UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
DEPARTAMENTO DE ECONOMIA E RELAÇÕES INTERNACIONAIS

JÚLIA RHEINHEIMER WALTER

UMA NOVA ESTIMATIVA DE ANOS MÉDIOS DE ESCOLARIDADE PARA O BRASIL, 1925 - 2015

Porto Alegre
2021

JÚLIA RHEINHEIMER WALTER

UMA NOVA ESTIMATIVA DE ANOS MÉDIOS DE ESCOLARIDADE PARA O BRASIL, 1925 - 2015

Trabalho de conclusão submetido ao Curso de Graduação em Ciências Econômicas da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título Bacharel em Economia.

Orientador: Prof. Dr. Sérgio Marley Modesto Monteiro.

Porto Alegre

2021

CIP - Catalogação na Publicação

Rheinheimer Walter, Júlia
Uma nova estimativa de anos médios de escolaridade para o Brasil, 1925 - 2015 / Júlia Rheinheimer Walter. -- 2021.
65 f.
Orientador: Sérgio Marley Modesto Monteiro.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Curso de Ciências Econômicas, Porto Alegre, BR-RS, 2021.

1. Capital humano. 2. Análise da educação. 3. Metodologia para coleta, estimativa e organização de dados macroeconômicos. 4. Acesso a dados. I. Marley Modesto Monteiro, Sérgio, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

JÚLIA RHEINHEIMER WALTER

UMA NOVA ESTIMATIVA DE ANOS MÉDIOS DE ESCOLARIDADE PARA O BRASIL, 1925 - 2015

Trabalho de conclusão submetido ao Curso de Graduação em Ciências Econômicas da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para obtenção do título Bacharel em Economia.

Aprovada em: Porto Alegre, _____ de _____de 2021.

BANCA EXAMINADORA:

Prof. Dr. Sérgio Marley Modesto Monteiro – Orientador
UFRGS

Prof. Dr. Leonardo Xavier da Silva
UFRGS

Prof. Dr. Thomas Hyeono Kang
ESPM

AGRADECIMENTO

Primeiramente, agradeço a Deus pela conquista. Sem Cristo não estaria aqui.

Agradeço a minha família que me deu todo o suporte, pai, mãe e mano, obrigada por todos os ensinamentos. Ao meu noivo, Carlos, agradeço por todo o seu incentivo e exemplo. Aos meus queridos parentes e amigos, obrigada.

Agradeço ao Samuel Pessôa, pelo suporte financeiro e pela idealização do trabalho, e ao professor Thomas H. Kang, que trabalhou lado a lado na concretização da pesquisa, obrigada por todo o apoio e paciência. Sem eles, não seria possível concretizar o trabalho. Agradeço ao professor Sérgio Monteiro e Leonardo Xavier por todo o suporte na minha trajetória acadêmica, tenho muita admiração por vocês. Obrigada a todos os professores que me auxiliaram no meu desenvolvimento, professores do colégio Dom Bosco no qual passei 14 anos e professores da Odontologia e Economia da UFRGS.

Deixo aqui também o meu pedido de desculpas, o peso de um curso não concluído. Eu pretendo retornar esses investimentos de alguma maneira.

Obrigada a todos que de alguma forma me auxiliaram para que eu chegasse até aqui.

RESUMO

Este trabalho visa estimar os anos médios de escolaridade da população de 15 a 64 anos de 1925 até 2015 através da utilização de série histórica de matrícula, microdados do censo, PNAD Contínua e dados populacionais do IBGE. O intuito secundário do trabalho é expandir a estimativa por gênero, por cor/raça (1925 a 2015) e por estado (1950 a 2015). Ademais, propõe-se a realizar uma comparação das estimativas apresentadas no trabalho com estudos de cunho internacional, sob a perspectiva brasileira, e a viabilizar os *scripts* do projeto, implementado na linguagem de programação R. O trabalho, de cunho metodológico, apresenta como principal resultado e contribuição disponibilizar uma nova base de dados de anos médios de escolaridade anualizada e mais precisa para estudos em nível nacional. Ela será útil para pesquisas na área econômica que necessitam de dados de *proxy* de capital humano e auxiliar na compreensão da análise do desenvolvimento educacional do país. A partir de uma breve comparação da estimativa de escolaridade apresentada neste trabalho com estudos internacionais que estimam a escolaridade do Brasil, o estudo de Lutz et al. (2018) é o que apresenta maior correlação com as estimativas.

Palavras-chave: Capital humano. Análise da educação. Metodologia para coleta, estimativa e organização de dados macroeconômicos. Acesso a dados.

ABSTRACT

The aim of this work is to estimate the average years of schooling of the population between 15 to 64 years old (1925 to 2015) using historical enrollment series, microdata from the census, PNAD Continuous, and population data from IBGE. Also, the secondary purpose is to expand the estimate by gender, race/color (1925 to 2015), and state (1950 to 2015). This work is proposed to carry out a comparison of the estimates presented in the work with international studies, from a Brazilian perspective, and to make available the project scripts, implemented in the R programming language. The work, of methodological nature, presents as the main result and contribution to make available a new database of annualized and more accurate average years of schooling for studies at the national level. It will be useful for economic research that provides human capital proxy data and to assist in understanding the analysis of the country's educational development. From a brief comparison of the schooling estimate presented in this work with international studies that estimate schooling in Brazil, the study by Lutz et al. (2018) is the one with the highest correlation with the estimates.

Keywords: Human capital. Analysis of education. Methodology for collecting, estimating, and organizing macroeconomic data. Data access.

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE EQUAÇÕES

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

ADF – Teste de Dickey-Fuller Aumentado

EF – Ensino Fundamental

EM – Ensino Médio

ES – Ensino Superior

PEA – População Economicamente Ativa

PIM – Método do Inventário Perpétuo

PNAD – Pesquisa Nacional por Amostra de Domicílios

SIDRA – Sistema IBGE de Recuperação Automática

IBGE – Instituto Brasileiro de Geografia e Estatística

ISCED – Classificação Internacional Normalizada da Educação

OCDE – Organização para a Cooperação e Desenvolvimento

UNESCO – Organização das Nações Unidas para Educação, Ciência e Cultura

UIS *Statistics* – UNESCO Institute for Statistics

SUMÁRIO

1 INTRODUÇÃO

Há na literatura diversas tentativas de mensuração do capital humano que visam quantificar a relação entre nível educacional e variáveis de resultado econômico e social (desde a análise do crescimento econômico até a aferição de impactos sobre mortalidade, fertilidade e distribuição de renda). Entretanto, devido ao fato de o capital humano ser multifacetado, ou seja, ser composto por diferentes atributos, apresenta a desvantagem de ser de difícil mensuração (BARRO; LEE, 2001).

Os principais erros das *proxies* de capital humano são a medição incorreta e a utilização de uma *proxy* imperfeita. Os anos médios de escolaridade são extensamente utilizados como *proxy* (WOESSMANN, 2003). Essa forma de mensuração é uma alternativa de estimação do campo educacional que supera a utilização da taxa de matrícula e de alfabetização como *proxy* para medição (BARRO; LEE, 1996). Apesar das limitações da metodologia a partir dos anos médios, ela possui a vantagem de apresentar uma ampla disponibilidade de dados. Através de sua utilização é possível estimar uma série histórica maior de escolaridade, a partir do início do século XX, diferentemente de outras abordagens, como a metodologia de estimação pela qualidade educacional, taxa interna de retorno e valor de mercado, que necessitam de dados mais difíceis de serem obtidos em registros históricos mais antigos.

De acordo com a classificação de Ludger Wossmann (2003), há predominantemente três metodologias principais de estimação dos anos médios de escolaridade: o método do inventário perpétuo (PIM), o de projeção e o de utilização de dados censitários como *benchmark*. Esta última tende a ser mais utilizada perante as demais, porém como as informações mais precisas de dados censitários de nível educacional só apresentam um registro histórico mais detalhado em períodos mais recentes, muitos estudos acabam mesclando dados de matrícula com informações censitárias, como as metodologias implementadas por Barro e Lee (1993, 2001, 2013) e Lee e Lee (2016). Suas estimações são extensamente empregadas em estudos em nível internacional e nacional como *proxy* para o capital humano e análise do comportamento educacional. Entretanto, devido ao fato de utilizarem dados de distribuição educacional da UNESCO, as estimativas de Barro e Lee, bem como as de Lee e Lee, apresentam inconsistências em suas medições (LUTZ et al. 2007), principalmente no contexto brasileiro, inviabilizando o seu aproveitamento no campo econômico e educacional.

À vista disso, a principal contribuição do trabalho é disponibilizar uma nova base de dados de anos médios de escolaridade anualizada para estudos à nível nacional. Além de mais

precisa, essa base possui dados desagregados por gênero, cor/raça e por região. O projeto servirá como subsídio para estudos na área econômica, que necessitam de dados de *proxy* de capital humano, e para análise do desenvolvimento educacional do país.

O objetivo da pesquisa quantitativa, portanto, consiste em criar uma base de dados de anos médios de escolaridade de 1925 a 2015, através da utilização de microdados do censo e PNAD Contínua do IBGE como *benchmark* e série histórica de matrícula do Kang et al. (2021). O intuito secundário do trabalho é realizar uma breve comparação dos dados obtidos com os outros estudos internacionais sobre anos médios de escolaridade e desagregação da estimação em gênero, cor/raça (1925 a 2015) e por região (1950 a 2015). Ademais, busca-se implementar, computacionalmente, o projeto na linguagem de programação R e disponibilizar os *scripts* utilizados, possibilitando a reprodutibilidade da pesquisa.

O respectivo trabalho, portanto, está dividido em três seções. Primeiro capítulo uma revisão de literatura que aborda uma breve descrição do histórico do capital humano, contextualização e definição. Da mesma forma, a exposição de uma das suas principais *proxies*: mensuração por anos médios de escolaridade. No segundo capítulo, é detalhado o modelo empírico para a estimativa de escolaridade no Brasil. Com base nos resultados obtidos, no último capítulo é feita uma breve comparação entre a base de dados apresentada e estudos internacionais e análise dos resultados. O último capítulo destaca as considerações finais da pesquisa.

2 CAPITAL HUMANO E ANOS MÉDIOS DE ESCOLARIDADE

Neste capítulo é feita uma contextualização da definição e da história do capital humano, é descrita a metodologia de estimação dos anos médios de escolaridade como *proxy* da educação e é apresentada uma breve análise dos estudos internacionais dessa área.

2.1 HISTÓRIA E DEFINIÇÃO

Uma das características dos fatores de produção, capital e trabalho, é apresentarem retornos decrescentes de escala. Portanto, em virtude disso, ao longo dos anos os países tenderiam a uma taxa de crescimento menor do PIB. Contudo, o cenário observado na década de 1950 foi o oposto para muitos países, havendo um crescimento persistente da renda, indicando que capital e trabalho não eram os únicos insumos determinantes para o crescimento (BECKER, 1964). Da mesma forma que, paralelo a esse fato, no âmbito microeconômico, era observada uma alta variação da renda per capita entre os indivíduos. À vista disso, os fatores de produção definidos na época eram insuficientes para explicar essas observações (MINCER, 1984).

Gary Becker (1964), Ted Schultz (1961) e Jacob Mincer (1958) foram pioneiros para a definição do conceito de capital humano, que segundo Jacob Mincer (1984), obteve um papel considerável para explicar essas observações macro e microeconômicas. Os economistas já sabiam da importância dos indivíduos na riqueza dos países, como consta nas obras de Adam Smith, Marx, Marshall e demais antecessores. No entanto, o fato de as pessoas investirem em si e esses investimentos serem elevados, ao ponto de terem altos impactos econômicos, só foi incorporado na economia formal na década de 1950, período relativamente recente. Muitos paradoxos da economia podem ser compreendidos com a adição do conceito de investimento humano (SCHULTZ, 1961). Ou seja, a diferença no nível de estoque de capital dos indivíduos tende a explicar, em parte, a variabilidade da renda per capita e a sua alta taxa de crescimento observada em 1950 (MINCER, 1984).

Pietro Garibaldi (2006, p.152) consegue descrever de forma sucinta o conceito do fator de produção, capital humano:

> *To understand how economists think about training, it is first necessary to understand how economist think about education in general. The most important view of education is the theory of human capital. Loosely speaking, human capital corresponds to any stock of knowledge or characteristics the works has (either innate or acquired) that contributes to his or her productivity. This definition is very broad, but enable us to think not only about years of schooling, but also about a variety of*

> *other characteristics as a part of human capital investment. Training done during the employment relationships is one of such characteristics[...]*

A título de exemplificação podem ser classificados como investimentos desse capital, gastos com educação, treinamentos, assistência médica, entre outros. Nessa análise, a educação e o treinamento entram como um dos pilares de investimento, visto que aumentam conhecimentos e habilidades, gerando ganhos para os indivíduos e uma elevação da produtividade. Assim dizendo, um dos principais conceitos da teoria do capital humano é a abordagem de que o investimento no ensino aumenta a produtividade (BECKER, 1964).

A teoria da sinalização de Spence (1973) surge em oposição à ideia descrita por Becker de aumento do investimento e concomitante elevação da produtividade. Nessa teoria o trabalhador tenderia a ganhar um maior retorno salarial por "sinalizar" que apresenta maior produtividade, e não por demonstrá-la de forma efetiva. Entretanto, conforme destacado por Psacharopoulos (1979), no momento da contratação há poucas informações que atestem o potencial produtivo do empregado. Após um tempo de serviço, caso o trabalhador não corresponda às expectativas, a tendência é que ocorra uma diminuição do seu prêmio salarial, minimizando o efeito da teoria.

Um dos principais desafios de análise do capital humano consiste na mensuração do seu estoque total e do seu investimento. Conforme descrito por Schultz (1961), diferentemente do capital físico, a estimação do investimento não se resume somente em computar os gastos para a produção de um bem. Haja vista que, no capital humano, há a dificuldade de distinção entre as despesas relativas ao consumo e ao investimento.

Como destacado por Barro e Lee (1993), estudos empíricos anteriores utilizavam a taxa de matrícula e a taxa de alfabetização como *proxy* de medida. Entretanto, apesar de serem dados extensamente divulgados, não medem de forma efetiva o seu estoque. A taxa de alfabetização restringe a sua análise somente para as fases iniciais de escolaridade. No caso da taxa de matrícula, representa uma medida do esforço de um país em alterar o seu estoque de capital humano (PSACHAROPOULOS; ARRIAGADA, 1986). Se explorada a sua taxa bruta, ela pode introduzir erros de mensuração relativos à repetência e à desistência (BARRO; LEE, 1993). Porém, destaca-se que no Brasil não há dados antigos confiáveis de repetência e evasão devido a erros nos dados oficiais (RIBEIRO, 1991). O ideal é utilizar a taxa líquida de matrícula, porém nem sempre esses dados estão disponíveis, principalmente em países em desenvolvimento (BARRO; LEE, 1993).

O nível educacional, expresso em termos de anos médios de escolaridade, supera a taxa de matrícula e a taxa de alfabetização, sendo mais representativo para mensurar o estoque de

capital humano agregado (BARRO; LEE, 1993). Desde 2010 os anos médios de escolaridade passam a ser um indicador contabilizado no índice de Desenvolvimento Humano (IDH), substituindo a taxa de analfabetismo como indicador educacional e a expectativa de vida escolar/anos esperados de escolaridade como um substituto das taxas brutas de matrícula (UNITED NATIONS DEVELOPMENT PROGRAMME, 2020). O indicador de anos médios de escolaridade é frequentemente utilizado para comparações *cross-country* e seu cálculo se torna complexo por algumas razões: as durações de níveis de escolaridade variam entre países, portanto há uma dificuldade de se estabelecer um padrão internacional; e o cálculo pode ser enviesado pelo número de indivíduos que não concluíram o curso (POTANCOKOVÁ et al., 2014). Diversas metodologias utilizadas para o cálculo dos anos médios de escolaridade realizam a conversão de distribuição educacional em anos médios. Conforme descrita pela fórmula de Barro e Lee (2013), o cálculo da média dos anos de escolaridade por faixas etárias, "a" representa uma coorte etária, é feito a partir de uma média ponderada da duração, "Dur", de cada nível de ensino, "j", com base no peso "h" que corresponde à fração do grupo de idade e seu respectivo nível de ensino em um período de mensuração "*t*" (Equação 1 - anos médios de escolaridade por coorte etária no período *t*). Um dos problemas na estimação dos anos médios de escolaridade de um amplo grupo de países é determinar a duração do nível educacional para a população de ensino incompleto, pois exige um grau de detalhamento nos dados que torna inviável a adoção em base de dados muito extensas. A título de exemplificação, no caso das pesquisas do Barro, ele atribui 4 anos para o ensino superior completo e 2 para o incompleto para todos os países como uma forma de simplificação metodológica, perdendo um pouco do seu grau de detalhamento.

$$s_t^a = \sum_j h_{j,t}^a \, Dur_{j,t}^a \tag{1}$$

A equação, aplicada a coorte etária de 15 a 19 anos (a = 15-19 anos), pode ser compreendida como o somatório da distribuição educacional dessa faixa-etária multiplicado pela respectiva duração do nível de ensino. Considere-se hipoteticamente que, em 1960, 53% da coorte tinha ensino fundamental (EF) incompleto; 3% EF completo; 3% ensino médio (EM) incompleto; 1% EM completo; 0,5% ES incompleto; 0,2% ES completo. Além disso, que a duração de cada nível de ensino seja de 4 anos; 8 anos; 10 anos; 11 anos; 12 anos; 14 anos. A partir desses dados, aplicando-se a Equação (1), é obtida uma média de 2,858 anos de escolaridade para a coorte de 15 a 19 anos. Da mesma forma, a escolaridade média da população de 15 a 64 anos pode ser encontrada a partir da média ponderada desse resultado aplicado a cada coorte etária.

Pelo fato de o capital humano ser multifacetado, a adoção de anos médios de escolaridade como *proxy* de medida apresenta imperfeições. Portanto, há na literatura três principais críticas à utilização dos anos médios de escolaridade como *proxy*. Em primeiro lugar, um ano a mais de estudo eleva o estoque de capital humano independentemente da série cursada, ou seja, assumem-se retornos constantes de escala. Em outras palavras, um ano de escolaridade adicional para o indivíduo que cursa o fundamental e outro que cursa o ensino superior acarretam a mesma elevação no estoque de capital humano, considera-se um ano adicional de educação como uma constante para qualquer nível de ensino. Em segundo lugar, essa abordagem pressupõe que um aumento de escolaridade em diferentes regiões com sistemas educacionais distintos é equivalente, ou seja, o aumento do estoque de capital humano independe da qualidade educacional (WOESSMANN, 2003). Em terceiro lugar, a metodologia presume que os trabalhadores são perfeitamente substitutos entre si, desconsiderando diferenças de habilidade entre indivíduos (MULLIGAN; SALA-I-MARTIN, 2000).

