Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências – Câmpus de Botucatu
Programa de Pós-Graduação em Biometria

Comparação do Desempenho de Modelos de Regressão em Conjuntos de Dados Espacialmente Distribuídos

Carolina Aparecida da Silva

Botucatu
2023

Carolina Aparecida da Silva

Comparação do Desempenho de Modelos de Regressão em Conjuntos de Dados Espacialmente Distribuídos

Dissertação de Mestrado apresentada ao Curso de Programa de Pós-Graduação em Biometria da Universidade Estadual Paulista "Júlio de Mesquita Filho" como parte dos requisitos necessários para a obtenção do título de Mestre em Biometria.

Orientador: Prof. Dr. José Silvio Govone

Botucatu
2023

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: MARIA CAROLINA A. CRUZ E SANTOS-CRB 8/10188

Silva, Carolina Aparecida.
Comparação do desempenho de modelos de regressão em conjuntos de dados espacialmente distribuídos / Carolina Aparecida Silva. - Botucatu, 2023

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: José Silvio Govone
Capes: 10202064

1. Estatística. 2. Modelos Lineares. 3. Análise espacial (Estatística).

Palavras-chave: Estatística espacial; Modelo SAR; Modelo SEM; Regressão linear.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Botucatu

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: Comparação do Desempenho de Modelos de Regressão em Conjuntos de Dados Espacialmente Distribuídos

AUTORA: CAROLINA APARECIDA DA SILVA

ORIENTADOR: JOSÉ SILVIO GOVONE

Aprovada como parte das exigências para obtenção do Título de Mestra em Biometria, pela Comissão Examinadora:

Prof. Dr. JOSÉ SILVIO GOVONE (Participaçao Presencial)
Departamento de Estatistica, Matematica Aplicada e Computacao / Instituto de Geociencias e Ciencias Exatas (IGCE) Campus de Rio Claro UNESP

Profª. Drª. LIDIA RAQUEL DE CARVALHO (Participaçao Presencial)
Departamento de Biodiversidade e Bioestatistica / Instituto de Biociências de Botucatu - UNESP

Profa. Dra. CARMEN MARIA ANDREAZZA (Participaçao Virtual)
IGCE / UNESP/Rio Claro (SP)

Botucatu, 28 de julho de 2023

Amanda Regina Sanches
Assistente Administrativo II
Instituto de Biociências de Botucatu

Agradecimentos

Esta dissertação de mestrado teve o apoio de muitas pessoas, e eu, primeiramente, gostaria de agradecer ao meu orientador, o Professor Dr. José Silvio Govone por toda a paciência, o empenho e a atenção com que me orientou ao longo deste trabalho e também nos anos de graduação.

Agradeço também ao 9° Comando de Policiamento do Interior, CPI-9, da PM do Estado de São Paulo, por ceder os dados que serviram de base para a pesquisa.

Desejo igualmente agradecer ao Programa de Pós-Graduação em Biometria da Universidade Estadual Paulista "Júlio de Mesquita Filho" de Botucatu que em parceria com a Pró-Reitoria de Graduação (PROGRAD) financiaram esta pesquisa.

E, por fim, agradeço também ao meu pai que, mesmo não me acompanhando nesta reta final, sempre me incentivou a estudar e nunca desistir dos meus sonhos. Ele é parte fundamental disso tudo e eu não poderia deixar de homenageá-lo.

O amor é o que o amor faz.
Bell Hooks

Resumo

Muitos métodos estatísticos vêm sendo desenvolvidos para auxiliar no estudo de variáveis distribuídas no espaço, com o avanço da tecnologia, modelos de estrutura complexa conseguem descrever melhor a realidade. O presente trabalho abordou a estatística espacial. Além da regressão linear clássica, dois modelos espaciais, que incorporam a autocorrelação espacial presente nos dados, foram definidos, sendo eles o modelo SAR(*Simultaneous Autoregressive Models*) e o SEM(*Simultaneous Error Models*). Através deles, foi possível analisar os dados de criminalidade da região do Comando de Policiamento do Interior-9 (CPI-9). Com base na literatura recente, diferentes abordagens espaço-temporais são analisadas e constituem as três técnicas propostas. Esses modelos visam mapear uma doença, entender seu comportamento ao longo dos anos, identificar áreas de alto risco e capturar a estrutura espacial e temporal. Cada técnica ajustou melhor cada uma das variáveis respostas, são elas roubos de carros, outros roubos e homicídios.

Palavras-chave: Estatística espacial. Modelos espaciais. Regressão linear. Modelo SAR. Modelo SEM. Homicídios. Roubos de carros. Roubos outros. CPI-9.

Abstract

Many statistical methods have been developed to assist in the study of variables distributed in space, with the advancement of technology, complex structure models can better describe reality. The present work addressed Spatial Statistics. In addition to the classical linear regression, two spatial models, which incorporate the spatial autocorrelation present in the data, were defined, namely the SAR(*Simultaneous Autoregressive Models*) model and the SEM(*Simultaneous Error Models*). Through them, it was possible to analyze the crime data of the region of the Interior Police Command-9 (CPI-9). Based on recent literature, different spatio-temporal approaches are analyzed and constitute the three proposed techniques. These models aim to map a disease, understand its behavior over the years, identify high-risk areas, and capture the spatial and temporal structure. Each technique best fitted each of the response variables, namely car thefts, other thefts and homicides.

Keywords: Spatial statistics. Spatio models. Linear regression. Model SAR. Model SEM. Homicides. Car thefts. Thefts others. CPI-9.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

SAR	*Simultaneous Autoregressive Models*
SEM	*Simultaneous Error Models*
SMA	*Spatial Moving Average*
SAC	*General Spatial Model*
AIC	*Critério de informação de Akaike*
CPI-9	*Comando de Policiamento do Interior-9*

Sumário

1 INTRODUÇÃO

A Estatística Espacial é o conjunto de métodos de análise de dados em que a localização espacial é de suma importância na análise e na interpretação de resultados (DRUCK S.; CARVALHO, 2004). Ela é a área da Estatística que compreende a distribuição espacial de dados, isto é, resultante de fenômenos ocorridos no espaço.

Utiliza-se "análise estatística espacial quando os dados são espacialmente localizados e se considera explicitamente a possível importância de seu arranjo espacial na análise ou interpretação dos resultados" (BAILEY; GATRELL, 1995).

As principais áreas de aplicação da Estatística Espacial são na epidemiologia, que consiste em coletar e analisar dados sobre ocorrências de doenças (DOMINGUES, 2017); o monitoramento da criminalidade, já que, um crime não acontece totalmente ao acaso, é preciso uma vítima em potencial e uma oportunidade e também o monitoramento ambiental de áreas geológicas e agrícolas, sendo que o monitoramento ambiental visa estimar a distribuição espacial relevante para a avaliação de potenciais fontes ambientais de problemas de saúde, como poluentes de rios, de lagos, do ar, da vegetação, entre outros (ANDRADE; MONTEIRO, 2007).

Com isso, surgem as seguintes questões analisadas pela Estatística Espacial, a distribuição geográfica dos casos de uma doença gera um padrão? Descrevendo o movimento de uma epidemia no espaço e no tempo, ela sugere formas de controle de combate? Há associação de algum fator ambiental? No tema criminalidade, qual é o perfil das vítimas? Qual o local de maior incidência dos crimes? Qual o período em que os atos infracionais ocorreram?

Através da aplicação de modelos espaciais, buscou-se analisar a distribuição espacial da violência nos municípios da região de Piracicaba, de modo a encontrar padrões que identifiquem as possíveis regiões de maior incidência da criminalidade.

1.1 Objetivos

O presente trabalho tem como objetivo comparar o desempenho de modelos de regressão a conjuntos de dados espacialmente distribuídos, particularmente dados de violência, selecionando o modelo mais adequado para cada caso estudado.

Para isso, comparou-se o modelo linear clássico, de mínimos quadrados, e os modelos

SEM (*Simultaneous Error Models*) e SAR (*Simultaneous Autoregressive Models*). Esses dois últimos modelos incorporam a dependência espacial existente entre os dados espacialmente distribuídos.

Com a aplicação da técnica, comparou-se os modelos através de métodos de comparação, como o Critério de Akaike e o erro quadrático médio. Deste modo, o modelo de melhor desempenho para cada varíavel foi selecionado.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Modelo linear clássico

A regressão linear é uma ferramenta muito útil para prever valores de uma variável quantitativa. Comparado a outros modelos estatísticos mais modernos, a regressão linear é um modelo mais simples porém, amplamente utilizado (JAMES, 2013).

A análise de regressão é uma ferramenta em que se tem uma variável dependente Y e um conjunto $X_1, X_2, ..., X_p$ de variáveis independentes e o modelo pode ser expresso como (WALLER L. A.; GOTWAY, 2004):

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.1}$$

em que $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$ é o vetor de parâmetros a ser estimado e $X = (X_0, X_1, X_2, ..., X_p)'$ é a matriz de variáveis independentes em que X_0 é um vetor coluna de 1's e ϵ representa o erro aleatório, não observável, sobre o qual recaem as seguintes suposições: $\epsilon \sim Normal(0, \sigma^2)$, isto é, erros são normalmente distribuídos com variância constante e são independentes entre si, de forma que a, $Cor(\epsilon_i, \epsilon_j) = 0$ para qualquer par de observações i, j, com $i \neq j$.

Para a sua adequeada aplicação, uma das suposições a de que os erros sejam independentes entre si, o que, em geral, não ocorre quando trabalhamos com dados espaciais, uma vez que, devido à proximidade espacial, geralmente ocorre uma correlação entre os pontos mais próximos. Por exemplo, em dados agrupados em áreas, os valores observados na área A estão relacionados com os valores observados nas áreas vizinhas (LASAGE; PACE, 2009).

Assim, a aplicação do modelo clássico de regressão linear, no geral, não é adequada quando os dados são espacialmente distribuídos. Para resolver esta situação, modelos de regressão que incorporam a dependência espacial foram propostos na literatura.

Apresenta-se na sequência, alguns desses modelos, os quais podem ser encontrados em Almeida (2012) e Druck & Carvalho (2004).

2.2 Modelo SAR (*Simultaneous Autoregressive Models*)

O modelo SAR é utilizado para modelar o resultado da interação espacial entre uma área e a sua vizinhança. O modelo SAR é dado por:

$$Y = \rho WY + X\beta + \epsilon \tag{2.2}$$

em que:

- Y é o vetor de observações;
- ρ é o coeficiente de autocorrelação espacial;
- W é a matriz de vizinhança;
- X é matriz de regressão;
- β é vetor de covariáveis;
- ϵ é o vetor de erros, em que ϵ_i independentes e identicamente distribuídos com $\epsilon_i \sim N(0, \sigma^2)$;
- WY representa a dependência espacial em Y.

  Desta forma:

$$Y - eWY = X\beta + \epsilon \Rightarrow (I - eW)Y = X\beta + \epsilon \tag{2.3}$$

$$\Rightarrow Y = (I - eW)^{-1}X\beta + (I - eW)^{-1}\epsilon, \tag{2.4}$$

  supondo $(I - eW)$ invertível.

A Figura 1, apresenta um esquema do modelo SAR, representando a interação espacial da variável Y em relação a outras variáveis. Aqui, nota-se que a variável dependente Y interage com a mesma variável em outra área.

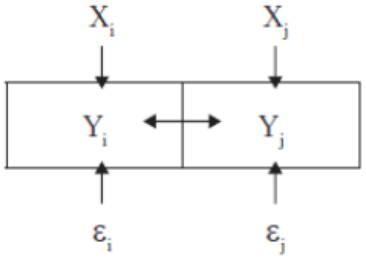


Figura 1 – Representação esquemática do modelo de defasagem espacial. Fonte: Almeida (2010).

Este modelo é conhecido como modelo de efeito espacial global, onde se supõe que seja possível capturar a estrutura da correlação espacial em um único parâmetro, o qual é adicionado ao modelo de regressão tradicional, sendo a autocorrelação espacial ignorada atribuída à variável dependente Y. A autocorrelação espacial é incorporada como uma componente do modelo (DRUCK S.; CARVALHO, 2004).

A hipótese nula para a não existência de correlação é que $e = 0$. Se esta correlação for significativa, então concluímos que parte da variação total da variável Y é explicada pela dependência de cada observação das observações vizinhas.

Se Y não for influenciado por uma vizinhança, e se houver algum agrupamento espacial que interfira em seu valor e na sua vizinhança, no entanto, esta característica não é observável. Neste caso, cabe considerar um modelo alternativo, como o modelo SEM.

2.3 Modelo SEM (*Simultaneous Error Models*) ou CAR (*Conditional Autoregressive*)

Neste modelo, o padrão espacial global considera um processo espacial autoregressivo ao termo do erro. Expressa-se o modelo SEM contendo o erro espacial autoregressivo de primeira ordem:

$$Y = X\beta + U \tag{2.5}$$

$$U = \lambda WU + \epsilon \tag{2.6}$$

em que:

- Y é o vetor de observações dependentes;
- λ é o coeficiente autoregressivo;
- ϵ é o erro aleatório com média zero e variância $\sigma^2 I$;
- X é a matriz das variáveis independentes;
- β é o vetor dos coeficientes de regressão;
- W é a matriz de vizinhança;
- WU é o componente do erro com efeito espacial.

Após algumas manipulações algébricas, a forma reduzida do modelo pode ser apresentada por:

$$Y = X\beta + (I - \lambda W)^{-1}U, \tag{2.7}$$

considera-se $(I - \lambda W)$ matriz inversível.

A Figura 2 apresenta o esquema de interação entre as variáveis das áreas A_i e A_j (VIEIRA, 2009).

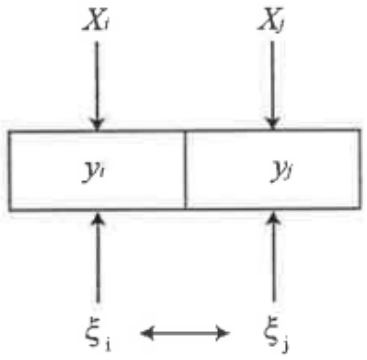


Figura 2 – Representação esquemática do modelo de erro espacial. Fonte: Vieira (2009).

Dada a incapacidade de se modelar toda a fonte de dependência espacial originada do processo estocástico, a parte da dependência não modelada se manifesta no padrão do erro aleatório entre regiões vizinhas, de forma que os erros tenham autocorrelação espacial (DARMOFAL, 2006).

