Universidade Federal do Pará
Instituto de Tecnologia
Programa de Pós-Graduação em Engenharia Elétrica

Ciência de Dados Aplicada em Dados Públicos: Estudos de Caso Acerca da Previdência Social Brasileira

Francisco Eguinaldo de Albuquerque Félix Júnior

DM 16/2020

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém - Pará - Brasil
2020

Universidade Federal do Pará

Instituto de Tecnologia

Programa de Pós-Graduação em Engenharia Elétrica

Francisco Eguinaldo de Albuquerque Félix Júnior

Ciência de Dados Aplicada em Dados Públicos: Estudos de Caso Acerca da Previdência Social Brasileira

Dissertação de Mestrado submetida à avaliação da Banca Examinadora aprovada pelo colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará e julgada adequada para a obtenção do Grau de Mestre em Engenharia Elétrica na área de computação aplicada.

UFPA / ITEC / PPGEE

Campus Universitário do Guamá

Belém - Pará - Brasil

2020

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

A345c Albuquerque Félix Junior, Francisco Eguinaldo de
Ciência de Dados Aplicada em Dados Públicos: Estudos de Caso Acerca da Previdência Social Brasileira / Francisco Eguinaldo de Albuquerque Félix Junior. — 2020.
82 f. : il. color.

Orientador(a): Prof. Dr. Marcelino Silva da Silva
Dissertação (Mestrado) - Programa de Pós-Graduação em Arquitetura e Urbanismo, Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2020.

1. Ciência de Dados. 2. Previdência Social. 3. Índice de Gini. 4. Dados Abertos. 5. Análise de Dados. I. Título.

CDD 006.312

Francisco Eguinaldo de Albuquerque Félix Júnior

Ciência de Dados Aplicada em Dados Públicos: Estudos de Caso Acerca da Previdência Social Brasileira

Esta Dissertação foi julgada adequada para obtenção do título de Mestre em Engenharia Elétrica, e aprovada na sua forma final, pela banca examinadora, designada pelo Programa de Pós-Graduação em Engenharia Elétrica do Instituto de Tecnologia da Universidade Federal do Pará em 20 de Março de 2020.

Banca Examinadora:

Prof. Dr. Marcelino Silva da Silva
Orientador

Prof. Dr. Diego Lisboa Cardoso
Membro Interno - PPGEE UFPA

Prof. Dr. Claudio Alberto Castelo Branco Puty
Membro Externo - FACECON UFPA

Prof. Dra. Liviane Ponte Rêgo
Membro Externo - FACECON UFPA

Visto:

Prof. Dra. Maria Emília de Lima Tostes
Coordenadora do PPGEE/ITEC/UFPA

Belém - Pará - Brasil
2020

Aos meus pais
Francisco Eguinaldo de Albuquerque Félix e Ivoneide Sales de Queiroz,
por sempre estarem comigo em todos os momentos.

Agradecimentos

Primeiramente à Deus, por ter me dado saúde e forças para enfrentar e superar todas as dificuldades encontradas nessa trajetória.

Aos meus pais, Francisco Eguinaldo e Ivoneide Sales, que me incentivaram durante toda a minha vida e estiveram do meu lado em todos os momentos que precisei, prestando apoio e sempre me confortando com palavras de amor e confiança.

Ao meu orientador prof. Marcelino Silva, pelo suporte, ajuda, confiança para a elaboração do trabalho e pelos conselhos acadêmicos que vou levar por toda a vida.

Aos meus amigos do LPO e LINC, Daniel Victor, Lucas, Ewerton, Sandio, Igor, Gean, Daniel Souza e Paulo que sempre me ajudaram nos trabalhos e dificuldades encontradas no dia a dia, em especial ao meu amigo André Augusto, companheiro no desenvolvimento desta pesquisa.

Aos professores que colaboraram para a minha formação acadêmica, em especial Fábio Lobato, Diego Lisboa, Cláudio Puty e Liviane Rêgo que sempre se dispuseram e estiveram presentes para me ajudar no desenvolvimento do trabalho.

À minha amiga Luanny Gomes, pela confiança, companheirismo, auxílio e colaboração na realização da minha pesquisa.

Aos meus amigos Max Weverton e Henrique Dias, pela parceria e ensinamentos de vida proporcionados no último ano.

À UFPA e ao PPGEE, por terem me acolhido como aluno de mestrado durante 2 anos, disponibilizando toda a infraestrutura necessária para a elaboração dessa dissertação.

À CAPES pelo fomento da pesquisa realizada.

"A imaginação é mais importante que a ciência,
porque a ciência é limitada, ao passo que a
imaginação abrange o mundo inteiro."
(Albert Einstein)

Resumo

A Ciência de Dados trata-se de uma área interdisciplinar relacionada a análise de dados, a qual objetiva a extração de conhecimento e possíveis tomadas de decisão acerca de problemáticas especificas. Nesse contexto, os dados abertos governamentais, que muitas vezes necessitam de pré-tratamentos e métodos computacionais para processamento acerca dos seus conjuntos de dados, apresentam-se como potenciais fontes de informações a serem exploradas na ótica de Ciência de Dados, possibilitando o desenvolvimento de estratégias cada vez mais eficientes e otimizadas em gestão pública. Diante disso, e aliado as recentes discussões relacionadas a reforma na previdência social brasileira, essa dissertação apresenta dois estudos de caso referentes a análises no sistema previdenciário nacional. O primeiro estudo utilizou os microdados referentes aos censos demográficos de 2000 e 2010, disponibilizados pelo IBGE, propondo avaliar a participação que as aposentadorias e pensões possuem na desigualdade de renda da população nos anos avaliados acerca dos estados e municípios brasileiros. Os resultados mostram que, embora os benefícios analisados contribuam para a concentração de renda no Brasil, a parcela correspondente até um salário mínimo contribui para a desconcentração da renda, e aquela acima de um salário contribui para a concentração, sendo um padrão repetitivo em todo o território nacional. Por outro lado, o segundo estudo propôs uma avaliação dos impactos que a reforma da previdência, proposta pela PEC 06/2019, causaria acerca das concessões dos benefícios de aposentadoria entre o período de 1995 a 2016. Observou-se que a PEC 06/2019 dificultaria o acesso aos benefícios, na qual, aproximadamente 83,28% das aposentadorias não haveriam sido concedidas se a mesma já estivesse em vigor desde 1995.

Palavras-chaves: Ciência de Dados. Previdência Social. Índice de Gini. Dados Abertos. Análise de Dados.

Abstract

Data Science is an interdisciplinary area related to data analysis, which aims to extract knowledge and possible decision-making about specific problems. In this context, open government data, which often need pre-treatments and computational methods to process their data sets, present themselves as potential sources of information to be explored taking the Data Science's perspective, allowing the development of strategies each time more efficient and optimized in public management. Given this, and allied to the recent discussions related to the reform in the Brazilian social security, this dissertation presents two case studies referring to analyzes in the national social security system. The first study used the microdata referring to the demographic censuses of 2000 and 2010, made available by IBGE, proposing to evaluate the participation that retirements and pensions have in the income inequality of the population in the years evaluated about Brazilian states and municipalities. The results show that, although the analyzed benefits contribute to the Brazil income concentration, the portion corres=ponding to a minimum wage contributes to the deconcentration of income, and the portion above one salary contributes to the concentration, being a repetitive pattern throughout the country. On the other hand, the second study proposed an evaluation of the impacts caused by the pension reform, which is proposed in PEC 06/2019. It was observed that PEC 06/2019 would hinder access to benefits, in which approximately 83,28% of the pensions would not have been granted had it been in effect since 1995.

Keywords: Data Science. Social Security. Retirements. Gini Index. Open Data. Data Analysis.

