Quantidade de micro-ondas, estado do Brasil e a nota do Enem

Esdras Cavalcanti
Ciência de Dados e Inteligência Artificial
Fundação Getulio Vargas
Rio de Janeiro, RJ
esdras2821@gmail.com

Abstract

Trabalho da disciplina de Modelagem Estatística. Através da análise dos microdados do Enem 2023, busca reforçar o problema brasileiro de desigualdade social, entre indivíduos e entre estados. Para isso, são utilizados modelos lineares junto com modelos multinível.

1 Introdução

A desigualdade social no Brasil é um fato corriqueiramente cobrado na redação do Enem. Costuma ser intuitivo que uma família pobre não terá o melhor acesso possível à educação, consequentemente os filhos permanecerão à margem da sociedade. Será que essa intuição está correta? Até que ponto critérios socioeconômicos estão atrelados ao desempenho acadêmico e, por consequência, profissional de um indivíduo? Será essa uma espécie de "ciclo da pobreza"[1] na educação?

As respostas para essas perguntas podem por um lado mostrar que está tudo correndo bem na sociedade brasileira, por outro podem revelar um grande abismo social de difícil transposição.

Além de uma desigualdade direta entre os indivíduos, costuma-se falar de uma desigualdade entre estados do país. Será que a influência do estado em que se vive e estuda é relevante?

Para obter tais respostas é necessário dados que relacionem resultado acadêmico com fatores socioeconômicos e geográficos. Além de rotineiramente abordar o tema da desigualdade, o Inep também utiliza a inscrição do Enem como forma de coletar dados socioeconômicos dos participantes, possuindo intrinsecamente dados geográficos de aplicaçãod da prova. Posteriormente, esses dados são disponibilizados[1] de forma anonimizada, para garantir a segurança e o sigilo.

Além disso, a nota do Enem é um bom indicativo da qualidade do ensino, fundamental e médio, que um estudante obteve. Apesar de não representar com perfeição o futuro profissional do indivíduo, é um bom parâmetro para medir o nível do aluno ao ingressar na universidade.

Com base nos microdados do Enem, o presente trabalho se propõe a buscar **relações entre os dados socioeconômicos e geográficos dos participantes e sua nota no exame** como forma de verificar a veracidade do padrão de ciclo da pobreza educacional muitas vezes tido como senso comum e a diferença desse padrão por estado brasileiro.

As implementações feitas em R e Python podem ser econtradas no repositório do GitHub[2].

[1] Microdados do Enem de 1998 a 2023 https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

[2] https://github.com/wobetec/Desigualdade_Social_e_Microdados_Enem

1.1 Microdados

Anualmente, alguns de pessoas se inscrevem para realizar o Enem por todos os estados do país. Em 2023, foram aproximadamente 4 milhões de inscritos e por volta de 2.7 milhões realizaram a prova. Como base para a análise desse trabalho serão utilizados os dados referente a esse ano de 2023. A utilização de múltiplos anos de forma agregada traria viés ao resultado, uma vez que o mesmo indivíduo pode realizar o exame em anos consecutivos.

Dentro dos microdados encontram-se alguns grupos de dados:

- Participante
- Escola
- Local de aplicação da prova
- Prova objetiva
- Redação
- Questionário socioeconômico

Como o exame possui 5 notas distintas (Uma para cada grupo do conhecimento mais a redação), é preciso agregar essas notas. Para não escolher apenas uma das notas, ou trabalhar cada uma de forma individual, vamos agregá-las com a média.

Entre as demais informações, atributos como sexo, cor, língua da prova, estado civil, idade, etc não contribuem para a pergunta as serem respondidas. Por isso, serão considerados apenas os dados referentes Questionário socioeconômico e informação geográfica.

Quando ao dado geográfico, há a opção de utilizar o estado do local de prova ou o estado da escola do indivíduo. Ao observar os dados, pode-se notar que há muitos dados faltantes de escola, fazendo com que desconsiderar as linhas com dados faltantes seja inviável para o problema. Apesar de não ser regra que o local de realização da prova e a escola onde o indivíduo cursou não sejam no mesmo estado, essa é uma suposição plausível e que retira pouco valor da análise.