Caso $\lambda = 0$, conclui-se que o termo do erro não é espacialmente correlacionado. Os estimadores dos parâmetros são obtidos através da maximização da função logarítmo da verossimilhança (ALMEIDA, 2012). Há certa dificuldade, na prática, de distiguir este modelo do modelo SAR, pois formalmente são muito próximos, embora sejam diferentes na sua formação.

2.4 Modelo SMA (*Spatial Moving Average*)

Os modelos já apresentados são aplicados onde a dependência espacial ocorre de forma global na área em estudo. Entretanto, há situações nas quais o alcance da dependência espacial é limitado, abrangendo apenas algumas regiões da área. Neste caso, é possível afirmar dizer que o alcance da dependência espacial é localizado, sendo o modelo SMA aplicado nesta situação. Como exemplo, considere uma fábrica que jogue poluentes no ar, prejudicando a produção agrícola na região de localização da fábrica, bem como nas regiões próximas, mas não em todas (ALMEIDA, 2012). Considerando a poluição um efeito não modelado na regressão, pois há dificuldade em sua medição, sua influência é justificada no termo do erro. Desta forma, em princípio seria aplicado o modelo SEM, porém, a situação agora é um pouco diferente, já que o efeito é mais localizado. Para a adequada aplicação do modelo, o efeito não modelado deve ser autocorrelacionado espacialmente, não estar correlacionado com quaisquer das variáveis explicativas e ter um alcance localizado. O modelo SMA consite num processo de média móvel de primeira ordem, sendo dado na mesma notação anteriormente usada:

$$Y = X\beta + U \tag{2.8}$$

$$U = \gamma We + \epsilon, \tag{2.9}$$

em que γ é o coeficiente que indica que a influência de efeitos não modelados tem um impacto localizado sobre a vizinhança, com a restrição $-1 < \gamma < 1$, para que os erros não apresentem um proceso explosivo.

O termo do erro é composto pelo choque na própria região (ϵ) e pelo choque nas regiões vizinhas (We). A suposição de um processo de média móvel espacial é que, em certas situações, pode ser mais realístico tratar a transmissão de choques como um fenômeno local ao invés de global (FLINGLETON, 2008). Após algumas manipulações algébricas, obtem-se o modelo transformado:

$$Y = X\beta + (I + \gamma W)\epsilon \tag{2.10}$$

Um indicativo para determinar até qual distância o impacto localizado produz efeito é a observação dos valores da matriz de variância e covariância do erro U, que leva em consideração a matriz W de ponderação, bem como o coeficiente de ponderação espacial. Estudos em maior profundidade dos modelos abordados, bem como de outros para efeitos espaciais globais ou locais podem ser encontrados em Almeida (ALMEIDA, 2012), Druck & Carvalho (DRUCK S.; CARVALHO, 2004) e Anselin (ANSELIN, 1988).

2.5 Modelo SAC (*General Spatial Model*)

O modelo SAC consiste em uma combinação dos modelos SAR e SEM, no qual o fenômeno em estudo apresenta uma dependência espacial tanto na variável dependente, quanto na forma de erros autocorrelacionados espacialmente. Um exemplo, (ALMEIDA, 2012) seja uma situação onde há um processo de difusão de nova técnica agrícola, na qual aparece um efeito não modelado, como uma praga na lavoura, sendo a intensidade de contágio decrescente, com taxa $(\lambda) < 1$. Neste caso, a praga não está relacionada com a variável explicativa e há interação tanto na variável Y de interesse, como no erro. Considerando as anotações dos modelos anteriores, o modelo é dado por:

$$Y = eW_1y + X\beta + U, \tag{2.11}$$

$$U = \lambda W_2U + \epsilon, \tag{2.12}$$

em que W_1 e W_2 são matrizes com pesos espaciais diferentes. Para evitar instabilidade no modelo, coloca-se as restrições $\mid e \mid < 1$ e $\mid \lambda \mid < 1$.

Após algumas manipulações algébricas, o modelo pode ser escrito como:

$$Y = (I - eW_1)^{-1}X\beta + (I - eW_1)^{-1}(I - \lambda W_2)^{-1}\epsilon \tag{2.13}$$

Discussões mais aprofundadas, bem como outros modelos de regressão espacial, podem ser encontrados em Almeida (2012).

2.6 Correlação espacial

É de grande interesse identificar a estrutura da correlação espacial que melhor descreve os dados. Muitas vezes estuda-se a estrutura da autocorrelação espacial de uma variável entre diferentes áreas em análise. É de grande valia conhecer o quanto o valor da variável numa região é dependente dos valores da mesma variável nas localizações vizinhas. Para dados distribuídos em áreas, as ferramentas utilizadas são o índice de Moran e o índice de Geary e, para contínuos, o variograma. Na sequência, os índices de Moran e Geary nos cálculos destes índices, utiliza-se a chamada matriz de proximidade espacial, W, de ordem $n \times m$, onde n é o número de áreas formada por elementos W, sendo $w_{i,j}$ a representação da proximidade entre as áreas A_i e A_j, para $i \neq j$ e $W_{cc} = 0$, $i = 1, 2, ..., n$; $j = 1, 2, ..., n$.

A matriz de proximidade espacial pode ser classificado de acordo com um critério geográfico ou socioeconômico (ALMEIDA, 2012). A matriz apoia-se na ideia de proximadade, que pode ser definida de acordo com a contiguidade ou a distância.

1. **Contiguidade:** Duas regiões são vizinhas quando compartilham de uma fronteira física em comum. Essas fronteiras podem ser definidas de três maneiras distintas, como pode ser observado na Figura 3.

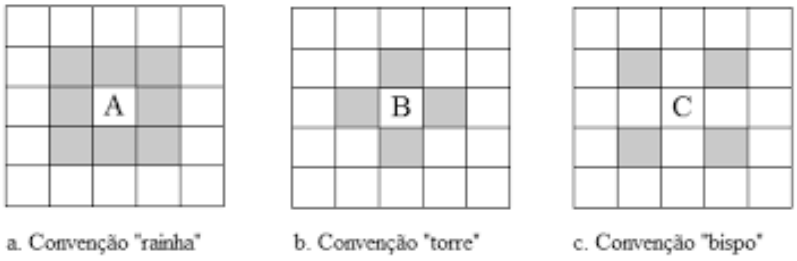


Figura 3 – Estruturas de vizinhança; Matriz de Contingência. Fonte: Almeida (2012)

- **QUEEN/RAINHA:** considera-se como vizinhas as unidades que possuem fronteiras ou vértices comuns, em que a unidade vizinha é definida da forma $w_{ij} = 1$, caso contrário $w_{ij} = 0$.
- **ROOK/TORRE:** considera-se como vizinhas as unidades que possuem fronteiras comuns, em que a unidade vizinha é definida da forma $w_{ij} = 1$, caso contrário $w_{ij} = 0$.

- **BISHOP/BISPO:** considera-se como vizinhas as unidades que possuem vértices comuns, em que a unidade vizinha é definida da forma $w_{ij} = 1$, caso contrário $w_{ij} = 0$.

A título de ilustração, a Tabela 1 mostra a matriz de proximidade espacial binária, cinco por cinco, das macrorregiões brasileiras segundo a convenção rainha ou torre, composta por 25 elementos (ALMEIDA, 2012).

Tabela 1 – Matriz binária de pesos espaciais para as macrorregiões brasileiras (Convenção rainha ou torre. Fonte: Almeida(2012))

	N	**NE**	**CO**	**SE**	**S**
N	0	1	1	0	0
NE	1	0	1	1	0
CO	1	1	0	1	1
SE	0	1	1	0	1
S	0	0	1	1	0

2. **Distância:** neste tipo de matriz de vizinhança considera-se um raio de distância ou número de vizinhos mais próximos, os k primeiros com $k \in \aleph$.

Há diversos critérios para a escolha dos valores w_{ij}, dentre eles:

- $w_{ij} = 1$ se o centróide de A_i está a certa distância especificada do centróide da área A_j; caso contrário, $w_{ij} = 0$;
- $w_{ij} = 1$, se A_i compartilha um lado comum com A_j; caso contrário, $w_{ij} = 0$;
- w_{ij} se o centróide de A_j é um dos k centróides mais próximos de A_i (sendo k um número natural especificado), caso contrário, $w_{ij} = 0$ (BAILEY; GATRELL, 1995).

Um exemplo de uma matriz de dois vizinhos mais próximos para macrorregiões brasileiras é dado pela Tabela 2. A distância de corte para definir os dois vizinhos mais próximos da região Norte foi diferente em relação a distância crítica adotada para se determinar os dois vizinhos mais próximos da região Sudeste e assim por diante (ALMEIDA, 2012).

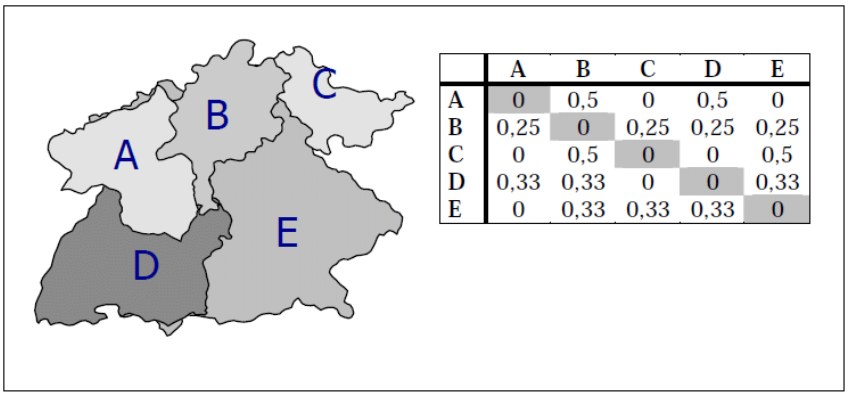
Tabela 2 – Matriz de dois vizinhos mais próximos para as regiões brasileiras. Fonte: Almeida (2012)

	N	**NE**	**CO**	**SE**	**S**
N	0	1	0	0	0
NE	1	0	0	0	0
CO	1	1	0	1	1
SE	0	0	1	0	1
S	0	0	1	1	0

É muito comum normalizar as linhas da matriz W, para que a soma dos pesos de cada linha seja igual a 1. Com isso, os cálculos dos índices de correlação são simplificados e a interpretação torna-se mais simples. Pode ser de interesse montar a Matriz W considerando vizinhos mais distantes (vizinhos dos vizinhos seria ordem 2 denotando a matriz por W^2; vizinhos dos vizinhos dos vizinhos seria ordem 3, denotado por W^3, e assim por diante, até uma ordem l genérica).

Como notação, w_{ij} representa os valores de W^1, enquanto w_{ij} representa os valores de W^l.

Um exemplo a Figura 4, na qual os elementos normalizados da matriz W^1 refletem o critério de adjacência utilizado.



	A	B	C	D	E
A	0	0,5	0	0,5	0
B	0,25	0	0,25	0,25	0,25
C	0	0,5	0	0	0,5
D	0,33	0,33	0	0	0,33
E	0	0,33	0,33	0,33	0

Figura 4 – Matriz de proximidade espacial normalizada pelas linhas. Fonte: Druck & Carvalho (2004)

Neste exemplo, foi utilizado a matriz W^2 o mesmo critério de adjacência é dada por:

$$\begin{pmatrix} A & B & C & D & E \\ 0 & 0 & 0,5 & 0 & 0,5 \\ 0 & 0 & 0 & 0 & 0 \\ 0,5 & 0 & 0 & 0,5 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} = w^2$$

2.7 Índice de Moran

Considerando os vizinhos de ordem 1, em uma localidade com n áreas, Z_i o valor da variável na i-ésima área, $\bar{Z}$ é o valor médio da variável na região de estudo e W_{ij} os elementos da matriz de proximidade espacial de ordem 1, o índice é dado por:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2 (\Sigma\Sigma_{i \neq j} w_{ij})} \tag{2.14}$$

Considerando a proximidade de ordem l, o índice pode ser generalizado:

$$I^l = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^l (y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}(y_i - \overline{y})^2 (\Sigma\Sigma_{i\neq j} w_{ij}^l)}, \tag{2.15}$$

sendo W_{ij}^l os elementos da matriz de proximidade espacial de ordem l.

Temos, $-1 < I < 1$, em que:

- $I \simeq -1$, indica forte correlação espacial negativa;
- $I \simeq 0$, indica ausência de correlação espacial;
- $I \simeq 1$, indica forte correlação espacial positiva.

Calculando o índice de Moran o próximo passo é testar a significância do valor encontrado. Dois testes são propostos na literatura, sendo um deles baseado na normalidade dos valores da variável, obtidos nas diferentes áreas.

Admitindo normalidade, se Z_i e Z_j forem espacialmente independentes para $i \neq j$, I tem distribuição amostral aproximadamente normal com esperança e variância:

$$\epsilon(I) = \frac{-1}{n-1}, \tag{2.16}$$

$$Var(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2}, \tag{2.17}$$

sendo $S_0 = \sum\limits_{i\neq j}\sum\limits_{i\neq j} w_{ij}$, $S_1 = \frac{1}{2}\sum\limits_{i\neq j}\sum\limits_{i\neq j}(w_{ij} + w_{ij})^2$, $S_2 = \sum\limits_{k}(\sum\limits_{j} w_{kj} + \sum\limits_{i} w_{ik})^2$

Observa-se que sendo a matriz W padrozinada pela soma dos pesos de cada linha igual a 1, então W_{ij} não necessariamente é igual a W_{ji}.

Quando a hipótese de normalidade não pode ser aplicada, uma possibilidade para o teste de pseudo-significância de permutação aleatória, no qual é gerado um número muito grande de diferentes permutações dos valores observados e associados às áreas. Para cada permutação, um valor I_p é calculado, e um gráfico de frequências desses valores I_p é plotado. Se o valor da correlação amostral I for de magnitude próxima a dos valores I_p, conclui-se pela não significância do índice de Moran. Caso contrário, considera-se o coeficiente significativo.

Como ilustração, segue a Figura 5, referente ás frequências dos valores I_p obtidos em um teste de permutação, com 999 permutações, sendo o valor amostral obtido, $I = 0,4365$, com $\epsilon(I) = -0,0051$ e $Var(I) = 0,001867$.

Conclui-se que o valor da correlação é significativo, pois ele se encontra distante dos valores obtidos supondo aleatoriedade.