Lista de ilustrações

Figura 1 – Processo de descoberta de conhecimento em bases de dados. 6
Figura 2 – Competências fundamentas para Ciência de Dados. 9
Figura 3 – Tipos de dados abertos 11
Figura 4 – Curva de Lorenz 13
Figura 5 – Variação das taxas de formalidade e fesemprego e do índice de Gini no Brasil, 2001 a 2014 23
Figura 6 – Fluxograma de desenvolvimento do primeiro estudo de caso 25
Figura 7 – (a) e (b): Índice de Gini das cidades brasileiras. (c) e (d): Rendimento médio domiciliar dos municípios brasileiros 33
Figura 8 – Participação percentual das rendas de aposentadorias e pensões no total dos rendimentos per capita das unidades federativas brasileiras, 2000 e 2010 34
Figura 9 – Progressividade (π_h) da parcela de rendimentos de aposentadorias e pensões por categorias para os municípios brasileiros, 2000 e 2010 . . . 36
Figura 10 – Matrizes de correlação das variáveis. (a) 2000 (b) 2010 37
Figura 11 – Fluxograma de desenvolvimento do segundo estudo de caso 40
Figura 12 – (a) Comparação entre os dados disponibilizados pela CPI da Previdência e as informações contidas no APES. (b) Erro percentual. 43
Figura 13 – Percentual de participação dos benefícios da base dados classificados por espécie. 44
Figura 14 – Fluxograma da simulação da reforma previdenciária PEC 06/2016 . . . 46

Lista de tabelas

Tabela 1 – Síntese dos principais trabalhos correlatos. 21
Tabela 2 – Variáveis selecionadas dos censos demográficos e suas descrições 26
Tabela 3 – Parâmetros de indexação e notação geral. 28
Tabela 4 – Descrição dos principais arquivos do projeto no GitHub referente ao primeiro estudo de caso 30
Tabela 5 – Índice de Gini (G) e renda média domiciliar (RD), Brasil 2000 e 2010. 31
Tabela 6 – Progressividade (π_h) da parcela de rendimentos de aposentadorias e pensões por categorias para os estados brasileiros, 2000 e 2010 35
Tabela 7 – Informações sobre os microdados disponibilizados pela CPI da Previdência de 2017. 41
Tabela 8 – Comparação entre os dados disponibilizados pelo DATAPREV e as informações contidas no APES: Quantidade de benefícios concedidos por grupos de espécies entre 2014/2016 43
Tabela 9 – Regras para obtenção de aposentadorias propostas pela PEC 06/2019. 45
Tabela 10 – Descrição dos principais arquivos do projeto no GitHub referente ao segundo estudo de caso 49
Tabela 11 – Idade média de entrada em benefício pelas regras atuais do RGPS. . . 50
Tabela 12 – Tempo (em anos) médio de contribuição pelas regras atuais do RGPS. 50
Tabela 13 – Número médio de contribuições por ano. 51
Tabela 14 – Proporção de pessoas que não atendem as regras da PEC na mesma data que deram entrada nos benefícios com as regras atuais. 52
Tabela 15 – Idade média estimada para a concessão das aposentadorias de acordo as diretrizes previstas pela PEC 06/2019. 53
Tabela 16 – Percentual médio a ser aplicado para cálculo do valor do benefício. . . 54

Lista de abreviaturas e siglas

AEPS	Anuário Estatístico da Previdência Social
AMC	Área Mínimas Comparáveis
ANFIP	Associação Nacional dos Auditores Fiscais da Receita Federal
BI	Business Intelligence
CPI	Comissão Parlamentar de Inquérito
ETL	Extract, Transfom and Load
GPL	General Public License
IBGE	Instituto Brasileiro de Geografia e Estatística
INCP	Índice Nacional de Preços ao Consumidor
INSS	Instituto Nacional de Seguro Social
KDD	Knowledge-Discovery in Database
LAI	Lei de Acesso à Informação
LDO	Lei de Diretrizes Orçamentarias
MD	Mineração de Dados
ONU	Organização das Nações Unidas
PAYG	Pay As You Go
PDI	Pentaho Data Integration
PEC	Proposta de Ementa Constitucional
PIB	Produto Interno Bruto
PNAD	Pesquisa Nacional por Amostra de Domicílios
RD	Rendimento Domiciliar
RGPS	Regime Geral da Previdência Social
SGBD	Sistema Gerenciador de Banco de Dados
SQL	Structured Query Language

Sumário

1 INTRODUÇÃO 1
1.1 Motivação e Justificativa 2
1.2 Objetivos Geral e Específicos 3
1.3 Organização da Dissertação 3

2 REFERENCIAL TEÓRICO 5
2.1 Considerações Iniciais 5
2.2 Análise de Dados 5
2.3 Ciência de Dados 8
2.4 Dados Abertos 9
2.4.1 Dados Abertos Governamentais 10
2.5 Técnicas Matemáticas Utilizadas 12
2.5.1 Decomposição do Índice de Gini 12
2.5.2 Correlação de Variáveis 14
2.6 Considerações Finais 16

3 TRABALHOS CORRELATOS 18
3.1 Considerações Iniciais 18
3.2 Correlatos 18
3.3 Considerações Finais 22

4 ANÁLISE DA PARTICIPAÇÃO DAS APOSENTADORIAS E PENSÕES NA DISTRIBUIÇÃO DE RENDA DO BRASIL, 2000 E 2010 23
4.1 Considerações Iniciais 23
4.2 Metodologia Utilizada 25
4.2.1 Origem dos Dados 25
4.2.2 Pré-Processamento dos Dados 26
4.2.3 Processamento 27
4.3 Resultados e Discussões 31
4.4 Considerações Finais 37

5 OS IMPACTOS DA REFORMA PREVIDENCIÁRIA ACERCA DAS CONCESSÕES DE BENEFÍCIOS DE APOSENTADORIA 39
5.1 Considerações Iniciais 39
5.2 Metodologia Utilizada 40

5.2.1 Origem dos Dados e Elaboração do *Data Warehouse* 40
5.2.2 Análise Exploratória 42
5.2.3 Simulação da Reforma Previdenciária (PEC06/2019) 44
5.3 Resultados e Discussão **50**
5.4 Considerações Finais **55**

6 CONSIDERAÇÕES FINAIS **56**
6.1 Dificuldades Encontradas **57**
6.2 Trabalhos Futuros **58**
6.3 Publicações **58**

REFERÊNCIAS **60**

APÊNDICES **67**

APÊNDICE A – ATRIBUTOS DOS MICRODADOS DISPONIBILIZADOS PELA CPI DA PREVIDÊNCIA **68**

1 Introdução

A Previdência Social é um programa de seguro público responsável por manter a fonte de renda dos trabalhadores e suas famílias quando estes perdem suas capacidades laborais provisoriamente, por acidentes, doenças ou maternidade, ou permanentemente, devido a morte, invalidez ou velhice. Esse programa é encarregado pelo ressarcimento de diversos benefícios sociais do trabalhador brasileiro, tais como aposentadorias, auxílios-doença, auxílios-acidente, pensão por morte, dentre outros (MINISTÉRIO DA FAZENDA, 2018). No cenário brasileiro, os trabalhadores formais, correspondentes aqueles que possuem carteira assinada, financiam mensalmente o fluxo de caixa dos repasses previdenciários, sendo comumente conhecido como regime de repartição simples ou *pay as you go* (PAYG). O Regime Geral da Previdência Social (RGPS), responsável por atender os trabalhadores da iniciativa privada e aos servidores públicos que não contam com regimes próprios de previdência, estrutura-se nesse modelo de repartição (CAMARANO; FERNANDES, 2016).

Os segurados da Previdência Social necessitam, como pré-requisito mínimo, contribuir regularmente para o Instituto Nacional do Seguro Social (INSS). Esse corresponde a uma autarquia do Governo Federal vinculado ao Ministério da Economia, possuindo como principal responsabilidade organizar as arrecadações e pagamentos dos benefícios previdenciários previstos por lei (INSS, 2020). O INSS trabalha em parceria com Empresa de Tecnologia e Informações da Previdência, o DATAPREV, encarregada por realizar o armazenamento e processamento dos dados gerados pelo instituto (DATAPREV, 2020).

Avaliando a Previdência Social em uma ótica mundial, observa-se em diversos países a realização de modificações em seus sistemas previdenciários tradicionais. Tais medidas, na maior parte das vezes, envolvem um acréscimo no tempo mínimo de contribuição e idade mínima como requisitos para acesso aos benefícios. Essas reformas, comumente, justificaram-se em uma relação de proporcionalidade entre as variáveis demográficas e financeiras, ou seja, o aumento da expectativa de vida e estagnação na densidade de população economicamente ativa proporcionaria gastos com beneficiários maiores do que arrecadações. Dessa forma, esse novo perfil etário causaria um desequilíbrio nas contas do sistema previdenciário (SILVA, 2017; COSTANZI; ANSILIERO, 2017).

Nesse contexto, a Organização das Nações Unidas (ONU) projeta que até 2100 a população idosa do mundo, correspondente a pessoas acima de 60 anos, tende a triplicar, estimando que essa categoria da população ultrapasse 3 bilhões de habitantes (UNITED NATIONS, 2017). A nível nacional, de acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), os índices de longevidade apresentam-se progressivos para os próximos anos convergindo para um crescente envelhecimento populacional (IBGE, 2018).

1.1 Motivação e Justificativa

Avaliando o cenário político brasileiro, observa-se que os debates relacionados a implantação de uma reforma na Previdência Social ganharam destaque novamente no final do ano de 2016. Devido à natureza macroeconômica desse tema, diversos debates foram acarretados acerca da sua dimensão econômica-fiscal, por compreender uma significativa parcela dos orçamentos públicos, e do seu caráter político-social, por refletir diretamente em mudanças nas regras previdenciárias sobre um conjunto grande da população, incluindo contribuintes/segurados e beneficiários.

