**UNIVERSIDADE ABERTA**

# O USO DA ANÁLISE ESPACIAL NO ESTUDO DA RELAÇÃO ENTRE A EXPECTATIVA DOCENTE E A PROFICIÊNCIA DE ESCOLARES DO 9º ANO DO ENSINO FUNDAMENTAL NO BRASIL

PAULO MARCOS RIBEIRO

Mestrado em Estatística, Matemática e Computação

na área de especialização de Estatística Computacional

**2023**

**UNIVERSIDADE ABERTA**


# O USO DA ANÁLISE ESPACIAL NO ESTUDO DA RELAÇÃO ENTRE A EXPECTATIVA DOCENTE E A PROFICIÊNCIA DE ESCOLARES DO 9º ANO DO ENSINO FUNDAMENTAL NO BRASIL

PAULO MARCOS RIBEIRO

Mestrado em Estatística, Matemática e Computação

na área de especialização de Estatística Computacional

Dissertação de Mestrado orientada por

Professora Doutora Maria do Rosário Olaia Duarte Ramos

**Janeiro de 2023**

## AGRADECIMENOS

Em primeiro lugar, a gradeço à minha mãe, Eunice Pedroso que soube transformar as dificuldades da vida em exemplo de superação e vitória e sempre me impulsionou a ir adiante em meus estudos.

Aos meus irmãos, Letícia e Marco Aurélio, e meu sobrinho Pedro Henrique, pelo convívio, apoio e amizade.

Ao Anderson, meu companheiro de todas as horas, agradeço a discussão e a colaboração nesse projeto de pesquisa e na vida. Obrigado pela paciência e incentivo permanente.

Agradeço ao Programa de Mestrado em Estatística, Matemática e Computação da Universidade Aberta de Portugal, pela oportunidade de realizar esse curso e à minha orientadora, Maria do Rosário Ramos, pelas importantes orientações e esclarecimentos.

Agradeço, também, ao projeto de extensão “Psico&Econo_METRIA” da Universidade Federal de Uberlândia, na figura de seu coordenador, professor Dr. Pablo Rogers, pelo olhar crítico no desenvolvimento da análise quantitativa dessa dissertação.

*Dedicatória*

**A todos os professores da educação pública brasileira e aos elaboradores de políticas públicas educacionais para redução das desigualdades.**

**Seguimos na luta.**

UNIVERSIDADE ABERTA
www.uab.pt

**DECLARAÇÃO DE INTEGRIDADE**

**STATEMENT OF INTEGRITY**

Declaro ter atuado com integridade na elaboração da presente dissertação/tese. Confirmo que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer outra forma de falsificação de resultados.

Mais declaro que tomei conhecimento integral do Regulamento Disciplinar da Universidade Aberta, publicado no Diário da República, 2.ª série, n.º 215, de 6 de novembro de 2013.

I hereby declare having conducted my thesis with integrity. I confirm that I have not used plagiarism or any form of falsification of results in the process of the thesis elaboration.

I further declare that I have fully acknowledged Disciplinary Regulations of the Universidade Aberta (regulation published in the official journal Diário da República, 2.ª série, N.º 215, de 6 de novembro de 2013).

Universidade Aberta, 09 de janeiro de 2023

Nome completo/Full name: Paulo Marcos Ribeiro

Assinatura/Signature:

DocuSigned by:
Paulo Marcos Ribeiro
8B0931338DB14C6...
manuscrita ou digital / handwritten or digital

**O uso da análise espacial no estudo da relação entre a expectativa docente e a proficiência de escolares do 9º ano do ensino fundamental no Brasil**


## Resumo

No Brasil, a qualidade educacional é medida bianualmente por meio de testes padronizados, aplicados aos alunos, e questionários contextuais, aplicados aos gestores, professores e alunos dos anos finais da escolarização básica, sob a responsabilidade do Sistema Nacional de Avaliação da Educação Básica (SAEB). Este sistema apura o domínio de habilidades esperadas para cada etapa educacional, chamada de proficiência, e os fatores a ela relacionados. As expectativas de professores sobre o desempenho dos estudantes, é um importante fator de sucesso, amplamente divulgado pelo SAEB, mas ainda pouco estudado no que se refere à sua distribuição espacial. Com a finalidade de contribuir para essa lacuna de conhecimento, este trabalho teve por objetivo avaliar a influência da expectativa docente sobre a proficiência por meio do uso de componente espacial. Utilizou-se dados secundários do questionário docente e os resultados da proficiência em Matemática e Língua Portuguesa dos alunos do 9º ano do ensino fundamental – estudantes com 14 anos de idade – extraídos das edições SAEB de 2013 e 2017, agregados por municípios. A partir da construção de três variáveis de expectativa docente, através de uma Análise de Componentes Principais (PCA) e do emprego da variação das variáveis, com fins de controlar os efeitos fixos que não mudam com o tempo, aplicaram-se técnicas da econometria espacial. A Análise Exploratória de Dados Espacial (AEDE) indicou forte dependência espacial nas variáveis de desempenho discente, assim como os modelos espaciais estimados Modelo Espacial Autoregressivo (SAR), Modelo de Erro Espacial (SEM) e Combinação Autoregressiva Espacial (SAC) captaram o contundente papel da proximidade para a composição da interação espacial entre os fenômenos em avaliação. Os resultados indicaram que municípios nos quais a percepção docente sobre a indisciplina dos alunos e alto nível de faltas é maior, o desempenho discente é menor. Em termos práticos, principalmente para pesquisadores da educação, as evidências apontadas no estudo direcionam para a importância da modelagem espacial no estudo do desempenho de alunos do ensino fundamental no Brasil.



**Palavras-chave:** Desempenho Escolar. Análise Exploratória de Dados Espaciais. Modelos Espaciais.

**The use of spatial analysis in relationship between teacher expectation and performance of 9th grade students in Brazil**


## Abstract

In Brazil, educational quality is biennially assessed through standardized tests administered to students, as well as contextual questionnaires given to administrators, teachers, and students in the final years of basic education, under the responsibility of the National System for the Evaluation of Basic Education (SAEB). This system measures the mastery of expected skills for each educational stage, referred to as proficiency, and the factors associated with it. Teachers' expectations of student performance are an important success factor widely emphasized by SAEB, yet they have been relatively understudied in terms of their spatial distribution. To address this knowledge gap, this study aims to evaluate the influence of teachers' expectations on proficiency using a spatial component approach. Secondary data from teacher questionnaires and the results of Mathematics and Portuguese Language proficiency tests among 9th-grade students – 14-year-old students – extracted from the 2013 and 2017 SAEB editions were aggregated at the municipal level. By constructing three teacher expectation variables using Principal Component Analysis (PCA) and employing variable variations to control for fixed effects that do not change over time, spatial econometric techniques were applied. The Exploratory Spatial Data Analysis (ESDA) revealed a strong spatial dependence in student performance variables, while the estimated spatial models – Spatial Autoregressive Model (SAR), Spatial Error Model (SEM), and Spatial Autoregressive Combination Model (SAC) – captured the significant role of proximity in the spatial interaction among the phenomena under evaluation. The results indicated that municipalities with higher teacher perception of student indiscipline and high absenteeism exhibited lower student performance. In practical terms, particularly for education researchers, the evidence presented in this study emphasizes the importance of spatial modeling in studying the performance of elementary school students in Brazil.



**Keywords:** School Performance, Exploratory Analysis of Spatial Data, Spatial Models.

**ÍNDICE GERAL**

## ÍNDICE QUADROS E TABELAS

## ÍNDICE DE FIGURAS

## LISTA DE SIGLAS E ABREVIATURAS

| | |
|---|---|
| 2SLS | Mínimos Quadrados em Dois Estágios |
| AEDE | Análise Exploratória de Dados Espaciais |
| AIC | Critério de Informação de Akaike |
| BIC | Critério de Informação Bayesiano |
| GMM | Métodos dos Momentos Generalizados |
| GNS | Modelo Geral de Aninhamento Espacial |
| KMO | Medida Kaiser-Meyer-Olkin |
| LISA | Indicador Local de Associação Espacial |
| LM | Teste Multiplicador de Lagrange |
| LR | Teste de Razão de Verossimilhança |
| ML | Máxima Verossimilhança |
| MSA | Medidas de Adequação de Amostragem |
| OLS | Mínimos Quadrados Ordinários |
| PCA | Análise de Componentes Principais |
| SAC | Combinação Autoregressiva Espacial |
| SAEB | Sistema Nacional de Avaliação da Educação Básica |
| SAR | Modelo Espacial Autoregressivo |
| SDEM | Modelo de Erro Espacial de Durbin |
| SEM | Modelo de Erro Espacial |
| VI | Variáveis Instrumentais |
| VIF | Fator de Inflação da Variância |

## INTRODUÇÃO

Estudo clássico da área educacional, realizado por Coleman e colaboradores nos Estados Unidos na década de 1960, foi pioneiro em relacionar fatores extraescolares, como renda, origem social ao desempenho escolar dos estudantes (Brooke & Soares, 2008). Posteriormente, uma importante linha de estudos, com base em abordagens estatísticas, foi sendo consolidada no tema dos fatores relacionados ao desempenho escolar.

O desempenho escolar, é importante explicitar de início, está relacionado com as estruturas educacionais construídas em torno de grupos de indivíduos, sejam eles famílias, escolas, bairros ou grupos de amigos. A partir desses agrupamentos, os indivíduos compartilham opiniões, atitudes ou realizações (Laros & Marciano, 2010). Nesse contexto, pode-se dizer que o desempenho escolar é determinado por diversos fatores: o que o aluno traz consigo e aquilo que a escola oferece em termos de ensino, de instalações, ambiente e, ainda, pela percepção de docentes quanto ao desempenho dos alunos (Vidal & Vieira, 2017). Uma das constatações da literatura de Psicologia e Educação (Alvidrez & Weinstein, 1999; Palardy, 1969; Rosenthal & Jacobson, 1968; Teixeira, 2020; Vidal et al., 2019; Xavier & Oliveira, 2020) é que as chances de sucesso no desempenho escolar estão diretamente relacionadas com as expectativas docentes sobre os alunos.

O famoso estudo experimental de Rosenthal e Jacobson (1968), conhecido como "*Pygmalion in the Classroom*" é precursor na análise da relação entre as expectativas docentes e o desempenho escolar. Nesse trabalho, os autores discutem os efeitos da manipulação das expectativas dos professores sobre a capacidade e aptidão dos alunos. Para tanto, forneceram informações falsas sobre o desempenho dos alunos em um teste inexistente, induzindo a formação de elevadas expectativas. Estudantes informados como mais capazes acabaram obtendo melhores desempenhos na escola, confirmando a expectativa positiva do professor. Essas descobertas forneceram importantes indicativos de que expectativas docentes imprecisas podem gerar as chamadas "profecias autorrealizáveis" (Vidal et al., 2019).

A partir da literatura especializada podem ser identificadas diversas variáveis contextuais relacionadas ao desempenho escolar (Jesus & Laros, 2004). Caso se queira pesquisar a complexidade de fatores relacionados ao desempenho escolar, é preciso dispor de instrumentos de modelagem que envolvam um nível comparável de complexidade. A partir desse quadro geral, o presente estudo tem a intenção de avançar nessa temática, ao

propor uma abordagem estatística não recorrente para avaliar a relação: expectativas docentes e desempenho educacional.

Assim, esta dissertação tem por objetivo relacionar a percepção e expectativa de docentes da educação básica ao desempenho escolar no nível agregado dos munícipios brasileiros, seguindo a abordagem da econometria (estatística) espacial. O campo da econometria espacial, segundo Vieira (2009), se diferencia da estatística espacial: a primeira se pauta em um modelo ou teoria em particular e tem, como foco a economia regional e urbana, enquanto a estatística espacial trata, de modo primordial, de fenômenos naturais, ligados, principalmente, a campos como a biologia e geologia (Vieira, 2009), com interesse em exame de superfícies contínuas.

Anselin (1988) coloca que a distinção entre esses dois campos é sutil, visto que os métodos de uma são utilizados pela outra. Segundo Anselin (1988), os pesquisadores que devem definir em quais dos dois campos seu trabalho se refere, e nesse sentido, devido ao uso de dados secundários sociodemográficos de municípios, a linguagem da econometria espacial torna-se mais condizente para os presentes fins.

Para expor a aplicação dos métodos relacionados com a econometria espacial na presente dissertação, foram utilizados dados secundários extraídos das bases de respostas dos docentes ao questionário do Sistema Nacional de Avaliação da Educação Básica (SAEB), de 2013 e de 2017 – uma avaliação em larga escala censitária, bianual, realizada pelo Ministério da Educação do Brasil –, e os resultados nos testes padronizados dos alunos do 9º ano do ensino fundamental em matemática e língua portuguesa.

Importante explicitar que o SAEB foi a primeira iniciativa brasileira, em termos de política educacional pública, para se conhecer a qualidade da educação deste país. Começou a ser desenvolvido no final dos anos de 1980 e foi aplicado pela primeira vez em 1990. Em 1995, o SAEB passou por uma reestruturação metodológica que possibilitou a comparação do desempenho escolar ao longo dos anos, ao utilizar a Teoria da Resposta ao Item, como base de suas medidas. Desde a sua primeira avaliação fornece dados sobre a qualidade dos sistemas educacionais do Brasil como um todo, das regiões geográficas e das unidades federadas (estados e Distrito Federal). Dois instrumentos são aplicados pelo SAEB:

a) Testes padronizados: direcionados aos alunos da alfabetização (aos 6 anos de idade); ensino fundamental I – 6º ano (aos 8 anos de idade); ensino

fundamental II – 9º ano (aos 14 anos de idade); e, ensino médio – 3º ano (aos 17 anos de idade)[1].

b) Questionários contextuais: direcionados aos gestores, professores, docentes e familiares dos estudantes avaliados[2].

Desses instrumentos são extraídas medidas de proficiência, traduzidas em escalas pedagógicas sobre o desempenho escolar e fatores associados ao desempenho. Tanto a proficiência quanto os fatores associados, em especial as expectativas docentes, constituem as variáveis de estudo desta dissertação.

Para calcular a proficiência, as respostas dos alunos são processadas, pelo SAEB, seguindo o modelo logístico unidimensional de três parâmetros da Teoria da Resposta ao Item (dificuldade, acerto ao acaso e discriminação), de maneira a relacionar os parâmetros dos itens e as proficiências dos alunos. Essa relação é expressa por meio de uma função monotônica crescente que indica que quanto maior a proficiência do aluno, maior será sua probabilidade de acertar o item. São realizadas, também, equalizações, de forma a colocar as proficiências dos alunos e parâmetros dos itens em escalas de proficiência que ordenam as médias obtidas em uma régua de desempenho de 0 a 500 pontos, dividida em intervalos de 25 pontos (Instituto Reúna, 2021).

Para Língua Portuguesa e Matemática, em todas as etapas avaliadas, a escala é única e cumulativa, ou seja, a escala aplicada para o 5º ano do Ensino Fundamental é a mesma para o 9º ano e para a 3ª série do Ensino Médio[3]. A lógica é a de que, quanto mais o estudante caminha ao longo da escala, mais habilidades terá acumulado. Assim, é esperado que alunos Ensino Fundamental I alcancem médias numéricas menores do que os do Fundamental II (9º ano) e estes alcancem médias menores do que as dos alunos do Ensino Médio.

Uma vez que cada área do conhecimento possui sua própria escala, elas são incomparáveis entre si. Ou seja, uma mesma escola que atinge 250, tanto para Matemática quanto para Língua Portuguesa, deve tratar os resultados de maneira distinta. O valor numérico pode ser o mesmo número, mas, pelo fato de estarem em escalas diferentes, as

[1] Conferir https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb

[2] Os questionário podem ser acessados pelos seguintes links: questionário do Professor 2013 < https://download.inep.gov.br/educacao_basica/saeb/aneb_anresc/quest_contextuais/2013/questionario_professor_2013.pdf> questionário do Professor 2017 < https://download.inep.gov.br/educacao_basica/saeb/aneb_anresc/quest_contextuais/2017/questionario_professor_2017.pdf>

[3] Exceto para a alfabetização, que segue escala própria definida em 2013 com a Avaliação Nacional da Alfabetização, com intervalos variando entre 0 e 1.000

proficiências não podem ser comparadas e, portanto, não permitem conclusões de que determinada proficiência em uma área do conhecimento é maior, menor ou igual a outra proficiência de outra área, nem tampouco entender se uma certa proficiência em uma área implica alguma interferência na outra, pois cada escala trará sempre uma informação diferente para cada um de seus níveis.

Os questionários contextuais, por sua vez, são analisados pelo SAEB, por meio de correlação, análise fatorial, análise multinível, Teoria da Resposta ao Item, entre outras, de maneira a estabelecer relações entre a proficiência dos alunos e fatores associados. Com base nos questionários contextuais, é possível, por exemplo, compreender a influência de cor, raça, gênero e nível socioeconômico na proficiência, relacionar as expectativas docentes e ação dos gestores ao desempenho dos alunos, entre outros fatores explicativos da desigualdade educacional brasileira (Instituto Reúna 2021).

No Brasil, os resultados do SAEB são apresentados para cada escola, município e estado, de forma a informar, à sociedade, padrões de qualidade educacionais alcançados pelo sistema[4].

Especificamente para este trabalho, justifica-se o foco no 9º ano do ensino fundamental tendo em vista a disponibilidade de dados, uma vez que nesta etapa houve questionários contextuais amplamente aplicados em nível nacional, e por ser essa a etapa crucial para o avanço da escolarização em direção ao ensino médio. Cabe considerar que os questionários contextuais, tal como informado pelo próprio SAEB, são aprimorados a cada edição. As edições de 2013 e 2017 apresentam as mesmas variáveis de análise, no que se refere à expectativa docente foco deste estudo.

Importante explicitar que a literatura já apresenta evidências favoráveis à modelagem espacial que justificam e dão pistas à exequibilidade do trabalho. Nesse sentido, Fujita et al. (2021) e Vernier (2016) encontraram forte dependência espacial, sugerindo que a estrutura espacial tem influência no desempenho escolar: o desempenho de um município está positivamente associado ao desempenho dos municípios vizinhos.

A proposta da presente dissertação reside, portanto, em ajustar modelos econométricos espaciais tendo como variável dependente (Y) a variação da proficiência em língua portuguesa e matemática dos alunos do 9º ano do ensino fundamental entre 2013 e

[4] Os resultados do SAEB podem ser acessados pelo seguinte link https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/resultados

2017 dos municípios brasileiros, e as variações das expectativas docentes como variável independente de interesse (X) nesse mesmo período, conforme propostas de mensuração apresentadas em Teixeira (2020) e Vidal et al. (2019), obtidas com base no questionário que avalia as percepções docentes.

Assim, buscar-se-á especificar, estimar e testar modelos teóricos influenciados pelos efeitos espaciais (dependência e/ou heterogeneidade espacial) usando os dados agregados do SAEB. Esses modelos buscam captar o papel da proximidade para o surgimento da interação espacial entre os fenômenos (Almeida, 2012).

Entretanto, antes de se avançar na modelagem do fenômeno em estudo, buscar-se-á fazer uma Análise Exploratória de Dados Espacial (AEDE), que, segundo Almeida (2012), são ferramentas que possibilitam manipular dados espaciais de diferentes formas e extrair conhecimento adicional como resposta ao dispor em mapas os possíveis padrões e relações entre variáveis. Esse tipo de análise inclui funções como consulta de informações espaciais dentro de áreas de interesse definidas, manipulação de mapas e a produção de alguns breves sumários estatísticos dessa informação; incorporando também funções como a investigação de padrões, agrupamentos e relacionamentos dos dados na região de interesse, buscando um melhor entendimento do fenômeno.

A partir do tema e objeto proposto no presente trabalho e do conceito geral da AEDE, a dissertação se propõe a investigar a distribuição espacial da percepção e expectativa docente, em relação ao desempenho dos estudantes, tendo por conjectura que fatores associados ao desempenho, tais como infraestrutura das escolas, sobrecarga de trabalho dos professores, características dos alunos, entre outros, podem apresentar aspectos de dependência ou heterogeneidade espacial entre municípios, passíveis de serem captadas pela AEDE.

Tendo em vista que as técnicas estatísticas da AEDE permitem descrever a distribuição das variáveis e identificar observações não só em relação ao tipo de distribuição, mas também em relação aos vizinhos e buscar a existência de padrões na distribuição espacial, estima-se ser possível estabelecer hipóteses sobre as observações, de forma a selecionar o modelo inferencial mais bem suportado pelos dados.

Considerando que a educação básica brasileira vem acumulando problemas de aprendizagem dos alunos, sobretudo aqueles ao fim da segunda etapa de escolarização, ou seja, no 9º ano, momento crucial para a entrada no ensino médio, justifica-se o tema aqui

proposto na medida em que os resultados podem contribuir para elucidar novas facetas acerca da complexa relação entre a percepção e expectativa docente e o desempenho escolar.

Estruturalmente essa dissertação está organizada em duas partes. A primeira parte, dividida em três capítulos, apresenta o referencial teórico e metodológico da Econometria Espacial utilizado no trabalho. No primeiro capítulo discute-se o conceito da Econometria Espacial, a dependência e heterogeneidade espacial, dados espaciais e seus desafios. No segundo capítulo, a primeira parte apresenta as matrizes de ponderação espacial. A análise exploratória de dados espaciais, em especial a autocorrelação espacial, é apresentada na segunda parte do capítulo. No terceiro capítulo, são apresentados os modelos de dependência espacial utilizados na parte empírica e seus métodos de estimação.

Na segunda parte dessa dissertação, discutem-se os resultados encontrados no estudo em quatro capítulos. No quarto discutem-se as variáveis do estudo, com atenção especial na construção das variáveis de percepções e expectativas docentes, devido a necessidade de emprego de uma Análise de Componentes Principais (PCA) para estimação dos fatores (variáveis) utilizados como variáveis independentes de interesse. No quinto e sexto capítulo os resultados são discutidos à luz do enquadramento teórico e metodológico utilizado, segregados na AEDE e nos modelos espaciais. No capítulo final, são apresentadas as análises conclusivas, além das limitações do estudo e perspectivas abertas.

Os dados utilizados nessa dissertação são padronizados, agregados por munícipios (n = 4.661), e foram analisados da seguinte forma pelos *softwares*:

i) SPSS para estimar os fatores do questionário "Problemas de Aprendizagem" respondido pelos professores durante a aplicação do SAEB. Os itens Q70 a Q82 foram selecionados para essa análise, e o método utilizado foi o PCA;

ii) GeoDa para análise exploratória dos dados espaciais, matriz de ponderação espacial e análise de autocorrelação espacial, e;

iii) GeoDaSpace utilizado para estimar modelos que controlam tanto a autocorrelação espacial quanto a heterocedasticidade espacial.

# PARTE I – REFERENCIAL TEÓRICO E METODOLÓGICO

Esta parte tem por objetivo apresentar os capítulos teóricos e metodológicos da dissertação, no âmbito do problema estudado, no que se refere à análise de possíveis modelos econométricos espaciais a serem utilizados. Essa parte da dissertação segue de perto a estrutura proposta por Almeida (2012). O capítulo adiante apresenta uma definição da Econometria Espacial e analisa a natureza e característica dos efeitos espaciais, notadamente importantes para a análise dos dados proposta nesse trabalho.

## 1. A Econometria Espacial

Segundo Almeida (2012), a Econometria Espacial, um ramo da Econometria, tem por objetivo especificar, testar, estimar e prever modelos teóricos a partir dos efeitos espaciais, por meio de dados em painel ou corte transversal. As observações, nesse sentido, referem-se ao estudo dos fenômenos em relação às regiões geográficas em que estes ocorrem. Do ponto de vista metodológico, a diferença entre econometria convencional e espacial reside na incorporação dos chamados efeitos espaciais na regressão, a saber, a dependência espacial, a heterogeneidade espacial e a imbricação dos efeitos espaciais (Almeida, 2012). Os efeitos que distinguem a econometria clássica da espacial são discutidos a seguir.

### 1.1. A Dependência espacial

Segundo Almeida (2012, p.22), a dependência espacial é um *cross sectional dependence*, que aparece quando as unidades de corte transversal, sejam indivíduos, instituições, regiões ou outros agregados, não são mais independentes entre si.

Importante considerar que todo o processo que se dá no espaço está sujeito à chamada Lei de Tobler, também conhecida como a Primeira Lei da Geografia, cujo enunciado pode ser estabelecido da seguinte forma: tudo depende de todo o restante, porém o que está mais próximo depende mais do que aquilo que está mais distante (Tobler, 1970). A Lei de Tobler, portanto, destaca o papel da proximidade para o segmento da interação espacial entre os fenômenos. Proximidade, nesse sentido, indica a noção de distância relativa entre as regiões e seus efeitos. O efeito “distância” deve ser tomado, nessa perspectiva, de

modo amplo, não apenas geográfico, mas também no sentido de fenômenos sociais, como a distância social, a distância econômica, a distância política, entre outras (Anselin, 1988).

A econometria espacial, de maneira geral, pauta-se em um modelo ou teoria em particular e tem, como foco, a economia regional e urbana. A abordagem da econometria espacial consiste basicamente em impor a estrutura do problema por meio da especificação de um modelo *a priori*, ao associá-lo a um teste de especificação com contrapartida em uma hipótese nula. De acordo com Almeida (2012), a dependência espacial implica que o valor de uma variável de interesse, $y_i$, em uma determinada região $i$, depende não apenas do valor dessa variável em regiões próximas, $y_j$, mas também de um conjunto de variáveis explicativas externas representadas pela matriz X. Essa ideia pode ser resumida como:

$$y_i = f(y_j, X), \qquad i, j = 1, \dots, n \quad e \quad i \neq j. \tag{1}$$

Observa-se, a partir da equação (1), que o comportamento da variável dependente ($y_i$) não se limita à influência dos fatores exógenos ($X$), mas também é afetado pelos valores dessa variável em regiões vizinhas, conforme o critério de vizinhança (Almeida, 2012, p.22). Segundo Almeida (2012), a dependência espacial pode ser representada por meio de um esquema simplificado que captura as interações entre as regiões. Por exemplo, se considerarmos duas regiões, uma única variável independente e uma forma linear, a equação (1) assume a seguinte forma:

$$y_1 = \rho_2 y_2 + \beta X_1 + \varepsilon_1. \tag{2}$$

É notável que a variável dependente na região 1 é afetada pela variável dependente na região vizinha 2, conforme indicado pelo coeficiente $\rho_2$,, além da presença da variável explicativa $X_1$ na região 1 e um termo de erro aleatório. De maneira semelhante, ocorre um evento análogo ao analisar o que acontece na região vizinha 2:

$$y_2 = \rho_1 y_1 + \beta X_2 + \varepsilon_2. \tag{3}$$

Na região 2, a variável dependente é afetada pela variável dependente na região 1, juntamente com a variável explicativa específica da região 2 ($X_2$) e um termo de erro aleatório. Isso resulta em uma situação de simultaneidade espacial, em que a variável y na região 1 é influenciada pela variável dependente na região 2, e, por sua vez, essa primeira

variável influencia a última (Almeida, 2012, p.22). Nesse contexto, essa situação pode ser melhor representada por meio de um sistema de equações simultâneas que refletem essa retroalimentação espacial, composto pelas equações (2) e (3).

Almeida (2012), nesse contexto, destaca a importância de reconhecer a multidirecionalidade dos processos espaciais em contraste com a unidirecionalidade dos processos temporais. Ao contrário do espaço, o tempo flui de forma unidirecional, do passado para o presente e do presente para o futuro. Isso implica que, devido a essa unidirecionalidade, $y_{t-1}$ influencia $y_t$, mas $y_t$ não influencia $y_{t-1}$. A influência no tempo ocorre apenas em direção ao futuro, enquanto no espaço, a influência pode ocorrer para frente, para trás, para cima ou para baixo. No domínio temporal, não surge o problema de simultaneidade presente nos fenômenos espaciais.

Ao generalizar a situação de simultaneidade espacial para $n$ regiões, descrita pelas equações (2) e (3), obtemos o seguinte sistema de equações simultâneas, considerando apenas uma variável explicativa para simplificar:

$$\begin{cases} y_1 = \rho_2 y_2 + \cdots + \rho_n y_n + \beta X_1 + \varepsilon_1 \\ \vdots \\ y_n = \rho_1 y_1 + \cdots + \rho_{n-1} y_{n-1} + \beta X_n + \varepsilon_n \end{cases} . \qquad (4)$$

A variável dependente observada na região 1, na primeira equação do sistema (1), é afetada pelas variáveis endógenas $y$ presentes nas outras regiões. De forma semelhante, isso se repete para cada uma das $n$ equações do sistema.

Almeida (2012, p. 23) ilustra essa simultaneidade com o exemplo da análise da criminalidade regional. De acordo com o autor, as taxas de criminalidade em uma região são explicadas por um conjunto de variáveis independentes, como a renda *per capita* da região, o nível de desigualdade de renda, a quantidade de policiais em cada região, a taxa de urbanização, entre outros. Além disso, a ocorrência de crimes em uma determinada região é influenciada pelas taxas de criminalidade nas regiões vizinhas. Por sua vez, a criminalidade em uma região vizinha é afetada pela taxa de crimes na região em questão.

Em relação às fontes primárias de dependência espacial, Almeida (2012) destaca três: a interação espacial, o erro de medida dos dados espaciais e a má especificação do modelo.

A primeira é de natureza teórica e está relacionada a uma variedade de processos espaciais, onde eventos ou circunstâncias em um lugar podem afetar as condições de outros lugares se houver interação entre eles (Anselin, 1988).

A segunda fonte de dependência decorre do erro de medição em dados espaciais. Um erro comum nesses dados é causado pela falta de correspondência entre o escopo do fenômeno em estudo e a divisão espacial das unidades disponíveis nos dados. Anselin (1988) alerta que os dados espaciais são coletados em níveis de agregação, como setores censitários, municípios, estados, distritos, e isso pode resultar em pouca correspondência entre o escopo espacial do fenômeno em estudo e o delineamento das unidades de observação. Essa falta de correspondência pode gerar erros de medição que tendem a se propagar de uma unidade espacial para outra, resultando em dependência espacial.

A terceira fonte de dependência é a má especificação do modelo, que pode ocorrer quando variáveis relevantes com padrão espacial são omitidas ou quando outliers espaciais ou pontos de alavancagem influenciam o modelo.

## 1.2. Heterogeneidade espacial

O princípio da heterogeneidade, conforme (Goodchild, 2004), pode ser expresso como:

$$y_i = f_i(X_i, \beta_i, \xi_i), \qquad \xi_i \sim (0, \Omega). \tag{5}$$

Nessa equação, $f_i$ representa a forma matemática funcional e $\xi_i$ é o termo de erro. O símbolo Ω denota a matriz de variância e covariância, em que a diagonal principal não consiste em constantes. A restante da notação utilizada é a mesma (Almeida, 2012, p.27).

Para Almeida (2012), a heterogeneidade espacial se manifesta na presença de instabilidade estrutural entre as regiões, resultando em diferentes respostas dependentes da localidade ou da escala espacial na forma de coeficientes variáveis ou regimes espaciais ($\beta_i$). Em um conjunto de dados espaciais, a relação entre a variável dependente e as variáveis independentes pode ser linear, enquanto em outro subconjunto, essa relação pode ser não linear.

As fontes de heterogeneidade espacial incluem as características da estrutura espacial, o erro de medição nos dados e a má especificação do modelo econométrico. Essas duas últimas fontes também são comuns à dependência espacial.

## 1.3. Dados espaciais e seus desafios

Conforme Almeida (2012, p. 48-49), na econometria espacial, os dados precisam ser considerados espaciais para serem incorporados aos modelos, o que implica ter uma ordem no domínio do espaço ou do espaço-tempo. Os dados espaciais possuem duas propriedades distintas: i) a magnitude da variação do atributo do fenômeno em cada estudo e ii) a natureza espacial, que indica a referência explícita em termos de localização geográfica desse atributo, revelando como os dados estão distribuídos no espaço. Por outro lado, os dados não espaciais possuem apenas a propriedade da variação do atributo do fenômeno.

Os dados espaciais apresentam desafios decorrentes das suas próprias características, tais como:

- Georreferenciamento: a posição relativa ou absoluta dos dados no mapa é importante, pois transmite informações valiosas.
- Multidirecionalidade: os dados observados em regiões interagem em todas as direções, o que leva à endogeneidade da interação espacial.
- Multidimensionalidade: o grau ou a dimensão da dependência espacial pode variar conforme a direção em que ocorre.

## 1.4. Problemas com dados espaciais

Os dados espaciais apresentam alguns problemas que podem ter impacto negativo na análise econométrica. Esses problemas incluem:

i) Falácia ecológica: refere-se aos erros resultantes da inferência de comportamentos individuais com base na análise de dados agregados, levando a resultados diferentes daqueles que seriam obtidos se dados individuais fossem utilizados. Esse termo foi introduzido por Robinson (1950) e destaca o risco de tirar conclusões sobre indivíduos com base em estudos com dados agregados. No contexto da econometria espacial, a preocupação recai sobre a agregação regional ou por área geográfica.

ii) Problema da unidade de área modificável: relaciona-se à sensibilidade dos resultados em relação à dimensão e à configuração da área de estudo. Isso significa que os resultados podem variar significativamente dependendo de como as unidades espaciais são definidas e agrupadas.

iii) Efeito de beirada: refere-se à influência das fronteiras geográficas na análise espacial. Regiões localizadas nas bordas podem ter características diferentes das regiões internas devido a interações com áreas vizinhas, o que pode distorcer os resultados.

iv) Influência de *outliers* espaciais: *outliers*, ou observações extremas, que exibem comportamento atípico em termos espaciais podem exercer influência desproporcional sobre os resultados da análise.

Esses problemas destacam a importância de considerar cuidadosamente os dados espaciais na análise econométrica, levando em conta as peculiaridades das relações e interações espaciais.