Ademais, em uma análise da série histórica de anos médios de escolaridade, Cohen e Soto (2007) observaram que as regiões que apresentavam taxa de crescimento da escolaridade mais elevada também foram as regiões que começaram com níveis muito baixos de educação, como no caso da África. Possivelmente, o fato de um país ter dobrado a escolaridade de 1 para 2 anos talvez não seja um indicador efetivo que a produtividade tenha aumentado na mesma proporção. Para ser uma *proxy* mais efetiva deveria haver um peso diferente dependendo da quantidade de anos de escolaridade acumulados e do tipo de sistema educacional do país (WOSSMANN, 2003).

Conforme destacado por Hanushek e Wossmann (2012), os anos de escolaridade são relevantes para análise do crescimento econômico somente quando acarretam uma elevação do conhecimento dos alunos. Os testes educacionais tendem a suprir esse lado qualitativo da análise do capital humano, contudo podem não expressar a escolaridade da população em idade ativa (BARRO, R. J. LEE, 2001). Essa metodologia apresenta a limitação de disponibilidade dos dados, assim como a técnica de aferição do capital humano em termos de valor de mercado, a partir do cálculo da taxa de retorno da educação. Essa aferição, expressa em unidades monetárias, costuma ser uma forma de medida mais efetiva que a utilização dos anos médios de escolaridade, contudo o seu cálculo também é limitado devido à necessidade de uma ampla base de dados salariais (BARRO, R. J. LEE, 2001), e ainda assim tende a não incorporar externalidades e fatores qualitativos.

Portando, em uma análise de uma série histórica em um intervalo grande de tempo, com dados do início do século XX, os anos médios de escolaridade são uma *proxy* razoável do capital

humano devido à maior disponibilidade de observações do período. Além do mais, a variável torna-se relevante principalmente na análise do desenvolvimento do Brasil, devido à existência de um histórico escolar defasado.

2.2 ANOS MÉDIOS COMO *PROXY* DO CAPITAL HUMANO

Segundo a classificação do autor Ludger Wössmann (2003), há na literatura três técnicas de cálculo dos anos médios de escolaridade. A metodologia do Inventário Perpétuo (PIM), abordada por Lau et al. (1991), Nehru et al. (1995), a partir de séries longas de matrículas. O PIM é uma metodologia de transformação de variáveis de fluxo (no caso, matrícula) em variáveis de estoque (anos de estudo) considerando uma defasagem educacional. Neste modelo de fluxo são requeridos muitos dados referentes à taxa de matrícula, repetência e mortalidade, sendo necessário muitas vezes fazer suposições e interpolações para preenchimento de lacunas, por falta de dados. A segunda técnica é abordada por Kyriacou (1991) que utiliza um método de conversão de taxa de matrícula em anos médios de escolaridade a partir de suposições paramétricas e técnica da projeção. Por fim, a terceira técnica, inicialmente adotada por Psacharopoulos e Arriagada (1986) e posteriormente refinada por Barro e Lee (1993; 1996; 2001), com utilização de dados censitários e pesquisas nacionais como *benchmark*, é considerada a mais elaborada. Porém, há uma dificuldade de implementação dessa técnica em períodos mais antigos, por isso estudos como o de Barro e Lee acabam mesclando metodologias, utilizando dados de matrícula e censos como *benchmark*.

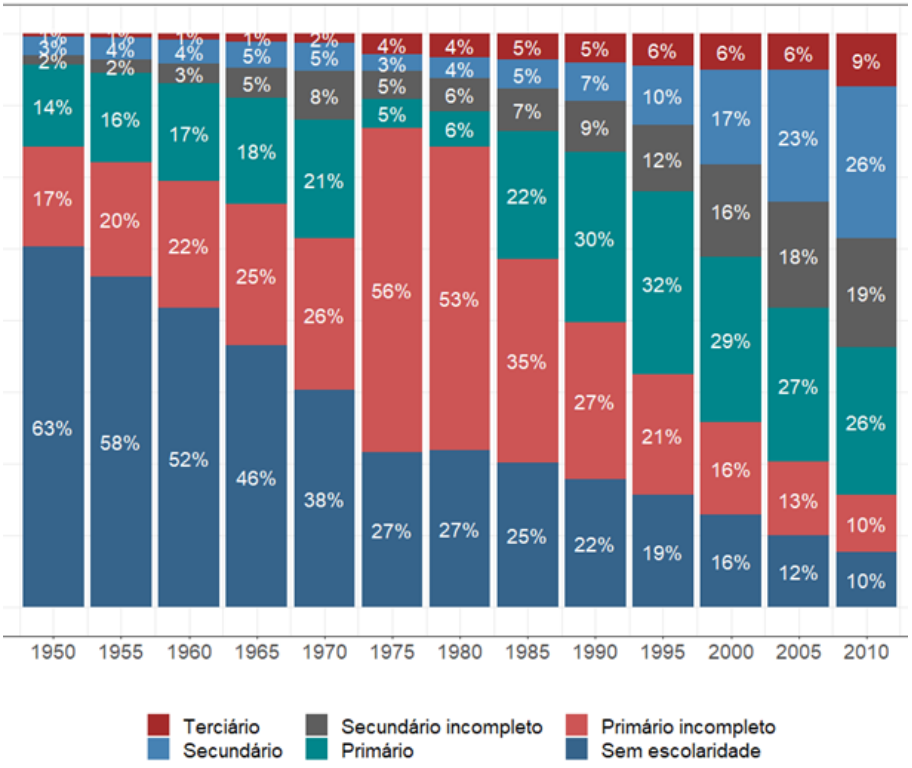
Barro e Lee (1993) preconizaram a utilização de censos e de pesquisas (UNESCO, U.N. *Demographic Yearbooks* e outros recursos) como *benchmark* em sua metodologia de estimação. [1] Com isso, maximiza-se a utilização de um maior número de observações de pesquisas e censos, para então preencher os dados remanescentes com informações de matrículas, mediante o uso da metodologia de inventário perpétuo (PIM). Essa técnica foi utilizada até o estudo de 1993, em que Barro e Lee estimaram o capital humano através da elaboração de uma base de dados de anos médios de escolaridade para 129 países de 1960 a 1985 em intervalos de cinco anos. Após este período, os autores empregaram a utilização de uma metodologia de projeção de dados educacionais a partir dos *benchmarks*.

[1] Estudos que utilizaram a base de dados de Barro e Lee (1993): Barro (1999); Easterly; Levine (1998); Hall; Jones (1999); Rajan; Zingales (1998); Ramey e Ramey (1995); Sachs e Warner (1995).

Em trabalhos subsequentes, Barro e Lee (1996, 2001, 2013) aperfeiçoaram sua metodologia e ampliaram a sua base de dados. A última atualização metodológica está contida em Barro e Lee (2013) e a base de dados mais recente é de junho de 2018. Nessa versão, o estudo engloba 146 países entre o período de 1950 a 2010, com estimativas a cada cinco anos. Parte das alterações entre as versões refletem as críticas de De La Fuente e Doménech (2006) e Cohen e Soto (2007) aos trabalhos anteriores. A título de exemplificação, no período de 1960, os bolivianos com 15 anos ou mais de idade apresentavam uma educação muito similar à dos franceses. Em 1980 os anos médios de escolaridade do Equador superavam a escolaridade da Itália, segundo Cohen e Soto (2007). De La Fuente e Doménech (2006) verificam uma mudança brusca nas séries históricas de Barro e Lee. Apesar de a base de dados ser utilizada por uma série de estudos conhecidos na literatura de crescimento, há graves imprecisões nos dados estimados para o Brasil.

Um dos problemas mais genéricos das estimativas de Barro e Lee está na confiança e precisão dos dados, além da falha na decomposição dos níveis de ensino entre concluintes e não concluintes (SPERINGER et al., 2015). Conforme pode ser observado no Gráfico 1, nos anos de 1975 a 1980, no Brasil, há uma inconsistência dos dados de nível educacional, uma elevação abrupta do primário incompleto com a respectiva diminuição do primário completo.

Gráfico 1 - Distribuição Educacional no Brasil, população de 15 anos ou mais, Barro e Lee (2018)

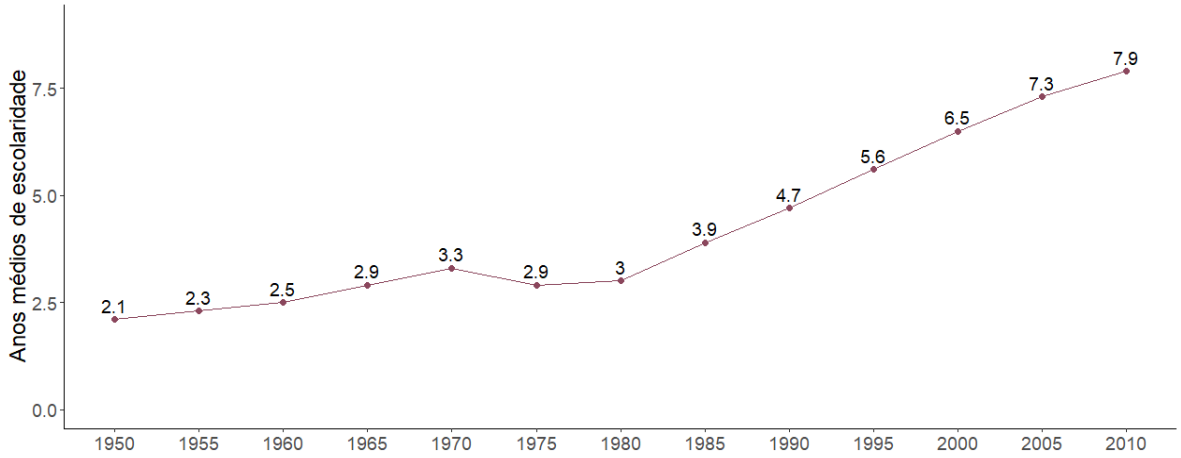


Fonte: Elaboração própria baseado no artigo Speringer et al. (2015) e dados de Barro e Lee (2018).

O Gráfico 2 mostra uma queda de 12% entre 1970 e 1975, uma redução na escolaridade que só é recuperada mais próximo do fim da ditadura militar em 1985. Esse resultado é implausível, uma vez que implica uma expressiva queda no estoque de capital humano na década de 1970 (aumento do número de matrículas, urbanização e crescimento econômico).

Uma possível justificativa para esses resultados seria a Lei 5.692/1971, reforma que alterou o currículo dos níveis de ensino primário e secundário. O antigo ensino primário (1ª a 4ª série) e o primeiro ciclo do ensino médio (5ª a 8ª série) se fundiram para originar o ensino de primeiro grau (ensino fundamental a partir de 1996), enquanto o segundo ciclo do ensino médio passou a ser denominado ensino de segundo grau (ensino médio a partir de 1996). Nos estudos de De La Fuente e Doménech (2006) ao analisarem a distribuição educacional do Canada a partir dos dados da Organização das Nações Unidas para Educação, Ciência e Cultura. (UNESCO), percebem um padrão bastante implausível, sendo sugerida a existência de uma variação dos critérios de classificação do nível educacional. Portanto, a essa incongruência observada nos dados de Barro para o Brasil não é necessariamente um problema metodológico de estimação, mas deve-se ao excesso de confiança dos autores nos dados da UNESCO, principal fonte de Barro e Lee. Se observada a distribuição educacional disponibilizada pelo UNESCO *Institute for Statistics* (UIS *Statistics*), os resultados apresentam o mesmo padrão do Gráfico 1. Ou seja, o problema está na conversão dos dados censitários em classificação *International Standard Classification of Education* (ISCED), padrão internacional de classificação de níveis de ensino, auxiliando na comparação de diferentes sistemas educacionais. Até 1970, possivelmente, a UNESCO considera como sendo primário incompleto todos os estudantes que cursaram a 1ª até a 3ª série (classificação do antigo ensino primário). Porém, após a reforma educacional, Barro e Lee adotam, erroneamente, uma nova classificação, sendo considerado primário incompleto da 1ª a 7ª série (correspondendo ao primeiro grau/ensino fundamental). Apesar de Barro e Lee (2001) procurarem levar em conta as diferenças educacionais de cada país, eles erraram por não adotarem uma padronização da classificação da distribuição educacional (considerar para todos os anos a classificação do antigo ensino primário, mesmo após a reforma de 1971). Essa hipótese explica o resultado encontrado por Pessôa et al. (2019) na análise dos dados de Barro e Lee (2001). No artigo eles verificam que o Brasil perde vantagem educacional, em termos de anos médios de escolaridade, em relação a outros países do mundo. Em 1960, o nível educacional do Brasil estava no patamar da média mundial, porém entre 1960 e 2000 o país ficou progressivamente para trás.

Gráfico 2 - Anos médios de escolaridade no Brasil, população de 15 anos ou mais, 1950-2010, Barro e Lee (2018)

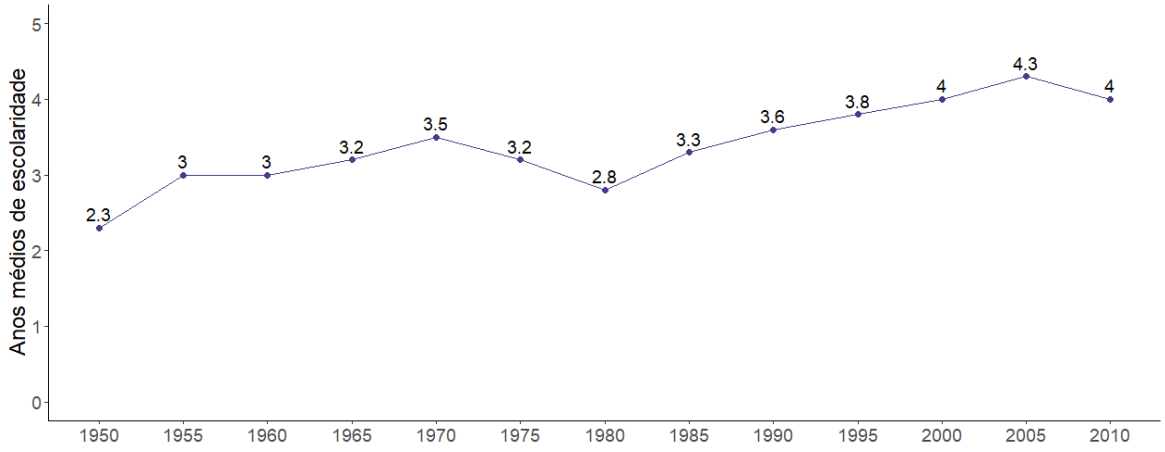


Fonte: Elaboração própria, dados de Barro e Lee (2018).

No Gráfico 3, percebe-se que essa diferença de classificação do nível de ensino se reflete no comportamento dos anos médios de escolaridade das coortes etárias. Há uma queda acentuada na escolaridade da população em 1980 na faixa etária de 45 a 49 em relação ao mesmo grupo populacional com 35 a 39 anos de idade em 1970 (ou seja, da mesma coorte), o que é implausível. O cenário esperado seria que a distribuição educacional fosse relativamente constante até 1995, quando a coorte estivesse na faixa etária de 60 a 64 anos. Passado esse período, haveria a atuação de diferenciais de mortalidade por escolaridade: o nível de escolaridade da população tenderia a aumentar pela mortalidade maior dos menos escolarizados após os 64 anos de idade. Porém não é o que ocorre.

Lee e Lee (2016) adotam a mesma metodologia de Barro e Lee (2013) e, portanto, apresentam as mesmas inconsistências para as estimativas do Brasil (ver Apêndice A). Contudo, o estudo de Lee e Lee (2016) apresenta a vantagem de estimar uma série mais longa de anos médios de escolaridade (a partir de 1870). A fim de construir uma série mais extensa, os autores adotaram a metodologia de fluxo baseado em matrícula/PIM, método semelhante ao utilizado por Morrisson e Murtin (2009) e Van Leeuwen e Van Leeuwen-Li (2014).

Gráfico 3 - Anos médios de escolaridade no Brasil, coorte etária de 15-19, Barro e Lee (2018), em 1950

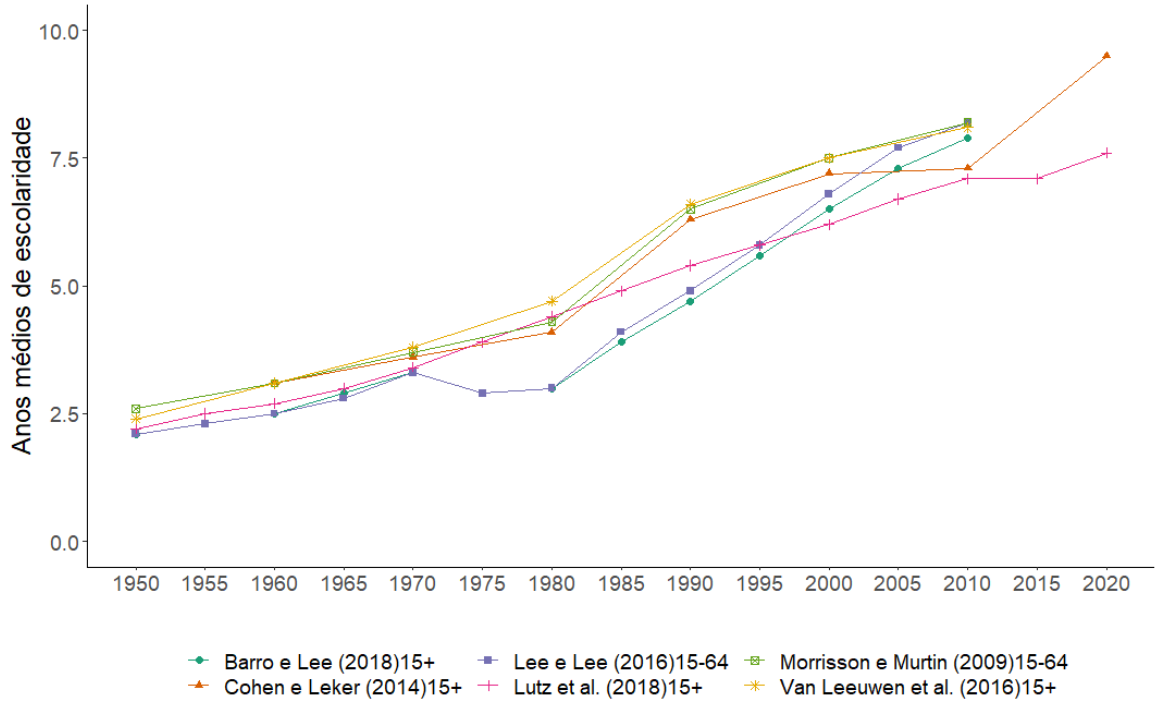


Fonte: Elaboração própria, dados de Barro e Lee (2018).