Por fim, dentre os dados existem indivíduos que estão realizando o exame, mas ainda não terem terminado o ensino médio. Outros faltaram pelo menos um dia de prova. Para garantir que estamos analisando indivíduos que concluíram seus estudos, bem como indivíduos que realizaram toda a prova, aplicaremos esse filtro sobre os dados.

1.2 Questionário Socioeconômico

O Inep realiza 25 perguntas no questionário:

- **Q001**: Até que série seu pai, ou o homem responsável por você, estudou?
- **Q002**: Até que série sua mãe, ou a mulher responsável por você, estudou?
- **Q003**: A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).
- **Q004**: A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).
- **Q005**: Incluindo você, quantas pessoas moram atualmente em sua residência?
- **Q006**: Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
- **Q007**: Em sua residência trabalha empregado(a) doméstico(a)?
- **Q008**: Na sua residência tem banheiro?
- **Q009**: Na sua residência tem quartos para dormir?
- **Q010**: Na sua residência tem carro?

- **Q011**: Na sua residência tem motocicleta?
- **Q012**: Na sua residência tem geladeira?
- **Q013**: Na sua residência tem freezer (independente ou segunda porta da geladeira)?
- **Q014**: Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)
- **Q015**: Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?
- **Q016**: Na sua residência tem forno micro-ondas?
- **Q017**: Na sua residência tem máquina de lavar louça?
- **Q018**: Na sua residência tem aspirador de pó?
- **Q019**: Na sua residência tem televisão em cores?
- **Q020**: Na sua residência tem aparelho de DVD?
- **Q021**: Na sua residência tem TV por assinatura?
- **Q022**: Na sua residência tem telefone celular?
- **Q023**: Na sua residência tem telefone fixo?
- **Q024**: Na sua residência tem computador?
- **Q025**: Na sua residência tem acesso à Internet?

As possíveis respostas são ser apresentadas de forma crescente, exemplo **Q016** a resposta 0 é nenhum, 1 é um micro-ondas, 2 são dois micro-ondas, 3 são 3 e 4 são 4 ou mais. Dessa forma, podem ser tomadas direto como tendo uma linha base.

É possível ver que a maioria delas é correlacionada com a renda familiar, por exemplo **Q010** é respondida como a quantidade de carros, que esta diretamente relacionada com o rendimento da família. Essa correlação pode vir a atrapalhar a indentificação das melhores variáveis explicativas.

Além disso, as 4 primeiras perguntas podem ser respondias com *Não sei*, resposta plausível caso o indivíduo não tenha pai ou mãe. Contudo, para simplificar a análise vamos desconsiderar esses casos, para garantir que temos apenas respostas completas.

1.3 Análise exploratória

Uma vez aplicando os filtros propostos e selecionando como variáveis explicativas o conjunto das 25 perguntas socioeconômicas mais a UF de realização da prova o resultado é uma base de dados com 797,357 linhas e 27 colunas. O filtro reduziu a base de 3,933,955 para apenas 20.27% das linhas, devido ao volume de dados, essa redução não irá prejudicar o objetivo da análise.

Ao observar a distribuição das notas(Figura 1) e como esperado pela teoria dos grandes números, devido ao volume de dados encontramos uma distribuição com formato de sino. Contudo, o objetivo do trabalho é entender a relação das notas com as variáveis socioeconômicas e geográficas.

Observando a distribuição dos dados com relação às variáveis geográficas (Figura 2) é possível ter uma noção do impacto que diferentes estados representam no valor final da nota de um indivíduo.

Além disso, olhando para os dados socioeconômicos, mais especificamente o número de micro-ondas da residência(Figura 3), encontra-se um padrão que se repete em dosa as variáveis. É possível observar essa tendência em todas as 25 questões.

Intuitivamente, ao observar os os gráficos já se começa a criar uma ideia do resultado que esperamos. Contudo, há algumas métricas que precisão ser avaliadas, considerando todo o conjunto de dados para entender se essa intuição está correta.

2 Métodos

Com o objetivo de verificar a influência das variáveis explicativas sobre a nota do Enem, optou-se por utilizar um modelo frequentista de regressão linear.