Gelman et al. (2001) propõem um modelo que representa a natureza da inferência ecológica. Consideremos $i = 1, \cdots, m$ indivíduos em uma região $j = 1, \cdots, n$. Seja $y_{ij}$ a variável dependente ou de resposta para o indivíduo $i$ na região $j$, $x_{ij}$ a variável explicativa correspondente, e $\varepsilon_{ij}$ um termo de erro com média zero e independente de $x_{ij}$, o que garante a identificação do modelo. Assim, o modelo é especificado como (Haining, 2003):

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij}. \quad (6)$$

O objetivo é estimar $\alpha_j$ e $\beta_j$, porém, não há dados individuais disponíveis, apenas médias para cada região, representadas por $\bar{y}_j$ e $\bar{x}_j$. Assumindo que o termo de erro seja próximo de zero, podemos estimar o seguinte modelo:

$$\bar{y}_j = \alpha_j + \beta_j \bar{x}_j; \quad (7)$$

A regressão ecológica é dada por:

$$\bar{y}_j = \alpha + \beta \bar{x}_j + \eta_j. \quad (8)$$

Sendo que $\eta_j$ é um termo de erro com média zero e não correlacionado com $\bar{x}_j$, independentes entre si, a incorrência na referência ecológica leva à análise dos parâmetros estimados $\alpha$ e $\beta$ na equação (8) como se fossem os parâmetros de interesse estimados $\alpha_j$ e $\beta_j$. O viés ecológico é caracterizado pelas diferenças $(\alpha_j - \alpha)$ e $(\beta_j - \beta)$.

De acordo com Almeida (2012, p. 60), existem várias razões pelas quais é necessário inferir, mesmo que parcialmente, o comportamento individual a partir de dados agregados. Um dos motivos é que alguns comportamentos individuais são influenciados pelo comportamento do grupo de indivíduos. Isso ocorre porque o indivíduo é afetado pelo ambiente externo e pelos efeitos de contexto, o que é particularmente relevante para a dissertação em questão. Por exemplo, o desempenho escolar de um estudante depende tanto das características individuais quanto de fatores do grupo ou da região em que a escola está localizada.

Em segundo lugar, devido à falta de dados em nível individual, alguns estudos só podem ser realizados utilizando dados agregados por área. Os dados eleitorais são um exemplo ilustrativo, uma vez que a disponibilidade de dados individuais de voto é impossibilitada por razões legais, devido ao direito ao voto secreto. Para Almeida (2012), qualquer solução para o problema da inferência ecológica depende da validade dos pressupostos assumidos ao aplicá-la a um problema específico.

Frequentemente, os resultados da análise de dados agregados dependem da definição do critério usado para a agregação espacial dos dados, e a inferência a partir de dados espaciais pode ser enganosa e levar a erros se os devidos cuidados não forem tomados com relação a esses problemas.

O problema da unidade de área modificável afeta tanto a análise univariada quanto multivariada, (Fotheringham & Wong, 1991). No contexto multivariado, esse problema gera incerteza em relação à validade dos resultados obtidos por meio da análise econométrica. Além disso, a unidade de área modificável restringe a possibilidade de replicação de um modelo em outras regiões de estudo, caso a escala e o zoneamento sejam diferentes dos utilizados na aplicação inicial.

Conforme destacado por Anselin (1988), a abordagem econométrica espacial pode abordar as questões de zoneamento e escala, pois cada uma delas está associada a um dos efeitos espaciais. O problema da unidade de área modificável está relacionado a um desafio

econometricamente relevante de zoneamento, que se refere ao efeito da heterogeneidade espacial.

Quando se aplica a metodologia econometricamente espacial, surge outro desafio com os dados espaciais conhecido como efeito de borda (*edge effect*), no qual as regiões dentro da área de estudo não capturam completamente a dependência espacial envolvida pelo fenômeno em análise. Isso resulta em uma correlação espacial entre as observações próximas à fronteira e regiões que estão além da área de estudo original (Darmofal, 2006). Em outras palavras, o efeito de borda ocorre quando as observações de regiões fora da área de estudo - porém suficientemente próximas para influenciar as regiões dentro da área de estudo - podem impactar a estimação e o teste de hipótese (Anselin, 1988).

Para lidar com o problema dos valores de fronteira e suas correções nos dados da amostra, Griffith (1983) propôs a criação de vizinhos artificiais para as regiões de fronteira, "dobrando o mapa" de forma que as regiões adjacentes à área de estudo se tornassem vizinhas diretas das regiões de fronteira oposta. No entanto, a questão principal é determinar se o pressuposto de que esses vizinhos artificiais apresentam interação espacial com as regiões de fronteira é válido. Como observado por Darmofal (2006), nesse caso, a solução pode ser pior do que o problema em si.

Uma possível correção para os efeitos de beirada nos resultados econométricos é expandir a área de estudo, considerando uma zona ao longo de toda a fronteira. Isso implica incluir as regiões adjacentes à área de estudo e próximas das suas fronteiras. Essa abordagem busca capturar a influência espacial das regiões vizinhas e mitigar o viés decorrente da falta de observações diretas nas fronteiras da área de estudo.

As técnicas de Análise Exploratória de Dados Espaciais (AEDE) desempenham um papel importante no desenvolvimento das etapas da modelagem estatística espacial. Essas técnicas são sensíveis ao tipo de distribuição dos dados, à presença de valores extremos e à falta de estacionariedade espacial. Elas são, em geral, adaptações das ferramentas tradicionais de análise exploratória de dados. Por exemplo, onde se utilizam histogramas ou *boxplots* para investigar valores extremos em análises convencionais, na análise espacial esses valores são investigados por meio de mapas (*boxmaps*), não apenas considerando o conjunto dos dados, mas também em relação aos seus vizinhos espaciais. Essa abordagem permite identificar padrões e detectar possíveis *outliers* espaciais, que podem ser relevantes para a compreensão dos processos espaciais em estudo.

Os *outliers* espaciais são observações em uma base de dados espaciais que possuem uma dependência espacial diferente das demais observações vizinhas. A presença de *outliers* e pontos de alavancagem pode estar relacionada a erros de medida durante a obtenção e armazenamento dos dados. No entanto, nem todos os *outliers* globais ou espaciais indicam erros grosseiros que precisam ser corrigidos ou descartados. Esses valores extremos podem, na verdade, fornecer informações relevantes sobre características legítimas do fenômeno em análise e merecem investigação adicional. Portanto, é importante considerar que os *outliers* espaciais podem oferecer *insights* valiosos e não devem ser automaticamente descartados, mas sim estudados para compreender melhor o fenômeno subjacente.

## 2. Análise Exploratória de Dados Espaciais (AEDE)

Este capítulo apresenta a Análise Exploratória de Dados Espaciais (AEDE), no âmbito da pesquisa proposta nesta dissertação. A AEDE é uma abordagem que agrupa várias técnicas para visualizar distribuições espaciais com o objetivo de identificar localidades atípicas, descobrir padrões de associação espacial (*clusters* espaciais) e sugerir regimes espaciais (heterogeneidade espacial) e outras formas de instabilidade espacial (Anselin, 2005).

Trata-se, portanto, de um ponto de partida para seguir às análises confirmatórias ou à modelagem econométrica, propriamente ditas. Conforme Fotheringham et al. (2003), antes de se fazer qualquer análise estatística sofisticada, é importante efetuar inicialmente alguma AEDE, de modo que esta análise preceda uma apropriada modelagem econométrico-espacial, pois auxilia no processo de especificação dos modelos em processos espaciais.

Uma primeira análise descritiva torna-se possível desenhar mapas com as unidades de observação divididas em faixas de valores para uma determinada variável (Tyszler, 2006) ou até mesmo a descrição da variabilidade dessa variável, tal como podemos visualizar para a variável Proficiência em Matemática em 2017 na Figura 2.1. Essas ilustrações são materializações de gráficos clássicos da estatística, no entanto, referenciados espacialmente. Por exemplo, parte dessas informações (Figura 2.1) também poderia ser originada de um histograma ou *boxplot*, como exemplificado na Figura: 2.2.

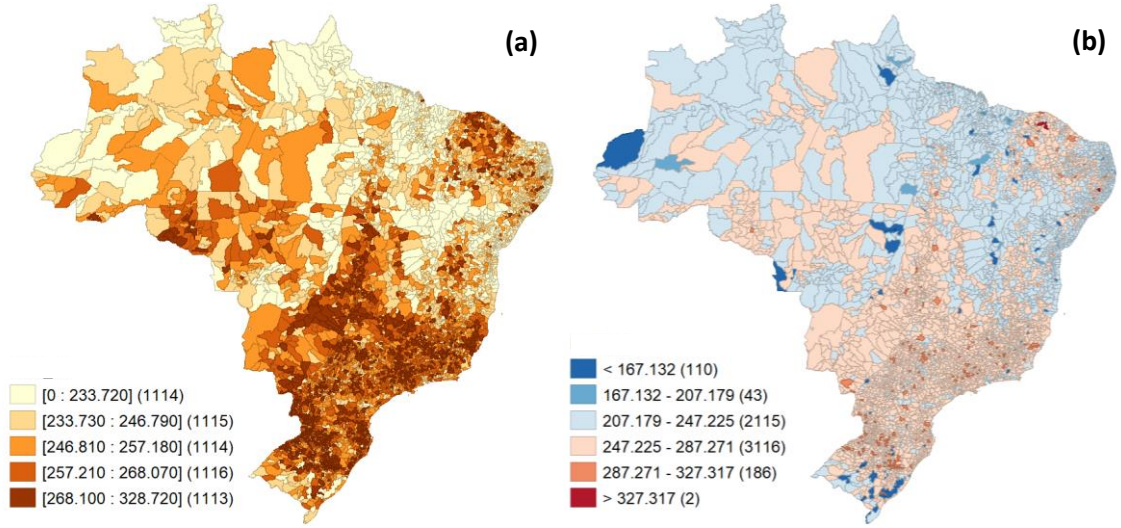


Figura 2.1: Exemplo de mapa quantílico (a) e de desvio-padrão (b) para a variável Proficiência em Matemática em 2017

Fonte: Autor

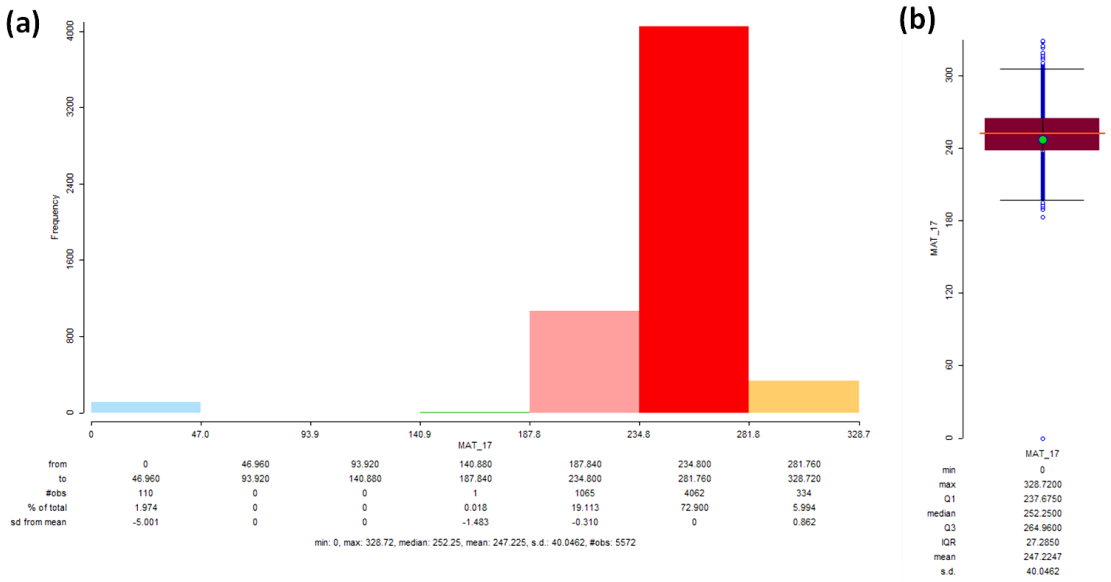


Figura: 2.2: Exemplo de histograma (a) e *boxplot* (b) para a variável Proficiência em Matemática em 2017
Fonte: Autor

Nos mapas do desempenho em matemática (Figura 2.1) é possível notar a concentração, já conhecida, de faixas mais altas nas regiões sul e sudeste, sendo as regiões com menor desempenho norte e nordeste. Esse tipo de conclusão não podemos ter nas ilustrações da Figura: 2.2. Entretanto, para algumas análises no contexto da AEDE deve-se antes definir a matriz de pesos espaciais (matriz $W$), que se discute, preliminarmente, na próxima seção.

## 2.1. Matrizes de ponderação espacial W

Conforme Almeida (2012), uma matriz de ponderação espacial é uma matriz quadrada de tamanho $n$ por $n$. Os pesos espaciais $w_{ij}$ representam a medida da conexão entre as regiões, levando em consideração critérios de proximidade, indicando a influência da região $j$ sobre a região $i$. A matriz de ponderação espacial, denotada por $W$, é útil para ponderar a influência que as regiões exercem umas sobre as outras. O grau de conexão expresso nas matrizes de peso espaciais pode ser classificado com base em critérios geográficos ou socioeconômicos.

Compreender a estrutura da vizinhança é um aspecto crucial na modelagem espacial, pois serve de base para a análise de dependências regionais e a modelagem das interações entre vizinhos (Kopczewska, 2021).

De acordo com Arbia (2014), uma das definições dessa matriz é a seguinte:

$$ {}_nW_n = \begin{bmatrix} w_{11} & \cdots & w_{n1} \\ \vdots & w_{ij} & \vdots \\ w_{1n} & \cdots & w_{nn} \end{bmatrix}, \tag{9} $$

em que cada elemento genérico é definido como

$$ w_{ij} = \begin{cases} 1 & se\ j\ \in N(i) \\ 0 & caso\ contrário \end{cases}, \tag{10} $$

$N(i)$ sendo o conjunto de vizinhos da localização $j$. Por definição temos que $w_{ii} = 0$.

Ainda segundo Almeida (2012), existem diferentes conceitos de vizinhança $N(i)$ que podem ser utilizados, variando desde a mera adjacência entre as unidades territoriais até critérios baseados em uma distância máxima (ou seja, $j \in N(i)$ se $d_{ij} < d_{max}$, onde $d_{ij}$ é a distância entre a localização $i$ e a localização $j$) ou no critério do vizinho mais próximo. Matrizes $W$ mais generalizadas podem ser especificadas, onde os pesos $w$ são funções (negativas) das distâncias geográficas, econômicas ou sociais entre as áreas, em vez de serem simplesmente caracterizados por valores binários, como na equação (10).

Em muitos casos, as matrizes $W$ são padronizadas para que a soma dos pesos em cada linha seja igual a um. Nesse caso, temos:

$$ w_{ij}^{*} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}; \quad w_{ij}^{*} \in W^{*}; \tag{11} $$

Essa padronização pode ser útil em diversas situações. Por exemplo, usando os pesos padronizados, define-se o produto da matriz,

$$ L(y) = W^{*}y, \tag{12} $$

em que cada elemento é igual a:

$$L(y_i) = \sum_{j=1}^{n} w_{ij}^{*}\, y_i = \sum_{j=1}^{n} \frac{w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}} = \frac{\sum_{j \in N(i)} y_j}{\#N(i)}, \tag{13}$$

com $\#N(i)$ representando a cardinalidade do conjunto $N(i)$. O termo da equação (13) representa a média da variável $y$ observada em todas as localidades vizinhas à localidade $i$, de acordo com o critério definido pela matriz $W$. Esse termo é chamado de valor espacialmente defasado de $y_i$ e é frequentemente denotado por $L(y)$, fazendo uma analogia com o operador de defasagem usado na análise de séries temporais.

Quanto às matrizes $W$ é importante considerar a tipologia da matriz, levando em conta a contiguidade e a distância geográfica. No caso da contiguidade, a vizinhança é modelada com base em uma fronteira comum entre as regiões. Nesse sentido, a matriz de pesos espaciais pode ser construída a partir de uma matriz binária que representa essa vizinhança. A matriz de pesos espaciais binários é criada de acordo com a ideia de contiguidade, em que duas regiões são consideradas vizinhas se compartilham uma fronteira física comum. Nesse caso, é atribuído o valor de 1 na matriz de pesos espaciais às regiões vizinhas, e valor nulo para as regiões não vizinhas, refletindo a ideia de maior interação espacial entre regiões contíguas.

A matriz de continuidade é a primeira matrize de ponderação espacial propostas na literatura, sendo introduzida nos estudos de (Geary, 1954; Moran, 1948). Essa matriz é composta por elementos binários, com valor 1 se as regiões possuírem um limite comum e valor 0 caso contrário. Uma característica importante dessa matriz é que ela é simétrica e quadrada, o que facilita diversos procedimentos de cálculo. Formalmente, pode ser representada da seguinte forma:

$$w_{ij} = \begin{cases} 1 & se\ i\ e\ j\ são\ contíguos \\ 0 & se\ i\ e\ j\ não\ são\ contíguo \end{cases}. \tag{14}$$

Convencionalmente, assume-se que $w_{ii} = 0$, ou seja, uma região não é considerada vizinha de si mesma. Isso implica que a matriz de contiguidade tem a sua diagonal principal composta por valores nulos. Embora esse conceito seja simples, existem várias possibilidades para definir a contiguidade de acordo com diferentes convenções. Isso

significa que diferentes critérios podem ser adotados para determinar se duas regiões são consideradas vizinhas com base na proximidade geográfica ou em outros critérios específicos.

Ainda seguindo Almeida (2012), além do critério de contiguidade, outro critério comumente utilizado na definição dos pesos espaciais é a distância geográfica. Nessa abordagem, regiões que estão geograficamente próximas são consideradas ter uma maior interação espacial. Para facilitar cálculos subsequentes e o processo de modelagem, os elementos binários podem ser convertidos em escalas. Esse procedimento também está relacionado ao uso do operador de defasagem espacial, onde o valor médio das variáveis nas regiões vizinhas é utilizado como um indicador de dependência espacial.

Por exemplo, o valor espacialmente defasado da variável $y$ para o local 1 $(y_1)$ representado como $W_{y_1}$ (para uma matriz padronizada por pesos) é uma média ponderada de observações adjacentes.

No entanto, uma matriz $W$ muito adotada na literatura é a matriz dos $k$ vizinhos mais próximos, $w_{ij}(k)$. Trata-se de uma matriz binária cuja convenção de proximidade é baseada na distância geográfica medida em quilômetros ou milhas. Formalmente:

$$w_{ij}(k) = \begin{cases} 1 & se\ d_{ij} \leq d_i(k) \\ 0 & se\ d_{ij} > d_i(k) \end{cases}, \qquad (15)$$

em que $d_i(k)$ é a distância de corte para a região $i$ especificamente, a fim de que esta região $i$ tenha $k$ vizinhos. De novo, assumindo que $w_{ii}(k) = 0$, por convenção. Mais precisamente, $d_i(k)$ é a menor distância para a região $i$ a fim de que ela possua exatamente $k$ vizinhos. Esta distância de corte varia de região para região, por isso, o subscrito $i$ em $d_i(k)$. Portanto, a expressão (15) indica que a proximidade é baseada num critério de distância de tal sorte que duas regiões são consideradas vizinhas, caso encontrem-se dentro de uma distância de corte necessária para que se tenha o número predeterminados de vizinhos.

A matriz dos k vizinhos mais próximos (KNN) é comumente utilizada para dados pontuais, pois examina diretamente os pontos individuais em vez de se referir a áreas. No entanto, é possível criar uma matriz KNN para dados de área ao determinar primeiro os centróides das regiões, que são os pontos que representam o centro de gravidade das

geometrias espaciais das regiões. Para dados pontuais, a matriz KNN é uma solução analítica natural, embora a escolha do número de vizinhos (k) geralmente seja baseada em modelagem ou experiência aleatória.

No caso de dados de área, a estrutura de vizinhança definida pela matriz KNN depende fortemente da forma e superfície das regiões para as quais os centróides foram atribuídos. Regiões estreitas e longas, por exemplo, podem ter centróides significativamente distantes uns dos outros e, portanto, podem não ser consideradas como vizinhos mais próximos de acordo com o critério de distância. No entanto, essas regiões ainda podem ser consideradas vizinhas de acordo com o critério de contiguidade, levando em conta os limites compartilhados. É importante considerar essas diferenças ao escolher a abordagem de matriz de vizinhos mais próximos, levando em conta a natureza dos dados e a estrutura espacial das regiões envolvidas.

### 2.1.1. Seleção de matrizes

A escolha da matriz de pesos espaciais determina os resultados da análise. Normalmente, a matriz *W* é determinada exogenamente, o que pode causar problemas de especificação (Florax & Rey, 1995). Mais comumente usadas são as matrizes de vizinhança definidas de acordo com o critério de contiguidade. Igualmente comuns são as matrizes que usam os vizinhos mais próximos (5 ou 10), vizinhos em um determinado raio e distância inversa.

A cada vez, no entanto, o pesquisador deve garantir que a matriz selecionada seja apropriada para o problema que está sendo analisado. Ainda há discussão na literatura sobre se a matriz de pesos espaciais *W* deve ser assumida *a priori* ou estimada com base em dados. Os proponentes da matriz *W* explicam sua abordagem como um teste de hipóteses sobre a extensão das interações espaciais entre os indivíduos. Isso também justifica as interações teóricas do modelo – por exemplo, apenas locais.

Os defensores da matriz *W* baseada em cálculo mostram um viés de coeficientes espaciais na situação de uma matriz *W* de pesos espaciais especificada incorretamente (Corrado & Fingleton, 2012). A abordagem da matriz de pesos espaciais estimados *W* está mais próxima de filtrar a relação espacial. A abordagem de usar a matriz *a priori W* está mais próxima de testar hipóteses sobre relações espaciais. Ter conhecimento prévio sobre a forma dos pesos é uma boa base para a escolha da matriz (Almeida, 2012).

Outros pesquisadores sugerem observar os dados do painel e rastrear suas relações para encontrar uma melhor correspondência de peso (Bhattacharjee & Jensen-Butler, 2013). A matriz apropriada também pode ser selecionada usando otimização – selecionando a melhor matriz de um conjunto predeterminado de potenciais candidatos. Isso é feito maximizando a estatística I de Moran (Kooijman, 1976), comparando estatísticas de dependência espacial ou critérios de informação, ou seja, Critério de Informação de Akaike (AIC) ou Critério de Informação Bayesiano (BIC), como em Golgher (2015). Baumont et al. (2004) propõem o seguinte procedimento em três passos:

1. Roda-se o modelo clássico de regressão linear por Mínimos Quadrados Ordinários (OLS);
2. Testam-se os resíduos para autocorrelação espacial por intermédio do valor da estatística de I de Moran, usando um conjunto de matrizes;
3. Define-se a matriz que tenha gerado o maior valor do I de Moran, significativo estatisticamente.

## 2.2. Autocorrelação espacial

Segundo Sabater et al. (2011), o primeiro aspecto para o estudo de AEDE é testar a hipótese de que os dados espaciais estejam distribuídos aleatoriamente, ou seja, se os valores de um atributo numa região não dependem dos valores destes atributos nas regiões vizinhas.

Importante considerar que um coeficiente de autocorrelação descreve um conjunto de dados que está ordenado numa certa sequência. Desse modo, um coeficiente de autocorrelação espacial descreve um conjunto de dados que está ordenado segundo uma sequência espacial. Qualquer coeficiente de autocorrelação é, nesse sentido, construído pela razão de uma medida de autocovariância e uma medida de variação total dos dados.

## 2.3. Estatística I de Moran

Usando a medida de autocovariância na forma de produto cruzado, Moran (1948) propôs a elaboração de um coeficiente de autocorrelação espacial, algebricamente demonstrado por:

$$I = \frac{n}{S_0} \frac{\sum_i \sum_i w_{ij} z_i z_j}{\sum_{i=2}^{n} z_i^2}, \tag{16}$$

ou matricialmente:

$$I = \frac{n}{S_0} \frac{z'Wz}{z'z}, \tag{17}$$

em que $n$ é o número de regiões, $z$ denota os valores da variável de interesse padronizada, $Wz$ representa os valores médios da variável de interesse padronizada nos vizinhos, definido segundo a matriz de ponderação espacial *W*. Um elemento dessa matriz, referente à região $i$ e a região $j$, é registrado como $w_{ij}$. $S_0$ é igual a operação $\sum \sum w_{ij}$, significando que todos os elementos da matriz de pesos espaciais *W* devem ser somados.

A estatística de I de Moran, nesse sentido, apresenta uma relação de autocovariância do tipo produto cruzado pela variância dos dados ($z'z$). A matriz de pesos espaciais, nesse caso, foi normalizada na linha, o termo $S_0$, ou seja, o duplo somatório no denominador da expressão ($S_0$) resulta em $n$. Desse modo, pode-se reescrever a equação (17) como:

$$I = \frac{z'Wz}{z'z}. \tag{18}$$

A hipótese nula sendo testada é a de aleatoriedade espacial. Conforme demonstrado por Cliff & Ord (1981), o I de Moran tem um valor esperado de $-\left[\frac{1}{(n-1)}\right]$, isto é, o valor que seria obtido se não houvesse padrão espacial nos dados. O valor calculado de *I* deve ser igual ao seu valor esperado, dentro dos limites da significância estatística, se $y_i$ for independente dos valores das regiões vizinhas. Valores de *I* que excedem o valor esperado indicam autocorrelação espacial positiva. Valores de *I* abaixo do valor esperado sinalizam uma autocorrelação negativa (Almeida, 2012).

Percebe-se que, ao contrário de um coeficiente de autocorrelação ordinário, a estatística não é centrada em zero. No entanto, à medida que o número de regiões aumenta, o valor esperado da estatística I de Moran aproxima-se de zero. Dessa forma, a estatística *I*

assemelha-se a um coeficiente de autocorrelação, porém, não é idêntico a ele, já que a média teórica (valor esperado) não é exatamente zero.

Quanto à interpretação das informações, de acordo com Fotheringham et al. (2003), se altos valores de um atributo tendem a ser agrupar juntos em certas partes de uma área de estudo e baixos valores tendem a se agrupar em outras, diz-se que o atributo exibe a autocorrelação espacial positiva (p. 103). Ou seja, a autocorrelação espacial positiva indica que, no geral, altos valores de uma variável de interesse ($y$) tendem a estar circundados por altos valores desta variável em regiões vizinhas $(W_y)$ e/ou baixos valores de $y$ tendem a estar rodeados por baixo valores também para $y$ em regiões vizinhas $(W_y)$. Esse é o padrão sistemático de distribuição dos valores da variável quando há um efeito de contágio ou efeito de transbordamento de um fenômeno em estudo. Nesse caso, a chance de se ter numa região vizinha um valor parecido com o que se tem numa determinada região é alta.

Por outro lado, uma indicação de autocorrelação espacial negativa revela que existe uma dissimilaridade entre os valores do atributo estudado e da localização espacial do atributo: se altos valores tendem a ser encontrados muito próximos a baixos valores e vice-versa, diz-se que o atributo exibe autocorrelação espacial negativa (Fotheringham et al., 2002, p. 103). A autocorrelação espacial negativa, nesse caso, indica que um alto valor da variável de interesse de uma região tende a estar rodeado por baixos valores desta mesma variável nas regiões vizinhas e/ou um baixo valor da variável de interesse da região tende a estar rodeado por alto valores desta variável de interesse em regiões vizinhas.

## 2.4. Indicador Local de Associação Espacial (LISA)

Uma abordagem alternativa para visualizar a autocorrelação espacial é baseada no diagrama de dispersão de Moran, que mostra a defasagem espacial da variável de interesse no eixo vertical e o valor da variável de interesse no eixo horizontal. Convém observar que a variável de interesse $(y)$ para sua defasagem espacial $(W_y)$ é padronizada de tal modo que tem a média zero e variância unitária, quando apresentada no diagrama, transformando-se em $z$ e $W_z$. O diagrama de dispersão de Moran é, na verdade, o gráfico da dispersão da nuvem de pontos representando as regiões, com a indicação da declividade da reta de regressão.

Para conseguir a declividade da reta, estima-se uma regressão linear simples por OLS, especificada como:

$$W_z = \alpha + \beta_z + \varepsilon, \tag{19}$$

em que $\alpha$ é a constante da regressão, $\beta$ é o coeficiente angular e $\varepsilon$ é um termo de erro aleatório.

Dessa forma, o coeficiente I de Moran pode ser interpretado como o coeficiente angular da reta de regressão (19) da defasagem espacial $(W_z)$ contra a variável de interesse $(z)$, estimado por OLS e representado pela linha de regressão

$$\hat{\beta} = I = \frac{z'Wz}{z'z}. \tag{20}$$

Percebe-se, pela equação (20), que o coeficiente β estimado é equivalente à fórmula do I de Moran em (18). Se o coeficiente angular da reta de regressão é positivo, há evidências de que a autocorrelação espacial é positiva. Se o coeficiente angular for negativo, existem evidências de que a autocorrelação espacial é negativa.

Do ponto de vista estatístico, a análise global da dependência espacial pode distorcer os resultados em nível local e esconder algumas particularidades presentes em determinadas localidades do conjunto geográfico considerado. Assim, as análises relacionadas com o território, normalmente, estão mais direcionadas para a identificação do comportamento local, bem como das características próprias de cada espaço analisado. Desse modo, o LISA, que nada mais é do que o I de Moran local, torna-se mais apropriado para verificar a autocorrelação espacial local (Sabater et al., 2011).

O LISA fornece o grau de autocorrelação espacial, estatisticamente significativo, em cada unidade regional. Combinando as informações do diagrama de Moran com o mapa LISA de significância, tem-se o mapa de *cluster*, que permite uma visualização geográfica mais adequada do grau de concentração das variáveis estudadas.

Quanto à tendência dos dados de agruparem no espaço o diagrama de dispersão de Moran, pode-se identificar quatro padrões de associação linear: Alto-Alto, Baixo-Baixo, Alto-Baixo e Baixo-Alto, conforme representado na figura a seguir:

- Alto-Alto (AA) – *High-High* (HH)
- Baixo-Baixo (BB) – *Low-Low* (LL)
- Alto-Baixo (AB) – *High-Low* (HL)
- Baixo-Alto (BA) – *Low-High* (LH)

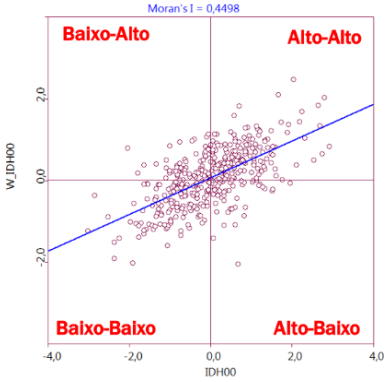


Figura: 2.3: Exemplo de diagrama de dispersão de Moran

Fonte: Adaptado de Vieira (2009)

Para detectar tais padrões locais, usam-se estatísticas tais como o $G_i$ de Getis-Ord local, o $O_i$ de Ord-Getis ou, principalmente, o coeficiente $I_i$ de Moran local (LISA). O mapa de *clusters* fornece os agrupamentos de dados na forma de associações Alto-Alto, Baixo-Baixo, Alto-Baixo, Baixo-Alto, estatisticamente significativos. O mapa de *clusters* é resultante da combinação da informação de dois outros mapas: o mapa de dispersão de Moran com o mapa de significância de LISA. Esse instrumento pode ser usado para uma análise tanto num contexto univariado quanto bivariado.

Os *outliers* espaciais causam efeitos danosos sobre os resultados da autocorrelação espacial. Com o uso do diagrama de dispersão de Moran é possível detectar se existem *outliers* espaciais, bem como pontos de alavancagem e qual a influência sobre o valor da estatística I. Outro recurso para a detecção de *outliers* espaciais é o mapa de *clusters*. *Outliers* espaciais bivariados também podem ser detectados por meio dos recursos do diagrama de dispersão de Moran bivariado e o mapa de cluster bivariado. A heterogeneidade espacial pode ser antecipada com as técnicas de AEDE em busca de regimes espaciais, instabilidade estrutural ou a detecção de *outliers* nos resíduos de uma regressão.

## 3. Modelos de Dependência Espacial

Do ponto de vista metodológico, a principal característica dos modelos espaciais reside na incorporação dos chamados efeitos espaciais na regressão: a dependência espacial, a heterogeneidade espacial e a imbricação dos efeitos espaciais (Almeida, 2012). No entanto, como aponta Anselin (1988), na maioria das vezes, os problemas gerados pela heterogeneidade espacial podem ser corrigidos com o uso de instrumentos fornecidos pela economia clássica (Tyszler, 2006; Vieira, 2009). A heterogeneidade espacial diz respeito à falta de estabilidade de comportamento ao longo do espaço, como, por exemplo, cidades ricas e pobres aglomeradas em diferentes regiões de um país. Em termos de modelagem econométrica, segundo Elhorst (2014), a heterogeneidade espacial significa que os parâmetros não são homogêneos ao longo do conjunto de dados, variando com a unidade; assim, se o objetivo for inferência sobre os parâmetros, podemos proceder a correção dos erros-padrão como na econometria clássica. Assim, na presente dissertação o foco será nos modelos de dependência espacial, pois a heterogeneidade espacial, se necessário, trataremos com os procedimentos usuais de correção de erro, tal como HAC e método de Kelejian e Prucha, como comentado adiante.

Para Elhorst (2014 p.14), uma estratégia comum em grande parte das análises espaciais reside em começar com um modelo de regressão linear não espacial, estimador por OLS, e depois testar se esse modelo precisa ou não contemplar a dependência espacial. O modelo básico assume a forma:

$$y = X\beta + \varepsilon, \tag{21}$$

onde *y* indica um vetor $N \times 1$ que consiste em uma observação na variável dependente para cada unidade na amostra $(i = 1, \dots, n)$, $X$ denota uma matriz $N \times K$ de variáveis explicativas exógenas com um vetor $N \times 1$ associado ao parâmetro de termo constante a ser estimado, $\beta$ é um vetor $K \times 1$ associado com parâmetros desconhecidos a serem estimados, e $\varepsilon$ é um vetor de termos de perturbação, onde $\varepsilon_i$ é assumido ser distribuído de forma independente e idêntica para todo $i$ com média zero e variância $\sigma^2$. Uma vez que o modelo de regressão linear é comumente estimado por Mínimos Quadrados Ordinários (OLS), muitas vezes é rotulado de modelo OLS.