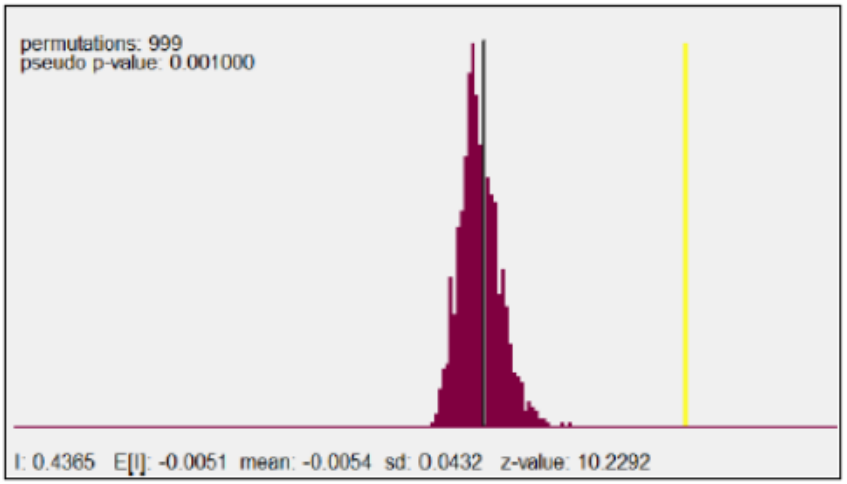


Figura 5 – Teste de permutação aleatória. Fonte: Domingues et al. (2016)

O índice de Moran apresentado é um índice global, pois é calculado um único valor para toda a região em estudo. Muitas vezes, entretanto, têm-se interesse em calcular o índice de Moran local para cada subconjunto de áreas da grande região de estudo.

Isso pode ser realizado quando suspeitamos que um único valor de correlação não representa toda a área, visto que cada subconjunto de área possui um valor de correlação diferente dos demais subconjuntos.

Assim, calculamos um valor específico para cada área, possibilitando a identificação de aglomerados espacias. Esse índice local pode facilmente ser encontrado na literatura, como na obra de Bailey & Gratell (1995).

2.8 Índice de Geary

O índice de Moran é muito empregado no estudo da correlação espacial. Entretanto, ele deve ser evitado quando não ocorrem estacionariedade de primeira ou segunda ordem, isto é, quando os dados apresentam uma tendência e/ou a variância não é constante na região como um todo. Neste caso, deve ser evitada a comparação de cada valor à média global $\overline{z}$. Uma alternativa é o chamado índice C de Geary, que compara os valores entre si, par a par. É dado pela seguinte expressão, utilizando a mesma notação do índice de Moran.

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(Z_i - Z_j)^2}{2\sum_{i=1}^{n}(Z_i - \overline{Z})^2(\sum_{i\neq j}\sum_{i\neq j} w_{ij})} \tag{2.18}$$

De maneira semelhante ao Índice de Moran, testes estatísticos podem ser aplicados para testar a significância do valor encontrado.

2.9 Dados espaciais

Dados espaciais são obtidos de uma variável distribuída espacialmente em uma área ou região e se caracterizam por serem, geralmente, correlacionados entre si. Basicamente, são distribuídos em quatro grandes categorias: dados pontuais, de superfície aleatória, de área e de interação espacial (ASSUNCAO, 2001). Para cada tipo de dado há um método específico de descrição de análise de dados.

2.9.1 Dados de processos pontuais

Os dados de processos pontuais são conjuntos de dados localizados em pontos representados por coordenadas geográficas, sendo o foco a localização de cada um desses pontos, bem como a verificação da forma em que tais pontos estão distribuídos no espaço: aleatória, uniforme ou por aglomerados. Ademais, no caso de ocorrência de aglomerados, de pontos em determinado local, verifica-se as possíveis fontes geradoras de tais aglomerados. Como exemplo, o estudo da distribuição espacial da ocorrência de dengue em uma cidade, num determinado período, através dos endereços dos pacientes. Em regiões de altos índices de ocorrências procura-se detectar uma causa, como a existência de criadouros de pernilongos.

2.9.2 Dados distribuídos em superfícies aleatórias

Em geral, são dados referentes à uma variável aleatória contínua, distribuída em uma área, como, por exemplo, o pH do solo em uma região de estudo. Os valores amostrados de pontos do solo, são equiespaçados quando possível. O estudo é feito pela geoestatística no qual procura-se ajustar uma função de dependência espacial (variograma) e, com base nessa função, obter estimativas da variável em diferentes pontos da região, através da chamada Krigagem.

2.9.3 Dados de área

Muitas vezes, a grande região em estudo é subdividida em áreas, de acordo com algum critério físico, político ou administrativo e o interesse é estudar a distribuição espacial dos dados da variável nestas áreas. Cada valor da variável está associado a uma determinada área, no qual associa-se algum ponto de referência dentro do polígono formado pela área para localizar o valor. Geralmente, o ponto de referência centróide da área. Deseja-se testar se o padrão dos pontos nas diferentes área é aleatório ou se há aglomerados de áreas em relação as ocorrências, bem como suavizar o mapa da distribuição espacial dos dados, calcular correlações entre as áreas e ajustar modelos de regressão. Como exemplo, estudar a distribuição das ocorrências de casos de dengue em certo período de tempo, em bairros de uma cidade, sendo os dados georeferenciados pelo endereço do paciente. Cada bairro é considerado uma área e, por exemplo, caso ocorram 15 casos em um bairro específico, atribuímos a localização de todas essas ocorrências no centróide do bairro.

2.9.4 Dados de interação espacial

Trata-se de uma situação parecida com os dados distribuídos em uma superfície aleatória. Todavia, agora, as posições são consideradas como pares ordenados (i, j), sendo o elemento (i) do par considerado como origem e o outro (j), como destino. O fluxo que trafega entre os pares, de (i) para (j) é considerado aleatório. Dados desta natureza são mais usados na Geografia e na Economia. Como exemplo, consideramos o fluxo de pacientes entre diferentes centros de atendimento, o fluxo migratório de pessoas, etc (ASSUNCAO, 2001).

2.10 Algumas aplicações na literatura

Encontra-se na literatura aplicações em que se busca comparar modelos de estatística espacial, identificando qual seria o melhor modelo para cada aplicação.

Domingues e Govone (2019) compararam os desempenhos dos modelos de regressão múltipla tradicional com os modelos espacias SAR e SEM, no estudo do número de casos de dengue em função de variáveis socioeconômicas, considerando os setores censitários da cidade de Rio Claro, SP. Seis variáveis foram analisadas e foram selecionadas as duas mais correlacionadas com o número de casos de dengue para entrarem nos modelos de regressão: médias de moradores por domicílio e número total de analfabetos responsáveis pelas famílias residentes em cada setor censitário. O melhor ajuste foi obtido pelo modelo SAR, o qual foi utilizado para indicar as regiões da cidade com maior incidência dos casos de dengue em função das duas variáveis (DOMINGUES; GOVONE, 2019).

Araújo (2014) analisou o modelo de regressão espacial autorregressivo misto (SAR) e o modelo do erro espacial (CAR) no intuito de investigar a associação entre a produtividade da soja e as variáveis agrometeorológicas relacionadas à precipitação pluvial, temperatura média e radiação solar global. Nesse estudo, verificou-se a correlação e a autocorrelação espacial entre a produtividade da soja e os elementos agrometeorológicos, por meio da análise espacial de área, usando técnicas como o índice I de Moran global e local e os testes de significância. O estudo pôde por fim, demonstrar que, por meio dos indicadores de desempenho utilizados, os modelos SAR e CAR ofereceram melhores resultados em relação ao modelo de regressão múltipla clássica (ARAUJO; URIBE-OPAZO; JOHANN, 2014).

Silva (2020) analisou a distribuição espacial da incidência da dengue no estado da Paraíba, entre 2007 e 2016, avaliando a existência de dependência geográfica e sua relação com fatores socioeconômicos e ambientais. Utilizou-se o índice de Moran global e local e a estatística C de Geary para avaliar a autocorrelação espacial da dengue e a associação com variáveis socioambientais. Ao analisar a distribuição de casos de dengue nos municípios da Paraíba, foi possível identificar o avanço da doença, que acomete o maior número de cidades a cada ano (SILVA et al., 2020).

Lima et al. (2005) estudaram a distribuição espacial da taxa de homicídios da população masculina, de 15 a 49 anos, residente no Estado de Pernambuco, entre 1995 e 1998. Foram oito variáveis socioeconômicas escolhidas como variáveis independentes, em relação aos municípios. Eles analisaram os modelos de regressão tradicional e os modelos espaciais CAR e Loess, como o modelo de detecção da tendência espacial. É relevante apontar que, o modelo de Loess é uma regressão linear ponderada. A conclusão foi a de que o modelo CAR foi o de melhor ajuste, o que confirmou a associação entre os índices de pobreza, analfabetismo e homicídio (LIMA et al., 2005).

Ceccato e Levine (2021) aplicaram um particular modelo CAR de regressão espacial para analisar a possível sobreposição entre os locais de ocorrência de acidentes de veículos e de crimes em rodovias da Suécia, onde foram consideradas as condições socioeconômicas da população, uso da terra e algumas características da vizinhança dos locais de ocorrência dos eventos. Aqueles de maior ocorrência de cada uma das variáveis (acidentes de veículos ou crimes) foram identificados e marcados, verificando a sobreposição de cada uma das variáveis (LEVINE; CECCATO, 2021).

Amoako (2021) analisou os efeitos dos fatores socioeconômicos em crimes violentos em distritos de Detroit. Neste caso, modelos de regressão envolvendo seis variáveis exploratórias foram aplicadas, sendo selecionadas as seguintes variáveis: heterogeneidade étnica e porcentagem de graduados, preditores significativos de crimes violentos naquela cidade (AMOAKO, 2021).

3 APLICAÇÕES

Nesta seção, desenvolveu-se um estudo espacial para os dados de violência na região do CPI-9, onde obteve-se, junto a CIA da Policia Militar, dados referentes aos casos de violência do ano de 2015, até junho de 2021. O estudo teve como objetivo analisar a distribuição da violência na região da CPI-9, além de analisar a relação da violência com fatores socioeconômicos.

3.1 Materiais e métodos

DEGEO (Diagnóstico Evolutivo Geoponderado), desenvolvido e aplicado inicialmente pela PM na cidade de São Paulo, há pouco mais de 15 anos, e foi aplicado no CPI-9, a partir de 2015 (MARTINS; GOVONE; AFFONSO, 2023). Essa técnica consiste na análise da evolução criminal e da produtividade da CIA PM (existente em todos os municípios) ao longo do tempo, fazendo uma comparação com períodos anteriores. Em relação à produtividade policial, são analisados mensalmente o número de procurados pela justiça, bem como o número de armas apreendidas. Quanto à violência, vários crimes são analisados, como homicídios dolosos e culposos, furtos e roubos de veículos, furtos e roubos outros; por roubos outros entende-se quaisquer tipos de roubos, com exceção de roubos ou furtos de carros. Cada CIA é periodicamente analisada, comparativamente em relação à sua produtividade e a violência no município, e um índice de desigualdade é calculado. Essas avaliações permitem uma atuação mais eficaz da polícia, diminuindo, ao longo do tempo os números de criminalidade (MARTINS; GOVONE; AFFONSO, 2023). De interesse do presente trabalho é utilizar alguns dados referentes aos índices de criminalidade ao longo do tempo e procurar um modelo de regressão mais adequado para relacionar essas varíaveis.

A aplicação das técnicas estudadas é realizada junto a dados de violência dos 52 municípios (Cias PM) que pertencem ao Comando de Policiamento do Interior-9 (CPI-9) da Polícia Militar do Estado de São Paulo, sediado na cidade de Piracicaba (PM, 2020).

A Figura 6 apresenta a região de estudo, no contexto do estado de São Paulo.

Figura 6 – Região dos 52 municípios pertencentes ao CPI-9.

A Figura 7 representa o mapa dos municípios abrangidos pela CPI-9, demarcados por cores diferentes, de acordo com os Batalhões que integram a região: são eles Americana, Limeira, Piracicaba, Rio Claro, São João da Boa Vista e Sumaré.

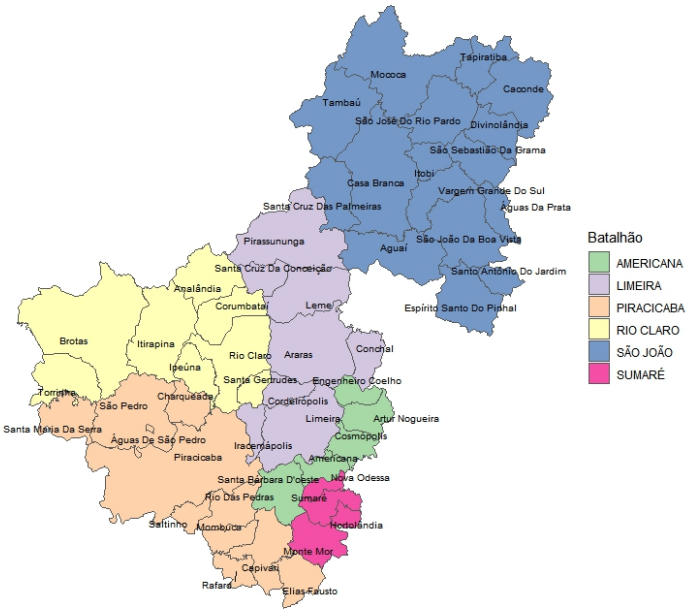


Figura 7 – Mapa com os 52 municípios analisados.

Dados mensais de variáveis de violência, no período 2015-2021, são disponibilizados, em cada um dos municípios, sendo notado que algumas formas de violência têm diminuído nos últimos anos, possivelmente devido á aplicação da técnica DEGEO. As aplicações buscaram, espacialmente, relacionar os dados de violência com fatores socioeconômicos e as variáveis obtidas no estudo foram obtidas pelo Censo Demográfico de 2010, realizado pelo IBGE.

Através dos modelos espaciais definidos na seção 2, sendo eles os modelos de regressão linear, SAR e SEM, buscou-se analisar a relação dos casos de violência com as variáveis

socioeconômicas. Elas foram obtidas pelo Censo Demográfico de 2010, realizado pelo IBGE, todas referentes aos 52 municípios que compõem a CPI-9. A Tabela 3 apresenta as variáveis consideradas no estudo.

Tabela 3 – Variáveis socioeconômicas.

