Universidade de Brasília
Departamento de Estatística

Avaliação de Fatores de Influência no Desempenho dos Estudantes no Distrito Federal:
Uma Abordagem Multinível

Matheus Silva Martinez

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Matheus Silva Martinez

Avaliação do Efeito Vizinhança no Desempenho dos Estudantes do Distrito Federal, a partir de Modelos de Regressão Multinível

Orientadora: Ana Maria Nogales Vasconcelos

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2025

Agradecimentos

Primeiramente, agradeço profundamente a Deus, por moldar meu caminho até aqui. O Senhor é infinito e bondoso.

Agradeço a minha família por todo o suporte durante esses mais de cinco anos. Ao meu pai, Luiz Henrique Alves Martinez, por me mostrar o que é resiliência, por me fazer essa pessoa curiosa e cheia de ânimo de viver. À minha mãe, Rejanne Karla D'Abbadia Silva, que sempre acreditou e apoiou todos os meus projetos. Sem todas as vezes que a senhora disse "*Calma, meu filho! Você vai conseguir.*", eu não teria conseguido chegar aqui. Agradeço à minha irmã, Ana Beatriz, por todas as vezes que me jogou pra cima nos momentos adversos. Obrigado também, Cecília Usai, por todas as vezes que me mostrou que eu podia chegar mais longe.

Agradeço a todos os professores com os quais tive contato durante a graduação. Obrigado por propagarem essa linda arte, que é a Estatística. Ao Professor Leandro Correia, pelo apoio que me deu durante toda a graduação. E, principalmente, à Professora Ana Maria Nogales, por quem vou nutrir uma profunda admiração para sempre. A senhora faz com que o aprendizado da Estatística seja apaixonante. Me sinto imensamente honrado por ter concluído este trabalho com a senhora, que é um dos meus grandes exemplos.

Não posso deixar de mencionar os grandes amigos que tive a oportunidade de conhecer durante minha trajetória universitária. Daniel Paranaguá, Arthur Rodrigues, Renan Menezes, Bruno Kawano, Tiago Sampaio e Gabriel Côrtes, obrigado por sempre sonharem comigo.

Agradeço ao Movimento Empresa Júnior e, especialmente, à ESTAT Consultoria Estatística, pela pessoa que me tornei. Sempre levarei essas vivências no meu coração. *Formar, por meio da vivência empresarial, lideranças comprometidas e capazes de transformar o país em um Brasil Empreendedor.*

Dedico este trabalho ao meu avô, Fausto D'Abbadia Silva, hoje não mais presente em corpo, mas certamente contente pela minha conquista.

Resumo

O presente estudo tem como objetivo analisar os fatores associados ao desempenho em Matemática dos estudantes do 5º ano do Ensino Fundamental das escolas públicas do Distrito Federal. Aprofundar essa análise requer uma abordagem que reconheça a estrutura hierárquica dos dados educacionais, onde os estudantes estão inseridos em escolas que, por sua vez, estão situadas em Regiões, com diferentes contextos socioeconômicos. Nesse sentido, torna-se essencial investigar a existência do "Efeito Escola", que reflete a influência do ambiente escolar no aprendizado, e do "Efeito Vizinhança", que capta o impacto das regiões das escolas, ou das residências dos estudantes.

Com o objetivo de contemplar variáveis dos três níveis de análise, foram utilizadas as bases de dados do Sistema de Avaliação da Educação Básica – Saeb (Prova Brasil 2019), Censo Escolar 2019 e Pesquisa Distrital por Amostra de Domicílios 2019 (PDAD). Assim, para captar da melhor forma os fatores individuais e contextuais que impactam no aprendizado, foram construídos Modelos de Regressão Multinível.

Com isso, foi encontrado que aproximadamente 8% da variação do desempenho dos estudantes é atribuída à escola (Efeito Escola), valor aquém do esperado ao comparar com outras Unidades Federativas (UFs). Porém, o resultado se torna sensato ao compreender que as escolas do DF são muito homogêneas entre si. Não foi encontrado efeito vizinhança a partir de um 3º nível da modelagem e, sim, por variáveis que indicam uma segregação socioeconômica no DF, ligadas à renda e conclusão do Ensino Superior.

Palavras-chave: Modelos Multinível; Indicadores Educacionais; Distrito Federal; Estatística; Efeito Escola; Efeito Vizinhança

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Indubitavelmente, a qualidade da educação é uma preocupação constante em qualquer sistema educacional, e a busca por entender os fatores que influenciam o desempenho dos estudantes é fundamental para promover melhorias significativas (MACHADO et al., 2014). A avaliação do desempenho dos estudantes é uma métrica crítica para medir o sucesso do sistema educacional. No entanto, a complexidade desse fenômeno envolve um conjunto diversificado de variáveis que vão muito além das características do aluno.

A partir da análise da base de dados do SAEB 1999, Albernaz, Ferreira e Franco (2002) determinaram que os fatores que mais influenciam o desempenho dos alunos podem ser agrupados em duas categorias: as características individuais e familiares e a categoria das variáveis escolares. Alguns autores classificam os fatores relacionados ao segundo grupo como o "Efeito Escola", que faz referência aos diversos recursos disponíveis para os estudantes, tais como infraestrutura escolar, equipamentos disponibilizados, e métodos de ensino. Ademais, de acordo com Alves e Franco (2008), "as condições das escolas, os recursos escolares, o acompanhamento dos professores e a gestão são características que possuem ligação com o desempenho dos alunos".

Outrossim, outra característica relevante para compreender fatores que influenciam no desempenho de estudantes é o "Efeito Vizinhança"(Maloutas et al., 2019), ou "Efeito de Lugar"(Bourdieu, 1997). Segundo Andrade e Silveira (2020), a partir da década de 1980, houve vários estudos que tiveram o objetivo de entender como o local onde as pessoas vivem afeta a maneira como estas acessam oportunidades. Os autores investigaram como a localização geográfica pode influenciar, dentre outros fatores, o desempenho escolar.

Também é possível refletir sobre o cenário de ocupação do Distrito Federal. O Índice de Desenvolvimento Humano (IDH) do DF, 0,824, é o maior do Brasil (IBGE, 2010). Todavia, de acordo com a Pesquisa Nacional por Amostra de Domicílios (PNAD) Contínua de 2021, o Índice de Gini, que mede a desigualdade de renda em uma escala de 0 a 1, sendo 0 a perfeita igualdade e 1 a máxima desigualdade, é 0,566. Essa desigualdade no Distrito Federal resulta em uma severa segregação socioespacial, que acarreta no polinucleamento do território (Paviani, 2011).

Portanto, este trabalho tem como objetivo central a investigação do impacto dos efeitos "Escola"e "Vizinhança"do Distrito Federal no desempenho dos estudantes do 5º ano do Ensino Fundamental, a partir de Modelos de Regressão Multinível (HOX;

MOERBEEK; SCHOOT, 2017). É importante salientar que o estudo se restringirá a estudantes de escolas públicas.

De acordo com Hox, Moerbeek e Schoot (2017, p.5) , ao considerar diferentes níveis, como Escola e Vizinhança, é possível que as observações individuais dos alunos não sejam completamente independentes. O autor destaca que "alunos na mesma escola tendem a ser semelhantes uns aos outros, devido a processos de seleção (por exemplo, algumas escolas podem atrair alunos de níveis mais elevados de status socioeconômico - INSE)". Nesse contexto, para lidar com a correlação intraclasse (que ocorre entre observações dentro de um mesmo nível), pode-se utilizar um Modelo de Regressão Multinível, ou Hierárquico, que permite ajustar as análises para esses diferentes níveis de agrupamento, levando em consideração a estrutura aninhada dos dados.

Diante desse cenário, este estudo tem como objetivo central investigar a influência dos efeitos "Escola" e "Vizinhança" sobre o desempenho dos estudantes do 5º ano do Ensino Fundamental das escolas públicas do Distrito Federal, utilizando Modelos de Regressão Multinível (HOX; MOERBEEK; SCHOOT, 2017). A análise considera diferentes níveis hierárquicos, reconhecendo que os alunos não estão isolados em seu processo de aprendizagem, mas sim inseridos em contextos que impactam diretamente sua trajetória educacional.

Para estruturar a análise, o trabalho está organizado da seguinte forma: no Capítulo 2, apresenta-se o referencial teórico, que discute conceitos fundamentais sobre Modelos de Regressão Multinível e suas aplicações em estudos educacionais. O Capítulo 3 detalha a metodologia empregada, incluindo a descrição das bases de dados utilizadas e a estruturação do modelo multinível. Em seguida, o Capítulo 4 apresenta os resultados obtidos a partir das análises, explorando a influência dos fatores individuais, escolares e da vizinhança sobre a proficiência em matemática dos estudantes.

A análise desse capítulo foi estruturada em três etapas principais. Primeiramente, foi realizada uma análise descritiva exploratória, na qual foram examinadas as características dos alunos, das escolas e das vizinhanças, utilizando tabelas, gráficos e medidas estatísticas. Essa etapa teve o propósito de fornecer um panorama inicial dos dados e identificar possíveis padrões nos diferentes níveis de análise. Na sequência, foram conduzidas análises bivariadas, com o objetivo de investigar associações entre as variáveis explicativas e a proficiência em matemática. Foram exploradas relações estatísticas entre variáveis individuais, escolares e de vizinhança com o desempenho acadêmico dos estudantes, utilizando testes de correlação e comparações de médias. Por fim, foi realizada a modelagem estatística, começando com um modelo de regressão linear simples para o

nível aluno (1º nível), seguido pelo modelo nulo (sem variáveis explicativas), que permitiu estimar o coeficiente de correlação intraclasse (ICC) e verificar a necessidade do uso da abordagem multinível. Posteriormente, foi apresentado um modelo de dois níveis, incluindo variáveis explicativas relacionadas à escola, e, por fim, um modelo de dois níveis aprimorado com variáveis da vizinhança das escolas.

Por fim, o Capítulo 5 traz as considerações finais, destacando as principais conclusões do estudo e as possíveis implicações dos resultados para a formulação de políticas educacionais.

2 Referencial Teórico

Nesta seção, serão apresentadas as técnicas estatísticas utilizadas neste relatório, com o objetivo de proporcionar uma melhor compreensão das análises realizadas.

2.1 Regressão Linear

Um modelo de regressão linear representa a relação entre uma variável resposta e uma ou mais variáveis explicativas (KUTNER et al., 2005). No caso em que este modelo possui duas variáveis explicativas, ele pode ser definido da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \tag{2.1.1}$$

em que,

i. Y_i é a variável resposta.

ii. X_{1i} e X_{2i} são as variáveis explicativas observadas no aluno i.

iii. β_0 é o intercepto da regressão linear, que representa o valor esperado da variável resposta (Y_i) quando todas as outras variáveis explicativas (X_1 e X_2) são iguais a zero. Entretanto, é importante notar que, em muitos contextos práticos, este valor pode não ser interpretável, especialmente se o ponto em que $X_1 = 0$ e $X_2 = 0$ não for significativo ou possível no contexto do estudo. Nesses casos, o intercepto serve mais como um ajuste matemático do modelo do que uma interpretação direta.

iv. β_1 e β_2 são coeficientes associados às variáveis X_1 e X_2, respectivamente. Estes coeficientes refletem como a média da variável resposta (Y_i) varia, dada a adição de uma unidade em X_1 ou X_2, mantendo a outra fixada.

v. ε_i representa o erro aleatório da regressão, que é a diferença entre o valor da variável resposta e o valor que se espera da variável resposta, através do modelo. Isso representa a variação não explicada pelos preditores escolhidos para o modelo.

É importante frisar que um Modelo de Regressão Linear tem que seguir os seguintes pressupostos fundamentais:

- Linearidade: A relação entre a Y_i e as variáveis independentes (X_1, X_2) deve ser linear.

- Os erros aleatórios (ε_i) não devem ser correlacionados.
- Homoscedasticidade: A variância dos erros ($Var(\varepsilon_i)$) deve ser constante.
- Normalidade: Os erros devem seguir distribuição normal ($\varepsilon_i \sim N(0, \sigma^2)$

Além disso, um possível problema de um Modelo de Regressão Linear é a existência de multicolinearidade, onde as variáveis explicativas (X_1, X_2) são altamente correlacionadas entre si.

Desse modo, o modelo de regressão linear definido acima só seria apropriado se este avaliasse relações de alunos que pertencem a um mesmo grupo, sendo assim, da mesma escola e vizinhança. Quando tratamos de situações de formato hierárquico, ocorre a violação do princípio de independência das variáveis resposta, de modo que a regressão linear não leva em consideração a separação em níveis e não acomoda adequadamente os efeitos correlacionados (Hox, 2010).

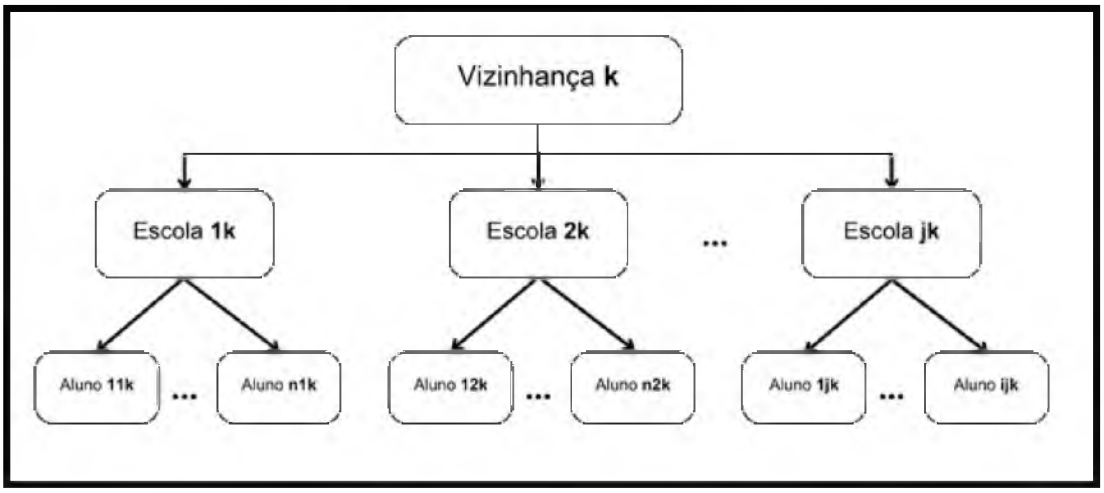
Além disso, outra motivação para a utilização do Modelo de Regressão Multinível é a adição de novas variáveis, como de infraestrutura escolar e características da vizinhança, que não são capturadas no primeiro nível de estudo (aluno) de forma apropriada. Ademais, como mencionado anteriormente, a estrutura hierárquica dos dados exige que levemos em consideração as relações entre alunos pertencentes à mesma escola ou vizinhança. Dessa forma, com o objetivo de capturar de forma efetiva as influências estudadas, os modelos de regressão multinível, que discutiremos a seguir, são a abordagem estatística mais apropriada.

2.2 Regressão Multinível

A regressão multinível, ou modelo hierárquico linear, é especialmente indicada para lidar com dados que apresentam uma estrutura hierárquica. Nesses casos, existe uma dependência natural entre os dados do estudo, o que torna inadequado o uso de técnicas de regressão tradicionais, que assumem independência entre as observações. (HOX; MOERBEEK; SCHOOT, 2017)

No contexto deste trabalho, que busca avaliar o impacto de diferentes fatores nas notas do SAEB de alunos do Ensino Fundamental no Distrito Federal, a regressão multinível se mostra adequada por possibilitar a análise simultânea das influências de variáveis em diferentes níveis (como as características individuais dos estudantes e os contextos escolares), respeitando as relações hierárquicas presentes nos dados.

Figura 1: Estrutura Hierárquica.



Elaboração própria

O referencial teórico dos Modelos Multinível neste trabalho está profundamente baseado em Hox, Moerbeek e Schoot (2017), Gavin (2004) e Goldstein (2003). Ademais, os estudos de Ferrão e Fernandes (2003) foram cruciais para o entendimento da aplicação destes modelos em contextos educacionais, especialmente no Brasil.

2.2.1 Detalhamento do Modelo Multinível

Matematicamente, o modelo multinível básico, com duas variáveis explicativas, pode ser expresso da seguinte forma:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \varepsilon_{ij} \tag{2.2.1}$$

em que,

i. Y_{ij} é a variável resposta.

ii. β_{0j} é o intercepto.

iii. β_{1j} e β_{2j} são os coeficientes de regressão, ou inclinação.

iv. ε_i representa o erro aleatório da regressão, que é a diferença entre o valor da variável resposta e o valor que se espera da variável resposta, através do modelo. Isso representa a variação não explicada pelos preditores escolhidos para o modelo.

É importante notar que em um modelo multinível, o segundo nível da análise é indexado por j, que varia de 1 a J, que representa o número de grupos que agrupamos

nossas observações. Já o primeiro nível continua sendo listado por i, que varia de 1 a n_j.

Para ficar mais claro, é possível fazer um paralelo com a análise que foi realizada neste trabalho. Neste caso, o índice i representa um aluno do 5º ano do Ensino Fundamental que realizou a prova SAEB em 2019, e o índice j uma das escolas públicas que passou pelo exame. Sendo J o número de escolas públicas que passaram pelo exame e n_j o número de estudantes que fizeram a prova na escola j.

Após visualizar as equações 2.1.1 e 2.2.1, é possível notar que a diferença entre elas é que o intercepto (β_{0j}) e os coeficientes de regressão (β_{1j} e β_{2j}) variam de acordo com as escolas (j).

Para compreender o modelo de regressão multinível, deve-se entender que a variação dos coeficientes de regressão é explicada ao introduzir variáveis explicativas do 2º nível (escolas).

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \tag{2.2.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \tag{2.2.3}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} Z_j + u_{2j} \tag{2.2.4}$$

onde, a equação β_{0j} prediz o valor médio da variável resposta no nível j em função de uma variável explicativa do 2º nível, no nosso caso, Z_j. Dessa forma, se o coeficiente γ_{01} for positivo, o valor médio da variável dependente é maior em grupos onde a variável explanatória do nível j tem valores mais altos. Por outro lado, se γ_{01} for negativo, o valor médio da variável dependente é menor em grupos com maiores valores da variável explicativa.