Outra abordagem é começar com um modelo mais geral, contendo um conjunto de modelos mais simples que idealmente deveriam representar as hipóteses econômicas alternativas. De maneira geral, alguns efeitos de interação podem explicar por que uma observação associada a um local específico pode depender de observações em outros locais. Assim, os efeitos de interação entre os termos de erro não requerem um modelo teórico para um processo de interação espacial, mas são consistentes com uma situação em que os determinantes da variável dependente omitida do modelo são autocorrelacionados espacialmente. Um modelo completo com todos os tipos de efeitos de dependência assume a forma:

$$y = \rho W y + X\beta + WX\tau + \xi,$$
$$\xi = \lambda W\xi + \varepsilon;$$

ou

$$\xi = \gamma W\varepsilon + \varepsilon, \tag{22}$$

onde $Wy$ denota os efeitos de interação endógena entre a variável dependente, $WX$ os efeitos de interação exógena entre as variáveis independentes e ($W\xi$ ou $W\varepsilon$) os efeitos de interação entre o termo de perturbação das diferentes unidades. Esse modelo refere-se ao Modelo Geral de Aninhamento Espacial (GNS) (Elhorst, 2014), uma vez que inclui todos os tipos de efeitos de interação, $\rho$ é chamado de coeficiente autorregressivo espacial, $\lambda$ é o coeficiente de autocorrelação espacial, $\gamma$ é o coeficiente de média móvel espacial, enquanto $\tau$, assim como $\beta$, representa um vetor $K \times 1$ de parâmetros fixos, mas desconhecidos a serem estimados. $W$ é uma matriz $N \times N$ não negativa que descreve a configuração espacial ou arranjo das unidades na amostra.

A Figura 3.1 representa essas relações e, em destaque, os modelos que serão foco no presente trabalho, uma vez que consideraremos para as análises espaciais apenas modelos que levem em consideração a variável dependente espacialmente defasada ($\rho Wy$) e/ou o erro espacialmente defasado ($\lambda W\xi$ ou $\gamma W\varepsilon$), isso porque, estima-se, e as evidências (intuição) sugerem (Vernier, 2016), não haver influência espacial (*spillover*) das variáveis independentes consideradas, principalmente aquelas relacionadas com o esforço e a percepção docente, condições extraescolares e do aluno, sobre a proficiência em Língua

Portuguesa ou Matemática do município. Em nosso entendimento seria muito forte a prerrogativa que as opiniões dos docentes sobre as escolas e alunos num município poderia influenciar a proficiência em Matemática ou Língua Portuguesa dos alunos em outro município. Tanto é verdade que nem as variáveis independentes em nível (sem ser espacialmente defasada) teve influência (apenas marginal) sobre as variáveis dependentes, como veremos na análise dos resultados.

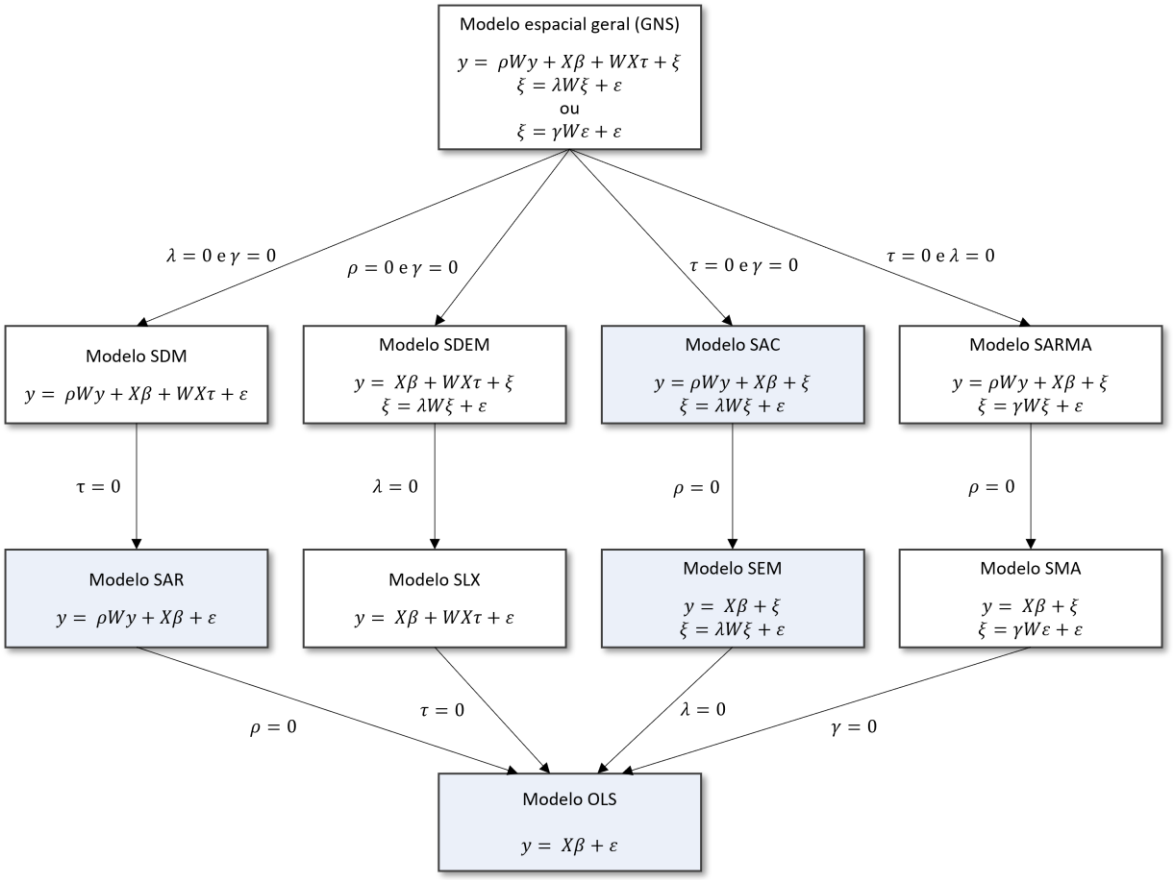


Figura 3.1: Tipologia dos modelos de dependência espacial

Fonte: Adaptado de Almeida (2012)

## 3.1.Taxonomia dos modelos de dependência espacial linear

A Figura 3.1 resume uma família de dez modelos econométricos espaciais lineares, entre os quais o modelo OLS na parte inferior e o modelo espacial geral (GNS) na parte superior. Cada modelo abaixo do modelo espacial geral pode ser obtido a partir desse modelo impondo restrições a um ou mais de seus parâmetros. As restrições são apresentadas ao lado das setas na Figura 3.1 (Elhorst, 2014).

É possível observar que existem modelos econométricos espaciais que são pouco considerados ou utilizados em pesquisas teórico-econométricas e empíricas. O Modelo de Erro Espacial de Durbin (SDEM), que contém efeitos de interação exógenos e efeitos de interação entre os termos de erro, é o melhor exemplo. A esse respeito, deve-se destacar que há uma grande lacuna no nível de interesse em diferentes tipos de efeitos de interação entre teóricos e empíricos. Os empíricos estão interessados, principalmente, nos modelos SAR e SEM, bem como no modelo SAC que combina efeitos de interação endógenos e efeitos de interação entre os termos de erro, e os teóricos, nos outros modelos devido aos problemas econométricos que acompanham sua estimação.

A razão pela qual alguns teóricos não se concentram em modelos econométricos espaciais com efeitos de interação exógena é porque a estimativa de tais modelos não apresenta problemas econométricos; técnicas de estimativa padrão são suficientes nessas circunstâncias. Consequentemente, o modelo SLX geralmente não faz parte da caixa de ferramentas de pesquisadores interessados na teoria econométrica de modelos espaciais. Os modelos mais utilizados na literatura são o SAC, SAR e SEM (Tyszler, 2006), considerados modelos de alcance global, por hospedarem a dependência espacial cujo alcance do transbordamento é refletido para todas as regiões da área de estudo. Estes serão os modelos utilizados na análise empírica dessa dissertação.

## 3.2. Modelo Espacial Autoregressivo (SAR)

De acordo com Carvalho & Albuquerque (2011), o modelo SAR é um dos modelos mais comumente utilizados para modelagem de correlação espacial. A lógica parte dos modelos de séries temporais, por meio da incorporação de um termo de *lag* espacial entre os regressores da equação. Na sua forma mais completa o modelo SAR pode ser estendido para incorporar variáveis exógenas na variável independente e tem a seguinte expressão:

$$y = \rho W y + X\beta + \varepsilon, \tag{23}$$

onde $y$ é um vetor coluna, contendo $n$ observações na amostra para a variável resposta $y_i$, o coeficiente escalar $\rho$ corresponde ao parâmetro autorregressivo, a matriz $X$ é uma matriz contendo as observações das variáveis exógenas de dimensão $n\ x\ k$, sendo $k$ o número de regressores, o vetor $\beta$ é um vetor coluna de coeficientes para as variáveis exógenas que possui dimensão $k\ x\ 1$ e $\varepsilon$ corresponde a um vetor coluna contendo os resíduos $\varepsilon_i$ da

equação. A matriz $W$ com dimensão *n x n* refere-se a algum tipo de matriz de vizinhança, ou pesos espaciais, discutidos na seção 2.1. Por exemplo, se ela for de contiguidade e normalizada na linha, o vetor $Wy$ corresponde a um vetor de médias simples das observações para a variável $y$ dos vizinhos. Geralmente, considera-se que os resíduos $\varepsilon_i$ são i.i.d. com distribuição normal, média zero e variância homogênea $\sigma^2$.

Se o parâmetro $\rho$ for positivo, mais comum nas pesquisas empíricas (Golgher, 2015), isso indica que existe autocorrelação espacial global positiva, ou seja, um alto (baixo) valor de $y$ nas regiões vizinhas aumenta (diminui) o valor de $y$ na região $i$ (Almeida, 2012, p.153). O valor do parâmetro $\rho$ não pode ser estimado por OLS, devido o vetor $Wy$ ser, por hipótese, correlacionado com os resíduos $\varepsilon_i$. Nessa situação as estimativas OLS são inconsistentes (Anselin, 1988; Carvalho & Albuquerque, 2011). Como alternativa pode-se utilizar a estimação via Máxima Verossimilhança (ML), que não sofre do problema de inconsistência, tal qual o estimador OLS, devido a endogeneidade do regressor $Wy$ (Carvalho & Albuquerque, 2011; Vieira, 2009), no entanto, as outras hipóteses gerais (resíduos $\varepsilon_i$ são i.i.d. com distribuição normal, média zero e variância homogênea $\sigma^2$) devem permanecer. No contexto do arcabouço ML pode-se recorrer aos testes Wald, o Teste de Razão de Verossimilhança (LR) ou o Teste Multiplicador de Lagrange (LM) e, assim, testar a significância do parâmetro $\rho$ (presença de dependência espacial das observações para a variável $y_i$), assim como discutido por Carvalho e Albuquerque (2011) e Tyszler (2006).

Pode-se também estimar os parâmetros do modelo SAR por métodos das Variáveis Instrumentais (VI) ou, especificamente, alguma derivação dos Métodos dos Momentos Generalizados (GMM), tal como o método de Mínimos Quadrados em Dois Estágios (2SLS). Dada a natureza multidirecional da dependência espacial, a presença do vetor $Wy$ é equivalente à introdução de uma variável endógena num sistema de equações simultâneas (Almeida, 2012, p.196), que comumente é tratada por variáveis instrumentais. Kelijian & Prucha (1998, 1999) propuseram retirar do conjunto de defasagens espaciais das variáveis explicativas exógenas ($WX$ e $W^2X$) os candidatos para instrumentalizar $Wy$, e adaptaram um conjunto de estimadores GMM para o contexto espacial. Esses estimadores servem tanto para o modelo SAR, quanto para os que discutiremos na sequência (SEM e SAC).

Ao contrário da estimação por ML, esses métodos prescindem do requisito da propriedade da normalidade do erro aleatório, constituindo, assim, numa alternativa robusta (Almeida, 2012, p.198). Além do mais, são computacionalmente mais eficientes e por não

necessitar inverter uma matriz $n\ x\ n$, podem incorporar outras variáveis endógenas além do vetor $Wy$. Assim, podemos tratar a presença da heterogeneidade e autocorrelação dos resíduos através de um estimador robusto: HAC (Carvalho e Albuquerque, 2011). Para a utilização desses estimadores robustos deve-se ainda definir uma função *kernel*, como discutidas em Carvalho e Albuquerque (2011) e Tyszler (2016), que busque descrever a variância assintótica robusta à heterocedasticidade e autocorrelação (HAC) nos resíduos em função da distância entre as observações (vizinhos).

A motivação para se ter um modelo SAR é o fato de representar um equilíbrio de longo prazo de um processo dinâmico, denotando decisões tomadas por agentes econômicos em períodos passados influenciando a decisão de agentes no presente (Almeida, 2012, p.156). Quando se admite a existência de transbordamentos espaciais, uma mudança na variável explicativa numa região qualquer afetará não apenas a própria região pelo efeito direto, mas pode afetar o valor da variável dependente em todas as regiões (ou seus vizinhos) por meio de um efeito indireto, via retroalimentação. Assim, o coeficiente $\beta$ no modelo SAR não tem o mesmo significado que no modelo OLS, cujos efeitos marginais são constantes e iguais à derivada parcial, ou seja, o próprio valor $\beta$. No modelo SAR deve-se levar em conta os efeitos diretos, indiretos e totais das variáveis explicativas sobre o $y$, sendo esse cálculo não trivial como num modelo OLS. Almeida (2012), Golgher (2015), Arbia (2014) e Elhorst (2014) discutem a interpretação dos coeficientes do modelo SAR e seus efeitos marginais, de forma a se ter o cálculo dos efeitos diretos, indiretos e indiretos para enriquecimento de uma análise econômico-espacial.

### 3.3. Modelo de Erro Espacial (SEM)

Assim como o modelo SAR parte da lógica dos modelos de séries temporais, por meio da incorporação de um termo AR entre os regressores da equação, o modelo SEM pode ser comparado ao modelo MA de séries temporais, por meio da especificação de um termo de média móveis para as observações no tempo (Carvalho & Albuquerque, 2011). O modelo SEM apresenta a seguinte especificação:

$$\begin{aligned} y &= X\beta + \xi, \\ \xi &= \lambda W\xi + \varepsilon, \end{aligned} \tag{24}$$

onde $y, X, \beta, W$ e $\varepsilon$ foram definidos na equação (23), e o coeficiente $\lambda$ indica a intensidade da autocorrelação espacial entre os resíduos da equação observada. Os resíduos $\xi$ da equação observada possuem uma estrutura autorregressiva: a autocorrelação espacial nos modelos SEM aparece nos termos de erro. Apesar de se costumeiramente aventar que os resíduos $\varepsilon_i$ são i.i.d. com distribuição normal, média zero e variância homogênea $\sigma^2$, e os modelos SEM serem estimados por ML, diferentemente dos modelos SAR, os modelos SEM (especificamente os $\beta$'s) também podem ser consistentemente estimados via OLS.

No entanto, como espera-se que os resíduos nos modelos SEM sejam correlacionados (Arbia, 2014), os estimadores OLS não são eficientes, tornando outros estimadores preferíveis, tais como os discutidos na seção anterior: métodos de VI ou, especificamente, alguma derivação GMM, tal como os métodos 2SLS propostos por Kelejian e Prucha (1998, 1999). Espera-se que parâmetro $|\lambda| < 1$, e no contexto do arcabouço ML pode-se recorrer aos testes Wald, LR ou LM e, assim, inferir sobre a existência de dependência espacial residual, assim como discutido por Carvalho e Albuquerque (2011) e Tyszler (2006).

O racional para se ter um modelo SEM reside no fato que os erros associados com qualquer observação são uma média dos erros nas regiões vizinhas mais um componente de erro aleatório. O modelo SEM informa que a influência sobre o vetor $y$ não é resultado apenas do choque, representado por $\xi$, específico da região, mas também de transbordamentos de choques de regiões mais conectadas ou menos conectadas pela matriz $W$ (Almeida, 2012, p.162). Assim, o efeito total de um choque não é apenas o choque que ocorreu na região $i$, mas também o efeito realimentador proveniente das outras regiões afetadas pelo choque original. Felizmente, nos modelos SEM o coeficiente $\beta$ tem a mesma interpretação que no modelo OLS, cujos efeitos marginais são constantes e iguais a derivada parcial, ou seja, o próprio valor $\beta$.

### 3.4. Combinação Autoregressiva Espacial (SAC)

Os modelos SAC é uma combinação dos modelos SAR e SEM cuja analogia com os modelos de séries temporais remetem aos modelos ARMA. Também são conhecidos como modelos Kelejian-Prucha (Golgher, 2015) ou modelo de defasagem espacial com erro espacial (SARAR) e possuem a seguinte especificação:

$$y = \rho W y + X\beta + \xi,$$
$$\xi = \lambda W \xi + \varepsilon, \tag{25}$$

onde $y$, $\rho$, $W$, $X$, $\beta$, $\lambda$, $\xi$ e $\varepsilon$ foram definidos nas seções anteriores. Nesse modelo a matriz de pesos $W$ na primeira equação pode ser diferente da segunda equação (Almeida, 2012; Carvalho e Albuquerque, 2011), no entanto, geralmente, até mesmo devido ao processo *ad hoc* da sua escolha, utiliza-se a mesma matriz $W$, por isso, nenhum subscrito diferente em $W$.

Se considerarmos que os resíduos $\varepsilon_i$ são i.i.d. com distribuição normal, média zero e variância homogênea $\sigma^2$, podemos utilizar a estimação via ML, assim como nos modelos SAR e SEM (Golgher 2015, Carvalho e Albuquerque, 2011). No entanto, como os modelos SAC são combinações dos modelos SAR e SEM, eles sofrem dos mesmos desafios de estimação dos parâmetros $\rho$, $\beta$ e $\lambda$. Não podemos utilizar OLS devido à falta de consistência em estimar o parâmetro $\rho$ devido o termo SAR, a despeito do termo SEM não trazer tantas complicações para estimar os parâmetros $\beta$'s, e por causa dos dois termos (SAR + SEM), a complexidade sobrepujada no caso de uma grande amostra pode fazer com que os estimadores ML falhem (falta de convergência). Computacionalmente estimar os modelos SAC é mais demandante, porque agora envolve procedimentos numéricos para estimar dois parâmetros simultâneos ($\rho$ e $\lambda$) com intuito de minimizar uma da função particular (Golgher, 2015, p.100).

Como colocado por Golgher (2015, p.100), a estimação dos modelos SAC envolve procedimentos similares aos descritos para os modelos SAR e SEM, pois pode obter a metodologia de estimação dos modelos SAC a partir da metodologia dos modelos SAR e SEM. Tendo isso em mente, os métodos 2SLS propostos por Kelejian e Prucha (1998, 1999) são computacionalmente mais simples em comparação ao ML e, adicionalmente, prescinde da hipótese da normalidade dos resíduos (Almeida, 2012, p.205). Esses estimadores são consistentes e como espera-se que os resíduos também sejam correlacionados, tal como no modelo SEM, esses procedimentos permitem a incorporação de correções para a presença de heterocedasticidade e autocorrelação residual nos termos de erro da regressão estimada (Carvalho e Albuquerque, 2011).

Geralmente, nas pesquisas empíricas, espera-se que o parâmetro $\rho$ seja positivo e o parâmetro $\lambda$ seja negativo: ambos entre zero e um (Arbia, 2014). Nesse caso, os testes Wald,

LR e LM também fazem presentes para testar tais hipóteses. Almeida (2012, p. 164) coloca que existe uma motivação para os modelos SAC porque às vezes o fenômeno a ser modelo pode requerer que a dependência espacial inerente seja mais intricada, manifestada tanto na forma substantiva de uma defasagem espacial da variável dependente quanto na forma de erros autocorrelacionados espacialmente. Como coloca Anselin (1988), um choque na região $i$ afeta todas as outras regiões por intermédio do multiplicador espacial $\rho$, amplificado pelo efeito multiplicador extra proporcionado por $\lambda$, que torna o padrão de interpretação dos coeficientes $\beta$'s mais complexo e não direto como no modelo SEM.

### 3.5. Diagnósticos e especificação de modelos espaciais

Antes de se cogitar utilizar qualquer um dos modelos discutidos anteriormente, deve-se testar a presença da dependência ou heterogeneidade espacial. Isso porque, se não verificada, podemos utilizar um modelo mais simples, tal qual o modelo linear estimado por OLS. Uma forma de se testar o fenômeno em análise reside nos testes já comentados nas seções anteriores no contexto do arcabouço ML: testes Wald, LR ou LM. São testes baseados nas distâncias das estimativas para o modelo irrestrito e as estimativas satisfazendo às restrições impostas pela hipótese nula, sendo utilizado, mais comumente, devido a conveniência computacional, o teste LM, por requerer apenas a estimação do modelo restrito (Tyszler, 2006).

Por exemplo, estima-se o modelo OLS (modelo restrito), e assim, a partir de seus resíduos, testa-se a significância dos parâmetros $\rho$, $\lambda$ ou $\rho + \lambda$, como discutido por Carvalho e Albuquerque (2011) e Tyszler (2006). Arbia (2014), Almeida (2012) e Tyszler (2006) discutem versões robustas do teste LM nas inferências sobre $\rho$ e $\lambda$, e Golgher (2015) apresenta uma estratégia para seleção de modelos baseada tanto na versão normal quanto na robusta do teste LM, ressaltando assim a importância das duas versões do teste LM para inferências sobre os parâmetros $\rho$ e $\lambda$. Isso porque, apesar da correção efetuada para má especificação do modelo no teste robusto (correção da não centralidade da distribuição qui-quadrado), o teste robusto LM para detectar $\lambda$ é menos poderoso do que o teste LM normal, quando $\lambda$ está presente.

No entanto, os testes Wald, LR ou LM são dependentes da especificação paramétrica para a forma de autocorrelação no espaço, ou seja, a partir do modelo *a priori*, SAR, SEM ou SAC, infere-se sobre a forma de dependência espacial observada nos dados.

Anselin (1988) denomina-os como testes focados, em detrimento de testes difusos, em que nenhuma indicação é fornecida no sentido de se detectar o tipo de autocorrelação espacial predominante na regressão, pois não são baseados numa especificação explícita sobre o processo gerador dos dados (Almeida, 2012, p.216). Os testes focados têm sofrido diversas críticas, mas os testes difusos não sofrem o mesmo ataque, e são relativamente bem aceitos na literatura (Carvalho e Albuquerque, 2011). Nesses últimos se enquadra a estatística I de Moran sobre os resíduos da regressão OLS, conforme já discutido na seção 2.3, e o teste de Kelejian-Robinson (KR), conforme discutido por Almeida (2012) e Carvalho e Albuquerque (2011). Diferentemente do teste I de Moran, o teste KR não pressupõe normalidade da variável $y_i$ ou dos resíduos $\xi_i$ da regressão.

Adicionalmente, quando se examinam os principais manuais de econometria espacial (Golgher, 2015; Arbia, 2014; Elhorst, 2014; Almeida, 2012) não se vê o detalhamento formal de outros testes diagnósticos importantes para modelos espaciais. Talvez porque esses manuais estão preocupados em formalizar as ferramentas construídas especificamente para a análise espacial. No entanto, alguns testes diagnósticos conhecidos da econometria clássica também devem ser aplicados no contexto dos modelos espaciais.

Por exemplo, para se testar a normalidade geralmente utiliza-se o teste Jarque-Bera também nos modelos espaciais. Para a heterocedasticidade da mesma forma, o teste Koenker-Basset se encontra disponível em alguns *softwares* de análise espacial (Anselin & Rey, 2014; Bivand et al., 2021) para avaliar a presença da heterogeneidade nos resíduos $u_i$ da regressão OLS. A multicolinearidade também se avalia nos modelos espaciais pelos métodos clássicos (Fator de Inflação da Variância – VIF), e torna-se particularmente relevante em modelos em que há variáveis exógenas defasadas espacialmente (Golgher, 2015; Almeida, 2012). Algumas medidas de ajuste comumente utilizadas em outras áreas da econometria também podem ser usadas, tais como os coeficientes de determinação ($R^2$) ou *pseudo* $R^2$, e os critérios e informação (AIC e BIC). Estes últimos, inclusive, são importantes na escolha do modelo que melhor explica o processo estocástico gerador dos dados. Quando os estimadores utilizados recaem no arcabouço GMM, os indicadores de ajuste se resumem ao *pseudo* $R^2$, pois boa parte das inferências e medidas de distância conhecidas requer uma função de verossimilhança.

Nessa esteira, o último capítulo do clássico manual de Anselin (1988) é dedicado ao processo de "incerteza da especificação" (Almeida, 2012, p.214) e escolha de modelos

espaciais, que em certa medida, segundo Anselin (1988), incorre em heurísticas filosóficas, principalmente, se acredita-se num processo científico indutivo ou dedutivo (Golgher, 2015). Na égide desse último processo deve-se partir, sempre, de considerações teóricas acerca do fenômeno sob estudo. Segundo Almeida (2012, p.214), na definição das defasagens espaciais a serem colocadas no modelo, a teoria ou evidências empíricas anteriores podem desempenhar papel preponderante se houver a presença de interação espacial no fenômeno sob investigação.

Entretanto, frequentemente os cientistas sociais se colocam num processo indutivo: observa e supõe por meio de dados para se chegar a uma conclusão. Na engrenagem desse processo, Golgher (2015), Almeida (2012) e Anselin (1988) discutem duas estratégias principais de especificação de modelos espaciais: geral-específica e específica-geral. A primeira é um procedimento baseado no princípio da metodologia de Hendry copiado da área de séries temporais. A partir dele estima-se um modelo de fator comum espacial, em que a especificação de todas as defasagens espaciais (inclusive das variáveis exógenas) são incluídas no lado direito da regressão (modelo geral), e testam-se as restrições aos parâmetros da equação para se chegar num modelo mais específico.

Por exemplo, Tyszler (2006) propugna, e valida através de simulação de Monte Carlo, uma busca alinhada com essa estratégia, em que a modelo espacial se inicie com o modelo SAC, sem necessidade de se adotar testes de autocorrelação espacial do tipo LM, LR ou Wald para auxiliar a especificação, mas apenas analisando a significância estatística dos parâmetros espaciais ($\lambda$ e $\rho$): i) se apenas $\rho$ for significativo especifica o modelo SAR; ii) se apenas $\lambda$ for significativo especifica o modelo SEM; iii) se ambos são significativos estima-se o modelo SAR e analisa-se a significância de $\rho$; se $\rho$ for diferente de zero, especifica o modelo SAC; caso $\rho$ não seja significativo, especifica-se o modelo SEM (Almeida, 2012, p.236).

Golgher (2015, p.120) coloca que as estratégias geral-específica e específica-geral apresentam resultados similares quanto à qualidade dos ajustes empíricos, mas orienta utilizar a estratégia específica-geral, pois ela é de certa forma superior à primeira em alguns aspectos. Nesse sentido, Golgher (2015, p. 145) e Almeida (2012, p.234) colocam que, em estudos de simulações, quando comparado com o procedimento de especificação baseado na metodologia de Hendry, são encontradas evidências empíricas de que o procedimento híbrido de especificação domina com respeito à detecção de dependência espacial. O

procedimento híbrido de especificação nada mais é do que o procedimento clássico (estratégia específico-geral), no entanto, substituindo os testes LM tradicionais pelas suas versões robustas. Os testes LM robustos são não enviesados, porém menos poderosos que os testes tradicionais, no entanto, segundo Anselin e Rey (2014), os ganhos em robustez dos testes robustos em muito compensam essa perda de poder. Assim, conforme Anselin (2005):

1. Estima-se o modelo básico OLS;
2. Testa-se a hipótese de ausência de autocorrelação espacial residual e na variável dependente por meio dos testes LM tradicionais;
3. Caso ambos não sejam significativos, considere o modelo OLS. Caso contrário, siga para o próximo passo;
4. Se uma das defasagens espaciais for significativa, estima-se o modelo apropriado. Por exemplo, se o parâmetro $\rho$ for significativo, estime o modelo SAR;
5. Caso ambas as defasagens são significativas, estima-se o modelo apontado como o mais significativo pelas versões robustas dos testes LM. Por exemplo, se o teste LMρ robusto for significativo, escolhe-se o modelo SAR como o mais adequado. Caso o teste LMλ robusto seja o significativo, adota-se o modelo SEM como o mais adequado.

A Figura 3.2 apresenta esquematicamente esse procedimento híbrido de especificação do modelo espacial (Almeida, 2012, p.233), que pode ser adaptado conforme a Figura 3.3 (Golgher, 2015, p.140) para englobar o modelo SAC, tendo em vista a proposta de Tyzler (2006) e que nas análises adiante não se consideram as variáveis exógenas espacialmente defasadas ($WX\tau$), pelos motivos discutidos no começo do capítulo.

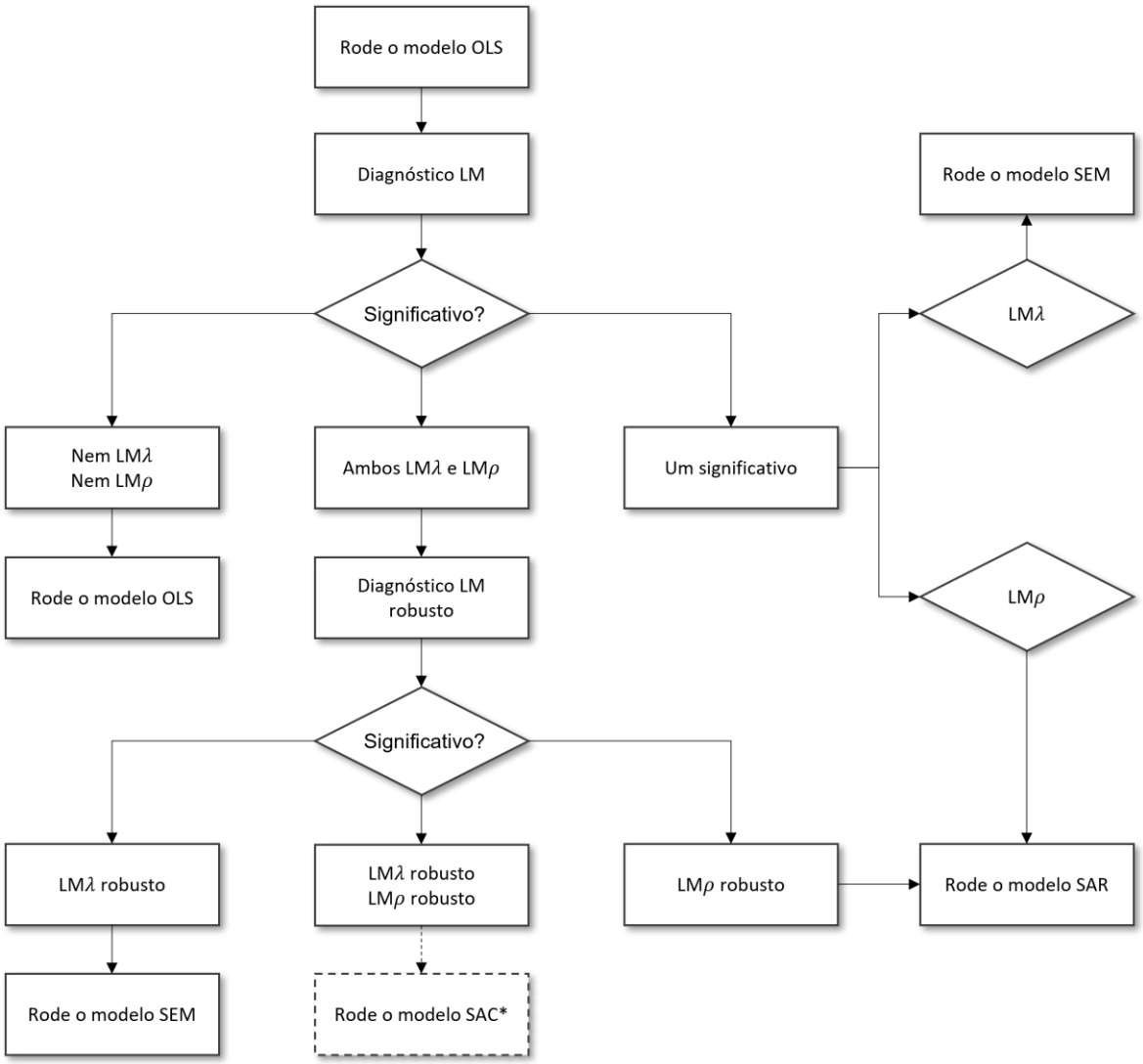


Figura 3.2: Procedimento híbrido de especificação de modelos espaciais

Fonte: Adaptado de Anselin e Rey (2014)

*Nota:* (*) A definição da escolha do modelo SAC ocorre mediante a análise cautelosa do resultado do teste SARMA.