Variáveis socioeconômicas	**Notação**
População	*pop*
IDHM	*IDHM*
Ocupação	*ocup*
Renda per capita	*renda*
Letramento	*let*
Índice de Gini	*gini*

Descrição:

- *pop*: número total de habitantes de cada município;
- *IDHM*: índice de desenvolvimento humano municipal;
- *ocupados*: percentual da população economicamente ativa (PEA) que esteja ocupada na semana de referência. Pessoas ocupadas podem ser empregados, empregadores, conta própria e não remunerados. Define-se como PEA a população entre 15 e 60 anos;
- *renda*: percentual da populacao residente com renda domiciliar mensal per capita abaixo de $R\$140$, a preços de 2010;
- *let*:percentual de indivíduos com mais de 18 anos que não sabem ler ou escrever;
- *gini*: valor do índice de Gini da renda domiciliar per capita das pessoas residentes em determinado espaço geográfico.

Para esta análise, foram considerados os dados de violência em 78 meses, que compreende o período de 2015 a 2021. A Figura 8 apresenta os municípios de acordo com o critério adotado, critério *queen* de montagem da vizinhança.

Figura 8 – Vizinhança de primeira ordem, critério de adjacência queen. Área urbana da CPI-9, dividida em 52 municípios.

Na Tabela 4 e na Figura 9 apresentamos a população de cada município, para o ano de 2022, segundo estimativas do IBGE. A partir dessa análise, notamos grande discrepância em relação aos tamanhos populacionais dos diferentes municípios.

Para padronizar os municípios, dividiu-se para cada variável analisada, as frequências mensais das ocorrências de cada município, pela população total do mesmo e multiplicou-se por 100 mil.

Tabela 4 – População estimada em 2022 dos municípios estudados. Fonte: IBGE (2022)

Batalhão	**Município**	**População**
Americana	Americana	243.674
	Arthur Nogueira	55.352
	Cosmópolis	59.715
	Engenheiro Coelho	20.119
	Santa Bárbara d'Oeste	183.447
Limeira	Araras	131.300
	Conchal	28.184
	Cordeirópolis	26.585
	Iracemápolis	21.768
	Leme	97.516
	Limeira	305.169
	Pirassununga	73.436
	Sta. Cruz da Conceição	4.179
Piracicaba	Águas de São Pedro	37.525
	Capivari	22.679
	Charqueada	15.739
	Elias Fausto	17.832
	Mombuca	3.724
	Piracicaba	434.432
	Rafard	9.333
	Rio das Pedras	31.503
	Saltinho	141.988
	Sta. Maria da Serra	5.753
	São Pedro	38.991
Rio Claro	Analândia	4.577
	Brotas	23.751
	Corumbataí	4.667
	Ipeúna	7.538
	Itirapina	16.157
	Rio Claro	206.950
	Sta. Gertrudes	23.721
	Torrinha	9.303

Batalhão	Município	População
São João da Boa Vista	Aguaí	36.648
	Águas da Prata	8.221
	Caconde	19.009
	Casa Branca	28.189
	Divinolândia	11.086
	Espírito Santo do Pinhal	44.471
	Itobi	7.852
	Mococa	68.980
	Sta. Cruz da Conceição	4.544
	Sto. Antônio do Jardim	5.940
	São João da Boa Vista	91.771
	São José do Rio Pardo	55.124
	São Sebastião da Grama	12.159
	Tapiratiba	12.950
	Tambaú	23.232
	Vargem Grande do Sul	43.110
Sumaré	Hortolândia	234.259
	Monte Mor	60.754
	Nova Odessa	60.956
	Sumaré	286.211

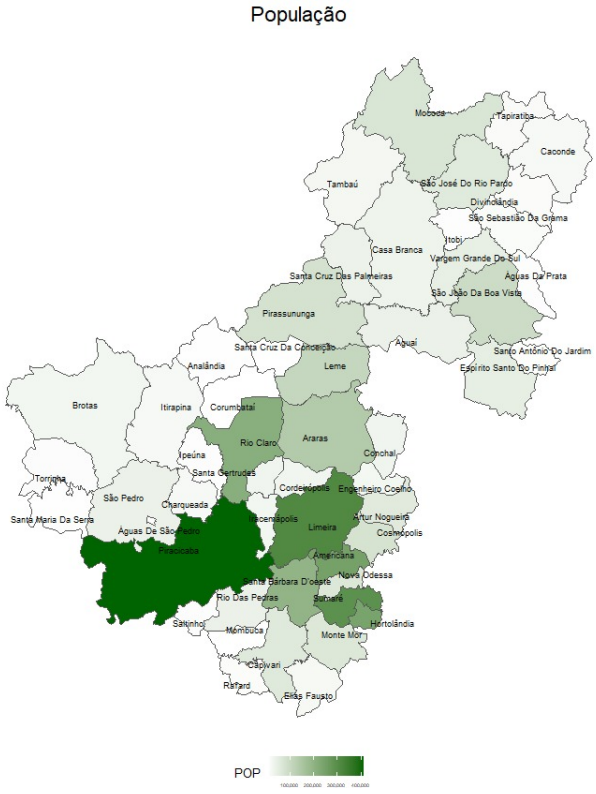


Figura 9 – Mapa da população em 2010 dos 52 Municípios.

4 RESULTADOS E DISCUSSÕES

Com o intuito de identificar a presença ou não de autocorrelação espacial entre os índices de violência e os municípios da CPI-9, calculou-se o índice de Moran e aplicou-se os três modelos apresentados no texto (regressão linear múltipla, modelo SAR e modelo SEM) para cada um dos casos analisados: roubos de carros, outros roubos e homicídios.

- Roubos de carros

Na figura 10, os pontos observados são plotados na diagonal inferior. Na diagonal, é possível observar um histograma de cada variável socioeconômica e do roubo de carros e, na diagonal superior, estão presentes os coeficientes da correlação de Pearson.

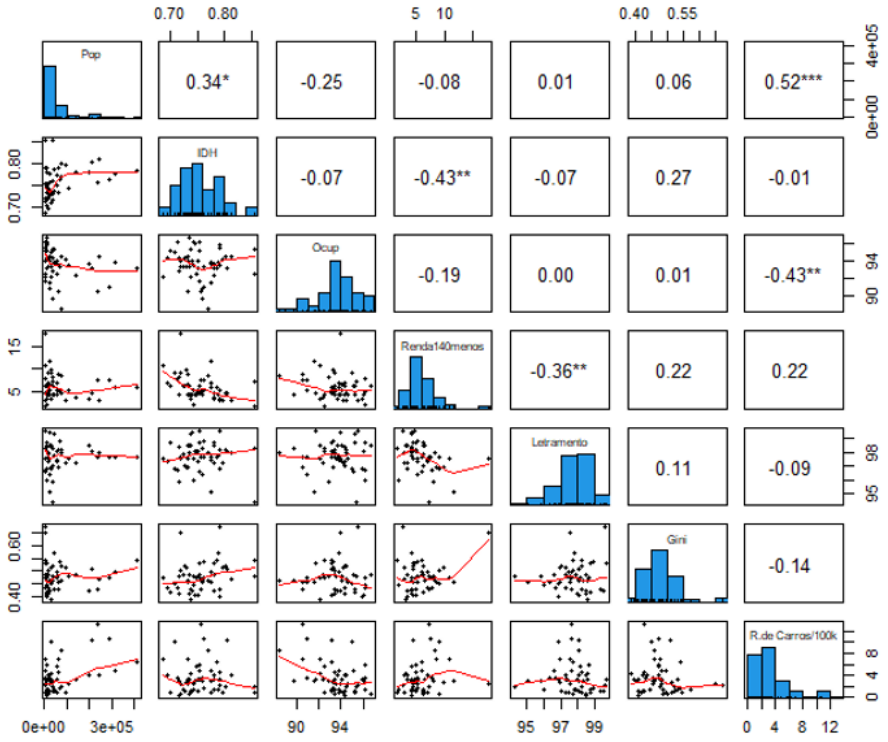


Figura 10 – Matriz de coeficientes da correlação de Pearson para roubos de carros.

A Figura 11 apresenta a distribuição dos casos de roubos de carros na região estudada.

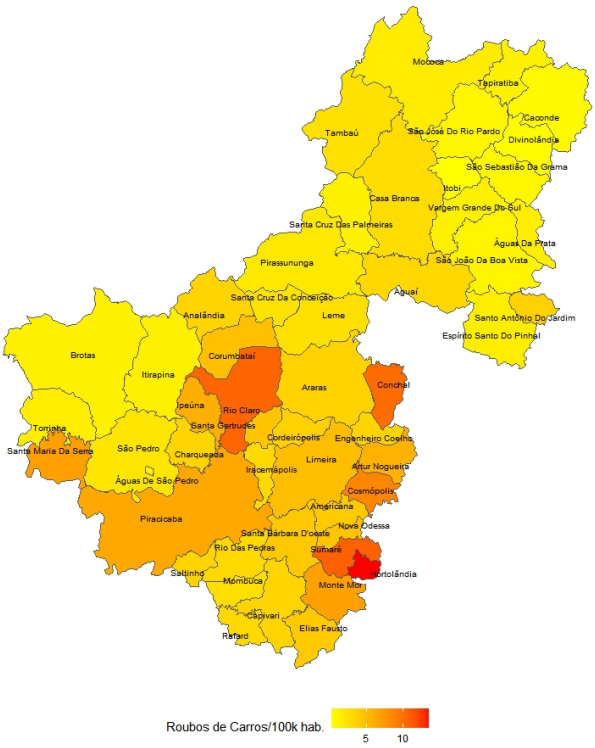


Figura 11 – Mapa de roubos de carros.

Na Tabela 5 apresenta os resultados da regressão para a variável roubos de carros.

Tabela 5 – Regressão linear múltipla para roubos de carros

	Estimativa	Erro padrão	Estatística-t	Pr(>|t|)
Intercepto	44.9	47.9	0.937	0.354
pop	0.0000159	0.00000390	4.08	0.000179
IDHM	-2.88	12.8	-0.224	0.824
ocup	-0.430	0.199	-2.16	0.0365
renda	0.261	0.174	1.49	0.142
let	0.0374	0.374	0.100	0.921
gini	-10.8	6.84	-1.58	0.121
R^2	0.450			
LIK	-113			
AIC	242			

Testou-se a normalidade da distribuição das variáveis escolhidas. O método utilizado nesta avaliação foi o método de Kolmogorov-Smirnov:

$$D_n = sup_x[\mid F(x) - S(x) \mid]$$

Este modelo é utilizado para testar o quão próximo os dados amostrais estão de uma

distribuição de probabilidade de referência, sob hipótese, $S(x)$, ou seja,

$$\begin{cases} H_0: & F(x) = S(x) \\ H_1: & F(x) \neq S(x) \end{cases}$$

Este teste observa a máxima diferença absoluta entre a função de distribuição acumulada S assumida para os dados, no caso a normal, e a função de distribuição empírica F dos dados. Como critério, foi comparada essa diferença com um valor crítico, para um dado nível de significância (RODRIGUES, 2023). Neste caso, o teste rejeitou a hipótese H_1, indicando que os erros seguem uma distribuição normal, com o valor de $p = 0,4996$.

O teste utilizado para heterocedasticidade espacial foi o teste de Breusch–Pagan que, tem como hipótese nula que as variâncias dos erros ao longo da reta de regressão ajustada são homocedásticas, e como hipótese alternativa à heterocedasticidade. No modelo em questão, aceita-se a hipótese nula ao nível de 5%, pois o valor *p* encontrado é de $0,1588$.

Um fator importante na observação é a existência de autocorrelação espacial nos resíduos. O índice de Moran calculado para os resíduos foi $I_{OLS} = 0,23687$ com $p = 0,006$ indicando presença de autocorrelação espacial significativa, ou seja, necessita-se considerar modelos espaciais para o ajuste.

É comum, no processo de elaboração de um modelo, ao coletar os dados, se deparar com a existência de multicolinearidade entre duas ou mais variáveis independentes, segundo Reynaldo: (REYNALDO, 1997) "em muitas análise de modelos de regressão deparamo-nos com o mau condicionamento da matriz de delineamento", ou seja, existe uma forte correlação, não desejável, entre as variáveis independentes de uma modelo de regressão. Ao realizar o teste não detectou-se inflação de variância no modelo.

Na Figura 12, é possível observar, através da análise do primeiro gráfico à esquerda, a linearidade do modelo; o primeiro gráfico à direita, encontra-se o gráfico da normalidade; no segundo gráfico à esquerda, refere-se à homocedasticidade e, por fim, o segundo gráfico à direita, trata-se de um gráfico de outliers do modelo.

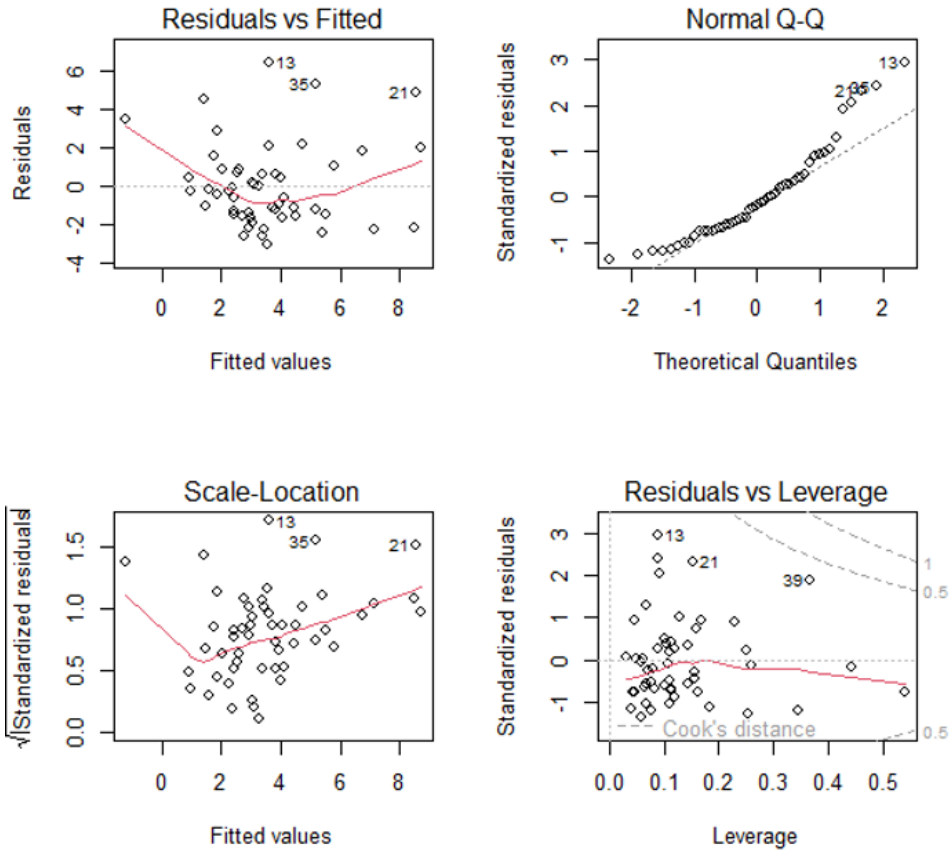


Figura 12 – Gráficos dos resíduos para roubos de carros.

