

Dener Rafael Quevedo de Matos

Regressão Geograficamente Ponderada: fatores que impactam o IDH em municípios do RS

Rio Grande, Rio Grande do Sul, Brasil

Fevereiro, 2025

Dener Rafael Quevedo de Matos

Regressão Geograficamente Ponderada: fatores que impactam o IDH em municípios do RS

Trabalho de Conclusão de Curso apresentado ao Instituto de Matemática, Estatística e Física da Universidade Federal do Rio Grande como requisito parcial para a conclusão do curso de Matemática Aplicada Bacharelado.

Universidade Federal do Rio Grande - FURG

Instituto de Matemática, Estatística e Física - IMEF

Curso de Matemática Aplicada Bacharelado

Orientador: Professora Doutora Raquel da Fontoura Nicolette

Rio Grande, Rio Grande do Sul, Brasil

Fevereiro, 2025

Dener Rafael Quevedo de Matos

Regressão Geograficamente Ponderada: fatores que impactam o IDH em municípios do RS

Trabalho aprovado

Documento assinado digitalmente
 **RAQUEL DA FONTOURA NICOLETTE**
Data: 10/02/2025 09:35:35-0300
Verifique em <https://validar.iti.gov.br>

**Prof^a Doutora Raquel da Fontoura
Nicolette**
(Orientadora - IMEF- FURG)

Documento assinado digitalmente
 **DEBORA SPENASSATO**
Data: 07/02/2025 12:23:24-0300
Verifique em <https://validar.iti.gov.br>

Prof^a Doutora Débora Spenassato
(Avaliadora - IMEF - FURG)

Documento assinado digitalmente
 **WASHINGTON LUIZ FELIX CORREIA FILHO**
Data: 07/02/2025 13:42:53-0300
Verifique em <https://validar.iti.gov.br>

**Avaliador - Dr. Washington Luiz Felix
Correia Filho**

Rio Grande, 06 de Fervereiro de 2025.

Rio Grande do Sul, Brasil

Agradecimentos

Agradeço à minha mãe, Deniz Matos, e ao meu irmão, Joel Matos, por me aturarem durante todo o meu tempo de faculdade.

Agradeço à minha namorada, Julia Sampaio, por todas às vezes que tentou estudar comigo.

Gostaria de agradecer também à minha orientadora, doutora Raquel Nicolette, que aceitou me orientar e muitas vezes me ajudou com conselhos que, com certeza, levarei para a vida.

Gostaria de agradecer também ao meu amigo, Rafael Jardim, que entrou no mesmo ano que eu e nunca desistiu – nossa frase acabou sendo: "Entramos juntos, sairemos juntos".

Agradeço também a todos os meus colegas, pelas conversas descontraídas e pelas trocas de conhecimento.

Agradeço a todos os professores que passaram pela minha vida acadêmica, pois alguns de seus ensinamentos com certeza serão levados para a vida toda.

*“Não importa a velocidade, siga em frente.”
(Confúcio)*

Resumo

O estudo apresenta o modelo de Regressão Geograficamente Ponderada (RGP) em alternativa ao modelo de Regressão Linear Múltipla (RLM), quando há presença de heterocedasticidade espacial. A RGP é utilizada para analisar os fatores que influenciam o Índice de Desenvolvimento Humano (IDH) nos municípios do Rio Grande do Sul. A pesquisa utiliza dados socioeconômicos obtidos no Atlas Brasil e cruzados com as coordenadas geográficas dos estados obtidos no IBGE, considerando o ano de 2017.

A metodologia inclui a aplicação dos modelos aos dados, a análise de variáveis preditoras e a avaliação da adequação de cada modelo à realidade espacial dos municípios. A RLM foi utilizada como modelo inicial, apresentando heterocedasticidade espacial, levando a utilização do modelo RGP. A RGP foi empregada para capturar a variação espacial dos fatores que impactam o Índice de Desenvolvimento Humano dos Municípios (IDHM), permitindo uma análise mais detalhada da heterogeneidade regional.

Os resultados indicam que a RGP oferece uma representação mais precisa das influências espaciais sobre o IDHM, enquanto a RLM apresenta limitações por não considerar a variabilidade geográfica. O estudo conclui que a escolha do modelo influencia significativamente a interpretação dos dados e destaca a importância de abordagens estatísticas que levem em conta a dimensão espacial no estudo de indicadores socioeconômicos.

Palavras-chaves: Regressão Geograficamente Ponderada; Regressão Linear Múltipla; Índice de Desenvolvimento Humano;

Abstract

The study presents the Geographically Weighted Regression (GWR) model as an alternative to the Multiple Linear Regression (MLR) model when spatial heteroscedasticity exists. GWR is used to analyze the factors that influence the Human Development Index (HDI) in the municipalities of Rio Grande do Sul. The research uses socioeconomic data obtained from the Atlas Brasil and crossed with the geographic coordinates of the states received from the IBGE, considering the year 2017.

The methodology includes the application of the models to the data, the analysis of predictor variables, and the assessment of the adequacy of each model to the spatial reality of the municipalities. The GWR was used as an initial model, presenting spatial heteroscedasticity, leading to the use of the GWR model. GWR was used to capture the spatial variation of the factors that impact the Human Development Index of the Municipalities (HDI-M), allowing a more detailed analysis of regional heterogeneity.

The results indicate that the RGP offers a more accurate representation of spatial influences on the HDI, while the RLM has limitations because it does not consider geographic variability. The study concludes that the choice of model significantly influences the interpretation of the data and highlights the importance of statistical approaches that take into account the spatial dimension in the study of socioeconomic indicators.

Keywords: Geographically Weighted Regression; Multiple Linear Regression; Human Development Index-Brazil.

Sumário

Introdução	8
1 OBJETIVOS	10
1.1 Objetivos Gerais	10
1.2 Objetivos Específicos	10
2 FUNDAMENTAÇÃO MATEMÁTICA	11
2.1 Modelos de Regressão	11
2.2 Modelo de Regressão Linear Múltipla	12
2.2.1 Mínimos Quadrados Ordinários(MQO)	12
2.2.2 Propriedades Numéricas do Estimador de MQO	14
2.2.3 Medidas de Ajustamento no Modelo de Regressão Linear Múltipla	15
2.2.4 Definição Matricial da RLM	16
2.2.5 Pressupostos do modelo RLM	17
2.3 Modelo de Regressão Geograficamente Ponderada (RGP)	17
2.4 Testes para verificação dos pressupostos do RLM	22
2.4.1 Teste de Kolmogorov-Smirnov	22
2.4.2 Teste de Durbin-Watson	24
2.4.3 Teste de Breusch-Pagan	24
2.4.4 Testes de comparação de modelos RLM e RGP	25
3 METODOLOGIA	27
3.1 Tipo de pesquisa	27
3.1.1 Dados	27
3.1.2 Período de análise	27
3.1.3 População e amostra	27
3.1.4 Critérios de Inclusão	28
3.2 Análise de Dados	28
3.2.1 Filtragem	28
3.2.2 Análise Exploratória de Dados	29
3.3 Suporte Computacional	29
4 RESULTADOS E DISCUSSÃO	30
4.1 Regressão Linear Múltipla	32
4.2 Regressão Geograficamente Ponderada	36
5 CONCLUSÃO	39

REFERÊNCIAS	40
A ANEXO - FIGURAS AUXILIARES	41
B ANEXO - CÓDIGO EM R	50

Introdução

Existem muitos questionamentos sobre onde a Matemática é usada, não há somente uma resposta certa, pois ela é utilizada em muitas áreas e uma delas é a do escopo desse trabalho: matemática aplicada ao geoprocessamento. Assim, são apresentadas a seguir alguns exemplos da sua aplicação na área:

- **Modelagem e Análise Espacial:** Utiliza-se a matemática para criar modelos que representam fenômenos geográficos. Por exemplo, a álgebra linear é usada para manipular matrizes de dados espaciais, enquanto o cálculo diferencial e integral pode ser aplicado na análise de superfícies e volumes.
- **Estatística:** Métodos estatísticos são essenciais para a análise de dados geográficos. Técnicas como interpolação espacial, regressão e análise de variância ajudam a entender padrões e tendências nos dados.
- **Geometria:** A geometria é usada para representar e manipular formas e estruturas espaciais. Por exemplo, a geometria computacional é aplicada em algoritmos de triangulação e na criação de mapas.
- **Algoritmos e Computação:** Muitos processos de geoprocessamento dependem de algoritmos matemáticos para realizar operações como a sobreposição de camadas, a análise de redes e a simulação de fenômenos naturais.
- **Transformações e Projeções Cartográficas:** A matemática é crucial para converter coordenadas geográficas em diferentes sistemas de projeção, permitindo a criação de mapas precisos e a comparação de dados de diferentes fontes.

O geoprocessamento é um conjunto de técnicas e ferramentas utilizadas para coletar, armazenar, processar, analisar e apresentar dados geográficos. Ele é amplamente utilizado em diversas áreas, como planejamento urbano, gestão ambiental, agricultura, entre outras. O objetivo principal do geoprocessamento é transformar dados espaciais em informações úteis para a tomada de decisões (ROSA; BRITO, 2013).

Este trabalho foca na comparação entre dois métodos de regressão: Regressão Linear Múltipla (RLM) e Regressão Geograficamente Ponderada (RGP). A RLM, amplamente utilizada na modelagem estatística, mas o modelo RLM apresenta limitações ao desconsiderar variações espaciais. Já a RGP incorpora a heterogeneidade espacial, oferecendo uma análise mais detalhada da influência dos fatores em diferentes localidades.

Queremos observar as relações entre variáveis socioeconômicas e o IDHM. Com essa comparação de métodos o intuito é analisar qual deles incorpora mais informação na análise da variável resposta IDHM.

1 Objetivos

1.1 Objetivos Gerais

Aplicar o modelo de Regressão Geograficamente Ponderada (RGP) em alternativa ao modelo de Regressão Linear Múltipla (RLM) quando os dados apresentam heterocedasticidade espacial.

1.2 Objetivos Específicos

- Aplicar ambos os modelos (RGP e RLM) aos dados socioeconômicos e geográficos dos municípios do Rio Grande do Sul, utilizando o IDHM como variável dependente.
- Identificar e interpretar as variáveis que mais contribuem para a variação do IDHM no modelo RGP com diferentes funções de ponderação de núcleo aplicadas.

2 Fundamentação Matemática

As causas subjacentes aos processos de mudança no uso e cobertura do solo variam significativamente conforme a região, o que torna difícil determinar a priori quais áreas possuem atividades semelhantes atuando de maneira homogênea. Portanto, a aplicação de métodos de análise local, como as Regressões Geograficamente Ponderada (RGP) (GWR, do inglês Geographically Weighted Regression), possibilita a identificação das variações espaciais desses processos. As RGP permitem reconhecer as flutuações espaciais entre uma variável dependente (por exemplo, taxa de ocupação urbana) e variáveis explicativas independentes (por exemplo, densidade populacional, área agrícola, entre outras).

O modelo de Regressão Geograficamente Ponderada (RGP) foi proposto por Brunsdon et al (1996), para explorar a não-estacionariedade espacial devido os modelos de regressão espacial global terem que contar com a forte estacionariedade para funcionarem corretamente (CARVALHO et al., 2006). O modelo RGP é uma extensão do modelo de regressão global baseado em regressão não-paramétrica e tem por objetivo analisar o efeito de variáveis preditoras na resposta à heterogeneidade espacial.

Este método foi escolhido por ser de regressão espacial local que considera a heterogeneidade espacial, ou seja, analisa a distribuição de um dado recurso em um ambiente determinado. A heterogeneidade espacial pode ser definida como quando uma única e igual variável explicativa resulta em respostas desiguais em diferentes regiões analisadas. Este método é utilizado para estudos de desenvolvimento humano, criminalidade e desempenho educacional, bem como uso e ocupação de solo, desmatamento e tantas outras áreas. Como estamos fazendo um estudo de desenvolvimento humano, esse foi um dos critérios adotados para a escolha do método, outro critério foi ele ser um modelo o qual podemos fazer o estudo sobre uma área específica, diferente do modelo global o qual veria uma cidade como um único ponto.

Com o intuito de se definir a Regressão Geograficamente Ponderada, primeiro será apresentado uma breve introdução a modelos de Regressão focando na Regressão Linear Múltipla que serve de base para se definir em detalhes RGP, foco deste estudo.

2.1 Modelos de Regressão

Os modelos de regressão são amplamente aplicados nas mais diversas áreas do conhecimento, como: saúde, economia, sociologia, ambiental, entre outras. O objetivo dos modelos de regressão é obter uma equação na qual a variável resposta possa ser descrita pelas variáveis preditoras.

2.2 Modelo de Regressão Linear Múltipla

Como nas estimações com uma única variável explicativa, é utilizado, nos modelos de regressão múltipla, o estimador de mínimos quadrados ordinários (MQO). Logo, considerando o modelo de regressão linear múltipla genérico para k variáveis explicativas como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i, \quad (2.1)$$

Nesse caso, temos $k + 1$ parâmetros, com:

- β_0 : intercepto;
- $\beta_1, \beta_2, \dots, \beta_k$: coeficientes, que podem ser interpretados como "inclinações", embora, na prática, elas não representem exatamente a inclinação da função;
- ϵ_i : termo de erro.

A hipótese-chave é definida como:

$$\mathbb{E}[\epsilon_i | X_{1i}, X_{2i}, \dots, X_{ki}] = 0. \quad (2.2)$$

A hipótese acima estabelece que o valor esperado do erro é o mesmo para qualquer combinação possível entre $X_1, X_2, \dots, X_k$. Desse modo, não existe correlação entre as variáveis no termo de erro ϵ_i e as variáveis explicativas $X_{1i}, X_{2i}, \dots, X_{ki}$ (CHEIN, 2019).

2.2.1 Mínimos Quadrados Ordinários(MQO)

O estimador empregado no modelo de regressão linear múltipla será obtido pelo método dos mínimos quadrados ordinários (MQO). Neste capítulo, as equações serão expandidas se utilizando 4 variáveis explicativas e no final apresentada uma generalização para n variáveis.

Assim, no caso de quatro variáveis explicativas, a equação estimada por MQO é a seguinte:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\epsilon}_i, \quad (2.3)$$

em que $\hat{\epsilon}_i$ é o resíduo da regressão.