A interpretação das equações seguintes (β_{1j} e β_{2j}) é um pouco mais complexa. A primeira equação (β_{1j}) indica que a relação entre a variável resposta Y e uma variável explicativa do 1º nível X_1 depende do valor da variável Z_j, do 2º nível. Dessa forma, se γ_{11} for positivo, o efeito de X_1 na variável dependente é maior para grupos com valores mais altos de Z_j. Inversamente, se γ_{11} for negativo, o efeito de X é menor para esses grupos. Ocorre da mesma forma para a equação de β_{2j}. Assim, pode-se afirmar que a variável explicativa de 2º nível, Z_j, atua como uma variável moderadora na relação entre Y e X_1 ou X_2.

Além disso, os termos u_{0j}, u_{1j} e u_{2j} nas equações acima são termos de resíduo (aleatórios) no 2º nível de análise (escola). Esses resíduos u_j são assumidos como tendo média zero e sendo independentes dos erros residuais e_{ij} no 1º nível (aluno). A variância dos erros residuais u_{0j} é especificada como σ^2_{u0}, e a variância dos erros residuais u_{1j} e u_{2j} é

especificada como σ^2_{u1} e σ^2_{u2}, respectivamente. As covariâncias entre os termos de resíduo são denotadas por σ_{u01}, σ_{u02} e σ_{u12}, que geralmente são assumidas como diferentes de zero.

Vale ressaltar que, nas equações 2.2.2, 2.2.3 e 2.2.4, os coeficientes de regressão γ são considerados constantes entre as escolas. Por esse motivo, não possuem o índice j, que indicaria uma variação entre as escolas.

Por fim, o modelo com duas variáveis explicativas no 1º nível (aluno) e uma variável explicativa no 2º nível (escola) pode ser escrito como uma única equação de regressão multinível completa:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}Z_j + \gamma_{11}X_{1ij}Z_j + \gamma_{21}X_{2ij}Z_j + u_{1j}X_{1ij} + u_{2j}X_{2ij} + u_{0j} + e_{ij} \tag{2.2.5}$$

onde,

- $\gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}Z_j + \gamma_{11}X_{1ij}Z_j + \gamma_{21}X_{2ij}Z_j$ são os coeficientes fixos do modelo.
- $u_{1j}X_{1ij} + u_{2j}X_{2ij} + u_{0j} + e_{ij}$ é chamada de parte randômica do modelo.

2.2.2 Estimação dos parâmetros

A estimação dos parâmetros em modelos multinível é fundamental para obter coeficientes confiáveis e interpretar corretamente as relações entre variáveis em diferentes níveis hierárquicos. Três principais de parâmetros são estimados: Efeitos fixos (γ), Efeitos aleatórios (β_j) e Componentes de variância (σ^2). O método mais utilizado na literatura é o da Máxima Verossimilhança (MV), garantindo estimativas consistentes. Dentro dessa técnica, pode-se seguir duas abordagens: Máxima Verossimilhança Completa (MVC), que estima conjuntamente os coeficientes de regressão e componentes de variância, e Máxima Verossimilhança Restrita (MVR) que estima os componentes de variância primeiro, corrigindo o viés presente na MVC.

Neste trabalho, foi utilizada a Máxima Verossimilhança Completa, visando obter estimativas mais precisas dos componentes de variância, fundamentais para uma análise multinível confiável. Essa escolha foi feita pois a MVC permite a comparação direta de modelos por meio de testes de razão de verossimilhança, facilitando a seleção do modelo mais adequado.

2.2.3 Estrutura para Análise

Seja p o número de variáveis explicativas incluídas no nível mais baixo e q o número de variáveis explicativas no nível mais alto. O modelo multinível geral de dois níveis pode ser expresso como:

$$Y_{ij} = \gamma_{p0} + \gamma_{p0}X_{pij} + \gamma_{0q}Z_{qj} + \gamma_{pq}X_{pij}Z_{qj} + u_{pj}X_{pij} + u_{0j} + e_{ij} \tag{2.2.6}$$

Para a modelagem multinível ideal, recomenda-se seguir um método estruturado em cinco passos:

Passo 1: Modelo Nulo e análise do Coeficiente de Correlação Intra-classe

O modelo nulo, ou também chamado de modelo somente com intercepto (*intercept-only model*), é o ponto de partida em uma análise de regressão multinível. Neste caso, o modelo não tem variáveis explicativas, portanto o preditor linear é composto apenas pelo intercepto. A utilização deste modelo é crucial para para decompor a variância total da variável dependente entre os diferentes níveis do modelo.

Portanto, com o β_{0j} definido na equação (2.2.2), pode-se definir o modelo multinível nulo como:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{2.2.7}$$

onde:

i. Y_{ij} é a variável dependente para o indivíduo i no grupo j;

ii. γ_{00} é o intercepto global (a média geral de todos os grupos);

iii. u_{0j} é o erro aleatório no nível do grupo j (a diferença entre a média de cada grupo e a média global);

iv. e_{ij} é o erro aleatório no nível do indivíduo i dentro do grupo j.

Esse modelo, sem variáveis explicativas, permite estimar a variância tanto dentro dos grupos quanto entre os grupos, dividindo a variância da variável resposta em duas partes:

- A variância dentro dos grupos (σ_e^2) — ou seja, a variância entre os indivíduos dentro do mesmo grupo;
- A variância entre os grupos (σ_{u0}^2) — ou seja, a variância entre as médias dos dife-

rentes grupos.

No contexto deste trabalho, onde i representa os alunos e j representa as escolas, o modelo nulo é útil para identificar quanto da variação no desempenho dos alunos pode ser atribuída às diferenças dentro das escolas e quanto é devido às diferenças entre as escolas.

Ademais, a avaliação do Coeficiente de Correlação Intraclasse (ICC) é uma parte crucial da formulação de um modelo multinível, pois tem como papel justificar seu uso. O ICC indica o grau de dependência entre as observações dentro de um mesmo grupo. Em outras palavras, ele mostra o quanto da variação total de Y_{ij} (a variável resposta) pode ser explicada pelas diferenças entre os grupos.

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_e} \tag{2.2.8}$$

onde ρ varia de 0 a 1.

É importante frisar que um ICC próximo de 0 indica que quase toda a variação ocorre dentro dos grupos, tendo pouca variação entre os grupos. Em contrapartida, um ICC próximo de 1 indica que a maior parte da variação ocorre entre os grupos, tendo pouca variação dentro dos grupos. Desse modo, se o Coeficiente for relativamente alto, é apropriado utilizar um modelo multinível para capturar essa estrutura hierárquica dos dados.

Por exemplo, neste trabalho, alunos estão aninhados em escolas, portanto o modelo nulo pode ser utilizado para verificar quanto à variação da proficiência dos alunos se deve às diferenças entre as escolas (2º nível), ou quanto é referente às diferenças entre os próprios alunos (1º nível). O ICC indicará a proporção da variância total da variável resposta que pode ser atribuída às diferenças entre as escolas.

O coeficiente de correlação intraclasse é amplamente utilizado em diferentes contextos de pesquisa para quantificar o efeito de grupos em várias situações. Em pesquisas educacionais, o ICC é comumente interpretado como o "Efeito Escola"(FERRÃO, 2004), em casos onde são estudados quanto da variação no desempenho dos alunos é atribuída às diferenças entre as escolas. O termo Efeito Vizinhança também é muito presente em estudos educacionais, de saúde pública, ou até de mercado de trabalho (OLIVEIRA, 2012), nestes caso, o ICC pode indicar o impacto da área de residência em diversos resultados.

Passo 2: Adição de variáveis do nível 1

Analisa-se um modelo somente com variáveis explicativas do primeiro nível de análise. Com isso, os componentes de variância dos coeficientes são fixados em zero.

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{pij} + u_{0j} + e_{ij} \tag{2.2.9}$$

onde X_{pij} são os valores das p variáveis explicativas do primeiro nível. Neste passo, testa-se a significância das variáveis adicionadas.

Passo 3: Adição de variáveis dos níveis 1 e 2

Aqui, são adicionadas variáveis explicativas do nível 2 que, neste trabalho, são informações a respeito das escolas.

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{pij} + \sum_{q=1}^{Q} \gamma_{0q} Z_{qj} + u_{0j} + e_{ij} \tag{2.2.10}$$

onde Z_{qj} são os valores das q variáveis explicativas do nível 2.

Os modelos apresentados nos passos 2 e 3 são comumente denominados na literatura como "modelos de componente de variância". Eles decompõem a variância do intercepto em diferentes componentes, para cada nível hierárquico, possibilitando comparar os modelos em relação à variância explicada.

Passo 4: Modelo de coeficientes randômicos

O próximo passo é avaliar se algum coeficiente de regressão do nível menor tem uma componente de variância significativamente diferente de zero, entre as escolas.

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{pij} + \sum_{q=1}^{Q} \gamma_{0q} Z_{qj} + \sum_{p=1}^{P} u_{pj} X_{pij} + u_{0j} + e_{ij} \tag{2.2.11}$$

onde u_{pj} são os resíduos do nível maior dos coeficientes das variáveis explicativas do nível menor, X_{pij}.

Passo 5: Modelo completo

No passo final, o modelo completo é dado pela adição das interações entre as variáveis do 2º nível e as variáveis que obtiveram uma variabilidade significativamente diferentes de zero, verificadas no passo anterior.

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{P} \gamma_{p0} X_{pij} + \sum_{q=1}^{Q} \gamma_{0q} Z_{qj} + \sum_{p=1}^{P} u_{pj} X_{pij} + \sum_{p=1}^{P} \sum_{q=1}^{Q} \gamma_{pq} W_{qj} X_{pij} + u_{0j} + e_{ij} \quad (2.2.12)$$

2.2.4 Qualidade do Modelo

É possível utilizar critérios de comparação entre diferentes modelos baseados na função de verossimilhança, como o *deviance*.

$$d = -2 * ln(L) \quad (2.2.13)$$

Geralmente, menores valores de d indicam um melhor ajuste do modelo. Quando um modelo mais simples (M1) pode ser obtido a partir de um modelo mais complexo (M2) removendo alguns parâmetros, os dois são considerados aninhados. Nesse caso, pode-se comparar os modelos utilizando um teste de razão de verossimilhança, que avalia se a inclusão dos parâmetros adicionais no modelo mais complexo resulta em uma melhoria significativa no ajuste. A estatística do teste segue uma distribuição qui-quadrado, sendo calculada pela diferença entre os valores de *deviance* dos modelos comparados.

$$D = d_1 - d_2 \quad (2.2.14)$$

Se os modelos não são aninhados, pode-se usar o AIC. Esta medida utiliza o *deviance* e penaliza de acordo com o número de parâmetros utilizados no modelo (k). O modelo que apresentar o menor valor de AIC é, geralmente, o que apresenta melhor ajuste.

$$AIC = d + 2k \quad (2.2.15)$$

Além do AIC, outra métrica utilizada para comparação de modelos multinível é o *Bayesian Information Criterion* (BIC), que adiciona uma penalização maior para modelos com mais parâmetros, considerando também o tamanho da amostra (n). Além disso, o BIC busca um equilíbrio entre ajuste e complexidade do modelo, o BIC tende a favorecer modelos mais simples quando o tamanho da amostra é grande. Assim, ao comparar modelos, um menor valor de BIC indica um ajuste mais parcimonioso.

$$BIC = d + k\ln(n) \tag{2.2.16}$$

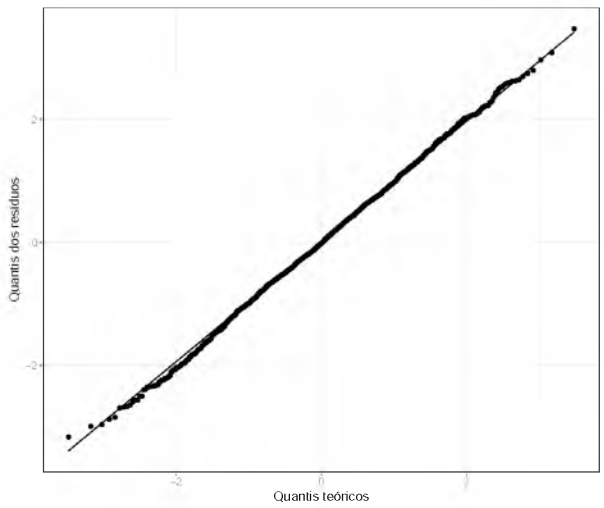
onde n é o tamanho da amostra.

2.2.5 Diagnóstico dos Resíduos

Por fim, após a escolha do melhor modelo, é crucial validar os pressupostos da Modelagem Multinível, que são sobre linearidade, normalidade e homocedasticidade dos resíduos, semelhante aos pressupostos de modelos de regressão tradicionais. No entanto, em modelos multinível, os resíduos possuem uma estrutura mais complexa, pois há diferentes componentes de erro associados aos efeitos aleatórios do modelo. A análise dos resíduos pode ser realizada tanto graficamente quanto por meio de testes estatísticos.

A normalidade dos resíduos pode ser verificada de forma visual, a partir de um histograma ou, principalmente, por gráficos quantil-quantil (*Q-Q plot*). Se os resíduos seguem distribuição normal, os pontos distribuem-se a longo da linha diagonal apresentada na Figura 2. Para verificar a condição de normalidade, também é comum utilizar testes de normalidade, como o de *Shapiro-Wilk*.

Figura 2: Gráfico de Probabilidade Normal

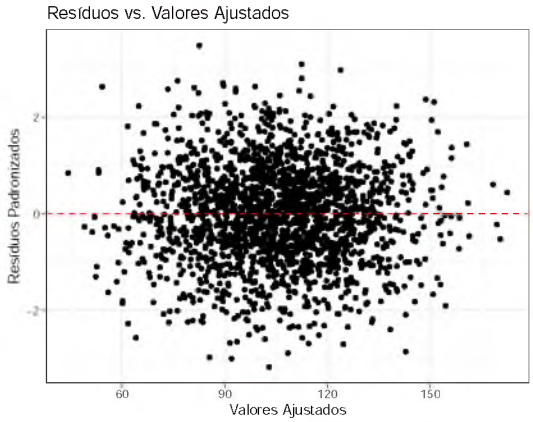


Fonte: Elaboração própria.

O gráfico de dispersão que compara resíduos padronizados (studentizados) e os valores preditos pelo modelo é o mais utilizado para verificar linearidade e homoscedasticidade dos resíduos. Se os pontos do gráfico estiverem visualmente distribuídos em torno do 0 e não apresentarem nenhum padrão nítido, os pressupostos foram atendidos. O teste

de *Breusch-Pagan* pode ser utilizado para verificar homoscedasticidade.

Figura 3: Resíduos studentizados *versus* valores preditos



Fonte: Elaboração própria.

A validação dos resíduos é fundamental para garantir a confiabilidade das inferências estatísticas realizadas a partir do modelo. Caso os pressupostos não sejam atendidos, estratégias como transformação de variáveis, inclusão de novos termos no modelo, atribuição de outra distribuição de probabilidade ou a utilização de modelos não lineares podem ser consideradas para aprimorar o ajuste e a adequação do modelo.

3 Metodologia

Nesta seção serão descritos os procedimentos metodológicos usados durante o estudo, bem como as bases de dados utilizadas, e as variáveis analisadas em cada nível hierárquico. Além disso, serão detalhados os métodos de tratamento, estruturação e análise dos dados.

3.1 Conjunto de dados

Como foi discutido anteriormente, as análises presentes neste estudo terão foco em avaliar a proficiência em matemática de estudantes do 5º ano do Ensino Fundamental das escolas públicas do Distrito Federal (DF) a partir de três níveis: Aluno, Escola e Vizinhança. Portanto, foi utilizado um banco de dados para cada um dos níveis.

É crucial informar que para a realização do estudo proposto, os dados dos níveis apresentados devem ser acessados de forma desagregada, para que assim seja possível cruzá-los. Desse modo, foi utilizada a Sala Segura do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), por meio do Serviço de Acesso a Dados Protegidos (SEDAP).

Por fim, é importante salientar que, de acordo com a Portaria nº 637/2019, do SEDAP: "só serão permitidas extrações de resultados cujo conteúdo não permita identificação, direta ou indireta, de pessoa natural". Portanto, todos os resultados presentes no relatório foram extraídos de forma agregada e sem identificação individual.

3.1.1 Prova do Sistema de Avaliação da Educação Básica (SAEB)

De acordo com Araújo e Lúzio (2005), o Sistema Nacional de Avaliação da Educação Básica (SAEB) representa uma das primeiras iniciativas do Brasil para compreender os resultados de aprendizado dos alunos. Atualmente, este é o instrumento mais abrangente para avaliar externamente a qualidade do desenvolvimento de habilidades e competências dos estudantes em todo o país e é reconhecido como um dos sistemas de avaliação em larga escala mais sofisticados e abrangentes da América Latina.

Atualmente, a aplicação do SAEB abrange escolas públicas e particulares dos 26 estados brasileiros e do Distrito Federal. A prova é aplicada em estudantes do ensino fundamental e médio, com o objetivo de avaliar suas proficiências em Língua Portuguesa e Matemática. É importante frisar que neste estudo serão utilizados os dados de estudantes

do 5º ano do ensino fundamental das escolas públicas do DF. A variável resposta, ou de interesse, é a proficiência em matemática. Além disso, é pedido que os estudantes preencham um questionário no cartão-resposta do SAEB, que aborda temas como nível socioeconômico, infraestrutura do bairro de residência, cor/raça, meio de deslocamento para a escola. As informações deste questionário também foram fornecidas pelo INEP e compuseram as variáveis explicativas do modelo multinível.

Tabela 1: Variáveis do nível aluno.