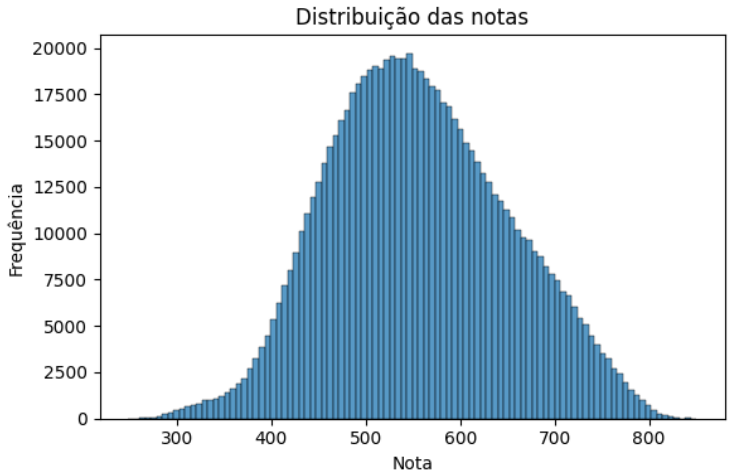

Figure 1: Distribuição das notas do Enem

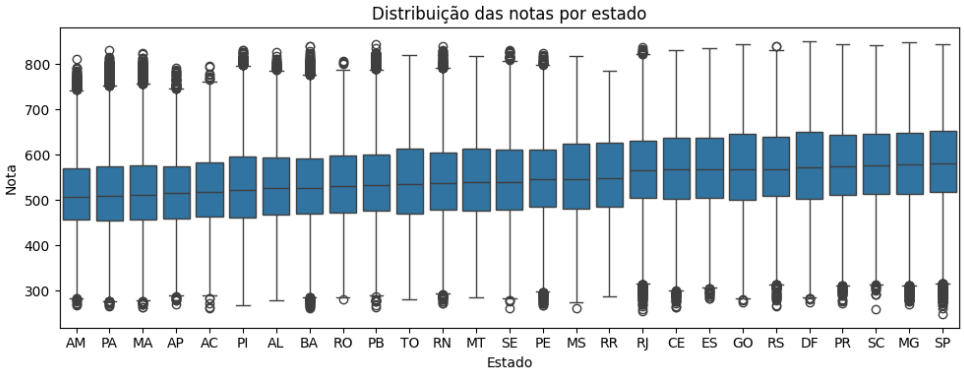

Figure 2: Distribuição das notas do Enem por estado

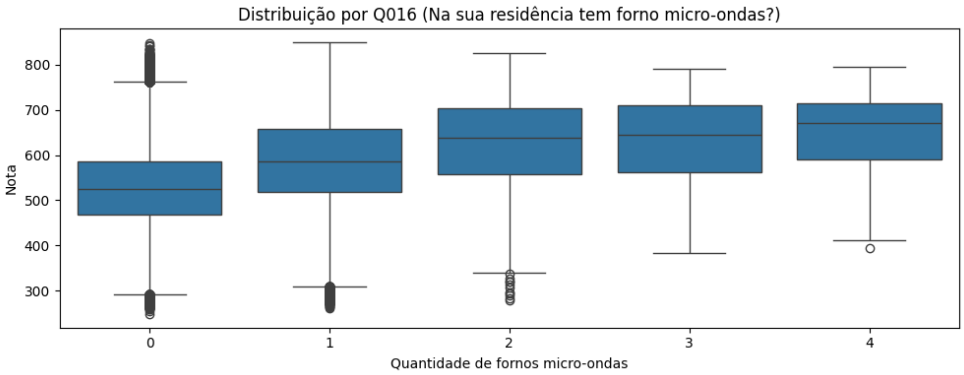

Figure 3: Distribuição das notas do Enem pela quantidade de micro-ondas na residência. *O valor 4 representa 4 micro-ondas ou mais.*

Esse tipo de análise permite ver a influência individual de cada covariável, além de ter custo relativamente baixo considerando o volume de dados utilizado. A análise será dividida em duas etapas explicadas a seguir.