O método de MQO escolhe os valores para os parâmetros desconhecidos que minimizam a soma dos quadrados dos resíduos da regressão. Com N observações de Y , X_1 ,

X_2 , X_3 , e X_4 , os parâmetros β_0 , β_1 , β_2 , β_3 , e β_4 são escolhidos simultaneamente, de modo a fazer com que o valor da expressão abaixo seja o menor possível:

$$\min \sum_{i=1}^N \hat{\epsilon}_i^2 = \sum_{i=1}^N \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \hat{\beta}_4 X_{4i} \right)^2. \quad (2.4)$$

A resolução matemática mostra que os estimadores de MQO satisfazem as equações normais:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_3 + \hat{\beta}_4 \bar{X}_4 \quad (2.5)$$

$$\sum_{i=1}^N Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^N X_{1i} + \hat{\beta}_1 \sum_{i=1}^N X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^N X_{1i} X_{2i} + \hat{\beta}_3 \sum_{i=1}^N X_{1i} X_{3i} + \hat{\beta}_4 \sum_{i=1}^N X_{1i} X_{4i} \quad (2.6)$$

$$\sum_{i=1}^N Y_i X_{2i} = \hat{\beta}_0 \sum_{i=1}^N X_{2i} + \hat{\beta}_1 \sum_{i=1}^N X_{1i} X_{2i} + \hat{\beta}_2 \sum_{i=1}^N X_{2i}^2 + \hat{\beta}_3 \sum_{i=1}^N X_{2i} X_{3i} + \hat{\beta}_4 \sum_{i=1}^N X_{2i} X_{4i} \quad (2.7)$$

$$\sum_{i=1}^N Y_i X_{3i} = \hat{\beta}_0 \sum_{i=1}^N X_{3i} + \hat{\beta}_1 \sum_{i=1}^N X_{1i} X_{3i} + \hat{\beta}_2 \sum_{i=1}^N X_{2i} X_{3i} + \hat{\beta}_3 \sum_{i=1}^N X_{3i}^2 + \hat{\beta}_4 \sum_{i=1}^N X_{3i} X_{4i} \quad (2.8)$$

$$\sum_{i=1}^N Y_i X_{4i} = \hat{\beta}_0 \sum_{i=1}^N X_{4i} + \hat{\beta}_1 \sum_{i=1}^N X_{1i} X_{4i} + \hat{\beta}_2 \sum_{i=1}^N X_{2i} X_{4i} + \hat{\beta}_3 \sum_{i=1}^N X_{3i} X_{4i} + \hat{\beta}_4 \sum_{i=1}^N X_{4i}^2 \quad (2.9)$$

Com intercepto:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 - \hat{\beta}_4 \bar{X}_4 \quad (2.10)$$

e inclinações:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (X_{1i} - \bar{X}_1)(Y_i - \bar{Y})}{\sum_{i=1}^N (X_{1i} - \bar{X}_1)^2} \quad (2.11)$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^N (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\sum_{i=1}^N (X_{2i} - \bar{X}_2)^2} \quad (2.12)$$

$$\hat{\beta}_3 = \frac{\sum_{i=1}^N (X_{3i} - \bar{X}_3)(Y_i - \bar{Y})}{\sum_{i=1}^N (X_{3i} - \bar{X}_3)^2} \quad (2.13)$$

$$\hat{\beta}_4 = \frac{\sum_{i=1}^N (X_{4i} - \bar{X}_4)(Y_i - \bar{Y})}{\sum_{i=1}^N (X_{4i} - \bar{X}_4)^2} \quad (2.14)$$

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i$$

Na interpretação do modelo de regressão linear múltipla, como em:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i}$$

tem-se que:

- $\hat{\beta}_0$ é o valor de Y quando $X_1 = X_2 = X_3 = X_4 = 0$
- $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ e $\hat{\beta}_4$ são efeitos parciais
- $\delta \hat{Y}_i = \hat{\beta}_1 \delta X_{1i} + \delta \hat{\beta}_2 X_{2i} + \delta \hat{\beta}_3 X_{3i} + \delta \hat{\beta}_4 X_{4i}$
- se X_4 é mantido fixo então $\delta X_4 = 0$
 $\delta \hat{Y}_i = \hat{\beta}_1 \delta X_{1i} + \delta \hat{\beta}_2 X_{2i} + \delta \hat{\beta}_3 X_{3i}$
- se X_3 é mantido fixo então $\delta X_3 = 0$
 $\delta \hat{Y}_i = \hat{\beta}_1 \delta X_{1i} + \delta \hat{\beta}_2 X_{2i} + \delta \hat{\beta}_4 X_{4i}$
- se X_2 é mantido fixo então $\delta X_2 = 0$
 $\delta \hat{Y}_i = \hat{\beta}_1 \delta X_{1i} + \delta \hat{\beta}_3 X_{3i} + \delta \hat{\beta}_4 X_{4i}$
- se X_1 é mantido fixo então $\delta X_1 = 0$
 $\delta \hat{Y}_i = \delta \hat{\beta}_2 X_{2i} + \delta \hat{\beta}_3 X_{3i} + \delta \hat{\beta}_4 X_{4i}$
- X_1, X_2, X_3, X_4 são as variáveis independentes.

Se X_2, X_3, X_4 são mantidas fixas, o efeito parcial de X_1 é dado por:

$$\frac{\partial \hat{Y}}{\partial X_1} = \beta_1$$

De maneira similar, os efeitos parciais de X_2, X_3 e X_4 seriam dados, respectivamente, por β_2, β_3 , e β_4 , mantendo as outras variáveis fixas.

2.2.2 Propriedades Numéricas do Estimador de MQO

A partir do modelo de regressão linear múltipla (2.31) e dos resíduos estimados (2.16), podemos elencar as propriedades numéricas do estimador de MQO.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} \quad (2.15)$$

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i \quad (2.16)$$

As principais propriedades numéricas são as seguintes:

1. A soma dos resíduos é igual a zero:

$$\sum_{i=1}^n \hat{\epsilon}_i = 0 \quad (2.17)$$

$$\bar{y} = \bar{\hat{y}}$$

(2.18)

2. A covariância amostral entre cada variável independente e os resíduos de MQO é zero:

$$\sum_{i=1}^n \hat{\epsilon}_i X_{ij} = 0, j = 1, \dots, k \quad (2.19)$$

3. A covariância amostral entre cada variável independente e o valor predito de MQO é zero:

$$\sum_{i=1}^n \hat{y}_i X_{ij} = 0, j = 1, \dots, k \quad (2.20)$$

4. O ponto $(\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k, \bar{Y})$ dos dados sempre está na reta de regressão de MQO:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \dots + \hat{\beta}_k \bar{x}_k \quad (2.21)$$

2.2.3 Medidas de Ajustamento no Modelo de Regressão Linear Múltipla

Para verificar o ajuste do modelo de regressão são usadas duas medidas: o R^2 e o erro padrão.

O coeficiente R^2 é dado por:

$$R^2 = \frac{\text{SQE}}{\text{STQ}} \quad (2.22)$$

O R^2 é um valor que varia entre 0 e 1 e mede qual a fração da variância da variável resposta é explicada pelas variáveis preditoras. Sendo o (SQE) Soma dos quadrados explicados pela regressão e o (STQ) Soma total dos quadrados.

Assim tem-se:

Soma total dos quadrados(STQ)

$$\text{STQ} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \quad (2.23)$$

Soma dos quadrados explicados pela regressão(SQE)

$$\text{SQE} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (2.24)$$

Soma do quadrado dos resíduos(SQR)

$$\text{SQR} = \sum_{i=1}^n \hat{\epsilon}_i^2 \quad (2.25)$$

A variação total em Y é a soma da variação que foi explicada pela regressão com a variação que não foi explicada:

$$\text{STQ} = \text{SQE} + \text{SQR} \quad (2.26)$$

Pode-se, então, definir o R^2 como o quadrado do coeficiente de correlação entre o valor atual de Y_i e o valor predito $\hat{Y}_i$.

$$R^2 = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}}))^2}{(\sum_{i=1}^n (y_i - \bar{y})^2)(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2)} \quad (2.27)$$

2.2.4 Definição Matricial da RLM

A partir do modelo de regressão linear múltipla

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \hat{\beta}_k X_{ki} \quad (2.28)$$

A sua equivalente matricial é dada por:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\varepsilon} \quad (2.29)$$

onde

$$\hat{\mathbf{Y}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 2 & X_{12} & X_{22} & \dots & X_{k2} \\ \dots & & & & \\ n & X_{1n} & X_{2n} & \dots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

Desta forma, o estimador de MQO para o vetor de coeficientes $\boldsymbol{\beta}$ o qual minimizará o erro quadrático total do modelos é dado por:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y}) \quad (2.30)$$

2.2.5 Pressupostos do modelo RLM

Os estimadores de MQO $\hat{\beta}$ serão considerados o melhor estimador linear não-viesado (MELNV) ou em inglês *Best Linear Unbiased Estimator* (BLUE) caso os pressupostos do Teorema de Gauss-Markov sejam válidos. Segundo *"The Cambridge Dictionary of Statistics"*, B. S. Everitt, 2ª Edição:

"Um teorema que prova que, se os termos de erro em uma regressão múltipla tiverem a mesma variância e não estiverem correlacionados, os critérios dos parâmetros no modelo produzido pela estimativa de mínimos quadrados são melhores (no sentido de ter menor dispersão sobre a média) do que qualquer outro estimador linear imparcial."

Os pressupostos a serem considerados são:

- Lineariedade : a v.a. Y_i é uma função linear das variáveis preditoras (X_{ij} , $j = 1, \dots, k$);
- $E(\varepsilon_i) = 0$: os erros, ou resíduos, possuem esperança condicional zero;
- Homocedasticidade: os erros, ou resíduos, ε_i são variáveis aleatórias de variância constante;
- $E(\varepsilon_i \varepsilon_j) = 0$ para $i \neq j$ os erros são não-correlacionados;
- Normalidade: os erros são normalmente distribuídos;
- Ausência de multicolinearidade: as variáveis preditoras $X_1, X_2, \dots, X_n$ são não-correlacionadas.

2.3 Modelo de Regressão Geograficamente Ponderada (RGP)

A RGP, é uma variação do modelo de regressão linear múltipla. Da mesma forma que apresentado anteriormente, o modelo será desenvolvido com 4 variáveis preditoras (BRUNSDON; FOTHERINGHAM; CHARLTON, 2016).

$$\hat{Y}_i = \hat{\beta}_0(u_i, v_i) + \hat{\beta}_1(u_i, v_i)X_{1i} + \hat{\beta}_2(u_i, v_i)X_{2i} + \dots + \hat{\beta}_k(u_i, v_i)X_{ki} \quad (2.31)$$

Sendo Y a variável resposta e X_i as variáveis preditoras. Tem-se $\beta_1, \beta_2, \beta_3$, e β_4 como coeficientes parciais de regressão, sendo U_i e V_i , nossas coordenadas latitude e longitude

A partir do modelo de regressão linear multivariada não-espacial, onde a variável resposta Y é representada como uma combinação linear de diferentes variáveis preditoras X_i , obtém-se o valor estimado de Y para um ponto i a partir como segue:

$$Y_i = \beta_0(u_i, v_i) + \sum_{k=1}^p \beta_k(u_i, v_i) \cdot X_{ki} + \varepsilon_i, \quad (2.32)$$

em que:

- x_{ki} : k -ésima variável explicativa na localização i ;
- β_0 : intercepto global (valor constante);
- β_k : k -ésimo coeficiente angular global referente a x_k (valor constante);
- ε_i : erros independentes.

Para estimativa dos coeficientes angulares e do intercepto em um modelo de regressão linear convencional, utiliza-se normalmente o método dos mínimos quadrados, resultando na seguinte equação dos estimadores em notação matricial:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y}) \quad (2.33)$$

em que:

- $\hat{\beta}$: vetor dos coeficientes (intercepto e coeficientes angulares);
- X : matriz de observações das variáveis explicativas;
- Y : vetor de observações da variável dependente (explicada).

A Regressão Geograficamente Ponderada é capaz de analisar o efeito das variáveis preditoras na resposta à heterogeneidade espacial que pode estar presente quando se analisa dados onde as variáveis preditoras dão uma resposta desigual em diferentes locais. Desta forma, a Regressão Geograficamente Ponderada (RGP) pode ser descrita como uma técnica simples de extensão da regressão linear simples que permite que parâmetros locais sejam estimados, deixando de lado as estimativas globais. Assim, para cada ponto i , as estimativas são específicas e o modelo pode ser definido como:

$$Y_i = \beta_{0,i}(u_i, v_i) + \sum_{k=1}^p \beta_{ki}(u_i, v_i) \cdot x_{ki} + \varepsilon_i, \quad (2.34)$$

onde:

- $\beta_{0,i}$: intercepto local específico da localização i ;

- β_{ki} : valor do k -ésimo coeficiente angular local específico da localização i .

A partir das equações (2.32) e (2.34), percebe-se que a principal diferença entre os modelos está na extensão e generalização obtida para cada localização i , deixando de ser meramente uma observação e passando a ser um modelo específico.

Desta forma, o método dos Mínimos Quadrados Ponderados (MQP), que é uma extensão do método dos Mínimos Quadrados Ordinários (MQO), empregado para ajuste de modelos de regressão linear, é usado para a estimativa no modelo de RGP. De forma genérica, o estimador MQO constitui um caso particular do estimador MQP, em que os pesos das observações são todos iguais.

O estimador de Mínimos Quadrados Ponderados (MQP) tem por característica a sua eficácia quando a variância dos erros não é constante ao longo da escala do modelo. Este estimador é capaz de lidar com dados que não cumprem o pressuposto da homocedasticidade, ou seja, quando os erros têm variância constante ao longo do modelo. Quando se trata de dados de natureza espacial, é possível que estes não cumpram o pressuposto da homocedasticidade, sendo os estimadores de MQP uma alternativa. De modo geral, no método de Mínimos Quadrados Ponderados, os fatores de ponderação são aplicados em cada diferença ao quadrado antes da minimização, de forma que a falta de precisão de alguns preditores acaba recebendo uma penalidade maior do que outros.

No modelo de regressão não espacial (2.31), todos os pesos atribuídos para as observações são constantes, ou seja, a estimativa dos parâmetros é a mesma para todo o espaço, obtendo-se uma estimativa global. Já na RGP, os mínimos quadrados são ponderados para cada observação de acordo com sua proximidade ao ponto i , o que pode ser observado no modelo apresentado em 2.3 dependendo da matriz de ponderação W .

$$\hat{\beta}_i = (\mathbf{X}^T \mathbf{W}_i \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{W}_i \mathbf{Y}) \quad (2.35)$$

em que, analogamente, $\mathbf{W}_i$ é uma matriz quadrada cujos elementos da diagonal principal correspondem aos pesos de cada observação com relação ao ponto i , e os elementos restantes são nulos. Considerando-se agora pesos para cada ponto i , deve-se estender também a notação. Por $w_{i,j}$, denota-se o peso atribuído à observação (ponto) j em relação ao ponto i . Assim:

$$\mathbf{W}_i = \begin{bmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{bmatrix}. \quad (2.36)$$

Função de Ponderação Espacial e Definição da largura de Banda

Para completar o modelo de RGP basta que sejam definidas as matrizes $\mathbf{W}_i$. Isto pode ser feito através da escolha adequada de funções de ponderação espacial.