Rótulo	**Descrição**	**Tipo**
proficiencia_mt_saeb	Proficiência em Matemática	Quantitativa
nu_idade	Idade	Quantitativa
tp_sexo	Sexo	Qualitativa
tp_cor_raca	Cor/Raça	Qualitativa
vl_inse_aluno	Nível Socioeconomico (INSE)	Quantitativa
incentivo_pais_q	Incentivo dos pais nos estudos	Qualitativa
pavimentada_q	Pavimentação na região onde mora	Qualitativa
iluminacao_q	Iluminação na região onde mora	Qualitativa
meiodescocamento_q	Meio de deslocamento para a escola	Qualitativa
reprovacao_q	Reprovação na escola	Qualitativa
abandono_q	Abandono na escola	Qualitativa

Fonte: Elaboração própria

3.1.2 Censo Escolar

O Censo Escolar é realizado anualmente pelo INEP. De acordo com o Órgão:

> O Censo Escolar é uma pesquisa estatística declaratória realizada anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), autarquia vinculada ao Ministério da Educação (MEC), em regime de colaboração entre a União, os estados, o Distrito Federal e os municípios, que tem por objetivo realizar um amplo levantamento sobre a educação brasileira. É o mais importante levantamento estatístico educacional sobre as diferentes etapas e modalidades de ensino da educação básica e da educação profissional. (INEP, 2022, p.2)

O aprimoramento da metodologia de coleta do Censo Escolar tem contribuído

para uma distribuição mais equitativa de recursos, uma vez que utiliza o número de alunos matriculados nas redes municipais e estaduais como critério. No contexto da formulação de políticas públicas, os dados coletados servem como base para a elaboração de um diagnóstico educacional do Brasil, com o intuito de desenvolver estratégias que promovam o acesso, a permanência e o sucesso dos estudantes na escola. Além disso, o Censo desempenha um papel fundamental ao fornecer os dados necessários para o cálculo de indicadores educacionais (LIMA; SOUSA, 2014, p.99) .

Desse modo, a base de dados do Censo Escolar de 2019 foi utilizada no estudo para o segundo nível hierárquico, a escola. É importante frisar que cada observação desta fonte de dados representa uma escola e as variáveis escolares disponibilizadas pelo INEP são resumidas às características da infraestrutura das escolas. Ademais, as mais de 50 variáveis de infraestrutura são qualitativas, em forma de *dummy*.

1. Prédio escolar próprio (IN_LOCAL_FUNC_PREDIO_ESCOLAR);
2. Abastecimento de água pela rede pública (IN_AGUA_REDE_PUBLICA);
3. Abastecimento de água por poço artesiano (IN_AGUA_POCO_ARTESIANO);
4. Abastecimento de água por cacimba (IN_AGUA_CACIMBA);
5. Abastecimento de água por fonte ou rio (IN_AGUA_FONTE_RIO);
6. Energia elétrica da rede pública (IN_ENERGIA_REDE_PUBLICA);
7. Esgoto conectado à rede pública (IN_ESGOTO_REDE_PUBLICA);
8. Esgoto por fossa (IN_ESGOTO_FOSSA);
9. Sistema de esgoto inexistente (IN_ESGOTO_INEXISTENTE);
10. Serviço de coleta de lixo (IN_LIXO_SERVICO_COLETA);
11. Descarte de lixo por queima (IN_LIXO_QUEIMA);
12. Descarte de lixo por enterro (IN_LIXO_ENTERRA);
13. Sala da diretoria (IN_SALA_DIRETORIA);
14. Sala dos professores (IN_SALA_PROFESSOR);
15. Laboratório de informática (IN_LABORATORIO_INFORMATICA);
16. Laboratório de ciências (IN_LABORATORIO_CIENCIAS);

17. Sala de atendimento especial (IN_SALA_ATENDIMENTO_ESPECIAL);
18. Quadra de esportes coberta (IN_QUADRA_ESPORTES_COBERTA);
19. Quadra de esportes descoberta (IN_QUADRA_ESPORTES_DESCOBERTA);
20. Quadra de esportes (IN_QUADRA_ESPORTES);
21. Cozinha (IN_COZINHA);
22. Biblioteca (IN_BIBLIOTECA);
23. Sala de leitura (IN_SALA_LEITURA);
24. Biblioteca ou sala de leitura (IN_BIBLIOTECA_SALA_LEITURA);
25. Parque infantil (IN_PARQUE_INFANTIL);
26. Banheiro exclusivo para educação infantil (IN_BANHEIRO_EI);
27. Banheiro adaptado para pessoas com deficiência (IN_BANHEIRO_PNE);
28. Secretaria (IN_SECRETARIA);
29. Banheiro com chuveiro (IN_BANHEIRO_CHUVEIRO);
30. Refeitório (IN_REFEITORIO);
31. Despensa (IN_DESPENSA);
32. Almoxarifado (IN_ALMOXARIFADO);
33. Auditório (IN_AUDITORIO);
34. Pátio coberto (IN_PATIO_COBERTO);
35. Pátio descoberto (IN_PATIO_DESCOBERTO);
36. Dormitório para alunos (IN_DORMITORIO_ALUNO);
37. Dormitório para professores (IN_DORMITORIO_PROFESSOR);
38. Área verde (IN_AREA_VERDE);
39. Outras dependências (IN_DEPENDENCIAS_OUTRAS);
40. Acesso à internet (IN_INTERNET);

41. Acesso à internet por banda larga (IN_BANDA_LARGA).

Além das variáveis de infraestrutura, a base de dados do Censo Escolar foi enriquecida com dados do INEP que dizem respeito à quantidade de professores concursados, terceirizados ou temporários nas escolas.

3.1.3 Pesquisa Distrital por Amostra de Domicílios (PDAD)

A Pesquisa Distrital por Amostra de Domicílios (PDAD) é conduzida pela Codeplan a cada dois anos, com o propósito de coletar dados demográficos, sociais, de renda, e características dos domicílios e da região que os cerca. De acordo com o *site* da Codeplan, essa pesquisa abrange mais de 97% da população de Brasília, abarcando todas as 33 Regiões Administrativas (RAs), e oferece uma análise detalhada da situação em cada uma delas.

A PDAD pode ser utilizada para compreender a desigualdade social do Distrito Federal (DF) a partir da segregação nas RAs. Portanto, os dados da PDAD 2021 serão utilizados para estudar o Efeito Vizinhança, terceiro nível do modelo proposto.

Neste banco de dados, as observações correspondem às Regiões Administrativas e as variáveis são suas respectivas características. Dessa forma, com o objetivo de entender os fatores relevantes causados pela desigualdade social no DF, as seguintes variáveis foram selecionadas e classificadas em "P1,P2,...":

1. Porcentagem da população preta ou parda(P1);
2. Proporção da população com posse de celular para uso pessoal(P2);
3. Porcentagem de arranjos domiciliares monoparentais femininos(P3);
4. Porcentagem de domicílios que acessaram internet todos os dias nos "últimos três meses"(P4);
5. Porcentagem de domicílios com acesso à internet por meio de computador/laptop(P5);
6. Porcentagem de domicílios que utilizaram internet por motivo de educação/cursos(P6);
7. Porcentagem de jovens entre 15 e 17 anos que frequentam escolas(P7);
8. Proporção de pessoas que trabalham na mesma RA que residem(P8);

9. Proporção da população jovem que tem um tempo de deslocamento de até 30 minutos até a unidade de ensino(P9);
10. Proporção da população com 25 anos ou mais de idade, com ensino superior completo(P10);
11. Proporção da população com tempo de deslocamento de até 30 minutos até o trabalho(P11);
12. Proporção da população que trabalha no setor de Administração Pública (P12);
13. Proporção da população que trabalha no setor de Serviços Domésticos(P13);
14. Porcentagem da população que utiliza ônibus como principal meio de transporte utilizado para o trabalho(P14);
15. Domicílios ocupados em lotes regularizados(P15);
16. Porcentagem de municípios com Coleta Seletiva Direta(P16);
17. Rua asfaltada ou pavimentada (%)(P17);
18. Rua de acesso aos domicílios com iluminação (%)(P18);
19. Rua de acesso aos domicílios com calçada (%)(P19);
20. Erosão nas cercarias dos domicílios (%)(P20);
21. Acumulo de entulho nas cercarias dos domicílios (%)(P21);
22. Esgoto a céu aberto nas cercarias dos domicílios (%)(P22);
23. Policiamento regular nas proximidades (%)(P23);
24. Utilização de serviços domésticos nos domicílios mensalistas(P24);
25. Porcentagem de domicílios com insegurança alimentar moderada e grave (P25).
26. Porcentagem de domicídios com rendimento bruto domiciliar maior que 5 salários mínimos (P26).

Por fim, a base de dados sobre as RAs, a PDAD 2021, foi enriquecida com as categorias de renda, que classificam as RAs entre Renda média, Renda alta e Renda baixa.

3.1.4 Base de dados final

Para consolidar as informações dos três níveis hierárquicos — Aluno, Escola e Vizinhança — foi realizada a agregação dos bancos de dados descritos anteriormente em uma única base de dados final. Este processo seguiu etapas rigorosas de organização e cruzamento, com o objetivo de garantir a consistência das informações.

Inicialmente, as bases de dados foram unidas utilizando identificadores únicos disponíveis em cada nível hierárquico. No caso dos alunos, as informações individuais do SAEB foram vinculadas às características das escolas, obtidas do Censo Escolar, utilizando o código da escola (*id_escola*) como chave. Em seguida, as informações das escolas foram agregadas aos dados das RAs, com base nas suas respectivas localizações geográficas.

Após a agregação das informações seguindo os critérios estabelecidos, o banco de dados final contou com 34.393 observações e 118 variáveis. Contudo, é importante destacar que algumas escolas não realizaram a coleta do questionário socioeconômico junto aos alunos. Por conta disso, as observações relacionadas a esses alunos continham apenas informações sobre as características das escolas e de suas respectivas vizinhanças. Além disso, diversos alunos não finalizaram suas provas e escolheram não declarar nenhuma informação no questionário do SAEB. Dessa forma, dado que a análise e interpretação das características individuais dos alunos também são tema central neste trabalho, optou-se por excluir as observações dos alunos que não haviam preenchido o questionário sociodemográfico. Essa decisão foi tomada para garantir a integridade e a completude das análises realizadas, assegurando que todas as dimensões propostas no modelo fossem contempladas, sem dados faltantes. A RA Lago Sul não foi utilizada para a modelagem, porque os alunos de sua (única) escola pública não responderam ao questionário.

Além disso, as variáveis categóricas, quando necessário, foram recodificadas em forma de *dummy*, a fim de facilitar a análise dos modelos multinível.

Por fim, a base final foi estruturada de forma hierárquica, com cada observação correspondendo a um aluno do 5º ano, incluindo as informações relacionadas à sua escola e à sua vizinhança. Essa estrutura permitiu respeitar a organização multinível dos dados, essencial para a aplicação adequada dos modelos estatísticos.

Para tornar mais evidente a perda de dados durante o processo de limpeza, será apresentada a quantidade de dados faltantes em cada variável:

- Não responderam o questionário sociodemográfico – 8.397

- Raça/Cor – 2.005
- Iluminação – 589
- Pavimentação – 301
- Tempo de deslocamento – 453
- Reprovação – 357
- Abandono – 194
- Meio de deslocamento – 907

Portanto, ao eliminar dados faltantes, a base de dados final contou com 21.140 alunos. Para aprofundar o entendimento da perda de observações no banco de dados, a Tabela 2 apresenta a quantidade de observações antes e depois do tratamento, discriminadas por Região Administrativa. Dessa forma, é possível identificar as regiões que sofreram maior impacto com a exclusão de alunos que não preencheram o questionário sociodemográfico ou cujas informações estavam incompletas.

Tabela 2: Comparação do número de observações antes e depois da exclusão de valores faltantes.

Região Administrativa	Antes	Depois	Perdidos (%)
Scia e Estrutural	701	331	52,78
Varjão	178	86	51,69
Fercal	291	152	47,77
Sobradinho II	614	327	46,74
Planaltina	3227	1761	45,43
Paranoá	1737	951	45,25
Sol Nascente/Pôr do Sol	449	249	44,54
Sudoeste e Octogonal	76	43	43,42
Cruzeiro	284	161	43,31
Candangolândia	168	96	42,86
Park Way	171	103	39,77
Lago Norte	63	38	39,68
Gama	2086	1271	39,07
Núcleo Bandeirante	322	198	38,51
São Sebastião	1678	1034	38,38
Recanto das Emas	2013	1251	37,85
Brazlândia	1265	796	37,08
Santa Maria	1832	1156	36,90
Guará	887	562	36,64
Itapoã	238	151	36,55
Arniqueira	301	191	36,54
Taguatinga	2359	1500	36,41
Ceilândia	5565	3550	36,21
Plano Piloto	2165	1395	35,57
Sobradinho	1288	835	35,17
Riacho Fundo II	782	510	34,78
Samambaia	3005	1981	34,08
Riacho Fundo	388	269	30,67
Vicente Pires	194	140	27,84
Jardim Botânico	66	52	21,21

Fonte: Elaboração própria.

3.2 Análise de dados e Modelagem

A análise de dados será conduzida em etapas, começando por uma exploração descritiva das variáveis referentes ao aluno e, após isso, sobre escola e vizinhança (Seção 4.1). Essa etapa inicial tem como objetivo compreender a base de dados utilizada, observar

o comportamento das variáveis em cada nível hierárquico e identificar padrões. Para isso, serão utilizados gráficos e tabelas descritivas.

Em seguida, na Seção 4.2 serão exploradas as relações entre as covariáveis e a variável resposta, que é a proficiência em matemática. Essas relações podem ser entendidas a partir de gráficos comparativos e testes estatísticos.

Na fase de modelagem (4.3), será apresentado um modelo de regressão linear múltiplo, que conta apenas com as variáveis do aluno. Após isso, a abordagem multinível será utilizada, onde será introduzido, passo a passo, o modelo de 2 níveis com variáveis da escola. O cálculo dos parâmetros no modelo de 3 níveis não convergiu, porém, o modelo final contou com variáveis que apresentam o impacto das caracterísiscas da vizinhança no aprendizado. É importante relembrar que, para motivar uma abordagem hierárquica, será analisado o modelo nulo, que permite o cálculo do Coeficiente de Correlação Intraclasse (ICC), que indica a necessidade de modelagem multinível.

A seleção do melhor modelo será realizada com base em critérios estatísticos, como análise de resíduos, AIC e *deviance*, que auxiliam na escolha de um modelo que equilibre ajuste e simplicidade. Os resultados obtidos serão interpretados no contexto educacional, destacando os efeitos estimados para cada nível e suas implicações no desempenho dos estudantes.

4 Resultados

4.1 Análise Descritiva

Indubitavelmente, antes de avaliar a existência dos efeitos de Escola ou Vizinhança no desempenho dos estudantes no Distrito Federal (DF), é crucial visualizar os dados fornecidos de maneira descritiva. A análise descritiva, ou exploratória, auxilia em uma compreensão prévia do fenômeno estudado, o que é indispensável, dada a vasta complexidade de realidades do DF.

4.1.1 Perfil do aluno

Como dito na seção de Metodologia, a população de interesse é formada por estudantes do 5º ano do ensino fundamental das escolas públicas do Distrito Federal. Todos os dados dos alunos são provenientes da Prova Brasil de 2019 e do questionário socioeconômico da avaliação. A partir dessas informações, é possível avaliar padrões nos dados.

A análise descritiva do primeiro nível, do aluno, busca explorar as características individuais dos estudantes, utilizando as variáveis do banco de dados do SAEB. Esse passo inicial permite uma visão geral do perfil sociodemográfico e acadêmico analisado, além de identificar padrões gerais na distribuição da proficiência em Matemática.

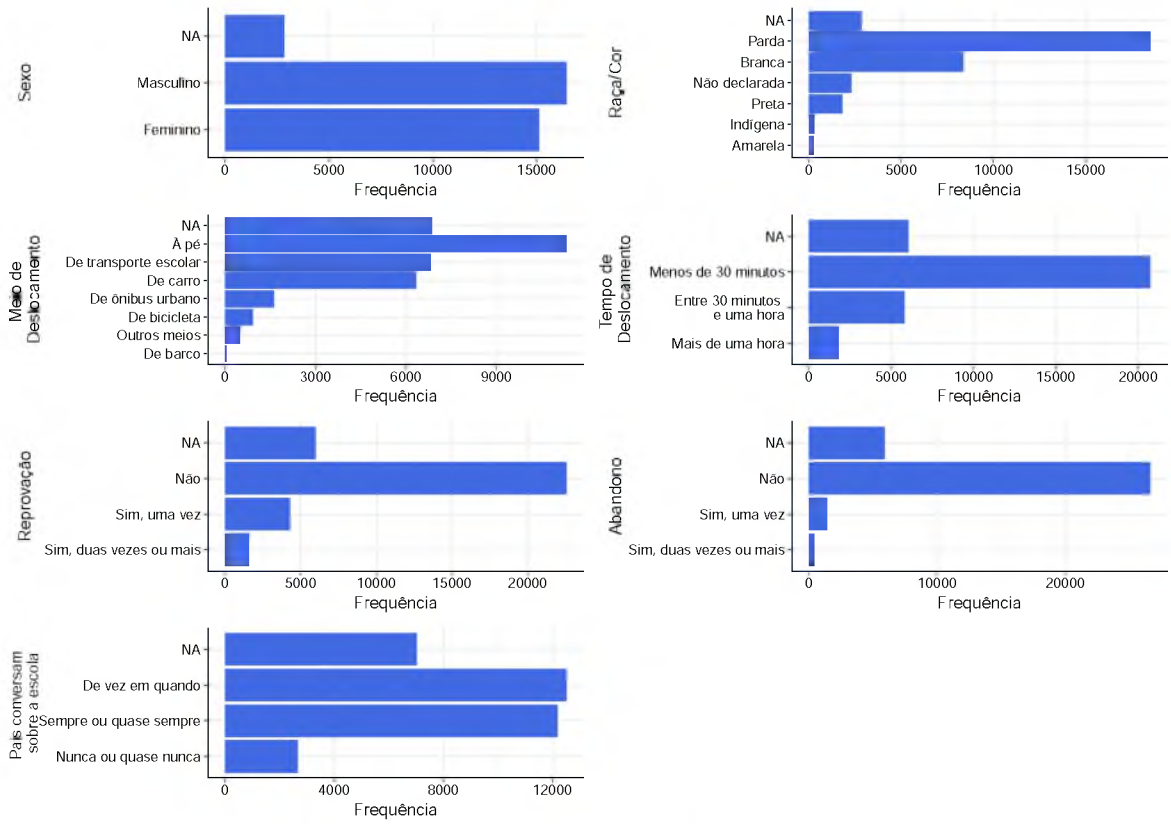
A partir das variáveis disponíveis, são apresentadas análises univariadas, que incluem frequências, proporções e estatísticas descritivas, como média, desvio-padrão e medidas de dispersão. Essas informações fornecem um panorama inicial sobre os estudantes do 5º ano do Ensino Fundamental nas escolas públicas do Distrito Federal, auxiliando na identificação de possíveis desigualdades e fatores relevantes para o desempenho acadêmico.