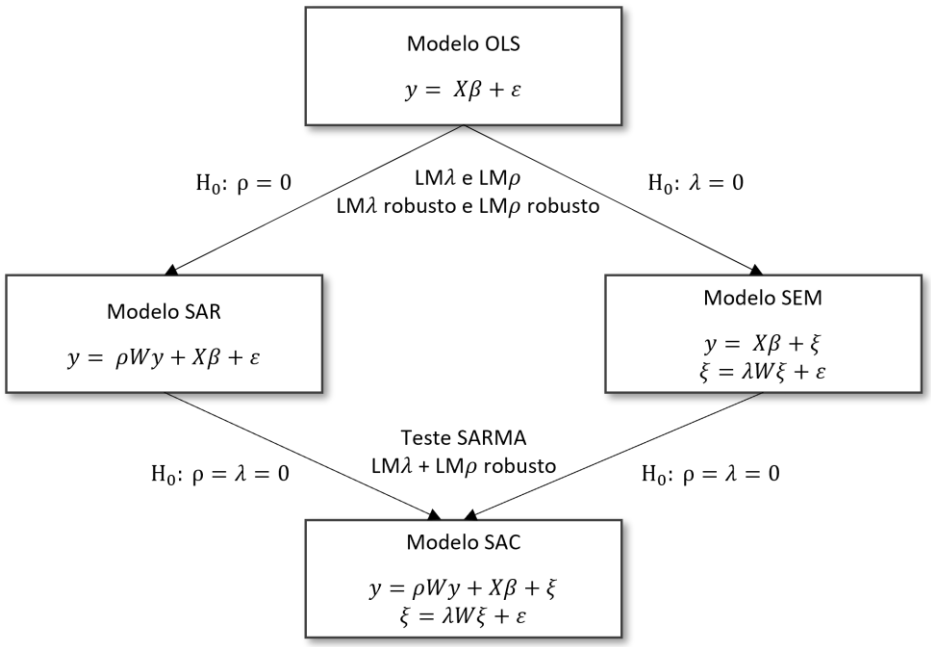


Figura 3.3: Estratégia específica-geral para comparar os modelos SAR, SEM e SAC.

Fonte: Adaptado de Golgher (2015).

*Nota:* O teste SARMA será significativo se qualquer um dos testes LM forem altamente significativos, assim, os resultados provenientes do teste SARMA devem ser avaliados com cautela, pois modelos de ordem superior (dois componentes espaciais defasados) só fazem sentido se há evidências de má especificação, ou seja, quando não se conseguiu eliminar completamente a autocorrelação espacial dos modelos SAR ou SEM (Anselin e Rey, 2014).

## PARTE II – APLICAÇÃO DO MÉTODO E ANÁLISE DOS RESULTADOS

Esta parte tem por objetivo apresentar as análises acerca da percepção e expectativa do docente da educação básica quanto ao desempenho escolar no nível agregado dos munícipios brasileiros, seguindo a abordagem da econometria espacial. Dentro da abordagem da econometria espacial utilizou-se a Análise Exploratória de Dados Espaciais (AEDE) e Modelos Espaciais, seguindo as orientações de Almeida (2012), Anselin (2005), Anselin & Rey (2014) e Golgher (2015), com utilização dos softwares GeoDa e GeoDaSpace (Anselin et al., 2006).

Seguimos a abordagem da econometria espacial neste trabalho, pois a literatura que relaciona desempenho escolar no nível agregado apresenta indícios de interação espacial em algumas pesquisas empíricas (Cavalcanti et al., 2020; Fujita et al., 2021; Vernier, 2016). Nesse sentido, discutiremos alguns estudos que tratam da relação entre o rendimento escolar, expressos pelos resultados dos testes educacionais, as características socioeconômicas dos alunos e os fatores escolares como as expectativas e percepções docentes, entre escolas de municípios vizinhos.

### 4. Evidências Empíricas

Meyer (1997) informa que os resultados dos testes educacionais podem estar contaminados pela mobilidade dos alunos entre escolas e por fatores extraescolares. Nesse sentido, Thieme et al. (2016) abordam variáveis contextuais e socioeconômicas que estão além do controle da escola e concluem, como Wodtke et al. (2011), que discentes que vivem em bairros de baixa renda, caracterizados pela alta pobreza, desemprego e recebimento de benefício social, famílias numerosas, chefiadas por mulheres e poucos adultos com nível escolar médio ou superior, durante todo o curso de vida da infância, têm um impacto importante sobre as chances de formação no ensino médio. Assim, em consonância, Tekwe et al. (2004) questionam se as escolas podem ser responsabilizadas pelos efeitos significativos de fatores sociodemográficos nos resultados obtidos pelos seus alunos em testes de avaliação.

No Brasil, foram poucos os estudos que procuraram avaliar a relação entre o desempenho escolar e as características socioeconômicas. Merecem destaque as pesquisas de Torres et al. (2003), na Região Metropolitana de São Paulo, e o de Cunha et al. (2009),

que abordaram a relação entre os resultados obtidos pelo Sistema de Avaliação de Rendimento Escolar do Estado de São Paulo (SARESP) e as condições socioeconômicas, de moradia e as estruturas das escolas de Campinas.

Como já ressaltado, uma das constatações da literatura de Psicologia e Educação (Alvidrez & Weinstein, 1999; Palardy, 1969; Rosenthal & Jacobson, 1968; Teixeira, 2020; Vidal et al., 2019; Xavier & Oliveira, 2020) é que as chances de sucesso no desempenho escolar também estão diretamente relacionadas com as expectativas docentes sobre os alunos, que segundo Vidal et al. (2019) trata-se de um tipo de "profecia autorrealizadora".

Por exemplo, Teixeira (2020) e Xavier & Oliveira (2020) constataram essa relação para escolares do ensino fundamental no Brasil. De acordo com os autores, as expectativas docentes possuem efeito diretos (grande influência) sobre o desempenho dos alunos, contrabalanceando os efeitos, muitas vezes perversos, da estrutura familiar que joga contra a qualidade da aprendizagem. Vidal et al. (2019) colocam que os problemas de aprendizagem dos alunos podem estar associados com o meio social, o nível cultural e a falta de assistência dos pais na vida escolar. Adicionalmente, conforme Xavier & Oliveira (2020), fatores de composição das turmas podem afetar a formação das expectativas.

O presente estudo tem a intenção de avançar nessa temática ao propor uma abordagem estatística de avaliar essa relação, que até o presente momento tem seguido a literatura de desempenho escolar, com a utilização de microdados e avaliação da aprendizagem ao nível individual (Jesus & Laros, 2004; Laros & Marciano, 2010; Teixeira, 2020; Vidal et al., 2019; Xavier & Oliveira, 2020). Assim, pretende-se relacionar a expectativa docente com o desempenho escolar no nível agregado dos munícipios brasileiros, seguindo uma abordagem espacial, cuja literatura encontrou evidências favoráveis a esse tipo de modelagem (Cavalcanti et al., 2020; Fujita et al., 2021; Vernier, 2016).

Por exemplo, mais recentemente Fujita et al. (2021) encontraram fortes evidências a favor de transbordamentos educacionais da presença de educação superior e qualidade das universidades nos municípios sobre o desempenho de alunos do ensino fundamental. Os resultados dos autores indicaram forte dependência espacial, sugerindo que a dimensão espacial influencia o desempenho escolar.

Adicionalmente, Cavalcanti et al. (2020) observaram predominância de *clusters* espaciais (dentro dos 5570 munícipios brasileiros), indicando a presença de autocorrelação

espacial positiva no desempenho escolar (proficiência média em português e matemática) para alunos do 9º ano do ensino fundamental. De acordo com os autores, o desempenho escolar de um munícipio está condicionado ao i) desempenho escolar anterior do próprio município; ii) dos municípios vizinhos; e iii) às características sociodemográficas dos municípios e dos seus vizinhos.

Vernier (2016) também encontrou forte dependência espacial, sugerindo que a estrutura espacial tem influência no desempenho escolar: o desempenho de um município está positivamente associado ao desempenho dos municípios vizinhos. Adicionalmente, Vernier (2016) relata que a formação dos professores também apresenta forte associação com o resultado escolar. Os resultados indicaram que o status sociodemográfico do aluno (escolaridade dos pais + cor do aluno) e características dos professores (conceito da universidade de formação e com pós-graduação) também influenciam o resultado do município e, principalmente, a heterogeneidade e autocorrelação espacial mostraram-se importantes na educação dos municípios brasileiros.

## 4.1. Variáveis do Estudo

Antes da modelagem espacial necessitou-se criar as variáveis de percepções e expectativas docentes a partir das perguntas do questionário do "Problemas de Aprendizagem" do Sistema de Avaliação da Educação Básica – SAEB 2013 e 2017. Os professores eram inqueridos a opinar sobre sua percepção de possíveis problemas de aprendizagem dos alunos devido aos itens Q70 a Q82. A descrição desses itens e as outras variáveis utilizadas na pesquisa encontram-se resumidas no Quadro: 4.1.

Quadro: 4.1: Resumo das variáveis do estudo

| CATEGORIA | CÓDIGO | DESCRIÇÃO |
|---|---|---|
| Identificação | COD_MUN | Código do município |
| | NOME_MUN | Nome do município |
| | COD_UF | Código do estado |
| | SG_UF | Sigla do estado |
| | LATITUDE | Latitude |
| | LONGITUDE | Longitude |
| | RESP_M | Nº de respondentes / Nº de matrículas |
| Desempenho dos alunos | MAT | Proficiência em Matemática 9º ano |
| | POR | Proficiência em Língua Portuguesa 9º ano |

| | | |
|---|---|---|
| Adequação da formação docente | AFD1 | Docentes com formação superior de licenciatura na mesma disciplina que lecionam, ou bacharelado na mesma disciplina com curso de complementação pedagógica concluído. |
| | AFD2 | Docentes com formação superior de bacharelado na disciplina correspondente, mas sem licenciatura ou complementação pedagógica. |
| | AFD3 | Docentes com licenciatura em área diferente daquela que leciona, ou com bacharelado nas disciplinas da base curricular comum e complementação pedagógica concluída em área diferente daquela que leciona. |
| | AFD4 | Docentes com outra formação superior não considerada nas categorias anteriores. |
| | AFD5 | Docentes que não possuem curso superior completo. |
| Indicador da complexidade da gestão | ICG1 | Porte inferior a 50 matrículas, operando em único turno e etapa e apresentando a Educação Infantil ou Anos Iniciais como etapa mais elevada. |
| | ICG2 | Porte entre 50 e 300 matrículas, operando em 2 turnos, com oferta de até 2 etapas e apresentando a Educação Infantil ou Anos Iniciais como etapa mais elevada. |
| | ICG3 | Porte entre 50 e 500 matrículas, operando em 2 turnos, com 2 ou 3 etapas e apresentando os Anos Finais como etapa mais elevada. |
| | ICG4 | Porte entre 150 e 1000 matrículas, operando em 2 ou 3 turnos, com 2 ou 3 etapas, apresentando Ensino Médio/profissional ou a EJA como etapa mais elevada. |
| | ICG5 | Porte entre 150 e 1000 matrículas, operando em 3 turnos, com 2 ou 3 etapas, apresentando a EJA como etapa mais elevada. |
| | ICG6 | Porte superior à 500 matrículas, operando em 3 turnos, com 4 ou mais etapas, apresentando a EJA como etapa mais elevada. |
| Índice de esforço docente | IED1 | Docente que tem até 25 alunos e atua em um único turno, escola e etapa. |
| | IED2 | Docente que tem entre 25 e 150 alunos e atua em um único turno, escola e etapa. |
| | IED3 | Docente que tem entre 25 e 300 alunos e atua em um ou dois turnos em uma única escola e etapa. |
| | IED4 | Docentes que tem entre 50 e 400 alunos e atua em dois turnos, em uma ou duas escolas e em duas etapas. |
| | IED5 | Docente que tem mais de 300 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas. |
| | IED6 | Docente que tem mais de 400 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas. |
| Variáveis intraescolares | Q70 | Carência de infraestrutura física. |
| | Q71 | Carência ou ineficiência da supervisão, coordenação e orientação pedagógica. |
| | Q72 | Conteúdos curriculares inadequados às necessidades dos alunos. |
| | Q73 | Não cumprimento dos conteúdos curriculares ao longo da trajetória escolar do aluno. |
| | Q74 | Sobrecarga de trabalho dos professores, dificultando o planejamento e o preparo das aulas. |
| | Q75 | Insatisfação e desestímulo do professor com a carreira docente. |
| Variáveis extraescolares | Q76 | Meio social em que o aluno vive. |
| | Q77 | Nível cultural dos pais dos alunos. |
| | Q78 | Falta de assistência e acompanhamento dos pais na vida escolar do aluno. |

| | | |
|---|---|---|
| Variáveis dependentes dos alunos | Q79 | Baixa autoestima dos alunos. |
| | Q80 | Desinteresse e falta de esforço do aluno. |
| | Q81 | Indisciplina dos alunos em sala de aula. |
| | Q82 | Alto índice de faltas por parte dos alunos. |

**Nota:** As duas variáveis de desempenho (MAT e POR) são escores com média 250 e desvio-padrão 50. As variáveis AFD1 a Q82 são frequências entre 0 e 1. As categorias das variáveis/perguntas de expectativas docentes da Prova Brasil (Q70 a Q82) foram baseadas, *a priori*, no trabalho de Vidal *et al.* (2019). Têm-se informações das variáveis disponíveis para 2013 e 2017, sendo que, quando necessário, os subscritos _13 e _17 indicam os respectivos anos: sem nenhum subscrito os códigos indicam variações de 2013 para 2017 (vide Tabela 4.4).

Fonte: Autor

## 4.2. Método acessório – PCA

Para criar as variáveis de percepções e expectativas docentes a partir dos itens Q70 a Q82 do questionário do professor "Problemas de Aprendizagem" do SAEB, conforme sugerido por Vidal et al. (2019), utilizou-se da Análise de Componentes Principais (PCA).

A PCA nesta pesquisa foi utilizada para obtenção de estimativas dos fatores (componentes), para que sejam utilizados como substitutos aos itens Q70 a Q82, visando a uma redução da quantidade de itens do estudo (Aranha & Zambaldi, 2008). O intuito dessa agregação assenta no fato que partirmos do trabalho de Vidal et al. (2019), que propuseram uma agregação teoricamente justificável dos itens Q70 a Q82 do questionário do professor "Problemas de Aprendizagem" do SAEB, no entanto, não evoluíram em nenhuma análise estatística. Dessa forma, aplicaremos a PCA com objetivo de simplificar o estudo, e relegaremos outros aspectos, tais como (i) mecanismo de construção e validação de escala; ou (ii) exploração de dados (Aranha & Zambaldi, 2008).

Para o ajuste do modelo PCA, este estudo segue as recomendações de Ferreira (2018) e Hair et al. (2014) sobre as seguintes estatísticas de ajuste: a) Teste de esfericidade de Bartlett significativo; b) Medida Kaiser-Meyer-Olkin (KMO) > 0,70; c) Medidas de Adequação de Amostragem (MSA) > 0,50 (no conjunto e para as variáveis individualmente); d) Comunalidades > 0,50 (ressalta-se que comunalidades superiores a 0,30 já são adequadas para efeito desse estudo em virtude do tamanho da amostra); e) porcentagem da Variância Explicada ao redor de 50%. Esses critérios buscam considerar o equilíbrio entre a obtenção de um modelo parcimonioso e o total de explicação da variação retida por esse modelo (Ferreira, 2018).

Hair et al. (2014, p.107) colocam que nas ciências sociais, onde as informações são menos precisas, é comum considerar uma solução com no máximo 60% de variância explicada (as vezes até menos) como satisfatória. Henson & Roberts (2006), por exemplo, fizeram uma revisão de 267 estudos na área de educação e reportaram que as soluções, em média, explicavam 52,03% da variância. Nesse mesmo sentido, Izquierdo et al. (2014), no levantamento que fizeram de 117 estudos que utilizaram PCA nos principais periódicos de psicologia, indicaram que a variância explicada média foi de 54%, sendo que alguns trabalhos reportaram 20% de variância explicada.

## 4.3. Resultados dos modelos PCA

A descrição dos itens utilizados nos modelos PCA para a percepção e expectativa docente encontram-se na Tabela 4.1. Os valores representam a frequência de respostas "sim" às perguntas do questionário. Assim, por exemplo, 81.4% dos professores em 2017 achavam que os problemas de aprendizagem de seus alunos estavam relacionados ao "Meio social em que o aluno vive" (Q76). Essa percepção era de 78.1% em 2013, ou seja, houve um aumento da percepção de problemas de aprendizagem advindos do meio social em que os alunos vivem.

Tabela 4.1: Descrição dos itens das expectativas docentes

| | 2017 (n=5.108) | | 2013 (n=5.071) | |
|---|---|---|---|---|
| | Média | Desvio-padrão | Média | Desvio-padrão |
| Q70 | 0,349 | 0,331 | 0,320 | 0,332 |
| Q71 | 0,160 | 0,254 | 0,189 | 0,279 |
| Q72 | 0,135 | 0,225 | 0,137 | 0,227 |
| Q73 | 0,264 | 0,294 | 0,258 | 0,296 |
| Q74 | 0,290 | 0,308 | 0,320 | 0,323 |
| Q75 | 0,268 | 0,298 | 0,315 | 0,321 |
| Q76 | 0,814 | 0,264 | 0,781 | 0,289 |
| Q77 | 0,810 | 0,261 | 0,785 | 0,285 |
| Q78 | 0,933 | 0,162 | 0,914 | 0,195 |
| Q79 | 0,772 | 0,287 | 0,753 | 0,299 |
| Q80 | 0,942 | 0,148 | 0,923 | 0,181 |
| Q81 | 0,702 | 0,310 | 0,694 | 0,320 |
| Q82 | 0,416 | 0,333 | 0,423 | 0,345 |

Fonte: Autor

Para estimação dos componentes buscou-se ajustar dois modelos independentes, um para cada ano (2013 e 2017). Todos os itens (Q70 a Q82) foram considerados

simultaneamente em um mesmo modelo. Inicialmente, optou-se por estimar o modelo com os dados de 2017. O primeiro modelo, ajustado conforme critérios indicados na seção anterior, apresentou bons índices de ajustes [MSA's > 0,65; KMO = 0,74; Bartlett`s ($\chi^2$) = 9,570 (p-valor < 0,001); variância explicada = 45%; e cargas > 0,40], exceto por uma carga cruzada do item Q80.

Na sequência, com a opção de exclusão do item Q80, o modelo resultante também se mostrou adequado [MSA's > 0,65; KMO = 0,73; Bartlett`s ($\chi^2$) = 8,538 (p-valor < 0,001); variância explicada = 47%; e cargas > 0.50], principalmente, por terem sido encontrados três componentes teoricamente justificáveis. Conforme se esperava *a priori* (Vidal *et al.*, 2019): i) os itens Q70 a Q75 se aglutinaram em um componente que se denominou de "Intraescolar" (INTRA), por se referir às variáveis da própria escola, ou seja, pode ser "descrito por meio dos professores, diretores, projeto pedagógico, insumos, instalações, estrutura institucional, 'clima' da escola e relações intersubjetivas no cotidiano escolar" (Vidal *et al.*, 2019) ii) os itens Q76 a Q79 se aglutinaram em outro componente que se chamou de "Extraescolar" (EXTRA), que dizem respeito às condições de vida dos alunos, de suas famílias e de seu contexto social, cultural e econômico (Vidal *et al.*, 2019); e iii) os itens Q81 e Q82 se aglutinaram em um componente que se denominou "Aluno" (ALUNO), por se referir às variáveis relacionadas ao próprio aluno associadas a comportamentos e atitudes em relação ao ambiente escolar e ao processo de ensino-aprendizagem (Vidal *et al.*, 2019).

Basicamente, as únicas diferenças entre as expectativas de Vidal *et al.* (2019) sobre a percepção docente das causas dos possíveis problemas de aprendizagem nas turmas são: i) exclusão do item Q80 das análises; e ii) o item Q79 ficou carregado no fator "Extraescolar" em vez do fator "Aluno". Cabe ressaltar que Vidal *et al.* (2019) não executaram nenhuma análise estatística para a proposta de categorização dos itens Q70 a Q82: tão somente, a partir de uma análise qualitativa, avaliaram os itens apenas em termos descritivos.

Em seguida, aplicaram-se as mesmas condições aos dados de 2013, sem entretanto: i) considerar o item Q80; e ii) fixando o número de componentes em três. A intenção é que os componentes, nos diferentes anos, sejam os mais comparáveis possíveis. O modelo PCA resultante para 2013 apresentou medidas de ajustes similares ao modelo de 2017 [MSA's > 0,65; KMO = 0,74; Bartlett`s ($\chi^2$) = 9,105 (p-valor < 0,001); variância explicada = 48%; e cargas > 0,50]. Com esses resultados (Tabela 4.2 e Tabela 4.3)[5], os escores fatoriais

[5] Vide informações adicionais nos Apêndices.

padronizados (média = 0 e desvio-padrão = 1), para cada um dos municípios com observação em 2013 e 2017, foram gerados através do método de regressão.

Tabela 4.2: Matriz de componentes rotacionados (cargas fatoriais)

| | 2017 | | | 2013 | | |
|---|---|---|---|---|---|---|
| | Extraescolar | Intraescolar | Aluno | Extraescolar | Intraescolar | Aluno |
| Q70 | 0,544 | | | 0,539 | | |
| Q71 | 0,608 | | | 0,619 | | |
| Q72 | 0,621 | | | 0,596 | | |
| Q73 | 0,578 | | | 0,567 | | |
| Q74 | 0,61 | | | 0,634 | | |
| Q75 | 0,646 | | | 0,632 | | |
| Q76 | | 0,781 | | | 0,792 | |
| Q77 | | 0,806 | | | 0,815 | |
| Q78 | | 0,642 | | | 0,700 | |
| Q79 | | 0,518 | | | 0,549 | |
| Q81 | | | 0,738 | | | 0,716 |
| Q82 | | | 0,714 | | | 0,727 |

**Nota:** Extração por componentes principais e rotação varimax com normalização de Kaiser

Fonte: Autor

Tabela 4.3: Matriz de pesos dos escores fatoriais

| | 2017 | | | 2013 | | |
|---|---|---|---|---|---|---|
| | Extraescolar | Intraescolar | Aluno | Extraescolar | Intraescolar | Aluno |
| Q70 | 0,255 | 0,045 | -0,097 | 0,252 | 0,049 | -0,109 |
| Q71 | 0,299 | 0,018 | -0,162 | 0,303 | 0,02 | -0,194 |
| Q72 | 0,297 | -0,003 | -0,089 | 0,283 | -0,058 | 0,008 |
| Q73 | 0,273 | 0,02 | -0,086 | 0,269 | 0,03 | -0,125 |
| Q74 | 0,261 | -0,12 | 0,218 | 0,285 | -0,098 | 0,205 |
| Q75 | 0,28 | -0,114 | 0,195 | 0,285 | -0,094 | 0,195 |
| Q76 | -0,014 | 0,404 | -0,081 | -0,023 | 0,396 | -0,13 |
| Q77 | 0,011 | 0,429 | -0,159 | -0,007 | 0,406 | -0,136 |
| Q78 | -0,048 | 0,32 | 0,022 | -0,043 | 0,324 | 0,018 |
| Q79 | -0,019 | 0,229 | 0,139 | -0,023 | 0,227 | 0,128 |
| Q81 | -0,085 | -0,02 | 0,555 | -0,043 | -0,008 | 0,53 |
| Q82 | -0,071 | -0,036 | 0,539 | -0,061 | -0,074 | 0,564 |

**Nota:** Extração por componentes principais e rotação varimax com normalização de Kaiser

Fonte: Autor

Como os valores dos componentes “Extraescolar”, “Intraescolar” e “Aluno” representam o número de desvio-padrão do município ao redor da média (escore

padronizado), a variação (diferenciação) de 2013 para 2017, em cada um dos referidos componentes, indica o quanto o município se aproximou ou distanciou da média global (Brasil).

O aumento dos escores fatoriais em cada ano é condizente com o aumento da frequência de respostas "sim" para a percepção docente de possíveis problemas de aprendizagem dos alunos no referido município, sendo esperado que suas variações positivas impactem negativamente o desempenho do aluno. Assim, espera-se que os componentes "Extraescolar", "Intraescolar" e "Aluno" tenham efeito negativo sobre a proficiência em Matemática ou Língua Portuguesa.

### 4.4. Análise descritiva das variáveis

Após a extração dos componentes/variáveis INTRA, EXTRA e ALUNO para os anos de 2013 e 2017, calculou-se a variação (2017 – 2013) para todas as variáveis utilizadas na pesquisa, ou seja, nas análises que seguem utilizam-se as primeiras diferenças das variáveis.

As estatísticas descritivas de todas as variáveis utilizadas na AEDE e nos modelos espaciais encontram-se dispostas na Tabela 4.4, que apresenta informação das observações disponíveis para cada uma das variáveis, no entanto, nas análises que seguem (AEDE e modelos espaciais) faz-se uso da matriz completa, sem *missings value* (n = 4.661).

Em termos médios, pode-se constatar que houve um aumento da proficiência de Matemática e Língua Portuguesa de 2013 para 2017: a proficiência em Matemática aumentou, em termos médios, em 7 pontos e a proficiência em Língua Portuguesa quase o dobro, em 13 pontos. Adicionalmente, a oscilação para cima ou para baixo também foi alta, por exemplo, tivemos município no qual o escore de proficiência em Matemática variou em 101 pontos para cima ou 55 pontos para baixo (a amplitude da proficiência em Língua Portuguesa foi menor: entre -43 e 67 pontos). Como as variáveis AFD, ICG e IED são proporções/frequências, os valores da Tabela 4.4 indicam variações em pontos percentuais, por exemplo, no caso da variável ICG2: a média de 0,0306 indica que de 2013 para 2017 os municípios com escolas de "porte entre 50 e 300 matrículas, operando em 2 turnos, com oferta de até 2 etapas e apresentando a Educação Infantil ou Anos Iniciais como etapa mais elevada" aumentou em 3,06 pontos percentuais.

Tabela 4.4: Estatística descritiva das variações (2017-2013) das variáveis utilizadas na pesquisa

| | n | Mínimo | Máximo | Média | Desvio-padrão |
|---|---|---|---|---|---|
| RESP_M | 4741 | -0,09 | 0,12 | 0,0069 | 0,01443 |
| MAT | 5295 | -55,64 | 101,62 | 7,0578 | 13,53147 |
| POR | 5295 | -42,89 | 66,97 | 13,2209 | 13,34891 |
| AFD1 | 5570 | -0,42 | 0,52 | 0,0383 | 0,09776 |
| AFD2 | 5570 | -0,38 | 0,38 | -0,0046 | 0,03901 |
| AFD3 | 5570 | -0,78 | 0,46 | 0,0016 | 0,10539 |
| AFD4 | 5570 | -0,32 | 0,50 | -0,0032 | 0,04573 |
| AFD5 | 5570 | -0,88 | 0,74 | -0,0321 | 0,10807 |
| ICG1 | 5570 | -0,70 | 0,67 | -0,0133 | 0,14127 |
| ICG2 | 5570 | -0,75 | 1,00 | 0,0306 | 0,14652 |
| ICG3 | 5570 | -1,00 | 0,67 | 0,0128 | 0,12139 |
| ICG4 | 5570 | -0,75 | 1,00 | -0,0084 | 0,12120 |
| ICG5 | 5570 | -0,67 | 0,78 | -0,0108 | 0,11982 |
| ICG6 | 5570 | -0,67 | 0,67 | -0,0109 | 0,05871 |
| IED1 | 5570 | -0,63 | 0,32 | -0,0080 | 0,06238 |
| IED2 | 5570 | -0,46 | 1,00 | 0,0032 | 0,08542 |
| IED3 | 5570 | -0,63 | 0,80 | 0,0064 | 0,10431 |
| IED4 | 5570 | -0,90 | 0,80 | -0,0019 | 0,12898 |
| IED5 | 5570 | -0,69 | 0,78 | -0,0001 | 0,09256 |
| IED6 | 5570 | -0,28 | 0,33 | 0,0005 | 0,05033 |
| INTRA | 4741 | -5,14 | 6,26 | 0,0054 | 1,23644 |
| EXTRA | 4741 | -6,28 | 5,32 | -0,0048 | 1,29443 |
| ALUNO | 4741 | -5,39 | 5,20 | 0,0176 | 1,26351 |
| n Válido (*listwise*) | 4661 | | | | |

**Nota:** A descrição das variáveis RESP_M a IED6 é apresentada no Quadro: 4.1. INTRA, EXTRA e ALUNO são as variações dos componentes estimados na seção anterior.

Fonte: Autor

No caso das variáveis INTRA, EXTRA e ALUNO, como são escores de média zero e desvio-padrão um, construídos independentemente para cada um dos anos, suas variações indicam mudanças relativas dos municípios na percepção docente sobre os problemas de aprendizagem, por exemplo, no caso da média de 0,0176 para a variável ALUNO, esse valor traduz que, em geral, a percepção do professor associada ao próprio aluno concernente a comportamentos e atitudes em relação ao ambiente escolar e ao processo de ensino-aprendizagem, piorou levemente de 2013 para 2017.[6]

[6] As perguntas Q81 e Q82 são negativas: "na sua percepção, os possíveis problemas de aprendizagem dos alunos das séries ou anos avaliados ocorrem, nesta escola, devido à: i) indisciplina dos alunos em sala de aula (1 - Sim/0 - Não); e ii) alto índice de faltas por parte dos alunos (1 - Sim/0 – Não)"; e como os pesos dos escores fatoriais são positivos, um aumento (respostas Sim = 1) indica uma piora no sentido pedagógico.

# 5. AEDE no Contexto do Estudo

Este capítulo apresenta os resultados da análise exploratória dos dados espaciais.

## 5.1. Especificação dos pesos espaciais

Apesar de algumas análises exploratórias no espaço não requererem a matriz W de vizinhança (matriz de pesos espaciais), optou-se por, inicialmente, especificá-la. Apesar de tratar-se de um processo *ad hoc*, seguiram-se os procedimentos recomendados pela literatura (Almeida, 2012; Anselin & Rey, 2014; Golgher, 2015), tal como buscar o tipo de matriz que maximiza a autocorrelação espacial (evidências de interação) presente nos dados (Vieira, 2009). Para isso, para cada uma das duas variáveis dependentes (MAT e POR), estimaram-se os modelos descritos na Tabela 6.1 e Tabela 6.3 por OLS, obtiveram-se os resíduos e, através deles, calcularam-se e avaliaram-se os índices de Moran global para diversas matrizes de pesos espaciais.

Começou-se com as matrizes contíguas, mais simples, Rainha e Torre, cujos achados foram similares e indicaram evidências de autocorrelação espacial positiva para as duas variáveis dependentes (Tabela 5.1). Essas duas matrizes geraram número de ilhas irrelevantes e número de vizinhos parcimoniosos. Ao redor da média de vizinhos desses dois tipos de matrizes espaciais (5,27 e 5,14, respectivamente), simulamos três outros tipos de matrizes baseados nos 4, 5 e 6 vizinhos mais próximos (*KNN*). Nesse tipo de abordagem não se geram ilhas, mas fixa o número de vizinhos: nessas três matrizes, os valores do I de Moran foram similares, indicando estabilidade dos resultados. Adicionalmente, simularam-se três extremos: 10, 15 e 20 vizinhos mais próximos; e os valores do I de Moran caíram ligeiramente.

Tabela 5.1: Especificação da Matriz W

| **Tipo** | **Variável Dependente** | **I de Moran** | **Z** | **Média Vizinhos** | **n Ilhas** | **Máximo Vizinhos** |
|---|---|---|---|---|---|---|
| Rainha | MAT | 0,221 | 23,168 | 5,27 | 2 | 23 |
| | POR | 0,179 | 18,628 | | | |
| Torre | MAT | 0,222 | 22,850 | 5,14 | 2 | 21 |
| | POR | 0,180 | 18,417 | | | |
| *KNN*-4 | MAT | 0,220 | 22,618 | 4,00 | 0 | 4 |
| | POR | 0,183 | 18,358 | | | |
| *KNN*-5 | MAT | 0,216 | 24,818 | 5,00 | 0 | 5 |
| | POR | 0,176 | 19,791 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| *KNN*-6 | MAT | 0,216 | 27,269 | 6,00 | 0 | 6 |
| | POR | 0,173 | 21,302 | | | |
| *KNN*-10 | MAT | 0,210 | 33,014 | 10,00 | 0 | 10 |
| | POR | 0,175 | 27,340 | | | |
| *KNN*-15 | MAT | 0,208 | 39,304 | 15,00 | 0 | 15 |
| | POR | 0,173 | 33,321 | | | |
| *KNN*-20 | MAT | 0,209 | 45,918 | 20,00 | 0 | 20 |
| | POR | 0,173 | 38,703 | | | |
| *KM*-282 | MAT | 0,151 | 83,924 | 290,65 | 2 | 584 |
| | POR | 0,126 | 69,730 | | | |
| *KM*-80 | MAT | 0,198 | 35,257 | 32,62 | 49 | 105 |
| | POR | 0,165 | 29,178 | | | |
| *KM*-40 | MAT | 0,206 | 21,314 | 8,78 | 312 | 39 |
| | POR | 0,170 | 17,630 | | | |

**Nota:** *KNN = K-nearest neighbors*, onde o valor após o hífen indica o número de vizinhos considerados; KM = quilômetros, onde o valor após o hífen indica a distância máxima entre os centroides dos municípios para ser considerados vizinhos; I de Moran univariado global, calculado a partir dos resíduos OLS dos modelos da Tabela 6.1 e Tabela 6.3, com inferência (z-valor e *pseudo* p-valor = 0,001) calculado a partir de 999 permutações (randomização).

Fonte: Autor

Para complementar a busca, simularam-se três outros tipos de matrizes baseados nos vizinhos até 282, 80 e 40 quilômetros de distância. O primeiro, KM-282, é o mínimo de distância requerida para não haver nenhuma ilha. Esse valor foi sugerido pelo próprio *software* GeoDa, baseado nas variáveis de latitude e longitude, e não destoa de outras evidências empíricas sobre a mesma variável dependente (Fujita et al., 2021) . Os dois outros tipos, KM-80 e KM-40, foram cogitados baseados na inspeção do correlograma espacial, como se apresenta nos Apêndices. Na primeira (KM-282), os valores do I de Moran reduziram substancialmente, e nas duas últimas (KM-80 e KM-40), apesar de isso não acontecer, o número médio e máximo de vizinho aumentou substancialmente. No caso da matriz KM-40 o número de ilhas torna-se inaceitável.

A partir dessas inspeções optou-se pela matriz W mais parcimoniosa: Rainha. Golgher (2015, p. 202) coloca que na dúvida deve-se escolher uma matriz simples, como a de contiguidade. O autor relata ainda que nos trabalhos empíricos em estudos regionais o número de conexões (vizinhos) varia entre 3,8 e 6,0 (Golgher, 2015, p. 203) e algumas evidências empíricas favoráveis às estimações com matrizes menos conectadas, como de contiguidade ou com os cinco ou sete vizinhos mais próximos (Golgher, 2015, p. 204-205). Em trabalho semelhante, Cavalcanti et al. (2020) consideraram uma matriz com 5 vizinhos

mais próximos, valor praticamente o mesmo da média de vizinhos (5,27) da matriz W considerada no presente trabalho.