Um ponto importante é a definição da função de ponderação espacial, ou seja, a função (kernel) sobre cada ponto do espaço a qual determina todos os pontos da regressão local que é ponderada pela distância. Desta forma, pontos mais próximos do ponto central tem maior peso.

Na prática, os resultados da RGP são relativamente insensíveis à escolha da função de ponderação, mas são sensíveis à largura de banda da função de ponderação específica escolhida, de modo que a determinação do valor ideal da largura de banda seja necessária como parte do modelo RGP (FOTHERINGHAM; CHARLTON; BRUNSDON, 1997).

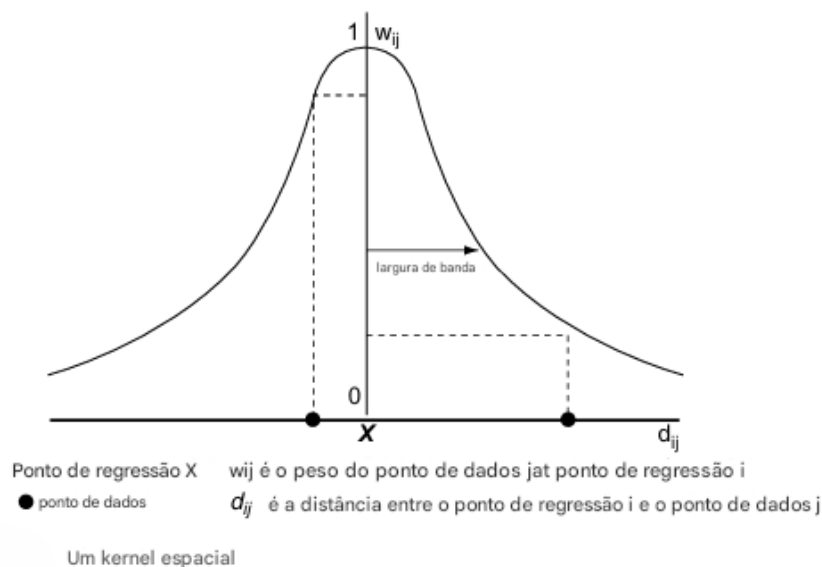


Figura 1 – Figura adaptada de Fotheringham, Charlton e Brunson (1997)

O modelo RGP, assume que pontos mais próximos do ponto de regressão tem maior chance de influenciá-lo, isso é efeito da função kernel espacial, que é uma função real contínua e simétrica, então a integral soma um. A função kernel define d_{ij} que é a distância entre duas localizações geográficas que representam regiões e um parâmetro de largura de banda que define um peso para w_{ij} entre esses dois pontos, que é relacionando inversamente a distância geográfica (SILVA, 2020).

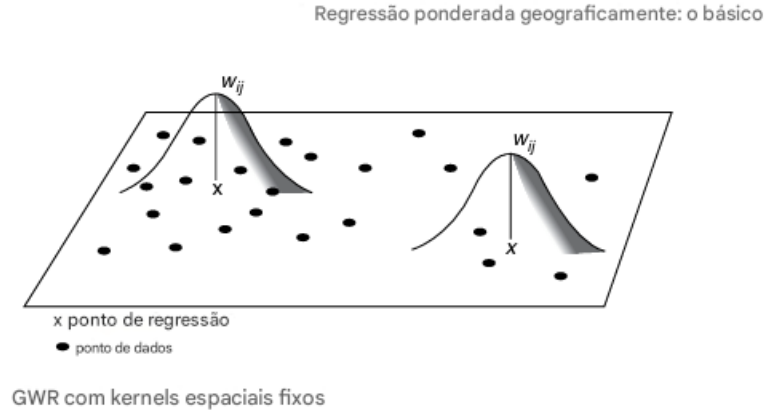


Figura 2 – Figura adaptada de [Fotheringham, Charlton e Brunson \(1997\)](#)

Considerando a matriz W no modelo de regressão não espacial, todas as observações têm peso unitário, o que corresponde a uma estimativa constante dos parâmetros ao longo do espaço, ou seja, ao modelo de regressão global.

$$w_{ij} = 1, \quad \forall j, i, \quad (2.37)$$

onde j representa um ponto específico no espaço no qual os dados são observados, e i , um ponto no qual os parâmetros são estimados.

Para que os pesos sejam ajustados de acordo com a natureza espacial dos dados é necessário que seja definida uma função de ponderação que é uma maneira de considerar pesos consistentes com a proximidade e, portanto, um modelo de características locais, o qual excluiriam no processo de calibração local, observações que estejam mais distantes do ponto em estudo do que uma determinada distância D , ou seja:

$$w_{ij} = \begin{cases} 1, & \text{se } d_{ij} < D, \forall A_{i,j} \\ 0, & \text{caso contrário.} \end{cases} \quad (2.38)$$

Seja d_{ij} a distância entre os pontos i e j ([FOTHERINGHAM; CHARLTON; BRUNSDON, 1997](#)), a inclusão de w_{ij} proporciona um modelo de regressão local ponderado geograficamente. Essa regressão local pode gerar problemas de descontinuidade, o que pode ser contornado definindo w_{ij} como uma função contínua da distância d_{ij} .

Uma escolha bastante popular é fixar uma função de núcleo gaussiana ([FOTHERINGHAM; CHARLTON; BRUNSDON, 1997](#)), pois quando i varia dentro da região de estudo, os coeficientes de regressão podem mudar significativamente, pois os pontos de observação podem ou não estar dentro da região circular ao redor de i .

Como a W_{ij} pode ser calculada pela função kernel existem 2 métodos para calcular a W_{ij} que são a Gaussiana e a bi-square (bi-quadrada), para cada um desses dois métodos existem as variações de ajustado e não ajustado (RANGKUTI; SUNUSI; TAHIR, 2023).

Gaussiana

$$W_{ij} = \exp \left(-\frac{1}{2} \left(\frac{d_{ij}}{b} \right)^2 \right)$$

bi-square

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

Para a escolha da largura de banda, Fotheringham et al. (2000) sugerem um processo de validação cruzada onde

$$SQE = \sum_{i=1}^n [y_i - y_{\neq i} \cdot (\beta)]^2$$

Onde:

- SQE - Soma dos quadrados dos erros;
- y_i - Valor observado da variável dependente no ponto i ;
- $y_{\neq i}$ - y_i ajustado sem o ponto i considerando o parâmetro (β ou $D_{influen\tilde{c}ia}$).

O valor escolhido de SQE é aquele que minimiza esta quantidade, ou seja, a escolha final dos parâmetros (β ou $D_{influen\tilde{c}ia}$) é aquela que minimiza o SQE .

2.4 Testes para verificação dos pressupostos do RLM

2.4.1 Teste de Kolmogorov-Smirnov

Teste de Kolmogorov-Smirnov foi originalmente proposto por Kolmogorov em 1933 (STEPHENS, 1992), enquanto a versão relacionada K_n^+ foi introduzida por Smirnov em 1939. Esses dois métodos são conhecidos como os testes Kolmogorov-Smirnov bilateral e unilateral, respectivamente.

Seja $X_1, X_2, \dots, X_n$ uma amostra aleatória composta por n observações provenientes de uma função de distribuição F , cuja forma é desconhecida. A função empírica de

distribuição associada a essa amostra é definida por:

$$F_n(x) = \begin{cases} 0 & \text{se } x < X_{(1)}, \\ \frac{i}{n} & \text{se } X_{(i)} \leq x < X_{(i+1)}, \\ 1 & \text{se } x \geq X_{(n)}, \end{cases}$$

onde $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ são as observações dispostas em ordem crescente.

Se houver uma função contínua H previamente definida, rejeitamos a hipótese de $F = H$ quando os valores de:

$$K_n = \sup_{-\infty < x < \infty} \sqrt{n} |F_n(x) - H(x)|$$

ou

$$K_n^+ = \sup_{-\infty < x < \infty} \sqrt{n} [F_n(x) - H(x)]$$

forem suficientemente altos. Especificamente, rejeitamos $F = H$ caso $K_n \geq d_n(\alpha)$, onde $P[K_n \geq d_n(\alpha) \mid F = H] = \alpha$, ou $K_n^+ \geq d_n^+(\alpha)$, com $P[K_n^+ \geq d_n^+(\alpha) \mid F = H] = \alpha$.

Esses testes, originalmente introduzidos por Kolmogorov e posteriormente ampliados por Smirnov, são conhecidos por avaliarem a qualidade do ajuste entre F_n e H . Informações mais detalhadas podem ser encontradas no trabalho de Darling, que também inclui uma extensa bibliografia.

A probabilidade de o teste K_n rejeitar $F = H$, caso a distribuição verdadeira seja G , é definida como:

$$P[K_n \geq d_n(\alpha) \mid F = G].$$

Quando consideramos uma sequência de distribuições alternativas $\{G_n\}$, o poder assintótico é expresso por:

$$\lim_{n \rightarrow \infty} P \left[\sup_{-\infty < x < \infty} \sqrt{n} |F_n(x) - H(x)| \geq d_n(\alpha) \mid F = G_n \right].$$

Para facilitar a análise, Doob propôs o uso do processo estocástico $Z_n(t)$, definido por:

$$Z_n(t) = \sqrt{n} [F_n(H_n^{-1}(t)) - t], \quad 0 \leq t \leq 1,$$

o que permite reescrever o poder assintótico como:

$$\lim_{n \rightarrow \infty} P \left[\sup_{0 < t < 1} |Z_n(t)| \geq d_n(\alpha) \mid F = G_n \right].$$

A condição $F = G_n$ pode ser omitida, já que todas as probabilidades pressupõem essa hipótese (QUADE, 1965); aonde as hipóteses desse teste são:

H0: Os resíduos do modelo são normalmente distribuídos

H1: Os resíduos do modelo não são normalmente distribuídos

Região Crítica H0 é rejeitado se valor $p < 0,05$.

2.4.2 Teste de Durbin-Watson

O teste de Durbin-Watson é uma ferramenta estatística usada para verificar a autocorrelação dos resíduos da regressão; definida por:

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

onde os e_t correspondem aos resíduos obtidos na regressão ajustada por meio do método dos mínimos quadrados ordinários.

Expandindo a equação obtemos:

$$d = \frac{\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 - 2 \sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$$

Para valores de n suficientemente grandes, O valor de d pode ser expresso de forma aproximada como $1 + 1 - 2r = 2(1 - r)$, em que r é o coeficiente de correlação entre e_t e e_{t-1} . Assim, o valor de d pode variar de zero (quando $r = 1$) até quatro (quando $r = -1$). Um d próximo de zero indica a presença de autocorrelação positiva nos resíduos, enquanto um d próximo de 4 sugere que os resíduos estão autocorrelacionados negativamente.

Hoje em dia, já existem softwares que calculam automaticamente a probabilidade caudal associada ao valor do teste de Durbin-Watson. Com isso, basta comparar essa probabilidade com o nível de significância escolhido para determinar se o resultado é significativo ou não, eliminando o risco de resultados inconclusivos (HOFFMANN, 2016)

As hipóteses para esse teste são:

H0: não há correlação nos resíduos

H1: há correlação nos resíduos

Região Crítica H0 é rejeitado se valor $p < 0,05$.

2.4.3 Teste de Breusch-Pagan

O teste de Breusch-Pagan (ou BP test) é um teste estatístico usado para verificar a presença de heterocedasticidade nos resíduos de um modelo de regressão linear. Heterocedasticidade ocorre quando a variância dos resíduos não é constante ao longo das observações, o que pode afetar a eficiência dos estimadores e a validade dos testes estatísticos. O teste foi proposto por Trevor Breusch e Adrian Pagan em 1979 (R Documentation, 2025). As hipóteses para esse teste são:

H0: não há heterocedasticidade nos resíduos

H1: há heterocedasticidade nos resíduos

Região Crítica H0 é rejeitado se valor $p < 0,05$.

2.4.4 Testes de comparação de modelos RLM e RGP

Os testes de Brunson, Fotheringham e Charlton (2002) e de Leung et al. (2000) são utilizados para realizar testes globais em modelos de Regressão Geograficamente Ponderada (RGP). Na biblioteca **spgwr** são disponibilizadas 4 testes de hipótese para a comparação dos modelos RLM e RGP: `lmz.f2GWR.test`, `lmz.f1GWR.test`, `BFC99.GWR.test`, `BFC02.GWR.test`. Especificamente, estes testes comparam a qualidade de ajuste entre um modelo de regressão linear e um modelo RGP, avaliando se o modelo RGP oferece uma melhoria significativa no ajuste dos dados em relação ao modelo global e cada um tem uma abordagem ligeiramente diferente para avaliar a significância da melhoria do RGP em relação ao outro modelo.

Os testes `lmz.f1GWR.test` e `lmz.f2GWR.test` são variações de testes F que comparam a soma dos quadrados dos resíduos (RSS) dos modelos dois modelos. A diferença entre eles está na forma como a estatística F é calculada. O teste `lmz.f2GWR.test` é uma versão mais robusta que leva em consideração a correlação espacial nos resíduos do modelo RGP, sendo sua estatística F ajustada para considerar a dependência espacial, o que pode ser mais apropriado para dados geográficos(?).

O `BFC99.RGP.test` é baseado no trabalho de **Brunson, Fotheringham e Charlton (1999)**. Ele é um dos primeiros testes propostos para comparar modelos RLM e RGP. A abordagem é a seguinte:

- **Hipótese nula:** O modelo linear é adequado (ou seja, não há vantagem em usar RGP).
- **Hipótese alternativa:** O modelo RGP é significativamente melhor que o linear.
- **Estatística de teste:**
 - Compara a soma dos quadrados dos resíduos (RSS) dos dois modelos.
 - Usa uma distribuição F para avaliar a significância.

Já o teste `BFC02.GWR.test` é uma evolução do `BFC99.GWR.test`, baseado no livro de **Brunson, Fotheringham e Charlton (2002)**. Esse Teste inclui melhorias para lidar com a heterogeneidade espacial e a dependência entre os resíduos, levando em conta que a variância dos resíduos pode não ser constante em todo o espaço (heterocedasticidade espacial), o tornando mais robusto para dados espaciais.