Tabela 3: Distribuição de frequências absolutas e relativas das variáveis qualitativas do 1º nível.

Variável	Categoria	Frequência Absoluta	Frequência Relativa (%)
Sexo	Feminino	15119	43,96
	Masculino	16428	47,77
	NA	2846	8,27
tp_cor_raca	Amarela	269	0,85
	Branca	8382	26,56
	Indígena	275	0,87
	Não declarada	2301	7,29
	Parda	18498	58,63
	Preta	1822	5,77
	NA	2846	8,27
meiodescocamento_q	De barco	52	0,15
	De bicicleta	914	2,66
	De carro	6336	18,42
	De transporte escolar	6812	19,81
	De onibus urbano	1613	4,69
	Outros meios de transporte	496	1,44
	À pé	11318	32,91
	NA	6852	19,92
descolamento_q	Entre 30 minutos e uma hora	5801	16,87
	Mais de uma hora	1805	5,25
	Menos de 30 minutos	20741	60,31
	NA	6046	17,58
reprovacao_q	Não	22557	65,59
	Sim, duas vezes ou mais	1580	4,59
	Sim, uma vez	4282	12,45
	NA	5974	17,37
abandono_q	Não	26630	77,43
	Sim, duas vezes ou mais	428	1,24
	Sim, uma vez	1413	4,11
	NA	5922	17,22
incentivo_paisa_q	De vez em quando	12513	36,38
	Nunca ou quase nunca	2664	7,75
	Sempre ou quase sempre	12195	35,46
	NA	7021	20,41
iluminacao_q	Não	1718	5,00
	Sim	26084	75,84
	NA	6591	19,16
pavimentacao_q	Não	4580	13,31
	Sim	23416	68,08%
	NA	6397	18,59%

Fonte: Elaboração própria.

Figura 4: Gráficos de barras das variáveis qualitativas do 1º nível.



Fonte: Elaboração própria.

A análise descritiva das variáveis qualitativas apresentadas na Tabela 3 fornece um panorama inicial sobre o perfil dos estudantes do 5º ano do ensino fundamental das escolas públicas do Distrito Federal, identificando possíveis padrões e desigualdades. Porém, é possível observar uma presença significante de dados faltantes. Por exemplo, nota-se que algumas variáveis, como abandono_q, deslocamento_q e incentivo_paisa_q, apresentam porcentagens consideráveis de dados ausentes (NA), variando de 17% a 20%. Esse fato reflete desafios relacionados à coleta ou preenchimento dos questionários socioeconômicos.

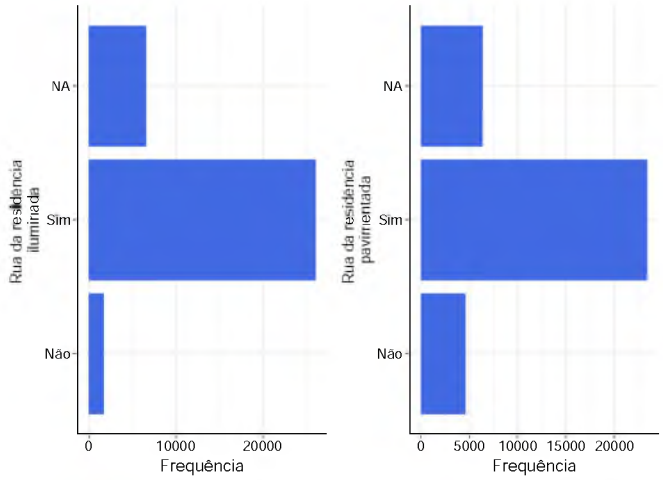
As variáveis relacionadas à diversidade étnico-racial (tp_cor_raca) indicam que a maioria dos estudantes se identificam como pardos (41,81%), seguido por Branco (19,74%). Já a proporção de sexo (tp_sexo) é equilibrada, com 43,96% do sexo feminino e 47,77% do sexo masculino. Quanto ao tempo de deslocamento até a escola (deslocamento_q), 60,31% dos estudantes reportaram trajetos inferiores a 30 minutos, refletindo uma possível proximidade geográfica da residência da maioria dos alunos com suas escolas. Outra variável que apresenta uma relação local entre as escolas e as comunidades atendidas é o Meio de Deslocamento, onde (meiodeslocament_q) mostra que 32,91% dos estudantes

chegam à escola a pé, seguido por 19,81% que utilizam transporte escolar e 18,42% que vão de carro.

Ademais, em relação ao desempenho escolar, a variável reprovacao_q evidencia que 65,59% dos estudantes nunca foram reprovados, mas 12,45% já repetiram ao menos um ano, apontando para potenciais barreiras ao progresso acadêmico. Na variável abandono_q, nota-se que 77,43% dos estudantes não reportaram histórico de abandono escolar, enquanto 5,35% indicaram algum nível de abandono, o que reforça a importância de investigar fatores associados à permanência escolar. Além disso, 35,46% dos estudantes assinalaram que os pais conversam sobre a escola "Sempre ou quase sempre". A minoria (7,75%) reportou que os pais "Nunca ou quase nunca"conversam sobre assuntos escolares.

É importante frisar que o questionário do SAEB também apresenta alguns itens relacionados a características da região onde os estudantes residem (Figura 5). Esses dados se tornam relevantes quando o conceito de Vizinhança é explorado.

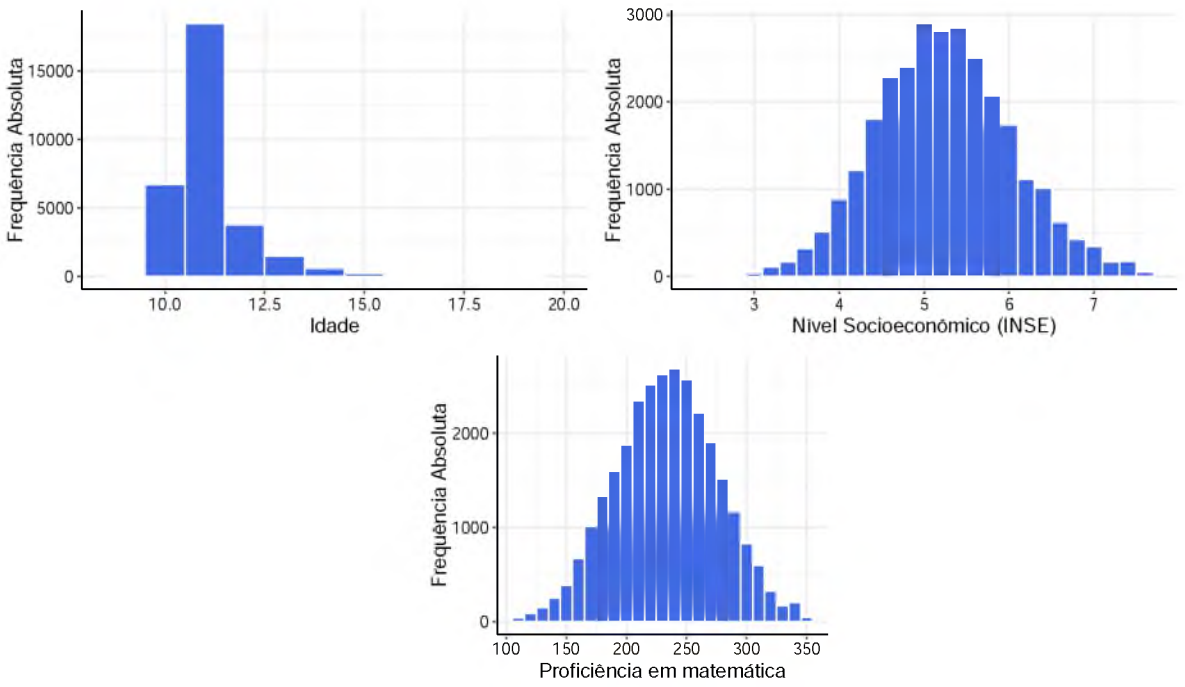
Figura 5: Gráficos de barras referentes a variáveis qualitativas atreladas à vizinhança dos alunos.



Fonte: Elaboração própria.

No que diz respeito à iluminação na rua que os estudantes residem (iluminacao_q), 75,84% dos estudantes indicaram possuir iluminação, uma variável que também pode ser interpretada como um indicador das condições socioeconômicas. Ainda nesse contexto, 68,08% (23.416) dos estudantes afirmaram que tem as ruas pavimentadas, enquanto 13,3% (4.580) afirmaram que são.

Figura 6: Histobramas das variáveis quantitativas do 1º nível.



Elaboração própria.

Tabela 4: Medidas descritivas das variáveis quantitatívas.

Variável	1º Quartil	Mediana	3º Quartil	Média	Desvio Padrão
INSE	4,680	5,210	5,760	5,235	0,801
Proficiência em matemática	204,4	233,8	262,3	233,3	42,502

Fonte: Elaboração própria.

A Tabela 4 apresenta os principais resumos descritivos das variáveis quantitativas INSE (Índice Socioeconômico) e Proficiência em Matemática dos estudantes do 5º ano das escolas públicas do Distrito Federal.

A variável INSE, que mede o nível socioeconômico dos estudantes, apresenta uma mediana de 5,210, com uma média ligeiramente superior de 5,235, indicando uma distribuição simétrica. O primeiro quartil (4,680) e o terceiro quartil (5,760) revelam uma amplitude interquartil relativamente estreita, o que sugere uma concentração dos valores em torno da mediana. O desvio padrão de 0,801 indica que as variações nos níveis socioeconômicos entre os estudantes não são muito acentuadas.

Já a Proficiência em Matemática apresenta uma mediana de 233,8, próxima da média de 233,3, sugerindo que a distribuição também é simétrica. O primeiro quartil

(204,4) e o terceiro quartil (262,3) mostram uma amplitude interquartil maior em comparação ao INSE, indicando maior dispersão nos níveis de desempenho em matemática. O desvio padrão de 42,502 confirma essa maior variabilidade, evidenciando diferenças significativas no desempenho acadêmico entre os estudantes.

Essa análise preliminar sugere a relevância de explorar como o contexto socioeconômico e outros fatores podem influenciar diretamente o desempenho acadêmico, o que será aprofundado na aplicação do modelo de regressão multinível.

4.1.2 Nível escola

Para dar continuidade à análise, é essencial explorar o contexto escolar. Nesta seção, serão analisadas variáveis relacionadas à infraestrutura das escolas, fornecidas pelo Censo Escolar, e variáveis criadas a partir do nível aluno, utilizando técnicas de agregação, que são habitualmente usadas em análises multinível, de acordo com Hox, Moerbeek e Schoot (2017). Essa abordagem permite captar aspectos coletivos que caracterizam o ambiente escolar e compreender como fatores contextuais, como condições estruturais e características médias dos estudantes, podem influenciar o aprendizado.

Tabela 5: Proporção de infraestrutura nas escolas.

Nome da Variável	Proporção 1 (%)	Proporção 0 (%)	Diferença (%)
IN_DORMITORIO_PROFESSOR	0,00	100,00	-100,00
IN_DEPENDENCIAS_OUTRAS	0,00	100,00	-100,00
IN_AGUA_FONTE_RIO	0,15	99,85	-99,71
IN_ESGOTO_INEXISTENTE	0,15	99,85	-99,71
IN_LIXO_ENTERRA	0,15	99,85	-99,71
IN_DORMITORIO_ALUNO	0,15	99,85	-99,71
IN_AGUA_CACIMBA	0,59	99,41	-98,83
IN_LIXO_QUEIMA	1,17	98,83	-97,66
IN_AGUA_POCO_ARTESIANO	8,20	91,80	-83,60
IN_AUDITORIO	13,62	86,38	-72,77
IN_LABORATORIO_CIENCIAS	14,49	85,51	-71,01
IN_ESGOTO_FOSSA	14,79	85,21	-70,42
IN_BIBLIOTECA	24,01	75,99	-51,98
IN_REFEITORIO	25,62	74,38	-48,76
IN_ALMOXARIFADO	31,48	68,52	-37,04
IN_QUADRA_ESPORTES_COBERTA	34,85	65,15	-30,31
IN_BANHEIRO_EI	40,41	59,59	-19,18
IN_QUADRA_ESPORTES_DESCOBERTA	40,70	59,30	-18,59
IN_PATIO_DESCOBERTO	48,17	51,83	-3,66
IN_PARQUE_INFANTIL	54,03	45,97	8,05
IN_AREA_VERDE	55,34	44,66	10,69
IN_BANHEIRO_CHUVEIRO	57,10	42,90	14,20
IN_LABORATORIO_INFORMATICA	64,57	35,43	29,14
IN_QUADRA_ESPORTES	66,91	33,09	33,82
IN_SALA_ATENDIMENTO_ESPECIAL	74,82	25,18	49,63
IN_PATIO_COBERTO	77,75	22,25	55,49
IN_SALA_LEITURA	78,62	21,38	57,25
IN_DESPENSA	79,65	20,35	59,30
IN_ESGOTO_REDE_PUBLICA	85,36	14,64	70,72
IN_BANHEIRO_PNE	85,51	14,49	71,01
IN_BIBLIOTECA_SALA_LEITURA	87,26	12,74	74,52
IN_AGUA_REDE_PUBLICA	92,83	7,17	85,65
IN_SECRETARIA	97,36	2,64	94,73
IN_INTERNET	97,51	2,49	95,02
IN_COZINHA	97,66	2,34	95,31
IN_BANDA_LARGA	97,90	2,10	95,80
IN_SALA_PROFESSOR	98,83	1,17	97,66
IN_SALA_DIRETORIA	99,41	0,59	98,83
IN_LOCAL_FUNC_PREDIO_ESCOLAR	99,71	0,29	99,41
IN_LIXO_SERVICO_COLETA	99,85	0,15	99,71
IN_ENERGIA_REDE_PUBLICA	100,00	0,00	100,00

Fonte: Elaboração própria.

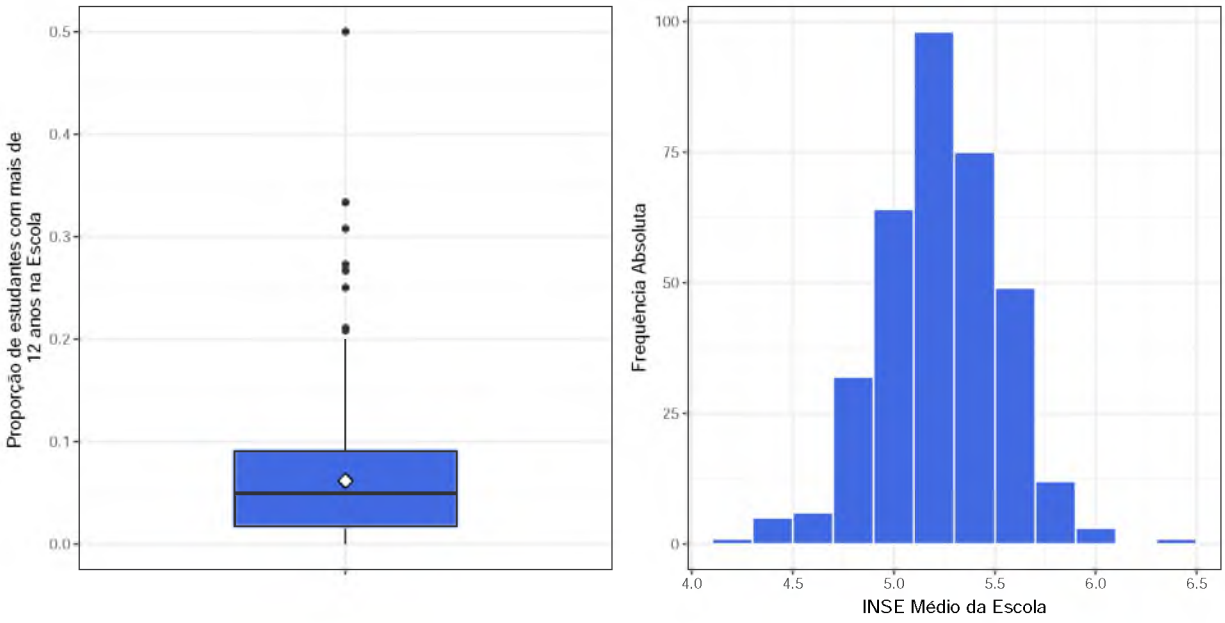
A Tabela 5 apresenta a distribuição percentual das variáveis relacionadas à infraestrutura das escolas públicas do Distrito Federal, com base nos dados do Censo Escolar. Essas variáveis, codificadas em 0 (infraestrutura inexistente) e 1 (infraestrutura presente), fornecem um panorama sobre as condições estruturais das escolas.

A visualização da Tabela mostra algumas disparidades em relação a existência de alguns ambientes. Por exemplo, apenas 14,49% das escolas possuem laboratórios de ciências, e 24,01% possuem bibliotecas completas.

Por outro lado, de maneira geral, os dados indicam um elevado padrão de homogeneidade em muitas variáveis de infraestrutura básica. Por exemplo, itens indispensáveis para o funcionamento das escolas, como energia elétrica da rede pública (100%) e coleta de lixo por serviço público (99,85%), estão presentes em praticamente todas as escolas. Estruturas administrativas, como sala de diretoria (99,41%) e secretaria (97,36%), também são amplamente distribuídas, reforçando a uniformidade nos aspectos fundamentais de organização escolar.

Os resultados da tabela reforçam uma hipótese de homogeneidade nas variáveis de infraestrutura básica entre as escolas do Distrito Federal. A elevada frequência de presença (códigos "1") em muitos desses itens sugere que praticamente todas as escolas atendem a um padrão mínimo de funcionamento.

Figura 7: Gráficos das variáveis quantitativas agregadas para o nível escola.



Fonte: Elaboração própria.

Tabela 6: Medidas descritivas das variáveis quantitatívas do 2º nível.

Variável	1º Quartil	Mediana	3º Quartil	Média	Desvio Padrão
INSE Médio	5,055	5,241	5,442	5,233	0,30
Defasagem idade-série	1,70%	4,96%	9,05%	6,15%	6,22%

Fonte: Elaboração própria.