Há na literatura outros estudos em nível internacional que estimam os anos médios de escolaridade para o Brasil, porém há bastante divergência entre as estimativas (ver Gráfico 4). Os estudos de Lutz et al. (2018), Van Leeuwen e Van Leeuwen-Li (2014) e Morrisson e Murtin (2009) aparentam ser os mais coerentes, pois apresentam um crescimento do estoque de escolaridade da população adulta ao longo dos anos. Por outro lado, os estudos divergem em termos de nível por apresentarem metodologias distintas de estimação: enquanto Lutz et al. (2018) projeta a população para trás, Van Leeuwen e Van Leeuwen-Li (2014) e Morrisson e Murtin (2009) utilizam um cruzamento de diferentes conjuntos de dados, dados de escolaridade de Cohen e Soto (2007) e dados de matrícula.

Vale também salientar a metodologia de Cohen e Soto (2007), que é bem semelhante à de Barro e Lee (2013), porém com um número de observações menor e utilização de outro método de estimativa para valores não observados. Segundo Cohen e Soto (2007), a diferença de seu estudo para os anteriores é utilizar informações de nível de escolaridade por idade, aspecto esse já adotado nas pesquisas mais recentes. Todavia, Barro e Lee (2001) destacam que Cohen e Soto (2007) usam fontes da Organização para a Cooperação e Desenvolvimento (OCDE) para países pertencentes à OCDE e dados da UNESCO para não pertencentes, sem considerar que há diferenças significativas entre as bases de dados.

Gráfico 4 - Comparação entre estudos, anos médios de escolaridade estimados para o Brasil, população de 15 anos ou mais



Fonte: Elaboração própria, dados de Barro e Lee (2018), Lee e Lee (2016), Morrisson e Murtin (2009), Cohen e Leker (2014), Lutz et al. (2018) e Van Leeuwen et al. (2016).

Nota: Lee e Lee (2016) e Morrisson e Murtin (2009) utilizam a faixa etária de 15 a 64 anos.

Em outro trabalho conhecido, Van Leeuwen e Van Leeuwen-Li (2014) empregam uma versão modificada de Cohen e Soto (2007) após 1960. Antes desse período, os autores utilizam o método PIM. Morrisson e Murtin (2009) aplicam essa mesma metodologia para estimar o nível educacional dos anos de 1960 a 1870 (dados de repetição e abandono da UNESCO) para 74 países e combinam essas informações com estimativas de Cohen e Soto (2007) para o período de 1960 a 2010. Como Morrisson e Murtin (2009) e Van Leeuwen e Van Leeuwen-Li (2014) utilizam dados do Cohen e Soto (2007) nas suas estimativas, os estudos podem apresentar as mesmas inconsistências apresentadas pelo último trabalho.

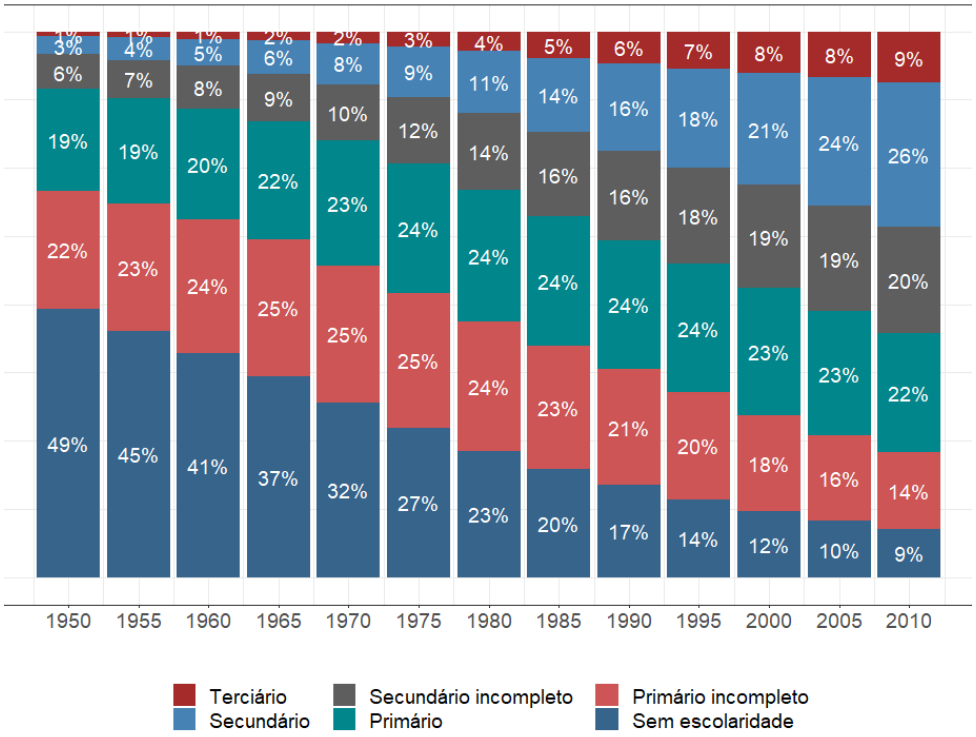
No estudo de Cohen e Leker (2014), que atualiza os achados de Cohen e Soto (2007), há somente duas observações para a estimativa de escolaridade no Brasil, cujas fontes são UNESCO (1980) e OCDE (1999). Além do mais, para alguns países, como no caso da China, os autores utilizaram exclusivamente dados de matrícula líquida, ou seja, não utilizam

benchmarks. Por conta do baixo número de observações, as estimativas não aparentam ser muito confiáveis.

Lutz et al. (2018) na estimativa da escolaridade no Brasil, por sua vez, apresentam uma distribuição educacional que gera os anos médios de escolaridade, garantindo uma certa consistência nos dados, assim como fazem Barro e Lee (2013). Estes estudos diferem de Morrisson e Murtin (2009), que não dão grande importância a uma distribuição de escolaridade confiável, desde que os estoques de escolaridade o sejam. Lutz et al. (2018) aprimoram os estudos de KC et al. (2010) e Lutz et al. (2007), calculando a distribuição educacional a partir do acompanhamento de coortes etárias e sem utilizar dados de matrícula. Pressupõe-se que a proporção de mulheres de 25 a 29 anos sem escolaridade em 1970 e de 55 a 59 anos em 2000 devem ser iguais por se tratar da mesma população. Os autores optaram por esse método e não pela utilização de dados de matrícula porque a última opção teria menos credibilidade. Ademais, há o viés da coleta de dados: escolas tenderiam a querer "aumentar" seus indicadores de matrícula para obtenção de mais recursos.

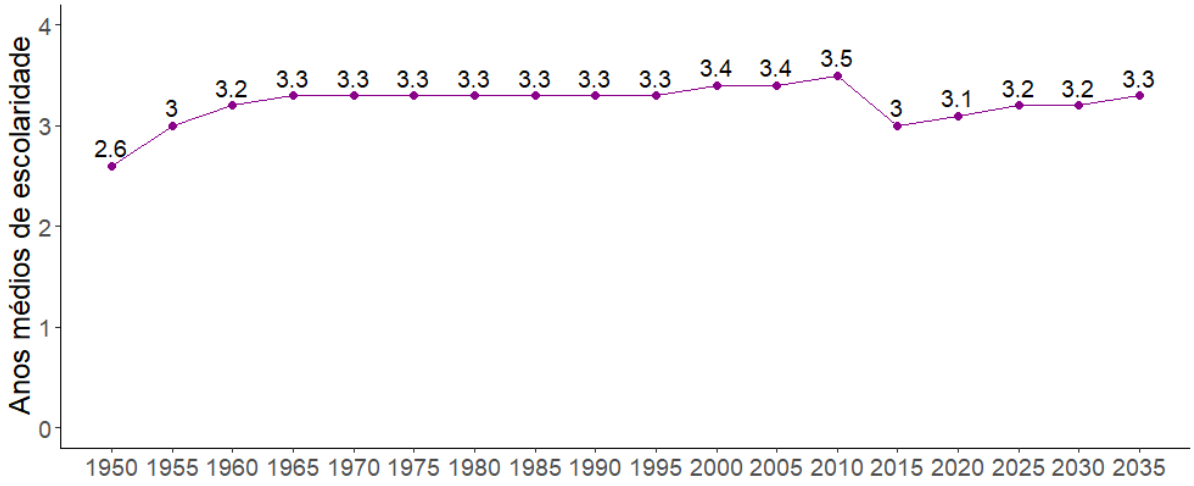
Destaca-se também que Lutz et al. (2018) utilizam preferencialmente bases de dados locais e não as disponibilizadas pela UNESCO. Por isso, esse trabalho não apresenta o mesmo erro metodológico das pesquisas concorrentes entre 1970 e 1980. Ainda assim, analisando por coorte etária, percebe-se que Lutz et al. (2018) apresenta uma incongruência nas estimativas posteriores a 2010. Parece haver uma implausível diminuição do estoque de capital humano: há mortalidade mais elevada da população mais instruída na faixa etária dos mais idosos, como ilustrado no Gráfico 6.

Gráfico 5 - Distribuição Educacional no Brasil, população de 15 anos ou mais, Lutz et al. (2018)



Fonte: Elaboração própria, dados de Lutz et al. (2018).

Gráfico 6- Anos médios de escolaridade no Brasil, coorte de 15-19 anos, Lutz et al. (2018), em 1950



Fonte: Elaboração própria, dados de Lutz et al. (2018).

A Tabela 1 contém uma síntese das principais diferenças metodológicas dos estudos mais relevantes: Barro e Lee (2013), Cohen e Leker (2014) e Speringer et al. (2015). À vista das divergências entre os estudos internacionais, nas próximas seções, estima-se uma série

anual de anos médios de escolaridade para o Brasil, com abertura por gênero, cor/raça e por estado, tendo como base informações do censo e dados de matrícula.

Tabela 1- Diferenças metodológicas para estimação dos anos médios de escolaridade nacional

	Barro e Lee (2013)	Cohen e Leker (2014)	Speringer et al. (2015)
Faixas Etárias	intervalo de 5 anos: 15-19; 20-24; … 75+	intervalo de 5 anos: 15-19; 20-24; … 80+	intervalo de 5 anos: 15-19; 20-24; … 85+
Indicadores Educacionais	Distribuição de escolaridade + Média de anos de escolaridade	Somente média de anos de escolaridade	Distribuição de escolaridade + Média de anos de escolaridade
Período	1950-2010 (intervalo de 5 anos)	1960-2020 (intervalo de 10 anos)	1970-2010 (intervalo de 5 anos)
Categorias Escolares	7 categorias - sem escolaridade; primário (total e completo); secundário (total e completo); terciário (total e completo).	Não mencionado. De acordo com os dados disponibilizados aparenta 7 categorias como as de Barro e Lee.	6 categorias - sem escolaridade; primário (incompleto e completo); secundário (*lower* e *upper*); terciário (completo).
Base de Dados	5 observações – 1950, 1970 (pesquisa/censo), 1976, 1980 e 2004. Nenhum dado foi coletado diretamente de fontes nacionais do país.	Cohen e Leker – 2 observações UNESCO – 1980 e OCDE 1999	IPUMS/censo - 1970, 1980, 1991; pesquisa nacional - 2000; 2010; UNESCO - 1970, 1976, 1980
Fonte	fonte: UNESCO; dados de matrícula	fonte: UNESCO ou OCDE; dados de matrícula	fonte: dados nacionais; não utiliza dados de matrícula
Metodologia	Dados censitários, preenchimento dos valores faltantes com projeção (*backward e forward*) / dados de matrícula (para população em idade escolar)	Dados censitários, preenchimento dos valores faltantes com projeção (*backward e forward*). No caso de não haver informação confiável utiliza pesquisas mais recentes ou dados de matrícula.	Acompanhamento de coortes etárias em intervalos de 5 anos a partir de um *benchmark*.

Fonte: Elaboração própria, tabela adaptada do artigo Speringer et al. (2015).

Nota: A metodologia da UNESCO (2013) não foi mencionada, pois diverge pouco ao do Barro e Lee (2013), dentre as principais modificações está a presença de 6 categorias (classificação é baseado no ISCED 1997 e 2011), os dados são coletados somente com fontes da UIS.

3 METODOLOGIA DE ESTIMAÇÃO

Neste capítulo é descrita a metodologia de estimação dos anos médios de escolaridade e da distribuição educacional do Brasil. A principal referência dessa seção são os microdados dos censos, da PNAD Contínua do IBGE e dados de matrícula de Kang et al. (2021), sendo adotada três metodologias de mensuração (*benchmark,* projeção do censo e estimativa de período intercensitário). A partir de pesquisas nacionais foi mensurado a escolaridade com a utilização de *benchmarks*. Devido ao fato dos microdados só estarem disponíveis a partir de 1960, foi adotada outra técnica de mensuração para o período anterior a 1950. Nesse caso, foi usada a metodologia de Lutz et al. (2007), a estimação é feita por meio de uma projeção da população do censo de 1960 para trás. (para estimativa de 1950 a 1925). Ademais, a implementação da metodologia modificada de Foldvári e Van Leeuwen (2009), possibilitou garantir uma série de anos médios anualizada e uma consistência entre a variável distribuição educacional (população sem escolaridade – SE, ensino fundamental – EF, médio – EM e superior – ES) e a variável anos médios de escolaridade.

3.1 DADOS E FONTES

As estimativas de escolaridade da população em idade ativa de 1960 a 2000 tomaram como base os microdados dos censos demográficos (registro das pessoas) do IBGE. Nessas bases de dados, as informações populacionais são desagregadas em nível individual, permitindo uma análise mais detalhada das variáveis educacionais. Para a realização da estimativa, foi necessária uma compatibilização da classificação dos anos médios entre os censos. Para o período mais recente, adotou-se também uma metodologia de compatibilização para o cálculo dos anos de 2012 a 2015, através da utilização dos microdados da PNAD Contínua (registro de pessoas). Para o cálculo dos anos intercensitários de 1950 a 2012, foi utilizada a metodologia de *backward* and *forward estimation* de Foldvári e Van Leeuwen (2009) com algumas modificações. A fim de se aplicar essa metodologia, foi necessária a utilização de dados de matrícula e informações populacionais baseadas nos censos demográficos, interpolada por uma função *spline* cúbica. Além disso, a técnica exigiu também a utilização de informações de duração média de anos de escolaridade por nível de ensino e distribuição populacional por nível educacional dos censos de 1960 a 2000. A partir disso, foi possível estimar uma série de distribuição educacional e anos médios de escolaridade anualizada.

Conforme destacado, visto que não há informação de microdados censitários no período anterior a 1960, optou-se por utilizar como referência o censo de 1960 projetado para trás. Com isso, estima-se o nível educacional a partir do acompanhamento de coortes ao longo do tempo, empregando conjuntamente dados populacionais do IBGE e de matrícula. A vantagem dessa metodologia é não necessitar de uma série histórica de matrícula tão extensa, ao contrário do cálculo via fluxo de matrícula/PIM. Como não é disponível uma ampla série histórica de matrículas no Brasil (dados de ensino secundário e terciário a partir de 1870), a metodologia de projeção é bastante útil para o caso brasileiro. Além disso, o cálculo via PIM não garante que haja uma distribuição educacional compatível ao ano que se pretende estimar a variável anos médios de estudo (MORRISSON; MURTIN, 2009). Para a estimação via PIM dos níveis de escolaridade da população de 15 a 64 anos em 1925, seriam necessários dados de matrícula a partir de 1868. Com a metodologia de projeção adotada no trabalho, há a necessidade de cálculo da taxa bruta de matrícula somente para a população em fase escolar, o que é viável. Ademais, a projeção permite estimar distribuição educacional e anos médios de escolaridade consistentes entre si.

3.2 *BENCHMARKS*: CENSO E PNAD CONTÍNUA, 1960-2015

Para a utilização dos microdados do censo e PNAD Contínua como *benchmarks*, é preciso compilar as informações de dados de instrução disponíveis em cada ano. Ou seja, é necessário agregar as informações de nível de ensino e série que o indivíduo frequentou/frequentava a escola, para então, poder realizar a conversão desses valores em anos de escolaridade para cada indivíduo. Como não há uma padronização nos censos para a coleta dos dados de instrução, é necessário realizar uma compatibilização. No Apêndice B, consta o detalhamento de cada conjunto de microdados (Censos 1960-2000 e PNAD Contínua 2012-2015). Além disso, os seguintes critérios foram utilizados para a atribuição da escolaridade:

a) indivíduos que frequentaram uma série adicional à requerida para completar o grau de ensino: não foi adicionado um ano a mais de escolaridade. Em 1960, a título de exemplo, quem cursou a 5ª série do elementar recebeu quatro anos de escolaridade e não cinco, uma vez que não há como saber, em muitos casos, a quantidade de anos cursados. O mesmo procedimento foi adotado para os demais níveis de ensino. No caso da PNAD Contínua, a Lei 11.274/2006 alterou a duração do ensino fundamental para 9 anos. Aos que cursaram o 9º ano, não foi atribuído um ano adicional de escolaridade, ou seja, independentemente da série

cursada, 9ª ou 8ª série, foi atribuído 8 anos de escolaridade. Isso é um problema mínimo, uma vez que, em 2015 (ano final da série de dados), os estudantes que entraram no EF de 9 anos em 2007 não o tinham terminado ainda;

b) alfabetização de adultos e ensino não seriado: foram excluídos os casos em que não era possível a classificação de uma série escolar, excetos vestibulandos e mestrandos, para os quais se contabilizaram 11 e 17 anos respectivamente;

c) limite máximo de anos de estudo: atribuiu-se limite máximo de 17 anos. Essa padronização foi feita a fim de evitar que houvesse distorções ao se comparar dados de censo com informações da PNAD Contínua, devido à maior especificação de série por nível de ensino;

d) diversas metodologias utilizadas para o cálculo dos anos médios de escolaridade realizam a conversão de distribuição educacional em anos médios. Conforme descrito por Barro e Lee (2013), o cálculo da média dos anos de escolaridade por faixas etárias é feito a partir de uma média ponderada da duração de cada nível de ensino (ver Equação 1).

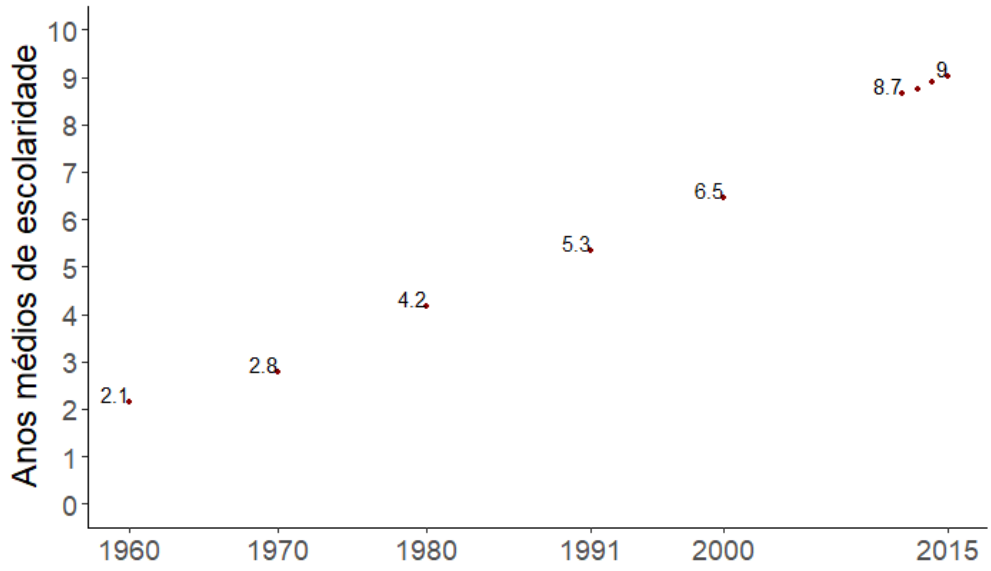
Diversas metodologias utilizadas para o cálculo dos anos médios de escolaridade realizam a conversão de distribuição educacional em anos médios. Conforme descrito por Barro e Lee (2013), o cálculo da média dos anos de escolaridade por faixas etárias é feito a partir de uma média ponderada da duração de cada nível de ensino (ver Equação 1).