Dessa forma, o governo federal, motivado pela desaceleração do crescimento econômico e elevação da dívida pública (ROSSI; MELLO, 2017), justificou a necessidade dessa reforma argumentando que, devido as elevadas taxas de longevidade registradas, o atual sistema tende a se tornar insustentável no decorrer dos próximos anos, e que Previdência Social pressiona a carga tributária nacional diminuindo o espaço para outros setores na composição do gasto público (SECRETÁRIA DE POLÍTICA ECONÔMICA, 2016). Com isso, foi apresentada a população brasileira a Proposta de Ementa Constitucional n. 287 de dezembro de 2016 (PEC 287/16) – posteriormente substituída pela PEC 06/2019 devido ao seu insucesso – propondo alterações nas regras previdenciárias em idade mínima e tempo de contribuição para acesso aos benefícios. Destaca-se que entre os itens da reforma as principais modificações ocorrem no RGPS.

Embora diversos cálculos atuarias estejam sendo realizados a fim de estimar, a curto e longo prazo, a solvência do sistema previdenciário atual (INSS, 2020; PLAMONDON et al., 2002), é necessário ponderar acima de tudo o impacto social que tais alterações acarretariam. Mediante a isso, o Ministério da Fazenda, através da Comissão Parlamentar de Inquérito (CPI) da Previdência, divulgou em junho de 2017 bases de dados relacionados aos números de benefícios de aposentadorias concebidos pelo RGPS, com o intuito de possibilitar a investigação dos resultados divulgados pelo programa.

Paralelamente, o advento da ciência de dados tem ganhado destaque no cenário interdisciplinar com a crescente densidade de informações geradas pelos diversos segmentos da sociedade (CARDOSO, 2019). Tal fator possibilita, sobretudo, o desenvolvimento de um novo campo de estudo para análises, simulações de cenários e extração de conhecimento válido acerca de bases de dados públicas. Nesse contexto, a utilização dos microdados previdenciários aliados a outras fontes de informação - demográficas e sociais - motivou o desenvolvimento dessa pesquisa, visto que possibilita a realização de avaliações críticas acerca do efeito que as aposentadorias causam na população brasileira e das consequências sociais que uma possível reforma nas regras para acesso aos benefícios causaria.

1.2 Objetivos Geral e Específicos

Diante disso, esta pesquisa objetiva, de forma geral, explorar a atual situação do sistema previdenciário brasileiro a partir da aplicação de técnicas e estratégias de Ciência de Dados em Dados Abertos Governamentais da Previdência Social. Para tanto, foram realizados dois estudos de caso visando: estimar a participação que o sistema previdenciário possui no cenário socioeconômico nacional; e simular os impactos que a reforma prevista na PEC 06/2019 causaria nas concessões de aposentadorias.

Especificamente, essa pesquisa tem como objetivos:

- Utilizar os microdados dos censos demográficos de 2000 e 2010, disponibilizados pelo IBGE, e metodologia de decomposição do coeficiente de Gini, para aferir matematicamente a participação das aposentadorias e pensões na distribuição de renda da população brasileira a nível estadual e municipal.
- Modelar uma estrutura de *Data Warehouse* para armazenar os microdados oficiais disponibilizados pela CPI da Previdência de 2017, objetivando - a partir da utilização de técnicas de processamento em lotes - simular os efeitos que a PEC 06/2019 causaria no cenário previdenciário entre 1995 a 2016, se a mesma já estivesse em vigor no momento em que cada benefício foi concedido.
- Aplicar estratégias e conceitos de ciência de dados para o tratamento, organização e processamento dos microdados utilizados, visando possibilitar a extração de conhecimento válido com simulações ou aplicação de medidas estatísticas acerca dos resultados gerados em cada análise proposta.

1.3 Organização da Dissertação

Este trabalho está organizado da seguinte forma:

- Capítulo 1: apresenta uma breve introdução do assunto abordado, a motivação para realização dessa pesquisa e os objetivos, geral e específicos, propostos por essa dissertação.
- Capítulo 2: apresenta a fundamentação teórica que norteou essa pesquisa, destacando os conceitos de análise de dados, ciência de dados e dados abertos, além de elencar os principais modelos matemáticas utilizados nas análises propostas.
- Capítulo 3: elenca os principais trabalhos correlatos a essa pesquisa, apontando as contribuições desta dissertação perante ao estado da arte da problemática abordada.

- Capítulo 4: apresenta o primeiro estudo de caso desta dissertação, responsável por avaliar a participação que as aposentadorias e pensões possuem na distribuição de renda da população brasileira.

- Capítulo 5: apresenta o segundo estudo de caso desta dissertação, responsável por simular os impactos que a reforma previdenciária prevista pela PEC 06/2019, avaliando o período de 1995 a 2016, causaria acerca das concessões de benefícios na hipótese da mesma já está em vigor quando cada aposentadoria foi concedida.

- Capítulo 6: apresenta as considerações finais desta pesquisa e as perspectivas de trabalhos futuros acerca dos estudos realizados.

2 Referencial Teórico

2.1 Considerações Iniciais

Este capítulo tem como finalidade conceituar teoricamente as aplicações e técnicas que fundamentaram a realização deste trabalho. Serão abordadas as definições de análise de dados e ciência de dados, destacando suas importâncias perante o cenário cientifico e tecnológico, além de conceituar o seu emprego na elaboração e organização de conjuntos de dados como os que nortearam o presente estudo. Também será abordada a importância das bases de dados governamentais disponibilizadas pelos órgãos públicos, finalizando o capítulo com uma fundamentação das técnicas matemáticas aplicadas nesta pesquisa.

2.2 Análise de Dados

Com a evolução de novas e mais modernas tecnologias, aliada ao advento da Internet e das mídias sociais, uma enorme quantidade de dados está sendo gerada e coletada diariamente a uma velocidade cada vez maior (PAL; MEHER; SKOWRON, 2015; GHANI et al., 2019). Celulares, notebooks, equipamentos hospitalares, automóveis, sensores, câmeras de segurança e diversos outros dispositivos eletrônicos presentes nos mais diversos segmentos da sociedade geram automaticamente informações que precisam ser armazenadas e muitas vezes processadas em tempo real. Entretanto, acompanhar, analisar e extrair conhecimento acerca desse fluxo de dados é uma tarefa substancialmente desafiadora, especialmente quando estes não estão em conformidade com as noções tradicionais de estrutura de dados (EMC, 2015).

Tal fenômeno, além de estar cada vez mais inserido no cotidiano da população, tem apresentado um crescimento exponencial. Segundo (BAESENS, 2014) aproximadamente 90% dos dados existentes no mundo foram criados nos últimos 2 anos[1]. Adicionalmente, o barateamento de ferramentas computacionais com maior poder de processamento e a onipresença da rede estão colaborando para que pessoas e empresas desenvolvam algoritmos acerca desses dados nas mais diversas áreas do conhecimento como: economia, engenharia, sociologia, saúde, marketing, dentre outras (PROVOST; FAWCETT, 2013; HAN; KAMBER; PEI, 2011)

Diante desse cenário, a utilização de técnicas computacionais para a aquisição inteligente de conhecimento acerca de dados sem representatividade passou a ser uma tarefa cada vez mais atrativa, uma vez que muda a forma como tais dados são anali-

[1] Informações referentes ao ano de 2016.

sados e contextualizados (REZENDE, 2003). Desse modo, Mineração de Dados (MD), Extração de Conhecimento de Base de Dados (mais conhecida como *Knowledge-Discovery in Database* – KDD) e Inteligência de Negócios (do inglês *Business Inteligence* – BI) são exemplos de métodos ou estratégias que buscam extrair informações e identificar padrões/tendências para a resolução de problemas específicos acerca de conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; BAESENS, 2014).

Embora diversos autores interpretem os processos KDD e MD como sinônimos (REZENDE, 2003; GORUNESCU, 2011), boa parte da literatura considera a mineração como uma etapa da descoberta de conhecimento, sendo essa composta pela seleção, pré-processamento, transformação, mineração de dados e interpretação (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAN; KAMBER; PEI, 2011). A Figura 1 elenca as etapas necessárias para a extração de conhecimento.

Figura 1 – Processo de descoberta de conhecimento em bases de dados.

