaciais (vizinhança)

S Matriz de projeção

r_i Resíduo padronizado (studentizado)

AIC_c Critério de Informação de Akaike Corrigido

GLOSSÁRIO

Outlier Valor atípico ou observação que se desvia significativamente do padrão geral da distribuição dos dados, podendo influenciar estimativas estatísticas

Dados de área Dados espaciais agregados em unidades geográficas definidas por polígonos (como municípios, bairros ou setores censitários), em contraste com dados pontuais ou geoestatísticos

Matriz de pesos espaciais Matriz quadrada que formaliza a estrutura de vizinhança entre as áreas, definindo quais locais são conectados e a intensidade dessa conexão

Autocorrelação espacial Fenômeno onde os valores de uma variável em um local são correlacionados com os valores da mesma variável em locais vizinhos (a primeira lei da geografia)

Resíduos A diferença entre o valor observado e o valor estimado pelo modelo de regressão. Em modelos espaciais, busca-se que os resíduos não apresentem autocorrelação

Arbovirose Doença causada por vírus transmitidos, principalmente, por artrópodes (como mosquitos). Exemplos incluem Dengue, Zika e Chikungunya

Shapefile Formato de arquivo de dados vetoriais geoespaciais usado para armazenar a localização, forma e atributos de feições geográficas

Taxa de incidência Medida da frequência com que novos casos de uma doença ocorrem em uma população durante um período específico, geralmente expressa por 100 mil habitantes

1 INTRODUÇÃO

A dengue é uma ameaça crescente. Nos últimos anos, essa arbovirose tem causado surtos cada vez mais intensos, levando a milhares de hospitalizações e mortes ao redor do mundo (WHO, 2023). No Brasil, os números são alarmantes e casos explodem a cada nova temporada, pressionando os sistemas de saúde (Brasil, 2023)

O vírus da dengue é transmitido pelo *Aedes aegypti* e sua propagação está ligada a fatores ambientais e urbanos (Zaparoli *et al.*, 2021). Para entender melhor sua disseminação, para agir de forma eficaz, é essencial utilizar métodos analíticos precisos. O uso de métodos estatísticos pode ajudar a identificar padrões e determinar áreas vulneráveis.

Modelos lineares tradicionais são amplamente utilizados, mas assumem que as observações são independentes, ignorando a autocorrelação espacial. Isso pode comprometer as estimativas quando há dependência geográfica (Song *et al.*, 2021). Como demonstrado por Garcia, Soares-Filho & Moro (2024), a utilização inadequada desses modelos pode levar a inferências equivocadas, o que, por sua vez, pode comprometer a formulação de políticas públicas ineficientes, resultando em ações mal direcionadas e desperdício de recursos financeiros e humanos.

Para solucionar essa limitação surgiram os modelos espaciais, como a Regressão Simultânea Autorregressiva (SAR) e a Regressão Condicional Autorregressiva (CAR) que capturam padrões espaciais e, conseqüentemente, corrigem a dependência espacial existente entre as regiões. No entanto, esses modelos de regressão são extremamente sensíveis a presença de valores atípicos nos dados, o que pode comprometer a precisão das análises (Atikah; Afifah; Kholifia, 2020).

Diante disso, a Regressão Geograficamente Ponderada (GWR) se destaca como uma possível alternativa pois permite que os coeficientes variem espacialmente e, assim, possibilita capturar as características locais (Zhou; Wang; Fang, 2019; Lin; Wen, 2011). No entanto, a GWR também pode ser afetada por valores extremos, exigindo uma abordagem ainda mais robusta.

Para lidar com essa questão, Harris *et al.* (2010) propõe o uso da Regressão Geograficamente Ponderada Robusta (RGWR) que, supostamente, aprimoraria a resistência a presença de *outliers* nos dados de áreas e garantiria estimativas mais confiáveis. Como apontado por Schmid & Münnich (2014), modelos robustos são mais eficazes no tratamento de valores atípicos, especialmente em dados espaciais, assegurando análises mais precisas e confiáveis.

Neste estudo foram utilizados métodos robustos de modelagem espacial para dados de áreas para investigar a incidência de dengue em Minas Gerais no ano de 2023, utilizando . O objetivo foi verificar se modelos espaciais robustos, capazes de lidar com valores extremos, podem melhorar a qualidade das estimativas e fornecer informações mais precisas para o monitoramento e controle da incidência da dengue.

A pergunta que se faz é: estamos preparados para enfrentar a dengue de maneira eficiente ou continuaremos reféns de análises imprecisas?

Não foram encontrados trabalhos aplicados à dengue utilizando modelos robustos para dados de áreas. Nas bases pesquisadas foram identificados poucos materiais escritos, todos em língua inglesa, nenhum abordando especificamente a aplicação desses métodos à incidência de dengue. Diante disso, o uso desse tipo de modelo pode ser fundamental para analisar casos de dengue, não apenas em um estado específico, mas de forma mais ampla, abrangendo todas as regiões com incidência alarmante da doença, inclusive em situações em que os valores se apresentam discrepantes em relação à realidade dos casos.

1.1 Objetivos específicos

- Investigar a dependência espacial da incidência de dengue nos municípios de Minas Gerais, identificando padrões de concentração e a distribuição geográfica.
- Aplicar e ajustar modelos de regressão espacial robustos para avaliar a influência de covariáveis socioambientais sobre a incidência de dengue e determinar os principais fatores envolvidos.
- Identificar e interpretar áreas críticas, com elevada incidência de dengue, com base nos resultados do melhor modelo ajustado, subsidiando intervenções direcionadas e políticas públicas de saúde.

1.2 Objetivos de Desenvolvimento Sustentável - ODS

ODS 3 – Saúde e Bem-Estar

Garantir o acesso à saúde de qualidade e promover o bem-estar para todos em todas as idades (ONU, 2025; OPAS e OMS, 2024).

Esta pesquisa contribui diretamente para o Objetivo de Desenvolvimento Sustentável 3 (Saúde e Bem-Estar), com ênfase nos alvos **3.3** e **3.b**. A utilização de métodos de análise espacial possibilita a identificação de padrões de doenças transmissíveis, como a dengue; o monitoramento de riscos ambientais associados à contaminação do solo; e o apoio a estratégias de vigilância epidemiológica e resposta territorial, promovendo ações efetivas de saúde pública com base em evidências. A seguir, têm-se detalhes sobre os alvos, ou seja:

- 3.3 Até 2030, acabar com as epidemias de AIDS, tuberculose, malária e doenças tropicais negligenciadas, e combater a hepatite, doenças transmitidas pela água, e outras doenças transmissíveis
- 3.b Apoiar a pesquisa e o desenvolvimento de vacinas e medicamentos para as doenças transmissíveis e não transmissíveis, que afetam principalmente os países em desenvolvimento, proporcionar o acesso a medicamentos e vacinas essenciais a preços acessíveis, de acordo com a Declaração de Doha, que afirma o direito dos países em desenvolvimento de utilizarem plenamente as disposições do acordo TRIPS sobre flexibilidades para proteger a saúde pública e, em particular, proporcionar o acesso a medicamentos para todos

Na Figura 1.1 apresenta a ODS 3 – Saúde e Bem-Estar, a qual orienta este trabalho na busca pela redução dos impactos ao meio ambiente e à população.



Figura 1.1 – ODS 3 – Saúde e Bem-Estar

Fonte: (ONU, 2025)

2 REFERENCIAL TEÓRICO

Nesta seção serão apresentados os fundamentos teóricos que justificam esta pesquisa. Assim, será feita uma breve apresentação sobre o *Aedes aegypti* e a dengue e serão abordados conceitos essenciais para a análise espacial e modelagem estatística. Além disso, serão discutidos os modelos iniciais utilizados como base, servindo de referência para a construção e aplicação do modelo mais adequado ao estudo e o critério para avaliação do modelo.

Será dividido em cinco tópicos:

- Vetor *Aedes aegypti*;
- Casos de Dengue;
- Estatística Espacial;
- Dados de Área;
- Modelos para a análise;
- Critérios para avaliação dos modelos.

2.1 Vetor *Aedes aegypti*

O mosquito *Aedes aegypti* é originário da África e se espalhou pelo mundo, inicialmente, da costa oeste do continente para as Américas e depois, provavelmente, da costa leste da África para a Ásia (Cruz, 2025).

O vetor foi descrito cientificamente, pela primeira vez em 1762, quando foi denominado *Culex aegypti*. *Culex* significa “mosquito” e *aegypti*, egípcio e, portanto, significando “mosquito egípcio”. O gênero *Aedes aegypti*, descrita anos antes, apresenta características morfológicas e biológicas semelhantes às de espécies do gênero *Aedes* e não às do já conhecido gênero *Culex*. Então, foi estabelecido o nome *Aedes aegypti* que significa “odioso do Egito”. (Cruz, 2025).

Na Figura 2.1 apresentamos a imagem do mosquito transmissor da dengue. A seguir, fornecemos uma breve descrição sobre suas características.



Figura 2.1 – Mosquito da Dengue

Fonte: (Cruz, 2025)

Observa-se na Figura 2.1 que o mosquito *Aedes aegypti*, vetor da dengue, possui listras brancas nas patas e corpo escuro. Essas características auxiliam em sua identificação e diferenciação de outros mosquitos.

Na Figura 2.2 são apresentados dois mosquitos da *Aedes aegypti*. À esquerda, está o macho e à direita, a fêmea. O mosquito da dengue possui comprimento entre 5 e 7 milímetros (Cruz, 2025).

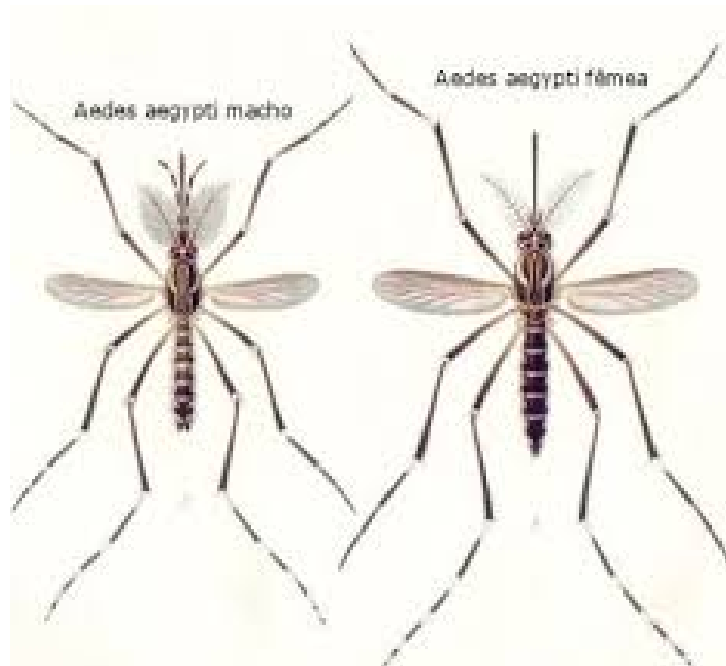


Figura 2.2 – Diferenças morfológicas entre o mosquito da dengue macho e fêmea.

Fonte: (Saúde, 2025)

A incidência de dengue tem aumentado significativamente em diversos países, especialmente nas Américas, representando um desafio considerável para os órgãos de saúde pública no combate a essa arbovirose. Apesar dos esforços intensificados dos órgãos de saúde pública para controlar a doença, observa-se uma grande heterogeneidade na distribuição dos casos, com números expressivos em algumas regiões.

A incidência é uma medida epidemiológica que representa o número de casos novos confirmados de dengue por 100 mil habitantes, considerando a população residente em um determinado espaço geográfico durante o ano analisado (Cruz, 2025).

Apesar dos esforços intensificados dos órgãos de saúde pública para controlar a dengue, a distribuição dos casos em Minas Gerais continua heterogênea, com algumas regiões registrando números expressivos de notificações. Essa variação pode estar associada a fatores espaciais, socioeconômicos e socioambientais que influenciam a proliferação do vetor e, consequentemente, a incidência da doença.

2.2 Casos de Dengue

O vetor da dengue é a fêmea do mosquito *Aedes aegypti*. Os vírus da dengue (DENV) estão classificados cientificamente na família Flaviviridae e no gênero Flavivirus. Até o momento são conhecidos quatro sorotipos; DENV-1, DENV-2, DENV-3 e DENV-4 (Brasil, 2023). As larvas do *Aedes aegypti* podem ser encontradas em diversos ambientes artificiais, caracterizando a domesticação e adaptação ao meio de vida humano, tornando-se antropofílico (Zaparoli *et al.*, 2021).

Para Biassoti & Ortiz (2017), o diagnóstico laboratorial da dengue torna-se muito importante, visto que diversas doenças febris agudas e exantemáticas apresentam diagnóstico clínico semelhante e podem ser confundidas com a infecção, como é o caso da leptospirose, malária, rubéola, influenza, febre maculosa brasileira e síndrome pulmonar por antivírus. Tauil (2001) afirma que a proliferação dos casos de dengue está diretamente relacionada às deficiências de saneamento básico, habitação e segurança pública que estão presentes nos aglomerados urbanos de países em desenvolvimento. Essas condições favorecem o surgimento de criadouros do *Aedes aegypti*. Ainda no estudo realizado por Biassoti & Ortiz (2017), foram considerados que os diagnósticos precisos da dengue configuram-se como fatores primordiais para o tratamento

clínico dos pacientes, diferenciação perante outras infecções e para controle das formas mais graves da doença.

Os mosquitos fêmea sugam sangue para produzir ovos. Se o mosquito da dengue estiver infectivo, poderá transmitir o vírus da dengue neste processo. Em geral, mosquitos sugam uma só pessoa a cada lote de ovos que produzem. O mosquito da dengue tem uma peculiaridade que se chama “discordância gonotrófica”, que significa que é capaz de picar mais de uma pessoa para um mesmo lote de ovos que produz, chegando a infectar até 5 pessoas (Cruz, 2025).

Conforme a OPAS e OMS (2024), o *Aedes aegypti* é o vetor que apresenta o maior risco de transmissão de arbovírus nas Américas e está presente em quase todos os países do hemisfério (exceto Canadá e Chile continental). É um mosquito doméstico (que vive dentro e perto das casas) que se reproduz em qualquer recipiente artificial ou natural que contenha água parada. Seu ciclo de vida pode completar, do ovo ao adulto, entre 7 e 10 dias; os mosquitos adultos geralmente vivem de 4 a 6 semanas. A fêmea do *Aedes aegypti* é responsável pela transmissão de doenças porque necessita de sangue humano para o desenvolvimento de seus ovos e para seu metabolismo. O macho não se alimenta de sangue. O comportamento de vôo e busca por alimento é mais intenso durante o início da manhã e ao anoitecer, períodos que representam maior risco de picadas.

No entanto, as fêmeas, que precisam continuar se alimentando, buscarão uma fonte de sangue em outros momentos. A fêmea *Aedes aegypti* se alimenta a cada 3-4 dias, porém, se ela não consegue tirar sangue suficiente, eles continuam a se alimentar a cada momento que podem. O *Aedes aegypti* prefere colocar seus ovos em recipientes artificiais contendo água, principalmente tambores, barris e pneus, dentro e ao redor de casas, escolas e locais de trabalho. Ovos dos *Aedes aegypti* podem resistir a condições ambientais secas por mais de um ano sendo esta uma das estratégias mais importantes que a espécie emprega para sobreviver e se espalhar (OPAS e OMS, 2024).

Na Figura 2.3 têm-se dois exemplares do mosquito *Aedes aegypti*, vetor da dengue, capturados pelo autor enquanto estudava na sala de aula do Departamento de Estatística da Universidade Federal de Lavras. A presença desses mosquitos em ambientes fechados reforça a importância de medidas preventivas para o controle da espécie, especialmente em locais com grande circulação de pessoas.



Figura 2.3 – Mosquitos *Aedes aegypti* capturados durante estudo em ambiente acadêmico.
Fonte: Autor (2024)

Como afirma OPAS e OMS (2024), na região das Américas, o número de casos de dengue notificados durante os primeiros seis meses de 2024 excedeu o maior número de casos notificados em um ano, de todos os anos anteriormente registrados. Até a semana epidemiológica 23 de 2024, 43 países e territórios da Região das Américas notificaram 9.386.082 casos de dengue, o dobro do número de casos notificados durante todo o ano de 2023, com 4.617.108 casos.

Barroso *et al.* (2020) afirmam que, atualmente, o Brasil é o país das Américas mais afetado em número de casos de dengue, sendo responsável por, aproximadamente, 70% dos casos notificados.

Na Figura 2.4 têm-se a média dos casos de dengue nos últimos cinco anos (2019-2023) e mostra que o Brasil evidenciou um aumento surpreendente nos casos de dengue em 2024. Nesse contexto, entender os padrões espaciais da incidência de doenças como a dengue é fundamental para o desenvolvimento de políticas públicas eficazes. A Estatística Espacial, portanto, torna-se uma ferramenta essencial, pois permite analisar como os fenômenos se distribuem no espaço geográfico e como essas distribuições podem influenciar o comportamento da arbovirose. A seguir, são explorados os conceitos e principais métodos da Estatística Espacial que são primordiais para entender a dinâmica dessa e de outras doenças em diferentes regiões.

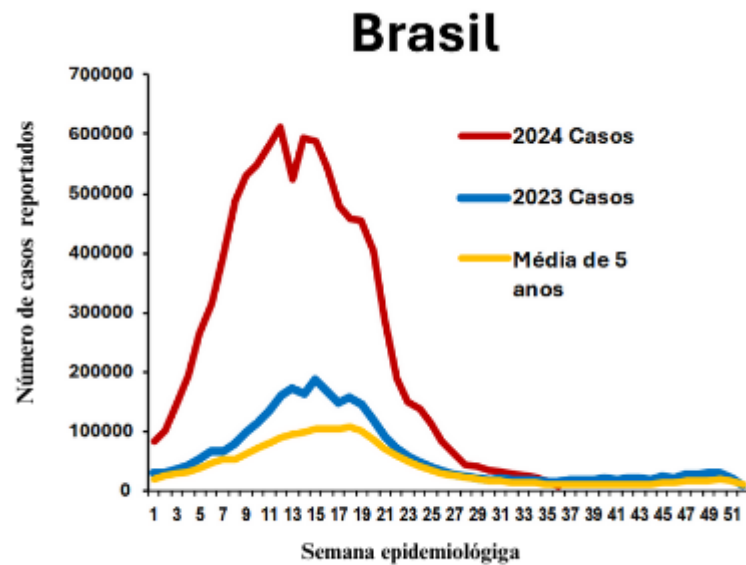


Figura 2.4 – Casos de dengue no Brasil
Fonte: (Saúde, 2024)

2.3 Estatística Espacial

De acordo com Scalón (2024), a Estatística Espacial é um ramo da Estatística que se concentra na análise de dados que possuem uma localização, representada por coordenadas geográficas e/ou regiões.

A Estatística Espacial fornece informações para responder a questões científicas em que a informação da localização espacial está presente nos dados e é relevante para as perguntas formuladas. O papel da teoria das probabilidades na Estatística Espacial é modelar a incerteza, tanto na teoria científica subjacente à questão, quanto nos dados espaciais provenientes das medições do processo espacial, que representa essa teoria científica (Cressie; Moores, 2023).

Conforme descrito por Scalón (2024), o que distingue a Estatística Espacial da clássica é a necessidade de que os dados possuam localizações geográficas que são utilizadas explicitamente na análise, característica ausente na Estatística Clássica. Essa distinção é fundamental para a análise de fenômenos que dependem da localização como é o caso das doenças transmitidas por vetores.

Dentro desse contexto, diferentes classificações dos dados espaciais foram propostas. De acordo com Cressie (1991), Hoef & Cressie (1993), Gatrell *et al.* (1996), Waller & Gotway (2004) e Scalón (2024), os dados espaciais podem ser classificados em três categorias princi-

país, com base na natureza estocástica de sua distribuição espacial. Essas classificações são essenciais para determinar os métodos apropriados para análise e interpretação de cada tipo de dado. São elas: Dados de processos pontuais; Dados de superfícies aleatórias (geoestatísticos); Dados de áreas.

De acordo com Scalon (2024), os dados espaciais podem ser caracterizados da seguinte forma:

- Dados de processos pontuais: são dados espaciais em que as próprias localizações dos eventos observados representam o principal objeto de interesse;
- Dados geoestatísticos: também chamados de dados de superfícies contínuas, representam variáveis observadas em locais específicos dentro de uma região de estudo, mas assumem continuidade espacial;
- Dados de área: também conhecidos como dados de *lattice* ou dados de superfície discreta, são dados espaciais indexados a sub-regiões que constituem uma partição de um domínio contínuo.

Esta dissertação utiliza exclusivamente dados de áreas e, portanto, exploraremos em detalhes os principais métodos utilizado na análise desse tipo de dados e suas aplicações na análise de fenômenos geográficos.

2.4 Dados de área

Dados de áreas, também conhecidos como dados de *lattice* ou ainda dados de superfície discreta. São dados de variáveis aleatórias associados a áreas contíguas e disjuntas de uma superfície (Scalon, 2024).

Conforme destacado por Scalon (2024), as sub-regiões (áreas) podem ser regulares ou irregularmente espaçadas, o que influencia diretamente a forma como os modelos espaciais interpretam a dependência e a variabilidade dos dados. Essas diferentes configurações espaciais exigem abordagens específicas para modelar as relações entre as unidades de análise, garantindo uma compreensão mais precisa dos padrões espaciais e, conseqüentemente, das dinâmicas que ocorrem dentro de cada região.

Nesse contexto, como afirmado por Silva *et al.* (2011), as técnicas de análise espacial visam identificar a semelhança ou dissimilaridade entre os valores de uma variável em uma

determinada região e seus vizinhos próximos ou distantes. Esse processo é fundamental para a detecção de estruturas espaciais que possam influenciar o fenômeno analisado e para a quantificação dessas relações. Assim, o uso de uma matriz de contiguidade (proximidade) espacial se torna indispensável. A matriz define e mede a proximidade entre as áreas analisadas, possibilitando uma análise mais robusta e eficaz.

2.4.1 Estrutura da dependência espacial

Para uma análise da estrutura espacial dos dados é necessário entender a dependência espacial, um conceito central na análise de dados espaciais. Conforme aponta Scalón (2024), espera-se que áreas que compartilham uma maior extensão de fronteira entre si apresentem uma maior interação. Essa interação determina a estrutura de vizinhança espacial que é essencial para a análise de dados espaciais.

A dependência espacial, por sua vez, refere-se à influência que as localizações geográficas vizinhas exercem sobre os valores de uma variável. Técnicas como o Índice de Moran são fundamentais para medir essa correlação, permitindo a análise de padrões e agrupamentos espaciais em dados agregados por área (Luzardo; Filho; Rubim, 2017). Esses métodos, como será abordado a seguir, são importantes para identificar a relação de dependência espacial.

2.4.2 Matriz de contiguidade espacial

A matriz de contiguidade expressa formalmente a estrutura espacial resultante, onde cada unidade espacial é representada tanto como uma linha quanto como uma coluna. Em cada linha, os elementos de coluna não nulos correspondem às unidades espaciais contíguas, refletindo as interações espaciais entre elas. Por convenção, uma célula não é contígua a si mesma, resultando em elementos diagonais zero (Anselin, 1988).

Essa matriz é uma ferramenta fundamental para caracterizar a dependência espacial e é amplamente utilizada na análise de dados espaciais de área, ajudando a identificar padrões de vizinhança e autocorrelação espacial, elementos essenciais para compreender as dinâmicas espaciais e aprimorar a interpretação dos dados, como detalhado por Anselin (1988).

Matrizes de contiguidade, como apresentadas na Figura 2.5, ilustram como a estrutura espacial é representada nas unidades de uma rede regular, evidenciando as interações espaciais

entre as áreas. A Figura 2.5 exibe a matriz de contiguidade detalhada por Anselin (1988), permitindo visualizar as relações de vizinhança entre as unidades espaciais.

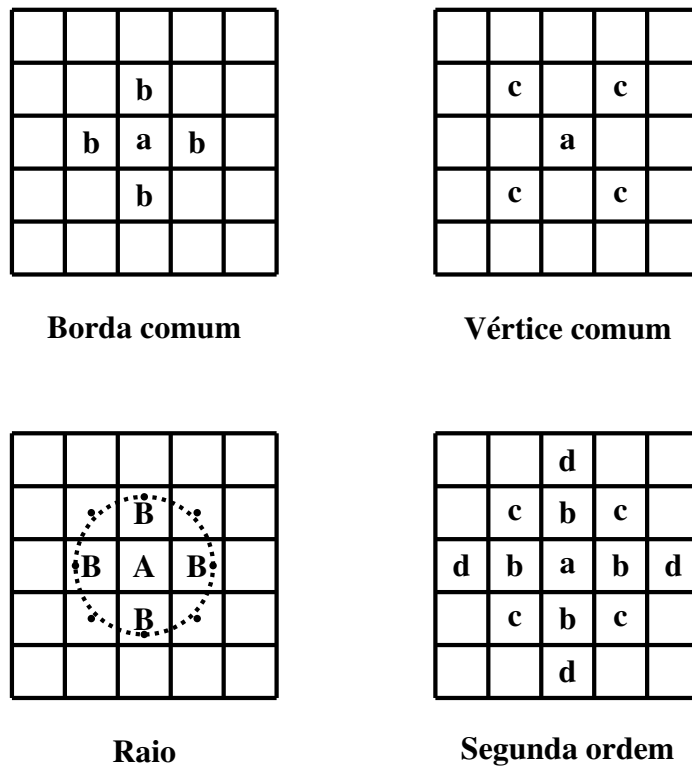


Figura 2.5 – Tipos de vizinhança espacial.

Fonte: Anselin (1988)

Como mostrado na Figura 2.5, a matriz de contiguidade ilustra a estrutura espacial das unidades geográficas, sendo um dos elementos-chave para a análise de dados espaciais (Anselin, 1988). A escolha dessa matriz é fundamental, pois ela determina o tipo de vizinhança a ser considerada, influenciando diretamente os resultados da análise espacial. Dependendo da configuração da matriz de contiguidade, diferentes padrões de interação espacial entre as unidades de análise são destacados, o que pode alterar a interpretação dos dados e os resultados modelados.

De acordo com Anselin (1988), a matriz de contiguidade para áreas irregulares pode ser representada a seguir:

$$w_{ij} = \begin{cases} 1 & \text{se a área } i \text{ e a área } j \text{ são vizinhas,} \\ 0 & \text{caso contrário,} \end{cases} \quad (2.1)$$

em que convencionou-se que $c_{ii} = 0$ para todo i (Cressie, 1991).

A Figura 2.6 apresenta a estrutura da matriz binária utilizada para representar áreas irregulares.

$$C = \begin{matrix} & \begin{matrix} \text{Áreas (j)} \\ 1 & 2 & 3 & \dots & n \end{matrix} \\ \begin{matrix} \text{Áreas (i)} \\ 1 \\ 2 \\ 3 \\ \vdots \\ n \end{matrix} & \begin{bmatrix} 0 & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & 0 & w_{23} & \dots & w_{2n} \\ w_{31} & w_{32} & 0 & \dots & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \dots & 0 \end{bmatrix} \end{matrix}$$

Figura 2.6 – Representação genérica de uma matriz de pesos espaciais C de ordem $n \times n$.
Fonte: Autor (2025).

De acordo com Anselin (1988), o uso de uma matriz de pesos espaciais tem como objetivo principal relacionar os valores de uma variável observada em um ponto específico do espaço com os valores dessa variável em outras unidades espaciais ao redor. Esse tipo de relação é essencial para a modelagem de dependência espacial, onde a proximidade geográfica entre as unidades influencia a variável de interesse, permitindo uma compreensão mais precisa das dinâmicas espaciais. Esse conceito de dependência espacial está diretamente relacionado à autocorrelação espacial, que será discutida a seguir.

A Figura 2.7 apresenta a relação entre a representação cartográfica e a representação matricial de uma malha irregular.

Conforme observado na Figura 2.7, o gráfico (A) destaca os polígonos geográficos, enquanto o gráfico (B) exibe a matriz de contiguidade correspondente em que o número 1 representa municípios vizinhos (apresentam bordas em comum). Para esta ilustração, foram selecionados sete municípios a fim de demonstrar como as conexões espaciais (vizinhança) são estruturadas em grades irregulares.