Neste trabalho, foi possível determinar a escolaridade populacional em cada *benchmark* (sem escolaridade, EF, EM, ES) e estimar com maior precisão os anos de escolaridade, pois é determinada a duração de cada nível de ensino considerando a população com ensino incompleto. Os estudos de Barro e Lee, por exemplo, consideram quatro anos para o ES completo e dois anos para ES incompleto, não garantindo uma maior precisão na estimativa. Esse ponto é particularmente importante no caso brasileiro, já que há relatos históricos de altas taxas de repetência nas séries iniciais (RIBEIRO, 1991). Em nível internacional, esse grau de detalhamento é de difícil implementação por conta da base extensa de países.

No Gráfico 7 constam os *benchmarks*: os anos médios de escolaridade estimados a partir dos censos (1960 – 2000) e dados da PNAD Contínua (2012 – 2015). No Gráfico 8, pode ser analisada a distribuição educacional por censo, com o percentual de concluintes e não concluintes por nível de ensino. Destaca-se que a distribuição educacional é calculada com base nos anos de escolaridade de cada indivíduo, podendo não resultar no valor exato da distribuição

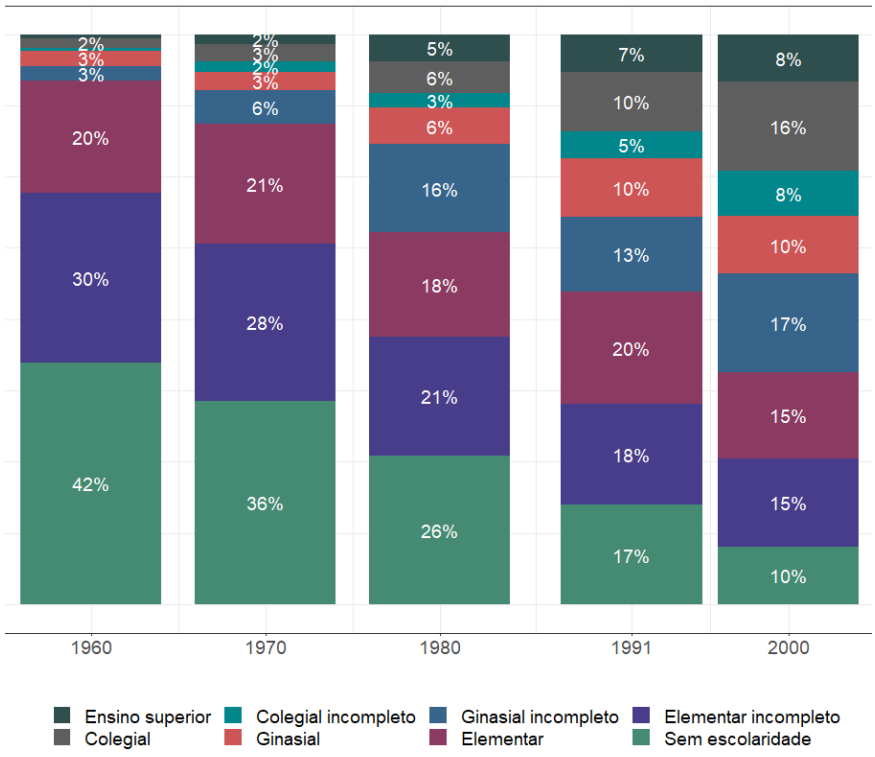
educacional analisando a pergunta do curso concluído/frequentado. Foram excluídos da base indivíduos que não informaram a última série de ensino concluída.

Gráfico 7 - Estimativas de *benchmarks* de anos médios de escolaridade no Brasil, população de 15 a 64 anos



Fonte: Elaboração própria baseada nos microdados do Censo do IBGE (1960 – 2000) e dados da PNAD Contínua (2012 – 2015).

Gráfico 8 - Distribuição educacional no Brasil, população de 15 a 64 anos, microdados



Fonte: Elaboração própria baseada nos microdados do Censo do IBGE (1960 – 2000).

3.3 ESTIMATIVA: 1925-1950

Devido à escassez de dados, pesquisas nacionais, nesse período, para a definição dos *benchmarks*, aplicou-se a metodologia de acompanhamento de coortes dos microdados de 1960 até 1925, conforme uma adaptação de Lutz et al. (2007). Para a mensuração, parte-se de algumas suposições:

a) a imigração não interfere no padrão educacional: não são consideradas diferenças educacionais entre imigrantes e locais;
b) a adoção dessa metodologia pressupõe distinção de mortalidade por faixa etária, porém é considerada a mortalidade homogênea entre grupos por nível de ensino. Não há a atuação de diferenciais de mortalidade, ou seja, nível escolar e taxa de mortalidade são variáveis independentes. Essa suposição é vista como aceitável para períodos mais antigos (COHEN; SOTO, 2007);
c) distribuição educacional constante fora da fase escolar (25 anos ou mais);
d) duração por nível de ensino de 1925 a 1960 é equivalente às durações estimadas em 1960 ("Dur"), ou seja, assume-se que o percentual de pessoas com ensino incompleto é o mesmo de 1925 a 1960;
e) taxa bruta constante de 1910 a 1933 para o EM, e o ES. Para o EF foi considerado o número de matriculados da 1ª a 4ª série para o cálculo da taxa bruta (1910-1933).

A metodologia adotada é semelhante às utilizadas por KC et al. (2010) e Lutz et al. (2007). Inicialmente, toma-se a população de 25 ou mais anos de idade do censo de 1960 para projetar as coortes no passado. Por exemplo, a coorte de 40 anos em 1960 com uma determinada escolaridade terá o mesmo nível educacional da coorte de 30 anos de idade em 1950. Essa metodologia foi aplicada em todas as coortes em fase não escolar, possibilitando uma estimação populacional para trás até 1925. Assumindo que somente a população em fase escolar altera a sua distribuição educacional, a escolaridade da população de 15 a 25 anos foi mensurada por meio da taxa bruta de matrícula, levando em consideração um período de defasagem. A título de exemplo, a coorte de 20 a 24 anos em 1950 cursou o EF em 1935 (defasagem de 15 anos), cursou o EM em 1940 (defasagem de 10 anos) e está na faixa etária para cursar o ES em 1950 (sem defasagem). Neste caso, são consideradas defasagens para elencar qual a taxa bruta de matrícula do período que equivale à distribuição educacional dessa coorte. O mesmo

procedimento é aplicado para a faixa etária de 15 a 19 anos que em 1950 estavam cursando o EM e que frequentavam o EF em 1945 (defasagem de 10 anos).

Para estimar a população de 1950 são considerados dados de matrícula (KANG et al., 2021) para as coortes em fase escolar (15 a 24 anos) e o censo de 1960 para a população de 25 a 64 anos. A partir dos resultados da distribuição educacional da população de 15 a 19 anos, 20 a 24 anos e 25 a 64 anos, é aplicada uma média ponderada para determinar a distribuição final. A fim de converter esses dados em anos médios de escolaridade, multiplica-se a distribuição educacional (EF, EM, ES) pela respectiva duração do nível de ensino, incluindo não concluintes, de 1960 (variável constante).

Além das vantagens metodológicas previamente citadas, essa técnica prescinde de dados de repetência, o que pode ser positivo, considerando que no Brasil não há dados antigos confiáveis de repetência e evasão devido a erros nos dados oficiais (RIBEIRO, 1991). Tendo em vista também a baixa disponibilidade de dados de matrícula líquida, utilizou-se a taxa bruta de matrícula para a estimação da distribuição educacional da população em faixa etária escolar. Essa escolha tende a superestimar os dados, assim como a adoção da duração por nível de ensino de 1960 como um parâmetro constante para a conversão da distribuição educacional em anos médios de escolaridade. Porém, desconsiderando que não haja diferencial de mortalidade para o período, há uma tendência de subestimação das variáveis, por adotar como referência a população de 25 ou mais anos de idade de 1960 para estimar a população de 25 a 64 anos de 1925 a 1950.

Testou-se também o método do inventário perpétuo (PIM), porém este só demonstrou ser efetivo para intervalos de séries temporais mais recentes. Trata-se de uma metodologia de difícil replicação para períodos mais antigos, pois exige majoritariamente uma série histórica extensa de matrículas.

3.4 PERÍODOS INTERCENSITÁRIOS, 1950 - 2012

Entre 1950 e 2012, com a intenção de criar uma série de anos médios anualizada, optou-se por aplicar a metodologia modificada de estimação baseada no PIM de Foldvári e Van Leeuwen (2009), uma versão modificada de Barro e Lee (1993, 2001). A técnica adotada pelos últimos apresenta a desvantagem de desconsiderar mortalidade diferencial e evasão, podendo resultar em algum viés nas estimativas em períodos mais recentes (FOLDVÁRI; VAN LEEWEEN, 2014). Desconsiderar o diferencial de mortalidade tende a subestimar os anos médios da população mais velha e o oposto ocorre ao se desconsiderar a evasão educacional.

Se for assumido que em uma década o percentual de desistência e mortalidade diferencial é constante, pode-se remover o viés através da média das *forward* (Equação 2) e *backward* (Equação 3) *estimations* de *benchmarks* equidistantes. Abaixo constam as equações extraídas do artigo de Foldvári e Van Leeuwen (2009):

$$h_{0,t} = H_{0,t}/L_t = h_{0,t-i}[1 - (L15_t.\, i/5.\, L_t)] + (L15_t.\, i/5.\, L_t).\,(1 - PRI_{t-i})$$
$$h_{1,t} = H_{1,t}/L_t = h_{1,t-i}[1 - (L15_t.\, i/5.\, L_t)] + (L15_t.\, i/5.\, L_t).\,(PRI_{t-i} - SEC_t)$$
$$h_{2,t} = H_{2,t}/L_t = h_{2,t-i}[1 - (L15_t.\, i/5.\, L_t)] + (L15_t.\, i/5.\, L_t).\, SEC_t - (L20_t.\, i/5.\, L_t).\, HIGH_t$$
$$h_{3,t} = H_{3,t}/L_t = h_{3,t-i}[1 - (L15_t.\, i/5.\, L_t)] + (L20_t.\, i/5.\, L_t).\, HIGH_t \quad (2)$$

$$h_{0,t-i} = \left(h_{0,t} - (L15_t.\, i/5.\, L_t).\,(1 - PRI_{t-i})\right)/[1 - (L15_t.\, i/5.\, L_t)]$$
$$h_{1,t-i} = \left(h_{1,t} - (L15_t.\, i/5.\, L_t).\,(PRI_{t-i} - SEC_t)\right)/[1 - (L15_t.\, i/5.\, L_t)]$$
$$h_{2,t-i} = \left(h_{2,t} - (L15_t.\, i/5.\, L_t).\, SEC_t + (L20_i.\, i/5.\, L_t).\, HIGH_t\right)/[1 - (L15_t.\, i/5.\, L_t)]$$
$$h_{3,t-i} = \left(h_{3,t} - (L20_t.\, i/5.\, L_t).\, HIGH_t\right)/[1 - (L15_t.\, i/5.\, L_t)] \quad (3)$$

Para ambos os conjuntos de equações, h é o nível educacional (h = 0, população sem escolaridade, h = 1, EF, h = 2, EM, h = 3, ES); H corresponde à população por nível de ensino; i é o número de anos entre o ano a ser estimado e o *benchmark*; L = população de 15 anos a 64 anos; L15 = população de 15 a 19; L20 = população de 20 a 24; PRI = taxa bruta de matrícula do EF; SEC = taxa bruta de matrícula do EM; HIGH = taxa bruta do ES.

A fim de calcular os anos médios de escolaridade de 1965, por exemplo, foi utilizado como base os censos de 1960 e 1970, aplicando a metodologia *forward* para o ano de 1960 e *backward* para o ano de 1970 e, em seguida, calculada a média dos resultados. Assim, obteve-se o percentual da população no ano de 1965 que não tinha escolaridade, que frequentou/frequentava o EF, o EM e o ES no período. A partir da obtenção da distribuição educacional de 1965, foi possível calcular a distribuição educacional de 1963 e assim por diante.

Uma adaptação adotada neste trabalho foi a determinação da duração de cada nível educacional, considerando não concluintes, para a conversão da distribuição em anos médios de escolaridade. A metodologia de Foldvári e Van Leeuwen (2009) não possibilita fazer distinção entre a população que completou ou não o seu nível de ensino. Para compensar essa falta de informação, a duração de cada nível educacional foi estimada com base nas durações dos censos utilizados como *benchmark*. Para o cálculo dos anos médios de escolaridade de 1965, cada distribuição educacional foi multiplicada por uma duração de nível de ensino correspondendo à média ponderada da duração dos censos de 1960 e 1970. Se a duração do nível educacional da população do EF de 1970 foi de 6 anos e de 1960 foi 5 anos, aplicou-se a

média ponderada das duas durações (com peso de 0,5, considerando que 1965 é equidistante a 1970 e 1960; se fosse 1964, haveria um peso maior no ano de 1960).

A partir dessa metodologia, prioriza-se que os pontos entre os *benchmarks*, posteriores e anteriores ao ano estimado, sejam equidistantes, pois isso possibilita que os vieses entre a estimação *backward* e *forward* se equalizem. Entretanto, nem sempre foi possível realizar estimativas equidistantes devido à distância dos *benchmarks*. Portanto, optou-se por dar um peso maior no *backward* nesses casos, superestimando o cálculo (exemplo: para estimar 1952, foram utilizados como *benchmarks* 1950 e 1955, distância de duas e três unidades, respectivamente). Contudo, essa diferença entre as distâncias nunca foi maior do que um.

3.5 ESTIMAÇÃO POR SUBGRUPOS

Também se estimaram os anos médios de estudo por gênero (homens e mulheres), cor/raça (amarelos, brancos, pardos e pretos) e por estado. No caso das estimativas por gênero, houve uma adaptação na metodologia para calcular a distribuição e anos médios de escolaridade de homens e mulheres. Inicialmente foi estimada a população de 1925 a 2015 e mantida a mesma taxa bruta, da população total. A partir desses dados foi possível calcular a escolaridade da população em faixa etária escolar de 1925 a 1950 e aplicar o método de Foldvári e Van Leeuwen (2009). Isto é, não foi utilizada taxa bruta específica por gênero, pois essas informações são mais limitadas para anos mais antigos. Neste caso, de 1925 a 1950 a escolaridade da população em fase escolar tende a estar subestimada para homens e superestimada para mulheres em períodos mais antigos, considerando que havia uma maior tendência de homens estarem matriculados na escola (a partir de uma análise dos dados de matrícula de 1933 a 1970, houve indícios de que a utilização de taxa bruta de matrícula similar para homens e mulheres é uma suposição válida para o período). O cálculo por cor/raça foi implementado de forma similar, com a exceção de 1970, pois não constam informações de cor/raça. Portanto, nesse caso foi utilizada a metodologia de Foldvári e Van Leeuwen (2009) entre o período de 1960 a 1980. Destaca-se que de 1960 para trás foi utilizada uma estimativa (com base no censo de 1960) da taxa bruta de matrícula por cor/raça do ensino fundamental; nas demais séries, optou-se por utilizar dados de frequência bruta (não interferindo muito no resultado devido à proporção populacional que cursou o ensino na época). Nos anos de 1960 a 2015 foi utilizada a mesma taxa bruta de matrícula da população total, pois foi detectado que o modelo não apresenta muita sensibilidade à variação de matrícula a partir desse período.

De maneira semelhante, o cálculo por estado também sofreu as mesmas modificações que as estimações de homens e mulheres. Nesse caso, foi utilizada a taxa bruta da população total e os estados de interesse foram selecionados de acordo com a classificação das Unidades da Federação em 1940. Também foi feita a estimativa agregada por macrorregiões (Norte, Nordeste, Centro-Oeste, Sudeste e Sul), conforme a classificação do censo de 2010.

3.6 ESTIMATIVA POPULACIONAL E DADOS DE MATRÍCULA

A estimativa populacional de 1950 a 2015 por estado (faixa etária de 5 a 14, 15 a 19, 20 a 24, 60 a 64 anos) foi feita através da interpolação com base na função *spline* cúbica dos dados populacionais coletados nos microdados (censo e PNAD Contínua) e dados do relatório de recenseamento de 1950 disponibilizados no site do IBGE. A partir desses dados foi calculada a população total do período de 1950 a 2015, por meio do somatório das populações estaduais. A população de 1925 a 1950 foi obtida a partir dos dados do Sistema IBGE de Recuperação Automática (SIDRA) interpolados. A estimativa da população feminina também foi calculada com base nos microdados até 1960. Nos demais anos censitários, estimou-se a população feminina a partir da multiplicação da proporção de mulheres sobre a população total disponibilizada por ano censitário no relatório de recenseamento de 1950 (as proporções foram equivalentes para todas as faixas etárias). A população masculina foi calculada por resíduo, sendo o mesmo procedimento adotado para as diferentes raças.

Para o cálculo da taxa bruta de matrícula, foram utilizados os dados de matrícula de Kang et al. (2021). A taxa bruta de matrícula correspondeu à divisão do número de matriculados por nível de ensino, de cada período: pela população em idade escolar de 5 a 14 anos para o EF; 15 a 19 anos para o EM; e 20 a 24 para o ES. Para as matrículas do ensino fundamental (um a quatro anos de estudo), são disponibilizados dados de 1872 até 2013, com alguns períodos faltantes que foram preenchidos por interpolação. Em relação ao nível médio e superior, estavam disponíveis dados a partir de 1933. Como eram necessários dados a partir de 1900, considerou-se que a taxa bruta de matrícula entre 1900 e 1933 foi constante (igual a de 1933). Para o ensino fundamental de 1900 a 1933, foi calculada a taxa bruta com base nos dados de matrícula disponíveis do antigo ensino primário (um a quatro anos de estudo), portanto, essa série tende a estar subestimada.