2.1 Escolhendo variáveis explicativas

Temos um alto número de covariáveis, são 25 perguntas do questionário socioeconômico mais o estado de realização da prova. Para tentar simplificar o modelo optou-se por filtrar essas variáveis utilizando o Akaike information criterion(AIC) dado por:

$$AIC = 2k - log(\hat{L}) \tag{1}$$

Com k sendo número de parâmetros e $\hat{L}$ o valor máximo da verossimilhança. Essa métrica é uma ferramenta útil por permitir comparar de forma mais justa modelos com diferente quantidade de covariáveis. Busca-se então o modelo com menor AIC, ou seja um modelo mais simples e ao mesmo tempo bom.

Para fazer essa seleção, é selecionado incrementalmente as variáveis de **Q001** a **Q025**, sendo adicionada ao pool de variáveis caso reduza o AIC, caso contrário é descartada. Esse processo, assim como a análise exploratória dos dados foi realizado em python. Com auxílio da biblioteca *statsmodels* verifica-se o AIC para cada modelo e selciona as variáveis.

Um segundo filtro para as covariáveis é o p valor. Como almeja-se apenas as que apresentam bom poder explicativo. Por isso, considera-se para o modelo final apenas as que não possuem o zero no seu intervalo de confiança.

2.2 Análise multinível

Uma vez tendo realizado a etapa anterior a nível de Brasil e obtido um modelo de regressão final, parte-se para analisar de forma multinível nos estados. Utilizando como base o capítulo 11 do Gelman[2], aplica-se um modelo multinível com intercepto variável sobre os estados.

A consideração de intercepto variável significa uma suposição de que dentro dos diferentes estados a distribuição das notas é semelhante, variando apenas sua localização. Variação essa que é dada pelo efeito (intercepto) aleatório dos estados.

A análise multinível é feita em R, com a biblioteca *lme4* e a função *lmer* para ajustar um modelo de efeito misto.

Por fim, com AIC, R^2 e MSE compara-se os dois modelos finais, verificando o impacto da suposição multinível nos estados.

3 Resultados

3.1 Seleção das variáveis explicativas

A seleção com AIC mostra que todas as variáveis devem ser consideradas. Com isso, ao ajustar o modelo a nível de Brasil duas variáveis são tidas como não explicativas (p valor muito baixo).

Table 1: Variáveis com p valor elevado no modelo nacional

Variável	Estimate	Std. Error	t value	Pr($>|t|$)
Q009	-0.03809	0.15093	-0.252	0.801
Q023	0.51764	0.31836	1.626	0.104

Essas variáveis são respectivamente "Na sua residência tem quartos para dormir?" e "Na sua residência tem telefone fixo?". Ao considerarmos amostras menores o mesmo ocorre. Vale ressaltar que eliminando essas variáveis há uma diminuição do AIC, tal variação não foi capturada pela nossa abordagem inicial devido à forma sequencial com que foi testada e não de forma gulosa testando todas as possíveis combinações.

Ao eliminar essas variáveis e ajustar novamente o modelo são obtidas as seguintes métricas:

O alta valor de AIC, bem como o baixo valor de R^2 ocorrem devido ao volume de dados e a quantidade de ruído presente. Além disso, está sendo considerado um subconjunto de covariáveis que não representa todas as possibilidades, nem as mais influentes para a prova do Enem.

Table 2: Métricas do modelo nacional

Métrica	Valor
AIC	9,193,925
R^2	0,345
MSE	5958.386

Contudo, essas métricas são boas métricas para compararmos modelos em cima do mesmo conjunto de dados.

3.2 Análise multinível nos estados

Ao considerar o modelo multinível de intercepto variável obtêm-se as seguintes métricas:

Table 3: Métricas do modelo multinível nos estados

Métrica	Valor
AIC	9,180,407
R^2_m	0.343
R^2_c	0.357
MSE	5856.549

A métrica R^2_m é o R^2 marginal, ele representa a métrica calculada considerando apenas os efeitos fixos, isto é, sem considerar o intercepto variável dos estados. Já o R^2_c é o condicional e leva em conta o efeito aleatório.

A diminuição do R^2_m comparado com o R^2 do modelo nacional representa que os estados não possuem exatamente a mesma distribuição de notas dado o questionário. Já o aumento do R^2_c, também em comparação com o R^2 nacional mostrar que assumir a influência dos estados aumentou a qualidade do modelo. Consequentemente levando à conclusão da existência dessa influência.