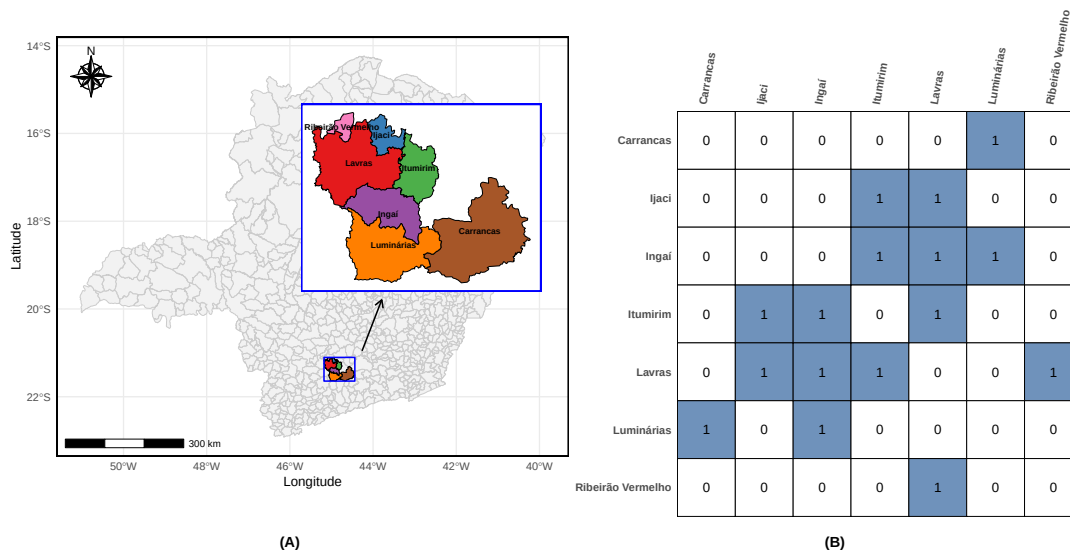


Figura 2.7 – Representação de uma área irregular (A) e sua respectiva matriz de adjacência binária W (B).

Fonte: Autor (2025).

2.4.3 Autocorrelação espacial

A autocorrelação espacial é uma medida que indica o grau de similaridade (dissimilaridade) de uma mesma variável no espaço. A ideia básica da autocorrelação espacial é determinar um indicador que mede a relação de uma determinada variável no espaço, ou seja, como essa variável se comporta numa determinada região geográfica (Manuel, 2011).

Os primeiros passos para a autocorrelação espacial (*spatial autocorrelation*) foram abordados por Moran (1948) e Geary (1954) que desenvolveram suas medidas de autocorrelação espacial (Anselin, 1988).

Para Anselin (1988), o problema da unidade de área modificável está relacionado à influência da definição das unidades espaciais sobre os resultados estatísticos em análises de dados geográficos. Esse problema pode impactar a interpretação da autocorrelação espacial, uma vez que medidas como o índice global e local de Moran, o índice de Geary e o variograma são sensíveis à maneira como as unidades espaciais são organizadas.

Conforme Druck *et al.* (2004), a autocorrelação espacial pode ser medida por meio de diversos índices, como o índice global e local de Moran, o índice de Geary e o variograma. Dentre esses indicadores, o Índice de Moran é o mais amplamente utilizado em análises de dados de áreas, pois oferece uma maneira robusta de quantificar a dependência espacial global,

ou seja, a relação entre os valores observados em diferentes unidades espaciais. Esse índice é fundamental para a detecção de padrões espaciais e é apresentado a seguir.

Índice de Moran global (I)

Conforme Waller & Gotway (2004), os índices globais de autocorrelação espacial, como o Índice de Moran (I), calculam a similaridade entre as regiões i e j com base no produto das diferenças entre os valores Y_i e Y_j , em relação à média global da variável aleatória Y . Este índice é amplamente utilizado para medir a autocorrelação espacial global, permitindo identificar padrões de agrupamento ou dispersão em dados de área. O Índice de Moran é essencial para entender como os valores observados em uma área estão relacionados com os valores em áreas vizinhas, fornecendo uma visão sobre a estrutura espacial dos dados.

A estatística do índice de Moran, conforme apresentada por Bivand, Pebesma & Rubio (2008), é estimada da seguinte forma:

$$\hat{I} = \frac{n}{\sum_{i=1}^n \sum_{j=1}^n W_{ij}} \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (2.2)$$

em que:

- W_{ij} : elemento da matriz de proximidade espacial entre os locais i e j ,
- Y_i : é o variável aleatória na área i ,
- Y_j : é o variável aleatória na área j ,
- $\bar{Y}$: valor médio amostral da variável aleatória em toda a região média de $\bar{Y} = \sum_{i=1}^n Z_i/n$ sobre todos os n locais de observação,
- n : número total de locais de observação.

O Índice de Moran (I) é uma estatística que varia, na maioria dos casos, entre **-1** e **1**, sendo que valores próximos de **1** indicam forte autocorrelação espacial positiva, ou seja, unidades geográficas vizinhas tendem a apresentar valores semelhantes. Valores próximos de **-1** indicam autocorrelação espacial negativa extrema, refletindo uma tendência de que áreas próximas apresentem valores contrastantes. Já valores de I próximos de zero sugerem ausência de

autocorrelação espacial, indicando uma distribuição aleatória dos dados no espaço. Essa estatística assume valores no conjunto dos números reais $\mathbb{R}$ e é amplamente utilizada para identificar padrões espaciais em fenômenos geográficos (Scalon, 2024).

Essa análise global da autocorrelação espacial oferece uma visão importante das interações gerais entre as unidades espaciais, mas ela não revela detalhes específicos sobre as relações locais. Para investigar essas relações em nível mais preciso é necessário recorrer ao Índice de Moran Local que permite examinar a autocorrelação espacial de forma mais detalhada, focando em áreas individuais dentro da região estudada. O Índice de Moran Local (LISA) foi desenvolvido justamente para essa finalidade, permitindo testar a autocorrelação local e identificar áreas espaciais que têm um impacto significativo no Índice de Moran Global.

Para verificar a significância da estatística de Moran, utiliza-se o teste de Wald, cuja estatística é dada por (Scalon, 2024):

$$Z = \frac{\hat{I} - \mathbb{E}[\hat{I}]}{\sqrt{\text{Var}[\hat{I}]}} \quad (2.3)$$

em que $\hat{I}$ representa a estatística de Moran observada, conforme definida na Equação 2.4.3, e $\mathbb{E}[\hat{I}]$ é o valor esperado da estatística $\hat{I}$.

O teste de Wald pressupõe a variável aleatória Y siga, aproximadamente, distribuição normal. Caso contrário, pode-se testar a significância estatística de $\hat{I}$ utilizando aleatorização (Scalon, 2024).

Índice de Moran Local (LISA)

O Índice de Moran Local (LISA) foi desenvolvido como uma estatística para analisar a autocorrelação espacial em nível local e identificar unidades espaciais que exercem influência sobre o Índice de Moran Global (I), para cada área i do estudo (Luzardo; Filho; Rubim, 2017).

De acordo com Scalon (2024), o estimador do índice local de autocorrelação espacial de Moran na i -ésima área ($\hat{I}_i$) é definido como:

$$\hat{I}_i = \frac{(Y_i - \bar{Y}) \sum_{j=1}^n W_{ij} (Y_j - \bar{Y})}{\sum_{j=1}^n (Y_j - \bar{Y})^2} \quad (2.4)$$

em que:

- Y_i é o valor da variável aleatória na área i ;
- Y_j é o valor da variável aleatória na área j ;
- $\bar{Y}$ é o valor da média amostral da variável aleatória em toda a região ($\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$)
- W_{ij} são os elementos da matriz de pesos espaciais, normalizados por linha.

A interpretação da equação de (2.4) é análoga à interpretação do índice global de Moran. Isso indica que o Índice de Moran Local (I_i) é útil para avaliar a homogeneidade ou a diversidade dos dados localmente conforme mencionado por Luzardo, Filho & Rubim (2017).

Complementando essa análise, Manuel (2011) destaca que o “LISA” permite verificar a similaridade ou dissimilaridade dos valores de algumas observações em relação às suas observações vizinhas. Essa medida local oferece uma visão detalhada dos padrões espaciais, revelando como as características das áreas analisadas se relacionam com as regiões ao seu redor.

Enquanto isso, Neves *et al.* (2000) explica que, ao contrário dos indicadores globais, como o índice de Moran, que fornecem um único valor representando a associação espacial para todo o conjunto de dados, os indicadores locais geram um valor específico para cada unidade espacial. Esse enfoque permite identificar com maior precisão agrupamentos de objetos com valores semelhantes (clusters), objetos anômalos (outliers) e diferentes regimes espaciais, facilitando a compreensão mais profunda da estrutura espacial dos dados.

Diagrama de Espalhamento de Moran

O Diagrama de Espalhamento de Moran, conforme descrito por Marques (2009), é uma ferramenta gráfica útil para visualizar o comportamento dos dados. Nesse diagrama, os valores dos desvios dos atributos em relação à média (Z) são representados no eixo X (abscissas), enquanto o valor da média dos vizinhos (WZ) está associado ao eixo Y (ordenadas), proporcionando uma representação clara das relações espaciais entre as unidades de análise, conforme mostra a Figura 2.8.

Conforme Marques (2009), a interpretação do diagrama a Figura 2.8 pode ser feita da seguinte forma:

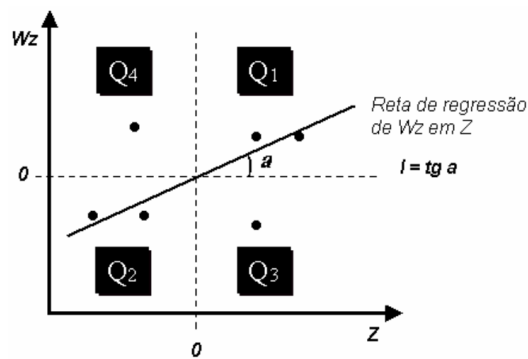


Figura 2.8 – Exemplo de diagrama de espalhamento de Moran

Fonte: Marques (2009)

- Q1 (valores [+], médias [+]) e Q2 (valores [-], médias [-]) indicam pontos de associação espacial positiva, no sentido que uma localização possui vizinhos com valores similares.
- Q3 (val. [+], médias [-]) e Q4 (val. [-], médias [+]) indicam pontos de associação espacial negativa, no sentido que uma localização possui vizinhos com valores dissimilares.

Cabe aqui salientar que os pontos localizados em Q3 e Q4 podem ser vistos como extremos, tanto por estar afastados da reta de regressão linear, como por indicar regiões que não seguem o mesmo processo de dependência espacial das demais observações. Estes pontos marcam regiões de transição entre regimes espaciais distintos (Marques, 2009).

2.5 Modelos de Regressão na análise de incidência da dengue: Regressão linear e Modelos com Dependência Espacial

Nesta seção são apresentados os modelos utilizados para a análise da incidência de dengue, incluindo tanto os modelos clássicos de regressão quanto aqueles que incorporam efeitos espaciais. São abordadas suas definições, métodos de estimação e pressupostos, fornecendo a base teórica para a modelagem dos dados.

A análise inicia-se com a aplicação do Modelo de Regressão Linear Múltipla com o objetivo de investigar a distribuição dos casos de dengue no Estado de Minas Gerais. Esse modelo fornece uma abordagem inicial para compreender as relações entre a incidência da doença e as covariáveis selecionadas. Contudo, a regressão linear múltipla pressupõe que os resíduos sejam normalmente distribuídos e que não existam efeitos espaciais significativos, o que pode limitar sua aplicabilidade em contextos nos quais essas premissas não são atendidas.

Um dos desafios desse modelo é a influência de *outliers* que podem distorcer os coeficientes de regressão e comprometer a precisão das estimativas. Segundo Hawkins (1980), um *outlier* é uma observação que se desvia significativamente das demais, podendo indicar um comportamento atípico ou erro nos dados. Além disso, Rousseeuw & Leroy (1987) destacam que a presença de *outliers* tanto na variável resposta Y quanto nas variáveis explicativas X pode impactar a qualidade dos ajustes e prejudicar a interpretação dos resultados.

Diante dessas limitações, torna-se necessário explorar metodologias mais robustas que possam lidar melhor com a presença de valores extremos e efeitos espaciais na distribuição dos casos de dengue. Assim, a análise será ampliada para considerar modelos que incorporam efeitos espaciais e técnicas robustas, permitindo uma avaliação mais precisa da influência dos fatores analisados.

2.5.1 Modelos de Regressão Linear Múltipla

A regressão linear é um modelo estatístico amplamente utilizado para modelar a relação entre uma variável dependente Y e uma ou mais variáveis independentes X_k . Segundo Su, Yan & Tsai (2012), essa abordagem é popular devido à sua simplicidade, facilidade de interpretação e fundamentação teórica. No caso da regressão linear simples, um único preditor é utilizado para explicar Y , enquanto na regressão linear múltipla, diversos preditores são considerados, permitindo um ajuste mais preciso do modelo.

Para Rencher & Schaalje (2008), a regressão múltipla assume uma relação linear entre Y_i e os preditores $X_{i1}, X_{i2}, \dots, X_{ik}$, usado no modelo de regressão múltipla sendo representada por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \quad (2.5)$$

em que:

- Y_i é a variável dependente,
- β_0 é o intercepto,
- $\beta_1, \beta_2, \dots, \beta_k$ são os coeficientes das variáveis independentes,
- $X_{i1}, X_{i2}, \dots, X_{ik}$ são as variáveis independentes,

- ε_i é o termo de erro, com $\varepsilon_i \sim N(0, \sigma^2)$.

O modelo matricial da regressão linear pode ser expresso da seguinte forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.6)$$

ou seja,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (2.7)$$

em que:

- Y é um vetor de dimensão $n \times 1$ contendo a variável dependente;
- X é uma matriz de dimensão n contendo as variáveis independentes (incluindo uma coluna de uns para o intercepto);
- $\boldsymbol{\beta}$ é um vetor contendo os parâmetros do modelo;
- $\boldsymbol{\varepsilon}_i$ é um vetor contendo os erros aleatórios não explicados pelo modelo.

Esse modelo matricial representa a relação entre a variável dependente e variáveis independentes na regressão linear múltipla.

Um método comumente utilizado para estimar os parâmetros do modelo de regressão é o método dos mínimos quadrados. No entanto, na presença de *outliers* ou quando os pressupostos do modelo são violados, essas estimativas podem tornar-se não confiáveis (Fontoura, 2015).

Após a aplicação de um modelo de regressão linear, é fundamental verificar se ele atende a certas condições estatísticas. Essas condições garantem que as estimativas obtidas sejam confiáveis e que as inferências realizadas a partir do modelo sejam válidas. A verificação dessas condições é feita por meio de testes estatísticos que avaliam as pressuposições do modelo ajustado.

A seguir, apresentamos as principais pressuposições que devem ser verificadas após o ajuste do modelo.

2.5.1.1 Análise das pressuposições sobre a variável aleatória erro

Os modelos de regressão linear baseiam-se em um conjunto de pressuposições estatísticas que garantem a validade das estimativas e inferências. Essas pressuposições referem-se principalmente ao comportamento dos erros do modelo, assumindo que:

- os erros sejam independentes;
- os erros possuam variância constante (homocedasticidade);
- os erros seguem uma distribuição normal com média zero e variância constante, isto é, $\varepsilon \sim N(0, I\sigma^2)$.

A verificação dessas condições é essencial, pois qualquer violação pode comprometer a validade dos resultados obtidos. Caso os pressupostos não sejam atendidos, as inferências realizadas com base no modelo podem ser viesadas ou incorretas, exigindo ajustes ou o uso de técnicas estatísticas mais robustas (Manuel, 2011).

A seguir, são descritos os pressupostos que o modelo de regressão linear deve atender para que ele seja ajustado de maneira adequada (Amador *et al.*, 2011).

- **Pressuposição de normalidade:** A hipótese de normalidade dos resíduos do modelo de regressão pode ser avaliada usando gráficos de probabilidade normal ou através de testes de normalidade, como o teste de Shapiro-Wilk, o teste de Jarque–Bera, dentre outros.
- **Pressuposição de homocedasticidade:** A homocedasticidade dos resíduos é uma das mais importantes pressuposições feitas neste tipo de modelo pois ela garante a eficiência dos testes de hipóteses. Para avaliá-la, podem ser usados vários testes estatísticos, como Hartley, Levene, Bartlett e Breusch-Pagan.
- **Pressuposição da independência:** Os resíduos do modelo devem ser independentes. Essa independência pode ser avaliada por métodos gráficos ou por meio de testes de hipótese como os testes de Durbin–Watson e Breusch–Godfrey.

2.5.2 O modelo espacial autoregressivo (Spatial Autoregressive Model - SAR)

A classe de modelos autorregressivos simultâneos (SAR), também conhecida como modelo de defasagem espacial ou *lag model* (*simultaneous autoregressive model*), utiliza a regressão dos valores observados em outras áreas para incorporar a dependência espacial. Dessa forma, a autocorrelação espacial é capturada por um único parâmetro adicional, incorporado ao modelo clássico de regressão (Scalon, 2024).

O modelo espacial autoregressivo (SAR) é descrito conforme Anselin (1988) por:

$$\mathbf{Y} = \boldsymbol{\delta} \mathbf{W} \mathbf{Y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.8)$$

em que:

- $\mathbf{Y}$ é o vetor da variável dependente,
- $\mathbf{X}$ é a matriz das variáveis independentes,
- $\boldsymbol{\delta}$ é o coeficiente de defasagem da variável dependente espacial,
- $\boldsymbol{\beta}$ é o vetor de parâmetros da regressão,
- $\mathbf{W}$ é a matriz de ponderação, de tamanho $\mathbf{n} \times \mathbf{n}$,
- $\boldsymbol{\varepsilon}$ é o vetor de erros, de tamanho $\mathbf{n} \times \mathbf{1}$ não correlacionados que seguem uma distribuição normal multivariada com média zero e variância constante, isto é, $\boldsymbol{\varepsilon} \sim N(0, I\sigma^2)$

A hipótese nula para a não existência de autocorrelação espacial é que $\boldsymbol{\delta} = 0$. A ideia básica nesse modelo é incorporar a autocorrelação espacial como componente do modelo. Pode-se observar que na ausência de autocorrelação espacial ($\boldsymbol{\delta} = 0$), o modelo espacial autoregressivo (equação 2.8) é o próprio modelo de regressão linear geral (equação 2.6).

2.5.3 Modelo de Erro Espacial (Conditional Autoregressive Model - CAR)

O modelo CAR considera que os efeitos espaciais são ruídos ou perturbações, ou seja, um termo que precisa ser removido. A especificação CAR baseia-se na distribuição condicional dos termos de erro espacial (Bivand; Pebesma; Rubio, 2008) .

O modelo de erro espacial, conforme Anselin (1988), é representado por:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.9)$$

em que $\mathbf{Y}$ é o vetor da variável dependente, $\mathbf{X}$ é a matriz de variáveis explicativas, $\boldsymbol{\beta}$ é o vetor dos parâmetros e $\boldsymbol{\varepsilon}$ representa o vetor de erros.

O vetor de erros $\boldsymbol{\varepsilon}$ é modelado da seguinte forma:

$$\boldsymbol{\varepsilon} = \boldsymbol{\lambda} \mathbf{W} \boldsymbol{\varepsilon} + \mathbf{u}, \quad (2.10)$$

em que $\boldsymbol{\lambda}$ é um coeficiente autorregressivo espacial, $\mathbf{W}$ é a matriz de proximidade de pesos espaciais e $\mathbf{u}$ é um termo de erro que satisfaz as suposições clássicas de distribuição independente e identicamente distribuída (i.i.d.), com variância constante σ^2 .

A hipótese nula para a não existência de autocorrelação espacial é que $\boldsymbol{\lambda} = \mathbf{0}$, ou seja, o termo de erro não é espacialmente correlacionado. Pode-se observar que, na ausência de autocorrelação espacial ($\boldsymbol{\lambda} = \mathbf{0}$), o modelo espacial autorregressivo da (equação 2.9) é o próprio modelo de regressão linear geral (equação 2.6).

Pressupostos dos modelos

Conforme Scalón (2024), ao ajustar os modelos de regressão e regressão espacial (SAR e CAR), considerando a autocorrelação espacial, parte-se da suposição de que as propriedades impostas pelo termo dos erros do modelo são verdadeiras. Consequentemente, essas suposições precisam ser validadas para garantir que os resultados inferenciais sejam confiáveis. Ainda segundo Scalón (2024), as suposições dos erros a serem testadas incluem independência, variância constante e normalidade nos resíduos. No entanto, quando essas pressuposições são violadas, as análises podem se tornar pouco confiáveis, uma vez que os estimadores tradicionais são sensíveis à presença de valores discrepantes. Como alternativa, métodos robustos, podem ser empregados para minimizar o impacto dos *outliers* e melhorar a qualidade do ajuste do modelo (Castelani *et al.*, 2019).

Testes de correlação espacial residual baseados na estatística de Moran

A autocorrelação espacial nos resíduos de um modelo de regressão pode indicar a presença de dependência espacial não capturada, o que compromete a validade das inferências estatísticas. Para lidar com essa questão, Anselin (1988) propôs um teste baseado no Índice de Moran I , que permite avaliar a estrutura espacial dos resíduos de um modelo estimado. Esse teste verifica se há padrões espaciais significativos nos resíduos, sugerindo possíveis omissões de variáveis espaciais relevantes ou especificações inadequadas do modelo.

Formalmente, a estatística de Moran I para os resíduos da regressão é definida como:

$$I = \frac{N}{S} \cdot \frac{\mathbf{e}'\mathbf{W}\mathbf{e}}{\mathbf{e}'\mathbf{e}} \quad (2.11)$$

em que:

- $\mathbf{e}$ é o vetor de resíduos da regressão estimada por Mínimos Quadrados Ordinários (OLS);
- $\mathbf{W}$ é a matriz de pesos espaciais, que define a estrutura de vizinhança entre as observações;
- N representa o número total de observações na amostra;
- S é um fator de padronização, correspondente à soma de todos os elementos da matriz de pesos espaciais.

A interpretação da estatística de Moran I nos resíduos segue a mesma lógica de sua aplicação em dados brutos. Valores significativamente positivos indicam a presença de autocorrelação espacial positiva, enquanto valores negativos sugerem padrões espaciais dispersos.

A significância do índice pode ser avaliada por meio de testes de permutação, que geram uma distribuição nula da estatística a partir de reamostragens aleatórias dos valores observados (Scalon, 2024). Caso a estatística seja significativa, recomenda-se revisar a especificação do modelo, incorporando efeitos espaciais por meio de modelos espaciais.

2.5.4 Modelo de Regressão Múltipla Robusta

Para Fox & Weisberg (2002), a regressão robusta é um método de estimação para modelos de regressão linear que é insensível a presença de outliers e pontos de alta alavancagem. Ela

fornece informações úteis mesmo quando algumas suposições do método não são aplicáveis. Esses métodos começaram a ser desenvolvidos em meados da década de 1960.

Para Oliveira (2020), os pontos de alta alavancagem, são pontos que possuem valores incomuns ou extremos nas variáveis preditoras (também chamadas de variáveis independentes). Esses pontos podem influenciar no ajuste do modelo porque podem alterar significativamente os coeficientes estimados. Embora esses pontos possam ser ajustados bem a um modelo linear, eles podem não ser representativos do comportamento global dos dados, o que pode comprometer a adequação do modelo para a maioria das observações.

Uma análise de mínimos quadrados pondera igualmente cada observação na obtenção das estimativas dos parâmetros. Porém, os métodos robustos permitem que as observações sejam ponderadas de forma desigual, ou seja, as observações que produzem grandes resíduos são desconsideradas em parte por um método de estimação robusta (Drapper; Smith, 1998).

Consideramos apenas o modelo linear (Fox; Weisberg, 2002)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \quad (2.12)$$

com representação matricial dada por:

$$Y_i = X_i \beta + \varepsilon_i \quad (2.13)$$

em que X_i é a i -ésima linha de $\mathbf{X}$, $i = 1, 2, \dots, n$, observações.

Dado um estimador b para β , o modelo ajustado é

$$\hat{Y}_i = a + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik} = X_i b \quad (2.14)$$

em que e_i são resíduos, que são dados por:

$$e_i = Y_i - \hat{Y}_i. \quad (2.15)$$

Conforme apresentado por Drapper & Smith (1998), as estimativas b são obtidas minimizando-se uma função objetivo específica em relação a todos os valores possíveis de b :

$$\sum_{i=1}^n \rho(e_i) = \sum_{i=1}^n \rho(Y_i - x_i' b) \quad (2.16)$$

em que a função ρ define a contribuição de cada resíduo para a função objetivo.

Uma função ρ deve possuir as seguintes propriedades:

- Sempre não negativa: $\rho(e) \geq 0$;
- Igual a zero quando seu argumento é zero: $\rho(0) = 0$;
- Simétrica: $\rho(e) = \rho(-e)$; e
- Monótona em relação ao valor absoluto de e_i : $\rho(e_i) \geq \rho(e'_i)$ para $|e_i| > |e'_i|$.

Essas propriedades são essenciais para assegurar que a função ρ seja apropriada à estimação robusta. Com base nesse princípio, define-se o conceito de M-estimador (Drapper; Smith, 1998).

M-Estimador

Para Drapper & Smith (1998), os *M-estimadores* são uma generalização dos estimadores de máxima verossimilhança. Suponha que os erros sejam independentes e identicamente distribuídos e sigam uma mesma distribuição $f()$. O estimador de máxima verossimilhança (MLE) de β , dado por $\hat{\beta}$, maximiza a função de verossimilhança:

$$L(\beta) = \prod_{i=1}^n f(Y_i - \mathbf{x}'_i \beta), \quad (2.17)$$

em que $\mathbf{x}'_i$ representa a i -ésima linha da matriz $\mathbf{X}$, para $i = 1, 2, \dots, n$, no modelo $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$.

Equivalentemente, o MLE de β pode ser obtido maximizando a função log-verossimilhança:

$$\ell(\beta) = \sum_{i=1}^n \ln f(Y_i - \mathbf{x}'_i \beta). \quad (2.18)$$

A forma específica da função de verossimilhança depende da distribuição assumida para os erros. Quando os erros seguem uma distribuição normal, $f()$ corresponde à densidade normal, e maximizar $\ell(\beta)$ equivale a minimizar a soma dos quadrados dos erros:

$$\sum_{i=1}^n (Y_i - \mathbf{x}'_i \beta)^2. \quad (2.19)$$

No entanto, a normalidade dos erros é uma suposição forte e pode não ser válida em diversas aplicações. Para dados com maior propensão a *outliers* ou desvios significativos da

normalidade, a distribuição exponencial dupla, também conhecida como distribuição de Laplace, pode ser uma alternativa mais adequada. Essa distribuição possui caudas mais pesadas que a normal, tornando-a mais resistente a valores extremos. Nesse caso, a função de verossimilhança leva à minimização da soma dos valores absolutos dos erros:

$$\sum_{i=1}^n |Y_i - \mathbf{x}_i' \beta|. \quad (2.20)$$

A regressão baseada na minimização dos valores absolutos dos erros, conhecida como regressão da mediana, apresenta maior robustez a presença de *outliers* em comparação ao método dos mínimos quadrados, uma vez que a penalização para grandes desvios é menor. Entretanto, ambas as abordagens podem ser interpretadas como casos particulares de uma formulação mais geral: os M-estimadores (Drapper; Smith, 1998).

Conforme Drapper & Smith (1998), essa ideia pode ser ampliada ao considerar uma função $\rho(u)$, definida para os resíduos u , juntamente com uma estimativa de escala s que não precisa ser a estimativa tradicional dos mínimos quadrados. Essa estimativa de escala oferece uma medida alternativa da variabilidade dos dados, tornando-se mais robusta a presença de *outliers* e valores extremos, especialmente em cenários onde as suposições dos métodos tradicionais não são atendidas.

Além disso, a necessidade de métodos robustos não se limita à suposição de normalidade. *Outliers* podem impactar negativamente a estimação de parâmetros em diversos modelos paramétricos, como aqueles baseados em distribuições exponencial, Weibull ou gama. Nessas situações, a utilização da máxima verossimilhança tradicional pode resultar em estimativas distorcidas, o que exige abordagens robustas para garantir a confiabilidade da análise (Maronna Ricardo A. e Martin, 2006).