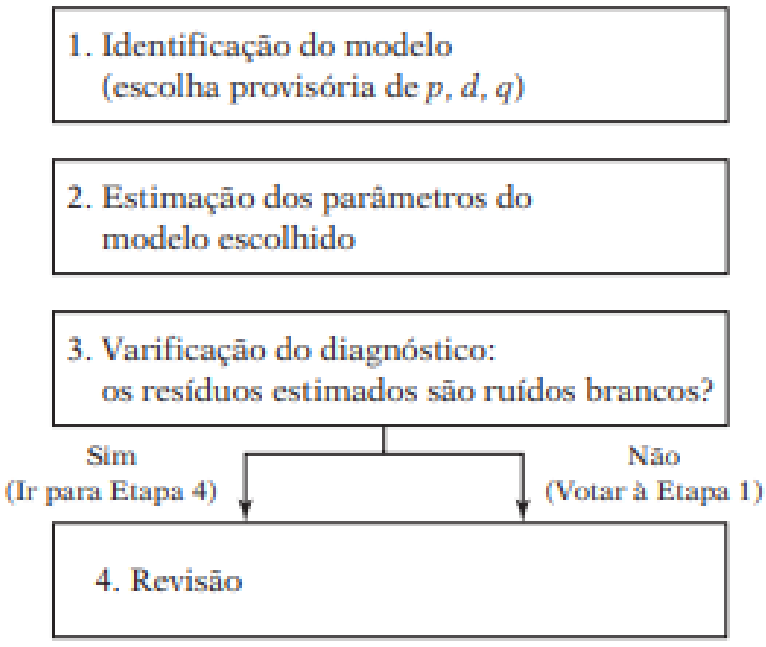
3.7 MODELO SOFTWARE R

Através do uso da plataforma *GitHub*, é possível criar um repositório no qual todos os códigos utilizados para o desenvolvimento da base de dados de anos médios de escolaridade ficam disponíveis para acesso. No Apêndice C está o detalhamento do repositório disponibilizado no link: https://github.com/juliarwalter/pesquisa-anos-medios-de-escolaridade e https://github.com/juliarwalter/pesquisa-anos-medios-de-escolaridade-complemento.

3.8 MODELO ARIMA

A partir da base de dados de anos médios de escolaridade da população de 15 a 64 de 1925 a 2015, com base na metodologia *Box-Jekins,* modelo ARIMA univariado, é possível realizar uma previsão da série de 2016 a 2025. Conforme descrito por Gujarati e Porter (2011), são 4 etapas para a implementação do modelo, sendo o objetivo desta modelagem converter os resíduos em ruído branco:

Figura 1 - Ilustração da metodologia *Box-Jekins*



Fonte: Gujarati e Porter (2011).

A partir do teste Dickey-Fuller aumentado (ADF) é analisada a estacionariedade da série temporal e verificada a necessidade de diferenciação do modelo. Após a determinação da ordem

de integração (“d”) e, posteriormente, do número de defasagens da variável (“p”), da ordem do modelo de média móvel (“q”), e feita a estimação dos parâmetros, é verificada qual é a modelagem mais adequada para a realização da previsão (significância dos coeficientes do modelo, o teste de *Ljung-Box*, o AIC, o *Hanna-n-Quinn* e o Critério de Schwarz). Posteriormente é analisada a normalidade e homocedasticidade dos resíduos e feita a previsão.

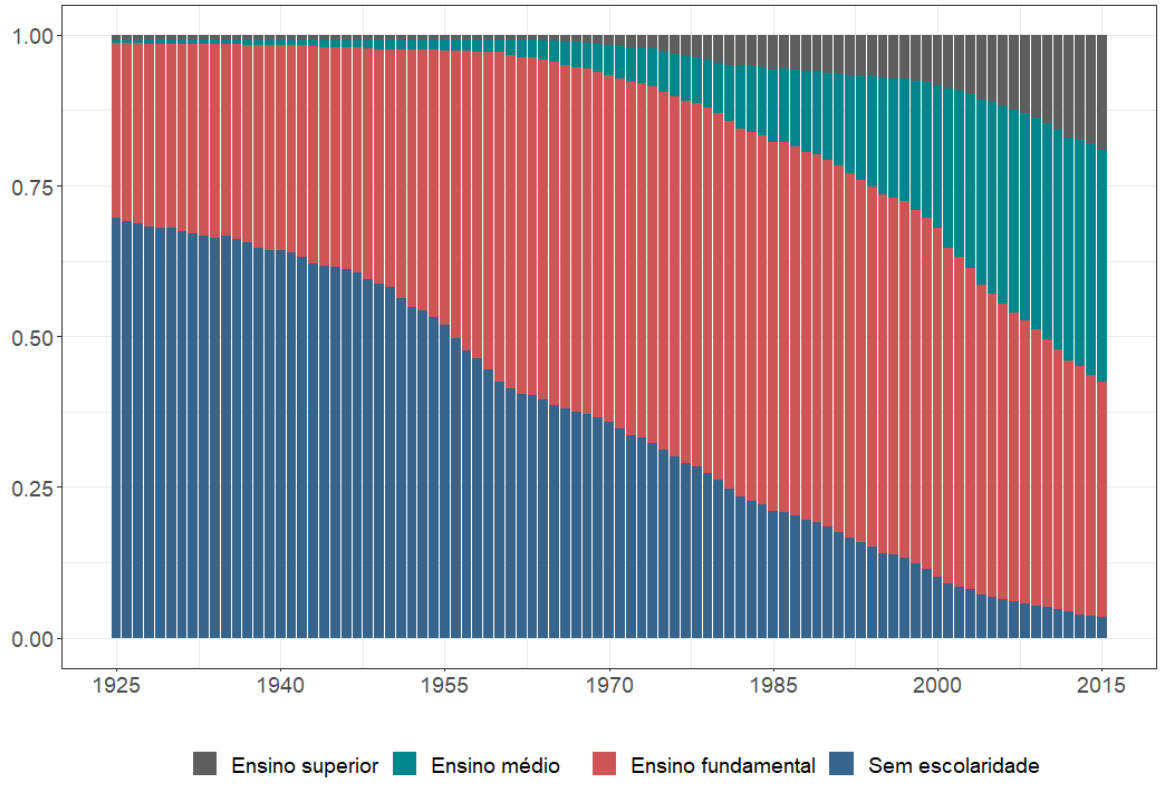
4 RESULTADO

Nesta seção é disponibilizado os resultados encontrados na pesquisa e feita uma breve análise e comparação com estudos internacionais que estimam a escolaridade brasileira.

4.1 ANOS MÉDIOS DE ESCOLARIDADE: POPULAÇÃO TOTAL

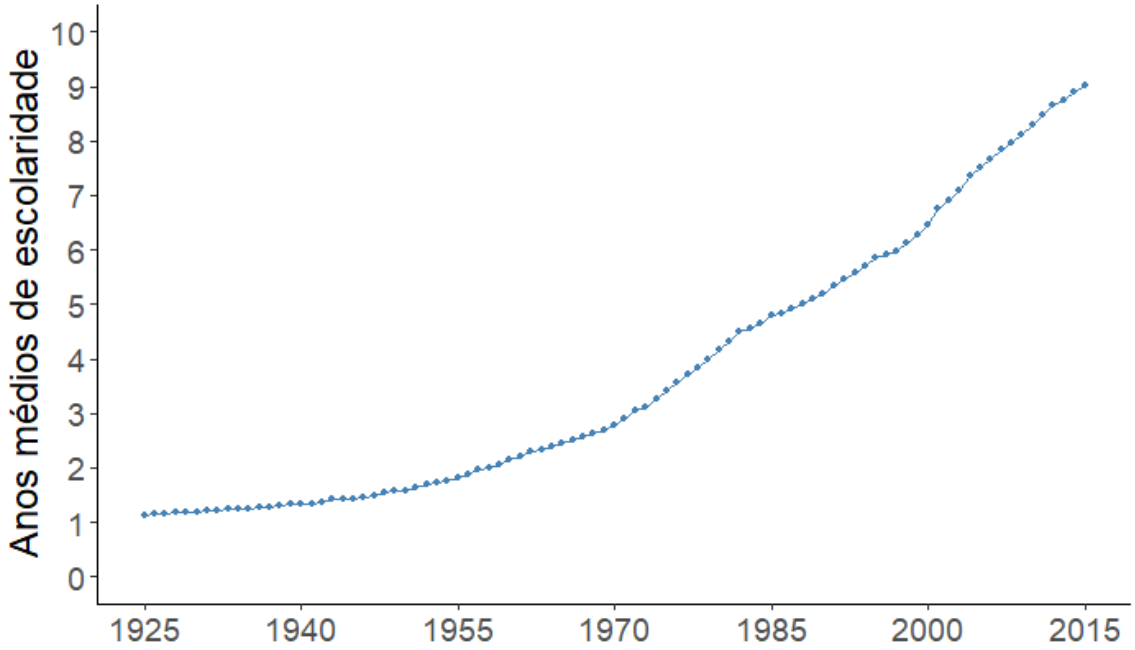
A metodologia utilizada resultou em uma série histórica de distribuição educacional anualizada para a população total, representada no Gráfico 9. Observa-se uma consistência na distribuição educacional ao longo dos anos. No mesmo gráfico, também se percebe um aumento da proporção de pessoas que passaram pelo EF (completo e incompleto) no início da década de 1950. A partir da distribuição educacional, foi possível determinar os anos médios de escolaridade da população após aplicar a Equação 1 (Gráfico 10), descrita no primeiro capítulo. Portanto, a partir das metodologias empregadas foi possível garantir uma consistência entre as duas variáveis (anos médios de escolaridade e distribuição educacional). Destaca-se que a distribuição educacional inclui também a população que não completou o nível (por exemplo, no resultado de EF, estão incluídos aqueles que têm EF incompleto).

Gráfico 9 - Distribuição educacional no Brasil, população de 15 a 64 anos



Fonte: Elaboração própria.
Nota: Distribuição em escala de 0 a 1. As estatísticas incluem também aqueles que não completaram o nível.

Gráfico 10: Anos médios de escolaridade no Brasil, população de 15 a 64 anos

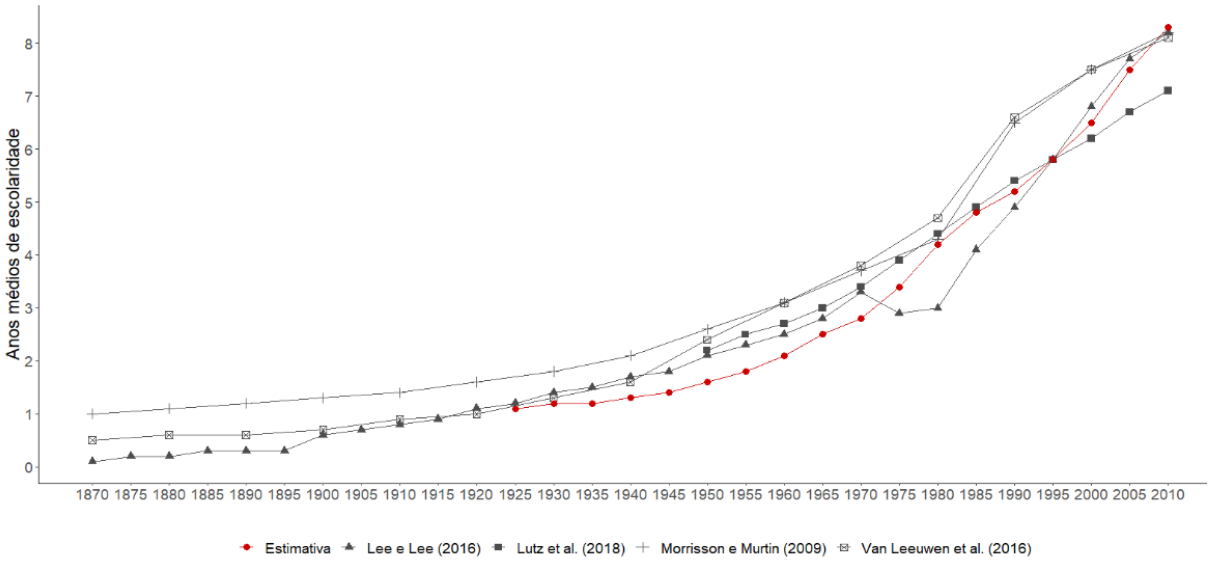


Fonte: Elaboração própria.

4.2 COMPARAÇÃO ENTRE ESTUDOS

O Gráfico 11 mostra que as estimativas deste trabalho divergem consideravelmente de Lee e Lee (2016) e Morrisson e Murtin (2009), levando em conta que ambos estimam anos de estudo para a população de 15 a 64 anos de idade. Esses trabalhos superestimam os valores, se comparados às estimativas calculadas nesta pesquisa. Van Leeuwen e Van Leeuwen-Li (2014) e Lutz et al. (2018) também tendem a superestimar os anos médios de escolaridade, apesar de utilizarem um intervalo populacional diferente (população de 15 ou mais anos de idade). Se fosse adotada uma faixa comparável, a série de anos médios de escolaridade desses estudos possivelmente apresentaria resultados ainda mais elevados.

Gráfico 11: Anos médios de escolaridade no Brasil, população de 15 a 64 anos, de 1870 a 2010, comparativo entre estudos



Fonte: Elaboração própria baseada nos dados de Lee e Lee (2016), Morrisson e Murtin (2009), Van Leeuwen et al. (2016) e Lutz et al. (2018).

Uma das justificativas para essa superestimação é que os estudos de Lutz et al. (2018) e Lee e Lee (2016) convertem distribuição educacional em anos médios de escolaridade a partir da Equação 1, porém não com uma duração exata que englobe a população com ensino incompleto. Van Leeuwen e Van Leeuwen-Li (2014) e Morrisson e Murtin (2009) apresentam problemas similares: esses trabalhos tendem a superestimar de 1960 para frente por utilizarem o estudo de Cohen e Soto (2007) como referência, que apresenta poucos *benchmarks* em suas estimativas (os autores não informam se foi utilizada alguma duração para conversão da distribuição escolar em anos médios de escolaridade). De 1960 para trás há uma possível inconsistência desses estudos devido à aplicação da metodologia PIM, afinal tomam-se taxas de matrícula líquida não confiáveis, em virtude da subestimação de dados oficiais de repetência e evasão do Brasil (RIBEIRO, 1991).

O coeficiente de determinação entre as estimativas dos estudos de Lee e Lee (2016), Morrisson e Murtin (2009), Van Leeuwen et al. (2016), Lutz et al. (2018) e às deste trabalho (R^2 da estimativa calculada no trabalho como função das outras estimativas) mostram que Lutz et al. (2018) é a série que mais se ajusta às nossas estimativas. Portanto, considerando apenas os dados brasileiros, Lutz et al. (2018) apresentam as estimativas mais fidedignas dentre os estudos internacionais, apesar das suas imprecisões em estimativas mais atuais (a partir de 2010) e de considerar um intervalo populacional diferente (população de 15 ou mais anos de idade).

Tabela 2 - Anos de escolaridade estimado x outros estudos (esta pesquisa = 1,00), 1920-2000

ano	Lee e Lee (2016)	Morrisson e Murtin (2009)	Van Leeuwen et al. (2016)	Lutz et al. (2018)
1950	1.33	1.61	1.48	1.39
1960	1.18	1.43	1.44	1.27
1970	1.18	1.32	1.35	1.23
1980	0.73	1.03	1.12	1.06
1990	0.95	1.26	1.28	1.04
2000	1.05	1.16	1.17	0.97
2010	0.99	0.99	0.98	0.85
R-Squared	0.94	0.96	0.96	0.98

Fonte: Elaboração própria baseada nos dados de Lee e Lee (2016), Morrisson e Murtin (2009), Van Leeuwen et al. (2016) e Lutz et al. (2018).

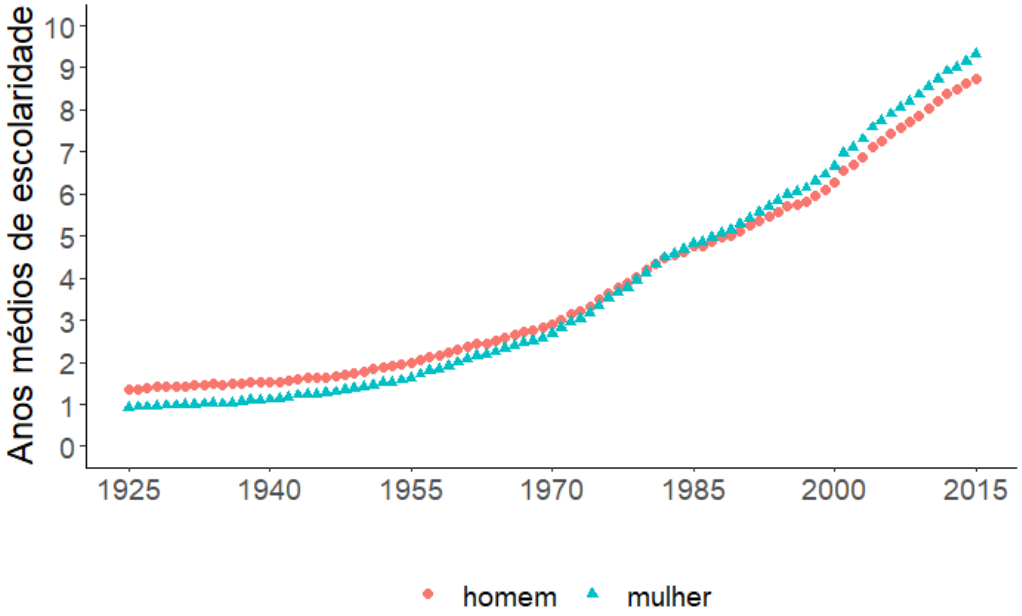
4.3 ESCOLARIDADE: GÊNERO E COR

O Gráfico 12 apresenta a discrepância educacional entre homens e mulheres. O Censo de 1991 já havia revelado que as mulheres tinham passado a ter mais escolaridade em média do que os homens (MELO; THOMÉ, 2018). De acordo com as estimativas anuais de anos médios de estudo calculadas no trabalho, as mulheres ultrapassam os homens em 1983. Esse marco é próximo ao ano de 1979, quando as mulheres passaram a usufruir dos mesmos direitos e deveres dos homens a partir da Assembleia Geral das Nações Unidas daquele ano (THE UNITED NATIONS, 1988). Um possível fator que pode estar associado com essa melhora educacional é a taxa de fecundidade brasileira, que passou de 4,35 filhos por mulher em 1980 para 2,89 em 1991. Se analisados os censos de 1960 a 2000, a década de 1980 foi o período de maior queda na taxa de fecundidade. É difícil saber qual o sentido da causalidade: tanto a queda no número de filhos pode ter possibilitado às mulheres uma maior inserção no mercado de trabalho e uma melhora no seu nível educacional, como o oposto pode ter ocorrido.