O único teste que apresenta a inclusão ou não de uma variável no modelo é o teste `LMZ.F3GWR.test` o qual é uma versão mais avançada dos testes F anteriores. Esse Teste é projetado para lidar com situações em que há heterogeneidade espacial e dependência

espacial nos resíduos do modelo RGP. Esse teste é particularmente útil quando se suspeita que a variância dos resíduos não é constante em todo o espaço (heterocedasticidade espacial) ou quando há correlação espacial entre os resíduos.

3 Metodologia

3.1 Tipo de pesquisa

3.1.1 Dados

O conjunto de dados utilizado para esta análise foi obtido no site do Atlas do Desenvolvimento Humano no Brasil “(<http://www.atlasbrasil.org.br/>)” e cruzados com as coordenadas geográficas dos estados obtidos no “(<https://www.ibge.gov.br/>)”. Ao incorporar considerações espaciais na estrutura de regressão, pretendemos capturar quaisquer dependências geográficas ou padrões espaciais que possam existir nos dados, fornecendo assim resultados de modelagem mais precisos e robustos.

3.1.2 Período de análise

Os dados disponíveis abrangem o período de 1991 a 2017, representando uma rica fonte de informações para análises temporais. No entanto, para a presente pesquisa, optamos por utilizar os dados referentes aos anos de 2010 e 2017. Essa decisão foi motivada pela presença de variáveis detalhadas nesses anos, as quais são fundamentais para o desenvolvimento e aprofundamento das análises propostas.

Por outro lado, os dados dos anos de 1991 e 2000 apresentavam uma limitação significativa: neles constava apenas o Índice de Desenvolvimento Humano Municipal (IDHM) já calculado, sem o detalhamento das variáveis que compõem o índice, como renda, longevidade e educação. Esse fato restringiu o potencial da análise para esses anos, uma vez que as variáveis desagregadas são cruciais para explorar dinâmicas sociais e econômicas mais complexas.

Dessa forma, o foco no ano de 2017 permitiu uma abordagem mais robusta e detalhada, garantindo maior consistência às conclusões e possibilitando uma análise comparativa mais rica sobre o período recente.

3.1.3 População e amostra

A população-alvo deste estudo abrange os municípios do estado do Rio Grande do Sul, localizados na região Sul do Brasil. Em 2010, a amostra utilizada contemplava um total de 496 municípios. No entanto, um evento administrativo relevante alterou esse panorama nos anos seguintes: o distrito de Pinto Bandeira, anteriormente vinculado ao município de Bento Gonçalves, iniciou em 2011 o processo formal de solicitação para sua emancipação política e administrativa.

Esse pedido foi aprovado em 2013, resultando na criação de um novo município. Assim, nos dados de 2017, Pinto Bandeira já aparece contabilizado como uma unidade municipal independente, elevando o total de municípios do Rio Grande do Sul para 497.

Essa mudança é relevante para a pesquisa, pois reflete as dinâmicas territoriais e administrativas do estado, além de impactar diretamente a análise longitudinal dos dados. Considerar essa transição é fundamental para garantir a consistência e a precisão dos resultados, especialmente em estudos que dependem da comparação de indicadores entre os períodos analisados.

Dessa forma, a inclusão de Pinto Bandeira como um município independente a partir de 2017 reforça a importância de compreender os contextos históricos e políticos que moldam a estrutura territorial do estado, assegurando que os resultados reflitam adequadamente a realidade de cada período.

3.1.4 Critérios de Inclusão

1 Localização Geográfica:

- Somente municípios do Rio Grande do Sul serão incluídos.

2 Municípios devem ter registros completos e confiáveis para todas as variáveis selecionadas

3.2 Análise de Dados

A análise de dados foi composta pelas seguintes etapas:

- Filtragem;
- Análise exploratória de dados;
- Construção e análise dos modelos.

3.2.1 Filtragem

O primeiro passo consistiu na leitura dos dados e a elaboração de uma planilha conjunta: dados do Atlas Brasil referente aos município do estado do Rio Grande do Sul acrescidos de uma coluna com as coordenadas das sedes administrativas de cada um desses municípios.

Com a base de dados consolidada, verificou-se no R a consistência dos mesmos. Verificação de valores ausentes, assegurando que as células vazias fossem corretamente

identificadas e representadas como NA, em conformidade com os padrões estabelecidos para valores faltantes.

3.2.2 Análise Exploratória de Dados

Os dados foram obtidos no site do Atlas do Desenvolvimento Humano no Brasil "<http://www.atlasbrasil.org.br/>" e cruzados com as coordenadas geográficas dos estados obtidos no IBGE.

3.3 Suporte Computacional

O trabalho utiliza o software R "<https://www.r-project.org/>" com a IDE Rstudio, e é desenvolvido no Laboratório de Estatística Ambiental (LEA) - IMEF.

Em questão de suporte computacional foi usado o software R que é um software livre de computação estatística e gráficos, onde qualquer usuário pode desenvolver um novo pacote e disponibilizá-lo para terceiros. Com intuito de analisar dados espaciais, como exemplo das bibliotecas usadas no R nesse projeto têm-se os pacotes `spgwr`, `readxl`, `car`, `lmtest` e `zoo`.

4 Resultados e Discussão

Do total de 497 municípios do estado do RS, após a exclusão de municípios com dados faltantes, restaram para a análise 490 município.

As tabelas 1, 2 e 3 representam todas as variáveis independentes com um total de 36 variáveis. A variável IDHM representa a variável Y e assim temos um total de 37 variáveis iniciais.

Tabela 1 – Descrição das Variáveis

Variável	Descrição
PANALF15	Percentual de analfabetos com 15 anos ou mais de idade no Cadastro Único.
2010	Índice de Desenvolvimento Humano Municipal de 2010.
PMATPUB_EF	Participação percentual das matrículas na rede pública do ensino fundamental no total de matrículas no ensino fundamental.
PMATPUB_EM	Participação percentual das matrículas na rede pública do ensino médio no total de matrículas no ensino médio.
DIST_EF_TOTAL	Participação percentual dos alunos que frequentam o ensino fundamental com atraso idade-série de dois anos ou mais no total de alunos que frequentam o ensino fundamental.
DIST_EM_TOTAL	Participação percentual dos alunos que frequentam o ensino médio com atraso idade-série de dois anos ou mais no total de alunos que frequentam o ensino médio.
DIST_EF_PUB	Participação percentual dos alunos que frequentam o ensino fundamental na rede pública com atraso idade-série de dois anos ou mais no total de alunos que frequentam o ensino fundamental na rede pública.
DIST_EM_PUB	Participação percentual dos alunos que frequentam o ensino médio na rede pública com atraso idade-série de dois anos ou mais no total de alunos que frequentam o ensino médio na rede pública.
PMATLINF_EF	Participação percentual dos alunos do ensino fundamental que estudam em escolas com laboratório de informática.
PMATLINF_EM	Participação percentual dos alunos do ensino médio que estudam em escolas com laboratório de informática.
PMATINT_EF	Participação percentual dos alunos do ensino fundamental que estudam em escolas com internet.
PMATINT_EM	Participação percentual dos alunos do ensino médio que estudam em escolas com internet.

Tabela 2 – Descrição das Variáveis

DOCSUP_EF_TOTAL	Participação percentual dos docentes do ensino fundamental com formação adequada.
DOCSUP_EM_TOTAL	Participação percentual dos docentes do ensino médio com formação adequada.
DOCSUP_EF_PUB	Participação percentual dos docentes na rede pública do ensino fundamental com formação adequada.
DOCSUP_EM_PUB	Participação percentual dos docentes na rede pública do ensino médio com formação adequada.
TXCOBPLP	Participação percentual da população coberta por planos de saúde suplementar no total da população.
PINTERSAP	Participação percentual de internações hospitalares por condições sensíveis à atenção primária no total de internações hospitalares.
PINTERDRSAI	Participação percentual de internações hospitalares por doenças relacionadas ao saneamento ambiental inadequado no total de internações hospitalares.
REN_TRPCBF_D	Valor médio mensal das transferências do programa Bolsa Família, dividido pela população total do município, em valores reais de 01/08/2010, corrigidos através do IPCA.
REN_TRPCBPC_D	Valor médio mensal das transferências do programa BPC, dividido pela população total do município. Valores em reais de 01/08/2010, corrigidos através do IPCA.
PDEFAGUA	Percentual de pessoas inscritas no Cadastro Único sem abastecimento de água adequado.
PDEFESGOTO	Percentual de pessoas inscritas no Cadastro Único sem esgotamento sanitário adequado.
PIND_POS	Percentual de extremamente pobres no Cadastro Único pós-Bolsa Família.
PMPOB_POS	Percentual de pobres no Cadastro Único pós-Bolsa Família.
PPOB_POS	Percentual de vulneráveis à pobreza no Cadastro Único pós-Bolsa Família.
PFLORA	Percentual da área do município que é coberta por vegetação natural. As florestas plantadas (Eucalipto e Pinus) não estão incluídas.
PFEM_VEREADOR	Participação percentual entre o número de vereadores do sexo feminino e o total de vereadores.

Tabela 3 – Descrição das Variáveis

IDHMRenda	Índice de Desenvolvimento Humano Municipal de Renda.
IDHMEdu	Índice de Desenvolvimento Humano Municipal de Educação.
IDHMLongevidade	Índice de Desenvolvimento Humano Municipal de Longevidade.
TXNASC7C	Percentual de nascidos vivos com pelo menos sete consultas de pré-natal.
TXBRUTAMORT	Taxa bruta de mortalidade.
REN_RENSF_D	Rendimento médio no setor formal.
PBF	Percentual de pessoas inscritas no Cadastro Único que recebem Bolsa Família.
POP_TOT	População total.
IDHM	Índice de Desenvolvimento Humano Municipal

4.1 Regressão Linear Múltipla

Como primeira alternativa usaremos a Regressão linear múltipla. O primeiro modelo incorpora as 36 variáveis independentes. O coeficiente de determinação R^2 , que é uma métrica que sintetiza o grau de ajuste da linha de regressão amostral em relação aos dados observados, foi calculado. Após fazermos esse primeiro modelo, observamos que o R^2 do modelo **formula0** explica 99,94%.

Foi verificada a multicolinearidade, a qual ocorre quando há uma forte correlação linear entre as variáveis independentes em um modelo de regressão. Essa condição pode aumentar as variâncias dos coeficientes estimados, resultando em instabilidade e dificultando a análise dos efeitos isolados de cada variável. Além disso, pode causar inconsistências na avaliação da significância estatística das variáveis independentes e do modelo como um todo. Um método amplamente utilizado para identificar a presença de multicolinearidade é o cálculo do Fator de Inflação da Variância (VIF) (PORTER, 2011). VIFs menores que 10 indicam que não há multicolinearidade entre as variáveis independentes.

Geralmente, um VIF acima de 4 ou uma tolerância abaixo de 0,25 indica que pode haver multicolinearidade, e uma investigação adicional é necessária. Quando o VIF é superior a 10 ou a tolerância é inferior a 0,1, há multicolinearidade significativa que precisa ser corrigida. Como houveram muitos valores acima de 10, o que significa uma multicolinearidade, decidiu-se retirar algumas variáveis, sendo as que apresentam multicolinearidade ou as que não possuem um sentido explícito para esse trabalho.

A tabela 4 representa as variáveis retiradas por apresentar multicolinearidade ou que complementavam outras variáveis.

Tabela 4 – Segue as siglas das variáveis removidas

POP_TOT	A2010	DIST_EF_PUB
DIST_EM_PUB	PMATLINF_EF	PMATINT_EF
PMATINT_EM	DOCSUP_EF_PUB	DOCSUP_EM_PUB
TXBRUTAMORT	PINTERSAP	PINTERDRSAI
IDHMEdu	REN_RENSF_D	REN_TRPCBPC_D
PFLORA	IDHMLongevidade	PFEM_VEREADOR
IDHMRenda	PMATLINF_EM	

Após a retirada dessas variáveis partiremos para um segundo modelo denominando-o de **formula1** para observarmos como as variáveis estão se comportando. Com isso obtivemos um $R^2=64,7\%$, ou seja, cerca de 64,7% da variação observada na variável dependente pode ser atribuída às variáveis independentes consideradas no modelo de regressão. Em outras palavras, o modelo consegue explicar uma parte significativa dessa variação. Isso indica que as variáveis independentes incluídas no modelo têm um bom poder explicativo e oferecem um ajuste satisfatório aos dados, sendo úteis para compreender as mudanças na variável dependente. Foi identificado ainda algumas variáveis com multicolinearidade. Foram retiradas da análise as variáveis **PPOB_POS** e **PMPOB_POS**, com isso partiremos para um terceiro teste para a verificação de que se realmente a multicolinearidade foi removida com sucesso das nossas variáveis.

Então agora partiremos para o modelo 3 que chamaremos de **formula2** e toda análise foi refeita. Cerca de 64,45% da variação observada na variável dependente pode ser atribuída às variáveis independentes consideradas no modelo de regressão. Em outras palavras, o modelo consegue explicar uma parte significativa dessa variação. Isso indica que as variáveis independentes incluídas no modelo têm um bom poder explicativo e oferecem um ajuste satisfatório aos dados, sendo úteis para compreender as mudanças na variável dependente. Neste conjunto de variáveis não foi constatada a multicolinearidade.

A tabela 5 é para representar as variáveis que ficaram após as análises da fórmula **formula2**. Estão descritas, novamente, as variáveis e suas descrições para maior fluidez na leitura.

Para exemplificar as relações do IDHM com as variáveis selecionadas, a figura 3 é apresentada. No eixo X (horizontal), temos o IDHM (Índice de Desenvolvimento Humano Municipal), que mede o nível de desenvolvimento dos municípios e no eixo Y (vertical), temos a variável TXNASC7C, que representa o percentual de nascidos vivos cujas mães

Tabela 5 – Variáveis finais sem Multicolinearidade

PMATPUB_EF	PMATPUB_EM
DIST_EF_TOTAL	DIST_EM_TOTAL
DOCSUP_EF_TOTAL	DOCSUP_EM_TOTAL
TXNASC7C	TXCOBPLP
REN_TRPCBF_D	PDEFAGUA
PDEFESGOTO	PANALF15
PIND_POS	PBF

realizaram pelo menos sete consultas de pré-natal. Como o IDHM é a variável dependente (ou seja, a que queremos explicar ou prever), este gráfico sugere que estamos analisando o impacto do acompanhamento pré-natal sobre o desenvolvimento humano dos municípios, cada ponto no gráfico representa um município do Rio Grande do Sul, o gráfico exibe uma grande dispersão, sugerindo que não há uma relação linear clara entre TXNASC7C e IDHM.