Calculou-se os modelos completos e reduzidos pelo AIC em todos os casos e como eles demonstraram ser indistinguíveis estatisticamente quanto à razão de verossimilhança, optou-se pelos modelos reduzidos em todos os casos e os demais modelos completos encontram-se no apêndice deste trabalho. A Tabela 6 apresenta a regressão linear reduzida pelo AIC.

Tabela 6 – Regressão linear reduzida para roubos de carros

	Estimativa	**Erro padrão**	**Estimativa t**	**Pr(>|t|)**
Intercepto	46.2	18.1	2.55	0.0141
pop	0.0000156	0.00000364	4.29	0.0000880
ocup	-0.426	0.191	-2.23	0.0308
renda	0.273	0.121	2.26	0.0284
gini	-11.3	5.61	-2.02	0.0496
R^2	0.449			
LIK	-113			
AIC	238			

O Índice de Moran não é utilizado apenas para o cálculo de autocorrelação espacial, ele também é capaz de identificiar especificações erradas no modelo, no entanto, não utiliza-se para sugerir qual modelo alternativo deva ser utilizado. Para isso, usa-se as estatísticas dos testes do multiplicador de lagrange (LM) (ANSELIN, 2005) que testam a hipótese de ausência de autocorrelação espacial devido a uma estrutura SAR ou a SEM. O teste busca indicar qual modelo deve ser considerado; para isso segue-se um algoritmo de decisões, são calculadas quatro estatísticas diferentes para o multiplicador de Lagrange LM-SAR e LM-SAR Robusto tomam o modelo SAR como alternativo e LM-SEM e LM-SEM Robusto referem-se ao modelo SEM como o alternativo (DOMINGUES; GOVONE, 2019). Vale ressaltar que os testes LM Robustos devem ser considerados quando as versões padrão (LM-SAR ou LM-SEM) são significativas. Quando não o forem, as propriedades das versões robustas podem não ser válidas. A rejeição da hipótese nula por ambas as estatísticas é uma situação comumente encontrada na prática, podendo então considerar as formas robustas das estatísticas (ANSELIN, 2005).

Conforme a Tabela 7, foram exibidas as quatro estatísticas diferentes para o multiplicador de lagrange e nota-se também que o LM do modelo SAR e LM do modelo SEM foram significativos.

Tabela 7 – Diagnóstico de dependência espacial para roubos de carros.

Diagnóstico para dependência espacial		
Diagnóstico - Testes	**Valor**	***p***
Multiplicador de Lagrange LM (lag)	9.88188	0.0000
LM Robusto (lag)	5.37391	0.0204
Multiplicador de Lagrange LM (erro)	4.97424	0.0000
LM Robusto (erro)	0.46628	0.4947

Anselin et al. (1996)(ANSELIN et al., 1996) propôs uma estratégia de especificação partindo da regressão linear simples e abrangendo testes de ML tanto em sua versão tradicional quanto em sua versão robusta conjuntamente para escolher entre esses dois modelos espaciais SAR ou SEM, com os seguintes passos:

1. Estima-se o modelo clássico de regressão linear;

2. Testa-se a hipótese de ausência de autocorrelação espacial devido a uma defasagem espacial ou a um erro espacial autoregressivo por meio das estatísticas ML_ρ e ML_λ;

3. Caso ambos os testes não sejam significativos do ponto de vista estatístico, estima-se o modelo clássico como o modelo mais adequado. Caso contrário, segue-se para o próximo passo;

4. Se um deles for significativo SAR ou SEM, estima-se o modelo espacial indicado pela hipótese alternativa do teste;

5. Caso ambos sejam significativos, estima-se o modelo apontado como o mais significativo pelas versões roubustas destes testes $ML_\rho Robusto$ e $ML_\lambda Robusto$.

A Figura 13 ilustra o processo descrito acima.

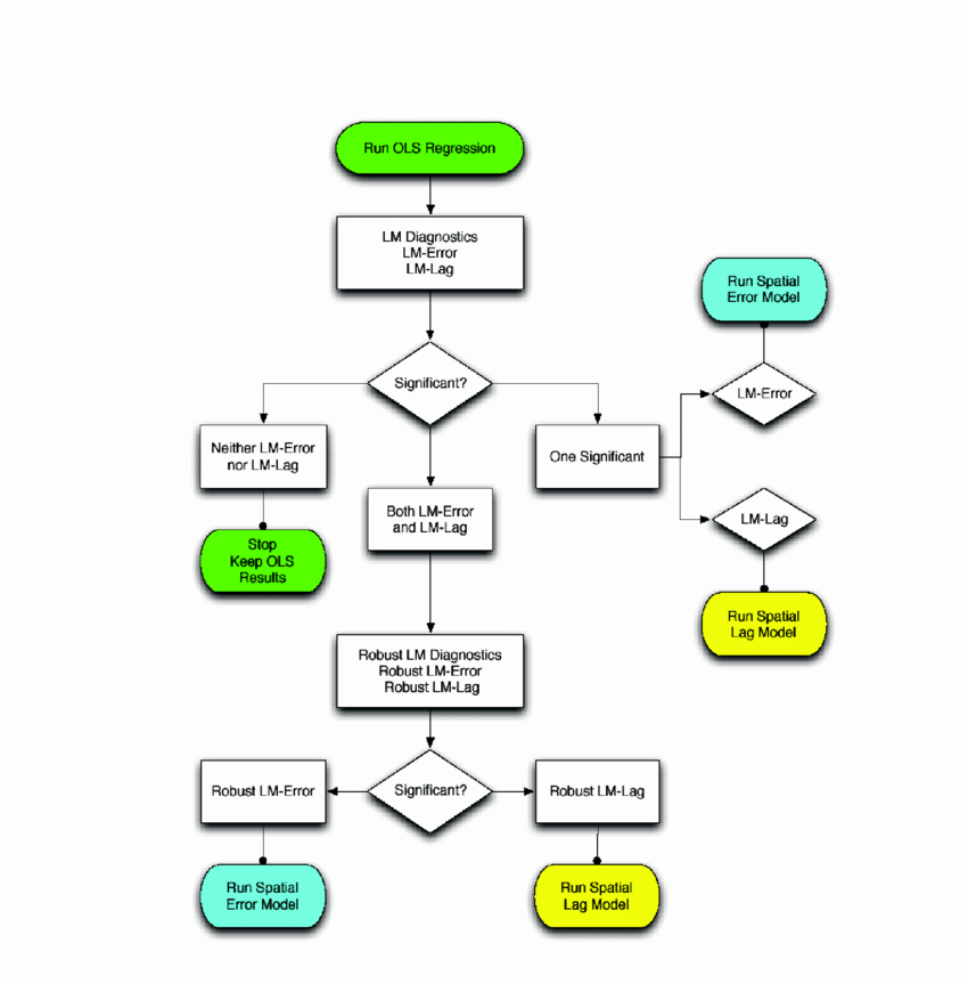


Figura 13 – Processo para regressão espacial. Fonte: Anselin (2005.

Utilizando o Software R (R Core Team, 2013), os valores encontrados para o modelo SAR e do modelo SEM estão apresentados na Tabela 8 e 9 respectivamente.

Tabela 8 – Modelo SAR reduzido para roubos de carros

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	2.9745e+01	1.5420e+01	1.9290	0.0537365
pop	1.0988e-05	3.1708e-06	3.4655	0.0005293
ocup	-2.8619e-01	1.6064e-01	-1.7816	0.748121
renda	2.6889e-01	1.0026e-01	2.6820	0.0073174
gini	-6.9767e+00	4.8741e+00	-1.4394	0.1500464
ρ	0.449			
LIK	-107.6014			
AIC	229.2			

O índice de Moran aplicado aos resíduos do modelo SAR, resultou em $I_{SAR} = 0.029029$ com valor $p = 0.282$, indicando que não há correlação espacial nos resíduos deste modelo.

Tabela 9 – Modelo SEM reduzido para roubos de carros

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	3.2251e+01	1.4357e+01	2.2463	0.0246834
pop	1.1361e-05	3.2838e-06	3.4599	0.0005404
ocup	-2.7479e-01	1.5002e-01	-1.8316	0.0670070
renda	2.9678e-01	9.2791e-02	3.1984	0.0013820
gini	-1.0703e+01	4.7516e+00	-2.2524	0.0242971
λ	0.56156			
LIK	-108.8490			
AIC	231,62			

O índice de Moran encontrado para os resíduos do modelo SEM foi de $I_{SEM} = 0.040944$ com valor de $p = 0.264$. Pode-se concluir então, que a aplicação dos modelos SAR e SEM produziram resíduos independentes entre si, removendo assim, a autocorrelação espacial dos mesmos, o que não acontece com o modelo de regressão clássico. Do ponto de vista estatístico, a escolha do modelo mais adequado é de extrema importância para a análise. O critério de seleção utilizado nesta análise foi o critério de cnformação de Akaike (AIC) e o logaritmo da verossimilhança (LIK). Na tabela 10, observa-se os ajustes e os critérios de comparação.

Tabela 10 – Critérios de comparação para roubos de carros

Modelos	**AIC**	**LIK**
Regressão linear	238	-113
SAR	**229.2**	**-107.6014**
SEM	231.62	-108.8094

Conclui-se da Tabela 10, que os dois modelos de regressão espacial apresentam melhores resultados que o modelo de regressão multivariada clássico. O modelo de regressão mais adequado para a variável roubos de carros ou furto de carros foi o modelo SAR. O mapa dos resíduos para o ajuste do modelo SAR é apresentado na Figura 14.

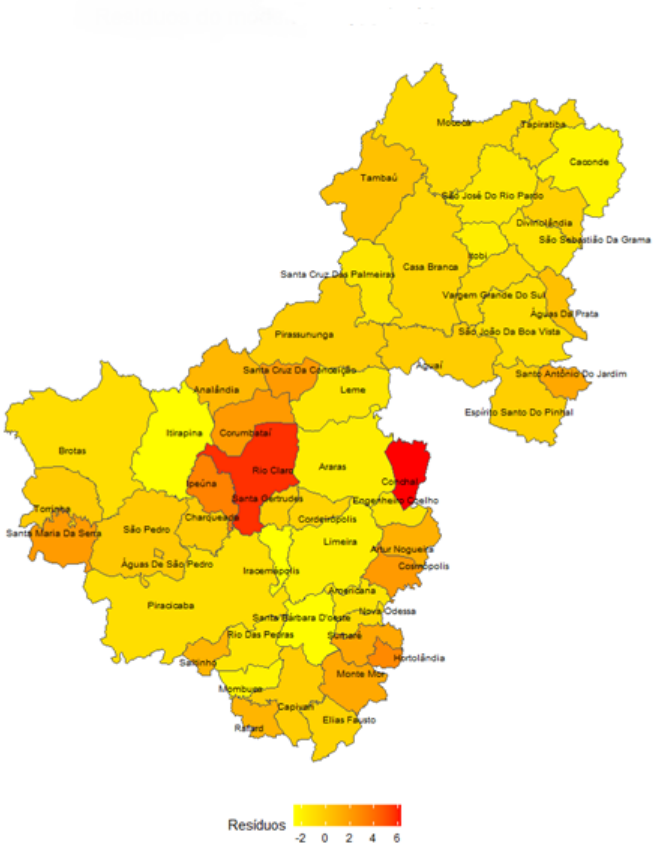


Figura 14 – Mapa dos resíduos do modelo SAR.

Para análise de outros roubos e homicídios, replicou-se o processo utilizado no caso de roubos de carros.

- Outros roubos

  A Figura 15 foi constuída de forma semelhante a Figura 10, agora para outros roubos.

  A Figura 16 apresenta a distribuição dos casos de outros roubos na região do CPI-9.

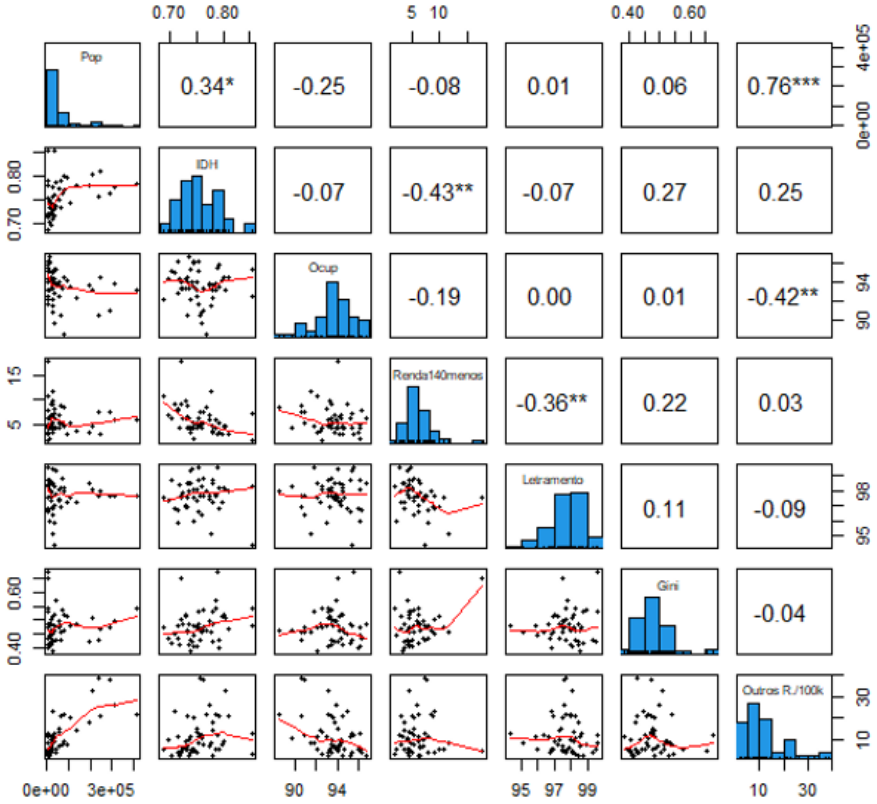


Figura 15 – Matriz de coeficientes da correlação de Pearson para outros roubos.