Lee e Lee (2016) também apresentam dados de escolaridade por homens e por mulheres para a população de 15 a 64 anos. Conforme as estimativas dos autores, as mulheres ultrapassam os homens em algum momento entre 1985 e 1990. Além disso, a diferença entre homens e mulheres no estudo não é tão proeminente quanto a das nossas estimativas. Essas mesmas análises podem ser observadas nas estimativas de Lutz et al. (2018), nas quais mulheres

ultrapassam a escolaridade masculina a partir de 1980 e apresentam pouca diferença de escolaridade se comparado aos homens.

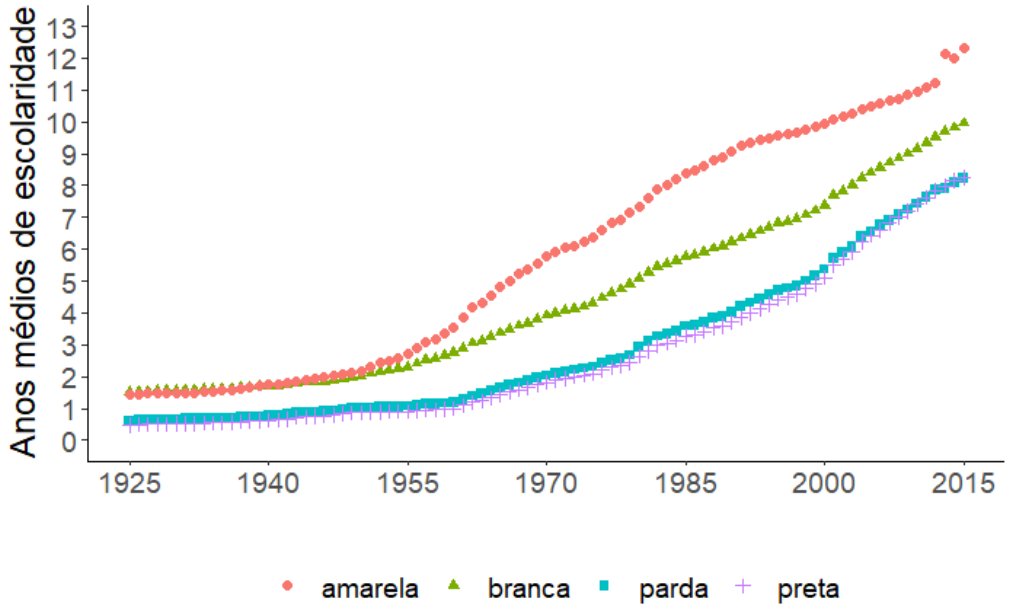
Gráfico 12: Anos médios de escolaridade de homens e mulheres no Brasil, população de 15 a 64 anos



Fonte: Elaboração própria.

No que diz respeito à análise educacional de diferentes raças/cores a diferença educacional é mais perceptível, sendo representada em três diferentes grupos: amarelos, brancos, pardos/pretos. A cor amarela (associada à cultura asiática) apresenta consideravelmente uma maior escolaridade, sendo um referencial de ensino. No ano de 2015, aproximadamente 50% da população cursava ou havia concluído o ensino superior. Porém essa raça é pouco representativa sobre o total populacional brasileiro. A população parda e preta, muito representativa na raça brasileira, possuem o menor grau de escolaridade. Em uma análise gráfica, não é percebida uma potencial convergência da educação entre cores/raças (Gráfico 13).

Gráfico 13: Anos médios de escolaridade entre cores/raças no Brasil, população de 15 a 64 anos

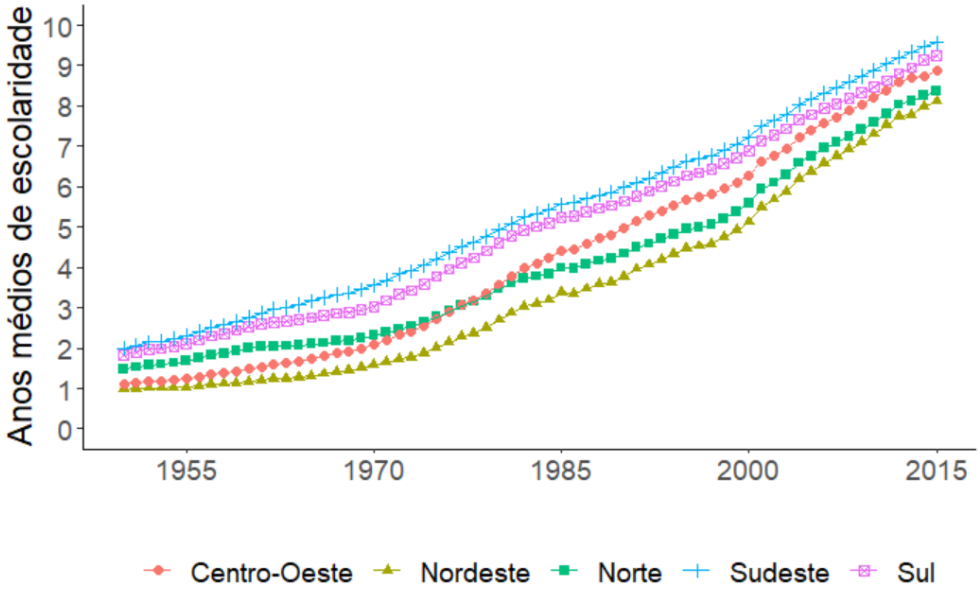


Fonte: Elaboração própria.
Nota: No período de 2013 na população amarela há uma quebra advinda dos dados da PNAD Contínua.

4.4 ESCOLARIDADE: ESTADOS E REGIÕES

A série histórica de anos médios de escolaridade por macrorregião revela uma certa estabilidade da escolaridade ao longo do tempo. A única mudança mais abrupta é a evolução da Região Centro-Oeste, que ultrapassou a Região Norte em termos de anos médios de estudo em 1977. É possível que esse rápido progresso educacional tenha relação com a criação de Brasília em 1962, apesar de o Distrito Federal não estar computado nessa análise com o intuito de evitar distorções. Além das externalidades da nova capital, incentivos governamentais levaram à ocupação do Centro-Oeste e à expansão da fronteira agrícola a partir da década de 1970. Considerando que grande parte dessa migração veio do Sul e do Sudeste, população mais escolarizada da época, esse fenômeno provavelmente elevou o nível médio de escolaridade do Centro-Oeste em relação ao Norte. Ademais, nota-se um elevado nível relativo de escolaridade nas regiões mais desenvolvidas, Sudeste e Sul, desde o início da série por regiões e estados em 1950. Esse diferencial se mantém até o período mais recente.

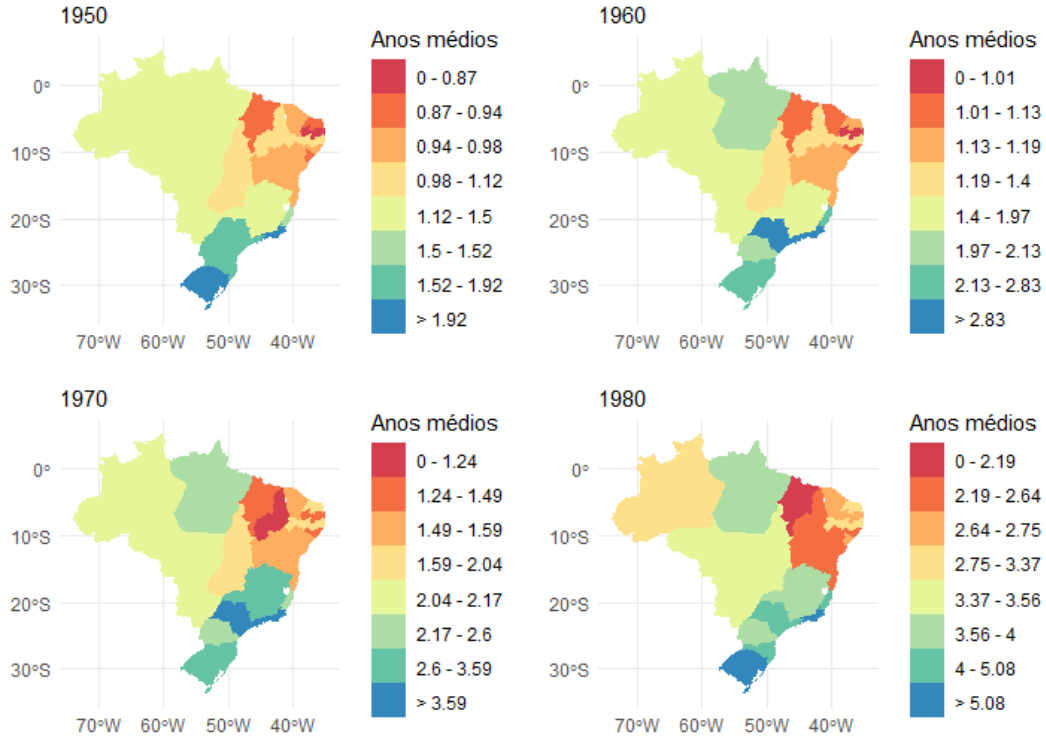
Gráfico 14: Anos médios de escolaridade por macrorregião brasileira, população 15 a 64 anos



Fonte: Elaboração própria.
Nota: Não foi contabilizado o Distrito Federal e Tocantins pertence ao Centro-Oeste. Eixo x em escala de 1950 a 2015.

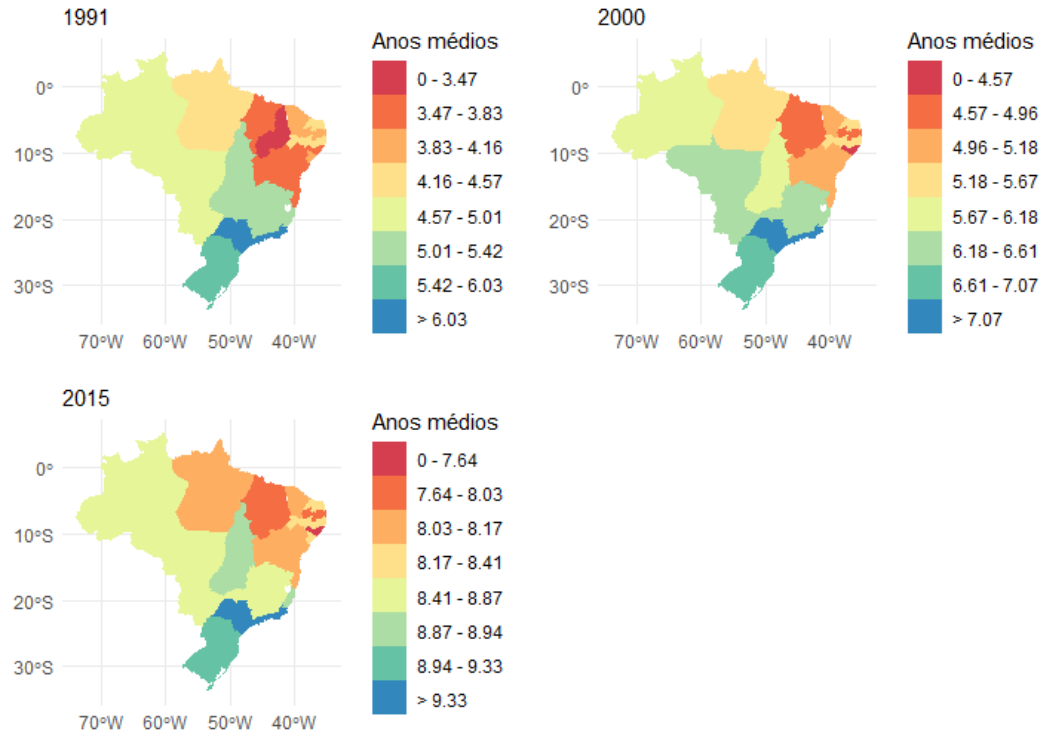
Na ótica da escolaridade por Unidade da Federação (Gráficos 15 e 16), de 1950 a 1980, percebe-se uma grande dispersão entre o estado com maior escolaridade, Rio de Janeiro (RJ), e o segundo colocado, São Paulo (SP). O RJ só foi superado em 2002 por SP. Um caso emblemático é o Rio Grande do Sul (RS), que estava parelho com SP de 1950 a 1980, mas passa a apresentar menor aceleração a partir de 1982. Em 2015, o RS havia atingido 9,2 anos médios de escolaridade (quinto estado mais instruído), enquanto SP já tinha chegado a 10,0 anos. À exceção de alguns casos, não houve expressiva alteração na distribuição de anos de escolaridade entre estados ao longo da série, como retratam os Gráfico 15 e 16.

Gráfico 15: Anos médios de escolaridade por estado brasileiro, população 15 a 64 anos (1950 a 1980)



Fonte: Elaboração própria.
Nota: Mapa em percentis: 15, 30, 45, 60, 75, 90, 100.

Gráfico 16: Anos médios de escolaridade por estado brasileiro, população 15 a 64 anos (1991 a 2015)



Fonte: Elaboração própria.
Nota: Mapa em percentis: 15, 30, 45, 60, 75, 90, 100.

4.6 ESTIMAÇÃO DOS ANOS MÉDIOS DE 2015 ATÉ 2025

Inicialmente foi feito o teste do Dickey-Fuller aumentado (ADF) para análise da estacionariedade da série temporal. Conforme observado graficamente e com base no teste, há a presença de tendência estocástica no gráfico de anos médios de escolaridade, portanto a série não é estacionária, possui raiz unitária. De acordo com o teste, a série necessitou de 2 diferenciações para se tornar estacionária, considerando um p-valor de 0,05. Baseado na metolodogia *Box-Jekings* o modelo que mais se ajustou à série temporal foi um ARIMA (0,2,1), os testes realizados estão no Apêndice D.

De acordo com esse modelo os anos médios de escolaridade de 2016 a 2025 tendem a apresentar os respectivos valores:

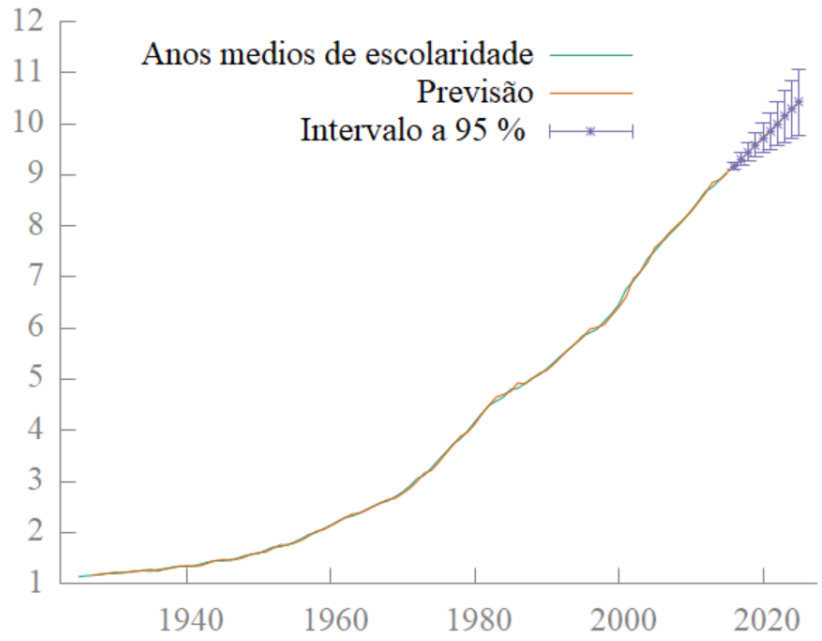
Tabela 3: Previsão de anos médios de escolaridade 2015 a 2025

Ano	Estimativa	Erro Padrão	Intervalo de Confiança
2016	9,17	0,037	9,10 - 9,24
2017	9,31	0,062	9,19 - 9,43
2018	9,45	0,089	9,27 - 9,62
2019	9,58	0,118	9,35 - 9,82
2020	9,72	0,149	9,43 – 10,02
2021	9,86	0,183	9,50 - 10,22
2022	10,00	0,218	9,57 - 10,43
2023	10,14	0,255	9,64 - 10,64
2024	10,28	0,294	9,70 - 10,85
2025	10,42	0,335	9,76 – 11,07

Fonte: Elaboração própria.

A projeção resultou na série de anos médios de escolaridade abaixo:

Gráfico 17 - Projeção dos anos médios de escolaridade para o Brasil, 2015-2025



Fonte: Elaboração Própria.

Nota: Eixo y iniciando no 1 e o eixo x em 1925.

Destaca-se que com o surgimento da COVID-19 e o agravamento da pandemia, segundo o Banco Mundial (2021) há uma tendência de possível retrocesso educacional, pois a pandemia é o maior choque educacional já registrado na história. Estimativas mostram que o fechamento de escolas por 10 meses pode resultar em uma perda de cerca de 1,3 ano de escolaridade e 71% dos estudantes ficariam abaixo do mínimo de proficiência requerido no PISA, na América Latina e Caribe. O impacto de longo prazo para 10 meses de fechamento é uma perda de ganhos de 1,7 trilhão de dólares. Estendendo mais 3 meses de fechamento a perda é cerca de 1,7 ano de escolaridade. Portanto há um indicativo que a projeção da escolaridade brasileira não siga a tendência da Tabela 3 (nota-se que o impacto na educação será sentido mais ao longo prazo, considerando que a série histórica de anos médios de escolaridade representa a população em idade ativa).

5 CONCLUSÃO

A principal contribuição do trabalho foi disponibilizar uma base de dados anualizada de escolaridade da população brasileira de 15 a 64 anos, bem como os dados desagregados em gênero, cor/raça e regiões (com os códigos de implementação do projeto disponibilizados no *Github*). Os dados são subsídios para futuras pesquisas na análise de desenvolvimento econômico do país, e através deles é possível compreender de forma mais clara a realidade educacional brasileira e análise de potenciais atrasos no ensino.

A partir do próximo censo é possível realizar uma atualização da base de dados, sendo prevista, para os próximos anos, uma provável piora educacional no país devido às consequências advindas da pandemia. Destaca-se que a não padronização dos censos dificulta a realização das estimativas. Ademais, a maneira de aferição educacional do censo de 2010 acabou inviabilizando uma análise mais aprofundada para o período.

Através desse trabalho é possível perceber que os dados de anos médios de escolaridade da população brasileira em âmbito internacional devem ser analisados com cautela, pois dependendo da fonte de dados utilizada, podem apresentar incongruências significativas nas estimativas, como o caso dos estudos de Barro e Lee com a utilização de fontes da UNESCO. Entre as pesquisas no âmbito internacional, a que mais se aproximou do comportamento esperado para o cenário brasileiro foi a pesquisa de Lutz et al. (2018), entretanto destaca-se que devem ser realizadas análises mais aprofundadas sobre essas estimativas. Por exemplo, verificar se para os demais países os anos médios de escolaridade estimados são representativos para cada região.

Ao observar a evolução educacional brasileira percebe-se um ganho educacional das mulheres, que ultrapassam os homens em 1983. Porém, quando analisada a escolaridade na ótica de cor/raça, ainda é percebida uma elevada desigualdade educacional. O mesmo cenário é observado na análise por regiões, uma certa manutenção da desigualdade, porém há uma maior variação educacional das macrorregiões Nordeste e Centro-Oeste.