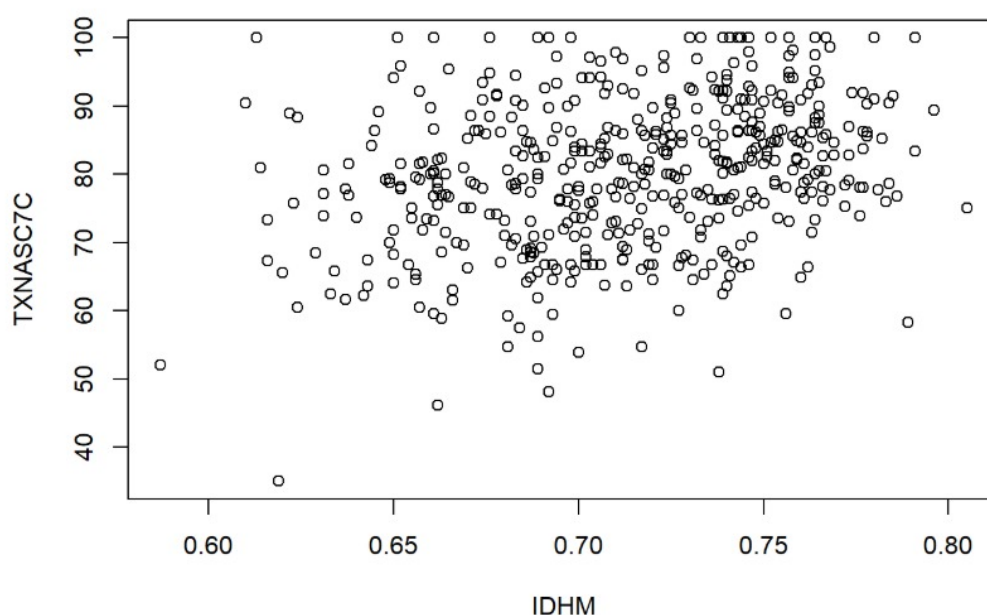


Figura 3 – TXNASC7C X IDHM

No entanto, é possível perceber uma leve tendência positiva: municípios onde um maior percentual de mães realizam pelo menos sete consultas de pré-natal tendem a ter um IDHM mais alto, isso faz sentido, pois o pré-natal adequado está relacionado a me-

lhores condições de saúde materno-infantil, o que pode contribuir para o desenvolvimento humano nos municípios. As demais figuras estão no Anexo - Figuras Auxiliares.

Sendo apresentado na tabela 6 as 14 variáveis que não apresentaram multicolinearidade e com seus devidos p-valores os quais apresentam o nível de significância para o modelo RLM, nosso (*) é para facilitar a localização das variáveis que apresentaram o p-valor < 0.05 .

Tabela 6 – Variáveis com p-valor do modelo RLM

Variáveis independentes	p-values
PMATPUB_EF	0.000227*
PMATPUB_EM	0.356116
DIST_EF_TOTAL	$3.83e^{-07}$ *
DIST_EM_TOTAL	0.005263*
DOCSUP_EF_TOTAL	0.004112*
DOCSUP_EM_TOTAL	0.035873*
TXNASC7C	0.444173
TXCOBPLP	$3.56e^{-10}$ *
REN_TRPCBF_D	$6.70e^{-09}$ *
PDEFAGUA	$1.72e^{-10}$ *
PDEFESGOTO	0.194887
PANALF15	0.190122
PIND_POS	0.002932*
PBF	0.027088*

Após a execução do modelo de RLM da **formula2**, a equação de regressão é apresentada a seguir:

$$Y = 2e-16 + 0.000227X_1 + 0.356116X_2 + 3.83e-07X_3 + 0.005263X_4 + 0.004112X_5 + 0.035873X_6 + 0.444173X_7 + 3.56e-10X_8 + 6.70e-09X_9 + 1.72e-10X_{10} + 0.194887X_{11} + 0.190122X_{12} + 0.002932X_{13} + 0.027088X_{14}$$

Após termos conseguido atender o pressuposto de não ter multicolinearidade entre as variáveis, o próximo passo é fazer a análise dos resíduos.

Um dos pressupostos necessários é a normalidade dos resíduos que pode ser verificado através do teste de Kolmogorov-Smirnov conforme apresentado na [2.4.1](#). Verifica-se que o p-valor = 0.9698 > 0.05 , desta forma não se rejeita a hipótese nula, conclui-se que, os resíduos apresentam distribuição normal.

Um segundo teste a ser verificado visa o pressuposto que não ha a autocorrelação nos resíduos, e para tal pode-se utilizar o teste de Durbin-Watson [2.4.2](#). Verificou-se que o p-valor = 0.5513 > 0.05 , desta forma não se rejeita a hipótese nula concluindo que os resíduos não são correlacionados.

O terceiro pressuposto na análise da heterocedasticidade espacial, ou seja, não há

heterocedasticidade dos resíduos, e para tal pode-se utilizar o teste de Breusch-Pagan 2.4.3. Verifica-se que o $p\text{-valor}=0.0005091 < 0.05$, desta forma rejeita-se a hipótese nula concluindo que os resíduos apresentam heterocedasticidade. Neste ponto o modelo RGP pode ser utilizado, pois nele não existe a necessidade de que os resíduos sejam homocedásticos.

Assim, vamos aplicar o modelo de Regressão geograficamente ponderada.

4.2 Regressão Geograficamente Ponderada

A RGP (Regressão Geograficamente Ponderada) serve como um método de modelagem espacial que oferece uma abordagem alternativa quando a regressão global (MLR) produz um modelo com erros heterocedásticos.

Ao incorporar as coordenadas espaciais de cada objeto, o RGP permite a criação de modelos localizados, permitindo uma compreensão mais sutil das relações espaciais e abordando questões de heterogeneidade espacial.

Cada ponto marcado neste recorte do Rio Grande do Sul é uma sede administrativa do município segundo o IBGE, podemos observar que em alguns lugares do mapa a uma grande concentração de pontos, o que indica ser algum centro populacional.

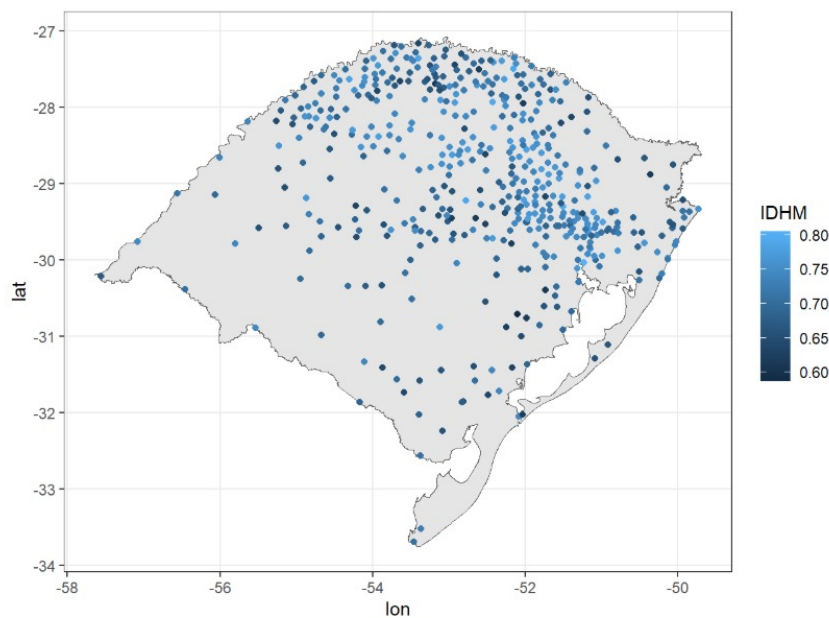


Figura 4 – sedes administrativas do RS

O primeiro passo é fazer a seleção do tamanho de banda; segundo é rodar o modelo de RGP; o terceiro analisar os resíduos e o quarto é verificar o ajuste do modelo.

Serão analisadas 4 situações:

- função Gauss não adaptativo
- função Gauss adaptativo
- função Bi-square (bi-quadrado) não adaptativo
- função Bi-square (bi-quadrado) adaptativo

O pacote **spgwr** possui um conjunto de testes para comparar o modelo de RLM e o modelo RGP sob uma estrutura inferencial. Esses testes são descritos em Fotheringham et al. (2002). A hipótese nula nestes testes é em favor do modelo sem a variável espacial, portanto, uma estatística de teste estatisticamente significativa indica que o RGP proporciona uma melhoria estatisticamente significativa em relação a um ML em termos da sua capacidade de corresponder aos valores observados.

Na tabela 7 são apresentados os testes de comparação dos modelos RLM e modelos RGP e seus p-values e largura de banda. Foi construída com o intuito de vermos qual método do RGP é mais adequado para esse trabalho, se olharmos para os 4 métodos de comparação do MLR com o RGP temos que o bi-square (bi-quadrado) adaptativo seria o que apresenta mais significância espacial, pois de quatro testes ele apresentou significância espacial em 3 deles. Sendo nosso (*) para facilitar a localização das variáveis que apresentaram o p-valor < 0.05 .

Tabela 7 – Testes de comparação dos p-values e largura de banda

Teste	GaussNAdap	GaussAdap	Bi-SquareNAdap	Bi-SquareAdap
LMZ.F1GWR	0.33520	0.43270	0.39310	0.22550
LMZ.F2GWR	0.03002*	0.03219*	0.07819	0.01442*
BFC99.gwr	0.00155*	4.492e ⁻⁰⁵ *	0.01101*	8.515e ⁻⁰⁵ *
BFC02.gwr	0.06429	0.29750	0.13470	0.00532*
Bandagwr.sel	1.28221	0.18771	3.48307	0.74489

Lembrando que os testes que apresentam p-values $< 0,05$ indicam significância espacial. Os primeiros 4 testes comparam o ajuste geral do modelo, enquanto o último teste dessa tabela é a largura de banda.

Como o teste **LMZ.F3GWR.test**, um dos mais atuais, é uma versão atualizada dos testes F1 e F2, e também ele sendo um dos únicos testes que apresenta a inclusão ou não de variáveis no modelo, optou-se por apresentar na tabela 8 os resultados.

Nessa tabela, pode-se ver que a função **Bi-squareAdap** e a função **GaussNAdap** apresentam o mesmo número de variáveis que têm significância espacial, mas que não são necessariamente as mesmas para as duas funções, então, para termos um teste ao qual nos referenciarmos, vamos usar para critério de desempate o teste **BFC02.GWP.test**, que é uma evolução do teste **BFC99.GWP.test**. O teste **BFC02.GWP.test** inclui melhorias para lidar com a heterogeneidade espacial e a dependência entre os resíduos.

Utilizando o teste **LMZ.F3GWR**, pode-se observar na Tabela 8 quais variáveis apresentam significância estatística espacial em cada modelo utilizando as diferentes funções, sendo nosso (*) para facilitar a localização das variáveis que apresentaram o p-valor < 0.05 .

Tabela 8 – Resultados do teste LMZ.F3GWR

LMZ.F3GWR	GaussNAdap	GaussAdap	Bi-SquareNAdap	Bi-SquareAdap
PMATPUB_EF	0.0420039*	0.0001276*	0.132966	0.001247*
PMATPUB_EM	0.4297557	0.5714361	0.538339	0.021412*
DIST_EF_T	0.0202911*	0.0131928*	0.527800	$4.025e^{-07}$ *
DIST_EM_T	0.9993229	0.9999999	0.990779	0.853715
DOCSUP_EF_T	0.9690880	1.0000000	0.974924	0.676327
DOCSUP_EM_T	1.0000000	1.0000000	0.046854*	0.998705
TXNASC7C	0.0041169*	0.1601974	0.084325	$7.103e^{-12}$ *
TXCOBPLP	0.9484079	0.6613277	0.997054	0.825057
REN_TRPCBF_D	0.8853837	0.3401401	0.629036	0.999216
PDEFAGUA	$7.783e^{-12}$ *	$2.2e^{-16}$ *	$4.808e^{-06}$ *	$2.2e^{-16}$ *
PDEFESGOTO	1.0000000	1.0000000	0.999999	1.000000
PANALF15	0.0003659*	0.1019869	0.003264*	0.069328
PIND_POS	0.6730028	0.0002346*	0.288916	0.998418
PBF	0.0001726*	$2.535e^{-09}$ *	$5.379e^{-06}$ *	$2.407e^{-08}$ *

Com essa tabela conseguimos identificar que apenas duas variáveis que foram significativas para todas as funções do RGP que foi o PDEFAGUA e o PBF e também temos as que não foram significativas em nenhum modelo DIST_EM_T, DOCSUP_EF_T, TXCOBPLP, REN_TRPCBF_D e PDEFESGOTO, não apresentaram significância em nenhuma das funções do modelo RGP.

5 Conclusão

Este trabalho apresentou como alternativa aos modelos de Regressão Linear Múltipla (RLM) os modelos Regressão Geograficamente Ponderada (RGP) na análise dos fatores que impactam o Índice de Desenvolvimento Humano Municipal (IDHM) nos municípios do Rio Grande do Sul, quando foi verificada a heterocedasticidade dos resíduos no modelo RLM.

A análise da heterocedasticidade revelou que os resíduos do modelo não apresentavam variância constante, tornando necessária a aplicação da RGP. A regressão geograficamente ponderada demonstrou ser mais adequada para capturar a variabilidade espacial do fenômeno estudado, possibilitando uma análise mais detalhada da relação entre as variáveis em diferentes municípios. O método Bi-Square adaptativo foi identificado como o mais apropriado, pois apresentou significância espacial no teste BFC02.gwp, utilizado como referência por ser o mais atual encontrado na bibliografia dos modelos RGP.

Os resultados sugerem que fatores como Percentual de pessoas inscritas no Cadastro Único sem abastecimento de água adequado e Percentual de nascidos vivos com pelo menos sete consultas de pré-natal apresentam relação significativa com o IDHM.