## 5.2. Mapas descritivos das variáveis dependentes

Como a proposta do trabalho é fazer uma análise espacial, principalmente, da variação da proficiência em Matemática e Língua Portuguesa entre os ciclos de avaliação do SAEB de 2013 e 2017, as análises descritivas concentram-se em ilustrações de mapas das variáveis dependentes, deixando para os Apêndices os mapas descritivos das variáveis independentes do estudo. A Figura 5.1 e Figura 5.2 indicam a distribuição das variações da proficiência em Matemática (Figura 5.1) e Língua Portuguesa (Figura 5.2) em seis classes uniformes para os valores observados da variável (a) e seu desvio-padrão (b). Esse mesmo tipo de mapa também é apresentado para as outras variáveis do estudo nos Apêndices.

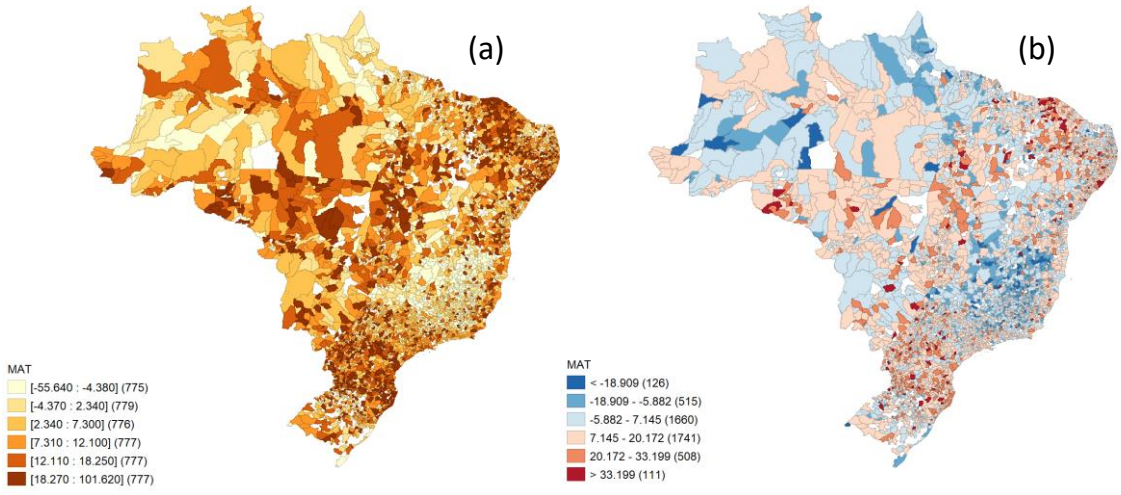


Figura 5.1: Mapa quantílico (a) e de desvio-padrão (b) da variação da proficiência em Matemática
Fonte: Autor

Sem pretensão de buscar padrões nos mapas elencados na Figura 5.1 e Figura 5.2, alguns fatos estilizados sobressaem, que até mesmo fazem sentido a partir da percepção empírica (Cavalcanti et al., 2020; Fujita et al., 2021), tal como: i) menores variações da proficiência em Matemática no noroeste de Minas Gerais e Bahia; ii) maiores variações da proficiência em Matemática no Sul, especificamente em Paraná e Santa Catarina; ii) a proficiência em Matemática e a proficiência em Língua Portuguesa parecem seguir um

padrão semelhante de variação: em municípios nos quais houve uma variação positiva em Língua Portuguesa também houve em Matemática; iv) o padrão de variabilidade (desvio-padrão) parece ser o mesmo tanto para a variação da proficiência em Matemática como da proficiência em Língua Portuguesa; v) os municípios do Norte são mais homogêneos (menor variabilidade) quanto às variações das proficiências, tanto para Língua Portuguesa quanto para Matemática.

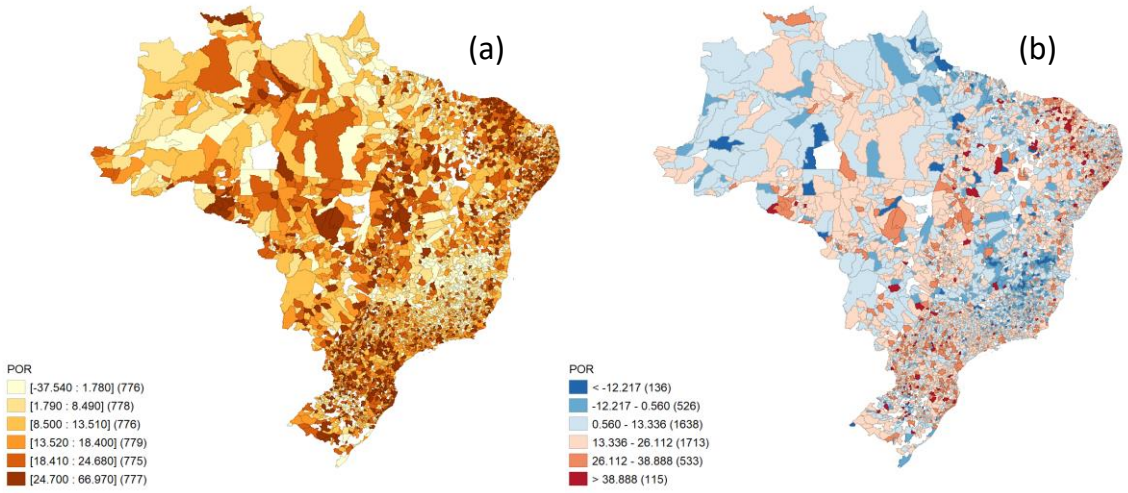


Figura 5.2: Mapa quantílico (a) e de desvio-padrão (b) da variação da proficiência em Língua Portuguesa
Fonte: Autor

## 5.3. Mapas LISA das variáveis dependentes

Para inferências mais contundentes sobre padrões, ou *clusters* espaciais, deve-se avaliar o indicador de correlação espacial de Moran local, materializados nos mapas LISA, como descrito na seção 2.4 dessa dissertação. As Figura 5.3 e Figura 5.4 apresentam o mapa LISA dos regimes espaciais (a) e da indicação de significâncias (p-valor < 0,01) dos indicadores de Moran local (b) para as variações das proficiências em Matemática (Figura 5.3) e em Língua Portuguesa (Figura 5.4). Os valores em vermelho no mapa LISA (a) são os *hot spots (high-high)*, onde variações altas da proficiência em Matemática/Língua Portuguesa são acompanhadas por variações altas dos vizinhos, e os valores em azul são os *cold spots (low-low)*, onde variações baixas da proficiência em Matemática/Língua Portuguesa são acompanhadas por variações baixas dos vizinhos. Geralmente, os regimes *low-high* ou *high-low* são considerados *outliers* espaciais (Anselin & Rey, 2014).

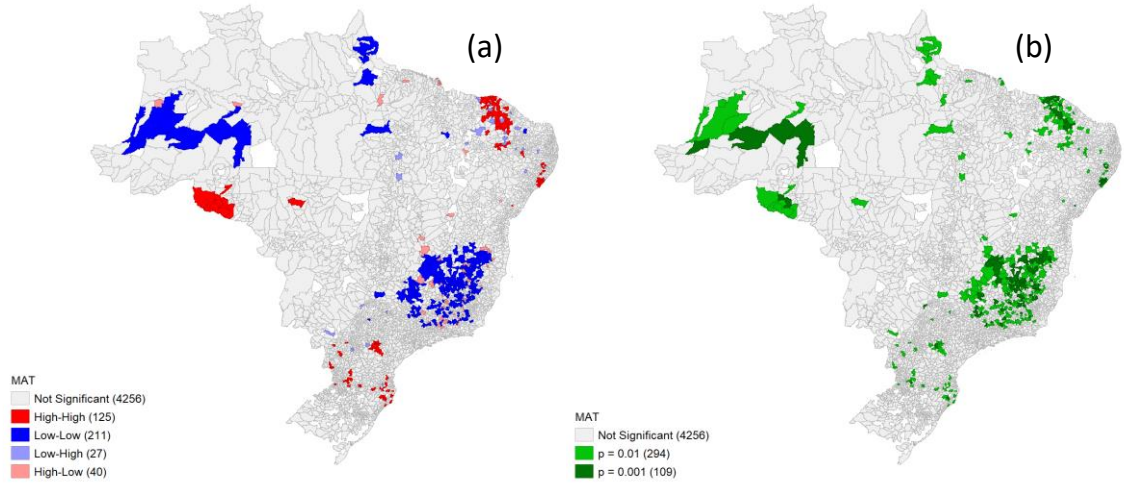


Figura 5.3: Mapa LISA (a) e de significância (p < 0,01) (b) da variação em MAT

Fonte: Autor

De uma forma geral, percebe-se mais pontos de *cold spots* do que de *hot spots*, mais evidente para a variação da proficiência em Matemática (n = 211 *versus* n = 125) do que para a variação da proficiência em Língua Portuguesa (n = 173 *versus* n = 130).

A maior concentração de pontos *low-low* parece se concentrar no nordeste de Minas Gerais e Bahia, no caso da proficiência de Língua Portuguesa (Figura 5.4). E no caso da variação da proficiência em Matemática, além dos *clusters* espaciais nessas mesmas localizações, também parece haver outro no Norte (Figura 5.3).

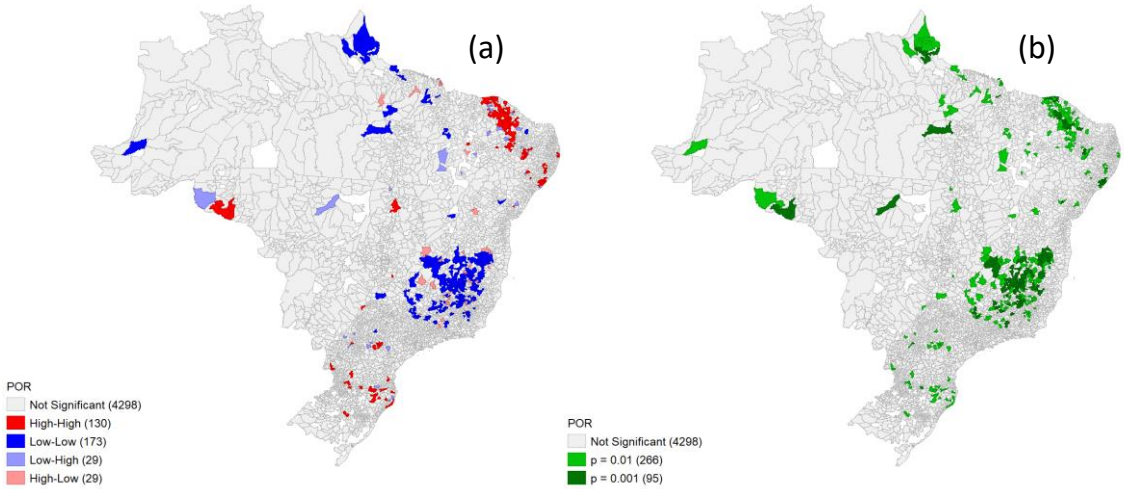


Figura 5.4: Mapa LISA (a) e de significância (p < 0,01) (b) da variação em POR

Fonte: Autor

Em relação aos *hot spots* percebe-se que há uma concentração de pontos *high-high* no Ceará, tanto para a variação da proficiência em Matemática como para Língua Portuguesa. No caso de variação da proficiência em Matemática, também existe um *cluster* espacial em Rondônia (Figura 5.4). Outros mapas LISA univariados[7] para as variáveis independentes são apresentados nos Apêndices. Neles encontramos poucos *clusters* espaciais, exceto no caso das variáveis AFD2, AFD3 e AFD4. Na variável AFD2, por exemplo, encontramos evidentes *hot spots* em Mato Grosso/Pará e *colds spots* em São Paulo/Minas Gerais, indicando haver um aumento (de 2013 para 2017) de municípios (e vizinhança) com "docentes com formação superior de bacharelado na disciplina correspondente, mas sem licenciatura ou complementação pedagógica" nas regiões centro-oeste/norte, e uma redução na região sudeste, pelo menos nos estados com os *clusters* espaciais identificados.

[7] Vide que discutimos apenas mapas LISA's univaridados e não bivariados. Como veremos adiante nos modelos espaciais, foram pouquíssimas relações significantes ao ponto de não fazer jus uma análise espacial bivariada detalhada.

## 6. Análise dos Modelos Espaciais

Cabe ressaltar que para estimação dos modelos utilizaram-se as variações das variáveis entre 2017 e 2013. Esse procedimento é recomendado quando se tem dados longitudinais com apenas dois pontos no tempo, e reside em eliminar as influências de fatores não observados que não mudam com o tempo, para melhor controle das variáveis omitidas. Com isso, de fato, se utiliza um tipo de estimador de efeito fixo, especificamente, o método de primeiras diferenças para dados em painel (Almeida, 2012), em que o resultado, por se ter apenas dois pontos no tempo, gera dados *cross section,* e assim, pode-se aplicar os procedimentos usuais (clássicos) para modelagem de dados espaciais (Almeida, 2012; Golgher, 2015).

Como descrito no capítulo 3, os componentes espaciais do modelo podem aparecer, basicamente, por meio de três formas: (1) na forma de defasagem espacial na variável dependente ($\rho Wy$), (2) na forma de defasagem nas variáveis explicativas ($WX\tau$), ou então (3) como defasagem no termo de erro ($\lambda W\xi$ ou $\gamma W\varepsilon$). Tais componentes podem aparecer de forma isolada ou em conjunto (Vieira, 2009). Como já comentado, serão considerados para as análises espaciais apenas modelos que levem em consideração a variável dependente espacialmente defasada ($\rho Wy$) e/ou o erro espacialmente defasado ($\lambda W\xi$ ou $\gamma W\varepsilon$), isso porque, estima-se, e as evidências sugerem, não haver influência espacial (*spillover*) das variáveis independentes consideradas, principalmente aquelas relacionadas com a percepção docente sobre a escola, condições extraescolares e do aluno, sobre a proficiência em Língua Portuguesa ou Matemática do município.

A partir dessa decisão, temos menor risco, devido excluir modelos espaciais mais complexos para estimar (SDM, SLX etc.), e ficamos restritos aos modelos SAR, SEM e SAC. Assim, seguindo o procedimento clássico (Almeida, 2012) de especificação de modelos espaciais, ou em outros dizeres, a estratégica específica-geral (Golgher, 2015), começando com o modelo OLS básico, estimaram-se os outros três modelos espaciais (SAR, SEM e SAC), um para cada variável dependente (MAT e POR), e a partir dos testes diagnósticos dos resíduos do modelo OLS e medidas de ajuste recomendadas, tomaram-se as decisões sobre os estimadores utilizados e modelos finais para interpretação.

De uma forma geral, algumas evidências foram comuns e contundentes em todas as análises (modelos) procedidas: i) rejeita-se veemente a hipótese de normalidade dos

resíduos, conforme apontado pelo teste Jarque-Bera; ii) rejeita-se a hipótese de homogeneidade dos resíduos, conforme evidenciado pelo teste Koenker-Basset; e iii) há evidências de algum tipo de dependência espacial, como apontado pelos outros testes diagnósticos (I de Moran dos resíduos e testes LM).

O primeiro e segundo ponto levam à decisão de considerar os Métodos 2SLS/GMM em detrimento do método ML para estimar os parâmetros dos modelos, um porque o método ML[8] requer normalidade, e outro devido à evidência de heterogeneidade, que somente nos métodos 2SLS/GMM, a partir das correções do erros-padrão HAC e o procedimento de Kelejian e Prucha, poderia ser tratada, já que optou-se seguir o caminho sugerido por Anselin (1988): tratar os problemas gerados pela heterogeneidade espacial com o uso de instrumentos fornecidos pela economia clássica, tal qual o uso de erros-padrão robustos (Vieira, 2009), em detrimento de um modelo de heterogeneidade espacial. O terceiro ponto indica a pertinência dos modelos espaciais aos dados em detrimento de um modelo OLS básico.

### 6.1. Proficiência em Matemática

Partindo do modelo específico, estimado por OLS, e diagnosticando seus resíduos, podemos constatar, conforme Tabela 6.1, alguns direcionamentos para ajustar o modelo espacial da variação da proficiência em Matemática: i) o teste difuso I de Moran dos resíduos indicou algum tipo de dependência espacial (I = 0,22; Z = 23,51; p-valor < 0,001); e ii) os testes focados LM direcionaram para um SAR ou SAC.

Essa última conclusão surgiu porque todos os testes LM, inclusive os robustos, foram altamente significativos, porém os testes LMρ e LMρ robusto foram mais significativos (maior valor da estatística $\chi^2$), direcionando assim para um modelo SAR. No entanto, como também os testes LMλ e LMλ robusto, inclusive o teste LM (SARMA) que testa, conjuntamente, $\rho = 0$ e $\lambda = 0$, foi significativo, direcionando porventura um modelo de ordem superior do tipo SAC, esse deve ser considerado apenas se houver ganhos efetivos em eliminar a dependência espacial que ficou nos resíduos (Anselin & Rey, 2014).

[8] De toda forma, também buscamos ajustar os referidos modelos por ML, no entanto, não tivemos sucesso devido a problemas de convergência.

Tabela 6.1: Modelo OLS com erros-padrão HAC para a proficiência em matemática

| MAT | **Coeficiente** | **Erro-padrão** | **T** | **p-valor** |
|---|---|---|---|---|
| Constante | 6,595 | 0,347 | 19,023 | 0,001 |
| AFD2 | -2,625 | 5,287 | -0,497 | 0,620 |
| AFD3 | -14,245 | 2,784 | -5,116 | 0,001 |
| AFD4 | -18,749 | 5,093 | -3,681 | 0,001 |
| AFD5 | -10,191 | 2,633 | -3,870 | 0,001 |
| ICG2 | -0,748 | 1,740 | -0,430 | 0,668 |
| ICG3 | 1,782 | 2,417 | 0,737 | 0,461 |
| ICG4 | -3,407 | 2,676 | -1,273 | 0,203 |
| ICG5 | 0,487 | 2,533 | 0,192 | 0,848 |
| ICG6 | 5,589 | 4,189 | 1,334 | 0,182 |
| IED2 | -2,625 | 4,541 | -0,578 | 0,563 |
| IED3 | 5,891 | 4,515 | 1,305 | 0,192 |
| IED4 | -4,148 | 4,238 | -0,979 | 0,328 |
| IED5 | -11,878 | 4,766 | -2,492 | 0,013 |
| IED6 | -7,242 | 6,071 | -1,193 | 0,233 |
| ALUNO | -0,352 | 0,172 | -2,045 | 0,041 |
| EXTRA | -0,288 | 0,163 | -1,766 | 0,077 |
| INTRA | 0,135 | 0,162 | 0,832 | 0,405 |
| RESP_M | 24,890 | 15,400 | 1,616 | 0,106 |
| Nº Observações: | 4661 | | LL: | -18526,666 |
| Adj. $R^2$ (%): | 0,018% | | AIC: | 37091,331 |
| estatística *F*: | 5,748*** | | SIC: | 37213,824 |

| **Teste Diagnóstico** | **Valor/DF** | **Estatística** | **p-valor** |
|---|---|---|---|
| I de Moran: | 0,2211 | 23,511 | 0,001 |
| LMρ: | 1 | 597,32 | 0,001 |
| LMρ robusto: | 1 | 101,79 | 0,001 |
| LMλ: | 1 | 549,01 | 0,001 |
| LMλ robusto: | 1 | 53,484 | 0,001 |
| LM (SARMA): | 2 | 650,804 | 0,001 |
| Jarque-Bera: | 2 | 947,574 | 0,001 |
| Koenker-Basset: | 18 | 108,118 | 0,001 |

* p-valor < 0,1; ** p-valor < 0,05; *** p-valor < 0,01. Multicolinearidade = 4,963; Adj $R^2$ = coeficiente de determinação ajustado; LL = Log da razão de verossimilhança; AIC = Critério de informação de Akaike; BIC = Critério de informação de Schwarz; LM = Multiplicador de Lagrange; DF = graus de liberdade. Modelo estimado por Mínimos Quadrados Ordinários (OLS) com erros-padrão ajustado para heterocedasticidade e autocorrelação (HAC), considerando uma função kernel uniforme e distância euclidiana, e matriz de pesos espaciais rainha, selecionada conforme especificação empreendida na Tabela 5.1.

Fonte: Autor

Os testes Jarque-Bera e Koenker-Basset rejeitaram a hipótese de normalidade e homogeneidade dos resíduos, respectivamente, e pelos valores da estatísticas $\chi^2$, a falta de normalidade e presença de heterogeneidade parece ser exacerbada, corroborando o uso de estimadores que não levem em conta a normalidade dos dados e que os erros-padrão possam

ser corrigidos para a presença de heterogeneidade. Assim, apresentamos todos os modelos espaciais estimados por 2SLS na Tabela 6.2, inclusive o modelo SEM para comparabilidade.

Os erros-padrão de todos os modelos, para as devidas inferências, foram ajustados quanto a heterogeneidade, seja pelo método HAC ou de Kelejian e Prucha. A única medida de ajuste existente quando se usam estimadores 2SLS/GMM, o *pseudo* $R^2$ ratifica as suspeitas iniciais: os modelos SAR e SAC são os melhores que se ajustam aos dados.

Tabela 6.2: Modelos espaciais 2SLS com erros-padrão robustos para a proficiência em matemática

| MAT | **SAR** | **SEM** | **SAC** |
|---|---|---|---|
| Constante | 0,591 | 6,853*** | 0,297 |
| AFD2 | -0,598 | -0,926 | -0,695 |
| AFD3 | -6,467** | -9,734*** | -3,903 |
| AFD4 | -6,789 | -11,965** | -4,145 |
| AFD5 | -4,436* | -6,792*** | -2,277 |
| ICG2 | 0,947 | 0,350 | 0,198 |
| ICG3 | 1,258 | 1,184 | 1,417 |
| ICG4 | -0,431 | -1,303 | -1,566 |
| ICG5 | 1,871 | 1,574 | 0,803 |
| ICG6 | 5,645 | 5,359 | 4,153 |
| IED2 | -5,451 | -4,907 | -2,543 |
| IED3 | 2,414 | 3,194 | 3,390 |
| IED4 | -4,162 | -4,690 | -1,866 |
| IED5 | -5,900 | -8,555** | -3,405 |
| IED6 | -4,412 | -5,630 | -1,772 |
| ALUNO | -0,248 | -0,255* | -0,281** |
| EXTRA | -0,022 | -0,110 | -0,041 |
| INTRA | 0,085 | 0,106 | 0,011 |
| RESP_M | 7,477 | 14,098 | 6,206 |
| $\rho$ | 0,898*** | | 0,944*** |
| $\lambda$ | | 0,456*** | -0,712*** |
| Nº Observações: | 4661 | 4661 | 4661 |
| Pseudo $R^2$ (%): | 15,84% | 2,06% | 15,70% |
| **Teste Diagnóstico** | **Valor/DF** | **Estatística** | **p-valor** |
| Anselin-Kelejian: | 1 | 113,366 | 0,001 |

* p-valor < 0,1; ** p-valor < 0,05; *** p-valor < 0,01. SAR = modelo de defasagem espacial estimado por Mínimos Quadrados Ordinários em Dois Estágios (2SLS) com erros-padrão ajustado para heterocedasticidade e autocorrelação (HAC); SEM = modelo de erro autoregressivo espacial estimado pelo método 2SLS espacial generalizado de Kelejian e Prucha; SAC = modelo de defasagem espacial com erro autoregressivo espacial estimado pelo método 2SLS espacial generalizado de Kelejian e Prucha. Para ajuste da heterocedasticidade nos modelos 2SLS considerou-se uma função kernel uniforme e distância euclidiana. A matriz de pesos espaciais utilizadas foi do tipo rainha, conforme especificação empreendida na Tabela 5.1.

Fonte: Autor

Lembrando que para as medidas AFD, ICG e IED não foram consideradas as *dummies* AFD1, ICG1 e IED1 para não incorrer em colinearidade perfeita, e assim, essas variáveis são as categorias-base. Desse modo, tomando o modelo SAR como o modelo final, preferido em detrimento do modelo SAC[9], tem-se que em municípios com maior proporção de "docentes com licenciatura **em área diferente** daquela que leciona, ou com bacharelado nas disciplinas da base curricular comum e complementação pedagógica concluída **em área diferente** daquela que leciona" (AFD3) em relação a "docentes com formação superior de licenciatura **na mesma disciplina** que lecionam, ou bacharelado **na mesma disciplina** com curso de complementação pedagógica concluído" (AFD1) a variação da proficiência em Matemática foi menor.

Nesse mesmo sentido, em municípios com maior proporção de "docentes que **não possuem curso superior completo**" (AFD5) do que "docentes com formação superior de licenciatura **na mesma disciplina** que lecionam, ou bacharelado **na mesma disciplina** com curso de complementação pedagógica concluído" (AFD1) a variação da proficiência em Matemática foi menor. Isso aconteceu, tanto para AFD3 e AFD5 no modelo SAR devido os coeficientes significativos e negativos. Esses achados estão condizentes com Fujita et al. (2021) e Vernier (2016), que relataram que a maior proporção de docentes pós-graduados nos municípios se relacionou positivamente com a performance dos alunos na escola. O valor dos coeficientes em modelos SAR não se interpreta em termos de variações marginais (derivadas parciais), como nos modelos OLS ou SEM, devido os efeitos *spillovers* de realimentação dos vizinhos, materializados no coeficiente $\rho$, e por isso optou-se por fazer esse tipo de análise somente na seção seguinte.

## 6.2. Proficiência em Língua Portuguesa

No caso da variação da proficiência em Língua Portuguesa, o modelo preterido também foi do tipo SAR, pelos mesmos motivos e enredo discutido na seção anterior: i) os testes Jarque-Bera e Koenker-Basset rejeitaram a hipótese de normalidade e homogeneidade dos resíduos (Tabela 6.3), respectivamente, e todos os modelos espaciais foram estimados por 2SLS, conforme Tabela 6.4: ii) os erros-padrão de todos os modelos foram ajustados quanto a heterogeneidade: e iii) o teste I de Moran dos resíduos, os testes focados LM e a

[9] Apesar da significância de $\lambda$ no modelo SAC o erro espacialmente defasado não melhorou a explicação da variação da proficiência em Matemática, e por isso, optou-se pelo modelo SAR, mais parcimonioso.

medida de ajuste *pseudo* $R^2$ direcionaram para os modelos SAR e SAC, sendo avaliado o primeiro (modelo final), pois apesar da significância de λ no modelo SAC o erro espacialmente defasado não melhorou a explicação da variação da proficiência em Língua Portuguesa, e por isso optou-se pelo modelo mais parcimonioso.

Tabela 6.3: Modelo OLS com erros-padrão HAC para a proficiência em língua portuguesa

| POR | **Coeficiente** | **Erro-padrão** | **T** | **p-valor** |
|---|---|---|---|---|
| Constante | 12,731 | 0,336 | 37,903 | 0,001 |
| AFD2 | -7,447 | 4,746 | -1,569 | 0,117 |
| AFD3 | -12,991 | 2,755 | -4,715 | 0,001 |
| AFD4 | -17,902 | 5,258 | -3,405 | 0,001 |
| AFD5 | -7,425 | 2,675 | -2,776 | 0,006 |
| ICG2 | -1,350 | 1,738 | -0,777 | 0,437 |
| ICG3 | 1,829 | 2,397 | 0,763 | 0,445 |
| ICG4 | -2,334 | 2,457 | -0,950 | 0,342 |
| ICG5 | -2,042 | 2,338 | -0,874 | 0,382 |
| ICG6 | 1,968 | 4,007 | 0,491 | 0,623 |
| IED2 | 2,293 | 4,314 | 0,532 | 0,595 |
| IED3 | 9,482 | 4,252 | 2,230 | 0,026 |
| IED4 | 0,543 | 4,009 | 0,135 | 0,892 |
| IED5 | -3,030 | 4,633 | -0,654 | 0,513 |
| IED6 | 0,710 | 5,963 | 0,119 | 0,905 |
| ALUNO | -0,446 | 0,164 | -2,716 | 0,007 |
| EXTRA | -0,349 | 0,160 | -2,182 | 0,029 |
| INTRA | 0,023 | 0,163 | 0,141 | 0,888 |
| RESP_M | 31,394 | 15,434 | 2,034 | 0,042 |
| Nº Observações: | 4661 | | LL: | -18441,973 |
| Adj. $R^2$ (%): | 0,016% | | AIC: | 36921,946 |
| estatística *F*: | 5,085*** | | SIC: | 37044,439 |

| **Teste Diagnóstico** | **Valor/DF** | **Estatística** | **p-valor** |
|---|---|---|---|
| I de Moran: | 0,1791 | 19,052 | 0,001 |
| LMρ: | 1 | 393,73 | 0,001 |
| LMρ robusto: | 1 | 77,41 | 0,001 |
| LMλ: | 1 | 360,15 | 0,001 |
| LMλ robusto: | 1 | 43,827 | 0,001 |
| LM (SARMA): | 2 | 437,556 | 0,001 |
| Jarque-Bera: | 2 | 142,703 | 0,001 |
| Koenker-Basset: | 18 | 140,665 | 0,001 |

* p-valor < 0,1; ** p-valor < 0,05; *** p-valor < 0,01. Multicolinearidade = 4,963; Adj $R^2$ = coeficiente de determinação ajustado; LL = Log da razão de verossimilhança; AIC = Critério de informação de Akaike; BIC = Critério de informação de Schwarz; LM = Multiplicador de Lagrange; DF = graus de liberdade. Modelo estimado por Mínimos Quadrados Ordinários (OLS) com erros-padrão ajustado para heterocedasticidade e autocorrelação (HAC), considerando uma função kernel uniforme e distância euclidiana, e matriz de pesos espaciais Rainha, selecionada conforme especificação empreendida na Tabela 5.1.

Fonte: Autor

A interpretação do resultado significativo da variável AFD3 se faz da mesma forma que na seção anterior, e o caso da variável AFD4, também significativa, tem-se que em municípios com maior proporção de "docentes com outra formação superior não considerada nas categorias anteriores" (AFD4) do que "docentes com formação superior de licenciatura **na mesma disciplina** que lecionam, ou bacharelado **na mesma disciplina** com curso de complementação pedagógica concluído" (AFD1), a variação da proficiência em Língua Portuguesa foi menor. Parece-nos que esse achado não tem muita relevância prática devido à definição da variável AFD4.

Tabela 6.4: Modelos 2SLS com erros-padrão robustos para a proficiência em língua portuguesa

| POR | **SAR** | **SEM** | **SAC** |
|---|---|---|---|
| Constante | 1,473 | 12,964*** | 0,939 |
| AFD2 | -0,771 | -3,585 | -2,303 |
| AFD3 | -6,295** | -9,589*** | -4,538** |
| AFD4 | -9,094* | -13,577*** | -6,197 |
| AFD5 | -3,363 | -5,253** | -2,614 |
| ICG2 | -0,130 | -0,726 | 0,088 |
| ICG3 | 1,601 | 1,488 | 1,603 |
| ICG4 | 0,588 | -0,684 | -0,761 |
| ICG5 | -0,931 | -1,414 | -0,311 |
| ICG6 | 2,630 | 2,098 | 1,667 |
| IED2 | -2,339 | -1,159 | 1,433 |
| IED3 | 4,445 | 5,987 | 5,819 |
| IED4 | -0,626 | -0,670 | 0,243 |
| IED5 | -0,514 | -2,174 | 1,023 |
| IED6 | 1,272 | 0,993 | 1,950 |
| ALUNO | -0,382** | -0,389** | -0,344*** |
| EXTRA | -0,196 | -0,252 | -0,147 |
| INTRA | 0,024 | 0,008 | 0,014 |
| RESP_M | 10,031 | 19,083 | 13,238 |
| $\rho$ | 0,875*** | | 0,912*** |
| $\lambda$ | | 0,389*** | -0,689*** |
| Nº Observações: | 4661 | 4661 | 4661 |
| Pseudo $R^2$ (%): | 11,680% | 1,860% | 11,570% |
| **Teste Diagnóstico** | **Valor/DF** | **Estatística** | **p-valor** |
| Anselin-Kelejian: | 1 | 83,865 | 0,001 |

* p-valor < 0,1; ** p-valor < 0,05; *** p-valor < 0,01. SAR = modelo de defasagem espacial estimado por Mínimos Quadrados Ordinários em Dois Estágios (2SLS) com erros-padrão ajustado para heterocedasticidade e autocorrelação (HAC); SEM = modelo de erro autoregressivo espacial estimado pelo método 2SLS espacial generalizado de Kelejian e Prucha; SAC = modelo de defasagem espacial com erro autoregressivo espacial estimado pelo método 2SLS espacial generalizado de Kelejian e Prucha. Para ajuste da heterocedasticidade nos modelos 2SLS considerou-se uma função kernel uniforme e distância euclidiana. A matriz de pesos espaciais utilizadas foi do tipo rainha, conforme especificação empreendida na Tabela 5.1.

Fonte: Autor

No entanto, a significância da variável ALUNO no modelo SAR (Tabela 6.4) e, diga-se de passagem, em todos outros modelos, inclusive nos modelos SEM e SAC da proficiência em Matemática (Tabela 6.2), com significância marginal (p-valor < 0,15) no modelo SAR (Tabela 6.2), traz implicações práticas interessantes, uma vez em todos os modelos o coeficiente da variável ALUNO foi negativo e com valores parecidos, indicando estabilidade dos resultados. Assim, num primeiro momento, podemos concluir que em municípios nos quais a percepção docente sobre a indisciplina dos alunos (Q81) e alto nível de faltas (Q82) é maior, a proficiência em Língua Portuguesa é menor. Isso também acontece no caso da proficiência em Matemática, mas essa avaliação é marginalmente significativa. Esses achados sobre a percepção docente coadunam com os de Vidal & Vieira (2017), que colocam que os fatores externos à escola são mais relevantes que os fatores internos nas deficiências de aprendizagem.