Diante desse cenário, os *M-estimadores* surgem como uma alternativa robusta, permitindo a escolha de diferentes funções de penalização para os resíduos. Ao contrário dos métodos tradicionais, que simplesmente minimizam a soma dos quadrados ou dos valores absolutos dos resíduos, os *M-estimadores* utilizam uma função de perda mais flexível e ajustável para reduzir a influência de *outliers*. Isso torna os *M-estimadores* particularmente úteis em situações onde a presença de valores atípicos pode afetar significativamente os resultados.

Assim, definimos um estimador robusto como aquele que minimiza a seguinte expressão:

$$\sum_{i=1}^n \rho(u_i) = \sum_{i=1}^n \rho\left(\frac{e_i}{s}\right) = \sum_{i=1}^n \rho\left(\frac{Y_i - \mathbf{x}_i' \boldsymbol{\beta}}{s}\right) \quad (2.21)$$

, em que o fator de escala s é definido por:

$$s = \frac{\text{mediana}|e_i - \text{mediana}(e_i)|}{0,6745}, \text{ ou Desvio Padrão Absoluto (MAD)}. \quad (2.22)$$

O fator de escala s desempenha um papel fundamental na regressão robusta, pois ele influencia diretamente como os resíduos são padronizados, impactando, portanto, a atribuição de pesos a cada observação no processo iterativo de estimação dos coeficientes. A escolha do denominador 0,6745 baseia-se no fato de que este valor corresponde ao quantil z com probabilidade acumulada de 75% na distribuição normal padrão. Dessa forma, sob a hipótese de normalidade dos resíduos, s atua como um estimador não tendencioso do desvio padrão dos resíduos. Essa padronização é essencial para a aplicação eficaz das funções de perda e peso que serão apresentadas a seguir.

Entre as funções de perda robustas, a função de perda de Huber se destaca por combinar a suavidade da soma dos quadrados dos erros, típica da regressão dos mínimos quadrados, com a capacidade de resistir a *outliers*. Essa característica torna a função de Huber particularmente útil em cenários nos quais os dados contêm valores atípicos, mas onde não se deseja que esses *outliers* afetem substancialmente o modelo, como ocorre nos métodos tradicionais de mínimos quadrados (Fox; Weisberg, 2002).

As funções de perda e peso de Huber, conforme descritas por Huber (1981), são expressas pelas seguintes equações:

$$\rho_H(e_i) = \begin{cases} \frac{1}{2}e_i^2 & \text{se } |e_i| \leq k \\ k|e_i| - \frac{1}{2}k^2 & \text{se } |e_i| > k \end{cases} \quad (2.23)$$

e

$$w_H(e_i) = \begin{cases} 1 & \text{se } |e_i| \leq k \\ \frac{k}{|e_i|} & \text{se } |e_i| > k \end{cases} \quad (2.24)$$

em que $\rho_H(e_i)$ é a função de perda e $w_H(e_i)$ é a função de pesos.

Nessas equações, e_i representa o erro ou resíduo de cada observação, que é a diferença entre o valor observado Y_i e o valor predito pelo modelo $\hat{Y}_i$, ou seja, $e_i = Y_i - \hat{Y}_i$. Já k é um parâmetro de limiar que determina o ponto de transição entre o tratamento suave dos resíduos (quando $|e| \leq k$) e o tratamento robusto (quando $|e| > k$). Resíduos pequenos, com valor absoluto inferior ou igual a k , são penalizados quadraticamente, enquanto resíduos maiores que k são penalizados de forma linear, o que diminui a influência de outliers no modelo.

Essas funções de perda e peso são essenciais para garantir uma estimação robusta dos parâmetros pois permitem que o modelo lide adequadamente com dados que contêm valores atípicos. No próximo item, será detalhado o procedimento de *M-estimação* que busca minimizar a soma das funções de perda e aplicar os pesos corretamente, ajustando os coeficientes do modelo de forma robusta.

O procedimento de M-estimação

O procedimento de *M-estimação* consiste em minimizar a equação (2.21) em relação aos parâmetros β_j , $j = 0, 1, 2, \dots, m$. Diferenciando essa equação em relação a cada parâmetro, obtemos $p = m + 1$ equações do tipo:

$$\sum_{i=1}^n x_{ij} \psi' \left\{ \frac{Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}}{s} \right\} = 0, \quad j = 0, 1, 2, \dots, m, \quad (2.25)$$

em que $\psi'(u)$ é a derivada da função $\rho(u)$. Como as equações geralmente não possuem solução explícita, é necessária uma solução numérica iterativa.

Os pesos são definidos por:

$$W_{i\beta} = \frac{\psi[(Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}})/s]}{(Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}})/s}, \quad i = 1, 2, \dots, n, \quad (2.26)$$

e, ao ponderar cada termo da Equação 2.25 pelos pesos $W_{i\beta}$ (definidos como 1 quando o resíduo é exatamente zero), obtemos:

$$\sum_{i=1}^n x_{ij} w_{i\beta} (Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}) = 0, \quad i = 0, 1, 2, \dots, k, \quad (2.27)$$

ou

$$\sum_{i=1}^n x_{ij} w_{i\beta} \mathbf{x}'_i \hat{\boldsymbol{\beta}} = \sum_{i=1}^n x_{ij} w_{j\beta} Y_i, \quad j = 0, 1, 2, \dots, k, \quad (2.28)$$

que podemos escrever na forma de matriz como

$$\mathbf{X}' \mathbf{W}_\beta \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{W}_\beta \mathbf{Y}, \quad (2.29)$$

em que $\mathbf{W}_\beta = \text{diag}(w_1, w_2, \dots, w_n)$. Estas equações são obviamente da forma de equações normais de mínimos quadrados ponderados. Mais precisamente, elas são equações do tipo mínimos quadrados ponderados porque $\mathbf{W}_\beta$ é uma matriz diagonal de pesos. A dificuldade em resolvê-las é que $\mathbf{W}_\beta$ depende de $\boldsymbol{\beta}$. Isto é contornado por meio de um processo estatístico tradicional. Para isso, de alguma forma, obtemos uma estimativa inicial $\hat{\boldsymbol{\beta}}_0$ de $\boldsymbol{\beta}$ (talvez usando mínimos quadrados), calculamos o valor $\mathbf{W}_0$, digamos, de $\mathbf{W}_\beta$ para aquele $\hat{\boldsymbol{\beta}}_0$ e resolvemos a equação 2.29 para obter a solução $\hat{\boldsymbol{\beta}}_1$. Usando $\hat{\boldsymbol{\beta}}_1$ em $\mathbf{W}_\beta$ obtemos $\mathbf{W}_1$. Então $\mathbf{W}_1$ é usado para obter $\hat{\boldsymbol{\beta}}_2$ e assim por diante. Tipicamente, a convergência ocorre rapidamente. O procedimento também pode ser interrompido após um número selecionado de etapas.

Podemos escrever a solução iterativa como:

$$\hat{\boldsymbol{\beta}}_{q+1} = (\mathbf{X}' \mathbf{W}_q \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}_q \mathbf{Y}, \quad q = 0, 1, \dots \quad (2.30)$$

em que, $\hat{\boldsymbol{\beta}}_{q+1}$ é o estimador do vetor de coeficientes na $(q+1)$ -ésima iteração; $\mathbf{X}$ é a matriz com dimensões $n \times (p+1)$, em que n é o número de observações e p é o número de variáveis independentes; $\mathbf{W}_q$ é a matriz de pesos na iteração q . Como é uma matriz diagonal, ela contém os pesos para cada observação na regressão. $\mathbf{W}_q$ é uma função dos coeficientes estimados na iteração anterior; $\mathbf{Y}$ é o vetor das observações da variável dependente de dimensão $n \times 1$; $(\mathbf{X}' \mathbf{W}_q \mathbf{X})^{-1}$ é a inversa da matriz resultante da multiplicação de $\mathbf{X}'$ (a transposta de $\mathbf{X}$), $\mathbf{W}_q$, e $\mathbf{X}$. Essa parte da equação ajusta o impacto das observações ponderadas na estimativa dos coeficientes; $\mathbf{X}' \mathbf{W}_q \mathbf{Y}$ é a multiplicação de $\mathbf{X}'$ (a transposta de $\mathbf{X}$) com $\mathbf{W}_q$ e $\mathbf{Y}$, que é uma forma de calcular a "soma ponderada" dos resíduos.

O procedimento iterativo ocorre da seguinte forma:

- **Início:** começa com uma estimativa inicial $\hat{\boldsymbol{\beta}}_0$, usualmente obtida por mínimos quadrados ordinários.
- **Iteração:** a cada passo q , os pesos W_q são recalculados, e a nova estimativa $\hat{\boldsymbol{\beta}}_{q+1}$ é obtida.

- **Convergência:** o processo converge quando as diferenças entre estimativas consecutivas forem pequenas o suficiente (por exemplo, abaixo de 0,1% ou 0,01%).

A Tabela 25.1 de Drapper & Smith (1998) (p. 532) apresenta diferentes escolhas para $\rho(u)$, ilustrando a variação da função peso $w(u)$. Além disso, a Figura 25.1 (p. 533), do mesmo livro, exibe os gráficos dessas funções para $u \geq 0$, destacando a influência da escolha do fator de escala.

A estimação robusta, embora vantajosa, apresenta desafios, como a dependência dos pesos dos parâmetros estimados, a escolha adequada da função de peso e a sensibilidade das estimativas à especificação do fator de escala (Drapper; Smith, 1998). Contudo, sua rápida convergência e a possibilidade de implementação por meio de mínimos quadrados ponderados tornam uma alternativa eficaz aos mínimos quadrados ordinários. A aplicação da estimação robusta é particularmente relevante para os modelos que discutiremos a seguir.

No entanto, essa abordagem não leva em consideração a autocorrelação espacial. Como nosso foco está em modelos que incorporam esse tipo de dependência espacial, a seguir, abordaremos os modelos de regressão geograficamente ponderada, bem como sua versão robusta.

2.6 Regressão Geograficamente Ponderada (GWR)

A análise de dados espaciais frequentemente nos indica que as relações entre variáveis não são constantes em toda a área de estudo, um fenômeno conhecido como não-estacionariedade espacial. Modelos de regressão globais, como os Mínimos Quadrados Ordinários (OLS), assumem estacionariedade e podem mascarar variações locais importantes, produzindo um único conjunto de parâmetros que representa uma média para toda a região.

Para superar essa limitação, a Regressão Geograficamente Ponderada (GWR) foi desenvolvida como uma técnica de modelagem local. A GWR estende a estrutura da regressão tradicional, permitindo que os parâmetros sejam estimados em cada localização, em vez de globalmente. O método utiliza uma janela móvel ponderada, onde observações mais próximas ao ponto de estimação recebem maior peso, permitindo que as relações entre as variáveis mudem continuamente no espaço.

2.6.1 O modelo GWR

A forma geral do modelo GWR é dada por Fotheringham, Brunson & Charlton (2002):

$$Y_i = \beta_0(u_i, v_i) + \sum_{k=1}^m \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (2.31)$$

em que:

- Y_i é a variável dependente na localização i ;
- (u_i, v_i) são as coordenadas geográficas da observação i ;
- x_{ik} é a k -ésima variável independente na localização i ;
- $\beta_0(u_i, v_i)$ é o intercepto local;
- $\beta_k(u_i, v_i)$ é o coeficiente local para a k -ésima variável;
- ε_i representa o termo de erro aleatório.

Os coeficientes locais são estimados por meio de mínimos quadrados ponderados, cuja expressão matricial é dada por:

$$\hat{\beta}(u_i, v_i) = \left(\mathbf{X}^\top \mathbf{G}(u_i, v_i) \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{G}(u_i, v_i) \mathbf{Y} \quad (2.32)$$

em que $\mathbf{G}(u_i, v_i)$ é a matriz diagonal de pesos geográficos para a localização i , que é fundamental para o modelo. O peso de cada observação é determinado por uma função kernel e, conseqüentemente, pela largura de banda *bandwidth* que define o alcance espacial da ponderação.

2.6.2 Calibração do Modelo e Seleção do *Bandwidth*

A escolha do *bandwidth* é uma etapa crítica pois define o grau de suavização das superfícies dos coeficientes. Valores muito grandes aproximam os resultados do GWR aos de um modelo OLS global, enquanto valores muito pequenos podem gerar estimativas instáveis. A otimização do *bandwidth* é geralmente realizada por meio da minimização do Critério de Informação de Akaike corrigido (AICc), dado por:

$$AIC_c = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n \left(\frac{n + \text{tr}(\mathbf{S})}{n - 2 - \text{tr}(\mathbf{S})} \right), \quad (2.33)$$

em que n é o tamanho da amostra, $\hat{\sigma}$ é a estimativa do desvio padrão do erro, e $\text{tr}(\mathbf{S})$ é o traço da matriz chapéu ($\mathbf{S}$), que atua como uma medida da complexidade do modelo, representando o número efetivo de parâmetros (Fotheringham; Brunson; Charlton, 2002).

2.6.3 Teste de Não-Estacionariedade Espacial

Para justificar o uso do GWR em detrimento de um modelo global, é necessário testar a significância da variação espacial dos coeficientes. A hipótese nula (H_0) é que os coeficientes são estacionários (constantes no espaço), enquanto a hipótese alternativa (H_1) é que eles variam espacialmente. Um método comum para isso é o teste de Monte Carlo, que compara a variância dos coeficientes locais observados com uma distribuição de variâncias geradas a partir de permutações aleatórias da localização dos dados Sulekan & Jamaludin (2020).

2.6.4 Regressão Geograficamente Ponderada Robusta (RGWR)

O modelo GWR padrão é sensível a presença *outliers* que podem distorcer as estimativas locais. Para mitigar esse problema, utiliza-se a Regressão Geograficamente Ponderada Robusta (RGWR). Conforme descrito por Fotheringham, Brunson e Charlton (2002), a abordagem mais comum é um procedimento iterativo que, em vez de remover os *outliers*, reduz o peso das observações com resíduos elevados.

O processo consiste em ajustar um GWR padrão, calcular os resíduos e, em seguida, calcular um "peso de robustez" ($w_{r,j}$) para cada observação. O modelo é então reajustado utilizando um peso final que é o produto do peso geográfico pelo peso de robustez, repetindo o processo até a convergência dos coeficientes. A matriz de pesos final do RGWR é detalhada na seção de Metodologia.

2.6.5 Regressão Geograficamente Ponderada (GWR)

A Regressão Geograficamente Ponderada (GWR) é uma extensão da regressão linear que permite estimar coeficientes locais, capturando variações espaciais nas relações entre variáveis dependentes e independentes. Dessa forma, a GWR permite identificar a não-estacionariedade espacial por meio do mapeamento desses coeficientes locais (Fotheringham; Charlton; Brunsdon, 1998). Segundo Bivand, Pebesma & Rubio (2008), a GWR utiliza uma janela móvel ponderada para estimar esses coeficientes localmente.

A forma geral do modelo GWR pode ser expressa como:

$$Y_i = \beta_0(u_i, v_i) + \sum_{k=1}^m \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \quad (2.34)$$

em que Y_i é a variável dependente na localização i ; X_{ik} é a k -ésima variável independente; (u_i, v_i) são as coordenadas geográficas da observação i ; $\beta_0(u_i, v_i)$ é o intercepto local; $\beta_k(u_i, v_i)$ são os coeficientes locais associados às variáveis independentes; e ε_i representa o termo de erro aleatório.

Como os dados são ponderados geograficamente, observações mais próximas têm mais influência na estimativa do conjunto local de coeficientes de regressão do que observações mais distantes. O modelo mede as características fundamentais ao redor de cada ponto de regressão i , onde cada conjunto de coeficientes de regressão é estimado por um método de mínimos quadrados ponderados.

A expressão matricial para essa estimativa é:

$$\hat{\boldsymbol{\beta}}_i = \left(\mathbf{X}^\top \mathbf{G}(u_i, v_i) \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{G}(u_i, v_i) \mathbf{Y} \quad (2.35)$$

onde $\mathbf{X}$ é a matriz das variáveis independentes com uma coluna de 1s para o intercepto; $\mathbf{Y}$ é o vetor da variável dependente; $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i0}, \dots, \hat{\beta}_{im})^\top$ é o vetor dos $m + 1$ coeficientes estimados de regressão locais; e $\mathbf{G}_i$ é a matriz diagonal que denota o peso geográfico de cada dado observado para o ponto de regressão i na localização (u_i, v_i) . O peso geográfico de cada observação é determinado por uma função de núcleo, que define como é a influência de cada ponto e esta diminui com a distância. A escolha da função de núcleo e da largura de banda (alcance espacial) é fundamental para o desempenho do modelo GWR.

A largura de banda ideal para a regressão GWR é determinada pela minimização de um critério de ajuste do modelo, como a validação cruzada *leave-one-out* (CV) ou o Critério de Informação de Akaike corrigido (AICc), que considera a parcimônia do modelo e leva em conta o tamanho da amostra. Na prática, o AICc penaliza mais severamente modelos com larguras de banda menores em comparação com aqueles que possuem larguras de banda maiores.

O Critério de Informação de Akaike (AIC) é utilizado para selecionar, dentre um conjunto de modelos, aquele que melhor equilibra o ajuste aos dados e a complexidade do modelo, minimizando a perda de informação em relação à realidade subjacente (Carvalho, 2024). Para o cálculo do AICc, é necessário definir uma matriz de projeção (S), que é a matriz de projeção que mapeia os valores observados y para os valores ajustados $\hat{y}$ (Hoaglin; Welsch, 1978).

A matriz de projeção S para o modelo de regressão linear é definida como

$$S = X(X'X)^{-1}X'. \quad (2.36)$$

Essa matriz é utilizada para calcular os valores ajustados em relação aos valores observados. Em seguida, apresenta-se a matriz de projeção para o caso do modelo GWR.

No contexto da GWR, devido à ponderação espacial, cada linha r_i dessa matriz é dada por

$$r_i = X_i (X^T G_i X)^{-1} X^T G_i. \quad (2.37)$$

em que X é a matriz de delineamento contendo as variáveis explicativas, G_i é a matriz de ponderação diagonal para o ponto i , X^T é a transposta de X e $(X^T G_i X)^{-1}$ é a inversa da matriz ponderada. Essa matriz é fundamental para calcular os valores ajustados $\hat{Y}$ no contexto da GWR, considerando as ponderações espaciais específicas de cada localidade.

Assim, o AICc para uma regressão GWR com largura de banda d pode ser calculado por:

$$AIC_c(d) = 2n \ln(\hat{\sigma}) + n \ln(2\pi) + n \left(\frac{n + \text{tr}(S)}{n - 2 - \text{tr}(S)} \right) \quad (2.38)$$

em que n é o tamanho da amostra (local) (de acordo com d); $\hat{\sigma}$ é a estimativa do desvio padrão do erro; e $\text{tr}(S)$ denota o traço da matriz S , que é a soma dos elementos da diagonal principal da matriz.

2.6.6 Regressão Geograficamente Ponderada Robusta (RGWR)

A Regressão Geograficamente Ponderada Robusta (RGWR) é uma extensão da GWR tradicional que foi desenvolvida para minimizar a influência de valores discrepantes nas estimativas locais dos coeficientes, o que é fundamental em estudos espaciais, como a análise da incidência da dengue.

Segundo Fotheringham, Brunson & Charlton (2002), há duas formas principais de aumentar a robustez da GWR. A primeira é a filtragem de observações com resíduos padronizados externos elevados, reestimando o modelo sem esses pontos atípicos, que podem distorcer as relações espaciais. A segunda abordagem, mais automatizada, reduz iterativamente o peso de observações discrepantes, permitindo que o modelo se ajuste de forma gradual, sem excluir completamente informações possivelmente relevantes.

Essas estratégias tornam a modelagem espacial mais precisa, garantindo que padrões regionais distintos sejam corretamente identificados, sem comprometer a estabilidade do modelo.

Nesta pesquisa, será adotado a segunda abordagem, que consiste em atribuir pesos menores às observações com resíduos elevados, ao invés de removê-las. Essa estratégia permite que as observações atípicas permaneçam no modelo, mas com menor impacto, garantindo maior controle sobre sua influência e evitando a perda de informações importantes.

Formalmente, o resíduo *estudentizado externo* padronizado externo global para cada localização i é dado por:

$$r_i = \frac{e_i}{\hat{\sigma}_{-i} \sqrt{q_{ii}}}$$

em que $e_i = z_i - \hat{z}_i$ representa o erro de predição na localização i , $\hat{\sigma}_{-i}$ é a estimativa *leave-one-out* da variância residual global, e q_{ii} é o i -ésimo elemento da matriz $(I - S)(I - S)^\top$, que mede a influência da observação i na estimativa.

Observações com $|r_i| > 3$ são consideradas discrepantes e filtradas, conforme o critério comum na literatura para identificação de outliers.

Além do estimador global, um estimador local pode ser obtido substituindo $\hat{\sigma}_{-i}$ por uma estimativa local $\hat{\sigma}_{-i}(u_i, v_i)$, calculada por meio de um suavizador de médias dos resíduos ao longo do espaço. Isso permite identificar *outliers* de forma mais contextualizada, levando em conta a distribuição espacial dos resíduos.

A abordagem robusta utilizada nesta pesquisa visa garantir que as estimativas locais dos coeficientes GWR sejam menos sensíveis a observações atípicas, proporcionando maior estabilidade nas inferências espaciais.

No segundo modelo robusto, a função de ponderação g_r reduz de forma iterativa o peso das observações com resíduos elevados, sendo definida como:

$$g_r(e_i) = \begin{cases} 1, & \text{se } |e_i| \leq 2\hat{\sigma} \\ \left[1 - (|e_i| - 2\hat{\sigma})^2\right]^2, & \text{se } 2\hat{\sigma} < |e_i| < 3\hat{\sigma} \\ 0, & \text{caso contrário} \end{cases}$$

em que $g_r(e_i)$ é uma função de ponderação que depende do resíduo e_i . Nesta função:

- $g_r(e_i)$ representa o peso atribuído à observação i no reajuste do modelo, ajustando sua influência de acordo com a magnitude do resíduo.
- $e_i = y_i - \hat{y}_i$ é o resíduo da observação i , ou seja, a diferença entre o valor observado y_i e o valor previsto $\hat{y}_i$ pelo modelo inicial.
- $2\hat{\sigma}$ corresponde a duas vezes o desvio padrão estimado dos resíduos, delimitando o intervalo dentro do qual os resíduos são considerados próximos do esperado.
- $3\hat{\sigma}$ corresponde a três vezes o desvio padrão estimado dos resíduos, delimitando o ponto a partir do qual as observações são consideradas *outliers* extremos e recebem peso zero.

Esta função fornece pesos que são utilizados em um reajuste do modelo. O modelo reajustado reduz a influência das observações cujos resíduos se encontram fora do intervalo esperado.

Quando o resíduo ajustado está entre $-2\hat{\sigma}$ e $2\hat{\sigma}$, os dados daquela observação recebem peso unitário (1). Portanto, se o modelo não apresenta resíduos superiores a duas vezes o desvio padrão, o modelo reajustado será praticamente equivalente ao modelo original.

Para resíduos situados entre $2\hat{\sigma}$ e $3\hat{\sigma}$, o peso decresce gradualmente, reduzindo a influência dessas observações no reajuste. Finalmente, se um resíduo exceder três desvios padrão ($|e_i| > 3\hat{\sigma}$), ele é completamente desconsiderado no ajuste, caracterizando-se como um *outlier* extremo. A definição desses limites é uma escolha do pesquisador, o que introduz um elemento de subjetividade no controle sobre a influência de *outliers* no modelo.

O método consiste em aplicar inicialmente um ajuste preliminar para identificar resíduos elevados e, em seguida, recalibrar o modelo, diminuindo a influência dessas observações.

No contexto da Regressão Geograficamente Ponderada (GWR), cada ajuste de regressão local já incorpora pesos definidos pela função kernel; portanto, o peso existente deve ser multiplicado por $g_r(e_i)$ durante o ajuste. Este procedimento pode ser repetido por diversas iterações: após cada ajuste ponderado, os resíduos são recalculados, os pesos atualizados e o modelo reajustado utilizando os novos pesos.

Esse processo fornece um mecanismo automatizado para lidar com outliers na calibração de modelos GWR. Computacionalmente, apresenta sobrecarga significativamente menor do que a abordagem baseada em resíduos estudentizados, uma vez que cada ciclo requer apenas o cálculo de n resíduos, sem a necessidade de manipular matrizes $n \times n$. Em muitos casos, uma única iteração é suficiente para mitigar as distorções mais severas causadas por outliers.

2.7 Matriz de Pesos do Modelo RGWR

A Regressão Geograficamente Ponderada Robusta (RGWR) é um procedimento iterativo que ajusta o modelo GWR padrão para reduzir a influência de observações atípicas (*outliers*). Isso é alcançado por meio de uma matriz de pesos final, aqui denotada como $\tilde{\mathbf{G}}(i)$ que é o resultado do produto elementar da matriz de pesos geográficos pela matriz de pesos de robustez.

Componente 1: Peso geográfico (Kernel espacial)

O primeiro componente é o peso geográfico padrão, g_{ij} , que atribui maior influência às observações (j) mais próximas do ponto de regressão (i). Neste trabalho, utiliza-se a função de kernel bi-quadrada, definida por (Fotheringham; Charlton; Brunson, 1998) como:

$$g_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{b} \right)^2 \right]^2, & \text{se } d_{ij} < b, \\ 0, & \text{caso contrário,} \end{cases} \quad (2.39)$$

$$d_{ij} = \sqrt{(t_i - t_j)^2 + (s_i - s_j)^2}, \quad (2.40)$$

onde $i, j = 1, 2, \dots, n$, d_{ij} representa a distância Euclidiana entre os locais de observação i e j , definidos pelas coordenadas espaciais (t_i, s_i) e (t_j, s_j) , e b denota a largura de banda do kernel (Gollini *et al.*, 2015; Rahman, 2024).

Componente 2: Peso de robustez (Resíduos)

O segundo componente é o peso de robustez, $g_{r,j}$, projetado para reduzir a influência de *outliers* com base nos resíduos (e_j) do ajuste da iteração anterior. A função de ponderação descrita por Fotheringham *et al.* (2002) opera em três estágios para desponderar observações com resíduos grandes:

$$g_{r,j} = \begin{cases} 1, & \text{se } |e_j| \leq 2\hat{\sigma}; \\ \left[1 - (|e_j| - 2)^2\right]^2, & \text{se } 2\hat{\sigma} < |e_j| < 3\hat{\sigma}; \\ 0, & \text{caso contrário.} \end{cases} \quad (2.41)$$

Estrutura da matriz de pesos final

A combinação dos dois pesos resulta em um peso final, $\tilde{g}_{ij}$, que compõe a diagonal da matriz $\tilde{\mathbf{G}}(i)$. A matriz tem a seguinte forma:

$$\tilde{\mathbf{G}}(i) = \begin{pmatrix} g_{i1} \cdot g_{r,1} & 0 & \cdots & 0 \\ 0 & g_{i2} \cdot g_{r,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & g_{in} \cdot g_{r,n} \end{pmatrix} \quad (2.42)$$

em que o cálculo de cada elemento da diagonal é dado pela função por partes que combina as equações 2.39 e 2.41:

$$\tilde{g}_{ij} = \begin{cases} \left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2, & \text{se } d_{ij} < h \text{ e } |e_j| \leq 2\hat{\sigma} \\ \left(1 - \left(\frac{d_{ij}}{h}\right)^2\right)^2 \cdot [1 - (|e_j| - 2)^2]^2, & \text{se } d_{ij} < h \text{ e } 2\hat{\sigma} < |e_j| < 3\hat{\sigma} \\ 0, & \text{caso contrário} \end{cases} \quad (2.43)$$

2.7.1 Estimação dos Parâmetros

Para cada localização i , os coeficientes locais do modelo RGWR são estimados pela solução do seguinte problema de minimização:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \arg \min_{\boldsymbol{\beta}} \sum_{j=1}^n \tilde{g}_{ij} \left(y_j - \mathbf{x}_j^\top \boldsymbol{\beta} \right)^2, \quad (2.44)$$

em que y_j é o valor observado da variável dependente (incidência da dengue por 100 mil hab) no município j , $\mathbf{x}_j$ é o vetor de variáveis explicativas do município j e $\tilde{g}_{ij}$ é o peso combinado espacial e robusto entre o ponto focal i e a observação j .