Fonte: (REZENDE, 2003)

A primeira fase do KDD é compreendida pelo conhecimento do domínio, ou identificação do problema, responsável por, a partir de avalição do domínio da aplicação a qual os dados pertencem, compreender e avaliar o processo como um todo e definir as metas e objetivos a serem alcançados. A segunda etapa é de fundamental importância para a realização de uma extração de conhecimentos válidos e potencialmente uteis sendo denominada de pré-processamento. Esta demanda aproximadamente 80% do tempo total do processo (KLEMETTINEN, 1999), e é responsável pela limpeza, correção de dados que ausentes, errôneos ou inconsistentes e possíveis transformações, objetivando facilitar posteriores identificações de padrões. Posteriormente, no estágio de extração de padrões,

ocorre a aplicação de algoritmos de Mineração de Dados específicos para o reconhecimento de padrões ou tendências acerca do conjunto de dados avaliado. (HAN; KAMBER; PEI, 2011) destaca 4 tipos de técnicas de inteligência computacional para a realização dessas propostas:

- Classificação: prediz a qual categoria ou classe um determinado item do conjunto de dados pertence;
- Agrupamento: também conhecido como *clustering*, identifica grupos de itens com propriedades e características semelhantes sem pressupor classes especificas;
- Associação: identifica grupos de dados que apresentem correlaçãoão entre si;
- Regressão: utiliza uma função preditiva para mapear valores associados aos dados em um ou mais valores reais.

Perante aos diversos algoritmos de mineração de dados existente, a seleção da estratégia especifica a ser utilizada não é uma tarefa trivial. Algumas das técnicas descritas acima consiste na aplicação de um determinado algoritmo enquanto outras propõe a utilização de diversos métodos objetivando uma maior confiabilidade no resultado final. Portanto, para a seleção do algoritmo apropriado, deve-se levar em consideração o objetivo final da análise ponderando possíveis hipóteses (WU et al., 2008). Em seguida, é realizada a etapa de pós-processamento acerca dos resultados encontrados. O sucesso dessa etapa determina a eficácia de todo o processo de KDD, avaliando e validando as informações geradas afim de identificar possíveis falhas nas fases anteriores (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Observando a Figura 1, é possível observar que o fluxo do KDD é um processo recursivo, ou seja, caso o resultado encontrado seja incoerente ou não aceitável é possível reiniciar o mesmo, analisando todas as etapas em busca de melhorar e/ou refazer o que for necessário, até que o conhecimento resultante seja julgado como satisfatório.

De forma semelhante, o processo de BI tem como finalidade coletar, organizar, compartilhar e analisar grandes volumes de dados empresariais projetando a realização de estratégias em gestão de negócios que visem a obtenção de vantagens competitivas (RUD, 2009; CHOO, 1996). Tal procedimento, embora tenha uma metodologia própria, se assemelha bastante a técnica de KDD visto que ambos objetivam a extração de conhecimento válido acerca de conjuntos de dados (MAHESHWARI, 2015).

No entanto, devido a diversos fatores históricos, como o crescente acumulo de dados, aumento da interdisciplinaridade das problemáticas abordadas, e até mesmo necessidade de padronização de uma área específica que englobe as diferentes vertentes de análises de dados, houve o surgimento de um novo campo de estudo, conhecido como ciência de

dados (VANDERPLAS, 2016; PROVOST; FAWCETT, 2013; CAO, 2016; CAO, 2017), tema abordado na próxima seção.

2.3 Ciência de Dados

Ciência de dados, do inglês *data science*, corresponde a uma área interdisciplinar relacionada a análise de dados que objetiva a extração de conhecimento, e possíveis tomadas de decisões, acerca de problemáticas especificas (GRUS, 2016; BAESENS, 2014). Embora tal conceito seja similar aos de KDD e BI, a ciência de dados pode ser compreendida como um domínio de estudo mais abrangente, englobando todas as definições e técnicas computacionais abordadas anteriormente em um novo campo de pesquisa (VANDERPLAS, 2016).

Em relação a definição de ciência de dados, observa-se no estado da arte dessa temática uma dificuldade na sua conceituação formal. Para (CAO, 2016), por exemplo, essa área pode ser entendida como um campo científico que desenvolve tecnologias e teorias relevantes para dados, desde a sua extração, pré-processamento, representação, armazenamento, pesquisa, compartilhamento, segurança, modelagem, aprendizagem, análise e visualização, até a sua integração com recursos heterógeneos, complexos e interdependentes para tomadas de decisão e abstração de valores significativos.

Por outro lado, de acordo com (JOURNAL OF DATA SCIENCE, 2019), a ciência de dados compreende praticamente tudo que tenha alguma relação com dados, desde a sua coleta até possíveis análises e modelagens. Contudo, destaca-se como um dos seus requisitos básicos a necessidade de aplicações desses dados na resolução de problemas, dos mais simples aos mais complexos, utilizando como instrumentos para auxilio técnicas matemáticas, ferramentas computacionais e avaliações críticas acerca das problemáticas abordadas. Além disso, a criatividade e intuição para elaborar a visualização dos resultados, também conhecido como *data visualization*, caracterizam os estudos associados a ciência de dados (EMBARAK, 2018).

Outra abordagem para ciência de dados é compreendida pela agregação de três competências, ou áreas, básicas do campo científico: ciência da computação, matemática e estatística e a especialização cientifica da problemática em questão (VANDERPLAS, 2016). A Figura 2 ilustra um diagrama de *venn* contendo tais competências e suas inter-relações, enfatizando a ciência de dados como um novo campo de estudo sendo produto de todas as áreas relatadas.

A respeito da utilização de ciência de dados, a pesquisa de (ROGATI, 2020) destaca que, embora o emprego de sistemas avançados de inteligência artificial mostre-se promissores em estratégias de negócios, a maioria das empresas não encontram-se preparadas para implementar tal abordagem. De acordo com a autora, há uma hierarquia de necessidades,

Figura 2 – Competências fundamentas para Ciência de Dados.

Fonte: Adaptado de (VANDERPLAS, 2016)

nas quais, requisitos básicos necessitam ser atendidos antes de obter possíveis benefícios a partir da utilização de modelos de inteligência artificial. Contudo, destaca-se que a aplicação de análises estatísticas simples ou extração de características acerca dados sem representatividade podem ser compreendidas como aplicações da ciência de dados.

Embora as principais aplicações da ciência de dados estejam fortemente relacionadas ao setor econômico e financeiro, devido principalmente ao fato de ter nascido neste ambiente para a resolução da problemática de grande quantidade de dados geradas pelas empresas (PROVOST; FAWCETT, 2013), a utilização desta área nos desafios encontrados pela pesquisa científica tornou-se uma alternativa plausível em virtude, essencialmente, da sua capacidade adaptativa e interdisciplinar (VANDERPLAS, 2016).

2.4 Dados Abertos

O conceito de dados abertos, do inglês *open data*¸ faz referência a dados que podem ser utilizados, reutilizados ou redistribuídos livremente por qualquer pessoa, restringindo, no máximo, a atribuição à fonte original e ao compartilhamento das mesmas licenças em que as informações foram apresentadas (OPEN DEFINITION, 2019). Dessa forma, a abertura de dados objetiva restringir o controle e restrições existentes em dados que forem publicados, possibilitando a exploração dos seus conteúdos de forma livre por pessoas (sejam essas físicas ou jurídicas) (ISOTANI; BITTENCOURT, 2015).

Assim sendo, em (OPEN DATA HANDBOOK, 2019) o conceito de dados abertos

é compreendido por três normas fundamentais:

- Os dados devem ser disponibilizados em um meio acessível, preferencialmente na internet, em um formato compreensível por máquina de tal forma que este possa ser acessado e modificado.
- Os dados devem ser disponibilizados perante a termos que permitam sua reutilização e redistribuição, inclusive a combinando com outros conjuntos de dados.
- Todos os usuários devem ser capazes de utilizar, reutilizar e redistribuir os dados, sem haver discriminação com áreas de atuação, pessoas ou grupos.

Uma vez que o movimento de dados abertos é estimulado e que as três regras supracitadas são atendidas, é aberta a possibilidade que diferentes pessoas, organizações e sistemas trabalhem de forma colaborativa. Tal fator é justificado pela interoperação dos dados que foram abertos, ampliando a comunicação existente e potencializando o desenvolvimento de sistemas cada vez mais complexos (ISOTANI; BITTENCOURT, 2015).

É importante ressaltar que os dados abertos não correspondem somente aqueles divulgados pelo governo, podendo originar de qualquer instituição contanto que atenda os requisitos básicos citados anteriormente, não apresentando restrições de uso (MANYIKA et al., 2013). Existem casos de dados abertos disponibilizados por diversas industrias ou setores, como transporte, produtos, sustentabilidade, mapas, economia, cultura, finanças, negócios, entre outros. A Figura 3 apresenta alguns exemplos de tipos de dados abertos comumente encontrados na internet (OKF, 2020).