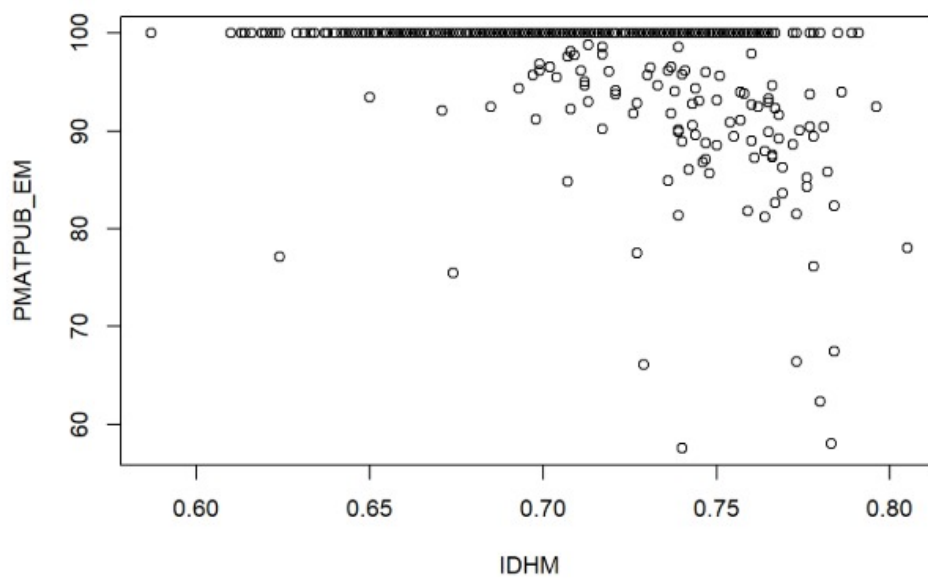
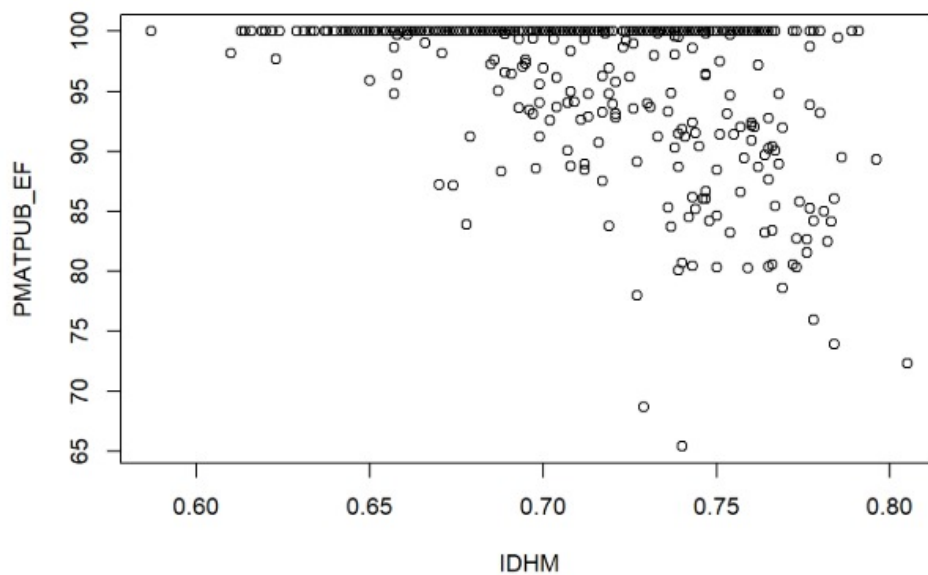
Dessa forma, conclui-se que, para a análise do IDHM nos municípios do Rio Grande do Sul, a Regressão Geograficamente Ponderada se mostrou como uma ferramenta robusta incorporando a heterocedasticidade espacial, uma vez que leva em consideração a influência espacial dos dados.

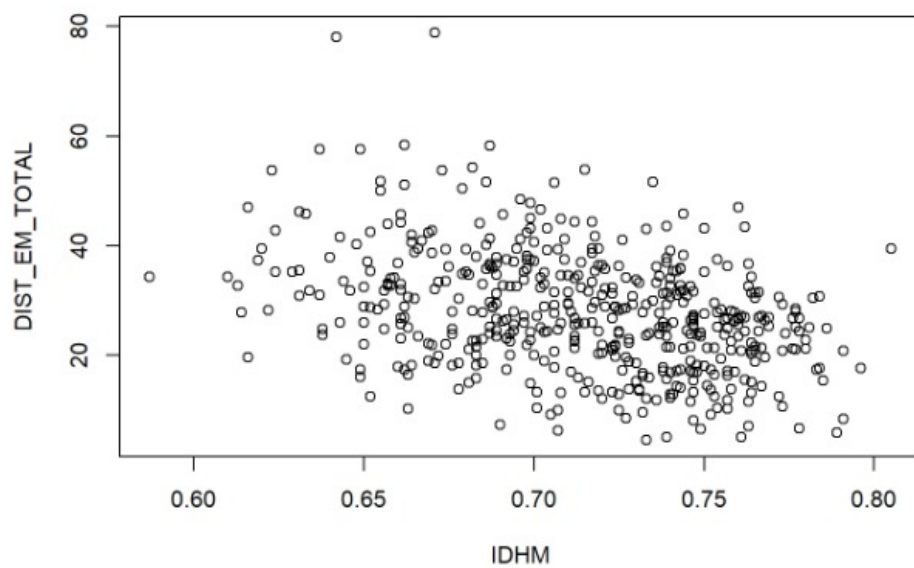
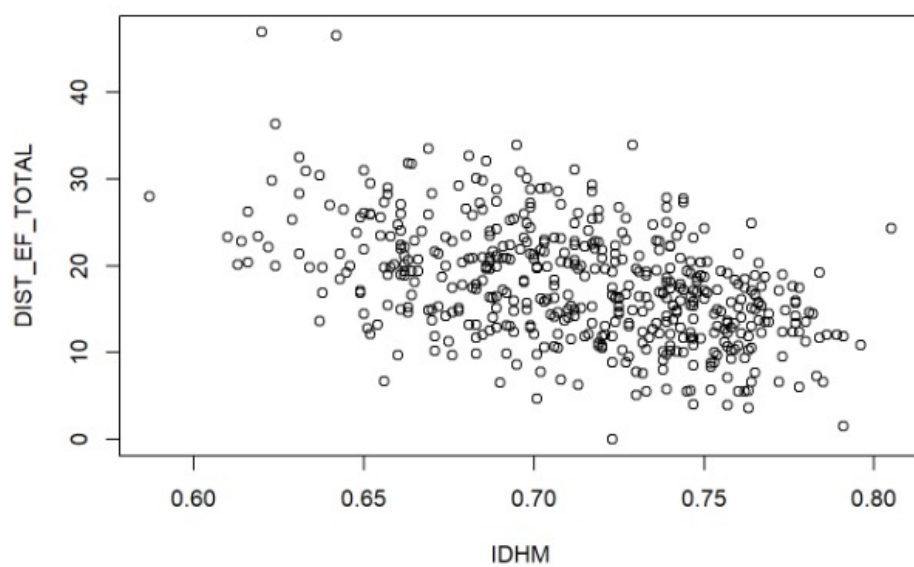
Um destaque final é da importante relevância da Matemática Aplicada e de profissionais qualificados na mesma, para a realização de trabalhos na área de dados socioeconômicos e demográficos. O domínio de técnicas matemáticas consiste na obtenção de métricas confiáveis e que possam ser replicadas com outros conjuntos de dados.

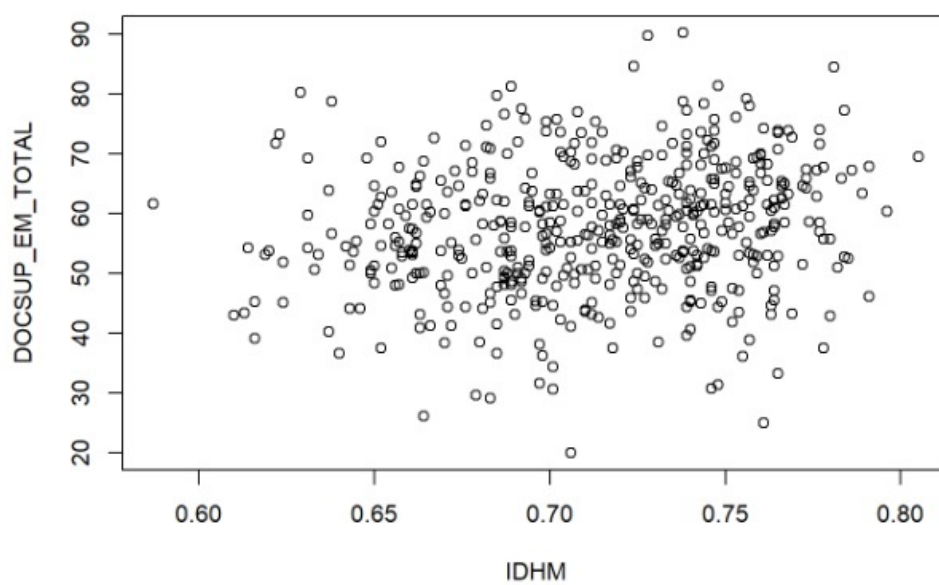
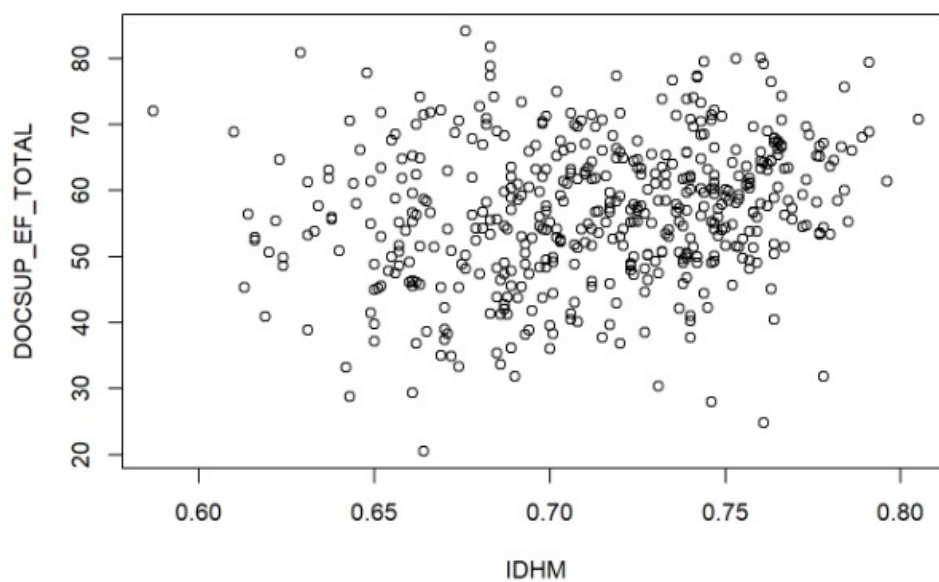
Referências

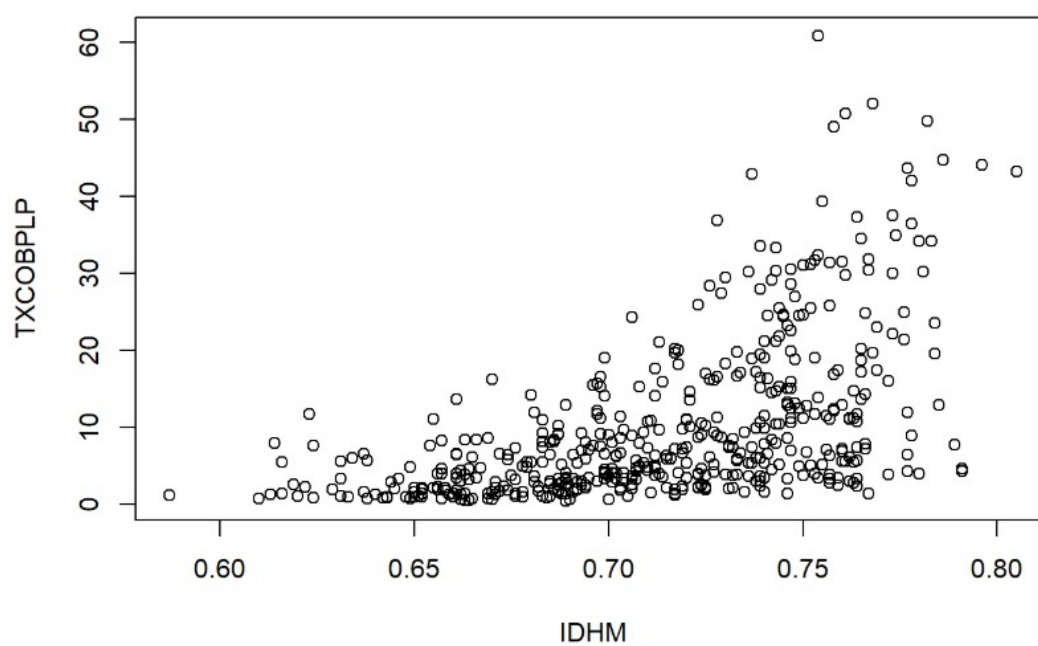
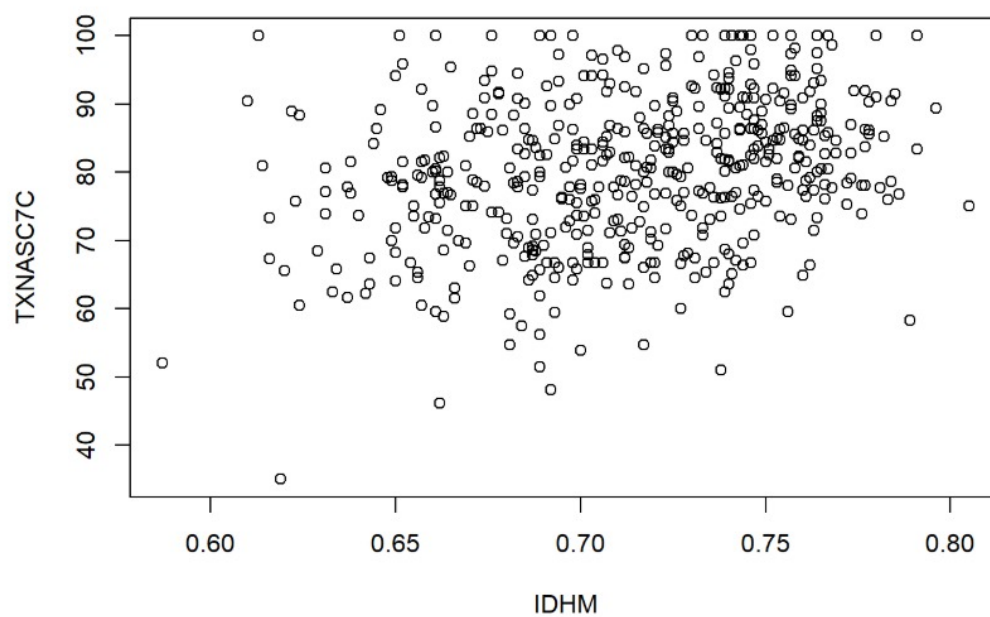
- BRUNSDON, C.; FOTHERINGHAM, S.; CHARLTON, M. Geographically weighted regression (gwr). *ResearchGate*, 2016. Disponível em: https://www.researchgate.net/publication/304623445_Geographically_Weighted_Regression_GWR. Citado na página 17.
- CARVALHO, L. E. X. et al. Regressão linear geograficamente ponderada em ambiente sig. *TRANSPORTES*, v. 14, n. 2, 2006. Citado na página 11.
- CHEIN, F. *Introdução aos modelos de regressão linear*. [S.l.]: Enap, 2019. Citado na página 12.
- FOTHERINGHAM, A. S.; CHARLTON, M.; BRUNSDON, C. Measuring spatial variations in relationships with geographically weighted regression. In: *Recent developments in spatial analysis: Spatial statistics, behavioural modelling, and computational intelligence*. [S.l.]: Springer, 1997. p. 60–82. Citado 2 vezes nas páginas 20 e 21.
- HOFFMANN, R. *ANÁLISE DE REGRESSÃO Uma Introdução à Econometria*. Portal de Livros Abertos da USP, 2016. Disponível em: <https://www.livrosabertos.abcd.usp.br/portaldelivrosUSP/catalog/view/73/64/310>. Citado na página 24.
- PORTER, D. N. G. D. C. *Econometria básica*. [S.l.]: AMGH Editora Ltda., 2011. Citado na página 32.
- QUADE, D. On the asymptotic power of the one-sample kolmogorov-smirnov tests. *The Annals of Mathematical Statistics*, JSTOR, v. 36, n. 3, p. 1000–1018, 1965. Citado na página 23.
- R Documentation. *Breusch-Pagan Test (bptest)*. 2025. Acesso em: 24 jan. 2025. Disponível em: <https://rdrr.io/cran/lmtest/man/bptest.html>. Citado na página 24.
- RANGKUTI, A.; SUNUSI, N.; TAHIR, A. W. Comparison of adaptive bisquare and adaptive gaussian weighting functions in poverty case modeling using geographically weighted regression in papua province. *International Journal of Academic and Applied Research (IJAAR)*, IJEAIS, v. 7, n. 2, p. 21–28, 2023. ISSN 2643-9603. Disponível em: <http://ijeais.org/wp-content/uploads/2023/2/IJAAR230204.pdf>. Citado na página 22.
- ROSA, R.; BRITO, J. L. S. *Introdução ao geoprocessamento. UFU: Apostila. Uberlândia*, 2013. Citado na página 8.
- SILVA, I. D. L. Geobia: Uma análise estatística e espacial entre a presença de piscinas e o nível socioeconômico da população da cidade de marília-sp. *repositorio.usp.br*, 2020. Citado na página 20.
- STEPHENS, M. A. Introduction to kolmogorov (1933) on the empirical determination of a distribution. In: *Breakthroughs in statistics: Methodology and distribution*. [S.l.]: Springer, 1992. p. 93–105. Citado na página 22.