Condicionalmente aos pesos da iteração corrente, a solução do problema em (2.44) pode ser expressa na forma matricial como:

$$\hat{\boldsymbol{\beta}}(u_i, v_i) = \left(\mathbf{X}^\top \tilde{\mathbf{G}}(i) \mathbf{X} \right)^{-1} \mathbf{X}^\top \tilde{\mathbf{G}}(i) \mathbf{y}, \quad (2.45)$$

em que $\mathbf{X}$ é a matriz das variáveis explicativas e $\mathbf{y}$ é o vetor da variável resposta.

Como os pesos robustos dependem dos resíduos estimados, o procedimento é realizado por meio do método dos Mínimos Quadrados Reponderados Iterativamente (*Iteratively Reweighted Least Squares – IRLS*) (Chen *et al.*, 2014). O algoritmo inicia-se com os resíduos obtidos a partir do modelo GWR clássico e é repetido até que a variação dos coeficientes entre iterações sucessivas seja inferior a um critério de convergência pré-estabelecido (Fotheringham; Charlton; Brunsdon, 1998; Harris *et al.*, 2010).

Para avaliar a significância estatística local dos coeficientes estimados pelo modelo RGWR, foram calculados os valores- t e os valores- p . Os valores- t correspondem à razão en-

tre os coeficientes locais estimados e seus respectivos erros padrão, sendo interpretados como *pseudo valores-t* no contexto da regressão geograficamente ponderada. Esses valores fornecem uma medida indicativa da evidência estatística de que os efeitos locais diferem de zero (Lu *et al.*, 2014). Os valores-*p* expressam a probabilidade de se obter um coeficiente local tão extremo quanto o estimado sob a hipótese nula de efeito nulo, sendo derivados dos *pseudo valores-t*. Em modelos GWR, esses valores devem ser ajustados para testes múltiplos, uma vez que são estimados simultaneamente para diversas localizações espaciais (Lu *et al.*, 2014).

A seguir, será descrito o critério para seleção do modelo ajustado aos dados, assim como suas definições.

2.8 Critérios para avaliação dos modelos

A avaliação da qualidade do ajuste e da adequação do modelo proposto é uma etapa essencial na análise estatística. Para isso, serão utilizadas diferentes métricas, que permitem comparar modelos e avaliar seu desempenho. Os critérios para a seleção do melhor modelo serão:

- O Critério de Informação de Akaike (**AIC**) que fornece uma medida relativa da qualidade do ajuste do modelo e penaliza sua complexidade. Este critério auxilia na seleção do modelo que melhor equilibra o ajuste aos dados e a parcimônia (Carvalho, 2024);
- O Critério de Informação de Akaike Corrigido (**AICc**) é uma versão ajustada do AIC que incorpora uma correção para amostras pequenas. Esse critério é particularmente útil quando o tamanho da amostra é limitado em relação ao número de parâmetros estimados (Fotheringham; Charlton; Brunson, 1998);
- O Critério de Informação Bayesiano (**BIC**), também conhecido como Critério de Informação de Schwartz (SCI), é uma estatística utilizada para avaliar o ajuste de modelos estatísticos. Ele é definido em função da verossimilhança do modelo (L), do número de parâmetros (k) e do tamanho da amostra (n) (Saefuddin; Setiabudi; Fitrianto, 2012). Assim como o AIC, valores menores de BIC indicam modelos mais adequados. No entanto, o BIC penaliza modelos com mais variáveis explicativas e tamanhos amostrais maiores de forma mais rigorosa do que o AIC e, assim, favorecendo modelos mais parcimoniosos (Saefuddin; Setiabudi; Fitrianto, 2012).

O Critério de Informação de Akaike (**AIC**) é uma medida utilizada para comparar modelos estatísticos, balanceando a qualidade do ajuste e a complexidade do modelo. Ele é definido por Akaike () como:

$$AIC = -2\ln(L) + 2a \quad (2.46)$$

em que:

- L é a verossimilhança do modelo,
- a é o número de parâmetros estimados no modelo.

O **AICc** para uma regressão GWR com largura de banda d pode ser calculado por (Fotheringham; Charlton; Brunson, 1998) como:

$$AIC_c(d) = 2n\ln(\hat{\sigma}) + n\ln(2\pi) + n\left(\frac{n + \text{tr}(\mathbf{S})}{n - 2 - \text{tr}(\mathbf{S})}\right) \quad (2.47)$$

em que n é o tamanho da amostra (local) (de acordo com d); $\hat{\sigma}$ é a estimativa do desvio padrão do erro; e $\text{tr}(\mathbf{S})$ denota o traço da matriz $\mathbf{S}$, que é a soma dos elementos da diagonal principal da matriz de projeção a matriz *hat*, que é uma função da largura de banda (*bandwidth*).

O **BIC** (Critério de Informação Bayesiano) é definido por Saefuddin, Setiabudi & Fitrianto (2012) como:

$$BIC = -2\ln(L) + a\ln(n) \quad (2.48)$$

em que a representa o número de parâmetros do modelo, n é o número de observações e L é a função de verossimilhança do modelo ajustado. O BIC penaliza modelos mais complexos para evitar o sobreajuste, sendo útil para seleção de modelos estatísticos. As métricas de erro quadrático médio (RMSE) e coeficiente de determinação (R^2) foram utilizadas de forma complementar, com o objetivo de avaliar o desempenho preditivo e a qualidade do ajuste dos modelos selecionados pelos critérios de informação AIC, AICc e BIC.

A utilização conjunta destes critérios proporcionará uma avaliação mais abrangente do desempenho dos modelos, permitindo identificar o ajuste mais adequado para os dados em análise.

3 MATERIAL E MÉTODOS

Nesta seção serão descritos os materiais utilizados, incluindo área de estudo, bases de dados, fontes de informação e ferramentas computacionais empregadas na análise.

3.1 Área de estudo

Na Figura 3.1 é apresentada a área de estudo do presente trabalho, destacando os municípios e as mesorregiões do Estado de Minas Gerais, Brasil.

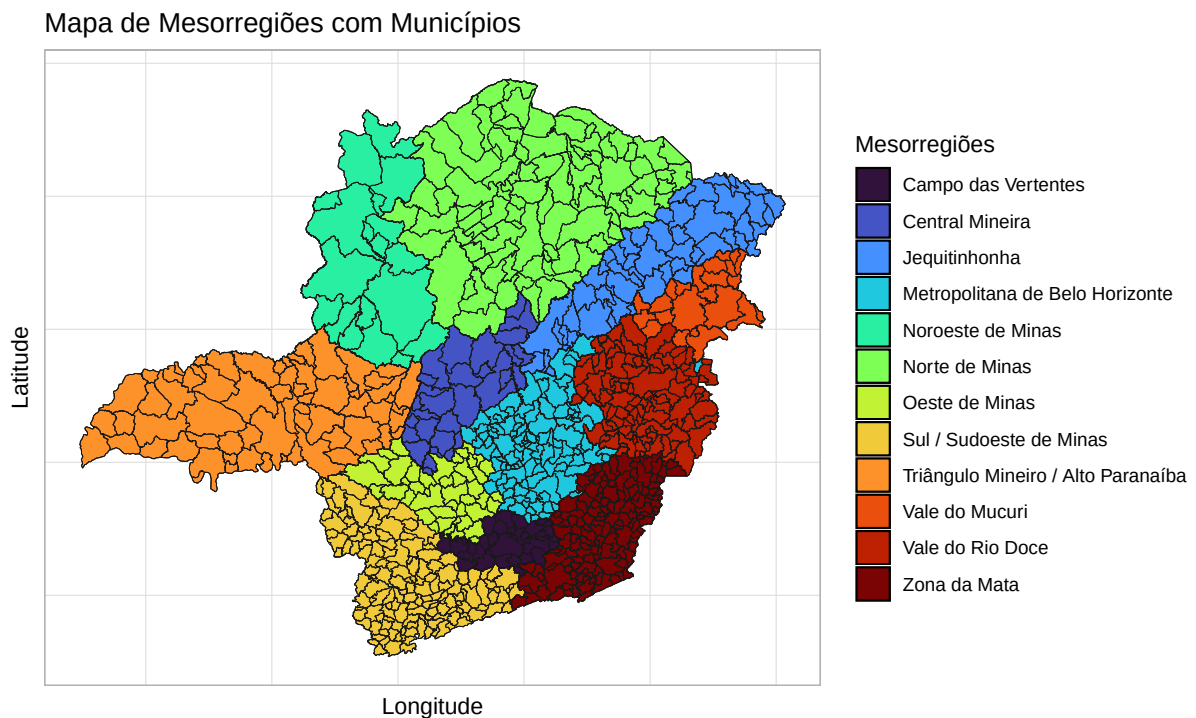


Figura 3.1 – Área de estudo sobre a incidência de dengue em Minas Gerais
Fonte: Autor (2025)

A Figura 3.1 detalha a área de estudo por mesorregiões do Estado de Minas Gerais, em que os municípios estão delimitados por seus respectivos polígonos. Essa representação permite visualizar a distribuição espacial dos municípios dentro de cada mesorregião, facilitando análises em nível regional.

3.2 Matriz de vizinhança espacial

A matriz de vizinhança espacial do tipo **rainha (Queen)** considera que duas regiões são vizinhas se compartilharem *uma borda ou um vértice* (Almeida; Oliveira, 2022). Qualquer município que toque outro, seja por uma fronteira comum ou por um vértice, é considerado vizinho. Este tipo de matriz é amplamente utilizado em análises espaciais para capturar a dependência espacial entre regiões.

A Figura 3.2 apresenta a matriz de vizinhança do tipo rainha (Queen) para os municípios de Minas Gerais na qual cada município é considerado vizinho daqueles com os quais compartilha bordas ou vértices

Matriz de Vizinhança do tipo (Queen)

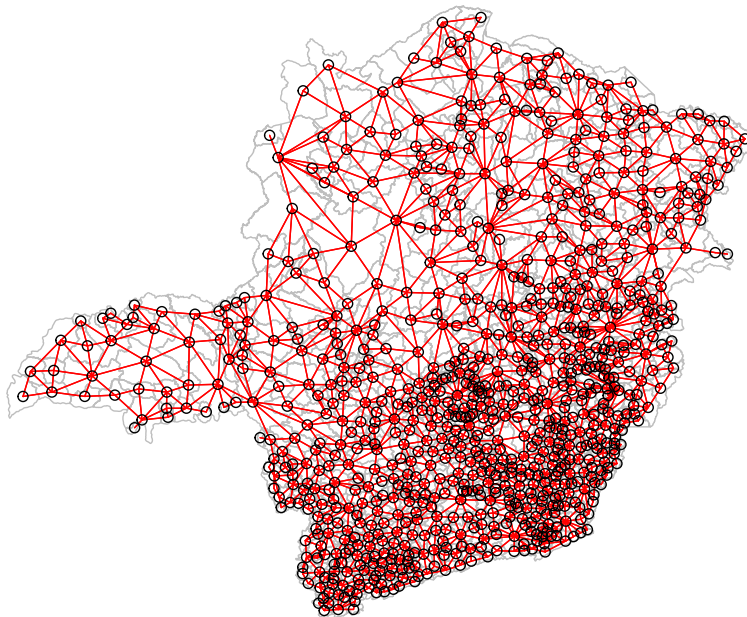


Figura 3.2 – Matriz de vizinhança do tipo rainha (Queen) para os municípios de Minas Gerais.
Fonte: Autor (2025)

3.3 Descrição dos dados e ferramentas computacionais

Neste estudo é analisada a incidência de dengue no Estado de Minas Gerais, utilizando dados obtidos por meio de solicitação formal à Secretaria de Saúde do Estado, via Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC), sob o protocolo nº 01320000248202404, no dia 10 de setembro de 2024. A resposta da Secretaria, contendo os dados solicitados, foi recebida em 18 de setembro de 2024, conforme despacho nº 171/2024/SES/SUBVS/SVE/DVDTI (protocolo e-SIC 01320000248202404) (Brasil, 2023).

A principal variável de interesse é a incidência de casos de dengue, enquanto as covariáveis analisadas são descritas a seguir: a urbanização de vias públicas, que representa o percentual de domicílios urbanos situados em vias com boca de lobo, pavimentação, meio-fio e calçada (%); a arborização de vias públicas, correspondendo à proporção de domicílios urbanos localizados em vias com cobertura arbórea (%); o esgotamento sanitário adequado, que indica a proporção da população residente em domicílios particulares permanentes com acesso a redes gerais de esgoto ou fossas sépticas (%); e a área urbanizada, que se refere à extensão territorial ocupada por edificações e infraestrutura urbana, expressa em quilômetros quadrados (km^2) (IBGE, 2024).

O estudo considera quatro variáveis independentes: esgotamento sanitário adequado (em %), urbanização de vias públicas (em %), arborização de vias públicas (em %) e área urbanizada (em km^2), além da variável dependente incidência da dengue.

Após o pré-processamento que incluiu organização e manipulação, os dados foram explorados utilizando estatísticas descritivas e gráficos usuais sem considerar a dependência espacial. Também foi realizadas análises para verificar os pressupostos de normalidade dos dados por meio do teste de Shapiro-Wilk e a detecção de *outliers* utilizando boxplots que identificam valores extremos com base no intervalo interquartil (IQR). Esse método considera como candidatos a *outliers* os valores que estão fora do intervalo: $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ em que $Q1$ e $Q3$ representam, respectivamente, o primeiro e o terceiro quartil dos dados.

A análise de dados foi realizada utilizando funções disponíveis e/ou desenvolvidas no *software* R (TEAM, 2024). Mais especificamente, para o tratamento, organização e análise dos dados foram utilizados as seguintes bibliotecas: tidyverse (Wickham *et al.*, 2019), funções de regressão linear (lm), índice de Moran, modelos espaciais SAR e CAR (Bivand; Piras, 2015; Cliff; Ord, 1981; Bivand; Hauke; Kossowski, 2013), GWR e GWR robusto (Fotheringham;

Charlton; Brunsdon, 1998), funções gráficas como hist, boxplot e plot (Becker; Chambers; Wilks, 1988; Chambers *et al.*, 1983a; Murrell, 2005; Cleveland, 1985), ggplot2 (Wickham *et al.*, 2019), sf (Pebesma *et al.*, 2024), procedimentos de detecção de outliers (Tukey, 1977; McGill; Tukey; Larsen, 1978; Velleman; Hoaglin, 1981; Emerson; Strenio, 1983; Chambers *et al.*, 1983b) e teste de normalidade Shapiro-Wilk (Royston, 1982b; Royston, 1982a; Royston, 1995).

Todas as análises foram realizadas em um notebook Inspiron 15 3511, com sistema operacional Windows 11 de 64 bits.

4 RESULTADOS E DISCUSSÃO

Nesta seção serão apresentados os resultados e discussões das análises exploratórias (estatísticas descritivas e gráficos) das variáveis e covariáveis. A análise exploratória inclui tanto uma abordagem estatística convencional quanto espacial, permitindo investigar as características dos dados, identificar padrões, tendências e possíveis heterogeneidades na incidência de dengue. A seguir são apresentados os resultados e discussões dos ajustes dos modelos de regressão Linear Múltiplo (LM), Linear Múltiplo Robusto (LM-MM), Modelos Espaciais SAR e CAR, além dos Modelos Geograficamente Ponderados GWR (padrão) e RGWR (robusto). Será dada uma atenção especial ao Modelo Geograficamente Ponderado Robusto.

4.1 Análise exploratória usual

O número de casos confirmados de dengue no ano de 2023 no Estado de Minas Gerais foi de 392.265. A Tabela 4.1 apresenta as estatísticas descritivas da variável dependente (incidência por 100 mil habitantes no ano de 2023) e das quatro variáveis independentes (covariáveis) selecionadas para análise.

Tabela 4.1 – Estatísticas descritivas das variáveis

Variável	Mínimo	1º Quartil	Mediana	Média	Máximo
Incidência por 100 mil hab.	0,00	54,1	360,6	1326,7	19099,1
Urbanização de vias públicas (%)	0,00	6,40	21,10	25,77	98,30
Arborização de vias públicas (%)	1,60	39,30	63,30	59,55	100,00
Esgotamento sanitário adequado (%)	0,00	42,10	64,00	59,09	98,40
Área urbanizada (km ²)	0,00	1,060	1,950	5,509	274,040

Na Tabela 4.1 podemos observar que a incidência apresenta uma grande variação, com valores que vão de 0 até 19.099,1 casos por 100 mil habitantes, indicando uma distribuição bastante assimétrica e com presença de valores extremos. A mediana de 360,6 sugere que pelo menos metade dos municípios apresenta incidências relativamente baixas, enquanto a média elevada reflete a influência de valores atípicos.

Entre as covariáveis explicativas nota-se que a urbanização de vias públicas apresenta valores que variam de 0% a 98,3%, com uma mediana de 21,1%, indicando que a maioria dos

municípios possui níveis reduzidos dessa infraestrutura. De forma semelhante, a arborização das vias públicas varia entre 1,6% e 100%, com uma mediana de 63,3%, sugerindo que, em muitos municípios, a presença de vegetação nas vias é relativamente alta. O esgotamento sanitário adequado também apresenta ampla variação com, um mínimo, de 0% e um máximo de 98,4%, e uma mediana de 64%, indicando que há municípios com acesso bastante limitado a essa infraestrutura. A covariável área urbanizada apresenta uma distribuição assimétrica, variando de 0 a 274,04 km², com uma mediana de apenas 1,95 km². Isso sugere que a maioria dos municípios analisados possui uma área urbanizada relativamente pequena, embora existam casos isolados de municípios com áreas significativamente maiores.

Essas características indicam que as variáveis possuem distribuições heterogêneas e podem apresentar diferentes padrões de associação com a incidência de casos, reforçando a necessidade da aplicação de modelos de regressão robustos para melhor compreensão dessas relações.

A Figura 4.1 apresenta um boxplot da variável incidência de casos de dengue no Estado de Minas Gerais no ano de 2023.

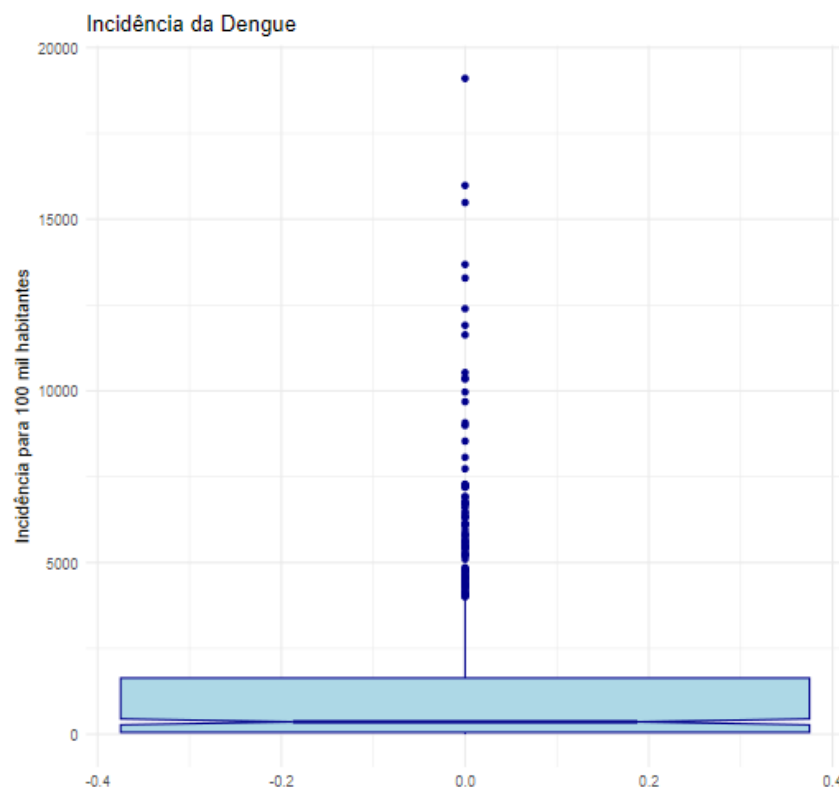


Figura 4.1 – Boxplot da incidência de dengue para o ano de 2023

Fonte: Autor (2025)

Na Figura 4.1 é possível identificar valores discrepantes no conjunto de dados. A análise foi realizada utilizando a função *boxplot.stats* do pacote *outliers*, permitindo a detecção de 87 valores considerados outliers, correspondendo a aproximadamente 10,20% do total de observações.

A Figura 4.2 apresenta o diagrama de dispersão que mostra a relação entre a variável dependente e as quatro variáveis independentes (covariáveis).

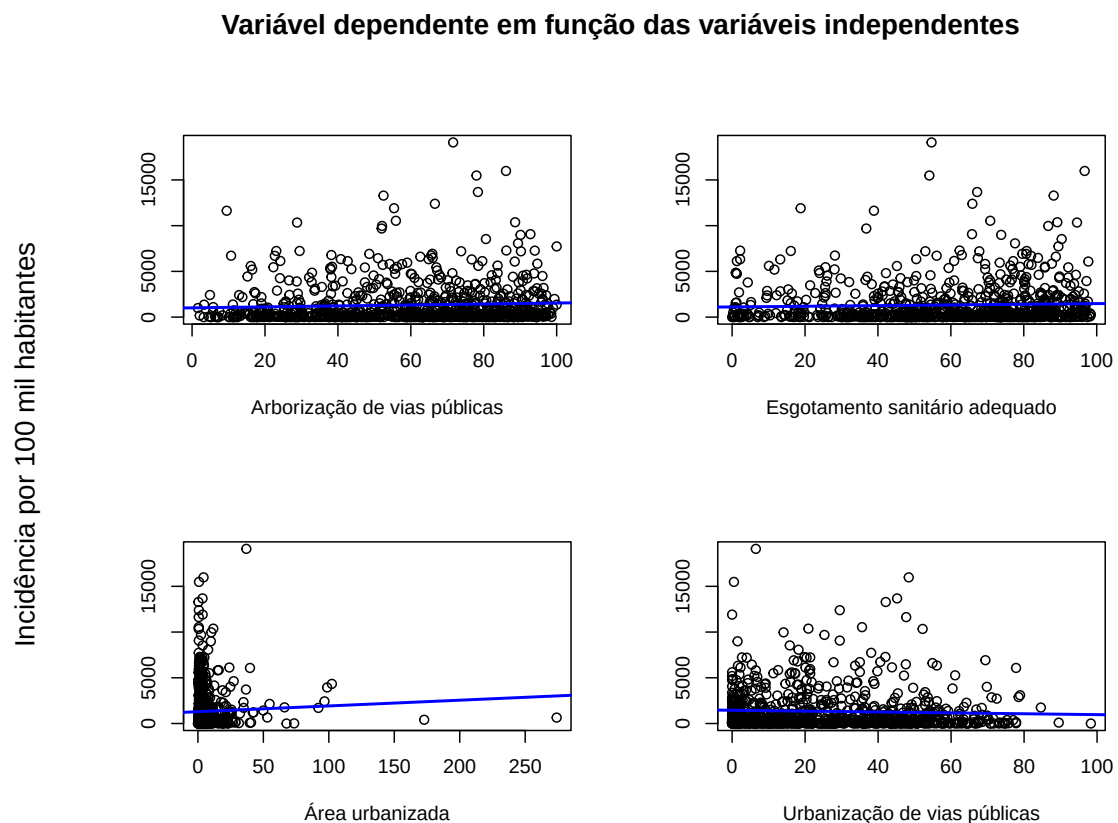


Figura 4.2 – Distribuição da variável dependente em função das covariáveis para o ano de 2023
Fonte: Autor (2025)

A Figura 4.2 apresenta uma análise preliminar da relação entre as covariáveis e a incidência de dengue por 100 mil habitantes. Os resultados indicam ausência de um padrão claro de associação, uma vez que os pontos apresentam alta dispersão em todos os gráficos, sugerindo a inexistência de tendência linear evidente. A linha de tendência ajustada corrobora essa observação, evidenciando efeito mínimo das variáveis explicativas sobre a incidência, com inclinações pouco acentuadas.

Além disso, a presença de *outliers* pode influenciar a interpretação dos resultados, reforçando a necessidade de abordagens mais robustas, como modelos de regressão robusta, para avaliar com maior precisão a relação entre as variáveis e identificar possíveis padrões subjacentes.

4.2 Modelo de regressão linear múltipla

Em qualquer modelagem estatística recomenda-se iniciar pelo modelo mais simples e conhecido antes de tentar modelos mais complexos. Assim, nesta seção apresentamos o ajuste do modelo de regressão linear múltipla e as análises de adequabilidade de ajuste deste modelo.

4.2.1 Ajuste do modelo

Na Tabela 4.2 são apresentados os resultados do modelo de regressão linear múltipla ajustado. Este modelo busca explicar a relação existente entre as variáveis independentes e dependente, permitindo avaliar o impacto de cada fator sobre o fenômeno em questão, considerando que não há dependência espacial.

Tabela 4.2 – Resultados da Regressão Linear Múltipla (OLS)

Variável	Coef.	Erro Padrão	Valor-t	p-valor
Intercepto	926,37	265,41	3,490	< 0,001***
Arborização de vias públicas	4,52	3,11	1,457	0,146
Esgotamento sanitário adequado	4,75	3,24	1,463	0,144
Área urbanizada	3,91	5,33	0,733	0,464
Urbanização de vias públicas	-6,63	3,82	-1,735	0,083
Coeficiente de determinação: $R^2 = 0,92\%$				

Nota: *** $p < 0,001$; ** $p < 0,01$; * $p < 0,05$.

A Tabela 4.2 mostra que o modelo apresentou um coeficiente de determinação muito baixo (R^2 de 0,92%) indicando que menos de 1% da variabilidade da incidência da dengue é explicada pelas variáveis independentes incluídas nos modelos.

Adicionalmente, nenhuma das covariáveis analisadas apresentou significância estatística ao nível de 5%, sugerindo que essas variáveis, não têm poder explicativo relevante sobre a incidência da doença.

4.2.2 Análise da adequabilidade do ajuste

Nesta seção são apresentadas análises estatísticas para verificação dos pressupostos de normalidade dos erros.

Análise da normalidade dos erros

A Tabela 4.3 apresenta o teste de normalidade de Shapiro-Wilk que foi utilizado nos resíduos para verificar o pressuposto de normalidade dos erros do modelo.

Tabela 4.3 – Resultado do teste de normalidade nos resíduos (Shapiro-Wilk)

Teste	Estatística W	p-valor
Shapiro-Wilk	0,6485	2,2e-16

A Tabela 4.3 mostra que o teste de normalidade de Shapiro-Wilk revelou um valor-p inferior a $2,2 \times 10^{-16}$, rejeitando a hipótese de normalidade dos erros. Isso indica que os erros não seguem uma distribuição normal multivariada, o que compromete a adequação do modelo linear para os dados e sugere a necessidade de modelos alternativos.

A Tabela 4.4 mostra que a assimetria dos resíduos foi de 3,1, indicando uma distribuição assimétrica à direita, e a curtose foi de 16,1, sugerindo caudas mais pesadas que a distribuição normal.

Tabela 4.4 – Assimetria e Curtose dos Resíduos do Modelo de Regressão Múltipla

Métrica	Valor
Assimetria	3,1
Curtose	16,1

Os resultados apresentados na Tabela 4.4 reforçam a presença de *outliers* e também indicam que o modelo não está adequadamente ajustado aos dados e que abordagens alternativas, como modelos de regressão robustos, podem ser necessárias.

A Figura 4.3 apresenta o histograma (direita) e o gráfico quantil-quantil (QQ-plot) de probabilidade normal (esquerda) dos resíduos padronizados do modelo de regressão linear.

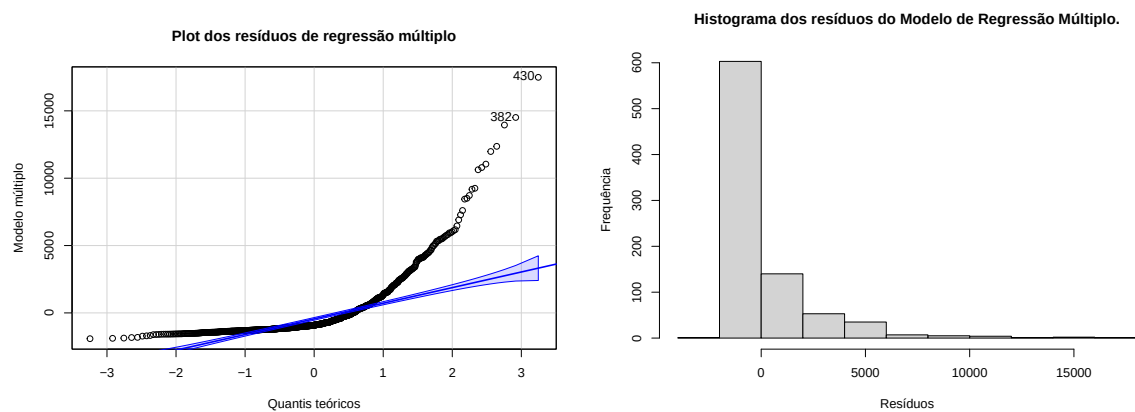


Figura 4.3 – Gráfico e Histograma dos resíduos do modelo
Fonte: autor (2025)

O QQ-Plot da imagem à esquerda da Figura 4.3 evidencia um desvio significativo dos resíduos em relação à linha de referência, indicando que eles não seguem uma distribuição normal. O histograma da imagem à direita, por sua vez, evidencia forte assimetria à direita, reforçando a ocorrência de *outliers* e confirmando a violação do pressuposto de normalidade dos erros. Os gráficos dos resíduos ainda mostram a presença de uma elevada proporção de outliers, o que pode indicar problemas na adequação do modelo, violando os pressupostos do termo do erro do modelo linear usual.