2.4.1 Dados Abertos Governamentais

O conceito de dados abertos governamentais está sendo cada vez mais disseminado, sobretudo após a divulgação de um memorando em 2008 do até então presidente dos Estados Unidos Barack Obama (OBAMA, 2008). Neste, é discutida a transparência de dados governamentais norte-americanos através da criação de uma plataforma web (USA, 2019) que disponibilizasse dados abertos para qualquer cidadão.

A abertura dos dados públicos objetiva, sobretudo, estimular que a sociedade desenvolva produtos, serviços e novas atividades econômicas utilizando essas informações. Além disso, destacam-se benefícios políticos relacionados, principalmente, a transparência governamental gerada através do acompanhamento das elaborações de políticas públicas. Tal fator, além de intensificar a responsabilidade e participação democrática dos cidadãos, aumenta a confiabilidade transmitida pelo governo (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Figura 3 – Tipos de dados abertos

Fonte: Adaptado de (OKF, 2020)

Nesse cenário, o governo brasileiro apresenta-se como um dos fundadores do *Open Government Partnership* (OPG, 2019) em 2011, uma iniciativa internacional que visa a promoção de governos abertos através da capacitação de cidadãos, combate a corrupção e aproveitamento de novas tecnologias. Além disso, a política de dados abertos governamentais no Brasil é consolidada pela Lei n. 12.527 de 2011, conhecida como Lei de Acesso à Informação (LAI), a qual garante acesso a qualquer informação de interesse público, desde que esta não seja imprescindível a segurança da sociedade e do Estado (Brasil, 2011). De maneira similar aos Estados Unidos, no Brasil há uma plataforma de acesso aos dados abertos governamentais (Brasil, 2019), sendo um portal que disponibiliza informações públicas seguindo os princípios de dados abertos relatados anteriormente.

No entanto, embora os dados abertos governamentais apresentem impactos extremamente positivos no governo, na sociedade e na economia, evidenciam-se adversidades relacionadas a sua implantação. Nesse contexto, o trabalho de (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012) divide esses obstáculos em seis grandes grupos: institucionais, complexidade da tarefa, utilização, legislação, qualidade da informação e técnicas. Dessa forma, a estimulação para a participação da população no uso dessas fontes de informações

apresenta-se potencialmente promissora acerca das vantagens abordadas nessa subseção.

2.5 Técnicas Matemáticas Utilizadas

2.5.1 Decomposição do Índice de Gini

O índice de Gini, também conhecido como coeficiente de Gini, é uma medida de dispersão estatística elaborado pelo italiano Corrando Gini que objetiva mensurar o grau de desigualdade de uma variável acerca de um conjunto de dados (GINI, 1912; YITZHAKI; SCHECHTMAN, 2012). Esse indicador é fortemente utilizado em econômica para aferir o grau de desigualdade de renda da população, no entanto, é possível aplicá-lo em diversos campos de estudo como educação, ecologia, ciências da saúde, sociologia, dentre outros (THOMAS; WANG; FAN, 2001). Sua interpretação baseia-se em um número variado entre 0 e 1, no qual 0 corresponde a total igualdade (todos recebem a mesma renda) e 1 a total desigualdade (uma pessoa detém de toda renda e os demais nada) (GINI, 1912; HOFFMANN, 2006).

Matematicamente, assume-se uma série discreta $x_i(i, ..., n)$ correspondente a renda da i-ésima pessoa de uma população com n pessoas. Admite-se que as rendas estão ordenadas, sendo:

$$x_1 \leq x_2 \leq ... \leq x_i \leq ... \leq x_n \tag{2.1}$$

Agregando as pessoas da mais pobre até a i-ésima posição da série 2.1, a proporção acumulada da população é dada por:

$$p_i = \frac{i}{n} \tag{2.2}$$

e a respectiva para proporção acumulada de renda é:

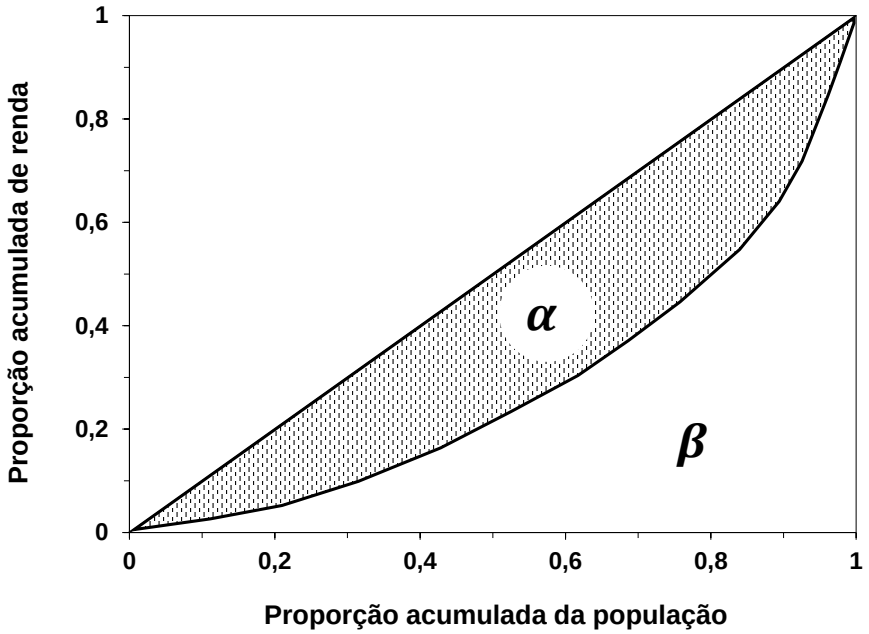
$$\varphi_i = \frac{1}{n\mu} \sum_{j=1}^{i} x_j \tag{2.3}$$

onde μ é a renda média, dada por:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.4}$$

A Curva de Lorenz é obtida pela relação entre os pares ordenados dos valores de p_i e φ_i, conforme indicado na Figura 4.

Figura 4 – Curva de Lorenz



Fonte: Adaptado de (YITZHAKI; SCHECHTMAN, 2012)

O coeficiente de Gini corresponde ao quociente da área α com a soma das áreas β e α, ou seja:

$$G = \frac{\alpha}{\alpha + \beta} = \frac{\alpha}{0,5} \tag{2.5}$$

Dessa forma pode-se reescrever a equação 2.5 como:

$$G = 1 - 2\beta \tag{2.6}$$

Similarmente, é possível considerar que a renda x_i é composta por k parcelas de rendimento de forma que:

$$x_i = \sum_{h=1}^{k} x_{hi} \tag{2.7}$$

onde x_{hi} representa o valor da h-ésima parcela de renda da i-ésima pessoa. Sendo assim, a média da h-ésima parcela é:

$$\mu_h = \frac{1}{n} \sum_{i=1}^{n} x_{hi} \tag{2.8}$$

e a proporção acumulada dessa parcela referente a série 2.1 é dado por:

$$\varphi_{hi} = \frac{1}{n\mu_h} \sum_{j=1}^{i} x_{hj} \tag{2.9}$$

Analogamente a equação 2.6 é possível definir a curva de concentração como sendo a relação dos pares ordenados de φ_{hi} e p_i. Ressalta-se que na construção da curva de concentração de x_{hi} é utilizada a ordenação de x_i e não a ordenação de x_{hi}, que pode ser diferente. Sendo assim, do mesmo modo que para o coeficiente de Gini, define-se a Razão de Concentração da parcela h (C_h) como:

$$C_h = 1 - 2\beta_h \tag{2.10}$$

onde β_h é a área entre a curva de concentração de x_{hi} e o eixo das abscissas.

Diante disso, (HOFFMANN, 2009) demonstra que o índice de Gini é definido pela média ponderada das razões de concentração das parcelas h. Então:

$$G = \sum_{h=1}^{k} \varphi_h C_h \tag{2.11}$$

como $\sum_{h=1} \varphi_h = 1$, pode-se reescrever a equação 1 como:

$$G = G - \sum_{h=1}^{k} \varphi_h (G - C_h) \tag{2.12}$$

com,

$$\pi_h = G - C_h \tag{2.13}$$

A equação 2.13 é conhecida como medida de progressividade, na qual $\pi_h > 0$ ($C_h < G$) simboliza uma parcela progressiva (contribuindo para o decréscimo do índice de Gini e diminuição da concentração de renda). De forma semelhante $\pi_h < 0$ ($C_h > G$) corresponde a uma parcela regressiva (contribuindo para o acréscimo do índice de Gini e aumentando a concentração de renda) (HOFFMANN, 2009).

2.5.2 Correlação de Variáveis

Em estatística, correlação compreender a interdependência existente entre duas variáveis, ou seja, o grau de relacionamento ou associação existentes entre as características de um dado. Em alguns casos essa relação é clara, como a massa corporal e a altura de uma pessoa, o consumo de combustível e a distância percorrida por um veículo ou o número de anúncios divulgados e a quantidade de produtos vendidos. Todavia, em muitos

casos essa interdependência não é intuitiva ou aparente, fazendo necessário, dessa forma, a utilização de estratégias matemáticas para aferir essa correlação (BUSSAB, 2017). Dentre os métodos mais comuns temos os índices de Pearson, Spearman e Kendall.