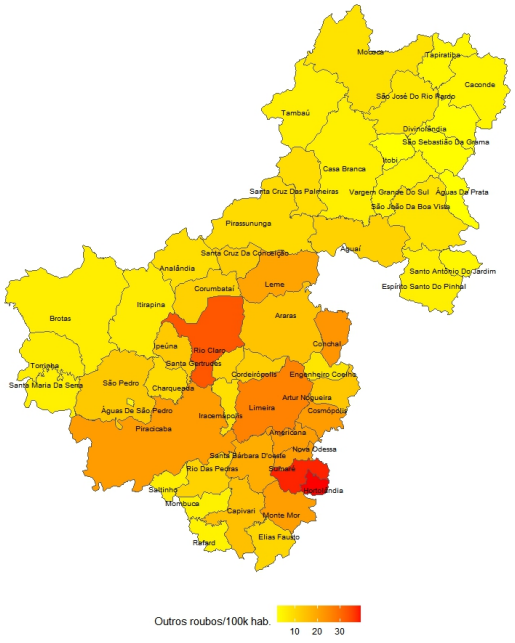


Figura 16 – Mapa de outros roubos.

Aplicou-se a regressão linear para o modelo completo e na Tabela 11, observa-se os resultados encontrados.

Tabela 11 – Regressão linear múltipla completa para outros roubos

	Estimativa	Erro padrão	Estimativa t	Pr(>|t|)
Intercepto	156	1117	1.33	0.190
pop	0.0000676	0.00000952	7.10	0.00000000725
IDHM	12.9	31.4	0.411	0.683
ocup	-1.13	0.487	-2.31	0.0253
renda	0.229	0.426	0.538	0.593
let	-0.475	0.915	-0.520	0.606
gini	-16.3	16.7	-0.978	0.333
R^2	0.643			
LIK	-159			
AIC	335			

Pelo teste de Kolomogorov-Smirnov, observou-se que os resíduos do modelo completo possuem distribuição normal, com o valor $p = 0.1999$ e, tendo aplicado o teste de Breush-Pagan, a heterocedasticidade foi significativa com valor $p = 0.0033124$.

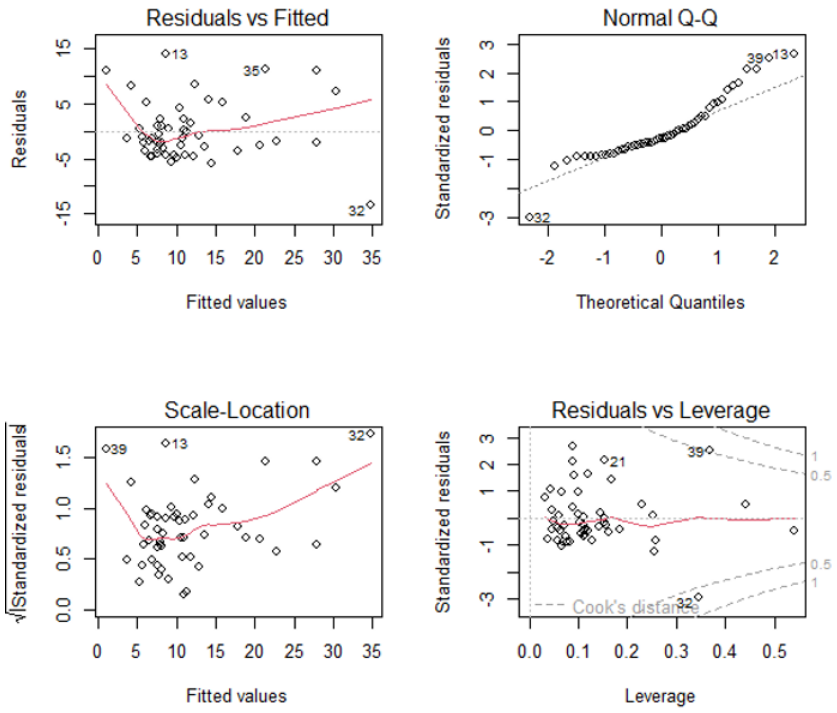
Para correção da heterocedasticidade no modelo utilizou-se uma transformação logarítmica. A Tabela 12 apresenta os resultados encontrados para o modelo da regressão transformada.

Tabela 12 – Regressão linear múltipla para outros roubos com transformação logarítmica

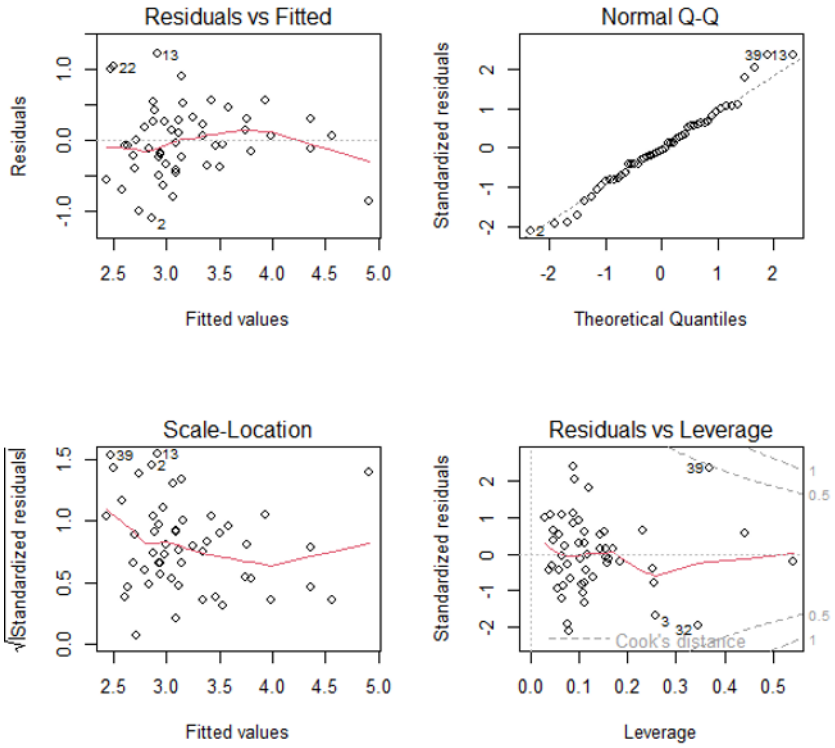
	Estimativa	**Erro padrão**	**Estimativa t**	**Pr(>|t|)**
Intercepto	27.6	11.2	2.48	0.0171
pop	0.00000479	0.000000909	5.27	0.00000373
IDHM	0.349	2.99	0.117	0.908
ocup	-0.117	0.464	-2.53	0.0150
renda	-0.0189	0.0406	-0.466	0.643
let	-0.142	0.0873	-1.63	0.110
gini	-0.0133	1.59	-0.00835	0.993
R^2	0.541			
LIK	-37.3			
AIC	90.7			

Pelo teste de Kolomogorov-Smirnov, observou-se que os resíduos do modelo completo possuem distribuição normal, com o valor $p = 0.9841$ e tendo aplicado novamente o teste de Breush-Pagan, a heterocedasticidade foi reduzida a um nível não significativo com $p = 0.12897$.

A Figura 17 apresenta os gráficos dos resíduos para dados transformados e não transformados.



(a) Regressão linear completa.



(b) Regressão linear com transformação logarítmica.

Figura 17 – Gráficos dos testes realizados para outros roubos

Manteve-se a aplicação reduzida dos modelos, e na Tabela 13, pode-se observar como o modelo da regressão linear reduzida com a transformação logarítmica.

Tabela 13 – Regressão Linear reduzida para outros roubos com transformação logarítmica

	Estimativa	**Erro padrão**	**Estimativa t**	**Pr(>|t|)**
Intercepto	25.5	7.91	3.22	0.00228
pop	0.00000491	0.000000832	5.90	0.000000350
IDHM	-0.111	0.431	-2.57	0.0134
ocup	-0.125	0.0698	-1.80	0.0787
R^2	0.536			
LIK	-37.6			
AIC	85.3			

Com o cálculo do índice de Moran para os resíduos do modelo de regressão, obteve-se um valor significativo com $I_{OLS} = 0.38612$ com $p = 0.001$ indicando autocorrelação entre os resíduos. Na Tabela 14, tem-se as quatro estatísticas do multiplicadores de Lagrange. Como, em ambos os modelos LM foi significativo, seguiu-se com o processo de regressão mantendo a transformação logarítmica.

Tabela 14 – Diagnóstico de dependência espacial para outros roubos.

Diagnóstico para dependência espacial		
Diagnóstico - Testes	**Valor**	***p***
Multiplicador de Lagrange LM (lag)	18.8274	0.0000
LM Robusto (lag)	6.09016	0.0136
Multiplicador de Lagrange LM (erro)	12.9142	0.0000
LM Robusto (erro)	0.17704	0.6739

Pela Tabela 14, nota-se que ambos LM do modelo SAR e LM do modelo SEM foram significativos.

Utilizando o Software R, os valores encontrados para o modelo SAR para outros roubos estão apresentados na Tabela 15.

Tabela 15 – Modelo SAR reduzido para outros roubos

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	1.7636e+01	6.3348e+00	2.7838	0.005373
pop	3.8933e-06	6.7613e-07	5.7583	8.49e-09
ocup	-5.8700e-02	3.4433e-02	-1.7048	0.088238
let	-1.1091e-01	5.4099e-02	-2.0501	0.040352
ρ	0.49778			
LIK	-28.08545			
AIC	68.171			

O índice de Moran aplicado aos resíduos do modelo SAR, resultou em $I_{SAR} = 0.019356$ com valor $p = 0.327$ e no caso do modelo SEM tem-se $I_{SEM} = -0.06364$ e $p = 0.67$. Novamente, os modelos de regressão espacial removeram a autocorrelação dos resíduos.

Na Tabela 16, observa-se resultados do modelo SEM reduzido para outros roubos.

Tabela 16 – Modelo SEM reduzido para outros roubos.

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	2.4658e+01	5.7323e+00	4.3015	1.696e-05
pop	3.7357e-06	6.1429e-07	6.0774	1.221e-09
ocup	-6.2629e-02	2.9910e-02	-2.0939	0.0362705
let	-1.6212e-01	4.6276e-02	-3.5033	0.0004594
λ	0.69904			
LIK	-27.86656			
AIC	67,733			

Na tabela 17, observa-se os ajustes e os critérios de comparação para outros roubos. E observa-se que neste caso, o modelo SEM foi o mais adequado na análise.

Tabela 17 – Critérios de comparação para outros roubos

Modelos	**AIC**	**LIK**
Regressão linear	85.3	-37.6
SAR	68.171	-28.08545
SEM	**67.733**	**-27.86656**

Conclui-se da Tabela 17, que os dois modelos de regressão espacial apresentam melhores resultados que o modelo de regressão multivariada clássico. O modelo de regressão mais adequado para a variável outros roubos foi o modelo SEM. A Figura 18 apresenta o mapa dos resíduos desse modelo.

Figura 18 – Mapa dos resíduos do modelo SEM.

- Homicídios

A Figura 19 foi construída de forma semelhante a Figura 10, agora apresenta-se correlações entre as variáveis no caso de homicídios.

Já a Figura 20 apresenta, como estão distribuídos os casos de homicídios na região estudada.

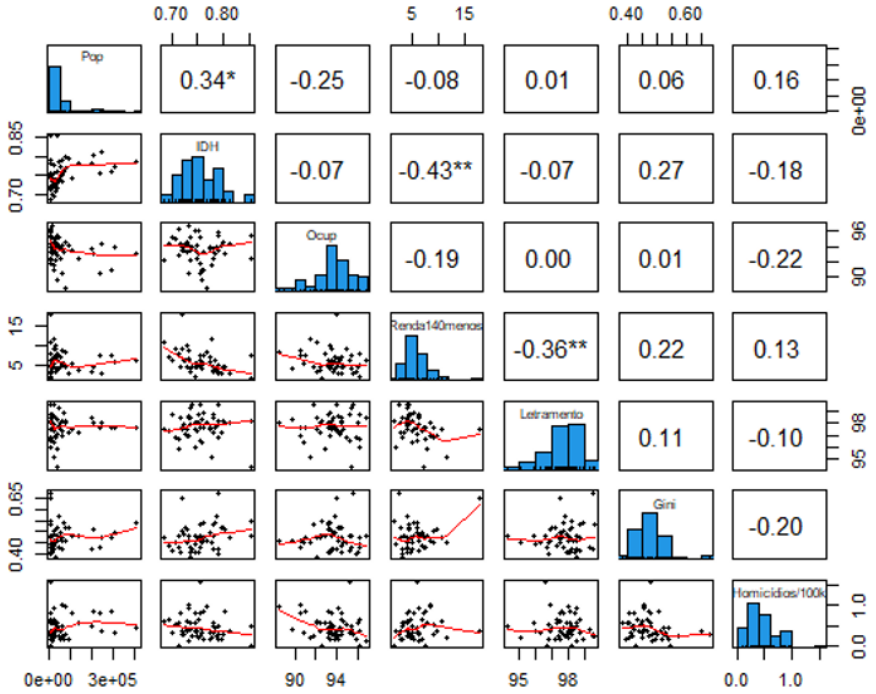


Figura 19 – Matriz de coeficientes da correlação de Pearson para homicídios.

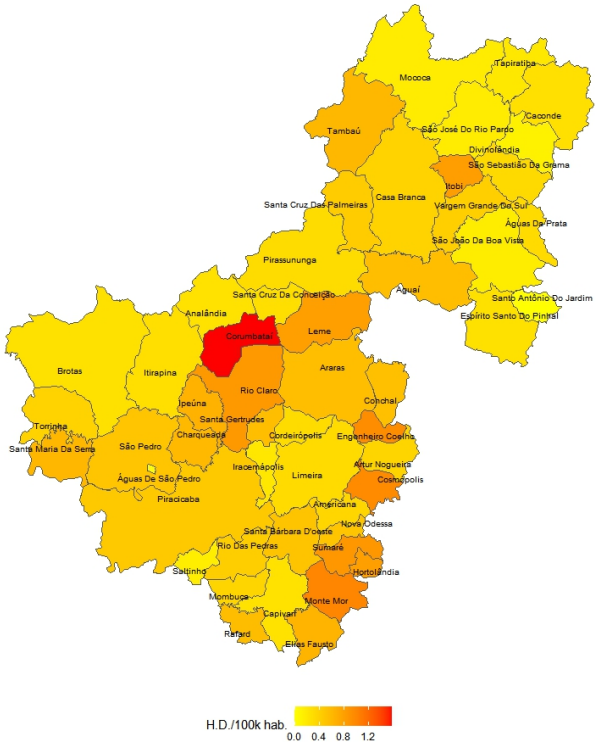


Figura 20 – Mapa de homicídios.