Teste de Homocedasticidade

A Tabela 4.5 apresenta o resultado do teste de Breusch-Pagan utilizado para verificar a hipótese de homocedasticidade dos erros.

Tabela 4.5 – Resultado do teste de homocedasticidade (Breusch-Pagan)

Teste	Estatística	GL	p-valor
Breusch-Pagan (BP)	6,2183	4	0,1834

O valor-p igual a 0,1834 observado na Tabela 4.5 mostra que não há evidências estatísticas para rejeitar a hipótese nula de homocedasticidade. Assim, pode-se considerar que a variância dos erros é constante.

Teste de Independência

A Tabela 4.6 apresenta o resultado do teste de Durbin-Watson, que avaliar a suposição de erros independentes.

Tabela 4.6 – Resultado do teste de independência (Durbin-Watson)

Teste	Estatística	p-valor
Durbin-Watson (DW)	1,6589	$3,021e - 7$

A Tabela 4.6 apresenta um valor-p extremamente baixo ($p < 0,001$), mostrando fortes evidências de autocorrelação positiva, sugerindo que os erros não são independentes. Diante desse resultado, recomenda-se a utilização de modelos que considerem a dependência espacial dos dados.

Os resultados obtidos nesta seção indicam que o modelo linear é inadequado para representar adequadamente os dados, evidenciando a necessidade de abordagens alternativas que considerem distribuições assimétricas e a presença de valores extremos.

Além disso, um estudo conduzido por Sousa & Barreto (2023) analisou o impacto da contaminação por outliers na identificação de sistemas e observou que, para modelos não robustos, a convergência máxima ocorre até 5% de contaminação, enquanto para modelos robustos, a convergência é mantida mesmo com até 25% de outliers. Isso sugere que a presença de discrepâncias nos dados pode afetar a estabilidade dos modelos estatísticos, sendo fundamental avaliar o uso de métodos robustos para evitar distorções nos resultados.

Por outro lado, Teixeira & Scalón (2013) destaca que a presença de *outliers* pode comprometer a precisão das estimativas e a interpretação dos resultados em análises estatísticas. Especialmente, quando presentes em grandes quantidades, os *outliers* podem influenciar negativamente os resultados, distorcendo as medições e a modelagem dos dados, exigindo cuidados especiais na sua detecção e tratamento para garantir a robustez e a confiabilidade da análise.

4.3 Modelo de regressão linear múltipla robusto

Como as análises realizadas anteriormente indicaram que o modelo de regressão linear múltipla violou os pressupostos básicos sobre os erros, foi ajustado um novo modelo utilizando uma abordagem robusta de regressão linear múltipla. Essa modelagem visa minimizar o impacto de possíveis outliers no ajuste, os quais podem comprometer os resultados do modelo tradicional sem robustez.

A Tabela 4.7 apresenta os resultados do modelo ajustado de regressão linear múltipla com robustez aplicados à incidência de dengue nos municípios de Minas Gerais.

Tabela 4.7 – Resultados do modelo robusto (MM-type)

Variável	Coef.	EP	Valor-t	p-valor
Intercepto	159,2571	59,0174	2,698	0,0071**
Arborização vias públicas	1,4506	0,6986	2,077	0,0381*
Esgotamento sanitário adequado	1,0301	0,7281	1,415	0,1575
Área urbanizada	2,2780	1,6190	1,407	0,1598
Urbanização vias públicas	-0,2559	0,8188	-0,313	0,7547
Coeficiente de determinação robusto: $R^2_{MM} = 2,007\%$				

Nota: *** $p < 0,001$; ** $p < 0,01$; * $p < 0,05$. Coef. = Coeficiente; EP = Erro Padrão.

A Tabela 4.7 mostra que o modelo robusto também apresentou um coeficiente de determinação muito baixo ($R^2 = 0,02007$), indicando que menos de 3% da variabilidade da incidência da dengue é explicada pelas variáveis independentes incluídas nos modelos.

Além disso, o modelo sem robustez (veja Tabela 4.2) mostrou que nenhuma das covariáveis foi estatisticamente significativa ao nível de 5%. Entretanto, o modelo com robustez (Tabela 4.7) apresentou apenas a variável arborização de vias públicas com significância estatística, o que sugere que, entre as variáveis analisadas, essa é a única com associação significativa com o aumento da incidência de dengue.

Foram conduzidas análises de resíduos para verificar a suposição dos erros do modelo ajustado. Os resultados (não apresentados aqui) obtidos indicaram que as suposições sobre os erros do modelo robusto também não foram observadas e, portanto, o modelo linear robusto também é inadequado para representar adequadamente os dados. Assim, nas próximas seções serão considerados modelos com estrutura espacial.

4.4 Análise exploratória espacial

Antes de ajustar modelos com estrutura de dependência espacial é necessário verificar se os dados apresentam dependência (autocorrelação) espacial. Essa verificação pode ser feita de várias formas conforme apresentaremos nesta seção.

4.4.1 Mapa da incidência de dengue

A Figura 4.4 apresenta um mapa que mostra a incidência de dengue por município do estado de Minas Gerais referente ao ano de 2023.

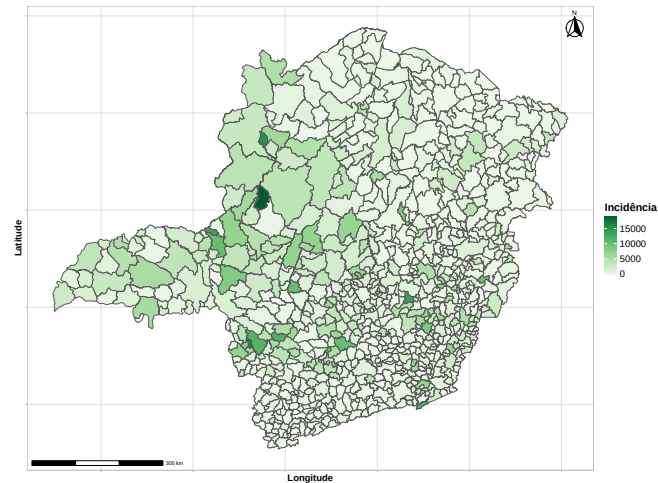


Figura 4.4 – Mapa da incidência de dengue para o ano de 2023

Fonte: Autor (2025)

Na Figura 4.4 observa-se que muitos municípios registraram zero casos, enquanto outros apresentaram valores muito elevados, indicando que a distribuição da doença não é homogênea para os municípios ao longo do território estadual. A Figura 4.4 mostra ainda que em 2023 houve maior concentração de casos nas regiões do Triângulo Mineiro e na Região Metropolitana de Belo Horizonte.

Índice de Moran global

Para avaliar a presença de autocorrelação espacial nos dados de incidência de dengue foi utilizado o Índice de Moran Global, cujo resultado é apresentado na Tabela 4.8.

Tabela 4.8 – Resultados do índice de Moran Global para a incidência de dengue para o ano de 2023

Índice	Estimativa	Valor Esperado	Erro Padrão	Z-score	p-value
Moran	0,2757	-0,0012	0,0004464	13,106	< 2,2e-16

Como é mostrado na Tabela 4.8, o índice de Moran de 0,2757 confirma a existência de autocorrelação espacial positiva, indicando que municípios com alta incidência estão geo-

graficamente agrupados. O valor-p inferior ao nível de significância de 5%, leva a rejeição da hipótese nula de distribuição aleatória da incidência de dengue. O Z-score elevado (13,106) reforça a significância estatística do padrão espacial observado, indicando que a distribuição da doença não ocorre de forma aleatória, mas sim pode estar sendo influenciada por fatores regionais.

4.4.2 Diagrama de dispersão de Moran

Na Figura 4.5 têm-se o diagrama de dispersão de Moran que é uma ferramenta essencial na análise de autocorrelação espacial.

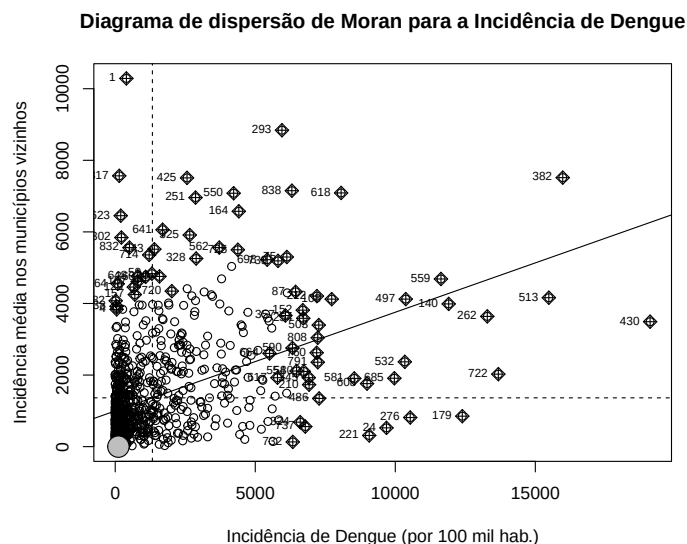


Figura 4.5 – Diagrama de dispersão
Fonte: Autor (2025).

Na Figura 4.5 pode-se observar a relação entre a taxa de incidência de Dengue (por 100 mil habitantes) em cada município (eixo x) e a média da incidência nos municípios vizinhos (eixo y), ponderada pela matriz de pesos espaciais.

Quadrantes Q1 (alto-alto) e Q2 (baixo-baixo) evidenciam clusters espaciais, ou seja, áreas em que municípios apresentam valores semelhantes de incidência de Dengue aos de seus vizinhos. Isso indica autocorrelação espacial positiva, sendo um padrão comum em regiões onde a propagação da doença ocorre de forma regionalizada (Scalon, 2024).

Por outro lado, os quadrantes Q3 (alto-baixo) e Q4 (baixo-alto) apontam para outliers espaciais, ou seja, municípios com valores discrepantes em relação aos seus vizinhos (Scalon,

2024). Essas diferenças de incidência de dengue pode nos indicar fatores locais específicos que influenciam a incidência da doença.

A distribuição dos pontos no gráfico permite, portanto, identificar não apenas o grau de autocorrelação espacial, mas também potenciais outliers que merecem atenção especial.

Índice local de autocorrelação espacial

O Índice local de autocorrelação espacial (LISA) foi utilizado para analisar a autocorrelação espacial local da incidência de dengue em Minas Gerais para o ano de 2023. Os resultados são apresentados na Figura 4.6.

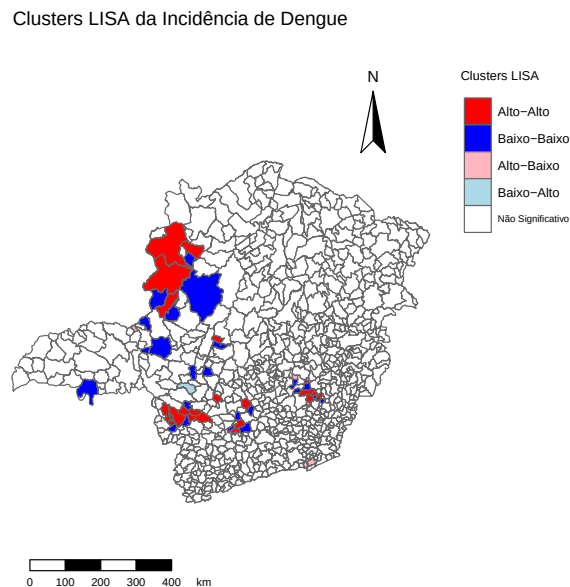


Figura 4.6 – Índice de Moran Local LISA
Fonte: Autor (2025)

A Figura 4.6 destaca padrões de autocorrelações espaciais locais estatisticamente significativos ($p < 0,05$). As áreas classificadas como Alto-Alto e Baixo-Baixo, correspondentes aos quadrantes Q1 e Q2 do diagrama de espalhamento de Moran, indicam locais onde a autocorrelação espacial é positiva. Por outro lado, as categorias Alto-Baixo e Baixo-Alto, associadas aos quadrantes Q3 e Q4, identificam áreas com autocorrelação espacial negativa, conforme ilustrado na Figura 2.8.

Além disso, outliers espaciais (Alto-Baixo e Baixo-Alto) aparecem de forma dispersa, enquanto a maior parte do território não apresenta autocorrelação espacial significativa. Esses

padrões espaciais sugerem que ações de vigilância e controle devem ser focadas, principalmente, nas áreas de clusters Alto-Alto, onde a incidência da doença é mais expressiva.

Mapa do intervalo interquartil

O mapa do intervalo interquartil considera como *outliers* aqueles municípios cuja incidência (*Incidência por 100 mil habitantes*) ultrapassa o limite superior definido por:

$$\text{Limite Superior} = Q3 + 1.5 \times IQR$$

. Na Figura 4.7 são apresentados os municípios identificados como *outliers* por meio do critério do intervalo interquartil (IQR).

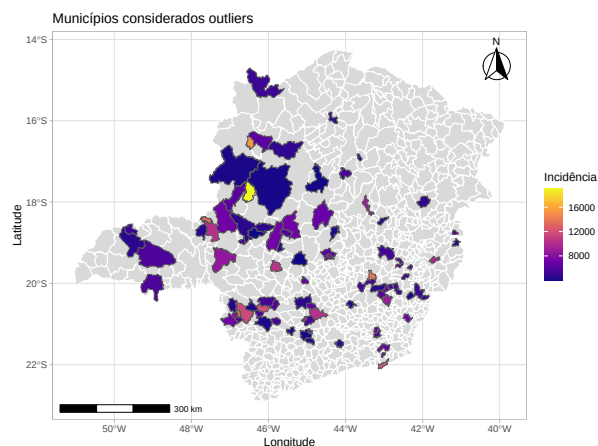


Figura 4.7 – Municípios que são considerados outliers

Fonte: Autor (2025)

A Figura 4.7 mostra que 87 municípios apresentam valores acima de 4011,104 e, portanto, são identificados como *outliers*.

Diante dos resultados apresentados nesta seção, torna-se essencial o ajuste de modelos robustos que incluem a estrutura de dependência espacial dos dados.

4.5 Modelos autorregressivo simultâneo (SAR) e Autoregressivo Condicional (CAR)

Iniciaremos com os modelos mais simples de regressão espacial não robustos (SAR e CAR) bem como seus respectivos diagnósticos.

4.5.1 Ajustes dos modelos SAR e CAR

Na Tabela 4.9 são apresentados os resultados dos modelos Modelo autorregressivo simultâneo (SAR) e Modelo Autoregressivo Condicional (CAR) ajustados com o objetivo de considerar a dependência espacial para explicar os casos de incidência de dengue em Minas Gerais em 2023. Ambos os modelos consideraram as quatro covariáveis do estudo, além do termo de autocorrelação espacial.

Tabela 4.9 – Comparação dos Modelos de Regressão Espacial: SAR (lag) e CAR (erro)

Fatores	SAR (lag)				CAR (erro)			
	Coef.	EP	z-valor	p-valor	Coef.	EP	z-valor	p-valor
Intercepto	466.68	247,80	1.883	0,060	1122.61	270,78	4.146	<0,001***
Arborização	2.90	2,83	1.022	0,307	2.26	2,86	0.789	0,430
Esgotamento	1.96	2,96	0.661	0,509	0.78	2,99	0.262	0,793
Área urbana	4.99	4,86	1.026	0,305	5.84	4,93	1.186	0,236
Urbanização	-2.90	3,49	-0.831	0,406	-1.48	3,52	-0.420	0,675
ρ/λ	0.454	0.042	10.744	<0.001***	0.457	0.042	10.821	<0.001***
Métricas								
AIC	15464				15465			
σ	2030,6				2031,3			

Nota: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$. EP = Erro Padrão; σ = desvio padrão dos resíduos.

ρ = coeficiente de autocorrelação espacial (SAR); λ = parâmetro de autocorrelação espacial (CAR).

Com base nos resultados apresentados na Tabela 4.9 observa-se que nenhuma das covariáveis foi estatisticamente significativa em ambos os modelos, com exceção do intercepto, o qual representa a média geral da variável resposta quando todas as covariáveis assumem valor zero. Isso indica que os modelos, por serem estimadores globais, não conseguiram captar adequadamente os valores extremos (*outliers*) presentes nos dados. Nota-se que as métricas dos dois modelos foram bastante semelhantes, indicando que ambos tiveram limitações em capturar toda a variação espacial presente nos dados, especialmente em relação aos valores atípicos (*outliers*) observados. Isso sugere que, na forma como foram especificados, os modelos SAR

e CAR não foram adequados para explicar a distribuição espacial dos casos de dengue nos municípios de Minas Gerais.

4.5.2 Análise de resíduos dos modelos SAR e CAR

A Figura 4.8 apresenta os histogramas e os gráficos QQ-plot dos resíduos dos modelos SAR e CAR.

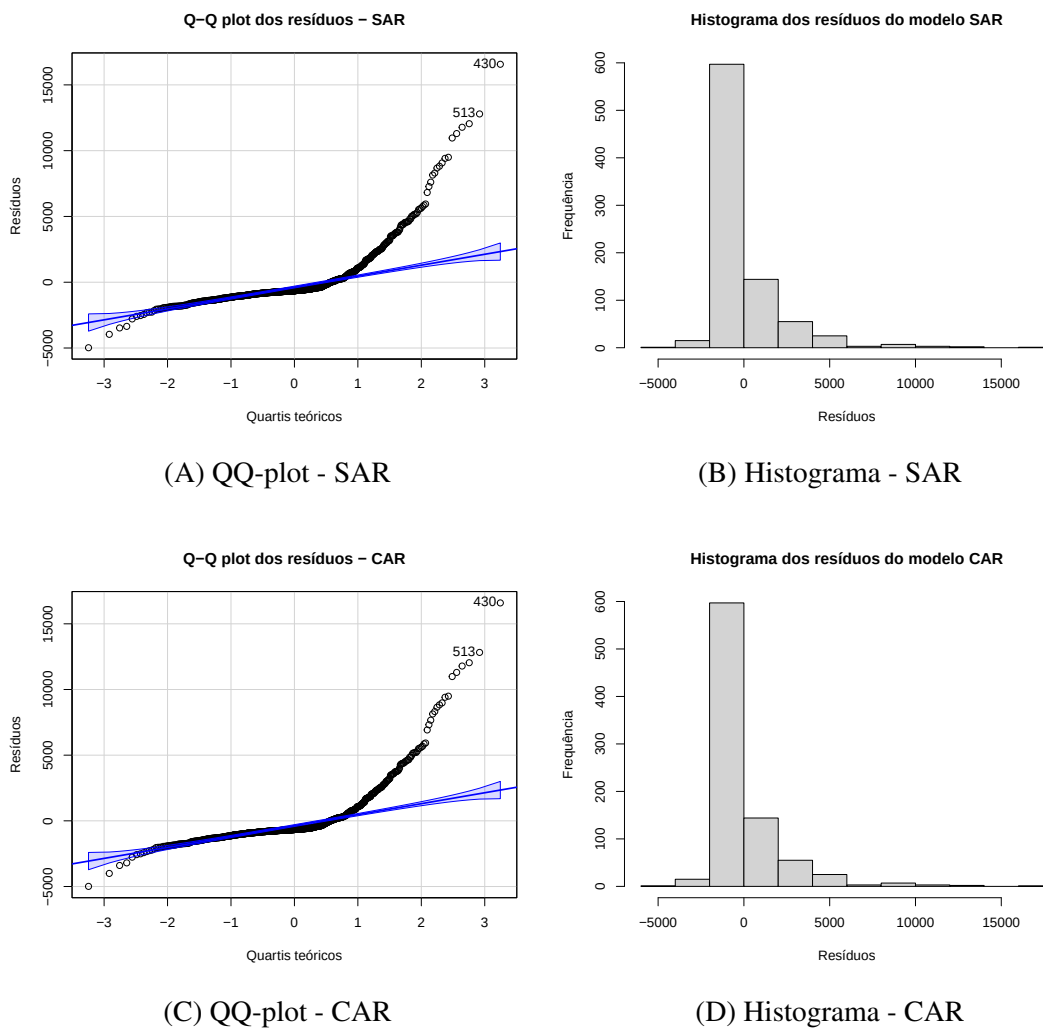


Figura 4.8 – Gráficos dos resíduos dos modelos SAR e CAR: QQ-plots e histogramas.

Fonte: Autor (2025).

Nos gráficos QQ-plot dos resíduos dos modelos SAR e CAR, representados nas subfiguras (A) e (C) da Figura 4.8, é possível notar desvios significativos nas caudas, com diversos valores fora do intervalo de confiança, indicando a presença de caudas mais pesadas. Ainda na

Figura 4.8, os histogramas dos resíduos dos gráficos (B) e (D) indicam uma assimetria à direita, o que reforça a indicação de que os erros dos modelos não seguem uma distribuição normal.

Além disso, foi realizado o teste de normalidade dos resíduos para complementar a análise visual, cujos resultados estão apresentados na Tabela 4.10.

Tabela 4.10 – Teste de Normalidade Shapiro-Wilk para os Resíduos dos Modelos SAR e CAR

Modelo	Estatística W	p-valor
SAR	0,678	< 2,2e-16
CAR	0,675	< 2,2e-16

Conforme evidenciado pelo teste de normalidade de Shapiro-Wilk (Tabela 4.10), observa-se que os erros de ambos os modelos não seguem uma distribuição normal, violando um dos pressupostos básicos para esses modelos. Essa violação indica que as estimativas dos modelos podem não ser tão precisas.

Análises para testar as hipóteses de independência e homocedasticidade (não apresentados aqui) dos erros mostraram que essas duas suposições foram aceitas.

4.6 Comparação dos quatro modelos ajustados

Nesta seção apresenta-se a comparação dos ajustes dos quatro modelos estimados até o momento.

A Tabela 4.11 apresenta algumas medidas resumo dos resíduos dos quatro modelos ajustados.

Tabela 4.11 – Resumo dos resíduos dos modelos ajustados

Modelo	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
Múltiplo Linear (LM)	-1913,8	-1225,4	-909,4	339,9	17488,2
Múltiplo Robusto (LM-MM)	-531,8	-229,6	38,2	1301,7	18697,1
Espacial SAR	-4979,3	-927,2	-650,5	193,2	16567,0
Espacial CAR	-4990,9	-923,0	-663,3	200,8	16592,5

Analisando os resultados apresentados na Tabela 4.11 observa-se uma grande variação entre os valores mínimos e máximos dos resíduos em cada modelo. No modelo linear simples

(LM), os resíduos variam de -1913,8 até 17488,2, enquanto no modelo linear com estimador robusto (LM-MM), os resíduos variam de -531,8 até 18697,1. Nota-se que o valor mínimo do modelo robusto é muito menor em magnitude, indicando que o estimador robusto reduz a influência dos valores atípicos extremos.

Já nos modelos espaciais SAR e CAR, os resíduos apresentam valores mínimos e máximos bastante semelhantes, indicando uma consistência entre eles. Essa proximidade sugere que ambos os modelos podem capturar de forma semelhante a variação espacial dos dados.

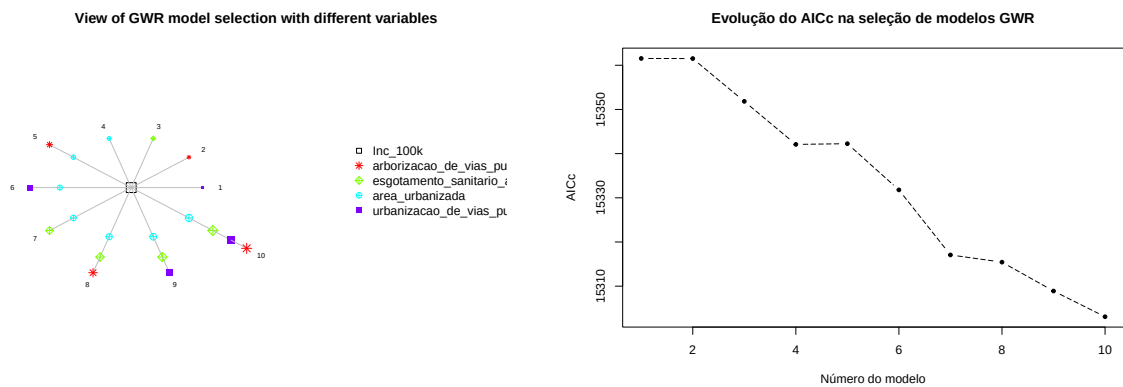
Na próxima seção descreve-se e discute os modelos com estimadores locais, os quais levam em conta a dependência espacial e os valores atípicos.

Resumindo, os quatro modelos ajustados anteriormente falham em modelar a incidência dos casos de dengue no estado de Minas Gerais. Portanto, sugere-se que, para os dados deste estudo, sejam adotados modelos robustos que atenuem a presença de valores discrepantes (*outliers*) mas que também considerem a estrutura de dependência espacial. Para tal, serão ajustados os modelos de Regressão Geograficamente Ponderada (GWR) e Regressão Geograficamente Ponderada robusta (RGWR).

4.7 Seleção de Variáveis para os Modelos GWR RGWR

Antes de ajustar os modelos GWR e RGWR foi realizado um procedimento de seleção de covariáveis para determinar o subconjunto de preditores que oferece o melhor equilíbrio entre ajuste e parcimônia. Para este fim, adotou-se a técnica de seleção passo a passo (*stepwise*) proposta por (Gollini *et al.*, 2015), disponível no pacote `GWmodel`. Este método avalia o desempenho de modelos GWR com diferentes combinações de variáveis utilizando o Critério de Informação de Akaike corrigido (AICc) como métrica de seleção.

Na Figura 4.9 apresenta-se uma visão em círculo (H) da sequência de inclusão das variáveis, enquanto o gráfico (I) exibe a queda no valor do AICc a cada passo, indicando a melhoria do modelo.



(H) Visão em círculo do processo de seleção

(I) Evolução do AICc por modelo

Figura 4.9 – Resultados do procedimento de seleção de variáveis para o modelo GWR.

Fonte: Elaborado pelo autor (2025).

O gráfico à esquerda (H) da Figura 4.9, denominado *visão em círculo (circle view)*, ilustra a sequência em que as variáveis independentes foram adicionadas ao modelo com base na maior redução do AICc em cada passo. A variável dependente (*Inc_100k*) está no centro e os raios representam a construção progressiva dos modelos, partindo dos preditores individuais até o modelo completo. O gráfico à direita (I) complementa a análise, exibindo a trajetória dos valores de AICc para cada um dos modelos ajustados sequencialmente. Evidencia-se uma tendência de queda contínua no valor do AICc à medida que mais variáveis são incorporadas, com o menor valor sendo alcançado no modelo final, que inclui todos os preditores. Os resultados da Figura 4.9 indicam que o modelo completo é a especificação que melhor equilibra o ajuste aos dados e a complexidade. Com este resultados tem-se a justificativa de usar todas as quatro covariáveis no modelo GWR e em sua versão robusta RGWR.

4.8 Modelos GWR e RGWR

Segundo Fotheringham et al. (1998), a regressão geograficamente ponderada (GWR) é uma extensão da regressão linear que permite estimar coeficientes locais, capturando variações espaciais (não-estacionariedade) existentes na relação entre a variável dependentes e variáveis independentes. A regressão geograficamente ponderada robusta (RGWR) é uma extensão da GWR desenvolvida para minimizar a influência de valores discrepantes nas estimativas locais dos coeficientes (Fotheringham et al., 2002). A ideia é que esses dois modelos possam, pelo menos, minimizar os problemas encontrados nos modelos anteriores.