Como já explicado, a inclusão e análise dessas variáveis agregadas no modelo multinível será essencial para compreender como características coletivas das escolas influenciam o desempenho dos estudantes. O INSE Médio oferece uma visão da composição socioeconômica de cada escola, o que pode influenciar diversos fatores. Ademais, a Defasagem idade-série é um indicador crítico muito utilizado, que é comumente usado em análises educacionais. Neste estudo, a defasagem idade-série é considerada como a proporção de estudantes com mais de 12 anos na escola.

Ao analisar a Tabela 6 e a Figura 7, pode-se notar que a variável INSE Médio, que representa a média do Índice Socioeconômico (INSE) dos estudantes por escola, apresenta uma mediana de 5,241 e uma média de 5,233, com desvio padrão de 0,30. Esses valores indicam pouca dispersão entre as escolas, sugerindo uma homogeneidade relativa nos níveis socioeconômicos médios das unidades escolares. O intervalo interquartil, com 1º quartil em 5,055 e 3º quartil em 5,442, reforça essa baixa variabilidade. Esse padrão é consistente com as análises de infraestrutura, evidenciando que as escolas do Distrito Federal compartilham condições socioeconômicas semelhantes em termos médios.

Além disso, a variável que indica defasagem idade-série, apresenta uma maior variabilidade. A mediana é de 4,96%, enquanto a média é de 6,15%, com um desvio padrão de 6,22%. O intervalo interquartil, entre 1,70% e 9,05%, sugere que, enquanto algumas escolas apresentam percentuais baixos de atraso escolar, outras enfrentam taxas elevadas, chegando até 50%. Essa variação pode indicar diferenças relevantes entre as escolas no que diz respeito à progressão dos alunos.

4.1.3 Nível vizinhança

Nesta seção, será realizada a análise descritiva univariada das variáveis que caracterizam os efeitos regionais nos dados utilizados neste trabalho. A análise baseia-se, principalmente, nos dados fornecidos pela Pesquisa Distrital por Amostra de Domicílios (PDAD), que disponibiliza indicadores em uma escala de 0 a 100. Esses indicadores representam a porcentagem de domicílios em cada Região Administrativa (RA) do Distrito Federal que possuem ou não determinadas características, como acesso a serviços públicos, condições de habitação, entre outros.

Com o objetivo de aprofundar a análise e explorar melhor a relação entre a vizinhança e os resultados escolares, foi decidido categorizar as RAs do Distrito Federal em quatro grupos econômicos: baixa renda, média baixa, média alta e alta renda. Essa categorização permitirá compreender como os fatores regionais se distribuem de forma diferenciada entre os grupos e como essas diferenças podem impactar o desempenho dos estudantes.

Antes de avançar para a análise descritiva dos indicadores regionais, é essencial compreender como as escolas estão distribuídas entre as RAs do Distrito Federal. Essa etapa inicial é fundamental para contextualizar a análise, destacando desigualdades regionais na distribuição das unidades escolares. É importante frisar que as escolas listadas são as que participaram da Prova Brasil 2019, e onde os alunos preencheram os questionários socioeconômicos.

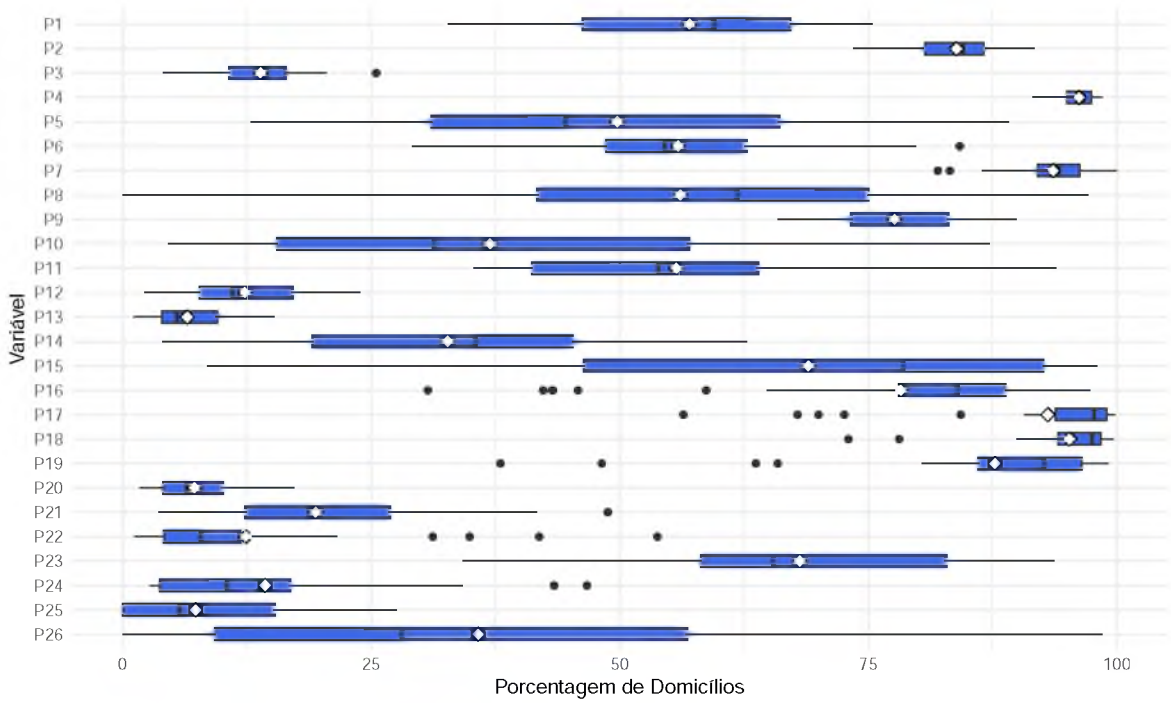
Tabela 7: Distribuição de escolas entre as Regiões Administrativas do DF.

Região Administrativa (RA)	Quantidade de escolas
Ceilândia	52
Planaltina	40
Plano Piloto	33
Gama	27
Taguatinga	25
Samambaia	23
Paranoá	19
Brazlândia	18
Recanto das Emas	14
Santa Maria	14
São Sebastião	14
Sobradinho	13
Fercal	7
Guará	6
Riacho Fundo II	6
Sobradinho II	5
Cruzeiro	4
Núcleo Bandeirante	4
Riacho Fundo	4
Park Way	3
Sol Nascente/Pôr do Sol	3
Arniqueira	2
Candangolândia	2
Lago Norte	2
Scia e Estrutural	2
Vicente Pires	2
Itapoã	1
Jardim Botânico	1
Sudoeste e Octogonal	1
Varjão	1

Tabela 8: Resultados dos indicadores da PDAD (%).

Variável	1º Quartil	Mediana	3º Quartil	Média	Desvio Padrão
P1	47,10	59,30	66,90	57,03	12,46
P2	80,70	84,40	86,30	83,87	4,55
P3	10,80	14,30	16,40	13,87	4,38
P4	95,00	96,60	97,40	96,20	1,69
P5	31,20	44,70	65,20	49,80	22,01
P6	48,60	54,80	62,80	55,88	10,98
P7	92,42	94,15	96,08	93,57	4,30
P8	42,20	61,00	74,90	56,12	23,23
P9	73,40	77,50	83,00	77,66	6,90
P10	16,00	34,00	56,90	36,94	25,41
P11	41,40	53,30	63,20	55,59	15,99
P12	7,75	11,20	16,75	12,27	6,23
P13	4,00	5,40	9,45	6,44	3,86
P14	19,35	34,40	44,95	32,73	16,60
P15	47,60	74,50	92,30	69,13	26,50
P16	78,30	83,80	88,40	78,40	18,06
P17	94,20	97,70	99,00	93,12	10,71
P18	94,30	97,50	98,40	95,27	5,75
P19	86,70	92,50	96,40	87,85	14,34
P20	4,15	6,15	9,90	7,17	4,19
P21	12,60	17,70	26,05	19,33	11,06
P22	4,20	7,80	10,70	12,23	12,86
P23	58,70	65,80	82,90	68,03	16,33
P24	3,82	9,85	16,80	14,00	13,54
P25	0,00	5,80	15,20	7,38	7,60
P26	9,30	28,05	56,77	35,77	30,45

Figura 8: Boxplots dos indicadores selecionados da PDAD



Elaboração própria

A análise descritiva dos indicadores da Pesquisa Distrital por Amostra de Domicílios (PDAD) revelou variações importantes entre os indicadores. De forma geral, os indicadores apresentam distribuições variadas, com algumas variáveis mostrando maior dispersão e diferenças notórias entre a média e a mediana, indicando assimetria nas distribuições.

Os indicadores P10 (proporção da população com 25 anos ou mais de idade com ensino superior completo) e P26 (proporção de domicílios com rendimento bruto domiciliar maior que 5 salários mínimos) destacam-se pela alta dispersão, provando certa desigualdade entre as Regiões Administrativas (RAs). Para P10, a média foi de 36,94%, enquanto a mediana foi de 34,00%, indicando uma distribuição levemente assimétrica. O desvio padrão elevado (25,41) reflete a heterogeneidade entre as regiões. De maneira similar, P26 apresentou uma média de 35,77% e mediana de 28,05%, com um desvio padrão de 30,45, reforçando a grande variabilidade nas condições de renda das RAs.

Alguns indicadores apresentaram pouca dispersão, como P4 (porcentagem de domicílios que acessaram internet nos últimos três meses), cuja média (96,20%) e mediana (96,60%) são muito próximas, e o desvio padrão é baixo (1,69), indicando uniformidade no acesso ă internet. Outro exemplo é P18 (rua de acesso aos domicílios com iluminação), com média de 95,27%, mediana de 97,50% e desvio padrão de 5,75, o que reflete um padrão

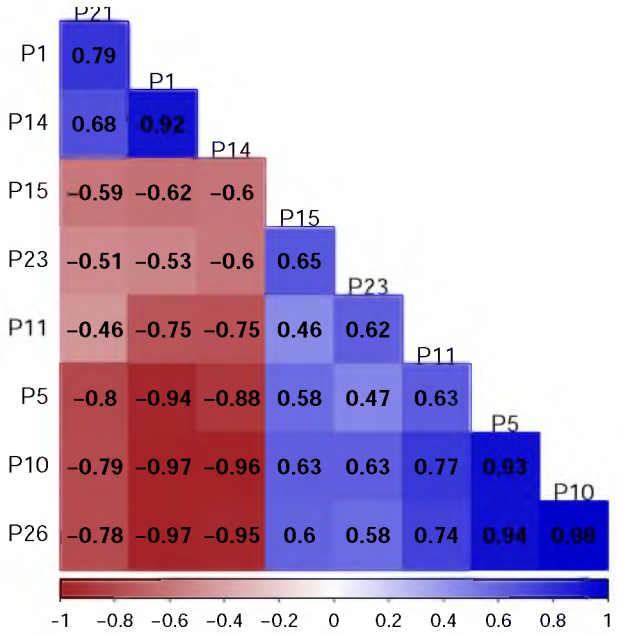
de infraestrutura em geral. Além disso, outros indicadores relacionados à infraestrutura urbana, como P17 (rua asfaltada ou pavimentada) e P22 (esgoto a céu aberto nas cercanias dos domicílios), mostram um contraste interessante. Enquanto P17 apresenta altos valores de média (93,12%) e mediana (97,70%), sugerindo boas condições gerais, P22 possui uma média relativamente baixa (12,2%) e grande dispersão (desvio padrão de 12,86), indicando que o problema do esgoto é mais localizado nos outliers apresentados na Figura 8, que correspondem às RAs Fercal (52%), Estrutural (41%), Sol Nascente/Pôr do Sol (34,9%) e São Sebastião (31%).

Por fim, é importante observar a consistência nos indicadores relacionados à educação e ao acesso à tecnologia, como P7 (porcentagem de jovens entre 15 e 17 anos que frequentam escolas), com média de 93,57% e desvio padrão 4,30. Já indicadores como P5 (acesso à internet por computador/laptop) mostram maior variabilidade, com desvio padrão de 22,01, refletindo possível desigualdades no acesso a tecnologia.

Essas análises destacam o contexto heterogêneo das Regiões Administrativas do Distrito Federal, especialmente em relação a fatores cruciais como renda, educação e infraestrutura. As variáveis P10 e P26, devido à sua alta dispersão e importância para o modelo multinível, merecem atenção especial nas análises subsequentes.

O correlograma da Figura 9 foi elaborado para explorar possíveis correlações entre algumas variáveis do conjunto de dados: P1 (Porcentagem da população preta ou parda), P5 (Porcentagem de domicílios com acesso à internet por meio de computador/laptop), P10 (Proporção da população com 25 anos ou mais de idade, com ensino superior completo), P11 (Proporção da população com tempo de deslocamento de até 30 minutos até o trabalho), P14 (Porcentagem da população que utiliza ônibus como principal meio de transporte para o trabalho), P15 (Domicílios ocupados em lotes regularizados), P21 (Acúmulo de entulho nas cercanias dos domicílios), P23 (Policiamento regular nas proximidades) e P26 (Porcentagem de domicílios com rendimento bruto domiciliar maior que 5 salários mínimos). No Correlograma apresentado, as cores representam a força da correlação: tons mais próximos do vermelho indicam associações negativas fortes, enquanto os tons azuis refletem correlações positivas significativas.

Figura 9: Correlograma de alguns indicadores da PDAD.



Fonte: Elaboração própria.

A seleção dessas variáveis foi baseada na análise de sua dispersão, indicando variabilidades relevantes para investigações mais aprofundadas. Além de serem variáveis correlacionadas de maneira moderada ou forte com os principais indicadores de vizinhança do modelo que será apresentado (P10 e P26).

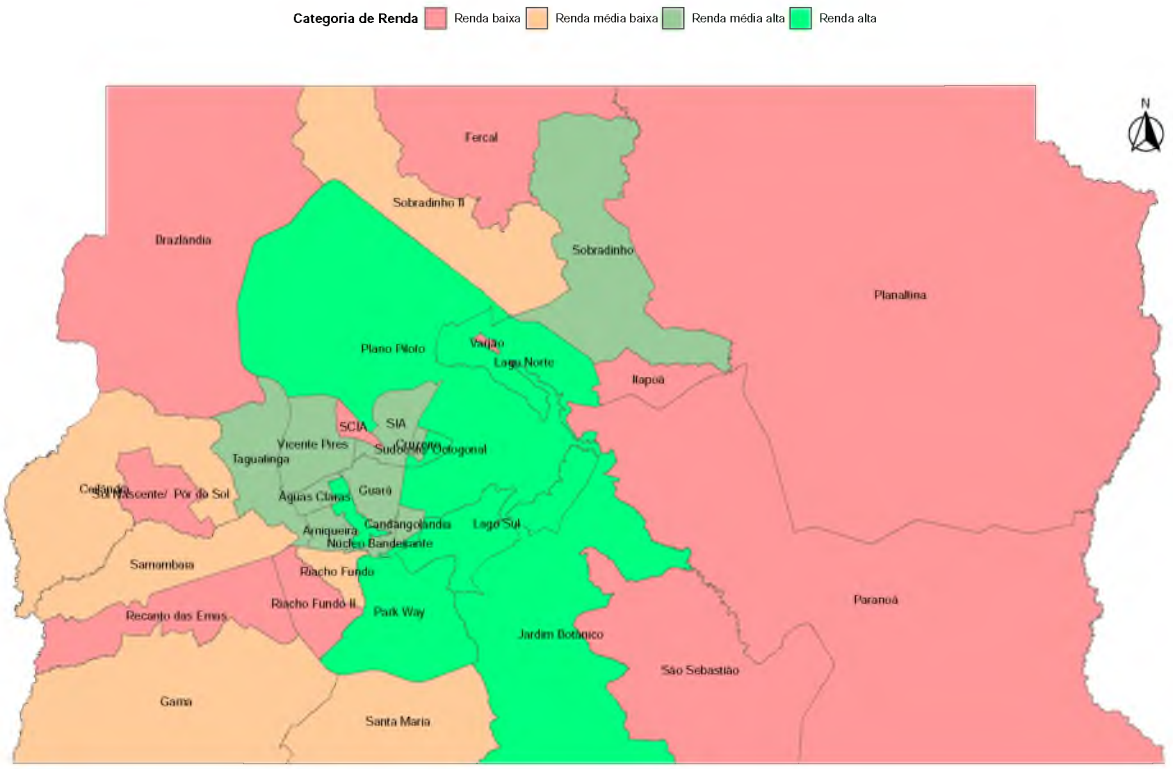
A partir da visualização da Figura 9, é possível notar uma forte correlação positiva (0,98) entre P10 e P26, indicando que regiões com maior proporção de pessoas com ensino superior completo apresentam uma concentração significativamente maior de domicílios com rendimentos mais elevados. Esse resultado reforça a relação direta entre escolaridade e renda.

Por outro lado, a proporção de população preta ou parda (P1) apresentou correlações negativas com P10 (-0,97) e P26 (-0,97), sugerindo que quanto maior a população de pretos e pardos na RA, menor será o acesso à educação superior e a melhores condições econômicas. Outro aspecto relevante é a relação entre P14 (uso de ônibus como principal meio de transporte) e P10/P26, que apresentou correlações negativas significativas (-0,96 e -0,95, respectivamente).

Além disso, a análise das correlações revelou que P5 (porcentagem de domicílios com acesso à internet por meio de computador/laptop) está positivamente associada a P26 (0,94), reforçando a ideia de que o acesso à tecnologia está concentrado em regiões com maior rendimento.

Esses achados deixam claro como diferentes variáveis sociais, econômicas e infraestruturais estão interligadas, criando um panorama de desigualdade no Distrito Federal, de acordo com a Região Administrativa. Neste contexto, dada a discussão deste trabalho sobre a segregação socioespacial, é importante visualizar como as Categorias de Renda estão distribuídas geograficamente no Distrito Federal a partir da Figura 10.

Figura 10: Mapa das Regiões Administrativas do DF e suas respectivas Categorias de Renda.