Devido os *spillovers* de realimentação o valor de -0,382 do parâmetro da variável ALUNO no modelo SAR deve ser avaliado em termos de efeito direto (ED)[10], efeito indireto (EI)[11] ou efeito total (ET)[12], como propõe Golgher (2015). Assim, ED = -1,630, EI = -1,426 e ET = -3,056. Considerando apenas ET, temos que em municípios que variou um desvio-padrão acima da média dos outros municípios no escore do componente ALUNO, indicando aumento da percepção docente da indisciplina e do nível de falta pelos alunos, a proficiência em Língua Portuguesa reduziu aproximadamente 3 pontos.

[10] $[\beta_j / n(1 - \rho^2)] \times (n - \rho^2)$
[11] $[\beta_j / n(1 - \rho^2)] \times (n\rho + \rho^2)$
[12] ED + EI

## 7. Conclusões

Estudos recentes (Teixeira, 2020; Vidal et al., 2019; Vidal & Vieira, 2017; Xavier & Oliveira, 2020) exibem evidências sobre a importância das expectativas docentes para resultados sociopsicológicos, comportamentais e, sobretudo, nos resultados de desempenho escolar. Alguns (Teixeira, 2020) apontam que, além de as expectativas cumprirem o papel de ser um processo escolar eficaz, ela é o processo escolar com maior potencial e relevância face a todos os outros processos escolares e tem alto poder de contrabalancear influências externas perversas sobre a qualidade do aprendizado discente, como por exemplo, características relacionadas ao *background* familiar e níveis socioeconômicos do aluno.

Nessa medida, outras evidências empíricas (Tekwe et al., 2004; Thieme et al., 2016; Wodtke et al., 2011) colocam os fatores extraescolares, tais como variáveis contextuais além do controle da escola e socioeconômicas (baixa renda, alta pobreza, desemprego, famílias numerosas, chefiadas por mulheres e poucos adultos com nível escolar médio ou superior), como a principal influência nos resultados obtidos pelos alunos em testes de avaliação. Assim, ao juntar essas duas lógicas, outras evidências empíricas (Vidal et al., 2019; Vidal & Vieira, 2017) concluem que as percepções docentes sobre os problemas de aprendizagem dos alunos estão mais relacionadas a fatores externos à escola, como meio social e a situação familiar e econômica.

O presente estudo buscou contribuir com essa discussão ao propor uma abordagem espacial (ecológica), com intuito de expurgar a influência desses fatores externos (à escola), que não mudam (ou mudam lentamente) com o tempo (efeitos fixos), sobre o desempenho em matemática e língua portuguesa de escolares do 9º ano no Brasil. Nesse sentido, cremos que avançamos nessa temática, porque até o presente momento a literatura de desempenho escolar tem utilizado microdados de avaliação da aprendizagem ao nível individual (Jesus & Laros, 2004; Laros & Marciano, 2010; Teixeira, 2020; Vidal et al., 2019; Xavier & Oliveira, 2020), e quando a pretensão residiu em relacionar a expectativa docente com o desempenho escolar no nível agregado dos munícipios (Cavalcanti et al., 2020; Fujita et al., 2021; Vernier, 2016), mesmo seguindo uma abordagem espacial, não houve um controle efetivo dos efeitos fixos que não mudam com tempo, e em muito, podem explicar o desempenho escolar (Teixeira, 2020). Assim,

cremos que inovamos ao atacarmos essas duas lacunas: uma abordagem espacial com controle dos fatores que não mudam com o tempo.

Assim como outros estudos espaciais (Cavalcanti et al., 2020; Fujita et al., 2021; Vernier, 2016), que encontraram forte dependência espacial nos resultados de proficiência em matemática e língua portuguesa do SAEB, também foi sugerido na presente pesquisa que a dimensão espacial influencia o desempenho escolar. Pontualmente, nossos achados indicaram que municípios com maior proporção de docentes com formação superior na mesma disciplina que lecionam possuem escores médios maiores na proficiência em matemática e português e que, em municípios nos quais a percepção docente sobre a indisciplina dos alunos e alto nível de faltas é maior, o desempenho discente é menor.

Nesse sentido, se acredita em algumas constatações da literatura de psicologia e educação (Alvidrez & Weinstein, 1999; Palardy, 1969; Rosenthal & Jacobson, 1968; Teixeira, 2020; Vidal et al., 2019; Xavier & Oliveira, 2020), frente às novas evidências encontradas, que as chances de sucesso no desempenho escolar também podem estar diretamente relacionadas com as expectativas docentes sobre os alunos (profecia autorrealizadora). Assim, os gestores escolares e formadores de políticas públicas devem ficar atentos para um processo de isenção profissional da função última e primordial da escola, que é zelar pelo ensino aprendizagem (Vidal & Vieira, 2017).

Uma visão pessimista sobre o aluno, por exemplo, pode estar arraigada na percepção docente e repercutir na sua prática, conduzindo a um baixo desempenho escolar (Vidal & Vieira, 2017). Assim, a melhoria de desempenho dos escolares brasileiros pode estar vinculada a uma nova postura dos professores diante do processo de ensino e aprendizagem, pois manifestar comportamentos (mesmo que involuntários) que serão absorvidos por seus alunos, pode influenciar a capacidade de desempenho dos mesmos (Teixeira, 2020). Se os docentes brasileiros percebem o desempenho dos alunos como dependentes de variáveis extraescolares, sendo irrelevante ou de menor impacto as variáveis intraescolares no âmbito da gestão escolar e ação docente, prevalece a visão de que as variáveis extraescolares definem o destino da trajetória escolar do aluno e de que sobre isso não há muito o que fazer. É essa postura que os gestores escolares e formadores de políticas públicas devem combater.

Em termos de limitações do estudo, não podemos extrapolar os resultados para o nível individual, correndo o risco de se cair na falácia ecológica. Ademais, apesar de

termos controlados os fatores que não mudam com o tempo, a partir da diferenciação das variáveis da pesquisa, outras variáveis de controle, que mudam com o tempo, poderiam ser consideradas, tal qual em Teixeira (2020), Cavalcanti et al. (2020) e Xavier e Oliveira (2020). Adicionalmente, apesar de termos uma justificativa teórica para não se colocar os valores das variáveis independentes defasadas espacialmente, poderíamos ter seguido a linha de Fujita et al. (2021), que justificaram a inclusão de tais variáveis em termos de controle para variáveis omitidas, sendo algumas, significativas.

Assim, enquanto perspectivas futuras para pesquisadores que desejam abraçar o mesmo tema, e avançar na relação entre desempenho discente e expectativas docentes, orientamos utilizar um período maior dos resultados do SAEB, e assim, com mais pontos no tempo, utilizar modelos de dados em painel espacial, sem necessidade, porventura, da diferenciação das variáveis e consequente perda da memória de longo prazo das séries. Adicionalmente, outras variáveis de controle poderiam ser pensadas, assim como, através delas, a incorporação de variáveis independentes defasadas espacialmente, e assim, outros modelos espaciais considerados.

**BIBLIOGRAFIA**


Almeida, E. (2012). *Econometria Espacial Aplicada*. Editora Alínea.

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology*, *91*(4). https://doi.org/10.1037/0022-0663.91.4.731

Anselin, L. (1988). *Spatial econometrics: methods and models* (Vol. 4). Kluwer Academic Publishers.

Anselin, L. (2005). *Exploring Spatial Data with GeoDa: A Workbook*. Center for Spatially Integrated Social Science. https://geodacenter.github.io/documentation.html

Anselin, L., & Rey, S. J. (2014). *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*, *38*(1), 5–22.

Aranha, F., & Zambaldi, F. (2008). *Análise Fatorial em Administração*. Cengage Learning.

Arbia, G. (2014). *A Primer for Spatial Econometrics with Applications in R*. Palgrave Macmillan.

Baumont, C., Ertur, C., & Gallo, J. (2004). Spatial Analysis of Employment and Population Density: The Case of the Agglomeration of Dijon 1999. *Geographical Analysis*, *36*(2), 146–176. https://doi.org/10.1111/j.1538-4632.2004.tb01130.x

Bhattacharjee, A., & Jensen-Butler, C. (2013). Estimation of the spatial weights matrix under structural constraints. *Regional Science and Urban Economics*, *43*(4), 617–634. https://doi.org/10.1016/j.regsciurbeco.2013.03.005

Bivand, R., Millo, G., & Piras, G. (2021). A Review of Software for Spatial Econometrics in R. *Mathematics*, *9*(1276). https://doi.org/10.3390/math9111276

Brooke, N.; Soares, F. S. Pesquisa em Eficácia Escolar: origem e trajetórias. Belo Horizonte: UFMG, 2008.

Carvalho, A. X. Y., & Albuquerque, P. H. M. (2011). Tópicos em econometria espacial para dados cross-section. In B. de O. Cruz, B. A. Furtado, L. Monasterio, & W. Rodrigues Júnior (Eds.), *Economia regional e urbana: teorias e métodos com ênfase no Brasil*. Instituto de Pesquisa Econômica Aplicada.

Cavalcanti, G. da S., André, D. de M., & Araújo, J. R. (2020). Transbordamentos Espaciais da Educação nos Municípios Brasileiros. *XVIII Encontro Nacional Da Associação Brasileira de Estudos Regionais e Urbanos (Enaber)*.

Cliff, A., & Ord, J. K. (1981). *Spatial processes, models and applications*. Pion.

Corrado, L., & Fingleton, B. (2012). Where is the economics in spatial econometrics? *Journal of Regional Science*, *52*(2), 210–239. https://doi.org/10.1111/j.1467-9787.2011.00726.x

Cunha, J. M. P. da, Jimenez, M. A., Perez, J. R. R., & Andrade, C. Y. de. (2009). Social segregation and academic achievement in state-run elementary schools in the municipality of Campinas, Brazil. *Geoforum*, *40*(5), 873–883. https://doi.org/10.1016/j.geoforum.2009.06.003

Darmofal, D. (2006). *Spatial econometrics and political science*.

Ferreira, D. F. (2018). *Estatística Multivariada* (3rd ed.). Editora UFLA.

Florax, R. J. G. M., & Rey, S. (1995). *The Impacts of Misspecified Spatial Interaction in Linear Regression Models* (pp. 111–135). https://doi.org/10.1007/978-3-642-79877-1_5

Fotheringham, A. S., Brunsdon, C., & Martin, C. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Fotheringham, A. S., & Wong, D. W. S. (1991). The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. *Environment and Planning A: Economy and Space*, *23*(7), 1025–1044. https://doi.org/10.1068/a231025

Fujita, L. D. V., Bagolin, I. P., & Fochezatto, A. (2021). Spatial distribution and dissemination of education in Brazilian municipalities. *Annals of Regional Science*, *66*(2), 255–277. https://doi.org/10.1007/s00168-020-01020-3

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, *5*(3), 115. https://doi.org/10.2307/2986645

Gelman, A., Park, D. K., Ansolabehere, S., Price, P. N., & Minnite, L. C. (2001). Models, assumptions and model checking in ecological regressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *164*(1), 101–118. https://doi.org/10.1111/1467-985X.00190

Golgher, A. B. (2015). *Introdução à Econometria Espacial*. Paco Editorial.

Goodchild, M. (2004). The validity and usefulness of laws in geographic information science and geography. *Annuals of the Association of American Geographers*, *94*(2), 300–303.

Griffith, D. A. (1983). The boundary value problem in spatial statistical analysis. *Journal of Regional Science*, *23*(3), 377–387. https://doi.org/10.1111/j.1467-9787.1983.tb00996.x

Haining, R. (2003). *Spatial data analysis: theory and practice*. Cambridge University Press.

Hair, J., Black, W., Babin, B., & Anderson, R. (2014). *Multivariate Data Analysis* (7th ed.). Pearson Education Limited.

Instituto Reúna: *Avaliações em larga escala no Brasil e no mundo*. Uma análise compara da de 14 experiências. https://www.institutoreuna.org.br/conteudo/avaliacoes-em-larga-escala-no-brasil-e-no-mundo

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. In *Educational and Psychological Measurement* (Vol. 66, Issue 3, pp. 393–416). SAGE Publications Inc. https://doi.org/10.1177/0013164405282485

Izquierdo, I., Olea, J., & Abad, F. J. (2014). El análisis factorial exploratorio en estudios de validación: Usos y recomendaciones. *Psicothema*, *26*(3), 395–400. https://doi.org/10.7334/psicothema2013.349

Jesus, G. R., & Laros, J. A. (2004). Eficácia Escolar: Regressão Multinível com Dados de Avaliação em Larga Escala. *Avaliação Psicológica*, *3*(2), 93–106.

Kelijian, H. H., & Prucha, I. R. (1998). *A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances*.

Kelijian, H. H., & Prucha, I. R. (1999). A Generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review*, *4*(2).

Kooijman, S. A. L. M. (1976). Some Remarks on the Statistical Analysis of Grids Especially with Respect to Ecology. In *Annals of Systems Research* (pp. 113–132). Springer US. https://doi.org/10.1007/978-1-4613-4243-4_6

Kopczewska, K. (2021). *Applied Spatial Statistics and Econometrics: Data Analysis in R*. Routledge.

Laros, J. A., & Marciano, J. L. P. (2010). Fatores que Afetam o Desempenho na Prova de Matemática do SAEB: Um Estudo Multinível. *Avaliação Psicológica*, *9*, 173–186.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*(3), 283–301. https://doi.org/10.1016/S0272-7757(96)00081-7

Moran, P. A. P. (1948). The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society*, *10*(2), 243–251.

Palardy, J. M. (1969). What Teachers Believe-What Children Achieve. *The Elementary School Journal*, *69*(7), 370–374. http://www.jstor.org/stable/1000271

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Americam Sociological Review*, *15*, 351–357.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: teacher expectation and pupil's intellectual development.* Holt, Rhinehat & Winston,.

Sabater, L. A., Tur, A. A., & Azorin, J. M. (2011). Análise Exploratória de Dados Espaciais (AEDE). In T. P. Dentinho, P. Nijkamp, & J. S. Costa (Eds.), *Compendio de Economia Regional* (Vol. 2, pp. 259–293). Principia.

Teixeira, O. H. (2020). *O processo de ensino-aprendizagem e suas relações com as expectativas docentes acerca do desempenho escolar* [Dissertação de Mestrado]. Universidade Federal de Viçosa.

Tekwe, C. D., Carter, R. L., Ma, C.-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics*, *29*(1), 11–36. https://doi.org/10.3102/10769986029001011

Thieme, C., Prior, D., Tortosa-Ausina, E., & Gempp, R. (2016). Value added, educational accountability approaches and their effects on schools' rankings: Evidence from Chile. *European Journal of Operational Research*, *253*(2), 456–471. https://doi.org/10.1016/j.ejor.2016.01.023

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141

Torres, H. da G., Marques, E., Ferreira, M. P., & Bitar, S. (2003). Pobreza e espaço: padrões de segregação em São Paulo. *Estudos Avançados*, *17*(47), 97–128. https://doi.org/10.1590/S0103-40142003000100006

Tyszler, M. (2006). *Econometria Espacial: Discutindo Medidas para a Matriz de Ponderação Espacial* [Dissertação]. Fundação Getúlio Vargas.

Vernier, L. D. S. (2016). *Crescimento Educacional Brasileiro: Uma Análise da Distribuição e Disseminação dos Efeitos Espaciais* [Tese de Doutorado]. Pontifícia Universidade Católica do Rio Grande do Sul.

Vidal, E. M., Galvão, W. N. M., Vieira, S. L., & Chaves, J. B. (2019). Expectativas docentes e aprendizagem: explorando dados do questionário da Prova Brasil 2015. *Educação e Pesquisa*, *45*. https://doi.org/10.1590/s1678-4634201945201657

Vidal, E. M., & Vieira, S. L. (2017). Professores da educação básica: perfil e percepções sobre sucesso dos alunos. *Estudos Em Avaliação Educacional*, *28*(67), 64. https://doi.org/10.18222/eae.v28i67.3936

Vieira, R. de S. (2009). *Crescimento econômico no estado de São Paulo: uma análise espacial*. Editora UNESP.

Wodtke, G. T., Harding, D. J., & Elwert, F. (2011). Neighborhood Effects in Temporal Perspective. *American Sociological Review*, *76*(5), 713–736. https://doi.org/10.1177/0003122411420816

Xavier, F. P., & Oliveira, V. C. de. (2020). Aprendizado, expectativas docentes e relação professor-aluno. Estudos Em Avaliação Educacional, x. https://doi.org/10.18222/eae.v0ix.6487

## APÊNDICES

PCA

Modelo Final 2017

*Communalities*

| | Initial | Extraction |
|---|---|---|
| Q70_17 | 1,000 | ,315 |
| Q71_17 | 1,000 | ,388 |
| Q72_17 | 1,000 | ,389 |
| Q73_17 | 1,000 | ,343 |
| Q74_17 | 1,000 | ,494 |
| Q75_17 | 1,000 | ,524 |
| Q76_17 | 1,000 | ,619 |
| Q77_17 | 1,000 | ,664 |
| Q78_17 | 1,000 | ,436 |
| Q79_17 | 1,000 | ,357 |
| Q81_17 | 1,000 | ,573 |
| Q82_17 | 1,000 | ,529 |

Extraction Method: Principal Component Analysis.

*Total Variance Explained*

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2,694 | 22,452 | 22,452 | 2,694 | 22,452 | 22,452 | 2,201 | 18,346 | 18,346 |
| 2 | 1,785 | 14,874 | 37,326 | 1,785 | 14,874 | 37,326 | 2,030 | 16,917 | 35,262 |
| 3 | 1,153 | 9,607 | 46,933 | 1,153 | 9,607 | 46,933 | 1,400 | 11,671 | 46,933 |
| 4 | ,968 | 8,065 | 54,998 | | | | | | |
| 5 | ,944 | 7,867 | 62,865 | | | | | | |
| 6 | ,796 | 6,635 | 69,500 | | | | | | |
| 7 | ,727 | 6,059 | 75,559 | | | | | | |
| 8 | ,689 | 5,739 | 81,298 | | | | | | |
| 9 | ,663 | 5,525 | 86,823 | | | | | | |
| 10 | ,650 | 5,414 | 92,237 | | | | | | |
| 11 | ,502 | 4,180 | 96,417 | | | | | | |
| 12 | ,430 | 3,583 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

Modelo Final 2013

*Communalities*

| | Initial | Extraction |
|---|---|---|
| Q70_13 | 1,000 | ,317 |
| Q71_13 | 1,000 | ,421 |
| Q72_13 | 1,000 | ,358 |
| Q73_13 | 1,000 | ,341 |
| Q74_13 | 1,000 | ,495 |
| Q75_13 | 1,000 | ,486 |
| Q76_13 | 1,000 | ,632 |

| | | |
|---|---|---|
| Q77_13 | 1,000 | ,677 |
| Q78_13 | 1,000 | ,521 |
| Q79_13 | 1,000 | ,387 |
| Q81_13 | 1,000 | ,566 |
| Q82_13 | 1,000 | ,537 |

Extraction Method: Principal Component Analysis.

*Total Variance Explained*

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 2,768 | 23,066 | 23,066 | 2,768 | 23,066 | 23,066 | 2,182 | 18,186 | 18,186 |
| 2 | 1,830 | 15,249 | 38,316 | 1,830 | 15,249 | 38,316 | 2,176 | 18,135 | 36,321 |
| 3 | 1,140 | 9,497 | 47,812 | 1,140 | 9,497 | 47,812 | 1,379 | 11,491 | 47,812 |
| 4 | 1,020 | 8,502 | 56,315 | | | | | | |
| 5 | ,909 | 7,576 | 63,891 | | | | | | |
| 6 | ,740 | 6,167 | 70,058 | | | | | | |
| 7 | ,716 | 5,963 | 76,021 | | | | | | |
| 8 | ,699 | 5,826 | 81,848 | | | | | | |
| 9 | ,641 | 5,344 | 87,192 | | | | | | |
| 10 | ,606 | 5,047 | 92,238 | | | | | | |
| 11 | ,507 | 4,222 | 96,460 | | | | | | |
| 12 | ,425 | 3,540 | 100,000 | | | | | | |

Extraction Method: Principal Component Analysis.

AEDE

Especificação Matriz W

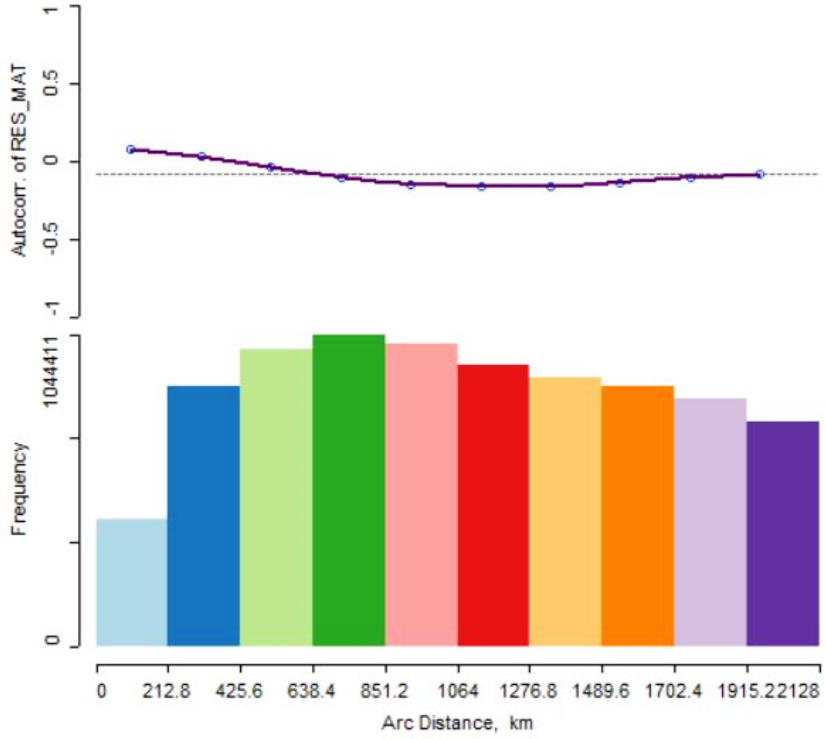


Figura 1 – Correlograma espacial dos resíduos da variável MAT (máx = 2128 km)

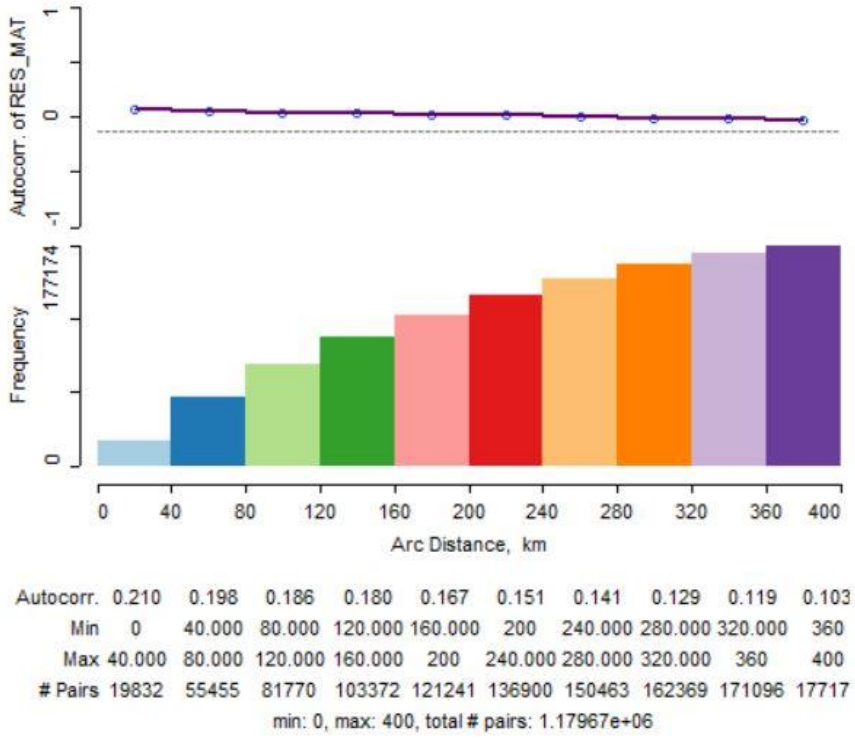


Figura 2 – Correlograma espacial dos resíduos da variável MAT (máx = 400 km)

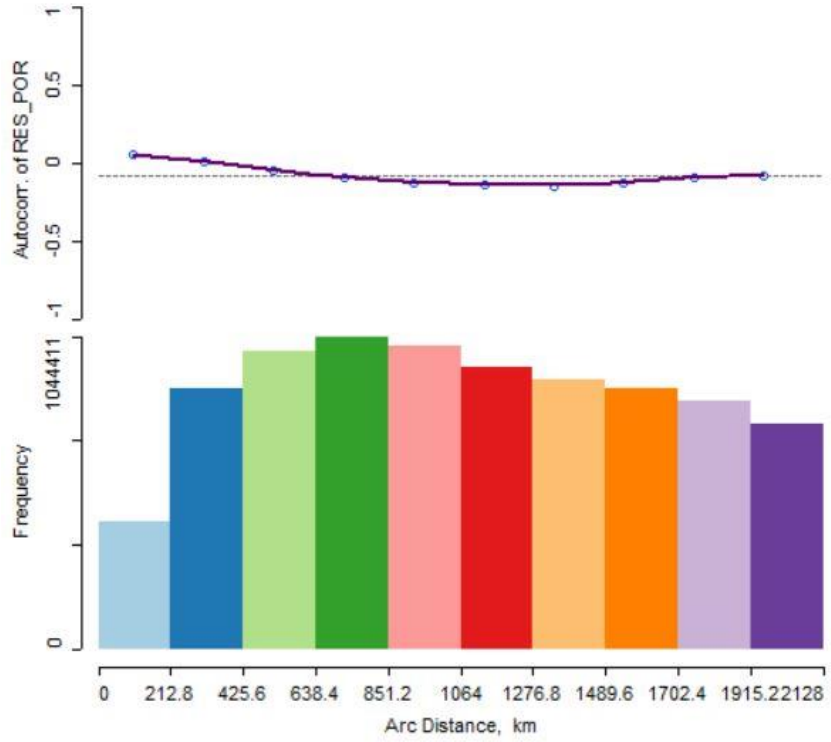


Figura 3 – Correlograma espacial dos resíduos da variável POR (máx = 2128 km)

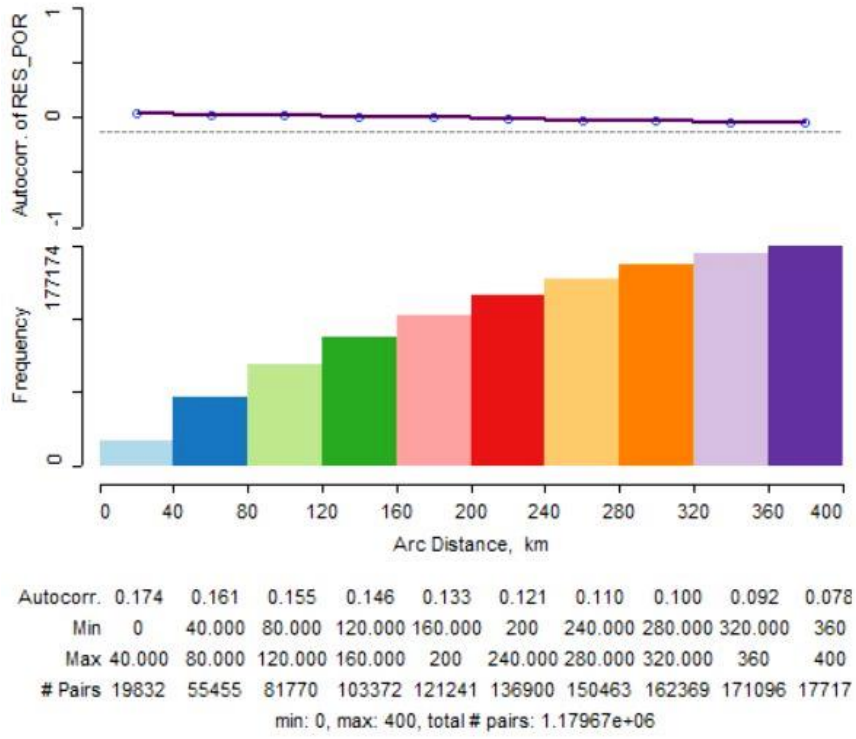


Figura 4 – Correlograma espacial dos resíduos da variável POR (máx = 400 km)

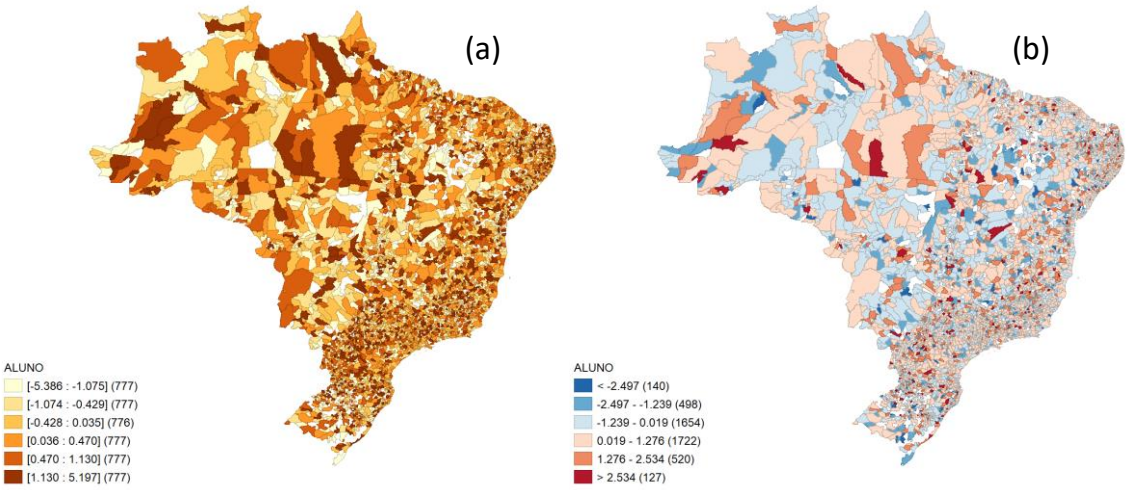


Figura 5 – Mapa quantílico (a) e de desvio-padrão (b) da variação do componente “aluno”

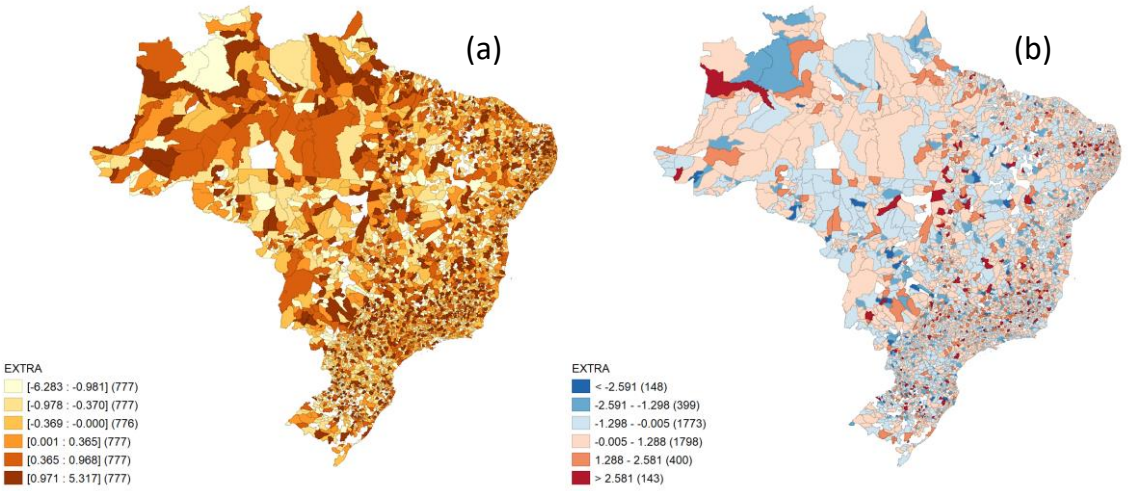


Figura 6 – Mapa quantílico (a) e de desvio-padrão (b) da variação do componente “extraescolar”

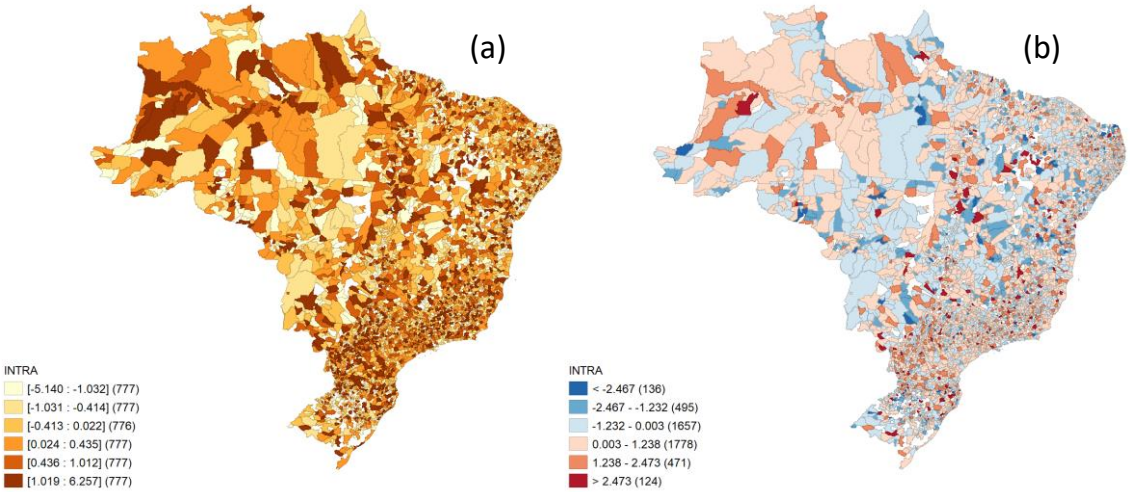


Figura 7 – Mapa quantílico (a) e de desvio-padrão (b) da variação do componente “intraescolar”