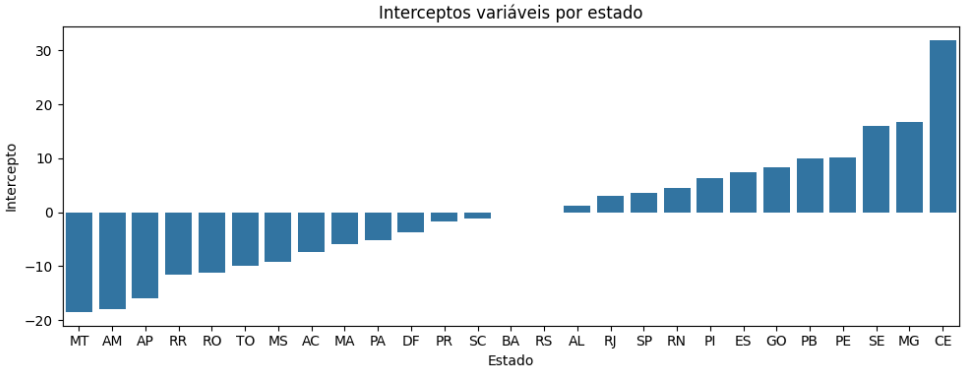

Figure 4: Interceptos variáveis por estado

Esses interceptos dizem que ao pegar dado todas as outra variáveis iguais, o estado que o indivíduo está aumenta ou diminui esse valor na nota.

Ao comparar as Figuras 2 e 4 é possível verificar diversas inversões na ordem dos estados. Ao mesmo tempo que grandes grupos continuam a aparecer mais ou menos na mesma região.

3.3 Comparação de modelos

O objetivo do trabalho é avaliar a influência das variáveis sobre a nota. Consequentemente, como foram considerados dois modelos distintos é importante que os coeficiêntes tenham a mesma direção.

Table 4: Comparação dos coeficientes para os dois modelos

Variável	Multinível	Nacional
(Intercept)	470.21	473.67
Q001	4.24	4.02
Q002	6.44	6.20
Q003	5.52	5.52
Q004	3.78	3.87
Q005	-5.05	-5.51
Q006	6.04	5.94
Q007	-5.51	-5.62
Q008	6.02	6.84
Q010	2.41	2.67
Q011	-3.11	-2.80
Q012	-8.20	-8.62
Q013	11.96	11.79
Q014	-2.85	-4.87
Q015	-5.70	-7.53
Q016	3.60	4.87
Q017	-5.56	-5.97
Q018	5.33	4.92
Q019	-2.27	-2.28
Q020	-3.45	-2.67
Q021	-9.31	-9.89
Q022	3.03	3.56
Q024	16.12	16.66
Q025	10.12	12.94

Como pode ser visto na Tabela 4, ambos os modelos consideram o mesmo sentido para as covariáveis analisadas.

Ao considerar o tipo de dado utilizado, partiu-se do pressuposto que a magnitude dos coeficientes não possuem tanto valor, explicativo para a maioria das variáveis. Como são variáveis categóricas que de forma direta apresentam uma relação de ordem, os saltos entre uma classe e outra não necessariamente possuem distância constante. Isto é, em **Q001** (Nível de ensino do pai) não se pode dizer que o salto ter quarta série completa e ter fundamental 2 completo é equivalente ao salto de ter médio completo e graduação completa.

Por outro lado, variáveis como **Q025** (Ter o ou não internet em casa) podem ser analisadas observando a magnitude do coeficiente. Uma vez que de forma direta, ter ou não ter acrescenta um certo número e pontos na nota do indivíduo.

Tendo isso em mente, a presente análise foca mais na direção da influência na maioria dos casos, em alguns poucos faz uso da magnitude.

4 Discussão

Primeiramente, as variáveis diretamente ligadas a renda(**Q003** e **Q004**) e educação(**Q001** e **Q002**) estão com coeficiente positivo. Tendo em vista que um maior valor dessas variáveis implica respectivamente em maior renda do do(a) pai/mãe e maior grau de ensino do(a) pai/mãe, observa-se que sim, pais que tiveram melhor educação, consequentemente melhor trabalho, influenciam diretamente nos resultados acadêmicos/profissionais dos filhos.