Ressalta-se que não há discordância na literatura da importância do papel educacional para o desenvolvimento do país. Esse estudo confirma que é imprescindível a adoção de medidas, políticas públicas e privadas, para uma educação melhor e mais igualitária.

REFERÊNCIAS

BANCO MUNDIAL. Agindo agora para proteger o capital humano de nossas crianças: Os custos e a resposta ao impacto de COVID-19 no setor de educação na América Latina e Caribe. Washington D.C.: **World Bank**, 2021. Disponível em: https://www.worldbank.org/pt/news/press-release/2021/03/17/hacer-frente-a-la-crisis-educativa-en-america-latina-y-el-caribe. Acesso em: 24 abr. 2021.

BARBOSA, R. J. **Instruções para o uso dos bancos de microdados das amostras dos Censos Demográficos Brasileiros (1960 a 2010).** Rio de Janeiro: Centro de Estudos da Metrópole, 2013.

BARRO, R. J.; LEE, J. W. International comparisons of educational attainment. **Journal of Monetary Economics**, Holanda do Norte, v. 32, n. 3, p. 363–394, 1993.

BARRO, R. J.; LEE, J. W. International measures of schooling years and schooling quality. **American Economic Review**, v. 86, n. 2, p. 218–223, 1996. Disponível em: https://www.jstor.org/stable/2118126. Acesso em: 21 jul. 2020.

BARRO, R. J.; LEE, J. W. International data on educational attainment: Updates and implications. **Oxford Economic Papers**, v. 53, n. 3, p. 541-563, 2001. Disponível em: http://www.jstor.org/stable/3488631. Acesso em: 20 jul. 2020.

BARRO, R. J.; LEE, J. W. A new data set of educational attainment in the world, 1950– 2010. **Journal of Development Economics**, v. 104, p. 184–198, 2013.

BARRO, R. J.; LEE, J. W. **Barro-Lee educational attainment dataset**. 2018. Disponível em: http://www.barrolee.com/. Acesso em: 02 abr. 2020.

BECKER, G. S. **Human capital:** a theoretical and empirical analysis with special reference to education. New York: Columbia University Press for National Bureau of Economic Research (NBER), 1964.

CARABETTA, J.; DAHIS, R.; ISRAEL, F.; SCOVINO, F. **Base dos Dados: Repositório de Dados Abertos**. Disponível em: https://basedosdados.org. Acesso em: 2 abr. 2021.

COHEN, D.; LEKER, L. Health and education: Another Look with the Proper Data. **CEPR Discussion** Paper No. DP9940, p. 1–25, 2014.

COHEN, D.; SOTO, M. Growth and human capital: Good data, good results**. Journal of Economic Growth**, v.12, n.1, p. 51-76, 2007.

DE LA FUENTE, A.; DOMÉNCH, R. Human capital in growth regressions: How much difference does data quality make? **Journal of the European Economic Association**, v. 4, n. 1, p. 1–36, 2006.

EASTERLY, W.; LEVINE, R. Troubles with the neighbours: Africa's problem, Africa's opportunity. **Journal of African Economies**, v. 7, n. 1, p. 120–142, 1998.

FOLDVÁRI, P.; VAN LEEUWEN, B. Average years of education in Hungary: annual estimates, 1920-2006**. Eastern European Economics**, v. 47, n. 2, p. 5–20, 2009.

FOLDVÁRI, P.; VAN LEEUWEN, B. Educational and income inequality in Europe, ca. 1870–2000. **Cliometrica**, v. 8, n. 3, p. 271–300, 2014.

GARIBALDI, P. **Personnel Economics in Imperfect Labour Markets**. Reino Unido: Oxford University Press, 2006.

GONZALEZ, C. A. G.; AIDAR, T. Análise de pseudo-coortes a partir dos censos demográficos no brasil: uma aproximação metodológica. Campinas: **Textos Nepo**, v. 71, out. 2015. Disponível em: http://www.nepo.unicamp.br/publicacoes/textos_nepo/textos_nepo_71.pdf. Acesso em: 05 abr. 2021.

GUJARATI, D. N.; PORTER, D. C. **Econometria básica**. 5. ed. Porto Alegre: AMGH, 2011.

HALL, R. E.; JONES, C. I. Why do some countries produce so much more output per worker than others? **The Quarterly Journal of Economics**, v. 114, n. 1, p. 83–116, 1999. Disponível em: http://www.jstor.org/stable/2586948. Acesso em: 02 nov. 2020.

HANUSHEK, E. A.; WOESSMANN, L. Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. **Journal of Economic Growth**, v. 17, n. 4, p. 267–321, 2012.

KANG, T.; PAESE, L.; FELIX, N. Late and unequal: measuring enrolments and retention in Brazilian education, 1933-2010. **Revista de História Económica / Journal of Iberian and Latin American Economic History**, 2021. Disponível em: https://www.sciencegate.app/source/94029 Acesso em: 20 fev. 2021.

KC, S. *et al*. Projection of populations by level of educational attainment, age, and sex for 120 countries for 2005-2050. **Demographic Research**, v. 22, n. 15, p. 383–472, 2010.

KYRIACOU, G. A. Level and growth effects of human capital: a cross-country study of the convergence hypothesis. **Working Papers**, Starr Center for Applied Economics, New York University, p. 91–26, 1991.

LAU, L. J.; JAMISON, D. T.; LOUAT F. F. Education and productivity in developing countries an aggregate production. **Policy Research Working Paper Series**, v. 1, n. 1, 1991.

LEE, J. W.; LEE, H. Human capital in the long run. **Journal of Development Economics**, v. 122, p. 147–169, 2016.

LUTZ, W. Reconstruction of populations by age, sex and level of educational attainment for 120 countries for 1970-2000. **Vienna Yearbook of Population Research**, v. 5, n. 1, p. 193–235, 2007.

LUTZ, W. Demographic and human capital scenarios for the 21st century: 2018 assessment for 201 countries. Luxemburgo**: Publications Office of the European Union**, 2018.

MELO, H. P.; THOMÉ, D. **Mulheres e poder**: histórias, ideias e indicadores. Rio de Janeiro: FGV, 2018.

MINCER, J. Investment in human capital and personal income distribution. **Journal of Political Economy**, v. 66, n. 4, p. 281–302, 1958.

MINCER, J. Human capital, and Economic Growth. **Economics of Education Review**, v. 3, n. 3, p. 195-205, 1984.

MORRISSON, C.; MURTIN, F. The century of education. **Journal of Human Capital**, v. 3, n. 1, p. 1-42, 2009.

MULLIGAN, C. B.; SALA-I-MARTIN, X. Measuring aggregate human capital. **Journal of Economic Growth**, v. 5, n. 3, p. 215–252, 2000. Disponível em: http://www.jstor.org/stable/40216033. Acesso em: 10 nov. 2020.

NEHRU, V.; SWANSON, E.; DUBEY, A. A New database on human capital stock in developing and industrial countries: sources, methodology, and results. **Journal of Development Economics**, v. 26, n. 2, p. 379-401, 1995.

PESSÔA, S. A.; SUMMERHILL, W. R.; VAREJÃO, Neto. E. S. **Economic consequences of educational backwardness in twentieth-century Brazil**. Rio de Janeiro: FGV, 2019. Disponível em: https://hdl.handle.net/10438/29220. Acesso em: 20 dez. 2020.

POTANCOKOVÁ M.; K.C. S.; GOUJON A. **IIASA Interim Report**: Global estimates of mean years of schooling: A new methodology. IIASA, Luxemburgo, Austria, 2014. Disponível em: http://pure.iiasa.ac.at/id/eprint/11261/. Acesso em 3 de jun 2020.

PSACHAROPOULOS, G.; ARRIAGADA, A. M. The educational composition of the labour force: an international comparison. **International Labour Review**, v. 125, n. 5, p. 561–574, 1986.

PSACHAROPOULOS, G.; ARRIAGADA, A. M. The educational composition of the labor force: an international update. **Journal of Educational Planning and Administration**, v. 6, n. 2, p. 141–159, 1992.

PSACHAROPOULOS, G. On the weak versus the strong version of the screening hypothesis. **Economics Letters**, v. 4, n. 2, p. 181-185, 1979.

RAJAN, G. R.; ZINGALES, L. Financial dependence and growth. **American Economic Review**, v. 88, n. 3, p. 559–586, 1998. https://doi.org/10.3386/w5758.

RAMEY, G.; RAMEY, V. A. Cross-country evidence on the link between volatility and growth. **American Economic Review**, v. 85, n. 5, p. 1138–1151, 1995. Disponível em: http://www.jstor.org/stable/2950979 Acesso em: 04 fev. 2021.

RIBEIRO, S. C. A pedagogia da repetência. **Estudos Avançados**, v. 12, n. 5, p. 07-21, 1991.

SACHS, J. D.; WARNER, A. M. Natural resource abundance and economic growth. **Journal of Development Economics**, v. 59, n. 1, p. 43-76, 1995.

SHULTZ, T. W. Investment in human capital. **The American Economic Review**, v. 51, n. 5, p. 1-17, 1961. Disponível em: https://www.jstor.org/stable/1818907. Acesso em: 10 fev. 2020.

SOLOW, R. M. A Contribution to the theory of economic growth. the quarterly. **Journal of Economics**, v. 70, n. 1, p. 65–94, 1956.

SPERINGER, M. et al. **Validation of the Wittgenstein Centre back-projections for populations by age, sex, and six levels of education from 2010 to 1970**. Luxemburgo: IIASA Interim Report, 2015.

SPENCE, M. Job market signaling**. Oxford University Press**, v. 87, n. 3, p. 355-374, ago. 1973.

THE UNITED NATIONS. Convention on the elimination of all forms of discrimination against Women. **Treaty Series**, v. 1249, p. 13, 1988.

UNESCO Institute for Statistics - UIS. **UIS Methodology for estimation of mean years of schooling**. Montreal: UNESCO Institute for Statistics, 2013.

UNITED NATIONS DEVELOPMENT PROGRAMME - UNDP. The next frontier Human development and the Anthropocene. New York: **Human Development Report**, 2020.

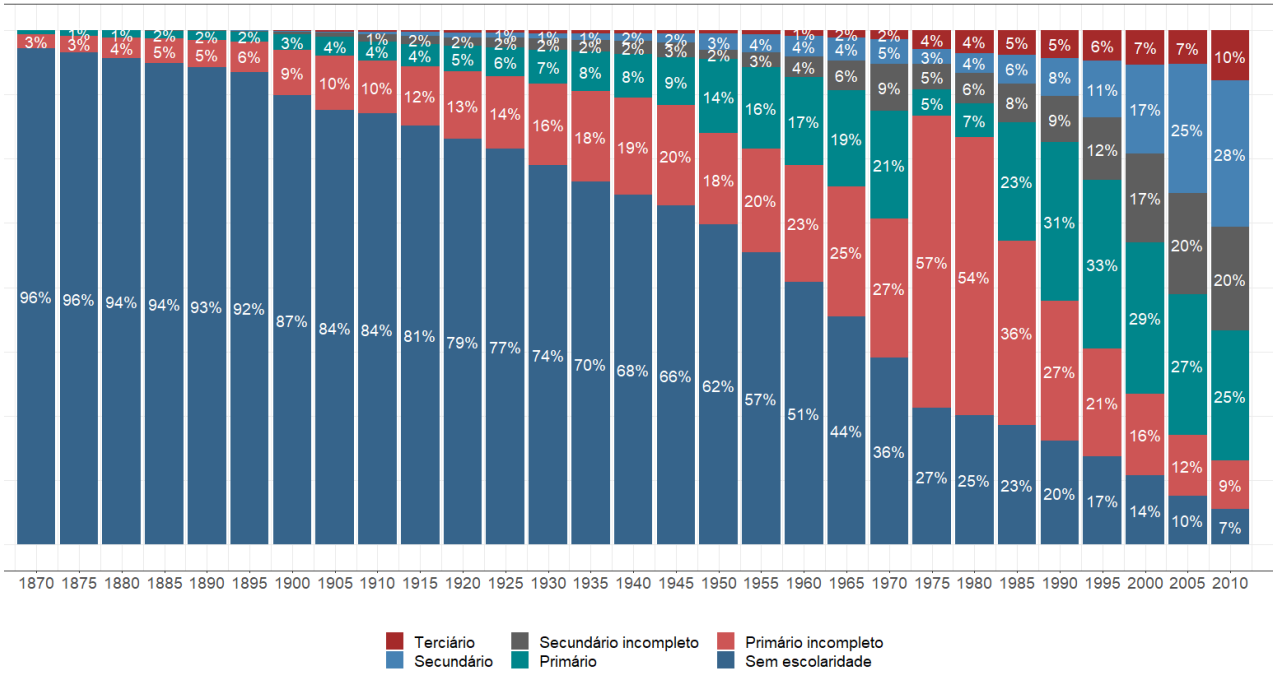
VAN LEEUWEN, B.; VAN LEEUWEN-LI, J. Education since 1820. In: ZANDEN, Jan Luiten van, et al. (eds.). **How was life?**: global well-being since 1820. Paris: OECD Publishing, 2014. p. 87-100.

VAN LEEUWEN, B.; VAN LEEUWEN-LI, J.; FOLDVARI, P. **Average years of education** (Average, total Population 15 years and older), 1850-2010, 2016. Disponível em: https://clio-infra.eu/Indicators/AverageYearsofEducation.html. Acesso em: mai. 2020.

WOESSMANN, L. Specifying human capital. **Journal of Economic Surveys**, v. 17, n. 3, p. 239–270, 2003.

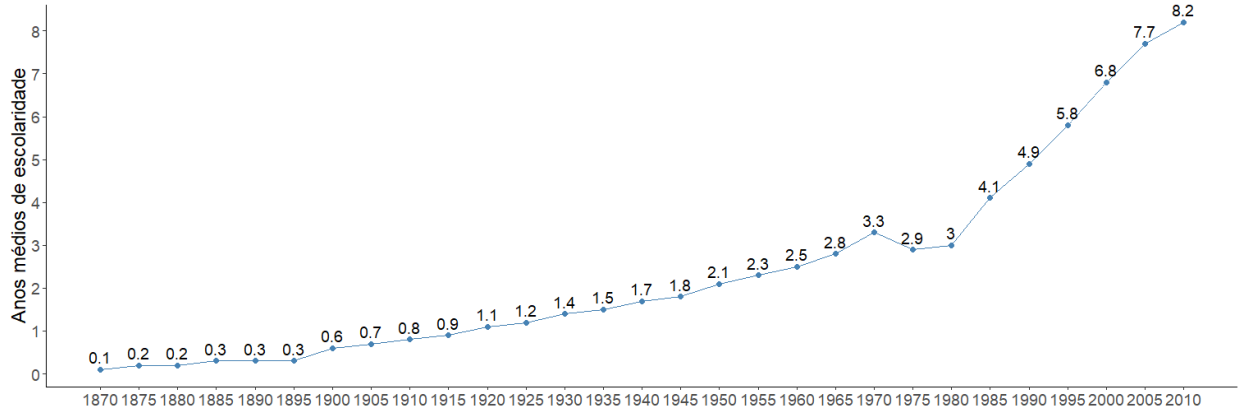
APÊNDICE A - Dados Lee e Lee (2016)

Gráfico 18 - Distribuição Educacional, população de 15 anos ou mais, Lee e Lee (2016)



Fonte: Elaboração própria, dados de Lee e Lee (2016).

Gráfico 19: Anos médios de escolaridade, população de 15 a 64 anos, Lee e Lee (2016)



Fonte: Elaboração própria, dados de Lee e Lee (2016).

APÊNDICE B – MICRODADOS

• Censo de 1960: Esses dados estão auto ponderados e apresentam uma fração amostral baixa, cerca de 1,25% da população, permitindo somente uma análise representativa para as unidades da federação. Neste censo foram coletados os dados dos moradores presentes, ausentes e não moradores dos domicílios. Caso os moradores estivessem ausentes no momento da entrevista, os moradores presentes forneciam as informações para a coleta. Por conta disso, excluem-se não moradores para a análise dos microdados, ou seja, indivíduos que estavam visitando o domicílio no momento da entrevista, para evitar dupla contagem de valores (Barbosa, 2013). Há no censo duas perguntas de interesse utilizadas para o cálculo dos anos médios de escolaridade, a última série concluída com aprovação e o grau da série concluída, destinadas para pessoas que frequentavam/frequentaram a escola.

• Censo de 1970: Apresenta uma amostra de 25% representativa também para regiões intramunicipais. Tal como em 1960, é necessário a exclusão dos não moradores para a realização da análise (Barbosa, 2013) e as perguntas aplicadas para a população que frequenta/frequentava a escola são semelhantes. Nesse censo, também são computadas provas de admissão, vestibular e art. 99/supletivo. Esse recenseamento tem a desvantagem de agregar todos os indivíduos que cursaram o 5º e o 6º ano em uma mesma variável, aos quais se atribuíram 17 anos de estudo.

• Censo de 1980: Também apresenta uma amostra de 25%. Neste ano e nos anos posteriores são somente coletados os dados de moradores ausentes e não moradores, ou seja, não é necessária nenhuma filtragem prévia das informações para a análise (Barbosa, 2013). A partir de 1980 há uma alteração das respostas do nível educacional para abranger a reforma do ensino (Lei 5.692 / 1971). A partir desse censo, estendendo-se aos demais, são aplicadas perguntas distintas para os indivíduos que frequentavam a escola, curso seriado e não seriado, no momento da entrevista e aqueles que já haviam frequentado. No caso de o indivíduo ter frequentado a escola em um curso inferior ou do mesmo grau a um curso já concluído, é necessário responder ambos os conjuntos de perguntas segundo o documento de instrução ao recenseador. Registro semelhante ocorre com os demais censos. Há a contabilização dos supletivos feitos via televisão ou rádio.

• Censo de 1991: Amostra de 25% da população. Tanto o censo de 1991 como o censo de 2000 são similares aos de 1980 em termos de coleta de dados de escolaridade.

• Censo de 2000: Amostra de 10% da população.

• Censo de 2010: Não foi possível aplicar a mesma metodologia de compatibilização, porque no questionário aplicado não há uma pergunta referente à série específica para a população que frequentou a escola (Gonzalez & Aidar, 2015). Seria necessário agrupar os anos médios de escolaridade em intervalos. Ao invés disso, optamos por utilizar dados da PNAD Contínua.

• PNAD Contínua, 2012 a 2015: Não foram utilizados dados da PNAD antiga por estarem descontinuados e por apresentarem flutuações amostrais e perguntas de escolaridade que variam entre períodos, exigindo alteração da classificação dos anos médios de escolaridade. A PNAD Contínua apresenta um plano amostral maior, menos sujeito a flutuações. Por isso, decidimos utilizá-la para estimar a escolaridade dos anos mais recentes. Para a estimação da distribuição educacional foram computados os dados do primeiro trimestre e utilizadas as perguntas relativas aos que cursavam/cursaram a escola no período.