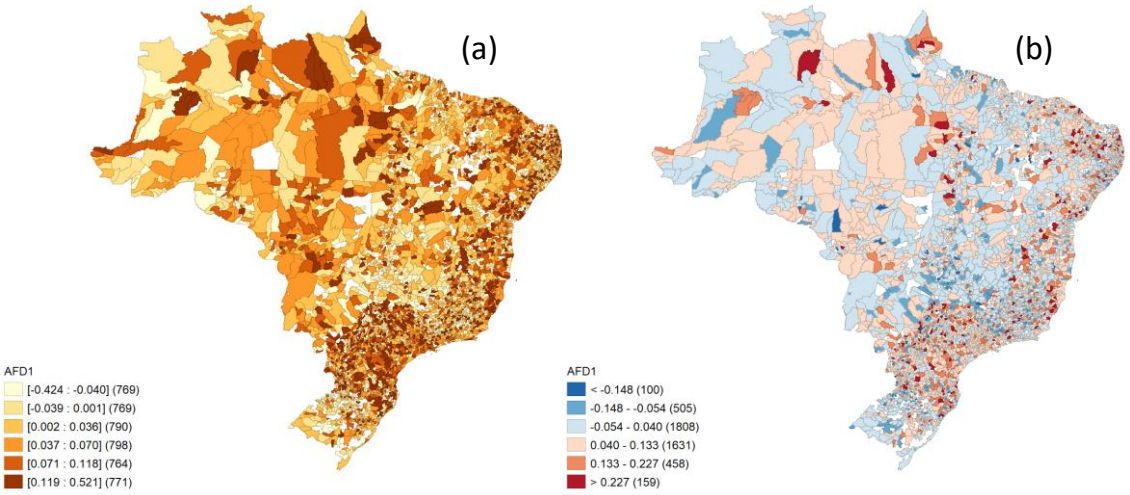


Figura 8 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável AFD1

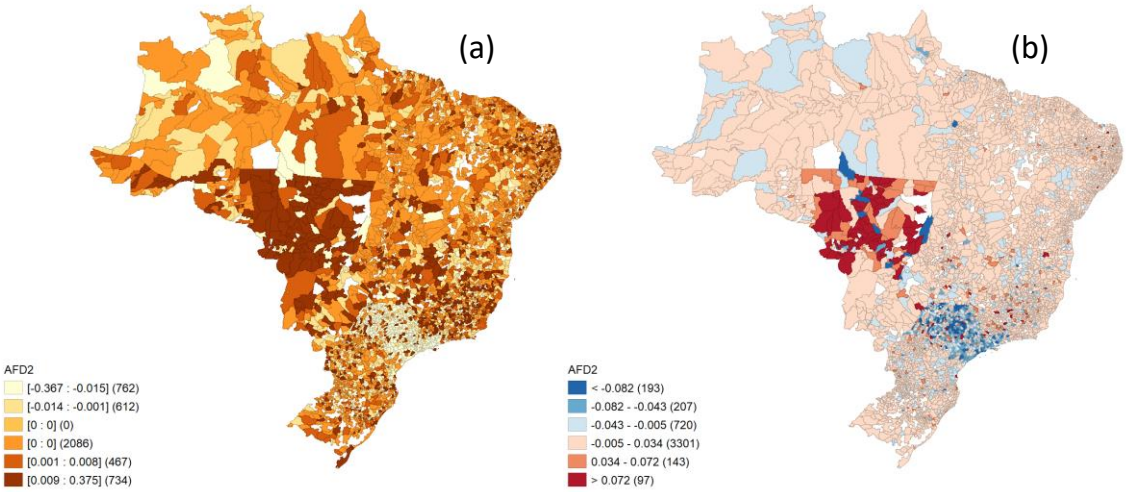


Figura 9 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável AFD2

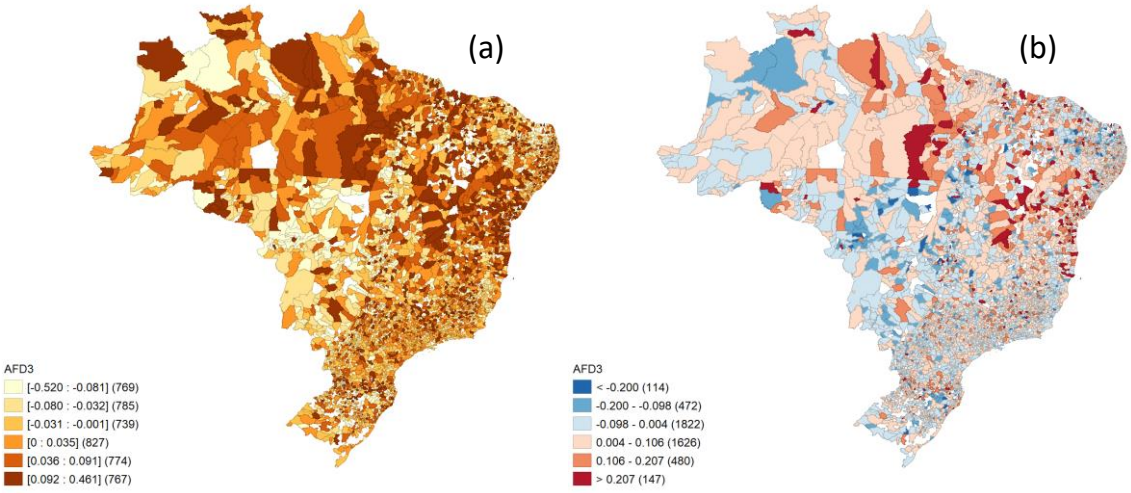


Figura 10 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável AFD3

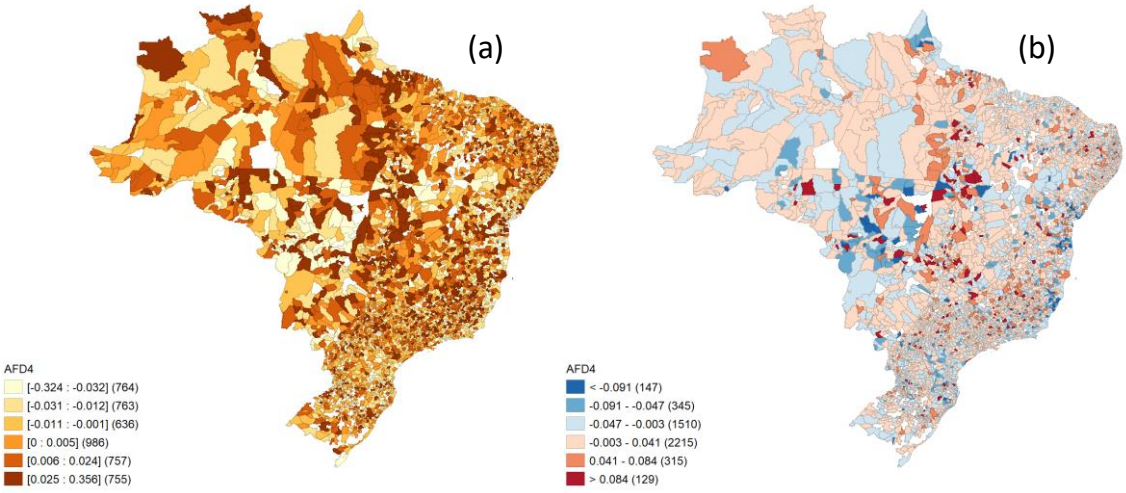


Figura 11 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável AFD4

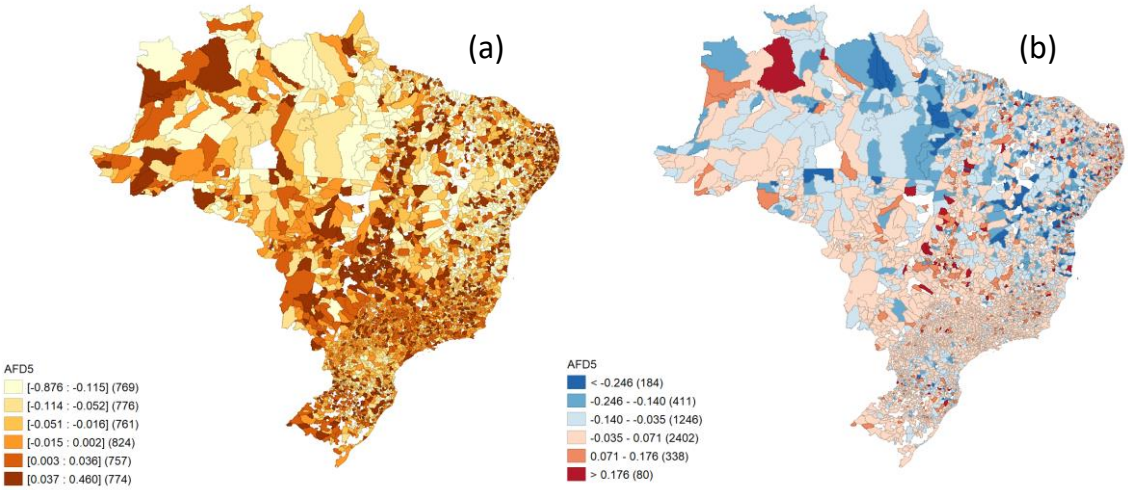


Figura 12 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável AFD5

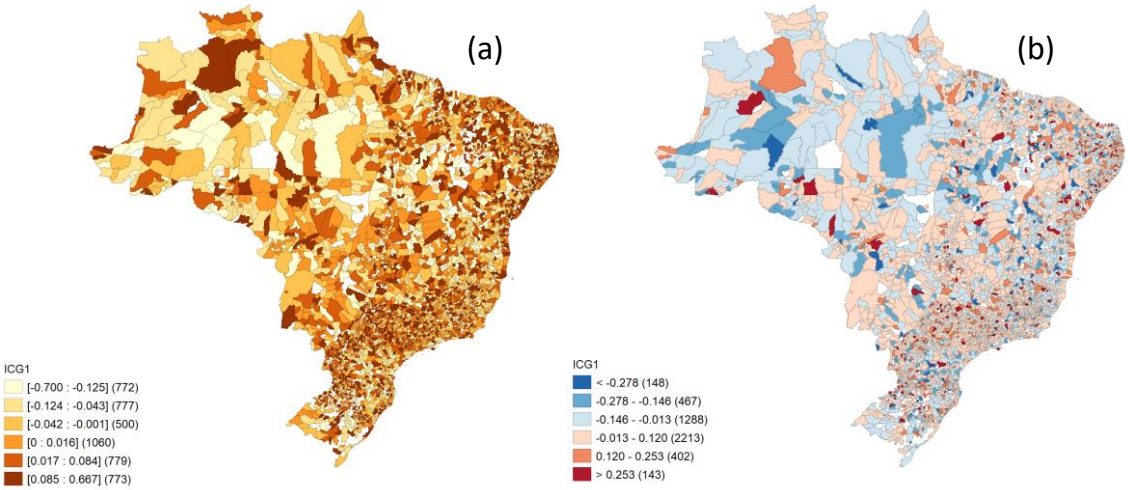


Figura 13 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável ICG1

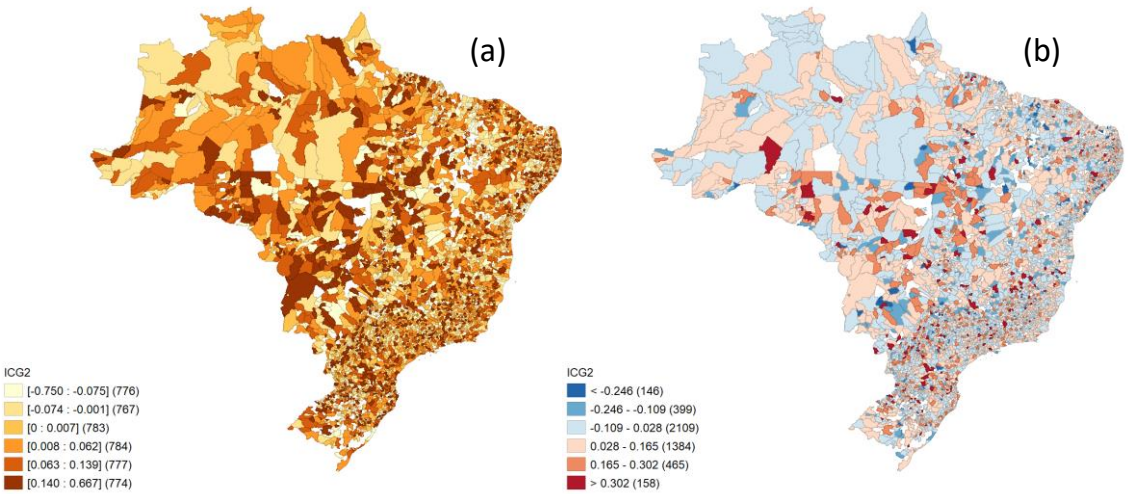


Figura 14 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável ICG2

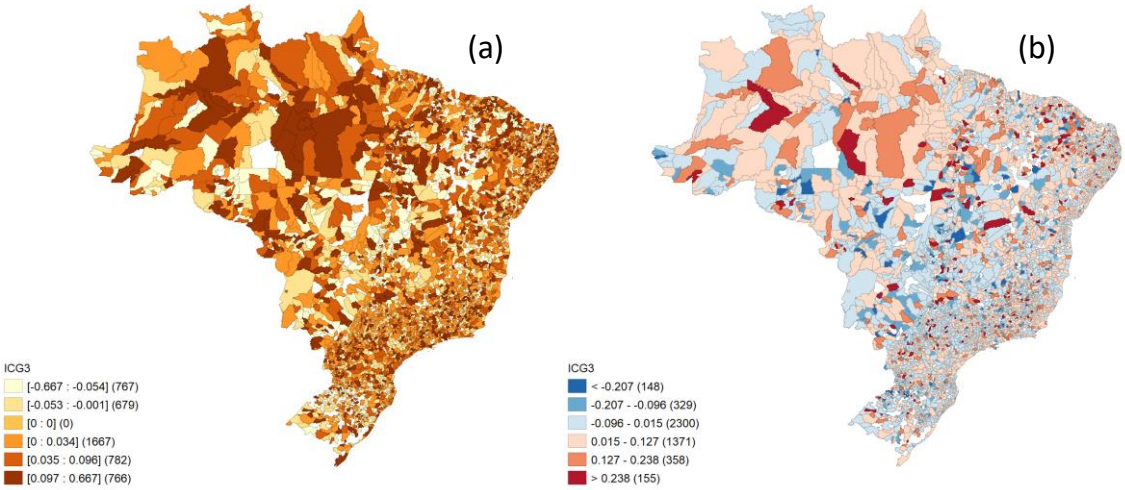


Figura 15 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável ICG3

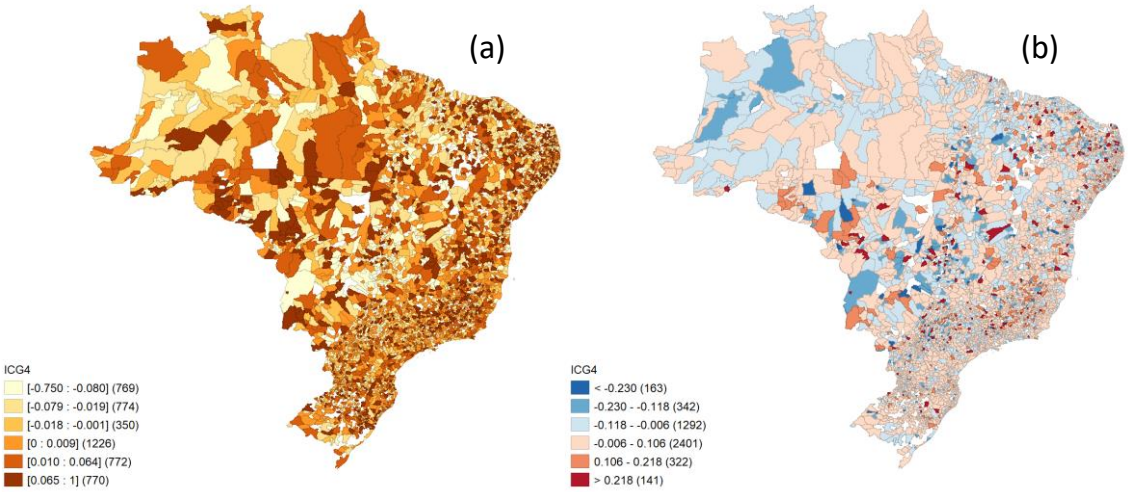


Figura 16 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável ICG4

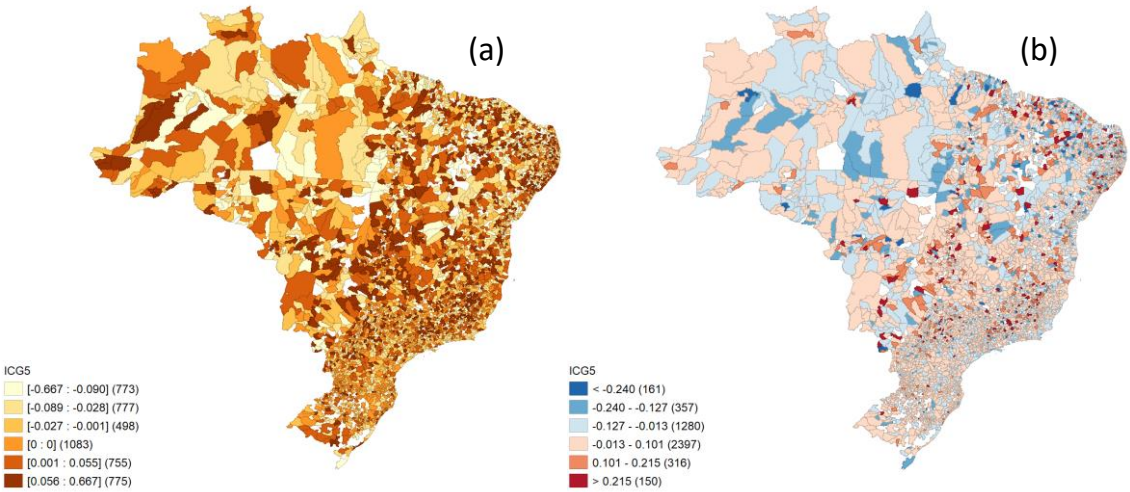


Figura 17 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável ICG5

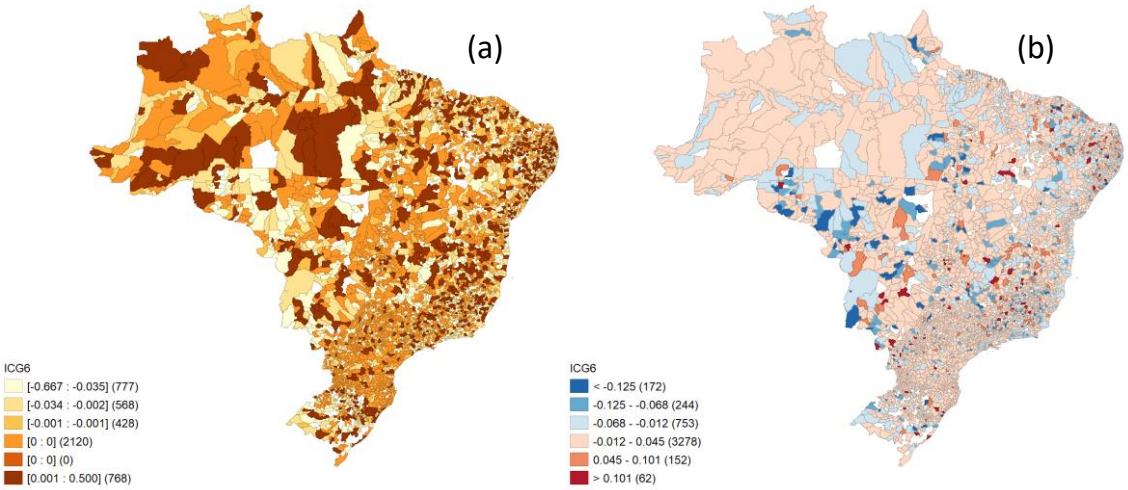


Figura 18 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável ICG6

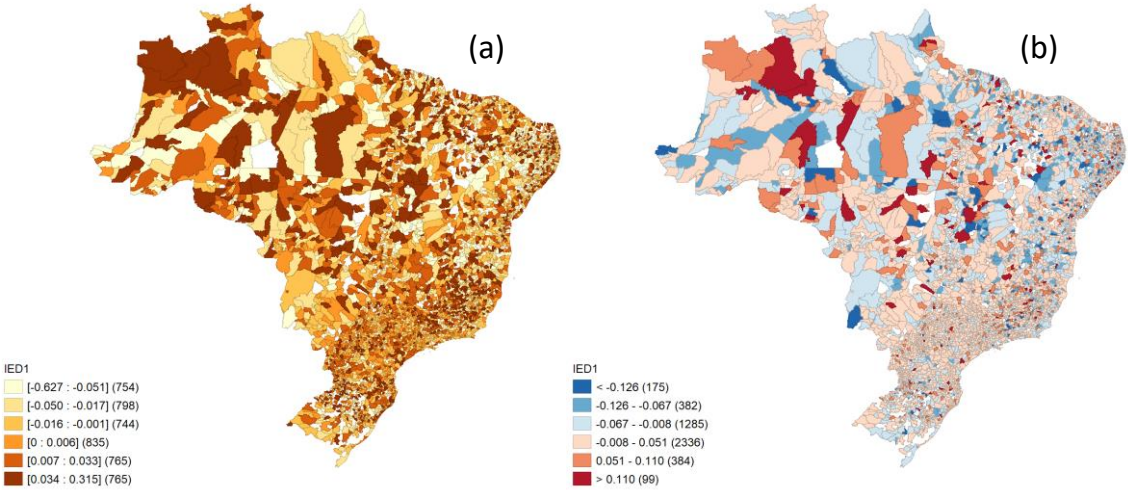


Figura 19 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável IED1

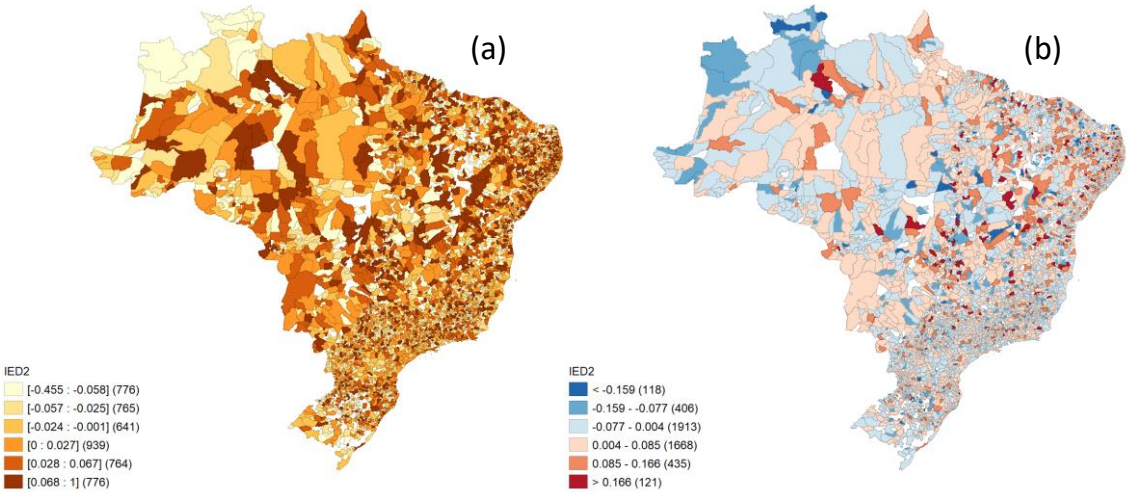


Figura 20 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável IED2

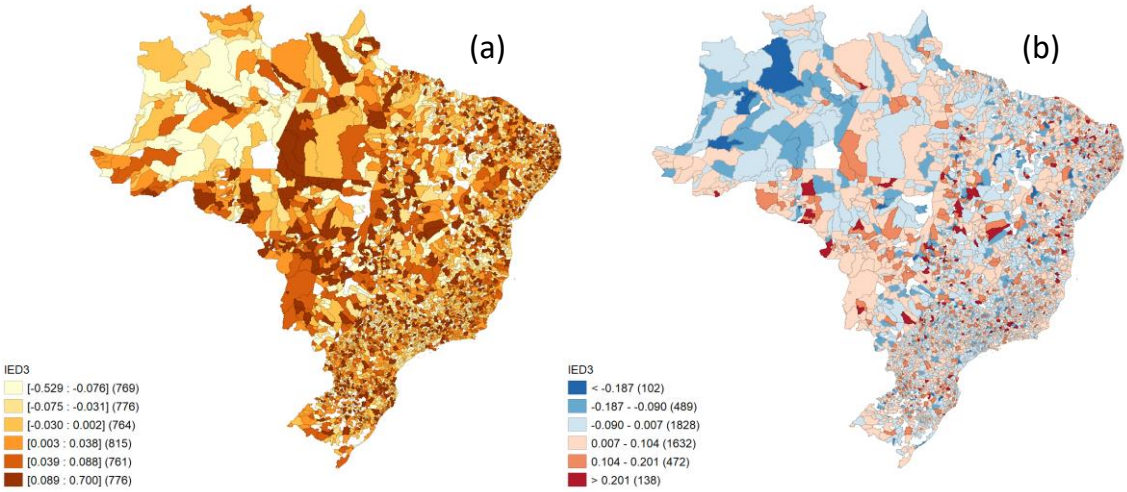


Figura 21 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável IED3

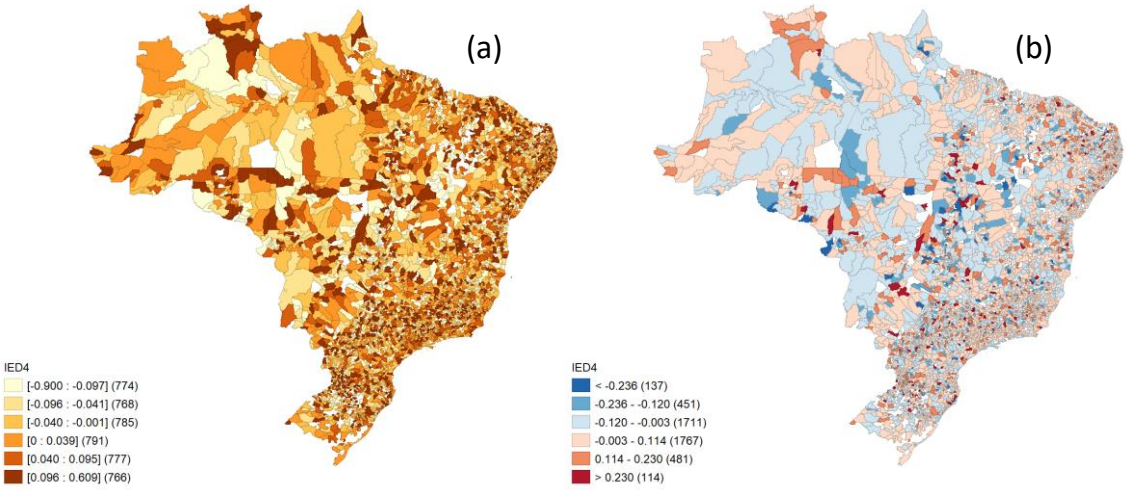


Figura 22 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável IED4

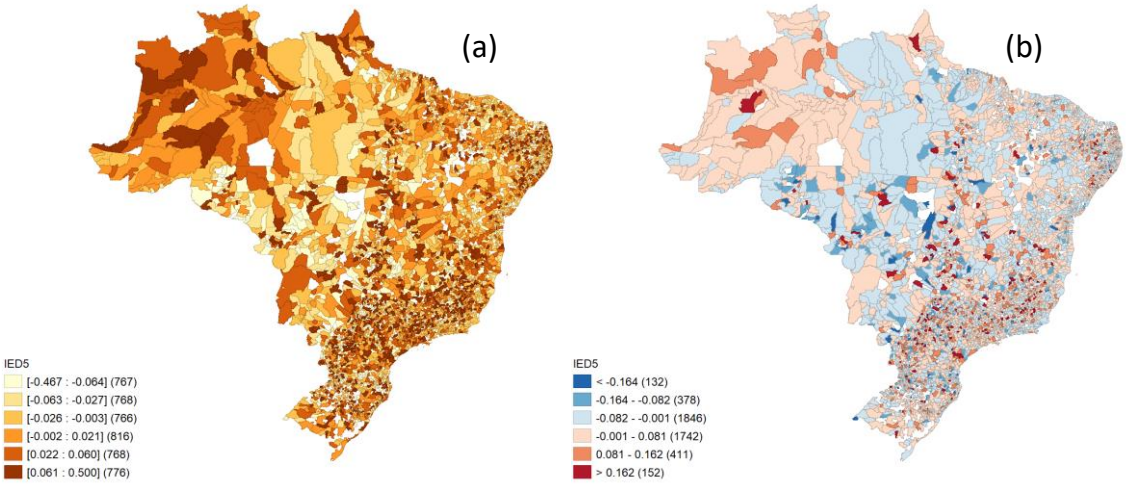


Figura 23 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável IED5

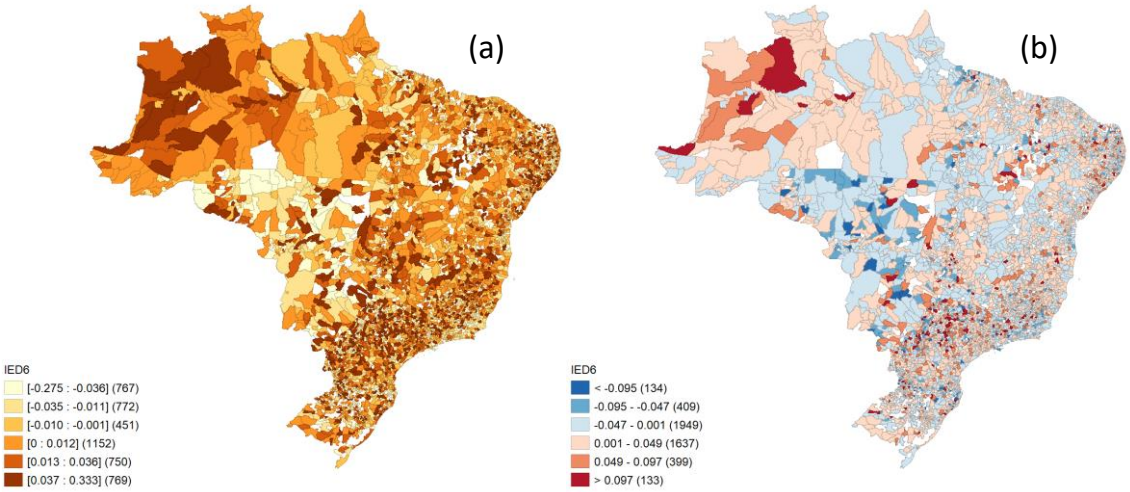


Figura 24 – Mapa quantílico (a) e de desvio-padrão (b) da variação da variável IED6

Mapas LISA[13]

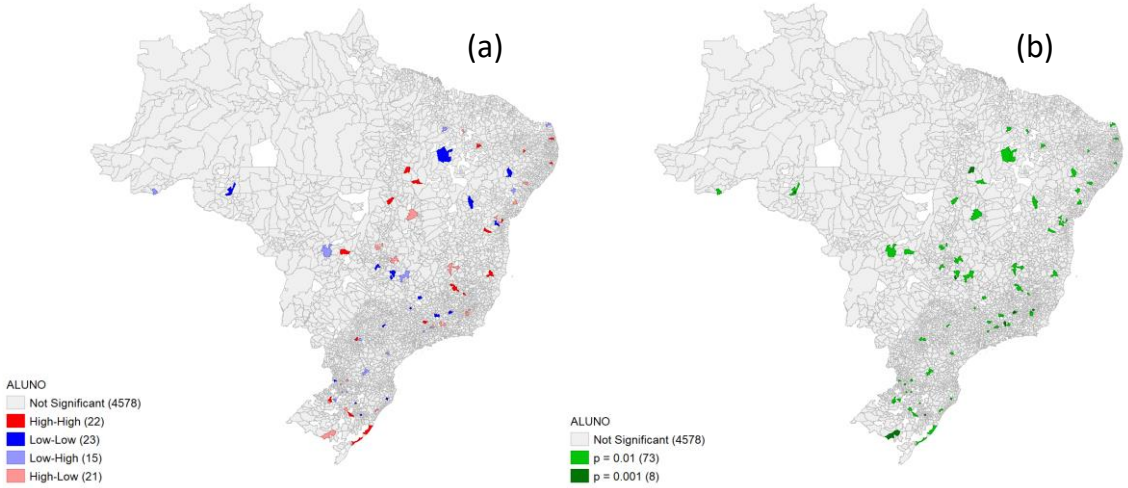


Figura 25 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ALUNO

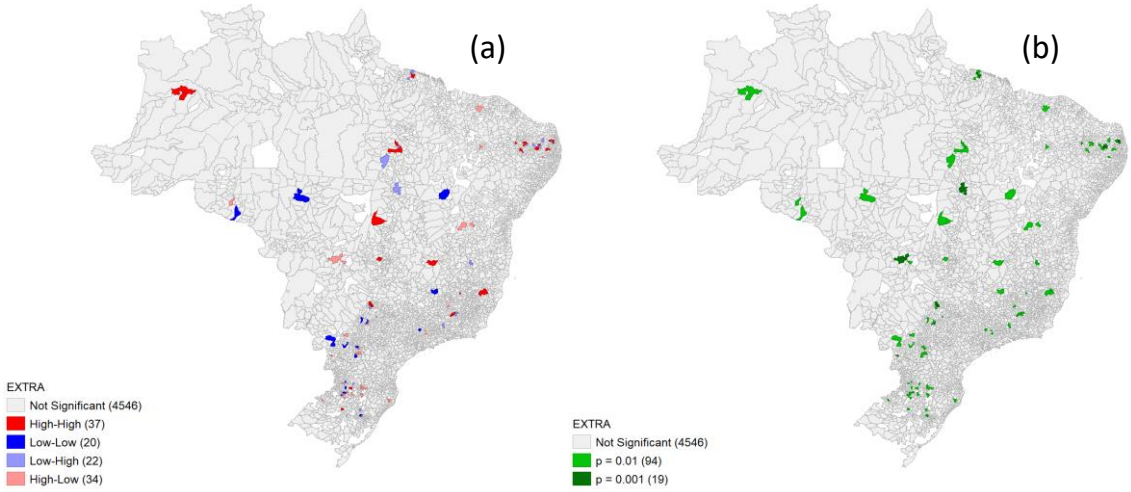


Figura 26 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável EXTRA

[13] *Local Indicator of Spatial Association* (I de Moran local)