Alguns valores demonstram ainda mais a questão da desigualdade. As perguntas **Q024** e **Q025** tratam sobre ter ou não computador e ter ou não internet em casa. Em um mundo tão conectado, pessoas de classes não tão altas já possuem ambos. Contudo, muitas famílias à margem não tem esse acesso, fato que diretamente influencia nas chances de sucesso na carreira profissional.

Por fim, entre as diversas variáveis, **Q016** (Quantidade de fornos micro-ondas) aparece de forma positiva, mostrando que quanto mais micro-ondas, mais chances de ser aprovado. Por outro lado, a quantidade de geladeiras (**Q012**) faz o contrário.

Em segundo lugar, ao observar os dados dos estados, apesar das diferenças entre as Figuras 2 e 4, focando em estados específicos como MA e AC, ambos estão na ponta direita dos dois gráficos. Em pesquisa de 2022 do Instituto de Pesquisa Econômica Aplicada (IPEA)[3], MA, AC figuram o top 2 de pobreza extrema no país. Contudo, não é regra, já que CE e PE também figuram o top 10, apesar de estarem na extrema direita da Figura 4.

5 Trabalhos Futuros

5.1 Intercepto variável

O presente trabalho focou em analisar apenas modelos multinível com intercepto variável, o que pode não representar a realidade dos dados, apesar de ser a suposição inicial. Dessa forma, seria cabível explorar tanto modelo com coeficientes variáveis quanto modelos mistos como apresentados por Gelma[2]

5.2 Agregação das notas

Uma linha de estudo interessante seria verificar a influência separada por grupo de conhecimento. A nota utilizada no trabalho é uma agregação das 5 áreas, mas seria possível fazer essa análise para cada um dos grupos: Matemática, Linguagens, Ciências da Natureza, Ciências Humanas e Redação.

5.3 Estados e municípios

Estados são formas de agregação muito extensas. A base de dados utilizada permite desagregar esses dados até ao nível de municípios. Uma abordagem que estava fora do escopo do presente trabalho seria comparar grandes centros com regiões menos favorecidas.

5.4 Outros anos

O Inep disponibiliza microdados desde 1998. Uma análise que tanto traria mais valor para esse trabalho, quanto permitira ver a evolução do país ao longo do tempos consistiria de fazer análise de forma individual de cada ano. Posteriormente comparar esses dados, possivelmente reforçando o padrão encontrado nos dados, ao mesmo tempo que seria possível ver se há melhora ao longo do tempo.

5.5 "Não sei"

Por fim, uma limitação fundamental que foi introduzida nos modelos é filtrar os dados para remover os dados que o indivíduo não possui pai ou mãe. Para as conclusões desejadas o impacto dessa decisão não é tão alto. Entretanto, caso em classes menos favorecidas haja maior incidência de filhos de mãe solteira por exemplo, os modelos não estariam considerando uma parcela importante dos dados. Portanto, uma possível sequência do trabalho seria construir modelos de forma a lidar com tais tipos de respostas.

6 Conclusão

Em suma, conseguiu-se responder as duas perguntas iniciais. Apresentando modelos que traduzem os padrões encontrados nos microdados do Enem.

Primeiramente, os critérios socioeconômicos tem influência direta no resultado acadêmico, apesar de como discutido, as métricas mostram que não são as únicas variáveis que influenciam. Foi possível observar a reação direta que a renda e escolaridade dos pais está diretamente ligada ao resultado acadêmico, mostrando sim a existência de um ciclo da pobreza.

[3] `http://www.ipeadata.gov.br`

Em segundo lugar, mostrou-se uma real influência do estado de realização da prova na nota final. Mostrando uma desigualdade social não apenas entre pessoas, mas entre estados do país.

Tais conclusões revelam, como dito na introdução, um grande abismo social de difícil transposição que o Brasil sempre enfrentou e está longe de ser resolvido.

Agradecimentos

Gostaríamos de agradecer ao professor Luiz Max pelo direcionamento quanto aos passos do trabalho. Sobretudo a sugestão para aplicar uma análise multinível, enriquecendo os resultados e possibilitando um entendimento ainda mais profundo da dinâmica socio-geográfica do Brasil.

Referências

[1] Costas Azariadis and John Stachurski. “Poverty traps”. In: *Handbook of economic growth* 1 (2005), pp. 295–384.

[2] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.