4.8.1 Ajuste dos Modelos GWR RGWR

Na Tabela 4.12 são descritos os resultados dos modelos GWR e RGWR ajustados aos dados de incidência de dengue no Estado de Minas Gerais em 2023.

Tabela 4.12 – Modelos GWR e RGWR (robusto) com AICc e AIC

Coeficientes	GWR					RGWR (robusto)				
	Mín.	1º Quartil	Mediana	3º Quartil	Máx.	Mín.	1º Quartil	Mediana	3º Quartil	Máx.
Intercepto	-416.71	359.05	963.82	1907.52	4687.58	-454.50	350.48	984.25	1924.23	4902.66
Arborização vias públicas	-20.15	-3.93	-0.10	3.63	22.20	-20.44	-3.97	-0.22	3.47	22.05
Esgotamento sanitário	-30.04	-4.38	2.04	6.47	35.20	-32.31	-4.33	2.07	6.66	34.90
Área urbanizada	-45.10	-4.36	4.50	23.03	146.85	-45.73	-4.44	4.04	23.58	147.94
Urbanização vias públicas	-51.40	-9.19	-4.01	0.72	31.72	-52.53	-9.32	-4.19	0.77	32.15
Métricas										
AICc			15460,76					15460,58		
AIC			15378,85					15376,39		

A Tabela 4.12 resume a distribuição estatística dos coeficientes locais estimados para cada variável explicativa ao longo dos municípios analisados e mostra a principal diferença entre esses modelos e os quatro modelos ajustados anteriormente: ao invés de termos estimativas pontuais dos coeficientes de regressão, temos distribuições. Assim, para o intercepto, os coeficientes locais estimados variam de -454,50 a 4902,66. Esses valores correspondem à incidência estimada de dengue por 100 mil habitantes quando as variáveis explicativas do modelo assumem valor zero em cada município. O intervalo interquartil varia de 350,48 a 1924,23, com amplitude de 1573,75, indicando variação espacial nos valores do intercepto estimados pelo modelo.

Considerando a variável arborização de vias públicas, a variabilidade espacial dos efeitos é analisada a partir das estimativas dos coeficientes locais obtidas pelo modelo RGWR. Para essa variável, expressa em porcentagem, os coeficientes locais estimados variam de -20,44 a 22,05. Esses resultados indicam que um aumento de 1 ponto percentual na arborização está associado, em algumas localidades, a uma redução de até 20,44 casos de dengue por 100 mil habitantes, enquanto, em outras, o mesmo aumento está associado a um acréscimo de até 22,05 casos por 100 mil habitantes.

Na covariável esgotamento sanitário adequado, o intervalo dos coeficientes varia de -32,31 a 34,90. A mediana positiva (2,07) apontando que uma tendência geral de que melhores

condições sanitárias estejam associadas a menores índices de dengue, embora haja localidades onde o efeito seja negativo. O intervalo interquartil vai de $-4,33$ a $6,66$, indicando que, em 50% dos municípios, os efeitos locais estão dentro dessa faixa.

Para a variável área urbanizada, os coeficientes locais apresentam ampla variação, de $-45,73$ a $147,94$. A mediana é de $4,04$, e o intervalo interquartil mostra que metade dos municípios apresenta efeitos entre $-4,44$ e $23,58$. Esses resultados indicam que a urbanização pode tanto aumentar quanto reduzir a incidência, dependendo da localidade.

A variável urbanização de vias públicas também apresenta forte variação espacial, com coeficientes entre $-52,53$ e $32,15$. A mediana negativa ($-4,19$) mostram que, na maioria das localidades, a melhoria da infraestrutura viária está associada a uma redução na incidência da doença. A variação interquartil vai de $-9,32$ a $0,77$.

A Tabela 4.12 também mostra que não existe diferenças entre os dois modelos com relação as métricas AICc e AIC.

Finalmente, pode-se observar na Tabela 4.12 que apenas o intercepto apresenta um intervalo interquartil que não contém o zero e, portanto, poderia ser considerado "significante". O mesmo não pode ser considerado para as estimativas dos coeficientes das variáveis independentes. Entretanto, esse modelos permitem a condução de análises mais precisas para verificar a significância estatística de cada um dos coeficientes de regressão, conforme veremos nas subseções a seguir. De qualquer maneira, esses resultados reforçam a principal vantagem dos modelos RGW e RGWR sobre os quatro anteriores, ou seja, identificar como os efeitos das covariáveis variam espacialmente. As análises mostram que, embora certas variáveis tenham efeitos positivos em algumas áreas, em outras podem ter impacto negativo. Isso evidencia a importância de abordagens que capturam a heterogeneidade espacial no contexto da saúde pública.

4.8.2 Coeficiente de determinação local

Uma das vantagens dos modelos GWR e RGWR é permitir uma avaliação da variabilidade local explicada pelo modelo através de um mapa dos coeficientes de determinação locais (Fotheringham; Brunsdon; Charlton, 2002). Na Figura 4.10 são apresentados os mapas desses coeficientes de determinação (R^2) locais para os modelos GWR (A) e RGWR (B).

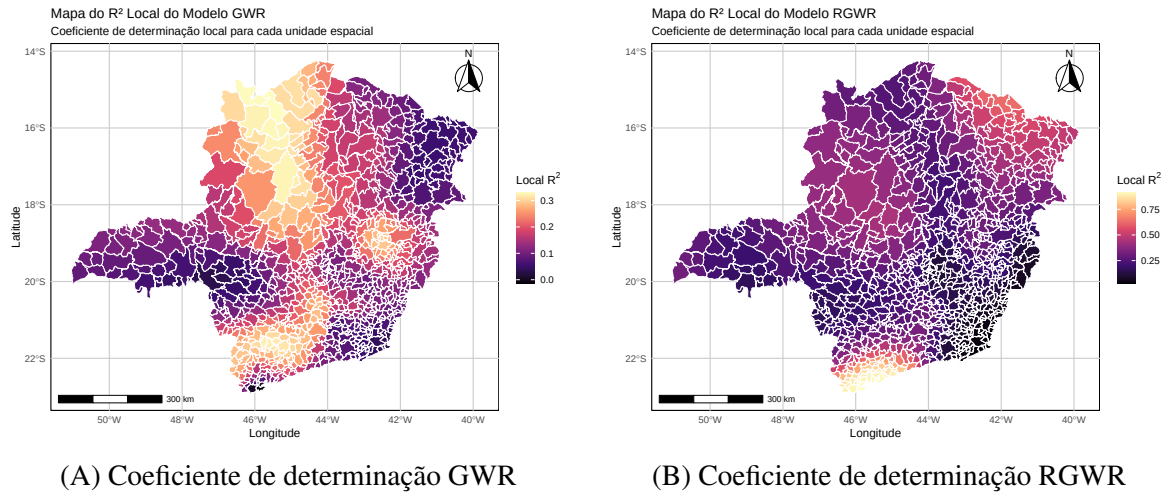


Figura 4.10 – Coeficientes de determinação locais para os modelos GWR (a) e RGWR(b).

Fonte: Autor (2025).

Ao comparar os dois mapas da Figura 4.10 nota-se uma diferença na distribuição espacial do R^2 local entre os dois modelos. O Modelo RGWR (Figura 4.10B) demonstra uma capacidade significativamente maior de explicar localmente a variação espacial na incidência de dengue ao incluir as covariáveis, com valores de R^2 local que alcançam até 0,75 (ou 75%) em algumas regiões. Já o Modelo GWR (Figura 4.10A) apresenta uma explicação localmente menor, com valores de R^2 local máximos de aproximadamente 0,3 (ou 30%). Principalmente no Sul de Minas Gerais, o modelo RGWR consegue explicar mais a variação total, o que nos indica que as covariáveis incluídas no modelo podem explicar até 75% da incidência de dengue nesses locais. Por outro lado, o modelo GWR, com um R^2 mais baixo, apresenta maior poder de explicação em algumas áreas do Sul, Norte e Centro de Minas Gerais, embora em menor magnitude. Podemos perceber que o modelo robusto consegue capturar melhor variação local.

4.8.3 Teste F1

Outra vantagem dos modelos GWR e RGWR é que eles permitem obter as somas dos quadrados dos resíduos para cada município. LEUNG, MEI & ZHANG (2000) apresentam uma estatística de teste (denominada de F1) e seu valor-p associado que permite comparar os dois modelos e indicar se a variabilidade espacial dos coeficientes é um fator relevante.

A Tabela 4.13 apresenta os resultados do teste F1.

Tabela 4.13 – Comparação dos Testes F1 de Leung et al. (2000) para GWR e RGWR

Modelo	F Estatística	GL Denom.	Valor-p
GWR	0,82696	848	0,0034***
RGWR	0,82538	848	2,08e-05***

Nota: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$.

Conforme apresentado na Tabela 4.13, para o modelo GWR e o RGWR, o valor-p foi inferior a 0,05, indicando que ambos os modelos proporcionam um ajuste estatisticamente significativo.

4.8.4 Teste F3

LEUNG, MEI & ZHANG (2000) também apresentam uma estatística de teste (denominada de F3) que pode ser utilizada para avaliar se os coeficientes de cada covariável analisada é estatisticamente significativa em toda a região de estudo, ou seja, para testar a variação espacial global desses parâmetros.

A Tabela 4.14 apresenta os resultados do Teste F3.

Tabela 4.14 – Teste F3 de LEUNG, MEI & ZHANG (2000) para os Modelos GWR e RGWR

Variável	GWR			RGWR (robusto)		
	F3 estatística	GL num.	p-valor	F3 estatística	GL num.	Valor-p
Intercepto	2,64322	279,64	$< 2e-16$ ***	2,62663	279,76	$< 2,2e-16$ ***
Arborização de vias públicas	0,80773	381,78	0,99131	0,79538	381,49	0,9987
Esgotamento sanitário adequado	1,19237	301,62	0,03045 *	1,20280	301,65	0,0090 **
Área urbanizada	1,50706	44,59	0,01939 *	1,55159	44,93	0,0103 *
Urbanização de vias públicas	1,00780	275,71	0,46240	0,99692	275,14	0,5031

Nota: *** $p < 0,001$, ** $p < 0,01$, * $p < 0,05$.

De acordo com a Tabela 4.14, observa-se que ambos os modelos GWR e RGWR identificaram que os coeficientes das covariáveis esgotamento sanitário adequado e área urbanizada variam significativamente no espaço. Para a variável esgotamento sanitário adequado, os valores-p obtidos foram de 0,03045 * no modelo GWR e de (0,0090 **) no RGWR. Já para área urbanizada, os valores-p foram de (0,01939 *) no GWR e de (0,0103 *) no RGWR. Esses valores são estatisticamente significativos ao nível de 5%, permitindo rejeitar a hipótese nula

de que os coeficientes são constantes em toda a área de estudo. O Teste F_3 , nos indica que os coeficientes dessas duas covariáveis apresentam variação espacial global estatisticamente significativa entre os municípios de Minas Gerais. O intercepto dos modelos também apresentou variação espacial altamente significativa, com valores-p inferiores a ($p < 2e-16$ ***). Em contrapartida, as covariáveis arborização de vias públicas e urbanização de vias públicas não apresentaram evidência de variação espacial significativa, com valores-p elevados, acima do nível de significância de (0,05).

4.8.5 Análise comparativa dos resíduos dos modelos GWR e RGWR

A Tabela 4.15 apresenta as estatísticas descritivas dos resíduos dos modelos RGWR (robusto) e GWR (padrão), permitindo avaliar seu desempenho relativo. Os resultados revelam padrões importantes sobre a precisão e robustez de cada abordagem.

Tabela 4.15 – Estatísticas descritivas dos resíduos dos modelos RGWR e GWR.

Estatística	RGWR (Robusto)	GWR (Padrão)	Interpretação
Média	12,98	14,77	RGWR apresenta erros médios 12,8% menores.
Mediana	-451,21	-451,65	Praticamente igual ($\Delta < 0,1\%$).
1º Quartil	-944,94	-936,34	RGWR reduz erros nos piores ajustes em 0,9%.
3º Quartil	150,54	172,80	RGWR diminui erros nos melhores ajustes em 12,9%.
Mínimo	-3.445,03	-3.342,35	RGWR apresenta melhores extremos (3,1%).
Máximo	13.845,65	13.856,23	Diferença muito pequena (0,08%).

Os resultados dos modelos indicam que o modelo RGWR apresentou um erro de 12,98, enquanto o modelo GWR apresentou um erro de 14,77. Embora a diferença seja pequena, o modelo RGWR apresentou menor erro. A mediana dos resíduos é praticamente idêntica para ambos os modelos (-451,21 para RGWR e -451,65 para GWR), sugerindo desempenho semelhante na maior parte das observações.

Ao analisar os quartis, observa-se que o RGWR reduz os erros nas observações com os piores ajustes (primeiro quartil: -944,94 contra -936,34), com melhoria de cerca de 0,9%. No terceiro quartil, o RGWR também supera o GWR, com resíduos de (150,54 contra 172,80),

indicando melhora de 12,9% nos melhores ajustes. Em relação aos valores extremos, o RGWR apresenta desempenho mais robusto nas subestimações mais intensas (mínimo de -3.445,03 contra -3.342,35, melhora de 3,1%). Já no valor máximo dos resíduos, a diferença entre os modelos é desprezível (13.845,65 no RGWR versus 13.856,23 no GWR), com variação inferior a 0,1%.

Esses resultados mostram que, além de apresentar desempenho semelhante ao GWR nas observações centrais, o RGWR se destaca pela maior robustez e precisão em situações de maior variabilidade, tanto em subestimações quanto em superestimações. Portanto, o modelo RGWR oferece ajustes mais estáveis frente às variações dos dados tanto na questão da estrutura de dependência espacial quanto na presença de *outliers*.

4.8.6 Verificação da Multicolinearidade

Para garantir a confiabilidade dos coeficientes estimados no modelo RGWR foi realizada a análise da multicolinearidade local por meio do cálculo do Fator de Inflação da Variância (VIF) para cada variável explicativa. Os resultados do resumo do VIF local estão apresentados na Tabela 4.16.

Tabela 4.16 – Resumo do VIF local para variáveis explicativas no modelo GWR

Variável	Min.	1º Quartil	Mediana	Média	3º Quartil	Máx.
Arborização de vias públicas	1.002	1.017	1.035	1.041	1.060	1.148
Esgotamento sanitário adequado	1.050	1.125	1.197	1.273	1.318	2.037
Área urbanizada	1.030	1.085	1.116	1.126	1.155	1.315
Urbanização de vias públicas	1.003	1.046	1.085	1.164	1.161	2.012

Os valores do VIF apresentados na Tabela 4.16 indicam que não há multicolinearidade significativa entre as variáveis analisadas. Todos os valores médios encontram-se muito abaixo do limiar comumente aceito de 5, o que sugere que a estimativa dos coeficientes do modelo GRWR não está comprometida por problemas de correlação excessiva entre os preditores.

4.9 Comparação de todos os modelos ajustados

Os resultados obtidos anteriormente não deixam muitas dúvidas que o modelo Geograficamente Ponderado robusto (RGWR) é o melhor para explicar a incidência de dengue no estado de Minas Gerais. De qualquer maneira, nesta seção vamos apresentar mais algumas análises que permitem realizar comparações entre os modelos Linear Múltiplo (LM), Linear Múltiplo Robusto (LM-MM), Modelos Espaciais SAR e CAR, além dos Modelos Geograficamente Ponderados GWR (padrão) e RGWR (robusto).

4.9.1 Estatística descritiva dos resíduos

Na Tabela 4.17 são apresentadas as estatísticas descritivas (medidas resumo) dos resíduos dos modelos ajustados, permitindo comparar a distribuição dos erros e a eficácia de cada abordagem na modelagem dos dados.

Tabela 4.17 – Medidas resumo dos resíduos dos modelos ajustados

Modelo	Mínimo	1º Quartil	Mediana	3º Quartil	Máximo
Múltiplo Linear (LM)	-1913,8	-1225,4	-909,4	339,9	17488,2
Múltiplo Robusto (LM-MM)	-531,8	-229,6	38,2	1301,7	18697,1
Espacial SAR	-4979,3	-927,2	-650,5	193,2	16567,0
Espacial CAR	-4990,9	-923,0	-663,3	200,8	16592,5
GWR (Padrão)	-3342,4	-936,3	-451,6	172,8	13856,2
RGWR (Robusto)	-3445,0	-944,9	-451,2	150,5	13845,6

O modelo LM (que não considera a estrutura espacial dos dados nem a presença de outliers) apresentou a maior variação nos resíduos, com valores entre -1.913,8 e 17.488,2. Essa ampla dispersão sugere baixa capacidade de ajuste em regiões com comportamento atípico, indicando a influência significativa de valores extremos.

A introdução de robustez no modelo LM-MM reduziu consideravelmente essa variabilidade, principalmente na cauda inferior, com resíduos variando de -531,8 a 18.697,1. A presença de uma mediana positiva (38,2) e um intervalo interquartilico mais estreito sugere maior simetria e menor influência de outliers.

Os modelos espaciais SAR e CAR apresentaram desempenhos semelhantes, com resíduos oscilando entre aproximadamente -4.990 e 16.590. Apesar de ainda apresentarem resíduos elevados, a menor amplitude em relação ao LM, indicando que a incorporação da dependência espacial contribui para um ajuste mais realista, especialmente em áreas com padrões espaciais bem definidos.

Do ponto de vista das medidas de tendência central, o modelo LM apresentou uma mediana negativa (-909,4) e um IQR em torno de 1.565,3, o que reforça a elevada dispersão dos erros e a presença de assimetria. Em contraste, o LM-MM apresentou maior equilíbrio entre os quartis (-229,6 a 1.301,7), sugerindo uma distribuição mais centrada dos resíduos.

Os modelos SAR e CAR, com medianas próximas a -650 e IQRs em torno de 1.120, demonstraram melhorias importantes no controle da variabilidade, embora ainda com leve assimetria negativa.

Adicionalmente, os modelos geograficamente ponderados (GWR e RGWR) demonstraram desempenho promissor em contextos com forte heterogeneidade espacial. O GWR padrão apresentou resíduos entre -3.342,4 e 13.856,2, com mediana de -451,6, enquanto o RGWR, ao incorporar robustez, exibiu leve melhora nas extremidades (mínimo de -3.445,0 e máximo de 13.845,6) e mediana praticamente idêntica (-451,2). Essa redução nos resíduos extremos sugere que o RGWR suaviza a influência de outliers, oferecendo um ajuste mais estável e equilibrado entre regiões com diferentes padrões espaciais.

Em síntese, os modelos LM-MM, SAR, CAR, GWR e RGWR apresentaram melhor desempenho em relação ao modelo linear clássico, com resíduos mais centrados e menor sensibilidade a valores extremos. A consideração de robustez e estrutura espacial, portanto, é fundamental para aprimorar a qualidade dos ajustes em dados com variações geográfica.

4.9.2 Intervalo interquartil dos resíduos

Uma forma robusta de avaliar a dispersão dos erros de cada modelo é através do Intervalo Interquartil (IQR), que mede a variabilidade dos 50% centrais dos resíduos e, portanto, sendo menos sensível a presença de *outliers*.

A Tabela 4.18 apresenta a comparação desta métrica entre os modelos ajustados.

Tabela 4.18 – Dispersão dos Resíduos (Intervalo Interquartil) por Modelo.

Modelo	1º Quartil	3º Quartil	IQR (3ºQ - 1ºQ)
OLS	-1225.4	339.9	1565.3
Robusto (MM)	-229.6	1301.7	1531.3
SAR	-927.2	193.2	1120.4
CAR	-923.0	200.7	1123.7
GWR	-936.34	172.80	1109.1
RGWR	-944.94	150.54	1095.5

Avaliando a Tabela 4.18 observa-se uma melhoria progressiva no controle da dispersão dos erros à medida que modelos mais complexos são aplicados. O modelo OLS apresenta o maior IQR (1565.3), confirmando uma alta variabilidade nos resíduos. Os modelos espaciais globais (SAR e CAR) conseguem reduzir significativamente essa dispersão para cerca de 1120. No entanto, os modelos geograficamente ponderados (GWR e RGWR) mostram-se ainda mais eficazes. O modelo final, **RGWR**, alcança o menor IQR de todos (1095.5), indicando que, ao modelar a não-estacionariedade espacial e controlar a influência de outliers simultaneamente, ele é o modelo mais eficiente em produzir erros com menor dispersão para a maioria das observações, reforçando sua superioridade de ajuste.

4.9.3 Mapas dos Resíduos

Uma etapa fundamental na avaliação e comparação dos modelos ajustados é a análise visual da distribuição espacial de seus resíduos. O objetivo é identificar se os modelos conseguem produzir erros aleatoriamente distribuídos no espaço, o que indicaria um bom ajuste.

A Figura 4.11 apresenta os mapas de resíduos para cada um dos seis modelos testados. As cores em vermelho indicam subestimação da incidência de dengue, enquanto as cores em azul indicam superestimação.

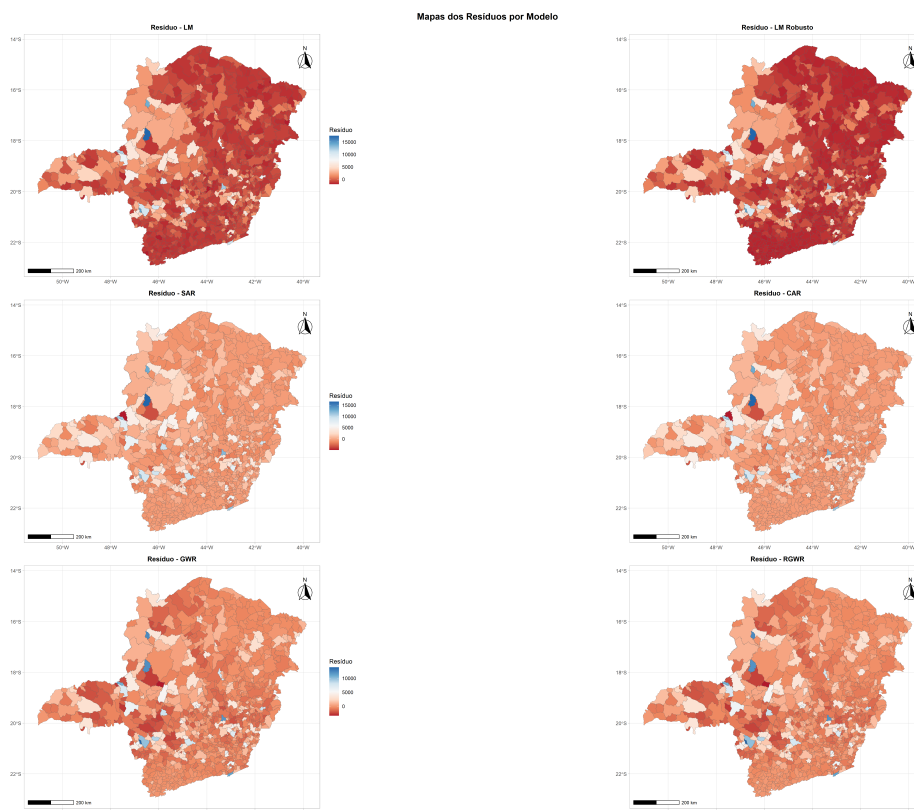


Figura 4.11 – Comparação da distribuição espacial dos resíduos para os seis modelos ajustados.

Fonte: Elaborado pelo autor (2025).

A comparação dos ajustes dos modelos nos mapas nos informa sobre a qualidade do ajuste dos modelos. A seguir, apresenta-se como cada modelo se comportou:

Mapa de resíduos do modelo de regressão linear múltipla: Esse mapa representa o pior cenário e serve como linha de base. Observa-se a formação de grandes e bem definidos clusters de resíduos positivos (vermelho) e negativos (azul), indicando uma forte autocorrelação espacial nos erros. O modelo falha de forma sistemática, subestimando a incidência em algumas regiões e superestimando em outras, o que justifica a necessidade de abordagens espaciais.

Mapa de resíduos do modelo LM robusto: O padrão espacial dos resíduos é visualmente muito semelhante ao do modelo LM. Embora a abordagem robusta possa ter controlado a influência de outliers nos coeficientes, ela não aborda a dependência espacial, que é a principal fonte de erro estruturado. Consequentemente, os resíduos permanecem espacialmente correlacionados.

Mapa de resíduos dos modelos SAR e CAR: Os modelos espaciais globais (SAR e CAR) apresentam uma melhora visível. Os clusters de erro são menos intensos e as manchas de cores quentes e frias começam a se desfragmentar. Ao modelar a autocorrelação espacial global,

esses modelos absorvem parte do padrão, resultando em erros com menor organização espacial. Contudo, os clusters ainda são aparentes, sugerindo que uma abordagem espacial global não é suficiente para capturar toda a heterogeneidade do processo.

Mapa de resíduos do modelo GWR (Geograficamente Ponderado): A transição para um modelo local representa a melhoria mais substancial. O mapa de resíduos do GWR está visivelmente mais fragmentado, aproximando-se de um padrão aleatório. Ao permitir que os coeficientes variem no espaço, o GWR conseguiu se adaptar às particularidades locais, explicando grande parte da não estacionariedade que os modelos globais ignoram.

Mapa de resíduos do modelo RGWR (GWR Robusto): Esse mapa apresenta o resultado mais próximo do ideal. Ele combina a flexibilidade do GWR em capturar a não estacionariedade com a capacidade de um modelo robusto de mitigar a influência de outliers locais. O resultado é o mapa de resíduos mais aleatório entre todos os modelos, com os menores e menos definidos clusters de erro, validando-o como o modelo mais bem ajustado e confiável para os dados em estudo.

Em suma, a sequência de mapas demonstra visualmente a superioridade do modelo RGWR. Partindo de um modelo inicial (LM), a análise mostra que a incorporação da dependência espacial (SAR/CAR) e, principalmente, da não-estacionariedade espacial (GWR), é fundamental. O refinamento final com a abordagem robusta (RGWR) produz os resíduos mais próximos do teórico, indicando sua seleção como o modelo final desta pesquisa.

Para investigar essa heterogeneidade espacial, a análise local (LISA), apresentado na Figura 4.6 foi fundamental, ao identificar *clusters* de alta incidência (Alto-Alto) nas regiões oeste e norte, e de baixa incidência (Baixo-Baixo) no centro-oeste.

4.9.4 Métricas para seleção do modelo

Na Tabela 4.19 são apresentadas as métricas utilizadas para a seleção do modelo mais adequado aos dados do estudo que relacionam as quatro covariáveis com a incidência de dengue no estado de Minas Gerais.

Tabela 4.19 – Comparação das métricas dos modelos ajustados

Modelo	AICc	AIC	BIC	RMSE	R²
Regressão Linear	15576.32	15576.00	15604.74	2224.00	0.009
Regressão Robusta	13319.65	13319.58	13343.25	422.80	0.020
SAR	15463.96	15464.00	15497.10	2030.60	0.169
CAR	15465.01	15465.00	15498.15	2031.30	0.169
GWR	15460.76	15378.85	15493.63	2022.81	0.264
GWR Robusto	15460.58	15376.39	14922.04	2020.47	0.267

Para a escolha do melhor modelo foram avaliadas as métricas AIC, AICc, BIC e RMSE, priorizando-se os menores valores bem como o maior valor de R². Com base nesses critérios é possível observar que o modelo GWR Robusto (RGWR) apresentou não apenas o melhor ajuste aos dados, mas também contribuiu para capturar as variações locais, minimizando o impacto de observações discrepantes.

Tempo de execução operacional de cada modelo proposto

A Tabela 4.20 apresenta os tempos de execução e o consumo de memória para os diferentes modelos estatísticos utilizados nesta dissertação.

Tabela 4.20 – Tempos de execução e consumo de memória dos modelos estatísticos

Modelo	Tempo (unid.)	Unidade	Memória (MB)
Regressão Linear (LM)	0.95	segundos	0.44
Regressão Linear Robusta (LMROB)	0.90	segundos	0.00
Modelo SAR	3.85	segundos	0.83
Modelo CAR	3.13	segundos	0.00
Modelo GWR	37.51	segundos	1.54
Modelo GWR Robusto	3.66	minutos	2.50

Os modelos de regressão linear e linear robusta apresentaram os menores tempos de execução. Os modelos espaciais globais (SAR e CAR) demandaram praticamente o mesmo tempo para serem executados. Mais lentos que os dois primeiros mas, muito mais rápidos

que os modelos locais que apresentaram maior tempo de processamento. O RGWR foi o que exigiu o maior tempo e consumo entre os modelos avaliados. Esses tempos não representam complexidade assintótica (*Big O*), mas sim valores empíricos obtidos durante a execução prática dos modelos nos dados utilizados. A análise desses tempos é útil para comparar a viabilidade computacional dos métodos propostos caso o número de variáveis independentes envolvidas e o número de áreas em análise aumentem.