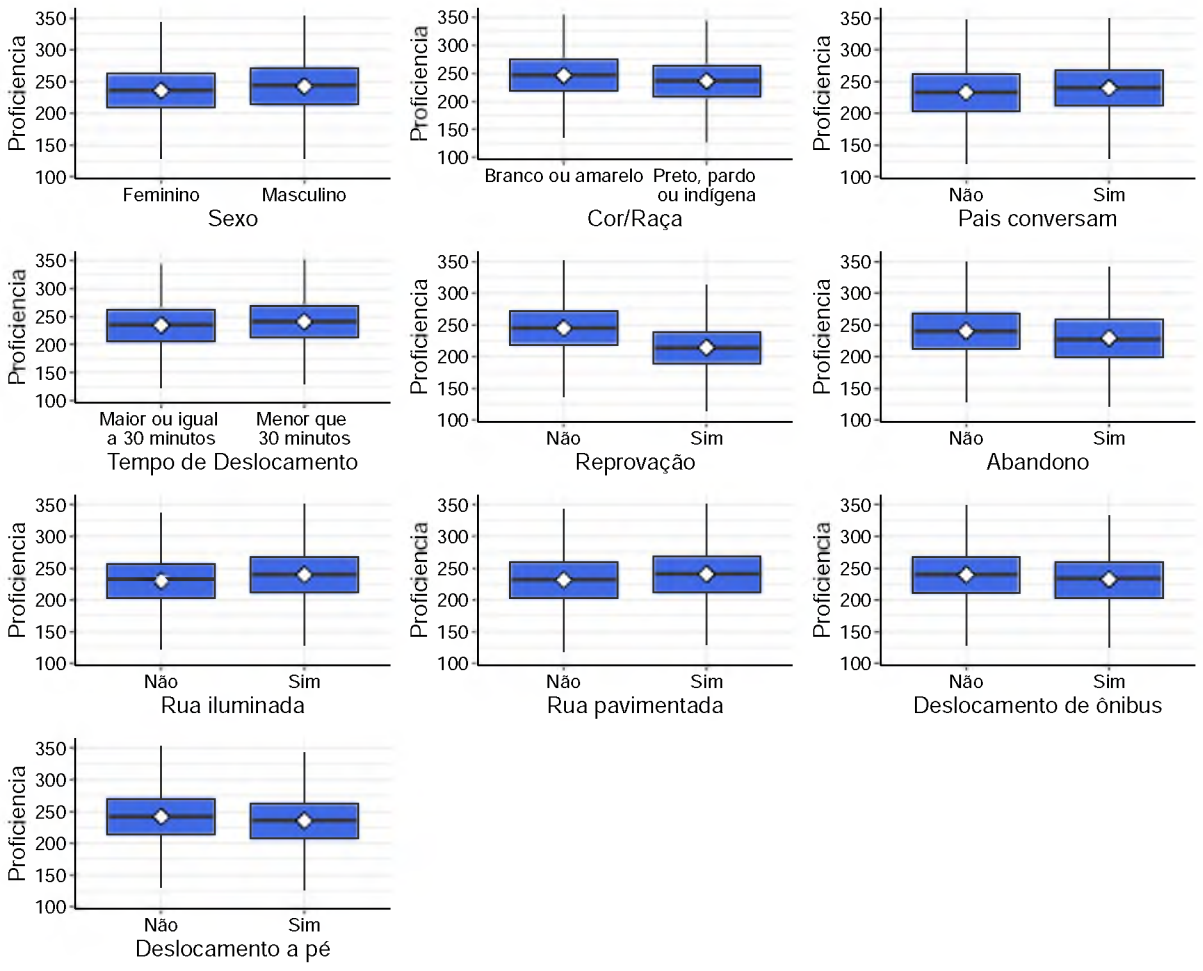


Fonte: Elaboração própria.

4.2 Análise Bivariada

Após a análise descritiva univariada, que forneceu um panorama geral das distribuições das variáveis nos três níveis do estudo, esta subseção foca na análise bivariada. O principal objetivo aqui é explorar as associações entre as variáveis explicativas identificadas como relevantes e a variável resposta, a proficiência em matemática. Essa etapa é essencial para compreender como diferentes fatores, como características individuais, escolares e de vizinhança, se relacionam com o desempenho acadêmico, possibilitando identificar padrões e direcionar hipóteses para os modelos estatísticos subsequentes. Além disso, as análises bivariadas permitem verificar a presença de associações estatisticamente significativas, auxiliando na construção do modelo multinível. Para isso, serão utilizadas técnicas de correlação e comparação de médias, além de gráficos exploratórios, considerando a classificação das variáveis (qualitativa ou quantitativa).

Figura 11: Boxplots das variáveis qualitativas do 1º nível, por proficiência em matemática.



Fonte: Elaboração própria.

Em relação ao Sexo, observou-se que os estudantes do sexo masculino apresentam maior média de proficiência (242,5) em comparação às estudantes do sexo feminino (235,8). Esse padrão também é consistente na mediana, com 243,9 para os homens e 236,0 para as mulheres. Ademais, quanto à Cor/Raça, alunos autodeclarados como "Branco ou amarelo"obtiveram média de proficiência mais alta (245,8) em relação aos estudantes classificados como "Preto, pardo ou indígena-- PPI (236,1).

A variável relacionada ao incentivo parental, representada por pais que conversam sobre a escola, mostrou-se relevante. Estudantes que relataram receber esse tipo de incentivo tiveram uma média de proficiência maior (239,8) do que aqueles que não o recebem (233,1).

No que se refere ao tempo de deslocamento até a escola, alunos que gastam menos de 30 minutos para chegar apresentaram média de proficiência mais alta (240,8) em comparação aos que têm um deslocamento maior ou igual a 30 minutos (234,6). Ainda em relação ao deslocamento,

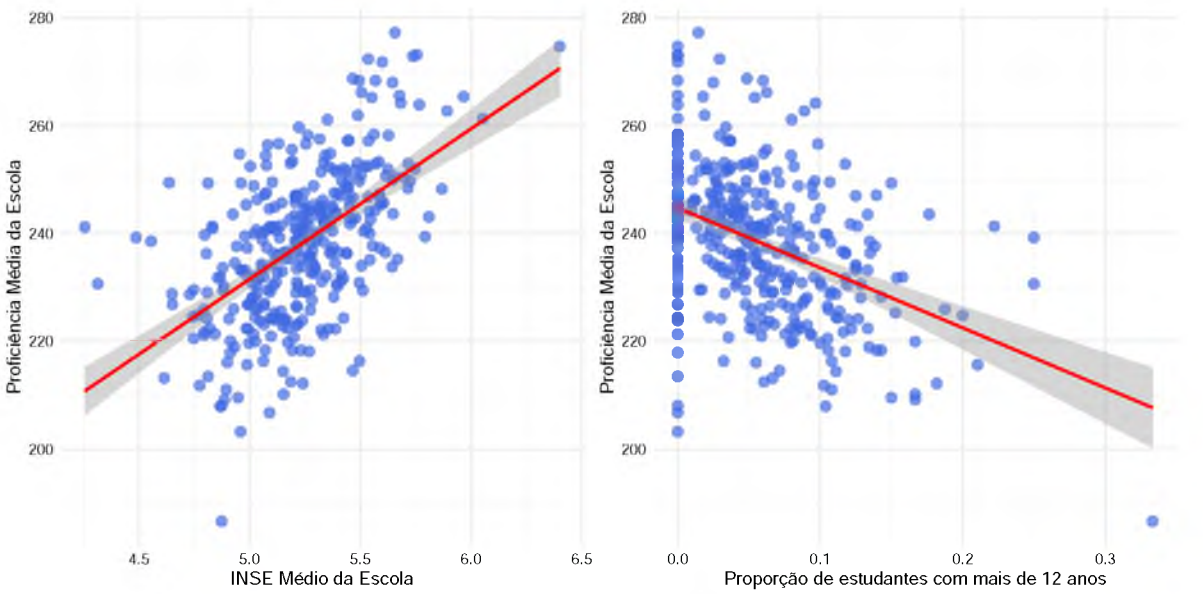
Além disso, as categorias das variáveis de reprovação e abandono escolar também apresentaram diferenças. Estudantes que nunca foram reprovados tiveram uma média de proficiência superior (244,7) àqueles que reprovaram pelo menos uma vez (214,3) e alunos que nunca abandonaram a escola apresentaram uma proficiência média de 239,7, enquanto aqueles que relataram abandono tiveram uma média menor (229,1).

Por fim, alunos que residem em áreas com iluminação pública apresentaram média de proficiência mais alta (239,7) em comparação aos que vivem em regiões sem iluminação (230,0). De maneira similar, estudantes de áreas com ruas pavimentadas apresentaram média superior (240,6) àqueles que vivem em locais sem pavimentação (231,5).

Para testar a significância estatística das diferenças observadas entre os grupos, foi aplicado o teste de Mann-Whitney Wilcoxon a todas as variáveis categóricas consideradas na análise. Em todos os casos, a hipótese nula de igualdade das distribuições foi rejeitada, indicando que as diferenças entre os grupos não ocorreram ao acaso. No entanto, é importante ressaltar que esse teste é muito sensível a grandes amostras, como as utilizadas neste estudo.

Nos gráficos a seguir (Figura 12), também é possível avaliar relação entre características quantitativas das escolas e a proficiência média dos estudantes. Neste contexto, foram utilizados os indicadores INSE Médio e Defasagem idade-série da escola, que foram explorados na Figura 7 e Tabela 6.

Figura 12: Gráficos de Dispersão de INSE Médio e Defasagem idade-série, por Proficiência Média.



Fonte: Elaboração própria.

Tabela 9: Resultados dos Testes de Correlação de Pearson.

Variáveis		Correlação	p-valor
Proficiência média	INSE Médio	0,561	<0,001
	Desafagem idade-série	-0,403	<0,001

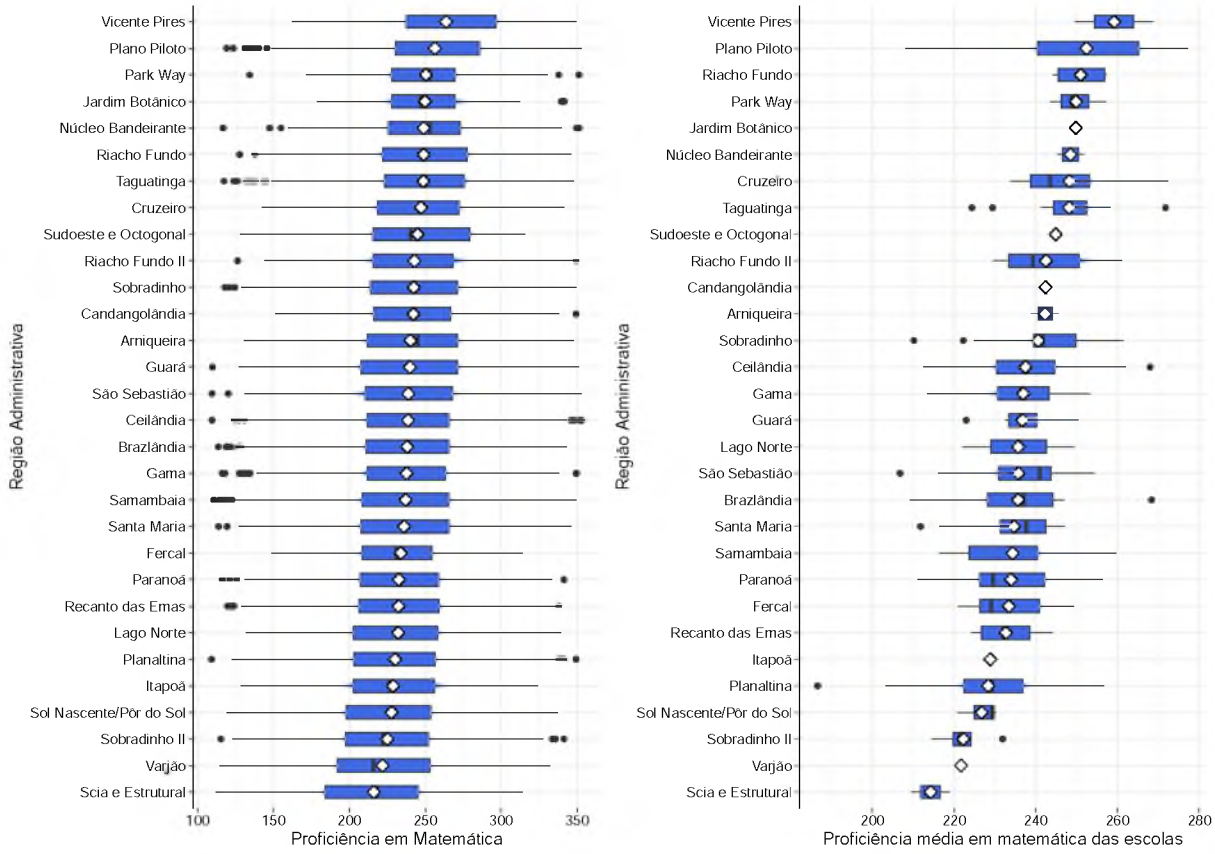
Fonte: Elaboração própria.

A análise dos gráficos de dispersão da Figura 12 sugere a existência de associação entre o INSE Médio dos estudantes da escola, a defasagem idade-série média da escola e a proficiência média em matemática. Logo, a partir do Teste de Correlação Linear de Pearson, as variáveis INSE Médio e a proficiência média tem uma correlação positiva moderada ($r = 0,561$, $p < 0,001$), indicando que escolas com um alunos com nível socioeconômico mais elevado tendem a ter melhores desempenhos. Por outro lado, a proporção de estudantes com defasagem idade-série apresenta uma correlação negativa moderada com a proficiência ($r = -0,403$, $p < 0,001$). Portanto, ao considerar um nível de significância de 5%, é possível afirmar que quanto maior o percentual de estudantes com mais de 12 anos no 5º ano do ensino fundamental, menor é a proficiência média em matemática na escola.

Dando continuidade à análise, é importante explorar mais o perfil das vizinhanças para compreender como as características socioeconômicas e demográficas das Regiões Administrativas (RAs) influenciam no desempenho dos estudantes. Essa abordagem dialoga

diretamente com o conceito de Efeito Vizinhança, apresentado neste trabalho. Logo, primeiramente, serão apresentados cruzamentos de proficiências por RA, onde uma conexão com a realidade econômica regional se tornará indispensável.

Figura 13: Boxplots das Proficiências médias em matemática das escolas, por Região Administrativa.



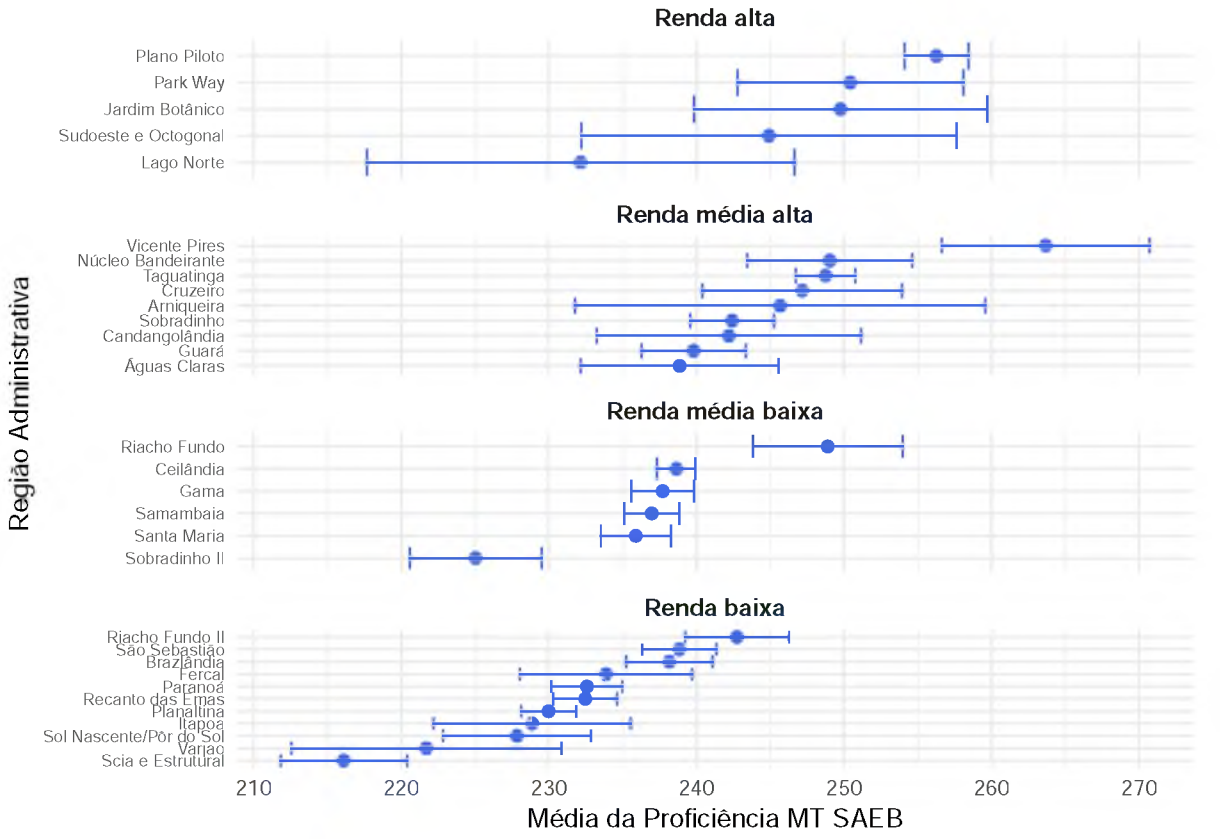
Fonte: Elaboração própria.

A partir do primeiro gráfico da Figura 13, é possível observar pequenas disparidades entre as proficiências médias dos alunos, por RA. As RAs com as maiores médias de proficiência foram Vicente Pires (264), porém com um desvio Padrão de 42,2 pontos de proficiência. Plano Piloto (com uma média de 256 e desvio padrão de 41,1), Park Way (250,39) e Jardim Botânico (com proficiência média de 250 pontos e desvio padrão de 35,7). Essas regiões estão associadas a melhores condições econômicas, como se pode observar na Figura 10.

Não obstante, as RAs com as menores médias de proficiência foram Scia/Estrutural, Varjão, Sobradinho II e Sol Nascente, com desvios padrão elevados. É importante frisar que essas regiões estão associadas a piores condições de renda, vide Figura 10. Esse padrão pode ser analisado para todas as Regiões.

A grande dispersão dos dados e o alto número de *outliers* apresentados no primeiro grăfico da Figura 13 motivaram a construção do segundo, que cruza a Proficiência média das escolas com a RA a que pertencem. Essa abordagem auxilia numa visualização mais clara das diferenças analisadas. De qualquer forma, a nova ordenação de RAs trouxe pouca diferença em relação ă anterior. Isso somado à notória influência da Categoria de Renda a qual a RA pertence, fortalece a hipótese da existência do Efeito Vizinhança, principalmente ligado ă Renda. Tal comportamento pode ser visualizado na Figura a seguir.