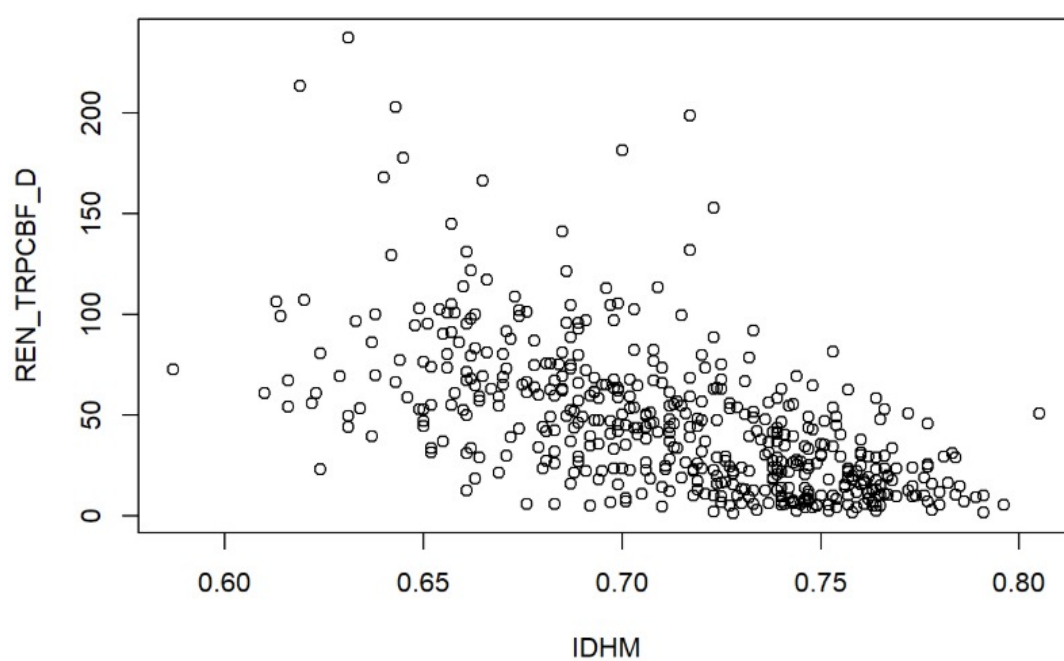
A Anexo - Figuras Auxiliares

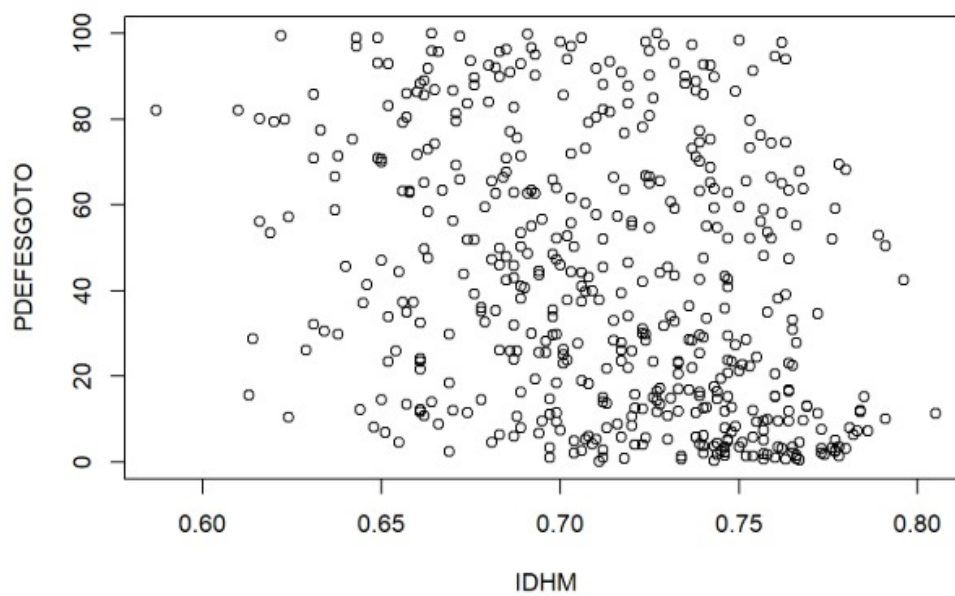
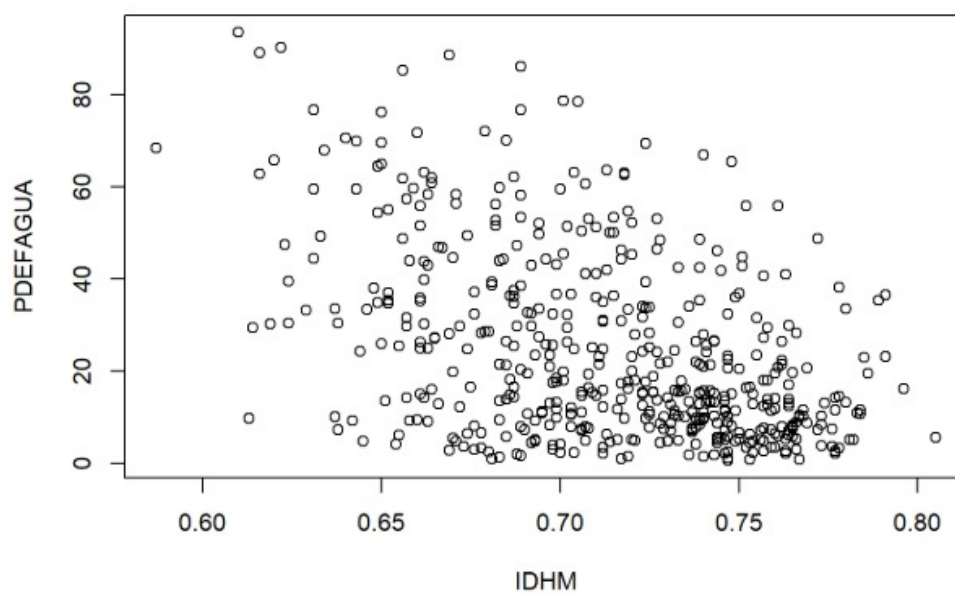


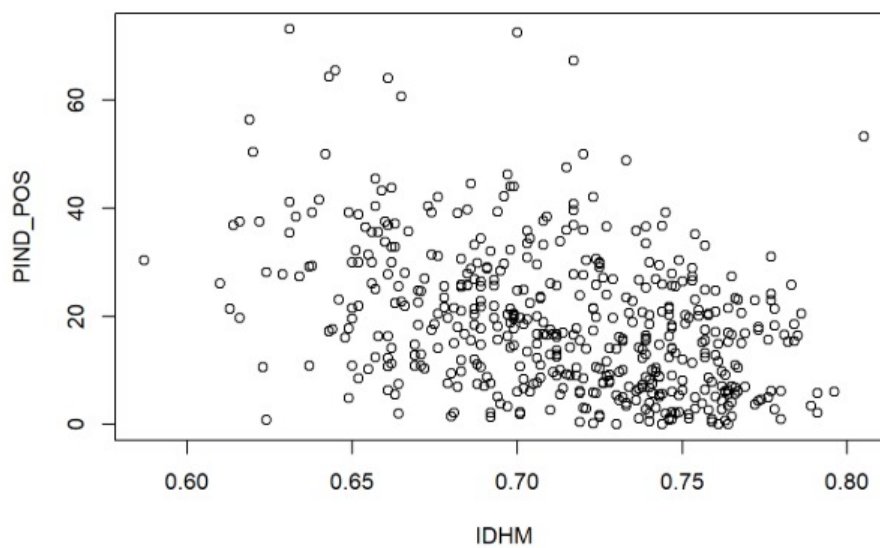
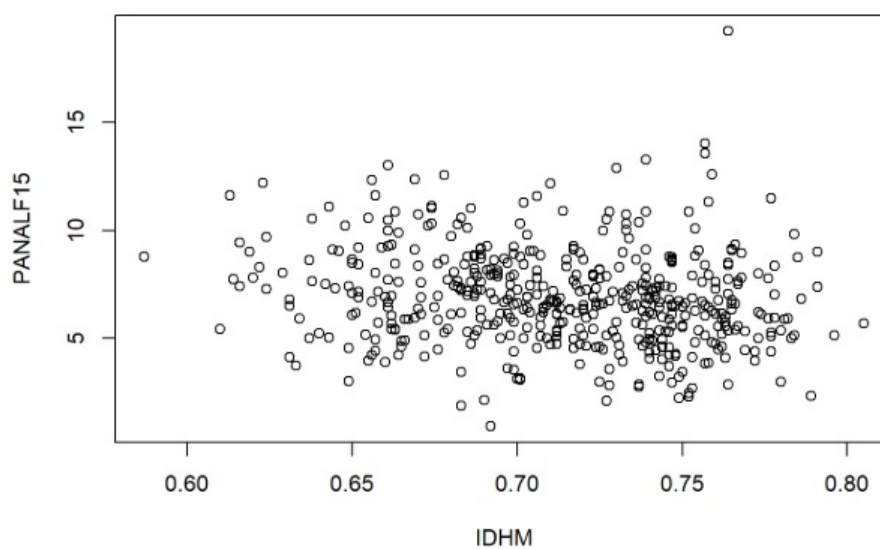