4.10 Análise visual do modelo RGWR por meio de mapas

Uma vez que as seções anteriores mostraram que o modelo RGWR é o melhor para explicar a incidência de dengue no estado de Minas Gerais, é possível explorar graficamente como os coeficientes de regressão estimados das variáveis independentes se comportam no estado de Minas Gerais. Nesta seção apresentaremos alguns desses mapas.

4.10.1 Mapas dos coeficientes: Arborização e esgotamento sanitário

A Figura 4.12 apresenta os coeficientes estimados pelo modelo RGWR para as variáveis independentes arborização de vias públicas (A) e esgotamento sanitário (B).

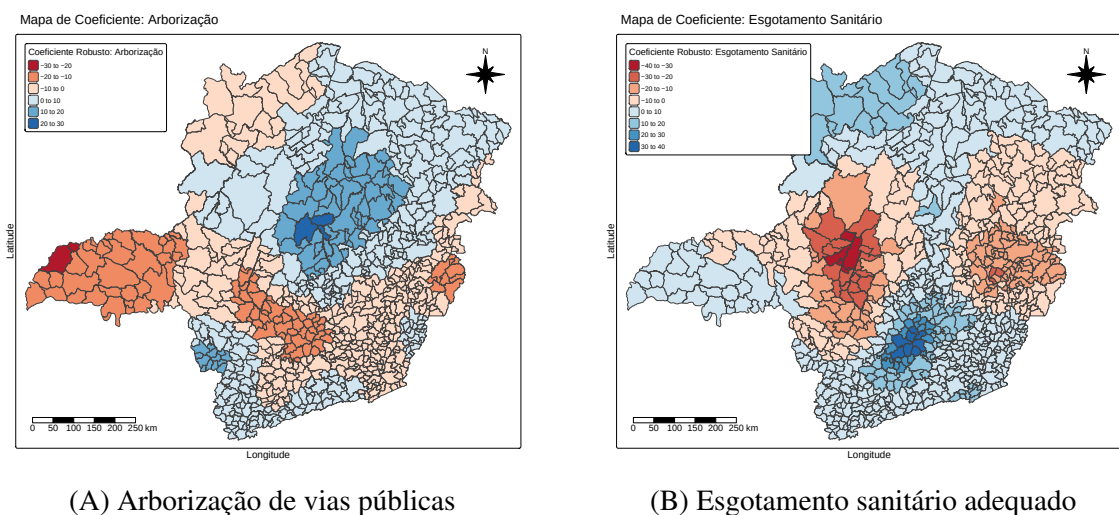


Figura 4.12 – Coeficientes locais para Arborização (a) e Esgotamento Sanitário (b).

Fonte: Elaborado pelo autor (2025).

A Figura 4.12 mostra como as variáveis arborização de vias públicas e esgotamento sanitário adequado se relacionam espacialmente com os casos de Dengue nos municípios analisa-

dos. Os resultados mostram que os efeitos dessas variáveis não são espacialmente homogêneos, ou seja, os impactos variam conforme a localização geográfica, podendo estar associados tanto ao aumento quanto à redução dos casos da doença.

Na Figura 4.12(A) observa-se que os coeficientes variam entre valores negativos e positivos, indicando diferentes relações entre arborização e a ocorrência de Dengue: Coeficientes positivos (0 a 30) predominam nas regiões centro e norte do estado, sugerindo que, nessas áreas, a maior arborização está associada a maior incidência de casos de Dengue. Esse efeito pode estar relacionado ao acúmulo de água em áreas sombreadas ou à dificuldade de controle vetorial em locais com vegetação densa. Coeficientes negativos (-30 a 0) concentram-se principalmente nas regiões sul e sudoeste, indicando que, nesses municípios, a arborização está relacionada a menor ocorrência da doença, possivelmente refletindo práticas urbanas que integram arborização com controle ambiental eficaz.

Na Figura 4.12(B) observa-se que os coeficientes também variam entre negativos e positivos, representando a influência espacial do esgotamento sanitário adequado sobre os casos de Dengue. Coeficientes positivos (0 a 40) aparecem em determinadas regiões, sugerindo que, paradoxalmente, melhores condições de saneamento podem estar associadas ao aumento nos casos de Dengue, possivelmente em áreas urbanizadas onde o crescimento desordenado supera a capacidade de controle do vetor. Coeficientes negativos (-40 a 0) indicam municípios onde o esgotamento sanitário adequado contribui para a redução da incidência da doença, reforçando a importância da infraestrutura básica no combate ao mosquito transmissor.

4.10.2 Mapa dos coeficientes: Área urbanizada e urbanização de vias

Na Figura 4.13 apresentam os coeficientes locais estimados para as variáveis Áreas urbanizadas (C) e Urbanização de vias públicas (D) em relação a incidência de Dengue.

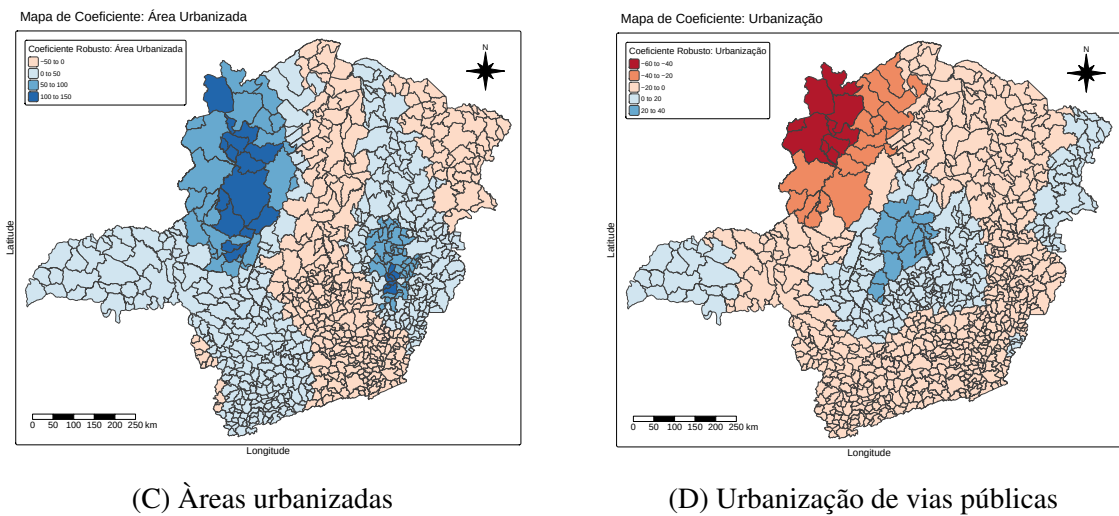


Figura 4.13 – Coeficientes locais para Áreas Urbanizadas (C) e Urbanização de Vias (D).

Fonte: Elaborado pelo autor (2025).

O mapa da esquerda da Figura 4.13(C) ilustra a relação entre o tamanho da áreas urbanizadas de um município e a incidência de dengue. Os coeficientes positivos, a maior parte do Estado de Minas Gerais, especialmente as regiões Noroeste, Central e Leste, apresenta coeficientes positivos. Nestas áreas, um aumento na área urbanizada está associado a um aumento nos casos de dengue. O pontos mais intenso (vermelho escuro) está localizado na região Leste. Os coeficientes negativos, mostram a diminuição dos casos de dengue nos municípios de estudo. O mapa da direita (Figura D) mostra a relação entre a urbanização de vias públicas e a incidência de dengue. Os coeficientes negativos nos indicam que na maior parte do território mineiro, com destaque para um grande e intenso cluster no noroeste e norte, exibe coeficientes negativos. Logo, podemos observar que nestes municípios, uma melhor urbanização das vias públicas funciona como um forte fator de proteção para o aumento dos casos de dengue. Os coeficientes positivos são observados na região Central do Estado de Minas Gerais, nestas áreas, uma melhor urbanização das vias está associada a mais incidência de dengue.

4.10.3 Mapa dos erros-padrão

A Figura 4.14 apresenta os mapas dos erros-padrão robustos estimados para os coeficientes das covariáveis no modelo RGWR. Os erros-padrão indicam o grau de incerteza na estimativa dos coeficientes locais. Assim, valores mais altos sugerem maior variabilidade e menor confiança na estimativa do efeito da covariável naquela região.

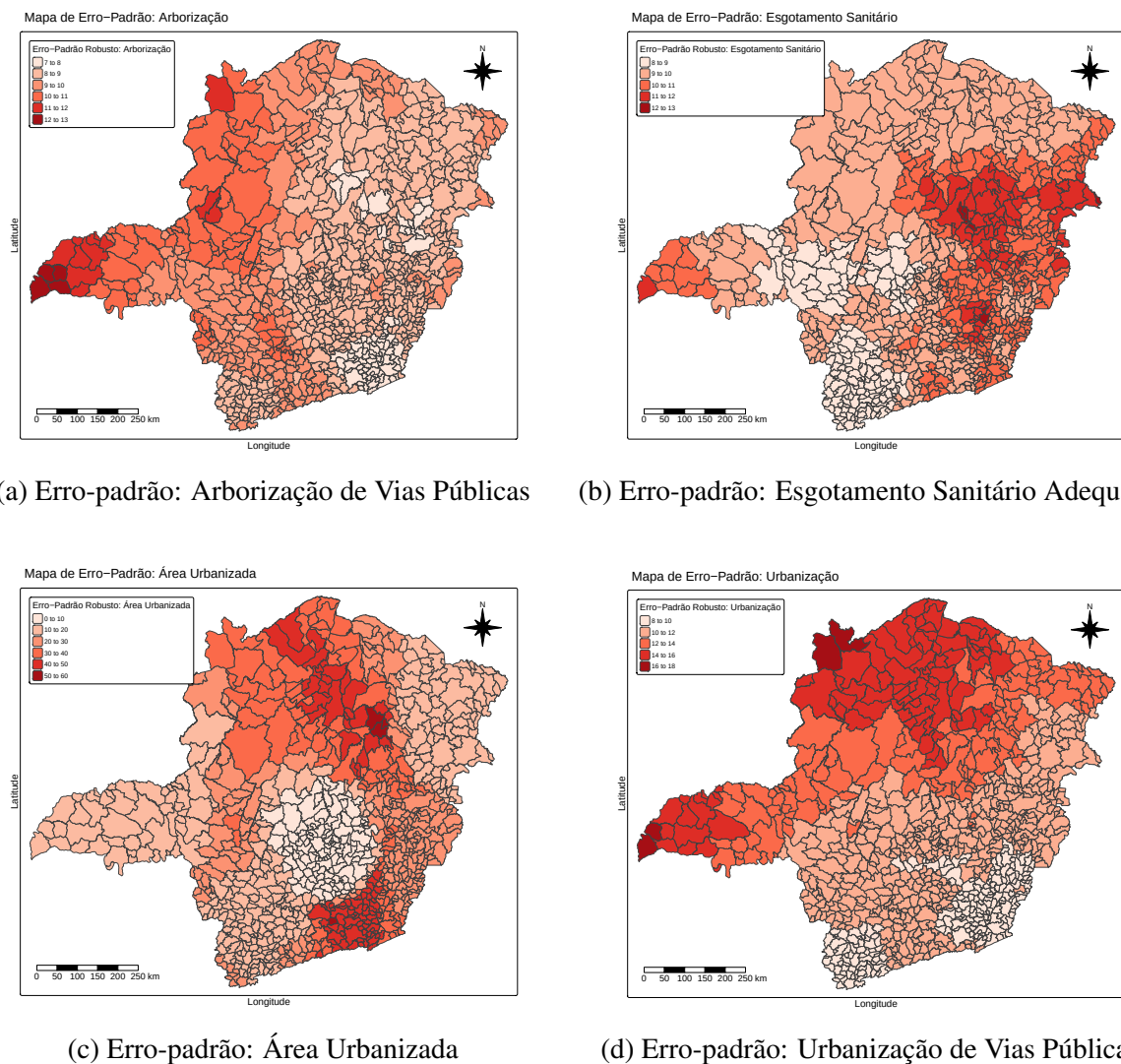


Figura 4.14 – Mapas dos erros-padrão dos coeficientes das covariáveis do estudo.

Fonte: Elaborado pelo autor (2025).

No caso da arborização (Figura 4.14a), os erros-padrão mais elevados concentram-se na região Oeste do Estado, especialmente no Triângulo Mineiro e parte do Norte, com valores entre 11 e 13. Já no Leste e Sul, os erros são menores, indicando maior precisão na estimativa do efeito dessa variável nessas áreas.

Para o esgotamento sanitário adequado (Figura 4.14b), os erros-padrão são mais elevados na porção Leste e Sudeste de Minas Gerais, destacando-se regiões como Zona da Mata e Vale do Rio Doce, com valores entre 11 e 13. Por outro lado, as menores incertezas ocorrem nas regiões Central e Sul do estado.

Com relação à área urbanizada (Figura 4.14c), observa-se uma grande variabilidade espacial. As regiões Norte e parte do Noroeste, especialmente a Zona da Mata, o Norte de Minas, o Noroeste e o Alto Paranaíba, apresentam os maiores erros-padrão, com valores que chegam a 60. Isso evidencia baixa precisão na estimativa dos efeitos locais dessa variável. Em contrapartida, a porção Sul do Estado apresenta erros mais baixos, indicando maior confiabilidade nas estimativas.

No caso da urbanização de vias públicas (Figura 4.14d), os maiores erros-padrão concentram-se na região Norte e no extremo Oeste, particularmente no Triângulo Mineiro e no Noroeste de Minas, com valores entre 16 e 18. Esses resultados sugerem maior incerteza quanto à influência dessa covariável sobre a incidência de dengue nessas áreas. Já nas regiões Central e Sul, os erros são menores, indicando maior robustez e precisão nas estimativas locais.

Em conjunto, os mapas indicam que a precisão das estimativas dos coeficientes varia espacialmente. Esses resultados reforçam a necessidade de utilizar modelos capazes de capturar essa heterogeneidade, como a Regressão Geograficamente Ponderada Robusta (RGWR). Ademais, as regiões com maiores erros-padrões exigem interpretação mais cautelosa, uma vez que os efeitos estimados das covariáveis nessas áreas são maiores.

4.10.4 Mapa dos valores-t

Os valores- t correspondem à razão entre os coeficientes locais estimados e seus respectivos erros padrão, sendo interpretados como *pseudo valores-t* no contexto da regressão geograficamente ponderada. Esses valores fornecem uma medida indicativa da evidência estatística de que os efeitos locais diferem de zero (Lu *et al.*, 2014).

Na Figura 4.15 são apresentados os mapas dos valores- t das covariáveis relacionadas à incidência de dengue para cada município de Minas Gerais. Esses mapas permitem identificar os efeitos locais e a significância estatística das variáveis explicativas no modelo GWR robusto.

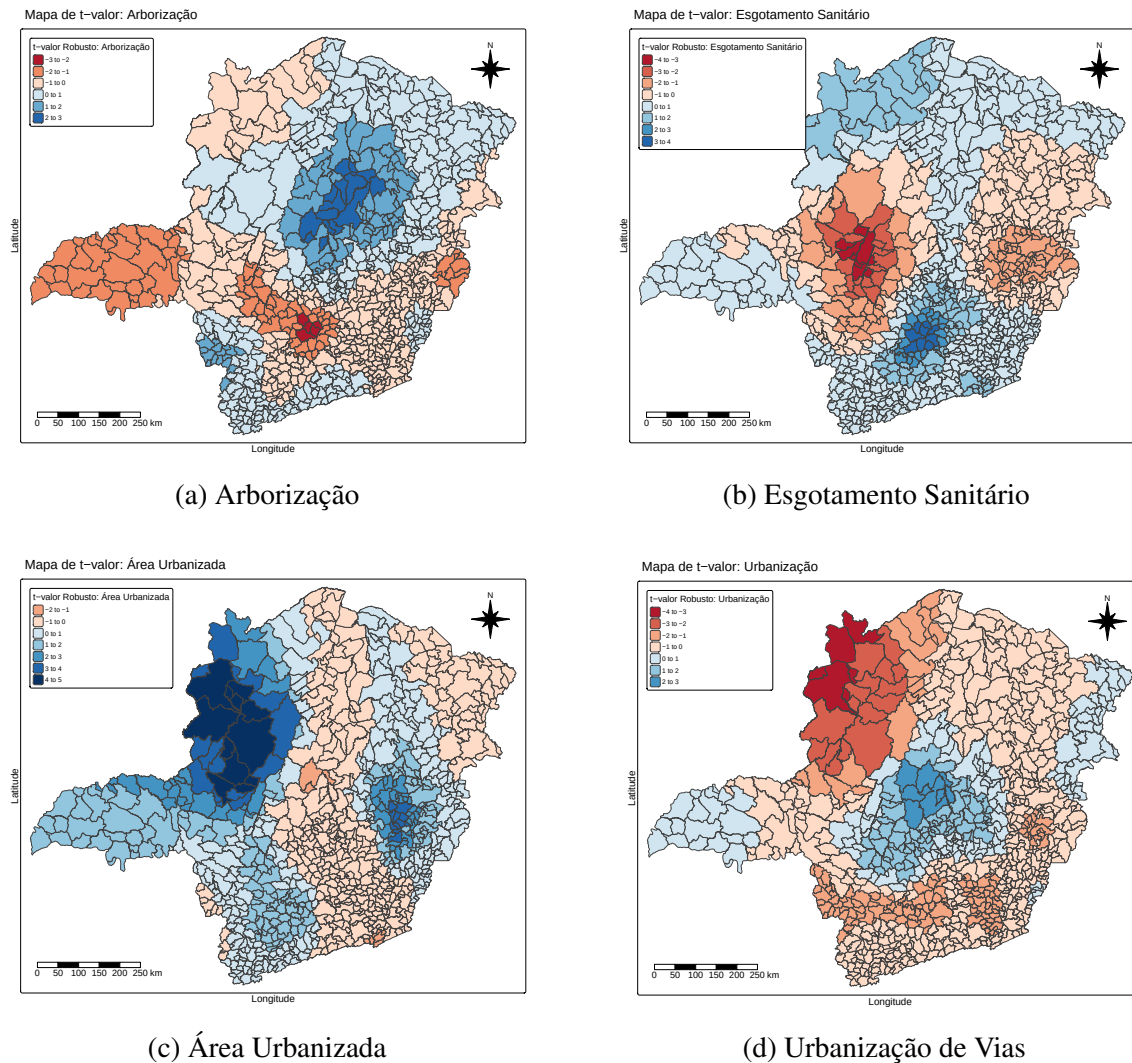


Figura 4.15 – Mapas de t-valor para Arborização (a), Esgotamento Sanitário (b), Área Urbanizada (c) e Urbanização de Vias (d).

Fonte: Elaborado pelo autor (2025).

Observa-se que a arborização de vias públicas (Figura 4.15a) apresenta valores-t entre -3 e 3. Nota-se um efeito positivo e estatisticamente significativo ($t > 2$) na região Norte e Central, enquanto na região Sul o efeito é negativo e também significativo ($t < -2$). Nas demais regiões do Estado, os valores-t estão indicando ausência de significância estatística.

No caso do esgotamento sanitário adequado (Figura 4.15b), os valores-t variam de -4 a 4. A associação é positiva e significativa no Sul do Estado ($t > 2$) e negativa e significativa no Oeste ($t < -2$), evidenciando que o impacto desta variável na incidência de dengue é espacialmente heterogêneo.

Para a área urbanizada (Figura 4.15c), os valores-t variam de -2 a 5. O destaque é a região Noroeste, onde os valores-t superiores a 4 indicam um forte efeito positivo sobre a

incidência de dengue. Em contrapartida, grande parte do estado apresenta valores próximos de zero, sugerindo ausência de significância estatística.

Na urbanização de vias públicas (Figura 4.15d) apresentam valores- t entre -4 e 3, com destaque para um efeito negativo e estatisticamente significativo nas regiões Norte e Noroeste ($t < -2$). Esses resultados indicam que, nessas áreas, a urbanização de vias pode estar associada à redução da incidência de dengue, possivelmente devido a melhores condições estruturais ou de drenagem.

Podemos destacar que os valores- t apresentados evidenciam padrões espaciais distintos, indicando a presença de heterogeneidade espacial nas covariáveis associadas à incidência de dengue. Isso significa que o efeito dessas variáveis não é uniforme ao longo do território, variando significativamente entre as diferentes regiões do estado.

4.10.5 Mapas dos valores- p

Os valores- p expressam a probabilidade de se obter um coeficiente local tão extremo quanto o estimado sob a hipótese nula de efeito nulo, sendo derivados dos *pseudo valores- t* . Em modelos GWR, esses valores devem ser ajustados para testes múltiplos, uma vez que são estimados simultaneamente para diversas localizações espaciais (Lu *et al.*, 2014).

Na Figura 4.16 são apresentados os mapas dos valores- p associados às variáveis explicativas consideradas no estudo, conforme metodologia descrita por (LU *et al.*, 2014). Cada gráfico (A, B, C e D) representa a distribuição espacial dos valores- p calculados localmente para uma das quatro covariáveis ambientais analisadas. Esses valores variam entre 0 e 1 e indicam o grau de significância estatística da relação entre cada covariável e a variável dependente (incidência de dengue nos municípios de Minas Gerais em 2023).

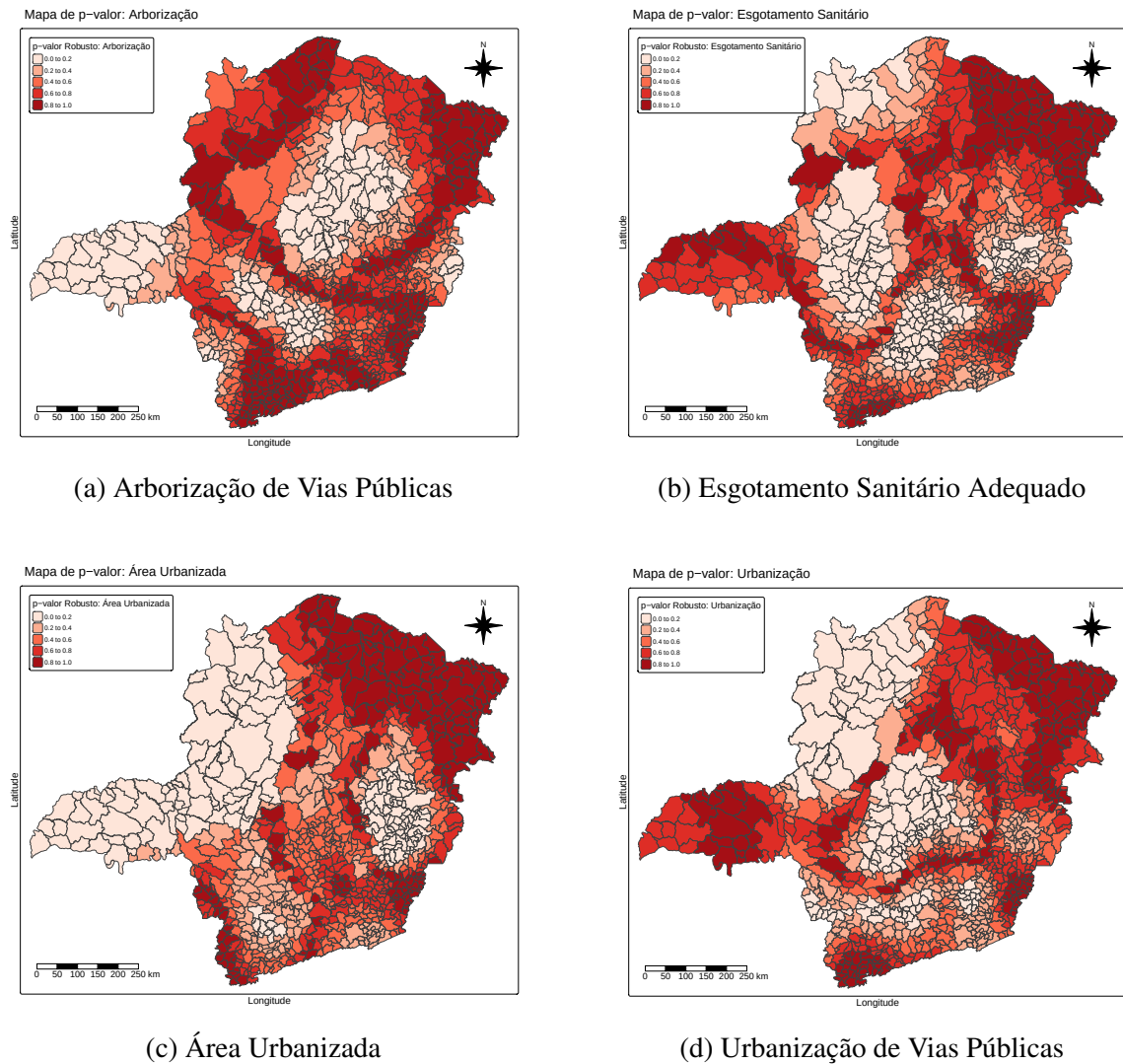


Figura 4.16 – Mapas de p-valor para as covariáveis ambientais selecionadas.

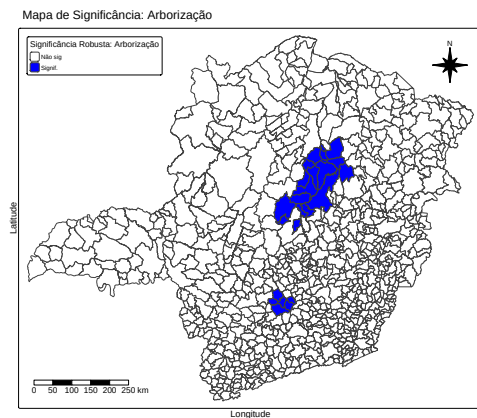
Fonte: Elaborado pelo autor (2025).

Observa-se que a significância das covariáveis apresenta variações espaciais em todo o domínio de probabilidades, o que pode dificultar a identificação dos municípios com significância estatística para cada uma das covariáveis. Uma maneira de facilitar essa identificação é estabelecer uma probabilidade arbitrária de corte e destacar no mapa, conforme apresentamos na próxima seção.

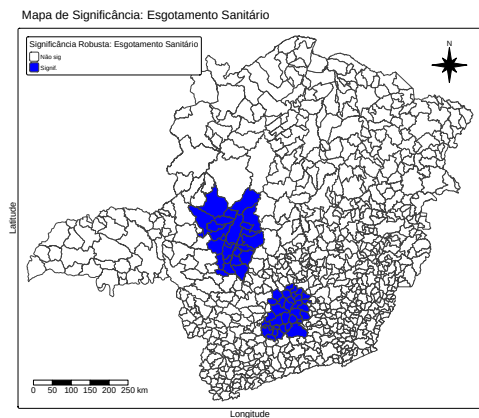
4.10.6 Mapas dos valores- p estatisticamente significativos

Nesta subseção são apresentados apenas os municípios cujos coeficientes locais apresentam evidência estatística de associação, conforme os valores- p estimados pelo modelo RGWR robusto. Adotou-se um nível de significância de 5% para este estudo, de modo que apenas os p -valores inferiores a 0,05 foram considerados estatisticamente significativos e evidenciados no mapa.

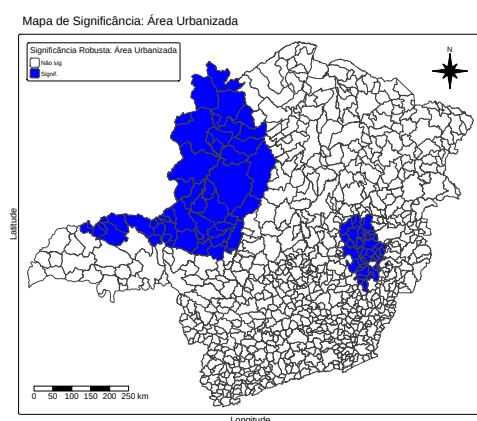
Na Figura 4.17 são realçados os municípios que apresentam covariáveis com significância estatística, ao nível de 5% de probabilidade, permitindo identificar as regiões onde os efeitos locais são relevantes para a incidência de dengue no Estado de Minas Gerais.