O coeficiente de correlação de Pearson, também conhecido como coeficiente de correlação produto-momento, é responsável por mensurar a interdependência, positiva ou negativa, existente entre duas variáveis. Essa estimativa, normalmente representada por ρ, assume valores entre -1 e 1, onde 1 corresponde a uma correlação perfeita positiva entre as variáveis, e -1 a uma correlação perfeita negativa, ou seja, se uma aumenta a outra sempre diminui. Por outro lado, quando o valor de ρ é igual 0 observa-se uma interdependência linear entre as variáveis, podendo, no entanto, existir uma dependência não-linear entre elas (BENESTY et al., 2009; MUKAKA, 2012).

O cálculo do coeficiente de correlação de Pearson pode ser realizado a partir da seguinte fórmula:

$$\rho = \frac{cov(X,Y)}{\sqrt{var(X).var(Y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.14}$$

onde, $x_1, x_2, ..., x_n$ e $y_1, y_2, ..., y_n$ correspondem aos valores assumidos por cada variável, e $\bar{x}$ e $\bar{y}$ são suas respectivas médias aritméticas.

A correlação de Pearson mensura a associação linear entre variáveis contínuas, estimando o quanto a relação dessas variáveis pode ser descrita por uma reta (BENESTY et al., 2009). Diante disso, pode ser interpretar o valor de ρ como:

- 0,9 a 1 positivo ou negativo indica uma correlação muito forte.
- 0.7 a 0.9 positivo ou negativo indica uma correlação forte.
- 0.5 a 0.7 positivo ou negativo indica uma correlação moderada.
- 0.3 a 0.5 positivo ou negativo indica uma correlação fraca.
- 0 a 0.3 positivo ou negativo indica uma correlação desprezível.

De maneira similar, o coeficiente de Spearman é comumente utilizado para aferir a interdependência existente entre variáveis aleatórias relacionadas monotonicamente[2], mas não de maneira linear. A forma deste coeficiente de correlação é a mesma do coeficiente de Pearson, mas as variáveis não correspondem mais aos valores originais coletados, substituindo-os por postos correspondes ao valor que cada variável assume em uma ordenação numérica (BUSSAB, 2017).

[2] Uma função entre dois conjuntos de dados é considerada monótona quando há a preservação da relação de ordem no seu domínio.

O coeficiente de correlação de Spearman pode ser aferido da seguinte forma:

$$\rho = 1 - \frac{6 \sum_{i-1}^{n} (X - Y)^2}{n(n^2 - 1)} \tag{2.15}$$

onde X e Y são as variáveis aleatórias variando de 1 a n, com n sendo a quantidade de elementos existentes no conjunto.

O coeficiente de Kendall, por sua vez, avalia a interdependência de postos ordenados, semelhante a correlação de Spearman. Essa medida é comumente eficiente para dados discretos, na qual (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) compreendem conjuntos de observações das variáveis aleatórias X e Y, tal que x_i e y_i sejam valores únicos. Qualquer par ordenado de observações, em que $i \neq j$, é considerado concordante se as classificações dos elementos concordarem entre si, ou seja, se $x_i > x_j$ e $y_i > y_j$, ou se $x_i < x_j$ e $y_i < y_j$. De maneira similar, eles serão discordantes se $x_i > x_j$ e $y_i < y_j$, ou se $x_i > x_j$ e $y_i < y_j$. Na hipótese de $x_i = x_j$ e $y_i = y_j$, o par não será nem concordante e nem discordante (KENDALL, 1938; BUSSAB, 2017). Dessa forma, é possível calcular o coeficiente de correlação de Kendall (τ) conforme a Equação 2.16:

$$\tau = \frac{(\textit{quantidade de pares concordantes}) - (\textit{quantidade de pares discordantes})}{n(n-1)/2} \tag{2.16}$$

É importante destacar que a correlação não indica necessariamente causa. Por exemplo, considere um estudante que teve boas notas nas disciplinas de química e física. O bom desempenho em química não pode ser necessariamente atribuído ao bom desempenho em física, mas ambas as notas podem ser explicas por um bom desempenho em matemática, fator não considerado no primeiro momento. Dessa forma, a correlação apresenta-se como uma importante ferramenta para análise de dados uma vez que, quando interpretada corretamente, possibilita a abstração de informações relevantes acerca de conjuntos de atributos interdependentes.

2.6 Considerações Finais

Nesse capítulo foi apresentada a fundamentação teórica que norteou o desenvolvimento desta dissertação, iniciando por uma contextualização acerca de análise de dados, a qual evidenciou o conceito de processo de extração de conhecimento (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996) – relatando a sua diferença em relação a mineração de dados (HAN; KAMBER; PEI, 2011) – e suas aplicações no cenário empresarial a partir do método de BI (CHOO, 1996; RUD, 2009). Com base nessas definições, foi realizada a conceituação de ciência de dados (PROVOST; FAWCETT, 2013; GRUS, 2016;

EMC, 2015), relatando sua relevância perante o atual cenário de produção de informações (BAESENS, 2014) e evidenciando sua capacidade adaptativa de trafegar entre as diversas áreas de conhecimento (VANDERPLAS, 2016).

Posteriormente, foi apresentado o conceito de dados abertos destacando sua importância nos mais diversos campos como ferramenta de acesso a informação para a sociedade (BAESENS, 2014). Nesse âmbito, enfatizou-se os dados abertos governamentais como mecanismo para a disponibilização de conteúdo público para a população, objetivando oferecer um sistema governamental mais transparente tanto no Brasil como em diversos países do mundo (OPEN DEFINITION, 2019).

Por fim, foram conceituadas algumas das técnicas matemáticas utilizadas nesta pesquisa. Primeiramente, fez-se uma definição acerca da decomposição do coeficiente de Gini (GINI, 1912), destacando sua utilização para aferir o grau de desigualdade de uma população acerca de uma determinada variável, como por exemplo a seu rendimentom e sua capacidade de mensurar o quanto cada parcela dessa variável influencia para diminuir ou aumentar a desigualdade existente. Em seguida, foram detalhados os conceitos de correlação entre variáveis interdependes, enfatizando sua eficiência para abstração de informação útil acerca de atributos em conjuntos de dados (BUSSAB, 2017).

3 Trabalhos Correlatos

3.1 Considerações Iniciais

O processo de extração de conhecimento válido acerca de bases de dados abertas públicas, seja pela ótica da ciência de dados ou análise de dados tradicional, apresenta-se como uma importante linha de pesquisa tanto na área da computação aplicada quanto nos demais campos interdisciplinares. Com isso, diversos trabalhos referentes a essa temática objetivam, de forma geral, compreender aspectos econômicos e sociais ou propostas governamentais baseando-se em simulações de cenários, análises estatísticas, aplicação de modelos de inteligência computacional e observações críticas acerca dos resultados, utilizando como objeto de estudos os dados relacionados a problemática abordada.

Diante disso, o presente capítulo apresenta um panorama geral dos estudos correlacionados a temática abordada nessa dissertação, explicitando os trabalhos julgados mais relevantes perante a problemática descrita no escopo desta pesquisa. Dessa forma, a Tabela 1 apresenta uma síntese acerca desses trabalhos correlatos, mostrando de maneira concisa as considerações realizadas acerca dos mesmos.

3.2 Correlatos

A aplicação da ciência ou análise de dados em dados abertos públicos apresenta-se promissora, e difusa, em meio aos diversos segmentos governamentais. Estudos utilizando os microdados da economia, educação e saúde pública, por exemplo, tendem a apresentar informações relevantes para a compreensão de alguns fenômenos sociais e podem servir como subsídio para auxiliar tomadas de decisões em gestão pública (SANTOS, 2019; BUZETI, 2017). Avaliando o estado da arte da previdência social brasileira, eixo temático desta pesquisa, evidenciam-se, no geral, trabalhos que objetivam compreender e discutir a sua contribuição social e a transparência dos modelos propostos pelo governo relativos à sustentabilidade do sistema.

Dessa forma, acerca dos impactos socioeconômicos causados pela previdência social nacional, o trabalho de (HOFFMANN, 2009) utilizou dados da PNAD de 2001 a 2007 para verificar quais variáveis de rendimento influenciam na redução ou aumento da desigualdade de distribuição renda – mensurada através da decomposição do índice de Gini, semelhante ao demonstrado na subseção 2.5.1 desta dissertação. Os resultados demonstraram que a parcela de aposentadorias e pensões, embora tenham colaborado para a redução do coeficiente de Gini com alterações ocorridas no decorrer dos anos, apresentava

regressividade, ou seja, colaboram para um aumento na concentração de renda. Além disso, destaca-se que embora o trabalho não seja recente, o mesmo apresentou importantes discussões que colaboraram para um melhor compreendimento da problemática abordada nessa pesquisa.