Figura 14: Boxplots das médias de Proficiência, com intervalo de confiança, por RA.

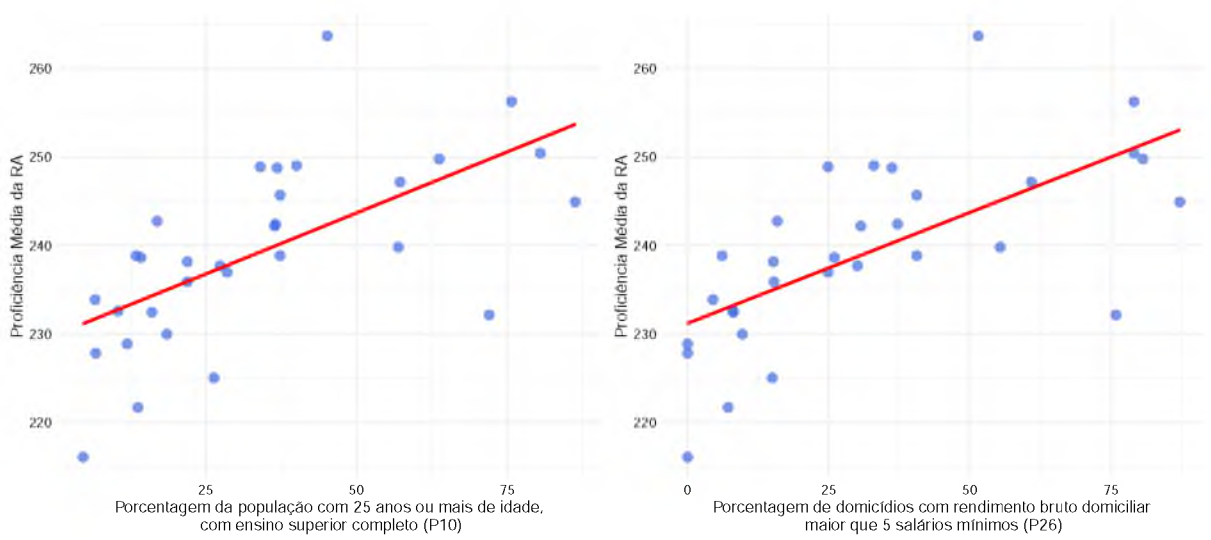


Fonte: Elaboração própria.

Ademais, é interessante compreender mais a fundo a influência de indicadores da PDAD com o desempenho dos estudantes que estudam em diferentes RAs do Distrito Federal. Dessa forma, foi feita uma anălise da correlação entre as variáveis P10 (proporção da população com ensino superior completo) e P26 (porcentagem de domicílios com rendimento superior a 5 salários mínimos) com a proficiência média dentro das RAs. Essas variáveis foram importantes para a construção do Modelo Multinível e refletem aspectos

importantes já discutidos na Subseção 4.1.3.

Figura 15: Gráficos de Dispersão da Proporção da população da RA com Ensino Superior Completo e Renda domiciliar maior que 5 salários mínimos por Proficiência Média.



Fonte: Elaboração própria.

Tabela 10: Resultados dos Testes de Correlação de Pearson.

Variáveis		Correlação	p-valor
Proficiência média	P10	0,626	<0,001
	P26	0,661	<0,001

Fonte: Elaboração própria.

Ao visualizar os gráficos de dispersão na Figura 15, é possível notar um que à medida que os valores de P10 e P26 aumentam, o desempenho médio em matemática na RA também aumenta. Dado isso, vale testar se há associação linear significativa entre os indicadores e a proficiência média. Dessa forma, a partir do Teste de Correlação de Pearson (Tabela 10, pode-se afirmar que a variável P10 e a proficiência média da RA têm uma correlação positiva moderada ($r = 0,626$, $p < 0,001$), indicando que em RAs com maior P10 (porcentagem da população com ensino superior completo) tem maiores notas médias. Outrossim, há também uma correlação positiva e moderada entre P26 e proficiência média da RA ($r = 0,661$, $p < 0,001$).

As análises anteriores destacam como as médias de proficiência tendem a variar entre as RAs de diferentes categorias de renda, evidenciando diferenças visualmente relavantes. Por fim, na etapa final desta Seção, as características dos estudantes das RAs

serão exploradas por gráficos de barras ordenadas pela proficiência média dos estudantes.

4.3 Modelagem

A partir da análise descritiva apresentada nas seções anteriores e com o passo a passo apresentado no Referencial Teórico, neste tópico será apresentado o modelo de regressão multinível proposto. É importante relembrar que a variável resposta é a proficiência em matemática dos estudantes do 5º ano do Ensino Fundamental das escolas públicas do Distrito Federal, metrificada em uma escala de 0 a 500 pontos. Além disso, todas as variáveis explicativas utilizadas já foram abordadas e descritas anteriormente neste relatório.

A construção do modelo final será dividida em três etapas. Primeiramente, será apresentado um modelo do 1º Nível (Aluno) utilizando a Regressão Linear Múltipla, sem uma abordagem multinível. Após isso, será incorporado o "Efeito Escola"(2º Nível), com suas respectivas variáveis explicativas (Efeitos Fixos). O Modelo Final, por fim, contará com a adição de variáveis das Regiões Administrativas das escolas, trazendo o impacto do "Efeito Vizinhança".

4.3.1 1º Nível - Aluno

O modelo de regressão linear múltiplo apresentado na Tabela 11 tem como objetivo analisar os fatores que influenciam a proficiência em Matemática no nível individual, sem a utilização da técnica multinível.

Os resultados do modelo de regressão linear abaixo revelam a influência significativa de diversas variáveis individuais na proficiência em Matemática já apresentadas durante o trabalho. Observa-se que ser do sexo masculino (8,301) e pertencer à raça/cor branca ou amarela (4,920) estão associados a maiores proficiências, assim como um maior INSE do aluno (4,258). Porém, reprovação (-24,060) e abandono (-2,832) demonstram forte impacto negativo no desempenho.

Ademais, é importante frisar que algumas das características dos estudantes que foram fornecidas estão estritamente ligadas com as "vizinhanças"de onde moram. Como, por exemplo, se o aluno tem pavimentação (7,240) e iluminação (4,140) nas proximidades de casa, ele tende a ter uma maior proficiência em matemática. Além disso, características como deslocamento menor que 30 minutos (4,714) e transporte por carro (2,226) estão positivamente relacionadas à proficiência, enquanto o transporte a pé (-3,755) ou de ônibus

(-0,035).

Tabela 11: Modelo de Regressão Linear com variaveis explicativas do 1º Nível - M1

Variáveis Explicativas	**Modelo do 1º Nível (M1)**		
	Estimativa	**Erro Padrão**	**p-valor**
Intercepto	233,299	4,531	<0,001
Sexo: masculino	8,301	0,535	<0,001
Cor/Raça: Branco ou Amarelo	4,920	0,583	<0,001
INSE do Aluno	4,281	0,358	<0,001
Idade do Aluno	-2,975	0,357	<0,001
Pais conversam sobre a escola	2,888	0,920	0,001
Deslocamento menor que 30 min.	4,714	0,646	<0,001
Já reprovou	-24,060	0,849	<0,001
Já abandovou	-2,832	1,187	0,017
Iluminação	4,140	1,176	<0,001
Pavimentação	7,240	0,762	<0,001
Transporte: Carro	2,226	0,768	<0,001
Transporte: Ônibus	-3,035	1,252	<0,001
Transporte: À pé	-3,755	0,688	<0,001

Fonte: Elaboração própria

4.3.2 2º Nível - Efeito escola

Compreendendo a possível dependência entre as informações dos alunos, em relação às escolas que estudam, é possível calcular o "Efeito Escola"a partir do Modelo Nulo – M0 (FERRÃO; FERNANDES, 2003). Como apresentado na subseção 2.2.3, o Modelo Nulo (M0) é o primeiro passo para a aplicação de um Modelo Multinível, pois, a partir de seus coeficientes, é possível avaliar a contribuição dos níveis na variável de resposta. Dessa forma, a Tabela 12 apresenta as estimativas para o Modelo.

Tabela 12: Modelo sem variáveis explicativas – Modelo Nulo (M0)

Variáveis Explicativas	**Modelo Nulo (M0)**		
Efeito Fixo	**Estimativa**	**Erro Padrão**	**p-valor**
Intercepto	238,53	0,709	<0,001
Efeito aleatório - nível 2			
Variância entre as escolas	138	11,7	<0,001
Efeito aleatório - nível 1			
Variância do resíduo	1.564	39,4	<0,001
Correlação intraclasse (ICC)	8,1%		
Deviance	216024		
Número de parâmetros	3		

O intercepto do Modelo Nulo (M0) é 238,53, valor que representa a média da proficiência em matemática dos estudantes. Além disso, a estimativa da variância entre as escolas, ou do intercepto, resultou em 138 e a variância entre os alunos em 1.564, valores os quais são estatisticamente significativos (p-valor $< 0,001$). A partir dessas estimativas, é possível calcular o Coeficiente de Correlação Intraclasse (ICC), comumente chamado de "Efeito Escola" neste contexto. Com base nos resultados apresentados, é possível inferir que 8,1% da variação do desempenho dos estudantes está relacionada à variabilidade entre essas escolas.

É relevante mencionar que este resultado foi aquém do esperado, dado os coeficientes apresentados em diferentes estudos em outras Unidades Federativas (UFs) do Brasil. Porém, tal achado pode ser entendido ao considerar que o estudo só abrange escolas públicas do DF, que, como visto na subseção 4.1.2, são consideravelmente homogêneas entre si. Ainda assim, o coeficiente justifica o emprego da abordagem multinível. Tal fato pode ser observado ao aplicar o teste de ANOVA entre o Modelo nulo (M0) e um modelo reduzido sem níveis.

O próximo passo (2.2.3) para aplicar as técnicas de modelagem multinível é analisar um modelo controlado pelas variáveis explicativas unicamente do nível 1 e, após isso, comparar com um modelo que conta com a inclusão de variáveis do nível 2.

Tabela 13: Modelo com variaveis explicativas do nível 1 - Modelo 2 (M2)

Variãveis Explicativas	**Modelo 2 (M2)**		
Efeito Fixo	**Estimativa**	**Erro Padrão**	**p-valor**
Intercepto	235,42	4,474	<0,001
Sexo: Masculino	8,444	0,522	<0,001
Cor/Raça: Branco ou Amarelo	3,914	0,574	<0,001
INSE do Aluno	3,149	0,350	<0,001
Idade do aluno	-2,568	0,350	<0,001
Pais conversam sobre a escola	2,239	0,899	0,012
Deslocamento menor que 30 min,	6,418	0,629	<0,001
Já reprovou	-23,730	0,833	<0,001
Já abandonou	-3,219	1,161	0,005
Iluminação	3,701	1,161	0,001
Pavimentação	7,080	0,776	<0,001
Transporte: Ônibus	-3,609	1,207	0,002
Transporte: À pé	-3,284	0,603	<0,001
Efeito Aleatório - nível 2			
Variância entre as escolas	95,7	9,78	<0,001
Efeito Aleatório - nível 1			
Variância do resíduo	1395,3	37,35	<0,001
Deviance	213154		
Número de parâmetros	15		
Variância do nível 1 explicada	10,79%		
Variância do nível 2 explicada	30,65%		

O Modelo 2 (M2), apresentado na Tabela 13, incorpora apenas variáveis explicativas relacionadas ao nível do aluno. Foram incluídas características previamente selecionadas com base nas análises anteriores. Do ponto de vista da variância, o modelo explica $10,79\%$ da variância no nível 1 (dentro das escolas) e $30,65\%$ no nível 2 (entre as escolas). A redução do *deviance* em relação ao modelo nulo foi substancial (deviance = 213.154), indicando que a inclusão dessas variáveis melhorou significativamente o ajuste do modelo.

Tabela 14: Modelo com variaveis explicativas do nível 1 e nível 2 - Modelo 3 (M3)

Variáveis Explicativas	**Modelo 3 (M3)**		
Efeito Fixo	**Estimativa**	**Erro Padrão**	**p-valor**
Intercepto	149,047	12,856	<0,001
Sexo: Masculino	8,428	0,522	<0,001
Cor/Raça: Branco ou Amarelo	3,777	0,574	<0,001
INSE do Aluno	2,784	0,353	<0,001
Idade do aluno	-2,498	0,350	<0,001
Pais conversam sobre a escola	2,303	0,899	0,010
Deslocamento menor que 30 min,	6,509	0,628	<0,001
Já reprovou	-23,632	0,833	<0,001
Já abandonou	-3,214	1,161	0,005
Iluminação	3,169	1,160	0,006
Pavimentação	6,501	0,775	<0,001
Transporte: Ônibus	-3,765	1,205	0,002
Transporte: À pé	-3,223	0,601	<0,001
INSE Médio da escola	17,179	2,290	<0,001
Porcentagem de defasagem idade-série	-0,312	13,422	0,02
Efeito Aleatório - nível 2			
Variância do intercepto	68	8,24	<0,001
Efeito Aleatório - nível 1			
Variância do resíduo	1.396	37,36	<0,001
Deviance	213074		
Número de parâmetros	17		
Variância do nível 1 explicada	10,74%		
Variância do nível 2 explicada	50,72%		

O Modelo 3, detalhado na Tabela 14, expande a análise ao adicionar variáveis do nível da escola, como o INSE médio da escola, a porcentagem de alunos com defasagem idade-série. Os resultados mostram que, além das variáveis do nível 1, para cada ponto adicional no INSE médio da escola, a proficiência esparada dos estudantes aumenta em 17,179 pontos ($\beta = 17,179, p < 0.001$). Em contraponto, a cada aumento de 1 ponto percentual na defasagem idade-série da escola, o desempenho dos alunos cai em 0.312 pontos ($\beta = -0,322, p = 0.02$).

Com a inclusão dessas variáveis, a variância entre as escolas (nível 2) foi ainda mais reduzida, explicando $50,72\%$, em comparação aos $30,65\%$ observados no M1, indicando como fatores socioeconômicos (INSE) e atraso escolar explicam boa parte nessa

variação. No entanto, a variância dentro das escolas (nível 1) permaneceu praticamente inalterada (10,74% no M2 versus 10,79% no M1), sugerindo que fatores individuais dos alunos ainda desempenham o papel central na explicação da variabilidade do desempenhos. A redução adicional no *deviance* (Deviance = 213.154 − 213.074) reforça a superioridade do M2 em relação ao M1.

4.3.3 Modelo Final

Inicialmente, buscou-se modelar a proficiência em matemática dos estudantes considerando um modelo hierárquico de três níveis (Aluno, Escola e Região Administrativa). No entanto, a tentativa de incluir o efeito da RA como um terceiro nível resultou na não convergência do modelo. Esse problema pode estar relacionado à complexidade do fenômeno estudado, ao grande número de parâmetros a serem estimados e à menor variabilidade explicada pelas Regiões Administrativas. Após o uso de diferentes técnicas de estimação, o Coeficiente de Correlação Intraclasse (ICC) de 3 níveis, que indicaria o efeito vizinhança, capturado pela variabilidade da proficiência entre as RAs, sempre foi inferior a 1%, sugerindo que a inclusão desse nível no modelo não traria ganhos significativos.

Apesar disso, a segregação socioespacial do Distrito Federal, fortemente associada à renda, demonstrou ser um fator influente nas proficiências dos estudantes. Como evidenciado ao longo deste trabalho, a renda impacta o desempenho acadêmico tanto individualmente (INSE do aluno), quanto coletivamente no ambiente escolar (INSE médio da escola). Além disso, a análise descritiva mostrou diferenças sistemáticas nas proficiências médias das Regiões Administrativas, organizadas por categoria de renda (Figura 14).

Diante desse contexto, optou-se por incorporar o efeito da vizinhança por meio da inclusão da variável P26 da PDAD (percentual de domicílios com renda superior a cinco salários mínimos). Com essa abordagem, o modelo final considera dois níveis hierárquicos (Aluno e Escola), incluindo variáveis individuais, ligadas a características individuais do aluno e ao lugar onde mora, além do contexto escolar e características socioeconômicas da vizinhança da escola.

Tabela 15: Modelo de dois níveis com variáveis de aluno, escola e vizinhança – Modelo 4 (M4).

Variáveis Explicativas	**Modelo 4 (M4)**		
Efeitos Fixos	**Estimativa**	**Erro Padrão**	**p-valor**
Intercepto	178,790	13,672	<0,001
Sexo: Masculino	8,432	0,521	<0,001
Cor/Raça: Branco ou Amarelo	3,756	0,573	<0,001
INSE do Aluno	2,785	0,353	<0,001
Idade do aluno	-2,489	0,349	<0,001
Pais conversam sobre a escola	2,243	0,898	0,012
Deslocamento menor que 30 min,	6,682	0,628	<0,001
Já reprovou	-23,644	0,832	<0,001
Já abandonou	-3,234	1,160	0,005
Iluminação	3,163	1,159	0,006
Pavimentação	6,354	0,773	<0,001
Transporte: Ônibus	-3,980	1,204	<0,001
Transporte: À pé	-3,130	0,600	<0,001
INSE Médio da escola	10,716	2,553	<0,001
Defasagem idade-série	-0,325	12,941	0,012
Famílias com renda maior que 5SM (P26)	0,149	0,029	<0,001
Efeito Aleatório - nível 2			
Variância do intercepto	60,1	7,75	<0,001
Efeito Aleatório - nível 1			
Variância do resíduo	1.390,1	37,35	<0,001
Deviance	213036		
Número de parâmetros	18		
Variância do nível 1 explicada	10,81%		
Variância do nível 2 explicada	56,44%		

O Modelo 4 (M4), apresentado na Tabela 15, incorpora variáveis do nível do aluno, da escola e do contexto socioeconômico da vizinhança. A inclusão dessas variáveis busca capturar não apenas características individuais e escolares, mas também efeitos da segregação socioeconômica do Distrito Federal.