No caso de homicídios, ao repetir o processo, notou-se que o modelo linear completo mostrou-se não significativo com $p = 0.2458$. O índice de Moran aplicado aos resíduos do modelo teve um valor significativo com $I_{OLS} = 0.15176$ e $p = 0.047$. A seleção de variáveis pelo AIC resultou em um modelo reduzido também não significativo com $p = 0.09951$, com índice de Moran $I = 0,098664$ e não significativo com $p = 0.119$. A Tabela 18 mostra os valores encontrados para a regressão completa.

Tabela 18 – Regressão linear completa para homicídios

	Estimativa	**Erro padrão**	**Estimativa t**	**Pr(>|t|)**
Intercepto	7.75	5.98	1.30	0.202
pop	0.000000655	0.000000487	1.35	0.185
IDHM	-1.90	1.60	-1.19	0.242
ocup	-0.0310	0.0249	-1.24	0.220
renda	0.000373	0.0218	0.0171	0.986
let	-0.0272	0.0467	-0.581	0.564
gini	-0.715	0.854	-0.838	0.406
R^2	0.155			
LIK	-4.86			
AIC	25.7			

Calculando-se o índice de Moran para os resíduos do modelo SAR e SEM obteve-se respectivamente $I_{SAR} = 0.0081194$ com $p = 0.372$ e $I_{SEM} = 0.0037639$ com $p = 0.388$, as Tabelas 19 e 20 apresentam os resultados dos modelos.

Tabela 19 – Modelo SAR completo para homicídios

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	7.3385e+00	5.3806e+00	1.3639	0.1726
pop	4.9113e-07	4.4086e-07	1.1140	0.2653
IDHM	-1.7532e+00	1.4425e+00	-1.2154	0.2242
ocup	-2.8786e-02	2.2389e-02	-1.2857	0.1985
renda	3.2863e-03	1.9621e-02	0.1675	0.8670
let	-2.7178e-02	4.2050e-02	-0.6463	0.5181
gini	-7.9698e-01	7.7309e-01	-1.0309	0.3026
ρ	0.27293			
LIK	-3.65161			
AIC	25.303			

Tabela 20 – Modelo SEM completo para homicídios

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	7.4982e+00	5.4325e+00	1.3803	0.1675
pop	4.9269e-07	4.4932e-07	1.0965	0.2728
IDHM	-1.1835e+00	1.4496e+00	-0.8164	0.4143
ocup	-2.9257e-02	2.1717e-02	-1.3472	0.1779
renda	9.3815e-03	1.8899e-02	0.4964	0.6196
let	-3.0365e-02	4.2153e-02	-0.7203	0.4713
gini	-1.0966e-01	7.7977e-01	-1.4062	0.1597
lamba	0.30711			
LIK	-3.562953			
AIC	25.126			

Tabela 21 – Critérios de comparação para outros roubos

Modelos	**AIC**	**LIK**
Regressão linear	25.7	-4.86
SAR	25.303	-3.65161
SEM	25.126	-3.56295

- Homicídio ajustado

Como os resultados não são significativos, pode-se concluir que não há correlação espacial no caso de homicídios. Diante desse resultado elaborou-se um novo modelo para homicídios incluindo os valores de roubos de carros por 100 mil habitantes e de outros roubos por 100 mil habitantes como variáveis explicativas dos homicídios.

Na Tabela 22 pode-se observar a regressão linear completa, considerando a inclusão dos valores de roubos de carros e de outros roubos por 100 mil habitantes. A Figura 21 mostra a correlação entre as novas variáveis.

Tabela 22 – Regressão linear considerando os novos ajustes

	Estimativa	**Erro padrão**	**Estimativa t**	**Pr(>|t|)**
Intercepto	4.02	5.40	0.745	0.461
carros100k	0.0238	0.0278	0.858	0.396
outros100k	0.0171	0.0114	1.50	0.141
pop	-0.000000878	0.000000652	-1.35	0.185
IDHM	-2.05	1.44	-1.43	0.160
ocup	-0.00152	0.0234	-0.0652	0.948
renda	-0.00975	0.0200	-0.487	0.629
let	-0.0200	0.0419	-0.477	0.636
gini	-0.0179	0.779	-0.230	0.819
R^2	0.366			
LIK	2.64			
AIC	14.7			

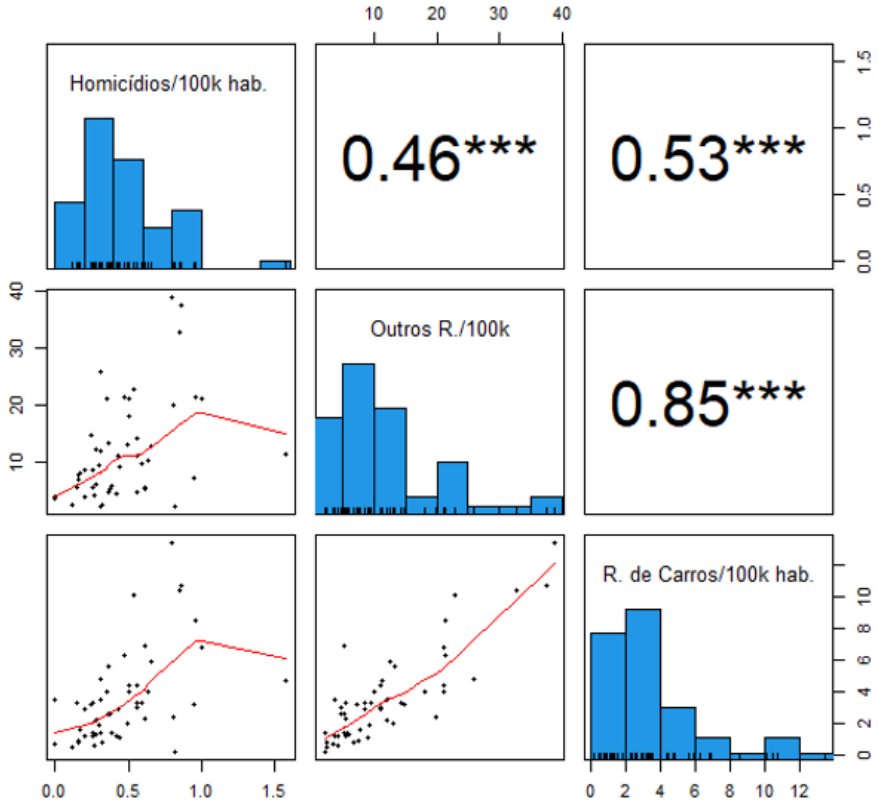


Figura 21 – Matriz de coeficientes da correlação de Pearson para os novos ajustes

Esse novo modelo foi significativo com $p = 0.0736$. Feito a seleção de variáveis pelo AIC, ela também resolveu o problema da colinearidade.

A seleção de variáveis pelo Critério de Informação de Akaike resultou em um modelo reduzido apresentado na Tabela 23.

Tabela 23 – Regressão linear reduzida considerando os novos ajustes

	Estimativa	**Erro padrão**	**Estimativa t**	**Pr(>|t|)**
Intercepto	1.85	0.760	2.44	0.0186
outros100k	0.0257	0.00592	4.34	0.0000722
pop	-0.00000109	0.000000593	-1.83	0.0730
IDHM	-2.14	1.01	-2.11	0.0400
R^2	0.347			
LIK	1.84			
AIC	6.32			

Os resíduos do novo modelo reduzido possuem distribuição normal pelo teste de Kolmogorov-Smirnov com $p = 0.767$ e pelo teste de Breusch-Pagan a heterocedasticidade não foi significativa com $p = 0.23397$. Não foi detectada inflação de variância no modelo. A Figura 22 apresenta os gráficos dos testes realizados. O índice de Moran calculado para este caso, foi de $I_{OLS} = -0.12466$ com $p = 0.872$. A Figura 23 apresenta o mapa dos resíduos dessa nova regressão linear. Uma vez que o índice de Moran foi baixo e sem significância estatística, o modelo linear múltiplo sem dependência espacial se mantém para esses dados.

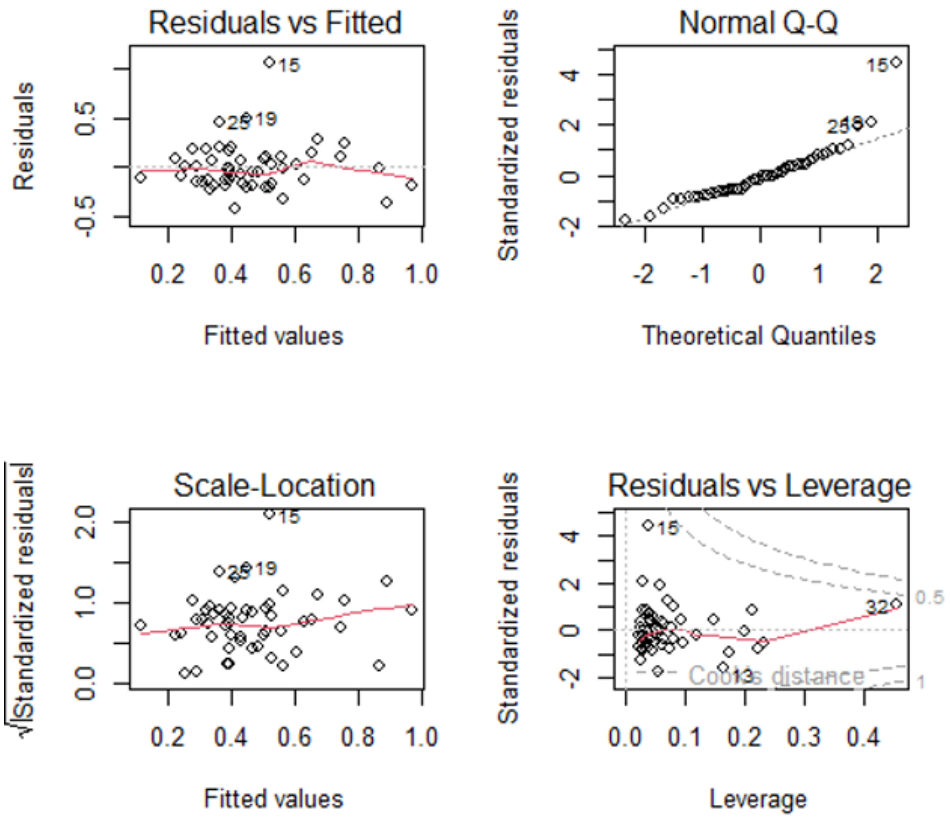


Figura 22 – Gráficos dos testes realizados.

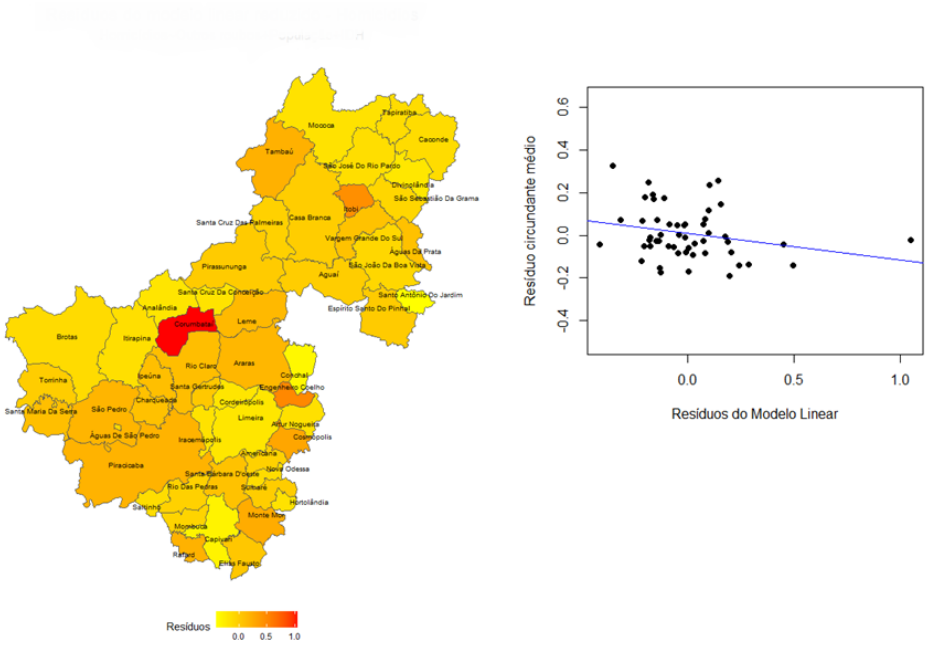


Figura 23 – Distribuição espacial dos resíduos da regressão linear para homicídios

Para verificar se houve alguma similaridade no caso de homicídios, usou-se uma análise de *cluster*.

Agrupamento hierárquico ou *Hierarchical clustering* no inglês é uma técnica de clusterização de dados que baseia-se no tamanho e distância dos dados em um conjunto, ou seja, a técnica agrupa pontos semelhantes de forma que os pontos do mesmo grupo sejam mais semelhantes entre si do que os pontos dos outros grupos. O grupo de pontos de dados semelhantes é chamado de *cluster*.

E para gerar esses clusters, segue-se alguns passos (PATLOLLA, 2018):

- Calcula-se a matriz de proximidade;
- Cada ponto de dados é um cluster;
- Repetir: fundir os dois clusters mais próximos e atualizar a matriz de proximidade, até que reste apenas um único cluster.

Para realizar esse processo são necessários dois recursos: Uma métrica de distância e um critério de ligação (LUCENA, 2019). A técnica de agrupamento hierárquico pode ser visualizada através de um dendrograma.

No caso de homicídios ajustado a análise de clusters, utiliza-se os números de ocorrência por 100 mil habitantes do caso de homicídios. O critério de ligação fez-se através da análise hierárquica simples, tendo como métrica a distância euclidiana.

Na Figura 24, observa-se o Dendograma gerado desse agrupamento hierárquico, nele mediu-se as distâncias entre todos os pares de cidades a partir dos dados de homicídios e essas

cidades vão sendo sequencialmente agrupadas na sequência das menores distâncias entre elas. É esse processo que gera o dendograma. Corumbataí é a última cidade agrupada e ela, é grupada isoladamente.