O trabalho de (NAKATANI-MACEDO, 2015), semelhante ao anterior, avalia a participação que os benefícios de aposentadorias e pensões, oficiais e não oficiais, apresentam na desigualdade de rendimento dos domicílios brasileiros. A autora utiliza os dados da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 2004 a 2017 e a metodologia de decomposição do índice de Gini, aplicando-os no Brasil e suas regiões, semelhante ao trabalho anterior. No entanto, seus resultados demonstraram que esses benefícios apresentam características diferentes quando divididos em categorias de salários, ou seja, as aposentadorias e pensões até um salário mínimo colaboram para a desconcentração de renda, enquanto as acima de um salário mínimo apresentavam característica oposta, contribuindo para aumentar o coeficiente de Gini.

Em (NAKATANI-MACEDO et al., 2015) foram utilizados os dados censitários, disponibilizados pelo Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES), dos anos de 1980, 1991, 2000 e 2010 para os 399 municípios paranaenses, e microdados de rendimento presentes nas PNADs de 1988 a 2012, objetivando compreender o processo de envelhecimento da população no estado e seus impactos socioeconômicos na distribuição de renda da população. O estudo demonstrou que no intervalo de 30 anos avaliado, a proporção de pessoas acima de 60 anos mais que triplicou, provocando um aumento paralelo no percentual de participação dos benefícios de aposentadorias e pensões na renda total, apontando, dessa forma, uma preocupação na formação de políticas públicas ao mercado de trabalho, sistema de previdência, saúde e lazer para o público analisado.

A pesquisa de (BARBOSA, 2015), por sua vez, propôs um modelo econométrico para compreender como a previdência social atua sobre as distribuições de renda regionais e estaduais no Brasil. Os resultados demonstraram que, em 2010, o sistema previdenciário foi um instrumento eficaz para a melhoria na distribuição de renda regional, mensurada através do índice de Gini, a qual apresentou impactos positivos diretos no PIB per capita brasileiro.

Em (SOLÓN, 2019) foram utilizados dados municipais relacionados a pagamentos de benefícios de aposentadorias, arrecadações e fundo de participação municipal, de 2017, para compreender o cenário previdenciário nacional. Mediante a isso, o estudo destaca a Previdência Social como principal instrumento assistencial para o combate à pobreza, relatando que, sem os benefícios pagos mensalmente aos pensionistas e aposentados, diversas áreas rurais dos municípios de pequeno porte estariam em situação de calamidade social.

Por outro lado, a literatura apresenta diversos trabalhos relacionados à sustentabi-

lidade do sistema previdenciário brasileiro e à transparência das políticas públicas que tangenciam essa temática. Com isso, o trabalho de (COSTANZI; ANSILIERO, 2017), por exemplo, buscou quantificar, usando um modelo de previsão a longo prazo, o impacto da demografia sobre as despesas da previdência social como proporção do Produto interno Bruto (PIB) para o Brasil. A pesquisa utilizou um modelo simples de projeção, demostrando que em 2060 as despesas do sistema compreenderiam aproximadamente 19% do PIB nacional. Entretanto, apesar da grande relevância dos resultados, o modelo é útil apenas para ilustrar os efeitos do envelhecimento da população em relação às despesas previdenciárias, não sendo apropriado para investigações acerca de propostas de reforma.

Em (SILVA et al., 2017), os autores, motivados pelas discussões políticas acerca da reforma da previdência prevista na PEC 287/16, investigaram os resultados divulgados pelo governo federal, os quais foram utilizados para justificar a proposta de política pública. O trabalho apontou que os dados oficiais apresentavam projeções tendenciosas a curto prazo e com erros elevados a longo prazo, uma vez que tais estimativas não consideravam intervalos de confiança. Além disso, a pesquisa relatou a impossibilidade de reprodução dos resultados divulgados pela Lei de Diretrizes Orçamentarias (LDOs), devido, principalmente, à falta de transparência nesses documentos oficiais, tanto na sua metodologia quanto nos dados utilizados.

A pesquisa de (COSTANZI et al., 2018) apresenta um panorama detalhado referente a atual situação da Previdência Social brasileira, elencando os principais pontos previstos pela PEC 287/16 e relatando os fatores econômicos utilizados para justifica-la. Os autores classificam a proposta citada como imprescindível para a garantia da sustentabilidade fiscal do país, uma vez que o acelerado envelhecimento populacional eleva os custos excessivos de financiamento para as gerações futuras. Semelhantemente, o trabalho de (GIAMBIAGI; PINTO; ROTHMULLER, 2018) afirma que na ausência de uma reforma, será praticamente impossível respeitar o teto do gasto público até o ano de 2026, último ano de vigência antes de uma possível revisão. Diante disso, o estudo discute a PEC 287/16 e outras duas reformas, sendo uma elaborada pelos autores, demonstrando que, caso aprovado, um reajuste nas diretrizes do sistema previdenciário possibilitaria um controle no teto de gastos, proporcionando condições para a continuidade e crescimento do programa.

Por outro lado, em (BASTOS et al., 2019) foram identificadas falsificações, ou possíveis erros, nos resultados utilizados pelo governo federal para justificar a reforma previdenciária PEC 06/2019. Utilizando as planilhas oficiais do Ministério da Economia referentes a Reforma da Previdência, obtidas através da LAI, foram refeitos os cálculos, demonstrando que os benefícios de aposentadorias dos trabalhadores mais pobre diminuem com a reforma proposta. Além disso, notou-se que as aposentadorias obtidas por tempo de contribuição – principal alvo da reforma –, pelas regras atuais, com idades mais novas, financiam os benefícios de menor valor dos trabalhadores que se aposentam com idade

mais avançada, fator que, possivelmente, colabora para redução da desigualdade social.

É importante destacar que, embora tenham participação direta na construção desta dissertação, nenhum dos trabalhos citados acima investigou a participação que as aposentadorias e pensões têm na distribuição de renda municipal do Brasil, e nem os efeitos que a uma reforma previdenciária causaria na concessão desses benefícios, principais objetivos desta dissertação. Dessa forma, objetivando esclarecer de forma resumida a abordagem de cada pesquisa, a Tabela 1 apresenta uma síntese dos trabalhos correlatos a esta pesquisa ressaltando as principais lacunas identificadas.

Tabela 1 – Síntese dos principais trabalhos correlatos.

Referência	**Objetivos**	**Lacunas**
(HOFFMANN, 2009)	Avaliar o impacto que as aposentadorias e pensões causam na distribuição de renda do Brasil entre 2001 e 2007	Concentra-se em uma avaliação nacional, não abordando unidades territoriais menores
(NAKATANI-MACEDO, 2015)	Avaliar o impacto que as aposentadorias e pensões causam na distribuição de renda do Brasil e regiões entre 2004 e 2014	Concentra-se em uma avaliação nacional e estadual, não abordando unidades territoriais menores
(NAKATANI-MACEDO et al., 2015)	Avaliar o impacto que as aposentadorias e pensões causam na distribuição de renda do estado do Paraná e regiões entre 1980 e 2010	Concentra-se em uma avaliação isolada de apenas um estado brasileiro
(BARBOSA, 2015)	Propor um modelo econométrico que compreenda como as aposentadorias implicam na distribuição de renda das regiões e estados brasileiros	Semelhante aos anteriores, não aplica uma avaliação acerca de unidades territoriais menores
(SOLÓN, 2019)	Utiliza dados previdenciários municipais de 2017 para avaliar a situação do sistema em cada cidade	Não estima o impacto que os benefícios causam na distribuição de renda da população de cada município
(COSTANZI; ANSILIERO, 2017)	Aplica um modelo de previsão a longo prazo para avaliar a a situação da Previdência Social em 2060	Não considera os impactos sociais que uma reforma provocaria e suas projeções não apresentam intervalos de confiança

(SILVA et al., 2017)	Recalcular as projeção realizadas pelo governo para verificar se os argumentos utilizadas para justificar a PEC 287/16 eram válidos	Não avaliou os possíveis impactos sociais que reforma a reforma proposta causaria
(COSTANZI et al., 2018)	Elencar os principais pontos da PEC 287/16 mediante a atual situação do sistema previdenciário nacional	Não pondera os possíveis pontos negativos que a proposta acarretaria no cenário social dos trabalhadores
(GIAMBIAGI;PINTO; ROTHMULLER, 2018)	Avaliar a PEC 287/16, e outras duas reformas, como alternativas para controlar os gastos públicos do sistema previdenciário nacional	Semelhante ao anterior, desconsidera os impactos sociais que as propostas poderiam causar
(BASTOS et al., 2019)	Recalcular os resultados divulgados pelo governo utilizados para justificar a reforma da previdência correspondente a PEC 06/2019	O estudo, embora enfatize os aspectos sociais negativos da reforma, não apresenta estimativas acerca do cenário previdenciário

Elaborado pelo autor.