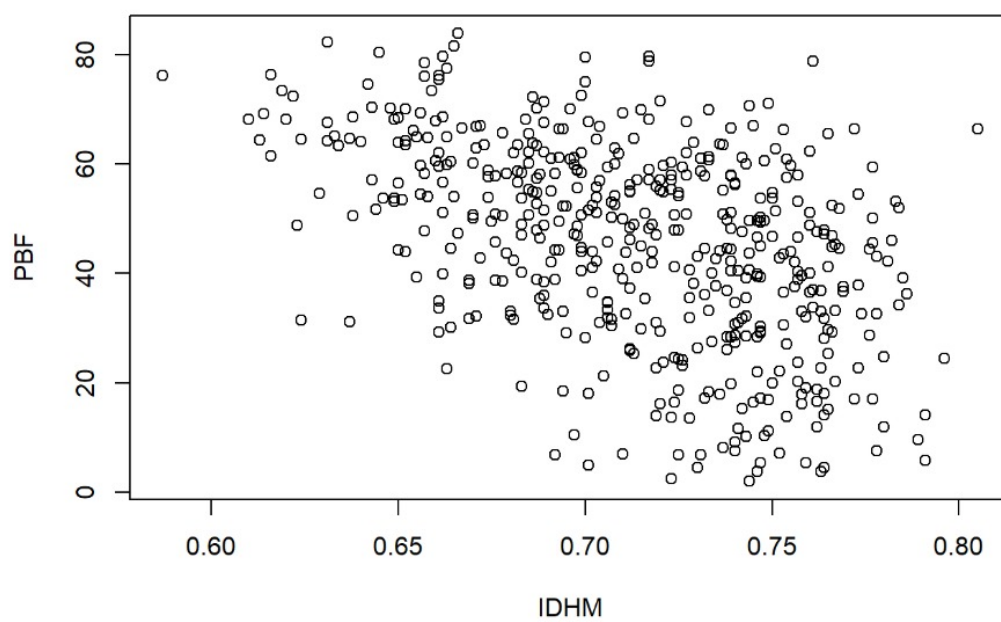












B Anexo - Código em R

```
## Ler os dados
message=FALSE, warning=FALSE, include=FALSE}
dados<-read.csv("dados_colunas_poucoNA.csv",header=TRUE, sep="," ,dec=".")
dados<-dados[, -1]
dados<-na.omit(dados)

for(i in 4:length(dados[1,]))
{ texto<-names(dados[i])
#png(filename = paste0(i,texto,'.png'))
plot(dados$IDHM,dados[,i],ylab=texto,xlab="IDHM")
i<-i+1
#dev.off()
}

## Seleção das variáveis

- Fazer uma tabela com todas as variáveis a serem exploradas e defini-las
- Definir um primeiro modelo _formula0_ incluindo todas as variáveis selecionadas

formula0 = dados$IDHM ~dados$POP_TOT+dados$A2010+dados$PMATPUB_EF+
  dados$PMATPUB_EM+dados$DIST_EF_TOTAL+dados$DIST_EM_TOTAL+
  dados$DIST_EF_PUB+dados$DIST_EM_PUB+dados$PMATLINF_EF+
  dados$PMATLINF_EM+dados$PMATINT_EF+dados$PMATINT_EM+
  dados$DOCSUP_EF_TOTAL+dados$DOCSUP_EM_TOTAL+dados$ DOCSUP_EF_PUB+
  dados$DOCSUP_EM_PUB+dados$TXNASC7C+dados$TXBRUTAMORT+
  dados$TXCOBPLP+dados$PINTERSAP+dados$PINTERDRSAI+dados$REN_RENSF_D+
  dados$REN_TRPCBF_D+dados$REN_TRPCBPC_D+dados$PDEFAGUA+
  dados$PDEFESGOTO+dados$PANALF15+dados$PIND_POS+dados$PMPOB_POS+
  dados$PPOB_POS+dados$PBF+dados$PFLORA+dados$PFEM_VEREADOR+
  dados$IDHMRenda+dados$IDHMEdu +dados$IDHMLongevidade

## Calcular o modelo linear para a _formula0_

message=FALSE, warning=FALSE}
mlr<-lm(formula = formula0)
summary(mlr)

#cálculo do VIF
valores_vif<-vif(mlr)

valores_vif

# Visualização do modelo
plot(mlr, which = 1, main = "Model Fit")

nomes<-c("POP_TOT", "A2010", "PMATPUB_EF",
  "PMATPUB_EM", "DIST_EF_TOTAL", "DIST_EM_TOTAL", "DIST_EF_PUB",
  "DIST_EM_PUB", "PMATLINF_EF", "PMATLINF_EM", "PMATINT_EF",
  "PMATINT_EM", "DOCSUP_EF_TOTAL", "DOCSUP_EM_TOTAL", "DOCSUP_EF_PUB",
  "DOCSUP_EM_PUB", "TXNASC7C", "TXBRUTAMORT", "TXCOBPLP", "PINTERSAP",
  "PINTERDRSAI", "REN_RENSF_D", "REN_TRPCBF_D", "REN_TRPCBPC_D",
  "PDEFAGUA", "PDEFESGOTO", "PANALF15", "PIND_POS", "PMPOB_POS",
  "PPOB_POS", "PBF", "PFLORA", "PFEM_VEREADOR", "IDHMRenda",
  "IDHMEdu", "IDHMLongevidade")
```

```
#Visualização do VIF
barplot(valores_vif, col = "skyblue", main = "Variance Inflation Factor (VIF)"
, names=nomes, las=2, cex.axis=1)

cor_matrix <- cor(dados[c("POP_TOT", "A2010", "PMATPUB_EF",
                          "PMATPUB_EM", "DIST_EF_TOTAL", "DIST_EM_TOTAL", "DIST_EF_PUB",
                          "DIST_EM_PUB", "PMATLINF_EF", "PMATLINF_EM", "PMATINT_EF",
                          "PMATINT_EM", "DOCSUP_EF_TOTAL", "DOCSUP_EM_TOTAL", "DOCSUP_EF_PUB",
                          "DOCSUP_EM_PUB", "TXNASC7C", "TXBRUTAMORT", "TXCOBPLP", "PINTERSAP",
                          "PINTERDRSAI", "REN_RENSF_D", "REN_TRPCBF_D", "REN_TRPCBPC_D",
                          "PDEFAGUA", "PDEFESGOTO", "PANALF15", "PIND_POS", "PMPOB_POS",
                          "PPOB_POS", "PBF", "PFLORA", "PFEM_VEREADOR", "IDHMRenda",
                          "IDHMEdu", "IDHMLongevidade")]))

image(cor_matrix, main = "Correlation Matrix", col = colorRampPalette(c("blue", "white", "red"))(20))

formula1 = dados$IDHM ~ dados$PMATPUB_EF +
  dados$PMATPUB_EM + dados$DIST_EF_TOTAL + dados$DIST_EM_TOTAL +
  dados$DOCSUP_EF_TOTAL + dados$DOCSUP_EM_TOTAL + dados$TXNASC7C +
  dados$TXCOBPLP + dados$REN_TRPCBF_D + dados$PDEFAGUA +
  dados$PDEFESGOTO + dados$PANALF15 + dados$PIND_POS + dados$PMPOB_POS +
  dados$PPOB_POS + dados$PBF

mlr <- lm(formula = formula1)

summary(mlr)
```

```
#cálculo do VIF
valores_vif <- vif(mlr)

valores_vif

# Visualização do modelo
plot(mlr, which = 1, main = "Model Fit")

#Apresenta multicolineariedade retirar algo
formula2 = dados$IDHM ~ dados$PMATPUB_EF +
  dados$PMATPUB_EM + dados$DIST_EF_TOTAL + dados$DIST_EM_TOTAL +
  dados$DOCSUP_EF_TOTAL + dados$DOCSUP_EM_TOTAL + dados$TXNASC7C +
  dados$TXCOBPLP + dados$REN_TRPCBF_D + dados$PDEFAGUA +
  dados$PDEFESGOTO + dados$PANALF15 + dados$PIND_POS + dados$PBF

mlr <- lm(formula = formula2)
summary(mlr)

#cálculo do VIF
valores_vif <- vif(mlr)

valores_vif
```

```
# Visualização do modelo
plot(mlr, which = 1, main = "Model Fit")

residuos<-residuals(mlr)

plot(mlr)

ks.test(residuos,"pnorm",mean(residuos),sd(residuos),alternative=c("two.sided"))

dwtest(lm(mlr$residuals~dados$PMATPUB_EF+
  dados$PMATPUB_EM+dados$DIST_EF_TOTAL+
  dados$DIST_EM_TOTAL+dados$DOCSUP_EF_TOTAL+dados$DOCSUP_EM_TOTAL+dados$TXNASC7C+
  dados$TXCOBPLP+dados$REN_TRPCBF_D+dados$PDEFAGUA+dados$PDEFESGOTO+dados$PANALF15
  +dados$PIND_POS+dados$PBF,data=dados))

bptest(lm(mlr$residuals~ dados$PMATPUB_EF+
  dados$PMATPUB_EM+dados$DIST_EF_TOTAL+dados$DIST_EM_TOTAL+
  dados$DOCSUP_EF_TOTAL+dados$DOCSUP_EM_TOTAL+dados$TXNASC7C+
  dados$TXCOBPLP+dados$REN_TRPCBF_D+dados$PDEFAGUA+
  dados$PDEFESGOTO+dados$PANALF15+dados$PIND_POS+dados$PBF,data=dados))
```

```
## Usando a função *Gauss* *não adaptativo*
```

```
b <- gwr.sel(dados$IDHM ~dados$A2010+ dados$PANALF15 +
  dados$REN_RENSF_D + dados$PBF +
  dados$TXNASC7C + dados$TXBRUTAMORT +
  dados$POP_TOT, coords=cbind(dados$lat, dados$lon),
  data=dados,
  adapt=FALSE,
  gweight=gwr.Gauss)

gwr <- gwr(dados$IDHM ~ dados$PMATPUB_EF+
  dados$PMATPUB_EM+dados$DIST_EF_TOTAL+dados$DIST_EM_TOTAL+
  dados$DOCSUP_EF_TOTAL+dados$DOCSUP_EM_TOTAL+dados$TXNASC7C+
  dados$TXCOBPLP+dados$REN_TRPCBF_D+dados$PDEFAGUA+
  dados$PDEFESGOTO+dados$PANALF15+dados$PIND_POS+dados$PBF,
  coords=cbind(dados$lat,dados$lon),
  bandwidth=b,
  data=dados,
  hatmatrix=TRUE,
  gweight=gwr.Gauss)

names(gwr$SDF)
```

```

t_X1=gwr$SDF$`dados$PMATPUB_EF`/gwr$SDF$`dados$PMATPUB_EF_se`
t_X2=gwr$SDF$`dados$PMATPUB_EM`/gwr$SDF$`dados$PMATPUB_EM_se`
t_X3=gwr$SDF$`dados$DIST_EF_TOTAL`/gwr$SDF$`dados$DIST_EF_TOTAL_se`
t_X4=gwr$SDF$`dados$DIST_EM_TOTAL`/gwr$SDF$`dados$DIST_EM_TOTAL_se`
t_X5=gwr$SDF$`dados$DOCSUP_EF_TOTAL`/gwr$SDF$`dados$DOCSUP_EF_TOTAL_se`
t_X6=gwr$SDF$`dados$DOCSUP_EM_TOTAL`/gwr$SDF$`dados$DOCSUP_EM_TOTAL_se`
t_X7=gwr$SDF$`dados$TXNASC7C`/gwr$SDF$`dados$TXNASC7C_se`
t_X8=gwr$SDF$`dados$TXCOBPLP`/gwr$SDF$`dados$TXCOBPLP_se`
t_X9=gwr$SDF$`dados$REN_TRPCBF_D`/gwr$SDF$`dados$REN_TRPCBF_D_se`
t_X10=gwr$SDF$`dados$PDEFAGUA`/gwr$SDF$`dados$PDEFAGUA_se`
t_X11=gwr$SDF$`dados$PDEFESGOTO`/gwr$SDF$`dados$PDEFESGOTO_se`
t_X12=gwr$SDF$`dados$PANALF15`/gwr$SDF$`dados$PANALF15_se`
t_X13=gwr$SDF$`dados$PIND_POS`/gwr$SDF$`dados$PIND_POS_se`
t_X14=gwr$SDF$`dados$PBF`/gwr$SDF$`dados$PBF_se`

```

```

gwr$SDF$`dados$PMATPUB_EF`=t_X1
gwr$SDF$`dados$PMATPUB_EM`=t_X2
gwr$SDF$`dados$DIST_EF_TOTAL`=t_X3
gwr$SDF$`dados$DIST_EM_TOTAL`=t_X4
gwr$SDF$`dados$DOCSUP_EF_TOTAL`=t_X5
gwr$SDF$`dados$DOCSUP_EM_TOTAL`=t_X6
gwr$SDF$`dados$TXNASC7C`=t_X7
gwr$SDF$`dados$TXCOBPLP`=t_X8
gwr$SDF$`dados$REN_TRPCBF_D`=t_X9
gwr$SDF$`dados$PDEFAGUA`=t_X10
gwr$SDF$`dados$PDEFESGOTO`=t_X11
gwr$SDF$`dados$PANALF15`=t_X12
gwr$SDF$`dados$PIND_POS`=t_X13
gwr$SDF$`dados$PBF`=t_X14

```

```

LMZ.F2GWR.test(gwr)
LMZ.F1GWR.test(gwr)
BFC99.gwr.test(gwr)
BFC02.gwr.test(gwr, approx=FALSE)
LMZ.F3GWR.test(gwr)

```

```

predlocal=gwr$SDF$`pred`
galatlocal=dados$IDHM-predlocal
galatlocalabs<-abs(galatlocal)
ks.test(galatlocal,"pnorm",mean(galatlocal),sd(galatlocal),alternative=c("two.sided"))

```

```

dwtest(lm(galatlocal~ dados$PMATPUB_EF+
  dados$PMATPUB_EM+dados$DIST_EF_TOTAL+dados$DIST_EM_TOTAL+
  dados$DOCSUP_EF_TOTAL+dados$DOCSUP_EM_TOTAL+dados$TXNASC7C+
  dados$TXCOBPLP+dados$REN_TRPCBF_D+dados$PDEFAGUA+
  dados$PDEFESGOTO+dados$PANALF15+dados$PIND_POS+dados$PBF, data=dados))

```

```

bptest(lm(galatlocal~dados$PMATPUB_EF+
  dados$PMATPUB_EM+dados$DIST_EF_TOTAL+dados$DIST_EM_TOTAL+
  dados$DOCSUP_EF_TOTAL+dados$DOCSUP_EM_TOTAL+dados$TXNASC7C+
  dados$TXCOBPLP+dados$REN_TRPCBF_D+dados$PDEFAGUA+
  dados$PDEFESGOTO+dados$PANALF15+dados$PIND_POS+dados$PBF, data = dados))

```

```

library(sf)
library(raster)
library(ggplot2)
shapeRsmun<-st_read("estado_rs_4326_20081217_1506.shp")
m<-shapedados_Final<-shapeRsmun$geometry
plot(m)

```

```

mapa<-ggplot(m) + geom_sf() + geom_point(aes(x = lon, y = lat, color = IDHM), data = dados)+ theme_bw()
mapa

```

Figura 5 – Pacotes utilizados

```
library(spgwr)  
library(readxl)  
library(car)  
library(lmtest)  
library(zoo)  
library(geobr)  
library(dplyr)
```



Universidade Federal do Rio Grande – FURG

Instituto de Matemática, Estatística e Física

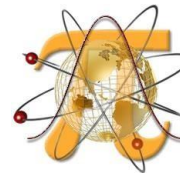
Curso de Bacharelado em Matemática Aplicada

Av. Itália km 8 Bairro Carreiros

Rio Grande-RS CEP: 96.203-900 Fone (53)3293.5411

e-mail: imef@furg.br

Sítio: www.imef.furg.br



Ata de Defesa de Monografia

No sexto dia do mês de fevereiro de 2025, às 14h, no Auditório do IMEF, foi realizada a defesa do Trabalho de Conclusão de Curso do acadêmico **Dener Rafael Quevedo de Matos**, intitulada **“Regressão Geograficamente Ponderada: análise do IDH nos municípios do RS”**, sob a orientação da Profa. Dra. Raquel da Fontoura Nicolette, deste instituto. A banca avaliadora foi composta pela Profa. Dra. Débora Spenassato (IMEF/FURG) e pelo Prof. Dr. Washington Luiz Felix Correia Filho. O candidato foi aprovado por unanimidade. Na forma regulamentar, foi lavrada a presente ata, que é abaixo assinada pelos membros da banca, na ordem acima relacionada.



Documento assinado digitalmente

RAQUEL DA FONTOURA NICOLETTE

Data: 10/02/2025 09:35:35-0300

Verifique em <https://validar.iti.gov.br>

Profa. Dra. Raquel da Fontoura Nicolette
Orientadora



Documento assinado digitalmente

DEBORA SPENASSATO

Data: 07/02/2025 12:23:24-0300

Verifique em <https://validar.iti.gov.br>

Profa. Dra. Débora Spenassato



Documento assinado digitalmente

WASHINGTON LUIZ FELIX CORREIA FILHO

Data: 07/02/2025 13:41:39-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. Washington Luiz Felix Correia Filho