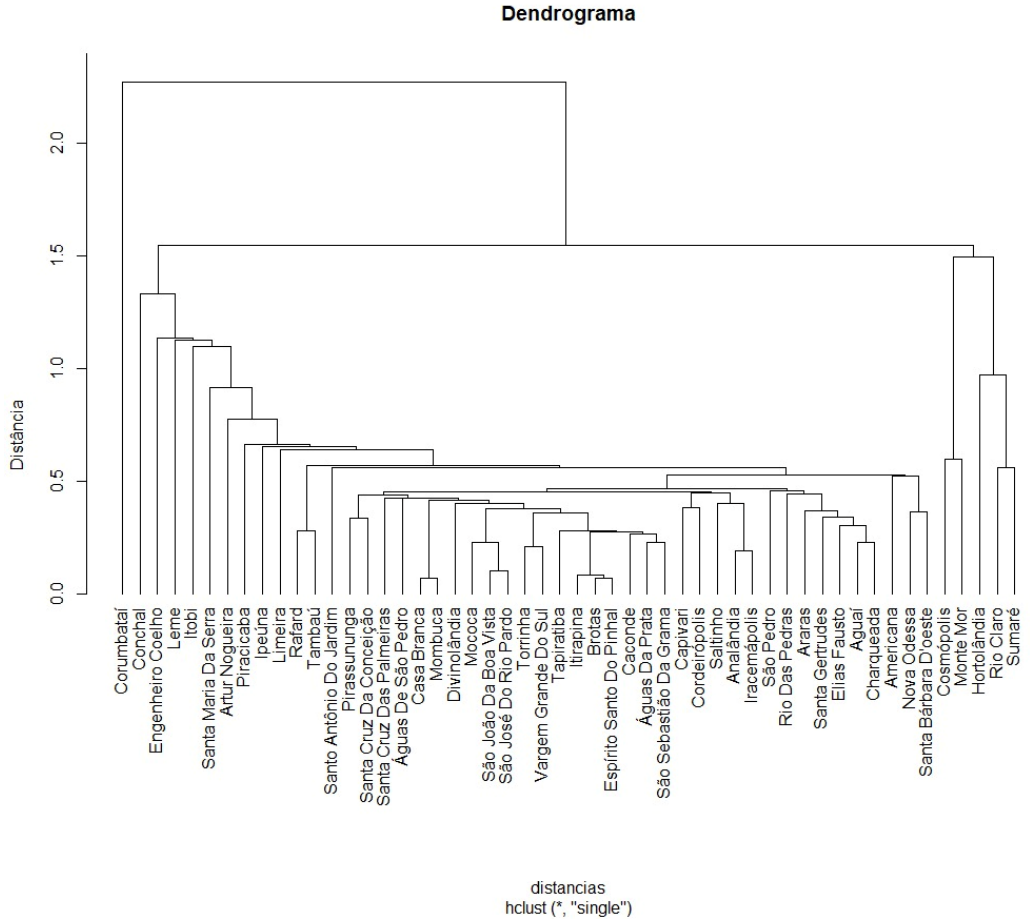


Figura 24 – Dendograma de homicídios ajustado.

Concluiu-se, dos resultados, que os modelos de regressão espacial produziram melhores resultados, em comparação com o modelo linear clássico quando analisa-se os casos de roubos de carros e outros roubos. A regressão clássica é o único resultado possível, quando analizado o caso de homicídios.

Tais conclusões estão coerentes com os resultados obtidos na literatura, como Domingues & Govone(2019), que compararam tais modelos no ajuste dos números de casos de dengue, por setores, na zona urbana de Rio Claro-SP, em função de variáveis socioeconômicas e concluíram que o modelo SAR apresentou melhor desempenho.

Analogamente, Araújo (2014) comparou os mesmos modelos na produtividade de soja de certa área, em função de variáveis climáticas, concluindo pelo melhor desempenho dos modelos SAR e SEM. Lima et al. (2005) concluíram que o modelo CAR (modelo SEM) apresentou melhor desempenho em relação a outros modelos, no estudo da distribuição espacial dos municípios

do estado de Pernambuco, quando ajustaram a regressão da taxa de homicídios em função de variáveis socioeconômicas.

5 CONCLUSÕES

Através das aplicações das técnicas de estatística espacial, pode-se identificar os municípios com maior incidência de casos de criminalidade, e também quais variáveis socioeconômicas estão relacionadas com esta incidência. Com o intuito de verificar a autocorrelação espacial, aplicamos o índice de Moran, o qual indicou autocorrelação entre as áreas. Ajustou-se três modelos de regressão: modelo linear clássico, modelo SAR e o modelo SEM, propondo um modelo que pudesse relacionar os índices de criminalidade com as variáveis socioeconômicas escolhidas para o estudo. No caso de roubos de carros e de outros roubos, os resultado obtidos na regressão linear clássica foram insatisfatórios por quebrarem as suposições usuais dos resíduos, além de não incorporarem a dependência espacial em sua estrutura. Assim, os modelos que incorporam estrutura espacial foram analisados, e se mostraram mais eficientes em relação a regressão linear clássica quando aplicados aos dados de roubos de carros e outros roubos. Ao comparar o modelo SAR e SEM com a regressão clássica, o melhor resultado obtido para o caso de roubos de carros foi o modelo que incorpora a dependência espacial, o modelo SAR. E no caso de outros roubos, o modelo SEM foi o mais adequado.

Já no caso dos homicídios, o cálculo do índice de Moran aos resíduos da regressão clássica, deu não significativo e buscou-se incluir os outros índices de criminalidade como variáveis explicativas neste caso, e mesmo assim, o índice permaneceu sendo não significativo. Com isso, o modelo linear clássico sem dependência espacial se manteve para o caso dos homícidios.

Concluiu-se, dos resultados, que os modelos de regressão espacial produziram melhores resultados, em comparação com o modelo linear clássico, acarretando independência dos resíduos.

Sugeriu-se como proposta para futuros trabalhos, testar os desempenhos de outros modelos de regressão espacial, como o modelo SAC - General Spatial Model (ALMEIDA, 2012), mais geral, que incorpora os modelos SAR e SEM num único modelo.

Referências

ALMEIDA, E. Econometria Espacial Aplicada. *Alínea Editora*, v. 1, p. 167–169, 2012. 6, 7, 8, 9, 43

AMOAKO, E. A. A spatial analysis of crime and neighborhood characteristics in Detroit census block groups. *Coleção PROPG Digital - UNESP*, v. 4, 2021. Disponível em: <https://ica-proc.copernicus.org/articles/4/5/2021/>. 15

ANDRADE; MONTEIRO. *Introdução à Estatística Espacial para a Saúde Pública.* 1ª. ed. [S.l.]: Brasil: Ministério da Saúde. Secretaria de Vigilância em Saúde. Fundação Oswaldo Cruz., 2007. v. 3. 1

ANSELIN, L. *Spatial econometrics: methods and models*. Boston: Kluwer Academic. ISBN: 90-247-3735-4, 1988. 7

ANSELIN, L. *Exploring spatial data with GeoDaTM : A workbook for spatial analysis.* [S.l.]: Laboratory Department of Geography - University of Illinois, UrbanaChampaign Urbana, IL 61801, Revised Version, March 6, 2005. 26

ANSELIN, L. et al. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, v. 26, n. 1, p. 77–104, 1996. ISSN 0166-0462. Disponível em: <https://www.sciencedirect.com/science/article/pii/0166046295021116>. 26

ARAUJO, E. C.; URIBE-OPAZO, M. A.; JOHANN, J. A. Modelo de regressão espacial para estimativa da produtividade da soja associada a variáveis agrometeorológicas na região oeste do estado do Paraná. *Engenharia Agrícola*, Associação Brasileira de Engenharia Agrícola, v. 34, p. 286–299, 2014. ISSN 0100-6916. Disponível em: <https://doi.org/10.1590/S0100-69162014000200010>. 14

ASSUNCAO, O. *Estatística espacial com aplicações em epidemiologia, economia e sociologia*. Ufscar, São Carlos: Minicurso da 7ª Escola de Modelos de Regressão, 2001. 13, 14

BAILEY, T.; GATRELL, A. *Interactive Spatial data analysis*. [S.l.]: Longman Scientific Technical, 1995. 1, 9

DARMOFAL, D. *Spatial econometrics and political sciense*. University of South Carolina, Colombia: Mimeo, Department of Political Sciense, 2006. 6

DOMINGUES, J. *Técnicas de processos espaciais e espaço-temporais com aplicações em dados de dengue.* [S.l.]: Botucatu: Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu, 2017. Dissertação-(mestrado). 1

DOMINGUES, J.; GOVONE, J. Avaliação de diferentes técnicas espaciais para análise da ocorrência de dengue em rio claro-sp. *Brazilian Journal of Biometrics*, v. 37(1), p. 1–16, 2019. 14, 26

DRUCK S.; CARVALHO, M. C. G. M. A. *Análise Espacial de Dados Geográficos*. [S.l.]: Brasília: EMBRAPA, 2004. 1, 5, 7

FLINGLETON, B. A generalized method of moments estimator for a spatial panel model with an endogenous spatial lag and spatial moving average errors. *Spatial Economic Analysis*, v. 3, p. 27–44, 2008. 7

JAMES, G. *An Introduction to Statistical Learning with Applications in R.* 1. ed. [S.l.]: New York: Springer, 2013. 3

LASAGE, J.; PACE, R. *Introduction to Spatial Econometrics*. New York: CRC Press Taylor & Francis Group, 2009. 3

LEVINE, N.; CECCATO, V. Malignant mixes: The overlap of motor vehicle crashes and crime in Stockholm, sweden. *Accident Analysis Prevention*, v. 161, p. 106361, 2021. ISSN 0001-4575. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0001457521003924>. 15

LIMA, M. L. C. d. et al. Análise espacial dos determinantes socioeconômicos dos homicídios no estado de pernambuco. *Revista de Saúde Pública*, Faculdade de Saúde Pública da Universidade de São Paulo, v. 39, n. 2, p. 176–182, Apr 2005. ISSN 0034-8910. Disponível em: <https://doi.org/10.1590/S0034-89102005000200006>. 15

LUCENA, W. *Agrupamento hierárquico*. 2019. <https://medium.com/@will.lucena> . Acesso em 20/06/2023. 40

MARTINS, W.; GOVONE, J.; AFFONSO, F. Estatísticas da variação temporal da criminalidade no comando de policiamento do interior-9, da polícia militar do estado de São Paulo, considerando-se a aplicação do diagnóstico evolutivo geoponderado(DEGEO) voltado aos municípios. *Trabalho submetido a publicação*, 2023. 16

PATLOLLA, C. R. *Understanding the concept of Hierarchical clustering Technique*. 2018. Disponível em: <https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec> . Acesso em 20/06/2023. 40

PM. *Divisão Operacional do CPI-9*. Piracicaba, São Paulo: PMESP, 2020. 16

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. Disponível em: <https://www.R-project.org/>. 27

REYNALDO, C. *Regressão "Ridge": Um Método Alternativo para o Mal Condicionamento da Matriz das Regressoras*. Dissertação (Mestrado) — Instituto de Matemática, Estatística e Computação Científica, UNICAMP, Campinas, 1997. 24

RODRIGUES, W. C. *Teste de Kolmogorov-Smirnov. DivEs - Diversidade de Espécies (AntSoft Systems On Demand)*. 2023. Disponível em: <https://www.ebras.bio.br/dives/guide_dives.aspx?IDTopic=920B6EF6-28679A59-A16B0B39&Topico=ST21> . Acesso em 10/06/2023. 24

SILVA, E. T. C. et al. Análise espacial da distribuição dos casos de dengue e sua relação com fatores socioambientais no estado da Paraíba, brasil, 2007-2016. *Saúde Em Debate*, v. 44(125), p. 465–477, 2020. 14

VIEIRA, R. d. S. Crescimento econômico no estado de São Paulo: uma análise espacial. *São Paulo: Cultura Acadêmica*, (Coleção PROPG Digital - UNESP), 2009. 6

WALLER L. A.; GOTWAY, C. A. *Applied Spatial Statistics for Public Health Data*. 1. ed. [S.l.]: New Jersey: John Wiley Sons, Inc, 2004. 3

Apêndices

A tabelas extras de roubos de carros

Tabela 24 – Modelo SAR completo

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	3.4778e+01	3.8464e+01	0.9042	0.3659063
pop	1.1605e-05	3.2450e-06	3.5764	0.0003484
IDHM	-7.5325e+00	1.0429e+01	-0.7223	0.4701284
ocup	-2.9713e-01	1.6154e-01	-1.8394	0.0658563
renda	2.2110e-01	1.4187e-01	1.5585	0.1191223
let	1.0473e-02	3.0112e-01	0.0348	0.9722548
gini	-5.1917e+00	5.7907e+00	-0.8966	0.3699543
ρ	0.4854			
LIK	-107.2703			
AIC	232.54			

Tabela 25 – Modelo SEM completo

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	2.7022e+01	3.9379e+01	0.6862	0.4925936
pop	1.1494e-05	3.3201e-06	3.4620	0.0005362
IDHM	-1.4684e+00	1.0505e+01	-0.1398	0.8888312
ocup	-2.6631e-01	1.5476e-01	-1.7208	0.0852938
renda	3.0001e-01	1.3351e-01	2.2471	0.0246313
let	5.6051e-02	3.0275e-01	0.1851	0.8531187
gini	-1.0643e+01	5.7659e+00	-1.8458	0.0649221
λ	0.56343			
LIK	-108.7602			
AIC	235.52			

B tabelas extras de outros roubos

Tabela 26 – Modelo SAR completo

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	2.3086e+01	8.2495e+00	2.7985	0.005134
pop	3.8944e-06	6.9209e-07	5.6270	1.834e-08
IDHM	-1.6923e+00	2.2564e+00	-0.7500	0.453237
ocup	-6.6896e-02	3.5022e-02	-1.9101	0.056117
renda	3.0002e-02	3.0165e-02	-0.9954	0.319533
let	-1.5155e-01	6.4217e-02	-2.3600	0.018274
gini	1.3886e+00	1.2347e+00	1.1247	0.260720
ρ	0.52232			
LIK	-27.32552			
AIC	72.651			

Tabela 27 – Modelo SEM completo

	Estimativa	**Erro padrão**	**Estimativa z**	**Pr(>|z|)**
Intercepto	2.3332e+01	7.8809e+00	2.9606	0.003070
pop	3.6407e-06	6.6823e-07	5.4483	5.085e-08
IDHM	1.4998e+00	2.1025e+00	0.7134	0.475620
ocup	-6.5639e-02	3.0829e-02	-2.1291	0.033243
renda	4.9439e-05	2.6427e-02	0.0019	0.998507
let	-1.5528e-01	6.0219e-02	-2.5785	0.009922
gini	-3.9018e-01	1.1667e+00	-0.3344	0.738057
lamba	0.70205			
LIK	-27.44971			
AIC	72.899			