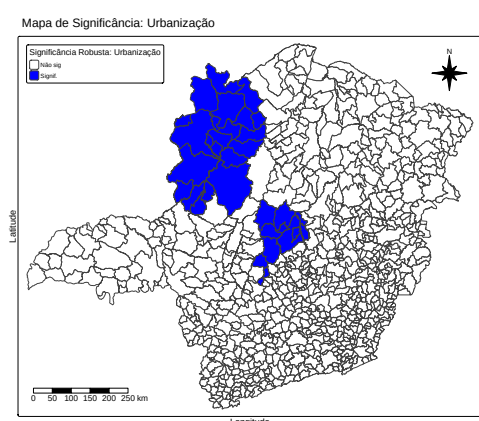
(e) Arborização



(f) Esgotamento Sanitário



(g) Área Urbanizada



(h) Urbanização de Vias

Figura 4.17 – Mapas dos p -valores locais dos coeficientes estimados pelo modelo RGWR robusto para as variáveis explicativas: Arborização (e), Esgotamento Sanitário (f), Área Urbanizada (g) e Urbanização de Vias (h).

Fonte: Elaborado pelo autor (2025).

Essa análise evidencia a heterogeneidade espacial da influência estatística das covariáveis sobre a incidência de dengue nos municípios de Minas Gerais. A distribuição espacial dos valores- p nos indica que os efeitos das covariáveis explicativas não são uniformes em todo o território. Apenas as variáveis área urbanizada e urbanização de vias públicas apresentam significância estatística coincidentes em alguns municípios do extremo oeste do estado.

A Tabela 4.21 apresenta o número de municípios com coeficientes locais estatisticamente significativos e não significativos para cada covariável relacionada à incidência de dengue, com base nos valores- p estimados segundo LU *et al.* (2014).

Tabela 4.21 – Resumo de valores significativos e não significativos por variável

Variável	Não Significativo	Significativo
Arborização	831	22
Esgotamento	792	61
Área Urbanizada	755	98
Urbanização	819	34

Podemos verificar na Tabela 4.21 que para a arborização de vias públicas, 22 municípios apresentaram associação estatisticamente significativa, indicando que a presença de árvores nas vias pode influenciar a incidência de dengue, possivelmente afetando o microclima ou a presença de criadouros do mosquito vetor. No caso do esgotamento sanitário adequado, 61 municípios mostraram coeficientes significativos, o que sugere que o saneamento básico tem papel importante na redução da incidência da doença, ao diminuir locais favoráveis à reprodução do mosquito. A área urbanizada apresentou 98 municípios com coeficientes significativos, evidenciando que o grau de urbanização está fortemente associado à incidência de dengue em diversas regiões, possivelmente em função da maior densidade populacional e condições ambientais propícias. Por fim, a urbanização de vias públicas teve 34 municípios com associação estatisticamente significativa, reforçando a influência das características urbanísticas locais na disseminação da doença.

Com base nos resultados obtidos neste capítulo é possível afirmar que o modelo GWR Robusto apresentou não apenas o melhor ajuste aos dados, mas também contribuiu para capturar as variações espaciais locais e minimizar o impacto de observações discrepantes. Estes resultados estão em consonância com os resultados de Harris *et al.* (2010) que não apenas pro-

puseram os métodos de filtragem utilizada no RGWR, como também demonstraram sua maior capacidade preditiva. Os resultados também corroboram os resultados obtidos por RAHAYU *et al.* (2023) que identificaram o RGWR como o modelo mais eficaz para analisar dados de saúde caracterizados por heterogeneidade espacial e presença de *outliers*. Em síntese, os resultados obtidos indicam que intervenções para controle da incidência da dengue devem considerar as especificidades locais (municipais) em termos de variáveis urbanas e sanitárias.

5 CONCLUSÃO

A análise inicial evidenciou as limitações dos modelos de regressão global para dados com estrutura espacial. A regressão linear múltipla, que assume que a relação entre as variáveis é estacionária ou uniforme no espaço, mostrou-se incapaz de capturar adequadamente da incidência da dengue em Minas Gerais. A violação de seus pressupostos, somada à presença de *outliers*, resultou em estimativas não confiáveis. A sua versão robusta e a inclusão da estrutura de dependência espacial nos modelos SAR e CAR também não produziram resultados satisfatórios. Diante desse cenário, utilizou-se os modelos de Regressão Geograficamente Ponderada (GWR) e sua versão robusta (RGWR). O modelo RGWR mostrou-se superior aos demais para entender a distribuição espacial da incidência de dengue no estado de Minas Gerais em função de variáveis urbanas e sanitária em dados de áreas (municípios) com a presença de *outliers* e não estacionariedade espacial. Espera-se que esses resultados possam contribuir para uma gestão mais eficiente e eficaz das ações de saúde pública ao combate ao mosquito da dengue.

REFERÊNCIAS

- AKAIKE, H. **Information Theory and an Extension of the Maximum Likelihood Principle. Selected Papers of Hirotugu Akaike**, [s.d.].
- ALMEIDA, A. F. M.; OLIVEIRA, F. d. Análise do pavement condition index aeroportuário com uso de estatística espacial. In: **Anais do Air Transportation Symposium (SITRAER). São José dos Campos, São Paulo**. [S.l.: s.n.], 2022.
- AMADOR, J. P. *et al.* Análise das pressuposições e adequação dos resíduos em modelo de regressão linear para individuais, ponderados e não ponderados, utilizando procedimentos do sas®. **Ciência e Natura**, v. 33, n. 2, p. 07–22, 2011.
- ANSELIN, L. **Spatial Econometrics: Methods and Models**. 1ed. ed. Dordrecht: Springer Netherlands, 1988. (Studies in Operational Regional Science).
- ATIKAH, N.; AFIFAH, D. L.; KHOLIFIA, N. Robust spatial regression model in city minimum wages (cmw) in east java 2018. In: **Proceedings of the 7th International Conference on Research, Implementation, and Education of Mathematics and Sciences (ICRIEMS 2020)**. [S.l.: s.n.], 2020.
- BARROSO, I. L. D. *et al.* Um estudo sobre a prevalência da dengue no brasil: Análise da literatura. **Brazilian Journal of Development**, v. 6, n. 8, p. 61878–61883, 2020.
- BECKER, R. A.; CHAMBERS, J. M.; WILKS, A. R. **The New S Language**. [S.l.]: Wadsworth & Brooks/Cole, 1988.
- BIASSOTI, A. V.; ORTIZ, M. A. L. Diagnóstico laboratorial da dengue. **Uningá Review**, v. 29, n. 1, 2017.
- BIVAND, R. S.; HAUKE, J.; KOSSOWSKI, T. Computing the jacobian in gaussian spatial autoregressive models: An illustrated comparison of available methods. **Geographical Analysis**, v. 45, n. 2, p. 150–179, 2013.
- BIVAND, R. S.; PEBESMA, E. J.; RUBIO, V. G. **Applied Spatial Data Analysis with R**. 2nd. ed. New York, NY: Springer, 2008.
- BIVAND, R. S.; PIRAS, G. Comparing implementations of estimation methods for spatial econometrics. **Journal of Statistical Software**, v. 63, p. 1–36, 2015.
- Brasil. **Dengue**. 2023. Acesso em: 10 set. 2024. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/d/dengue>.
- CARVALHO, L. R. **Modelos para descrever a curva de lactação de ovelhas Santa Inês e mestiças Santa Inês x Lacaune**. 69 p. Dissertação (Mestrado) — Universidade Federal de Lavras, Lavras, 2024.
- CASTELANI, E. V. *et al.* **A robust method based on LOVO functions for solving least squares problems**. 2019.
- CHAMBERS, J. M. *et al.* **Graphical Methods for Data Analysis**. [S.l.]: Wadsworth & Brooks/Cole, 1983.

CHAMBERS, J. M. *et al.* **Graphical Methods for Data Analysis**. [S.l.]: Wadsworth & Brooks/Cole, 1983.

CHEN, C. *et al.* Fast iteratively reweighted least squares algorithms for analysis-based sparse reconstruction. **Medical Image Analysis**, v. 49, 11 2014.

CLEVELAND, W. S. **The Elements of Graphing Data**. Monterey, CA: Wadsworth, 1985.

CLIFF, A. D.; ORD, J. K. **Spatial processes**. [S.l.]: Pion, 1981. 21 p.

CRESSIE, N. **Statistics for Spatial Data**. Revised edition. [S.l.]: Wiley-Interscience, 1991. 920 p. (Wiley Series in Probability and Statistics).

CRESSIE, N.; MOORES, M. T. Spatial statistics. In: **Encyclopedia of mathematical geosciences**. [S.l.]: Springer, 2023. p. 1362–1373.

CRUZ, F. O. **Curiosidades sobre a Dengue**. 2025. Acesso em: 23 fev. 2025. Disponível em: <https://www.ioc.fiocruz.br/dengue/textos/curiosidades.html>.

DRAPPER, N. R.; SMITH, H. **Applied Regression Analysis**. New York: Wiley, 1998. 407 p.

DRUCK, S. *et al.* **Análise Espacial de Dados Geográficos**. Brasília: EMBRAPA, 2004.

EMERSON, J. D.; STRENIO, J. Boxplots and batch comparison. In: HOAGLIN, D. C.; MOSTELLER, F.; TUKEY, J. W. (Ed.). **Understanding Robust and Exploratory Data Analysis**. [S.l.]: Wiley, 1983. cap. 3.

FONTOURA, G. F. Regressão linear robusta: o método de telbs e uma aplicação a dados de e-commerce. 2015.

FOTHERINGHAM, A.; BRUNSDON, C.; CHARLTON, M. Geographically weighted regression: The analysis of spatially varying relationships. **John Wiley & Sons**, v. 13, 01 2002.

FOTHERINGHAM, A.; CHARLTON, M.; BRUNSDON, C. Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. **Environment and Planning A**, v. 30, p. 1905–1927, 02 1998.

FOX, J.; WEISBERG, S. Robust regression. **An R and S-Plus companion to applied regression**, v. 91, p. 6, 2002.

GARCIA, R. A.; SOARES-FILHO, B.; MORO, S. Modelagem espacial do desmatamento amazônico. **Cadernos do Leste**, v. 24, n. 24, 2024.

GATRELL, A. C. *et al.* Spatial point pattern analysis and its application in geographical epidemiology. **Transactions of the Institute of British Geographers**, p. 256–274, 1996.

GOLLINI, I. *et al.* Gwmodel: an r package for exploring spatial heterogeneity using geographically weighted models. **Journal of statistical software**, v. 63, p. 1–50, 2015.

HARRIS, P. *et al.* The use of geographically weighted regression for spatial prediction: an evaluation of models using simulated data sets. **Mathematical Geosciences**, v. 42, p. 657–680, 2010.

HAWKINS, D. M. **Identification of outliers**. [S.l.: s.n.], 1980. v. 11.

- HOAGLIN, D. C.; WELSCH, R. E. The hat matrix in regression and anova. **The American Statistician**, Taylor & Francis, v. 32, n. 1, p. 17–22, 1978.
- HOEF, J. M. V.; CRESSIE, N. Multivariable spatial prediction. **Mathematical Geology**, v. 25, p. 219–240, 1993.
- HUBER, P. J. **Robust Statistics**. [S.l.]: Wiley, 1981. (Wiley series in probability and mathematical statistics).
- IBGE. **Estimativas da população residente com data de referência 1º de julho de 2024**. 2024. <https://www.ibge.gov.br>.
- LEUNG, Y.; MEI, C.-L.; ZHANG, W.-X. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. **Environment and planning A**, v. 32, n. 1, p. 9–32, 2000.
- LIN, C.-H.; WEN, T.-H. Using geographically weighted regression (gwr) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. **International Journal of Environmental Research and Public Health**, v. 8, p. 2798–2815, 2011.
- LU, B. *et al.* The gwmodel r package: further topics for exploring spatial heterogeneity using geographically weighted models. **Geo-spatial Information Science**, Taylor & Francis, v. 17, n. 2, p. 85–101, 2014. Disponível em: <https://doi.org/10.1080/10095020.2014.917453>.
- LU, B. *et al.* The gwmodel r package: further topics for exploring spatial heterogeneity using geographically weighted models. **Geo-spatial Information Science**, v. 17, n. 2, p. 85–101, 2014.
- LUZARDO, A. J. R.; FILHO, R. M. C.; RUBIM, I. B. Análise espacial exploratória com o emprego do índice de moran. **GEOgraphia**, v. 19, n. 40, p. 161–179, 2017.
- MANUEL, L. **Modelos de regressão linear com efeitos espaciais na análise da mortalidade infantil**. Dissertação (Mestrado) — Programa de Pós-graduação em Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, Lavras, 2011.
- MARONNA RICARDO A. E MARTIN, D. R. e. Y. V. J. **Robust Statistics: Theory and Methods**. Chichester, England: John Wiley & Sons Ltd., 2006.
- MARQUES, A. J. Aplicações de técnicas de autocorrelação espacial para avaliação da organização das habitações nas mesorregiões do estado do paran . **Observatrium: Revista Eletr nica de Geografia**, v. 1, n. 3, p. 89–105, 2009.
- MCGILL, R.; TUKEY, J. W.; LARSEN, W. A. Variations of box plots. **The American Statistician**, v. 32, p. 12–16, 1978.
- MURRELL, P. **R Graphics**. [S.l.]: Chapman & Hall/CRC Press, 2005.
- NEVES, M. C. *et al.* An lise explor t ria espacial de dados s cio-econ micos de s o paulo. **Salvador: GIS Brasil2000**, p. 1–11, 2000.
- OLIVEIRA, T. D. Identificando pontos de alavancagem em li  es aprendidas de projetos: uma perspectiva sist mica. 2020.

ONU. **Objetivos de Desenvolvimento Sustentável**. 2025. <https://brasil.un.org/pt-br/sdgs>. Acesso em: 16 set. 2025.

OPAS E OMS. **Dengue**. 2024. <https://www.paho.org/pt/topicos/dengue>. Acesso em: 19 set. 2024.

PEBESMA, E. *et al.* **sf: Simple Features for R**. 2024.

RAHAYU, L. *et al.* Unraveling geospatial determinants: Robust geographically weighted regression analysis of maternal mortality in indonesia. **Infolitika Journal of Data Science**, v. 1, n. 2, p. 73–81, 2023.

RAHMAN, A. S. **Robust Geographically Weighted Regression Modeling Using the Methods of MM Estimation and Least Absolute Deviation**. Dissertação (Master's thesis) — Universitas Hasanuddin, Makassar, Indonesia, 2024.

RENCHE, A. C.; SCHAALJE, G. B. **Linear Models in Statistics**. 2nd. ed. Hoboken, NJ: John Wiley & Sons, 2008.

ROUSSEEUW, P. J.; LEROY, A. M. **Robust regression and outlier detection**. [S.l.]: John Wiley & Sons, 1987. (Wiley series in probability and mathematical statistics).

ROYSTON, P. Algorithm as 181: The w test for normality. **Applied Statistics**, v. 31, p. 176–180, 1982.

ROYSTON, P. An extension of shapiro and wilk's w test for normality to large samples. **Applied Statistics**, v. 31, p. 115–124, 1982.

ROYSTON, P. Remark as r94: A remark on algorithm as 181: The w test for normality. **Applied Statistics**, v. 44, p. 547–551, 1995.

SAEFUDDIN, A.; SETIABUDI, N. A.; FITRIANTO, A. On comparison between logistic regression and geographically weighted logistic regression: with application to indonesian poverty data. **World applied sciences journal**, v. 19, n. 2, p. 205–210, 2012.

SAÚDE. **Plisa health information platform for the americas: dengue indicators portal**. Washington, D.C., 2024. Acesso em: 23 set. 2024. Disponível em: <https://www3.paho.org/data/index.php/en/mnu-topics/indicadores-dengue-en.html>.

SAÚDE. **Fotos do mosquito da dengue**. 2025. Acesso em: 23 fev. 2025. Disponível em: <https://www.mdsaude.com/doencas-infecciosas/fotos-mosquito-dengue/>.

SCALON, J. D. **Análise de Dados Espaciais com Aplicações em R**. 1. ed. Lavras, MG: Editora UFLA, 2024.

SCHMID, T.; MÜNNICH, R. T. Spatial robust small area estimation. **Statistical Papers**, v. 55, p. 653–670, 2014.

SILVA, N. *et al.* Análise de dados de área aplicada a dois indicadores econômicos de mesorregiões do estado de minas gerais. **Revista Brasileira de Biometria**, v. 29, n. 3, p. 369–395, 2011.

- SONG, Y. *et al.* Robust variable selection with exponential squared loss for the spatial autoregressive model. **Computational Statistics & Data Analysis**, v. 155, p. 107094, 2021.
- SOUSA, I. R.; BARRETO, G. A. Identificação arx l_{pv} robusta a outliers via estimador-m. In: **Simpósio Brasileiro de Automação Inteligente-SBAI**. [S.l.: s.n.], 2023. v. 1, n. 2.
- SU, X.; YAN, X.; TSAI, C.-L. Linear regression. **Wiley Interdisciplinary Reviews: Computational Statistics**, Wiley Online Library, v. 4, n. 3, p. 275–294, 2012.
- SULEKAN, A.; JAMALUDIN, S. S. S. Review on geographically weighted regression (gwr) approach in spatial analysis. **Malays J Fundam Appl Sci**, v. 16, n. 2, p. 173–7, 2020.
- TAUIL, P. L. Urbanização e ecologia do dengue. **Cadernos de Saúde Pública**, v. 17, p. S99–S102, 2001.
- TEAM, R. C. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>.
- TEIXEIRA, M. B. dos R.; SCALON, J. D. Comparac ao entre estimadores de semivari^ancia. **Rev. Bras. Biom**, v. 31, n. 2, p. 248–269, 2013.
- TUKEY, J. W. **Exploratory Data Analysis**. [S.l.]: Addison-Wesley, 1977. Section 2C.
- VELLEMAN, P. F.; HOAGLIN, D. C. **Applications, Basics and Computing of Exploratory Data Analysis**. [S.l.]: Duxbury Press, 1981.
- WALLER, L. A.; GOTWAY, C. A. **Applied Spatial Statistics for Public Health Data**. [S.l.]: John Wiley & Sons, 2004. (Wiley Series in Probability and Statistics).
- WHO. **Dengue and severe dengue**. 2023. Acessado em: 20 jan. 2025. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>.
- WICKHAM, H. *et al.* **tidyverse: Easily Install and Load the 'Tidyverse'**. 2019. R package version 1.3.1. Disponível em: <https://CRAN.R-project.org/package=tidyverse>.
- ZAPAROLI, I. C. V. B. *et al.* Resposta dos casos de dengue em função do clima no estado de são paulo. **Brazilian Journal of Development**, v. 7, n. 3, p. 28572–28587, 2021.
- ZHOU, Q.; WANG, C.; FANG, S. Application of geographically weighted regression (gwr) in the analysis of the cause of haze pollution in china. **Atmospheric Pollution Research**, v. 10, p. 835–846, 2019.

APÊNDICES

APÊNDICE A – APÊNDICE

.1 Códigos para Modelagem

.1.1 Script em R

Listing 1 – Código R para modelagens de dengue

```
# Codigos para fazer as modelagens dos modelos
#-----Pacotes
library(scales)
library(ggspatial)
library(gridExtra)
library(grid)
library(ggplot2)
library(skimr)
#----- Carregando os dados
-----
#
=====

source("juncao_todas_bases.R")

# Para nao alterar os dados originais
dados_23 <- base_final

# Visualizacao descritiva da base de dados
boxplot(dados_23$Inc_100k)
hist(dados_23$Inc_100k)
hist(dados_23$area_urbanizada)
hist(dados_23$urbanizacao_de_vias_publicas)
hist(dados_23$arborizacao_de_vias_publicas)
hist(dados_23$esgotamento_sanitario_adequado)
skim(dados_23)
plot(
  as.numeric(dados_23$total),
  na.rm = TRUE,
  type = "l",
  xlab = "Municipios",
  ylab = "Casos_notificados",
```

```

main = "Casos_de_dengue_notificados_por_municipio_de_Minas_
      Gerais",
col = "red"
)

# Verificando os outliers
iqr <- IQR(dados_23$Inc_100k, na.rm = TRUE)

# Determinar os limites superior e inferior
q1 <- quantile(dados_23$Inc_100k, 0.25, na.rm = TRUE)
q3 <- quantile(dados_23$Inc_100k, 0.75, na.rm = TRUE)
lower_limit <- q1 - 1.5 * iqr
upper_limit <- q3 + 1.5 * iqr
# Identificar outliers
outliers <- dados_23$Inc_100k[dados_23$Inc_100k < lower_
  limit | dados_23$Inc_100k > upper_limit]
# Exibir os outliers
length(outliers)
#
-----

#---- Adicionar os outliers no conjunto de dados
dados_23$outliers <- ifelse(dados$Inc_100k < lower_limit |
  dados$Inc_100k > upper_limit, 1, 0)

# Plotando os outliers

ggplot(dados_23) +
  geom_sf(aes(fill = outliers)) +
  scale_fill_distiller(palette = "Greens", direction = 1, name
    = "Incidencia") +
  annotation_north_arrow(style = north_arrow_fancy_
    orienteering(), location = "tr") +
  labs(title = "Casos_considerados_extremos_acima_de_4012_por_
    municipio_", x = "Longitude", y = "Latitude") +
  annotation_scale() +
  theme(axis.text.x = element_blank(), axis.text.y = element_
    blank())

#-----Modelo linear multiplo
-----
modelo_lm_2 <- lm(Inc_100k ~ arborizacao_de_vias_publicas +
  esgotamento_sanitario_adequado +
  area_urbanizada +
  urbanizacao_de_vias_publicas,
  data = dados_23)
summary(modelo_lm_2)

```

```

qqPlot(modelo_lm_2$residuals, main = "Q-Q_plot_dos_residuos_
      log_normal", col = "orange")
vif(modelo_lm_2)
summary(modelo_lm_2)
# Teste de normalidade dos residuos
library(lmtest)
library(car)
shapiro.test(residuals(modelo_lm_2))
# teste de homocedasticidade dos residuos
bptest(modelo_lm_2)
# teste de independencia dos residuos
dwtest(modelo_lm_2)
#-----Modelo robusto com MM-Estimador
-----

## regressao robusta com MM-Estimador pelo metodo iterativo
IRLS
library(MASS) # Para usar a funcao rlm
library(robustbase)

# Ajustar o modelo robusto
<- lmrob(Inc_100k ~ arborizacao_de_vias_publicas +
esgotamento_sanitario_adequado +
area_urbanizada +
urbanizacao_de_vias_publicas,
data = dados_23,
method = "MM")
summary(modelo_rb)

# -----Matriz de vizinhanca
-----

nb <- poly2nb(dados_23$geom, queen=TRUE) # Criacao da matriz
de vizinhanca do tipo Rainha
lw <- nb2listw(nb, style="W", zero.policy=TRUE) # Criacao da
matris de pesos espaciais
plot(nb, dados_23, col="red", lwd=2) # Visualizacao da
matriz de vizinhanca
moran.test(dados_23$Inc_100k, lw, randomisation = TRUE)
moran.plot(dados_23$Inc_100k, lw)

# 1. Rodar o Moran Plot
mp <- moran.plot(dados_23$Inc_100k, lw, labels = dados_23$NM
_MUN, pch = 19)

# 2. Extrair os residuos padronizados e valores de defasagem
espacial
resid_pad <- scale(dados_23$Inc_100k)
lag_val <- lag.listw(lw, resid_pad)

```

```

# 3. Identificar possiveis outliers
outliers <- which(abs(resid_pad) > limite | abs(lag_val) >
  limite)

# 4. Criar uma coluna no seu objeto sf indicando outlier
dados_23$outlier <- "Normal"
dados_23$outlier[outliers] <- "Outlier"
table(dados_23$outlier)
# 5. Plotar o mapa
ggplot(dados_23) +
  geom_sf(aes(fill = outlier), color = "white") +
  scale_fill_manual(values = c("Normal" = "grey80", "Outlier"
    = "red")) +
  labs(title = "Outliers_do_Moran_Plot_no_mapa",
    fill = "Classificacao") +
  theme_minimal()

#----- Plotar os clustes - LISA
local_moran1 <- localmoran(dados_23$soma, lw)
dados_23$local_moran <- local_moran1[, 1]
dados_sf <- st_as_sf(dados_23, coords = c("x", "y"), crs =
  4326)

# Visualizar o Indice local de Moran no mapa
tm_shape(dados_sf) +
tm_bubbles(col = "local_moran", palette = "-RdBu", size =
  0.2, alpha = 0.6) +
tm_layout(
  title = "Indice_Local_de_Moran_da_Incidencia_de_Dengue",
  legend.title.size = 1.2, # Ajusta o tamanho do titulo da
    legenda
  legend.position = c("left", "bottom") # Posicao correta da
    legenda
)
tm_shape(dados_sf) +
tm_bubbles(col = "local_moran",
  palette = "RdYlBu", # Melhor paleta para contraste
  size = 0.3, # Aumentar tamanho dos pontos
  border.col = "black", # Adicionar borda para contraste
  border.lwd = 0.2, # Ajustar largura da borda
  alpha = 0.8) + # Melhor transparencia
tm_layout(
  title = "Indice_Local_de_Moran_da_Incidencia_de_Dengue",
  title.position = c("center", "top"), # Melhorar
    posicionamento do titulo
  title.size = 1.5, # Aumentar o tamanho do titulo
  legend.title.size = 1.2,

```



```

legend.text.size = 0.9, # Ajustar tamanho do texto da
  legenda
legend.position = c("right", "bottom") # Melhor posicao para
  legenda
)

# Criar categorias de clusters
dados_sf$cluster <- factor(
  with(dados_sf, ifelse(local_moran > 0 & p_valor < 0.05 & Inc_100k > mean(Inc_100k), "Alto-Alto",
ifelse(local_moran > 0 & p_valor < 0.05 & Inc_100k <= mean(
  Inc_100k), "Baixo-Baixo",
ifelse(local_moran < 0 & p_valor < 0.05 & Inc_100k > mean(
  Inc_100k), "Alto-Baixo",
ifelse(local_moran < 0 & p_valor < 0.05 & Inc_100k <= mean(
  Inc_100k), "Baixo-Alto", "Nao_Significativo")))),
levels = c("Alto-Alto", "Baixo-Baixo", "Alto-Baixo", "Baixo-
  Alto", "Nao_Significativo")
)

#-----Spatial Autoregressive Model (Spatial Lag/SAR)
-----
modelo_sarl <- lagsarlm(formula = Inc_100k ~ arborizacao_de_
  vias_publicas +
  esgotamento_sanitario_adequado +
  area_urbanizada +
  urbanizacao_de_vias_publicas,
data = dados_23,
listw = lw,
zero.policy = TRUE)

#-----Modelo CAR
-----
modelo_car <- errorsarlm(formula = Inc_100k ~ arborizacao_de
  _vias_publicas +
  esgotamento_sanitario_adequado +
  area_urbanizada +
  urbanizacao_de_vias_publicas,
data = dados_23,
listw = lw,
zero.policy = TRUE)

#-----Modelo GWR
-----

# 5. Bandwidth otimo para GWR
# -----
baw.gwr <- bw.gwr(Inc_100k ~ arborizacao_de_vias_publicas +
  esgotamento_sanitario_adequado +
  area_urbanizada +
  urbanizacao_de_vias_publicas,

```

```

data = dados_23,
approach = "AICc",
kernel = "bisquare",
adaptive = TRUE)

# 6. GWR basico
gwr.cv_basic <- gwr.basic(Inc_100k ~ arborizacao_de_vias_
  publicas +
  esgotamento_sanitario_adequado +
  area_urbanizada +
  urbanizacao_de_vias_publicas,
data = dados_23,
bw = baw.gwr,
kernel = "bisquare",
adaptive = TRUE,
F123.test = TRUE)
gwr.cv_basic

# 8. GWR robusto (mantendo outliers)
rgwr.res_robust <- gwr.robust(Inc_100k ~ arborizacao_de_vias_
  _publicas +
  esgotamento_sanitario_adequado +
  area_urbanizada +
  urbanizacao_de_vias_publicas,
data = dados_23,
bw = baw.gwr,
kernel = "bisquare",
adaptive = TRUE,
F123.test = TRUE,
filtered = FALSE, # Abordagem automatica
cut1 = 2,          # Primeiro corte mais conservador
cut2 = 3,          # Segundo corte padrao
maxiter = 20)      # Numero suficiente de iteracoes

rgwr.res_robust

```