APÊNDICE C – DETALHAMENTO DO REPOSITÓRIO NO GITHUB

Link do repositório: https://github.com/juliarwalter/pesquisa-anos-medios-de-escolaridade.

Documentação do README:

Objetivo: Disponibilizar metodologia aplicada para a estimação da escolaridade da população brasileira de 15 a 64 anos de 1925 a 2015, bem como a estimativa da escolaridade desagregada em gênero, cor/raça, estados (1950 a 2015) e macrorregiões (1950 a 2015).

Tecnologias utilizadas: Software R e Excel.

Base de dados utilizada: Microdados censo disponibilizados em: https://basedosdados.org/; Dados de matrícula Kang et al. (2021); Microdados PNAD Contínua, acesso via library ('PNADcIBGE'); SIDRA (estimativa populacional).

Instrução para implementação do projeto:

a) Necessário entrar no site: https://basedosdados.org/ e baixar os microdados (download CSV ou consulta no Big Query);

Figura 2 - Microdados de 1970: baixando arquivo CSV do site e selecionando variáveis de interesse

```
df <- read.csv("E: microdados_pessoa_1970.csv", header = TRUE, sep = ",", quote = "\"")
df <- df %>%
       select('sigla_uf', 'id_municipio', 'ordem','v023', 'v003', 'v022', 'v036',
       'v035', 'v030', 'v029', 'v024', 'id_domicilio', 'v026', 'v027', 'v037', 'v038', 'v054')
saveRDS(df, "/data/external/df_1970_comp.rds")
```

Fonte: Elaboração Própria.

Figura 3 - Microdados de 1980: baixando arquivo CSV do site e selecionando variáveis de interesse

```
df <- read.csv("microdados_pessoa_1980.csv", header = TRUE, sep = ",", quote = "\"")
df <- df %>%
       select('sigla_uf', 'id_municipio', 'numero_ordem','v519', 'v520', 'v521', 'v511',
       'v604', 'v606', 'v523', 'v522', 'v524', 'v525', 'v501', 'v509', 'v211', 'v508')
saveRDS(df, "/data/external/df_1980_comp.rds")
```

Fonte: Elaboração Própria.

Figura 4 - Microdados de 1991: baixando arquivo CSV do site e selecionando variáveis de interesse

```
df <- read.csv("microdados_pessoa_1991.csv", header = TRUE, sep = ",", quote = "\"")
df_ <- df %>%
       select('sigla_uf','v3071', 'v3072', 'v3073', 'v0323', 'v0324', 'v0325', 'v0326',
       'v0327', 'v0328', 'v3241', 'v0329', 'v7301', 'v0309', 'v0301', 'v0310')
saveRDS(df_, "/data/external/df_1991_comp.rds")
```

Fonte: Elaboração Própria.

Figura 5 - Microdados de 2000: baixando arquivo CSV do site e selecionando variáveis de interesse

```
df <- read.csv("microdados_pessoa_2000.csv", header = TRUE, sep = ",", quote = "\"")
df_ <- df %>%
       select('sigla_uf', 'id_municipio', 'id_mesorregiao','area_ponderacao', 'estr',
       'estrp', 'v4752', 'v0401', 'marca', 'v0402', 'v0403', 'v4070', 'v0419', 'v0428',
       'v0429', 'v0430', 'v0431', 'v0432', 'v0433', 'v0434', 'v4355', 'v4300', 'p001', 'v0408', 'v4090')
saveRDS(df_, "/data/external/df_2000_comp.rds")
```

Fonte: Elaboração Própria.

b) Rodar todos os arquivos “classificação” da pasta feature e posteriormente o arquivo “calculo_pop_.Rmd” no software R;
c) Rodar os arquivos da pasta estimation (resultará no output desejado: pasta final_data).

Abaixo segue um breve resumo da estrutura de arquivos do diretório:

├── README.md <- Arquivo com instruções iniciais sobre o projeto.
├── data
│ ├── external <- Microdados censo
│ ├── df_1960.rds <- Microdado censo 1960 *
│ ├── df_1970_comp.rds <- Microdado censo 1970 *
│ ├── df_1980_comp.rds <- Microdado censo 1980 *
│ ├── df_1991_comp.rds <- Microdado censo 1991 *
│ ├── df_2000_comp.rds <- Microdado censo 2000 *
│ ├── interim <- Base de dados transformadas.
│ ├── matricula_calculado.xlsx <- Planilha com dados de matrícula interpolados (output do arquivo: calculo_pop_.Rmd)
│ ├── pop_ajustado.xlsx <- Planilha com dados populacionais por estado (output do arquivo: calculo_pop_.Rmd). Dados colados no arquivo pop_calculado.xlsx, aba “dados_1960_2019”

│ ├── pop_calculado.xlsx <- Planilha com dados populacionais (input do arquivo: calculo_pop_.Rmd)
│ ├── processed <- Base de dados processadas para a implementação do projeto.
│ ├── df_1960.rds <- output do arquivo classificação_1960.Rmd e input para calculo_pop_.Rmd e pasta feature*
│ ├── df_1970.rds <- output do arquivo classificação_1970.Rmd e input para calculo_pop_.Rmd e pasta feature*
│ ├── df_1980.rds <- output do arquivo classificação_1980.Rmd e input para calculo_pop_.Rmd e pasta feature*
│ ├── df_1991.rds <- output do arquivo classificação_1991.Rmd e input para calculo_pop_.Rmd e pasta feature*
│ ├── df_2000.rds <- output do arquivo classificação_2000.Rmd e input para calculo_pop_.Rmd e pasta feature*
│ ├── df_2012.rds; df_2013.rds; df_2014.rds; df_2015.rds <- outputs do arquivo classificação_2012-2015.Rmd e input para calculo_pop_.Rmd e pasta feature*
│ ├── matricula_.xlsx <- Dados de matrícula e taxa bruta (Kang et al., 2021), input pasta feature
│ ├── matricula_amarelos.xlsx <- Dados de matrícula e taxa bruta, input estimativa_anos_medios_cor.Rmd
│ ├── matricula_brancos.xlsx <- Dados de matrícula e taxa bruta, input estimativa_anos_medios_cor.Rmd
│ ├── matricula_pretos_pardos.xlsx <- Dados de matrícula e taxa bruta, input estimativa_anos_medios_cor.Rmd
│ ├── pop_et.xlsx <- Dados de população por cor, input estimativa_anos_medios_cor.Rmd
│ ├── pop_h_m.xlsx <- Dados de população por gênero, input estimativa_anos_medios_mh.Rmd e estimativa_anos_medios.Rmd.
│ ├── pop_macro.xlsx <- Dados de população por macrorregião, input estimativa_anos_medios_macro.Rmd
│ ├── pop_uf.xlsx <- Dados de população por estado, input estimativa_anos_medios_regiao.Rmd
├── notebooks
│ ├── features
│ ├── classificação_1960.Rmd <- Padronização do censo de 1960

```
│       ├── classificação_1970.Rmd <- Padronização do censo de 1970
│       ├── classificação_1980.Rmd <- Padronização do censo de 1980
│       ├── classificação_1991.Rmd <- Padronização do censo de 1991
│       ├── classificação_2000.Rmd <- Padronização do censo de 2000
│       ├── classificação_2012-2015.Rmd <- Padronização da PNAD Contínua (2012-
2015)
│       ├── calculo_pop_.Rmd <- Estimativa populacional e interpolação de dados de
matrícula
│   ├── estimation
│       ├── estimativa_anos_medios.Rmd <- Modelo que estima a escolaridade da
população total de 15 a 64 anos de 1925 a 2015 (output escolaridade_total.xlsx)
│       ├── estimativa_anos_medios_cor.Rmd.<- Modelo que estima a escolaridade por
cor (output escolaridade_et.xlsx)
│       ├── estimativa_anos_medios_macro.Rmd.<- Modelo que estima a escolaridade por
macrorregião (output escolaridade_macro.xlsx)
│       ├── estimativa_anos_medios_mh.Rmd <- Modelo que estima a escolaridade por
gênero (output escolaridade_h_m.xlsx)
│       ├── estimativa_anos_medios_regiao.Rmd.<- Modelo que estima a escolaridade por
estado (output escolaridade_região.xlsx)
├── final_data
│   ├── escolaridade_cor.xlsx <- database escolaridade por cor
│   ├── escolaridade_h_m.xlsx <- database escolaridade por gênero
│   ├── escolaridade_macro.xlsx <- database escolaridade por macrorregião
│   ├── escolaridade_regiao.xlsx <- database escolaridade por regiao
│   ├── escolaridade_total.xlsx <- database escolaridade da população total
```

Arquivos sinalizados com * devem ser baixados, microdados censo (pasta data/external), ou são output dos arquivos "classificação" pasta "feature" (pasta data/processed).

Link do repositório: https://github.com/juliarwalter/pesquisa-anos-medios-de-escolaridade/complemento.

Instrução para implementação do projeto:

a) Rodar arquivo “graficos_1_2.Rmd” para gerar gráficos da seção 1, 2 do artigo e apêndice;
b) Rodar arquivo “graficos_3_4.Rmd” para gerar gráficos da seção 3 e 4 do artigo.

Abaixo segue um breve resumo da estrutura de arquivos do diretório:

├── README.md <- Arquivo com instruções iniciais sobre o projeto
├── data
│ ├── dados_graficos_1_2.xlsx <- Anos médios e distribuição de escolaridade dos estudos: Barro e Lee (2018); Lee e Lee (2016); Morrisson e Murtin (2009); Van Leeuwen et al. (2016); Lutz et al. (2018); Cohen e Leker (2014) e estimativas calculadas no artigo.
│ ├── dados_graficos_3_4.xlsx <- Dados estimados no artigo
├── notebooks
│ ├── graficos_1_2.Rmd <- Gráficos gerados na seção 1, 2 do artigo e apêndice
│ ├── graficos_3_4.Rmd <- Gráficos gerados na seção 3, 4 do artigo

APÊNDICE D – MODELAGEM ARIMA

Figura 6 - Teste ADF, modelo sem diferenciação

```
Teste Aumentado de Dickey-Fuller para d_anos_medios
testar para baixo a partir de 5 defasagens, critério AIC
tamanho da amostra: 85
hipótese nula de raiz unitária: a = 1

  teste com constante
  incluindo 4 defasagens de (1-L)d_anos_medios
  modelo: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0,100719
  estatística de teste: tau_c(1) = -1,47447
  p-valor assintótico 0,5468
  coeficiente de 1ª ordem para e: 0,001
  diferenças defasadas: F(4, 79) = 7,571 [0,0000]

  com constante e tendência
  incluindo 3 defasagens de (1-L)d_anos_medios
  modelo: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0,558908
  estatística de teste: tau_ct(1) = -3,59796
  p-valor assintótico 0,02993
  coeficiente de 1ª ordem para e: 0,026
  diferenças defasadas: F(3, 80) = 3,125 [0,0304]

  com constante e tendência linear e quadrática
  incluindo 3 defasagens de (1-L)d_anos_medios
  modelo: (1-L)y = b0 + b1*t + b2*t^2 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0,559997
  estatística de teste: tau_ctt(1) = -3,60557
  p-valor assintótico 0,08838
  coeficiente de 1ª ordem para e: 0,026
  diferenças defasadas: F(3, 79) = 3,125 [0,0304]
```

Fonte: Elaboração própria.

Figura 7- Teste ADF, modelo com uma diferenciação

```
Teste Aumentado de Dickey-Fuller para anos_medios
testar para baixo a partir de 5 defasagens, critério AIC
tamanho da amostra: 85
hipótese nula de raiz unitária: a = 1

  teste com constante
  incluindo 5 defasagens de (1-L)anos_medios
  modelo: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): 0,00430295
  estatística de teste: tau_c(1) = 1,1388
  p-valor assintótico 0,9978
  coeficiente de 1ª ordem para e: 0,006
  diferenças defasadas: F(5, 78) = 7,039 [0,0000]

  com constante e tendência
  incluindo 4 defasagens de (1-L)anos_medios
  modelo: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0,00826162
  estatística de teste: tau_ct(1) = -1,36263
  p-valor assintótico 0,8717
  coeficiente de 1ª ordem para e: 0,026
  diferenças defasadas: F(4, 79) = 4,765 [0,0017]

  com constante e tendência linear e quadrática
  incluindo 5 defasagens de (1-L)anos_medios
  modelo: (1-L)y = b0 + b1*t + b2*t^2 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -0,134167
  estatística de teste: tau_ctt(1) = -3,11192
  p-valor assintótico 0,2453
  coeficiente de 1ª ordem para e: -0,007
  diferenças defasadas: F(5, 76) = 5,903 [0,0001]
```

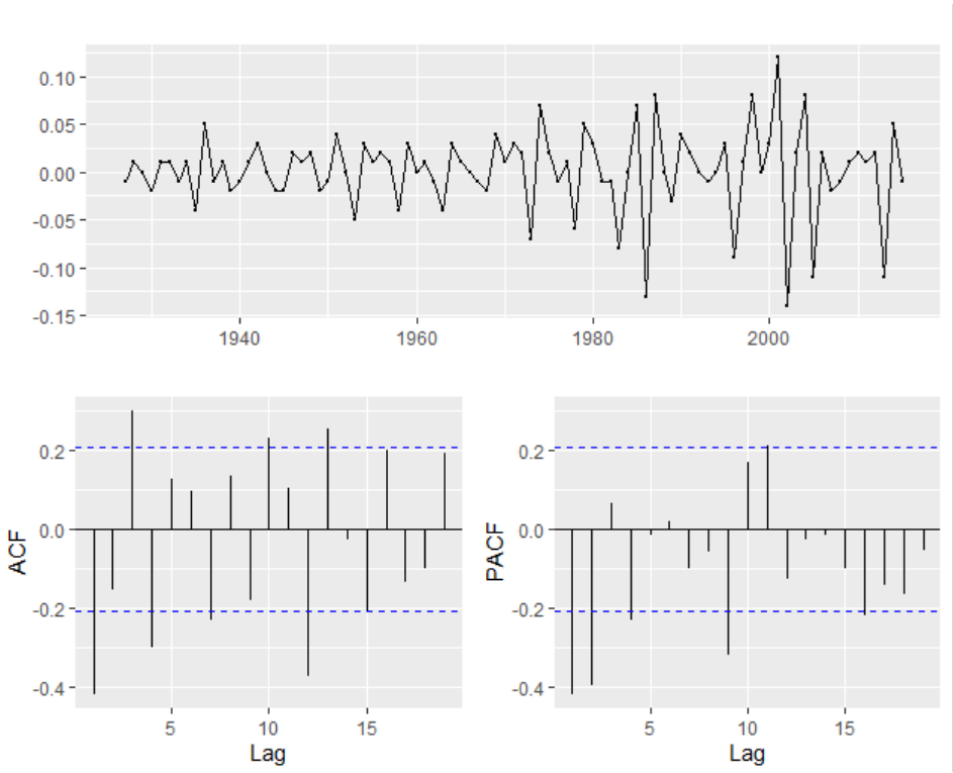
Fonte: Elaboração Própria.

Figura 8 - Teste ADF, modelo com duas diferenciações

```
Teste Aumentado de Dickey-Fuller para d_d_anos_medios
testar para baixo a partir de 5 defasagens, critério AIC
tamanho da amostra: 85
hipótese nula de raiz unitária: a = 1

  teste com constante
  incluindo 3 defasagens de (1-L)d_d_anos_medios
  modelo: (1-L)y = b0 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -2,29036
  estatística de teste: tau_c(1) = -6,6666
  p-valor assintótico 2,542e-009
  coeficiente de 1ª ordem para e: -0,002
  diferenças defasadas: F(3, 80) = 7,268 [0,0002]

  com constante e tendência
  incluindo 3 defasagens de (1-L)d_d_anos_medios
  modelo: (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -2,29221
  estatística de teste: tau_ct(1) = -6,63738
  p-valor assintótico 2,607e-008
  coeficiente de 1ª ordem para e: -0,003
  diferenças defasadas: F(3, 79) = 7,223 [0,0002]

  com constante e tendência linear e quadrática
  incluindo 3 defasagens de (1-L)d_d_anos_medios
  modelo: (1-L)y = b0 + b1*t + b2*t^2 + (a-1)*y(-1) + ... + e
  valor estimado de (a - 1): -2,32894
  estatística de teste: tau_ctt(1) = -6,69425
  p-valor assintótico 2,88e-008
  coeficiente de 1ª ordem para e: -0,003
  diferenças defasadas: F(3, 78) = 7,416 [0,0002]
```

Fonte: Elaboração Própria.

Figura 9 – Correlograma da série com duas diferenciações



Fonte: Elaboração Própria.

Tabela 4 - Modelos ARIMA e principais resultados

Modelos	Coeficientes	Teste de Ljung-Box	AIC	Hannan-Quinn	Critério de Schwarz
ARIMA(0,2,2)	Não significativo	-	-	-	-
ARIMA(1,1,0)	Significativo	Não passou	−314,9364	−311,9122	−307,4370
ARIMA(1,2,1)	Significativo	Não passou	−331,7874	−327,7750	−321,8329
ARIMA(2,2,0)	Não significativo	-	-	-	-
ARIMA(2,2,1)	Não significativo	-	-	-	-
ARIMA(2,2,2)	Não significativo	-	-	-	-
ARIMA(3,2,1)	Não significativo	-	-	-	-
ARIMA(4,2,1)	Não significativo	-	-	-	-
ARIMA(2,2,5)	Não significativo	-	-	-	-
ARIMA(0,2,1)	Significativo	Passou	-331,9406	-329,9344	-316,9633
ARIMA(2,2,7)	Significativo	Não passou	−339,0140	−335,0016	−329,0595

Fonte: Elaboração Própria.

Nota: ARIMA(0,2,1) passou no teste de heterocedasticidade e normalidade dos resíduos.

Figura 10 - Resultado do modelo ARIMA (0,2,1)

```
Funções calculadas: 34
Cálculos de gradientes: 13

Modelo 6: ARIMA, usando as observações 1927-2015 (T = 89)
Estimado usando AS 197 (Máxima verossimilhança exata)
Variável dependente: (1-L)^2 anos_medios
Erros padrão baseados na Hessiana

            coeficiente   erro padrão     z       p-valor
  ---------------------------------------------------------
  theta_1     -0,624777     0,0877227    -7,122   1,06e-012 ***

Média var. dependente   0,001236   D.P. var. dependente    0,043218
Média de inovações      0,003614   D.P. das inovações      0,036552
R-quadrado              0,999777   R-quadrado ajustado     0,999777
Log da verossimilhança  167,9703   Critério de Akaike     -331,9406
Critério de Schwarz    -326,9633   Critério Hannan-Quinn  -329,9344

                      Real  Imaginária    Módulo  Frequência
  ------------------------------------------------------------
  MA
    Raiz  1           1,6006     0,0000     1,6006     0,0000
  ------------------------------------------------------------
```

Fonte: Elaboração Própria.