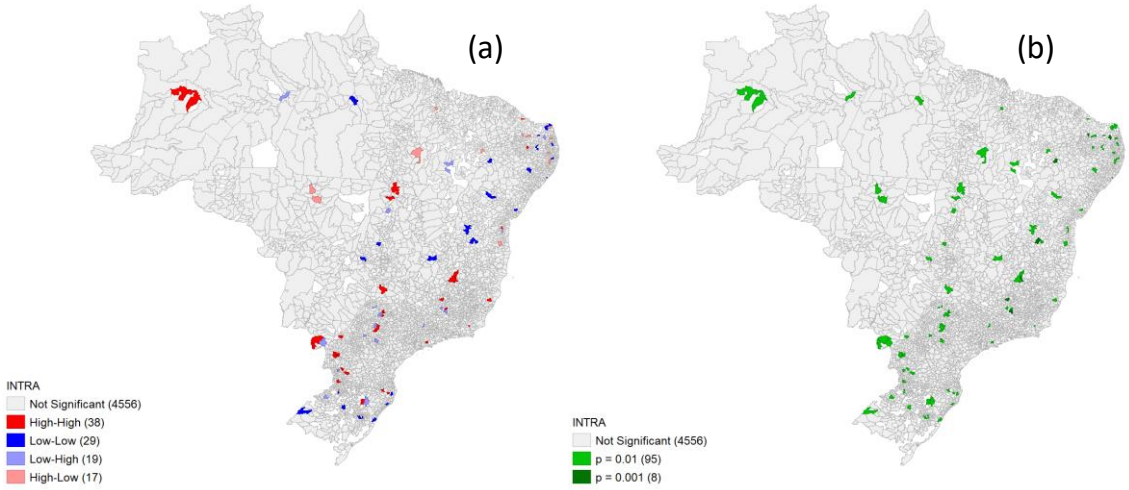


Figura 27 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável INTRA

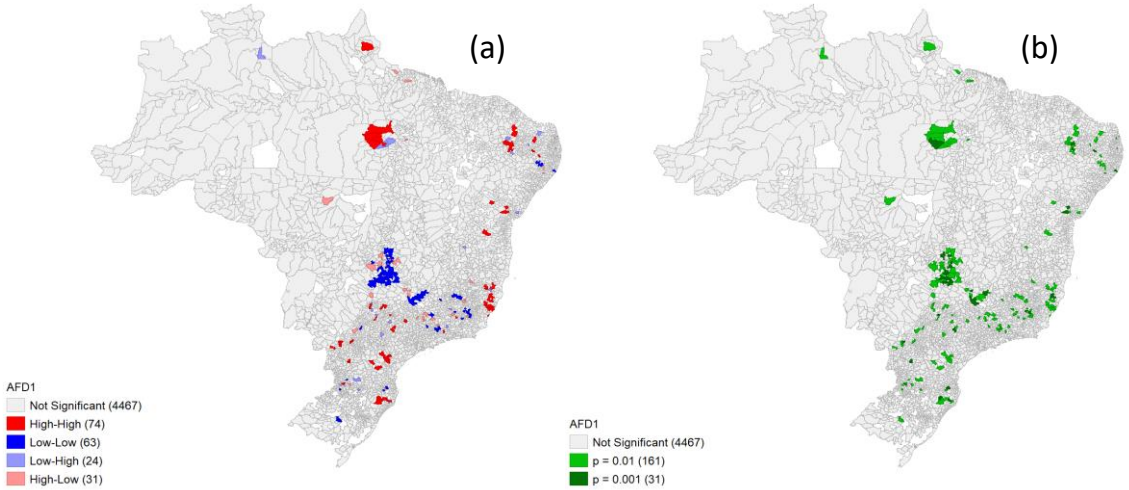


Figura 28 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável AFD1

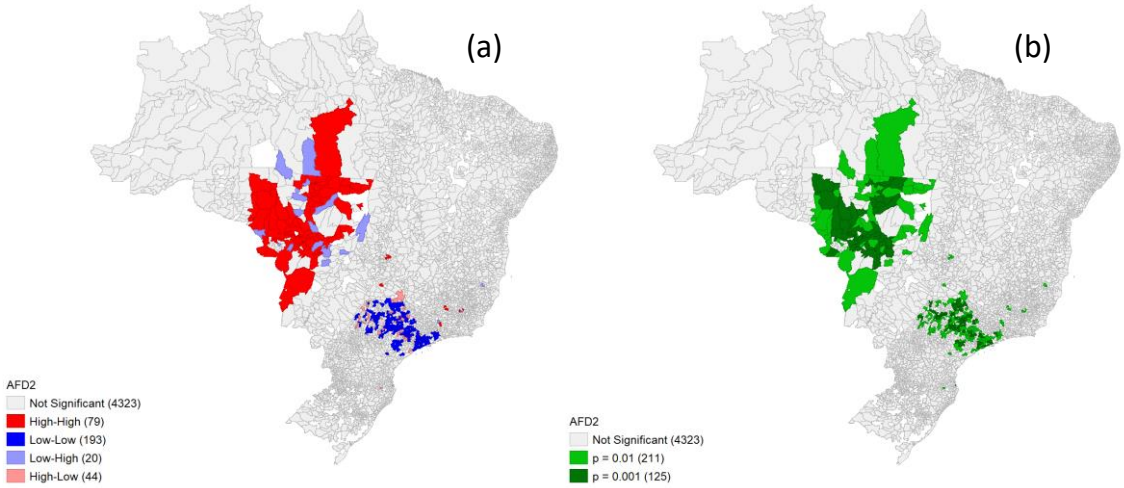


Figura 29 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável AFD2

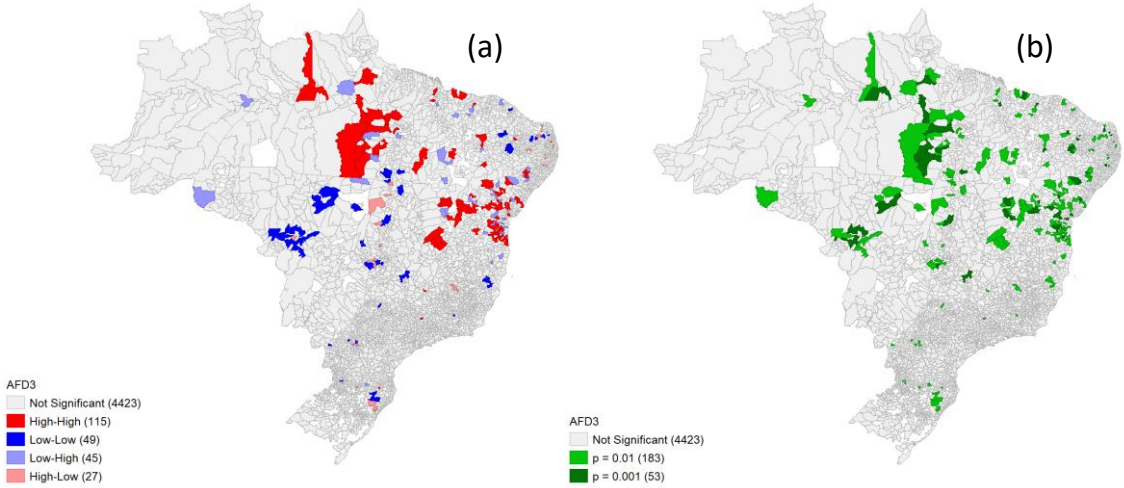


Figura 30 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável AFD3

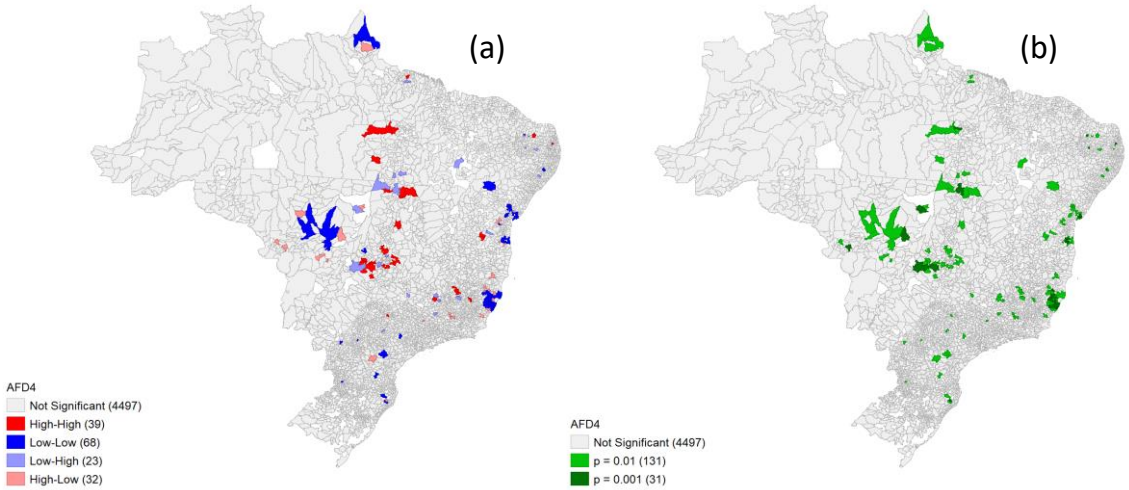


Figura 31 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável AFD4

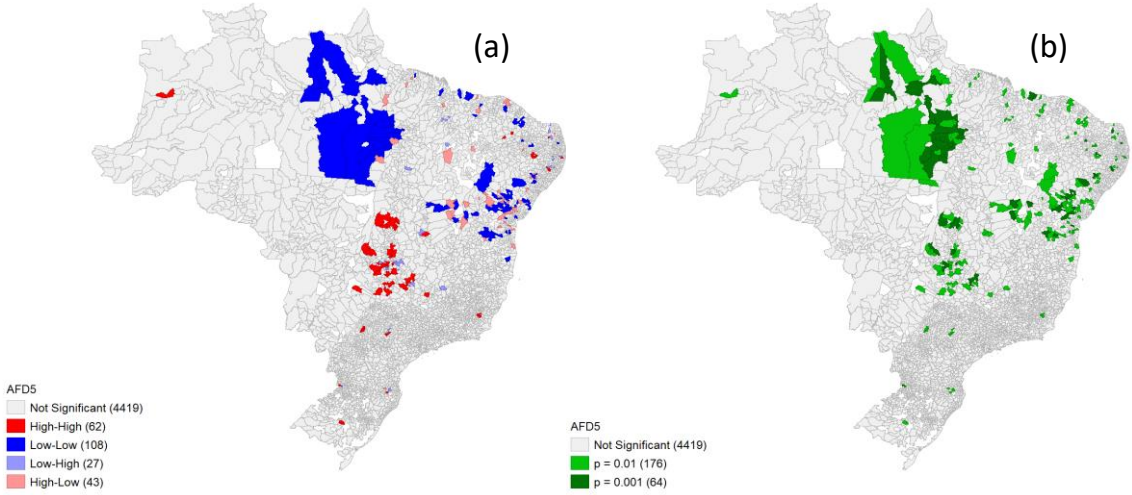


Figura 32 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável AFD5

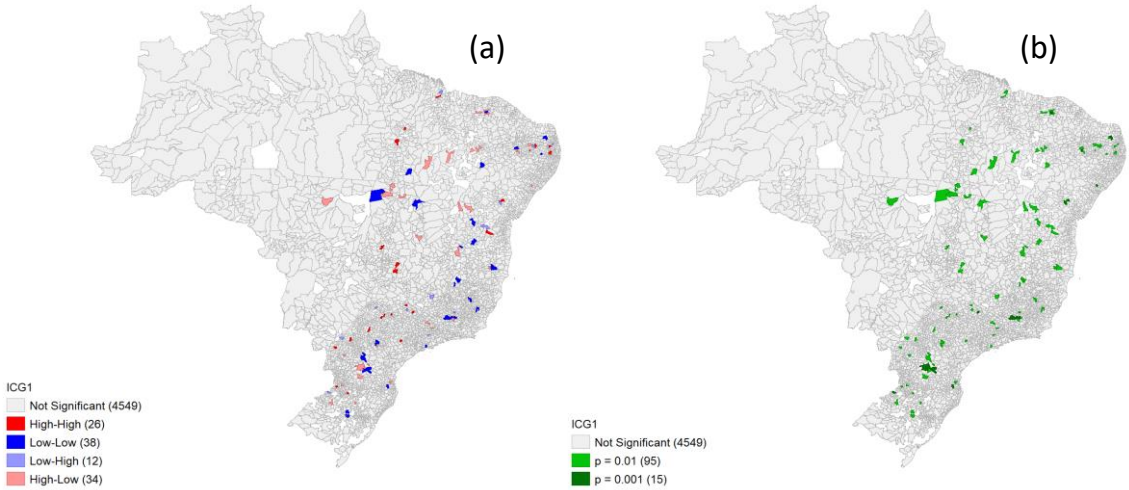


Figura 33 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ICG1

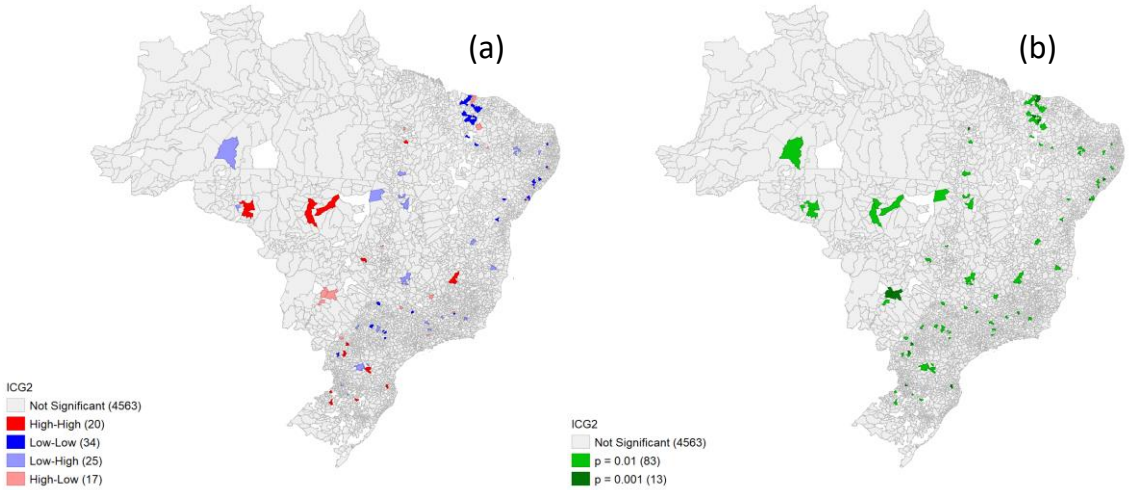


Figura 34 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ICG2

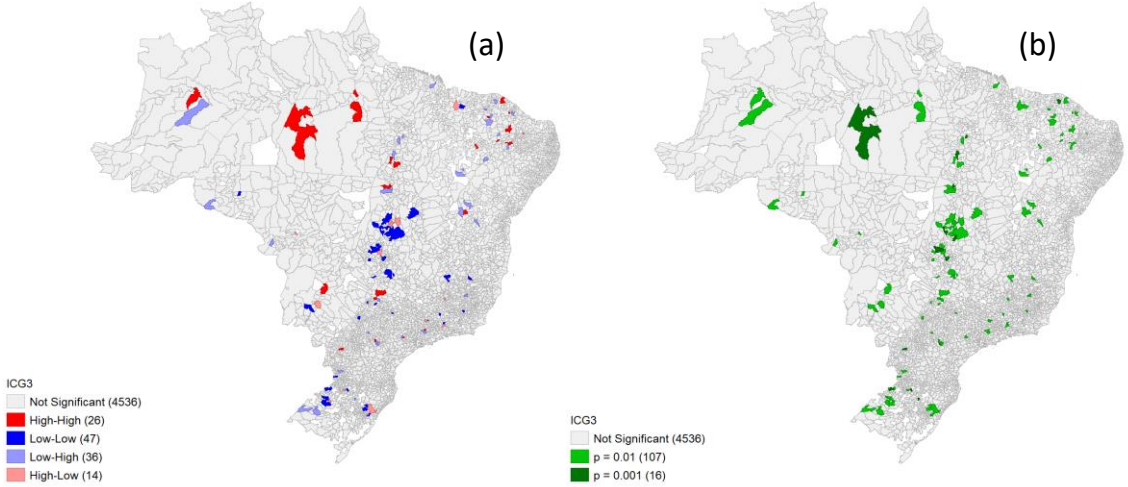


Figura 35 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ICG3

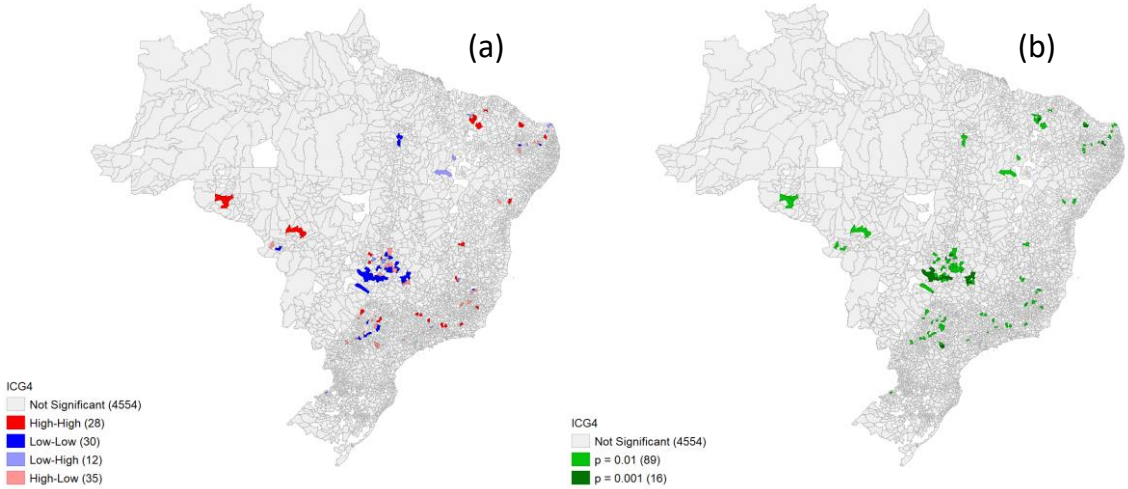


Figura 36 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ICG4

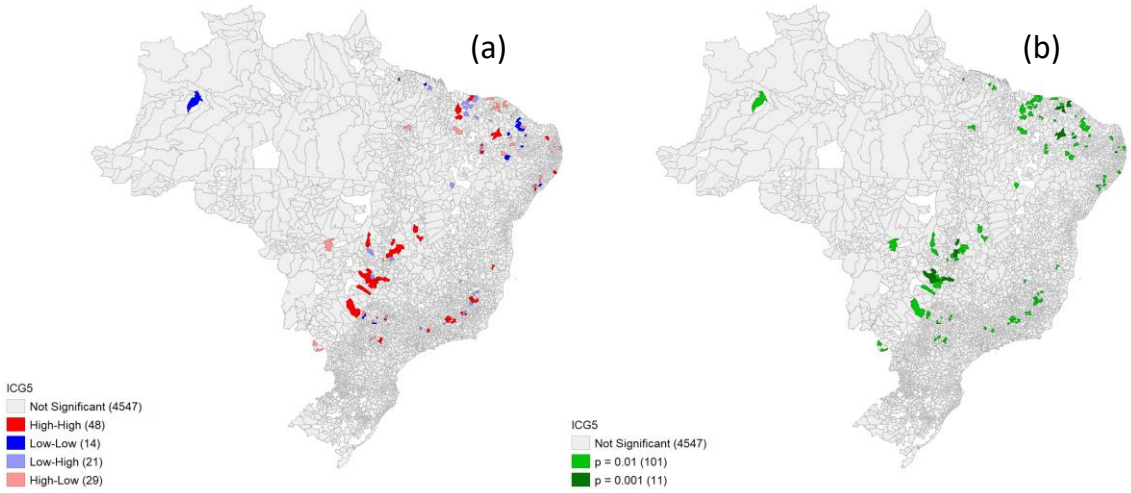


Figura 37 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ICG5

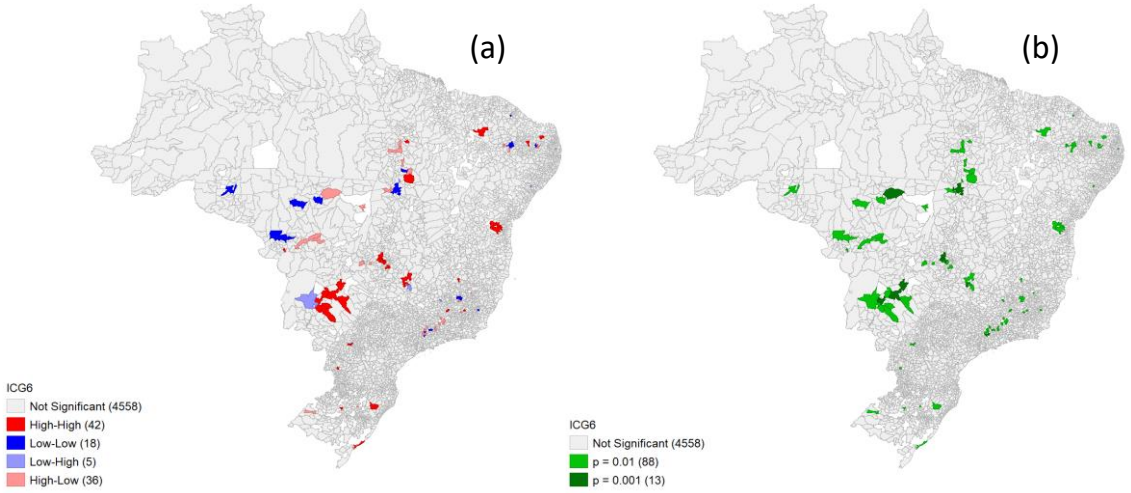


Figura 38 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável ICG6

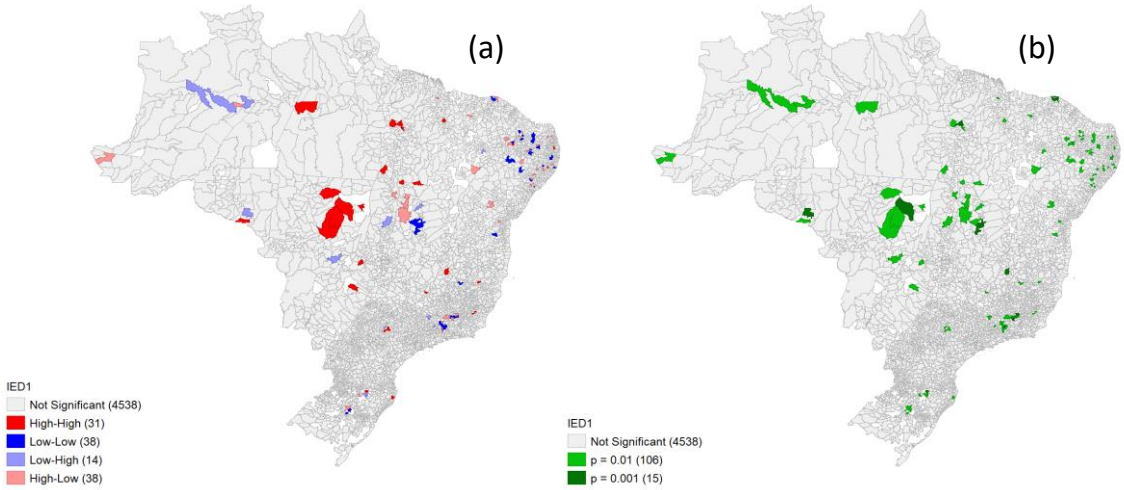


Figura 39 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável IED1

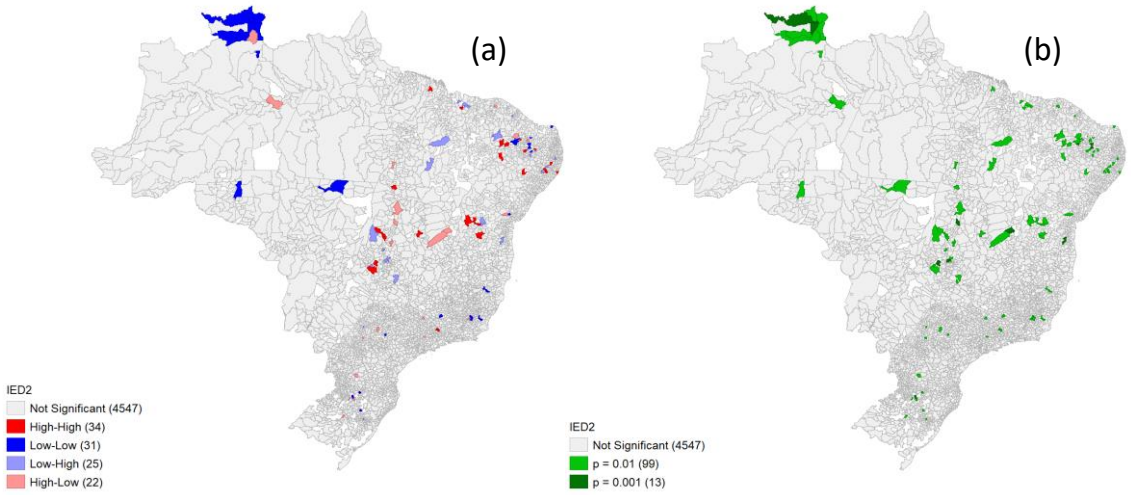


Figura 40 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável IED2

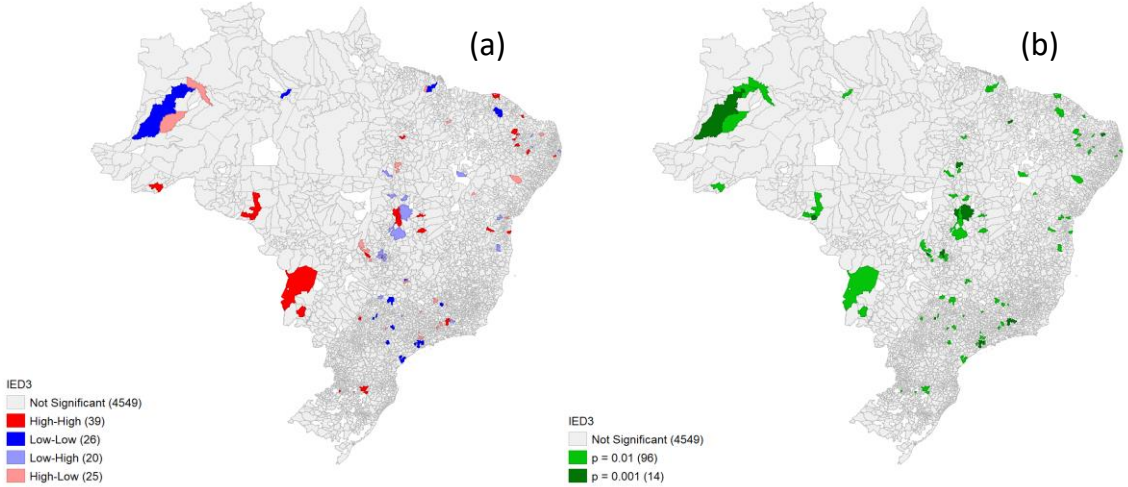


Figura 41 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável IED3

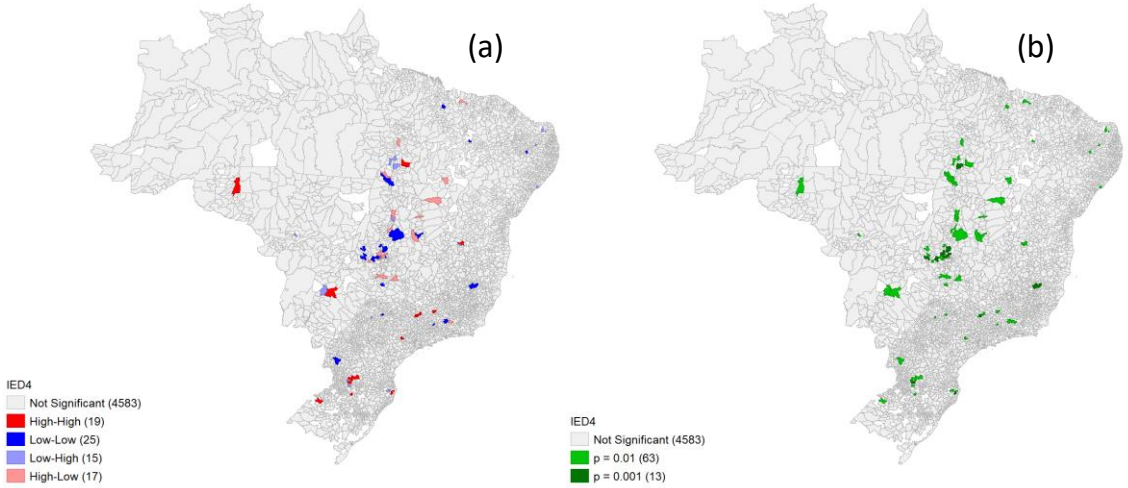


Figura 42 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável IED4

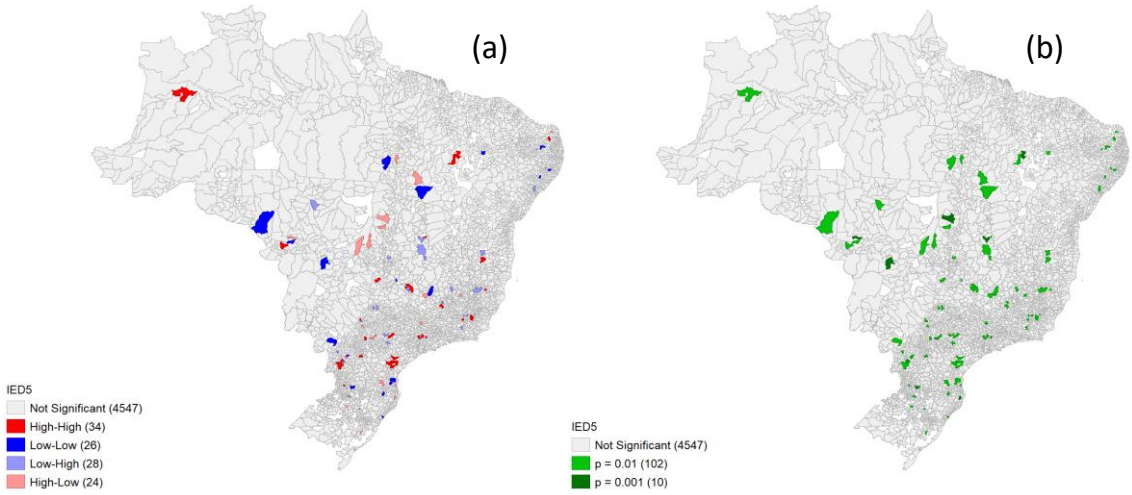


Figura 43 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável IED5

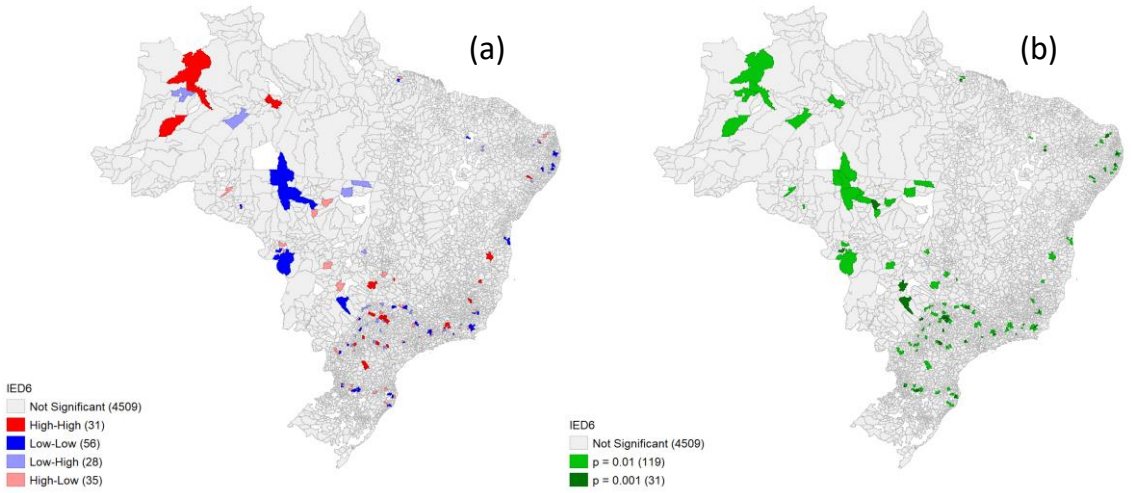


Figura 44 – Mapa LISA (a) e de significância (p < 0,01) (b) da variável IED6