No nível individual, as variáveis explicativas que tinham demonstrado impacto significativo nos modelos anteriores continuaram relevantes, com pequenas variações em seus coeficientes. Em relação ao Sexo, considerando a categoria de referência como feminino, os alunos do sexo masculino apresentaram, em média, um desempenho 8,43 pontos maior ($\beta = 8,432, p < 0,001$). Com relação à Raça/Cor, considerando a categoria de

referência como P.P.I., os alunos brancos ou amarelos obtiveram uma proficiência média de 3,76 pontos superior ($\beta = 3,756, p < 0,001$).

O nível socioeconômico individual (INSE do aluno) apresentou um coeficiente positivo de 2,79 pontos por unidade. Ou seja, para cada unidade de aumento no INSE do aluno, sua proficiência esperada aumenta, em média, em 2,79 pontos. Em relação a idade do estudante, alunos mais velhos tendem a ter um menor desempenho, que é um reflexo da defasagem idade-série e possíveis reprovações ou abandono escolar. Neste contexto, estudantes que já reprovaram apresentaram, em média, um desempenho de aproximadamente 23 pontos menor. Ademais, os alunos que abandonaram a escola tiveram uma proficiência 3,23 pontos menor. Por fim, alunos que conversam com os pais sobre a escola apresentaram um desempenho, em média, 2,24 pontos maior.

Em relação ao contexto escolar, os efeitos seguem praticamente os mesmos. Ao aumentar o nível socioeconômico médio dos estudantes da escola em uma unidade, a proficiência média dos alunos tende a crescer em 10,72 pontos. Além disso, a cada aumento de 10 pontos percentuais da defasagem idade-série, o desempenho cai em 3,25 pontos.

Finalmente, no que se refere ao efeito da vizinhança, para cada aumento de 10 pontos percentuais na proporção de famílias com maior renda na Região Administrativa, a proficiência média dos estudantes tende a aumentar em 1,49 pontos. Esse resultado reforça a influência da condição econômica da vizinhança sobre o desempenho escolar, além dos efeitos já capturados pelo INSE do aluno e da escola.

Além disso, duas variáveis de 1º nível que também reforçam o efeito vizinhança são as condições de infraestrutura nas proximidades das residências dos alunos. A presença de iluminação na rua da casa do estudante é associada a um aumento médio de 3,16 pontos na proficiência, enquanto a pavimentação resulta em um acréscimo de 6,35 pontos.

Por fim, a inclusão da variável P26 resultou em uma melhora na explicação da variabilidade entre as escolas. No total, as variáveis do modelo explicam 56,44% da diferença entre as escolas. Enquanto isso, no nível 1, a variância explicada permaneceu praticamente a mesma (10,81%), o que sugere que, embora fatores externos tenham impacto relevante, a maior parte da variação do desempenho dos alunos ainda está ligada a fatores individuais.

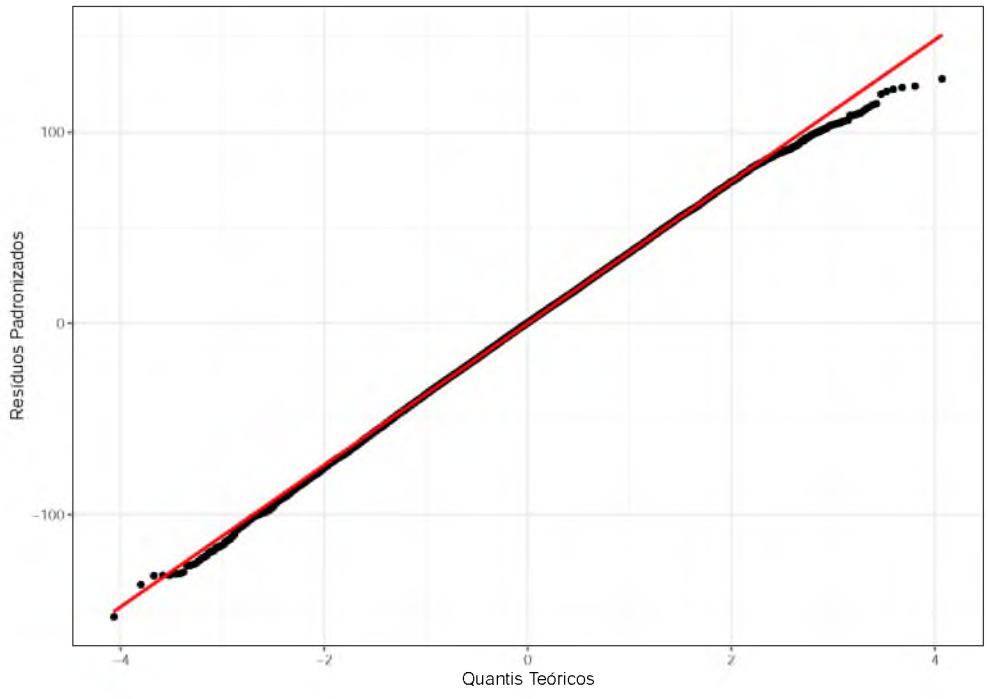
Tabela 16: Tabela com Deviance e AIC dos modelos apresentados.

Modelo	Deviance	AIC	Número de parâmetros
M0	216024	216030	3
M1	31276.300	213949	14
M2	213171	213154	15
M3	213074	213084	17
M4	213031	213071	18

A comparação com os modelos anteriores mostra que a adição de informações da escola e vizinhança melhoram o ajuste do modelo. A redução do *deviance* para 213.031 reforça a superioridade do M4 em relação aos modelos anteriores.

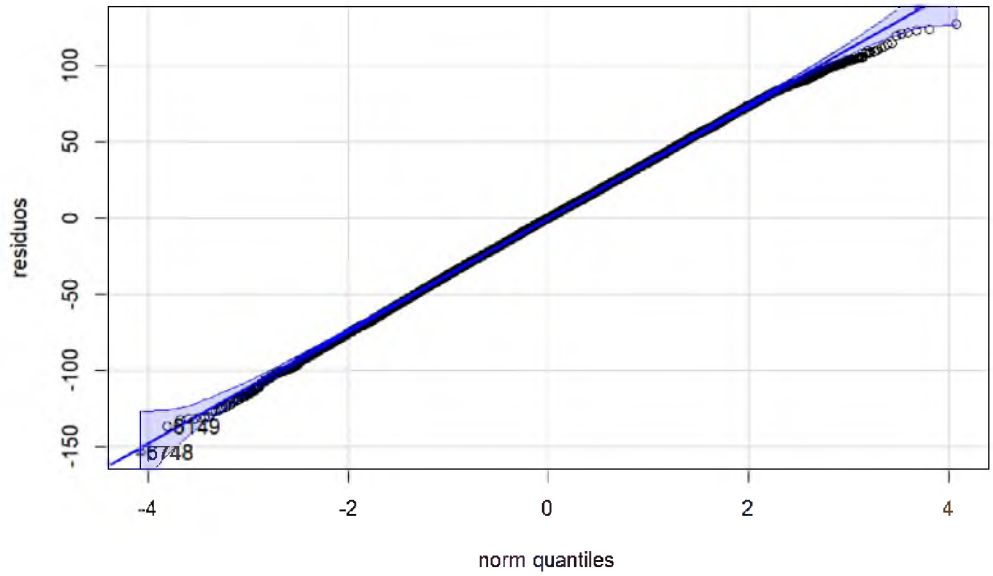
Por último, os pressupostos de independência, normalidade e homoscedasticidade dos resíduos também podem ser validados a partir dos gráficos a seguir.

Figura 16: Q-Q Plot dos Resíduos.



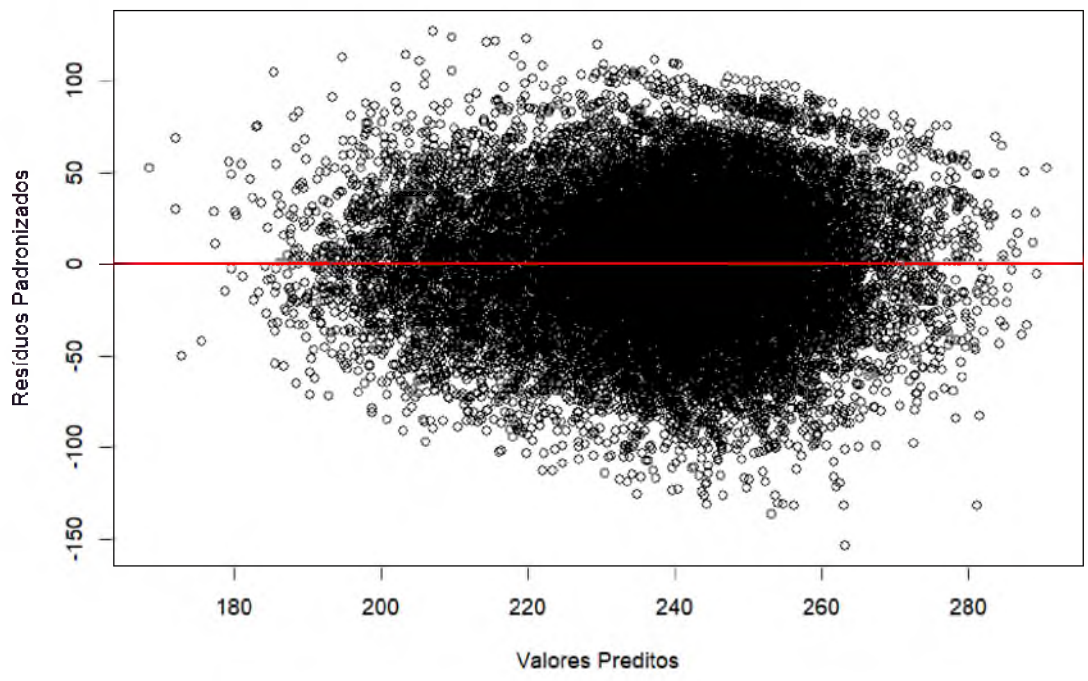
Fonte: Elaboração própria.

Figura 17: Q-Q Plot dos Resíduos, com envelopes.



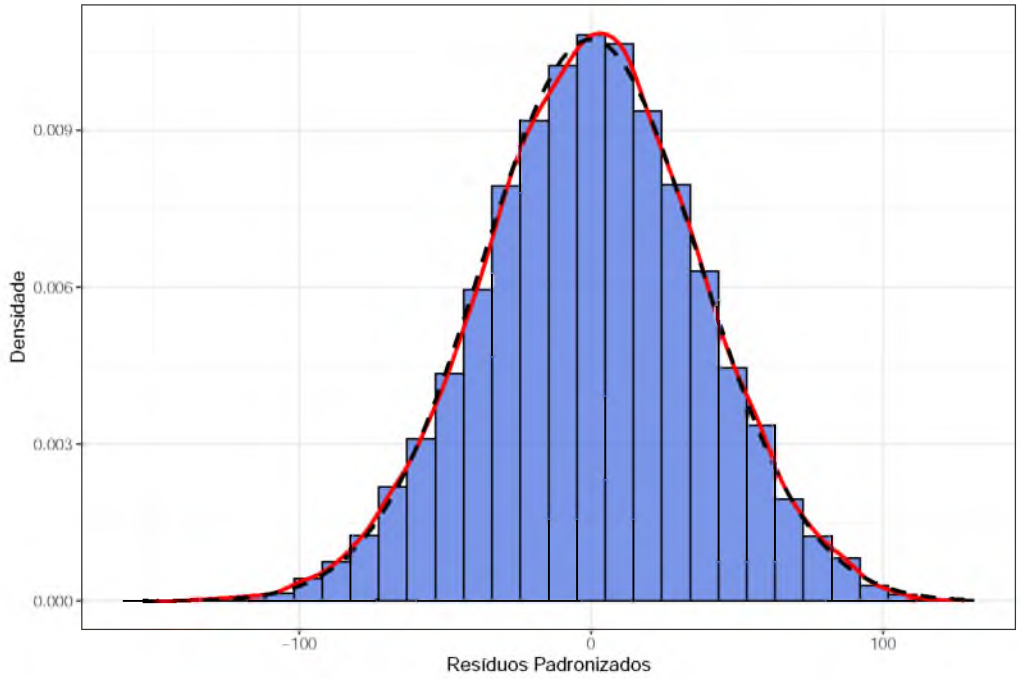
Fonte: Elaboração própria.

Figura 18: Gráfico de dispersão dos Resíduos Padronizados e Valores Preditos.



Fonte: Elaboração própria.

Figura 19: Histograma dos resíduos.



Fonte: Elaboração própria.

A variável P10 (proporção de indivíduos com ensino superior completo) apresenta uma forte correlação com P26 ($r = 0,95$), pois ambas refletem aspectos socioeconômicos relacionados à desigualdade econômica. Dessa forma, um modelo que inclui P10 como variável no 3º nível também apresentou um bom ajuste, embora inferior ao modelo M4, conforme indicado pelos valores de deviance e pelo Teste de Razão de Verossimilhanças. Da mesma maneira, um modelo alternativo que utilizava a categorização de renda das RAs (Figura 10) como variável explicativa do 3º nível também foi testado, mas descartado pelos mesmos motivos.

5 Conclusão

O trabalho teve como objetivo avaliar os fatores, principalmente ligados à escola e vizinhança, que influenciam a proficiência em matemática dos estudantes do 5º ano do Ensino Fundamental das escolas públicas do Distrito Federal (DF), a partir de modelos de regressão multinível. Os resultados apresentados revelam que as características individuais são as que desempenham o papel mais significativo no aprendizado em matemática. O sexo masculino, cor branca ou amarela e maior nível socioeconômico (INSE) estão associados a maiores medidas de desempenho. Por outro lado, fatores como reprovação, abandono escolar e defasagem idade-série tiveram impactos negativos consideráveis na proficiência. Além disso, a presença de pais que conversam frequentemente sobre a escola também se mostrou um fator positivo para o aprendizado.

Em relação às escolas, suas infraestruturas não se mostraram significativas para o aprendizado em matemática. Isso pode ser explicado pelo fato de que as escolas públicas se demonstraram muito homogêneas entre si, em questão de gestão e estrutura física. Dessa forma, somente variáveis vindas de "agregações"de características dos alunos foram significativas. A partir delas, foi identificado que o nível socioeconômico médio dos alunos tem uma influência positiva no desempenho. Enquanto isso, aquelas com maior taxa de defasagem idade-série enfrentam desafios na aprendizagem.

Ao analisar a vizinhança, não foi possível aplicar um modelo hierárquico de três níveis, por problemas de convergência. Porém, de qualquer forma, as variáveis ligadas à vizinhança tiveram destaque no trabalho, quando ligadas com as proficiências dos estudantes. Principalmente fatores ligados a áreas economicamente mais favorecidas, como a proporção de famílias com renda superior a cinco salários mínimos, porcentagem de pessoas com ensino superior completo, além de iluminação e pavimentação das ruas.

Além disso, observou-se que o tempo de deslocamento até a escola e o meio de transporte utilizado pelos estudantes também exercem influência no desempenho. Alunos que gastam menos tempo no trajeto e que utilizam meios de transporte mais estáveis, como transporte particular ou escolar, tendem a apresentar melhores resultados. Por outro lado, aqueles que se deslocam a pé ou enfrentam longos períodos no trajeto demonstram proficiências mais baixas.

Apesar das contribuições deste estudo, algumas limitações devem ser consideradas. A base de dados utilizada apresenta uma quantidade considerável de dados faltantes, especialmente nas variáveis socioeconômicas dos estudantes, o que pode impactar

a robustez das análises. Além disso, algumas características do ambiente escolar, como metodologias de ensino e engajamento dos professores, não foram consideradas devido à indisponibilidade de dados.

Referências

ALBERNAZ, Â.; FERREIRA, F. H.; FRANCO, C. *Qualidade e eqüidade na educação fundamental brasileira*. [S.l.], 2002.

ANDRADE, L. T.; SILVEIRA, L. S. Efeito-território: explorações em torno de um conceito sociológico. *Civitas-Revista de Ciências Sociais*, SciELO Brasil, v. 13, p. 381–402, 2020.

ARAÚJO CARLOS HENRIQUE, L. *Avaliação da educação básica em busca da qualidade e eqüidade no Brasil*. [S.l.]: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2005.

FERRÃO, M. E.; FERNANDES, C. O efeito-escola e a mudança-dá para mudar?: Evidências da investigação brasileira. *REICE: Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación*, Red Iberoamericana de Investigación Sobre Cambio y Eficacia Escolar, v. 1, n. 1, p. 5, 2003.

FERRãO, M. E. Introdução aos modelos de regressão multinível em educação. *Cadernos de Pesquisa*, v. 34, n. 121, p. 248, abr. 2004. Disponível em: ⟨https://publicacoes.fcc.org.br/cp/article/view/499⟩.

GAVIN, M. Hierarchical linear models: Applications and data analysis methods. *Organizational Research Methods*, SAGE PUBLICATIONS, INC., v. 7, n. 2, p. 228, 2004.

GOLDSTEIN, H. *Multilevel Statistical Models*. E. Arnold, 2003. (Kendall's advanced theory of statistics and kendall's library of statistics). ISBN 9780340595299. Disponível em: ⟨https://books.google.com.br/books?id=7yGFLgAACAAJ⟩.

HOX, J.; MOERBEEK, M.; SCHOOT, R. Van de. *Multilevel analysis: Techniques and applications*. [S.l.]: Routledge, 2017.

INEP. *BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Escalas de proficiência do SAEB. Brasília, DF*. 2020. Disponível em: ⟨https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/escalas_de_proficiencia_do_saeb.pdf⟩.

KUTNER, M. H. et al. *Applied linear statistical models*. [S.l.]: McGraw-hill, 2005.

LIMA, A. A. A. A. de; SOUSA, F. P. de. Censo escolar da educação básica: Uma referência para elaboração de políticas públicas e transferência de recursos para educação pública. *Revista Com Censo: Estudos Educacionais do Distrito Federal*, v. 1, n. 1, p. 99, 2014.

MACHADO, D. C. d. O. et al. Análise de fatores associados ao desempenho escolar de alunos do quinto ano do ensino fundamental com base na construção de indicadores. 2014.

MALOUTAS, T. et al. Residential and school segregation as parameters of educational performance in athens. *Cybergeo: European Journal of Geography*, CNRS-UMR Géographie-cités 8504, 2019.

OLIVEIRA, C. S. *Efeito vizinhança sobre a escolha do indivíduo no mercado de trabalho em Fortaleza*. 2012. Disponível em: ⟨http://repositorio.ufc.br/handle/riufc/5409⟩.