3.3 Considerações Finais

Este capítulo apresentou os trabalhos correlatos que nortearam a realização dessa pesquisa. Avaliando o estado da arte da problemática abordada nessa dissertação, observa-se um conjunto de estudos referente à análise dos impactos socioeconômicos gerados pelo sistema previdenciário brasileiro, e à avaliação da sustentabilidade de propostas governamentais. Todavia, destaca-se que grande parte desses estudos não aborda nas suas metodologias as aplicações de técnicas, ou estratégias, de análise de dados, normalmente utilizadas para a abstração dos resultados discutidos.

Nesse contexto, os estudos de caso propostos por essa pesquisa realizam uma abordagem diferente da comumente encontrada na literatura. Além da avaliação crítica realizada acerca das análises propostas, será detalhado o procedimento metodológico utilizado - desde a extração dos dados sem representatividade, até a abstração de conhecimento válido acerca dos resultados gerados. Tal abordagem apresenta-se promissora, dentre outros fatores, por possibilitar a replicação do trabalho pela comunidade científica, proporcionando, dessa forma, o desenvolvimento de uma pesquisa colaborativa mais eficiente.

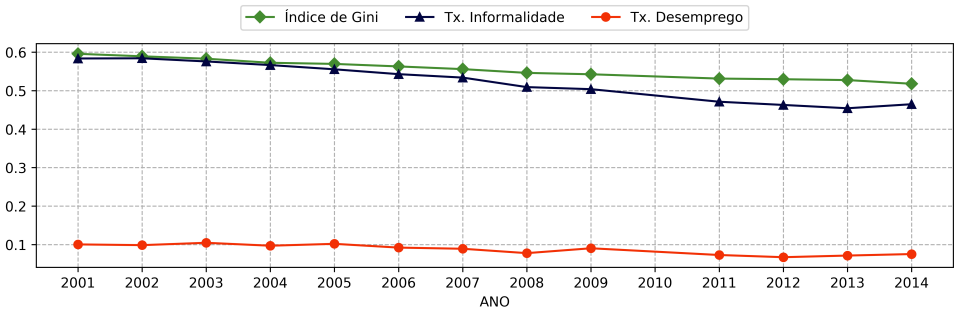
4 Análise da Participação das Aposentadorias e Pensões na Distribuição de Renda do Brasil, 2000 e 2010

4.1 Considerações Iniciais

A distribuição de renda, ou distribuição de riquezas, de um país é um tópico de inquestionável importância em todos os seguimentos da sociedade devido, principalmente, ao fato de estar diretamente relacionado com as condições de vida da população. Tal assunto apresenta-se em constante evidencia no cenário brasileiro uma vez que esse apresenta-se como um dos países mais desiguais do mundo (SOUZA; MEDEIROS, 2017; UNITED NATIONS, 2015). De acordo com dados da (THE WORLD BANK, 2019), o Brasil ocupa a 156ª posição[1] no ranking de índices de Gini de mundo, corroborando para que grande parte da população nacional tenha rendimentos insuficientes para atender suas necessidades básicas.

Entretanto, informações anuais divulgadas pela PNAD revelam que a desigualdade de renda nacional sofreu uma substancial redução nos últimos anos (IPEA, 2019) - ainda que se apresente muito elevada comparada a outros países do mundo. Paralelamente a esse fator, outros indicadores do mercado de trabalho, como taxas de informalidade e desemprego, demonstram-se regressivos nos últimos anos, apontando uma possível contribuição para a diminuição do índice de Gini registrado. A Figura 5 ilustra o comportamento dessas métricas no Brasil entre os anos de 2001-2014.

Figura 5 – Variação das taxas de formalidade e fesemprego e do índice de Gini no Brasil, 2001 a 2014



Fonte: Elaborado pelo autor a partir de dados contidos em (IPEA, 2019)

[1] Não há dados atualizados de índice de Gini em diversos países, indicando dessa forma que, possivelmente, o Brasil não ocupe exatamente a posição informada.

Paralelamente, os benefícios de aposentadorias e pensões compreendem uma importante parcela de rendimento da população, apresentando-se cada vez mais ativos na composição de renda dos domicílios brasileiros (HOFFMANN, 2009; PNAD, 2019). Esses benefícios, em sua grande maioria, correspondem a repasses garantidos pela Previdência Social, objetivando assegurar o rendimento dos contribuintes quando esses não detêm mais de suas capacidades laborais (MINISTÉRIO DA FAZENDA, 2018).

Diante desse fator, oberava-se na literatura uma divergência de pontos de vista relacionados a contribuição dessa parcela de benefícios para o aumento ou diminuição da distribuição de renda no Brasil. Alguns trabalhos como (FERREIRA, 2006; HOFFMANN; LEONE, 2004; NICHOLSON, 2007) indicam que as aposentadorias e pensões colaboram para o crescimento do coeficiente de Gini tanto em escala nacional quanto regional. No entanto, as pesquisas de (BAPTISTELLA; SOLANGE; FERREIRA, 2011; NAKATANI-MACEDO, 2015) demonstram que essa parcela de benefícios colabora para a redução da desigualdade de renda no Brasil.

Revisando o estado da arte relacionado a essa problemática, nota-se que grande parte dos estudos abstraem seus *insights* a partir dos microdados anuais divulgados pela PNAD. Entretanto, tal estratégia restringe as possibilidades de avaliações a serem realizadas em virtude dessa pesquisa corresponder a uma amostra domiciliar efetuada nas unidades federativas brasileiras, impossibilitando a flexibilidade para análises em unidades territoriais menores como municípios ou setores censitários. Uma avaliação municipal da proporcionalidade entre a distribuição de renda e os repasses de aposentadorias e pensões, apresenta-se como uma importante forma de compreender o real impacto que esses benefícios causam no coeficiente de Gini nacional. Tal importância destaca-se no trabalho de (SOLÓN, 2019), o qual compreende a previdência social como instrumento assistencial que combate à pobreza em diversos municípios brasileiros.

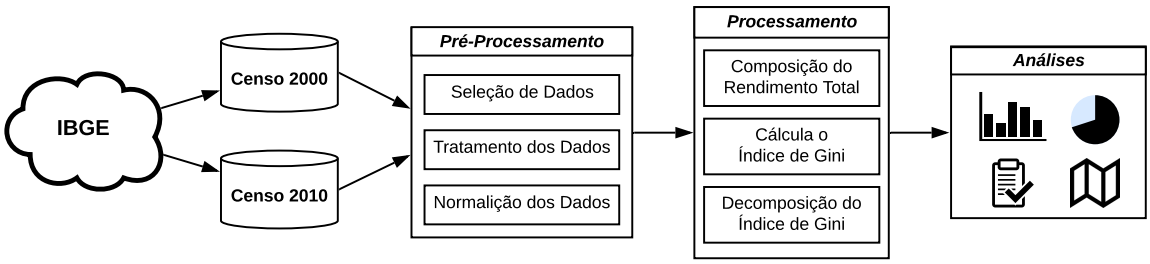
Diante disso, neste estudo de caso é realizada uma avaliação da participação que os rendimentos de aposentadorias e pensões tem na distribuição de riquezas dos estados e municípios do Brasil. Para tanto, optou-se por utilizar estratégias e conceitos de ciência de dados acerca dos microdados dos censos demográficos de 2000 e 2010, e a metodologia de decomposição do índice de Gini - conforme descrito no capítulo 2 - para aferir o grau de contribuição das parcelas desejadas. A adoção dessa métrica se deu pelo fato dela ser comumente utilizada na literatura para aferir o nível de desigualdade de renda de uma população, possibilitando comparações com resultados encontrados em outros estudos.

Dessa forma, esse capítulo foi dividido em 4 subseções incluindo esta introdução. Na subseção 4.2, será abordada a metodologia utilizada para o desenvolvimento das análises propostas, demonstrando detalhadamente os tratamentos e processamentos efetuados acerca dos dados. Na subseção 4.3 serão discutidos os principais resultados encontrados, finalizando o estudo na subseção 4.4 com as considerações finais.

4.2 Metodologia Utilizada

O procedimento metodológico para a realização do estudo proposto foi sistematizado mediante a uma sequência de etapas de análise de dados: extração, pré-processamento, aplicação de modelos matemáticos/computacionais e abstração de informações acerca dos dados de entrada. A Figura 6 ilustra o fluxograma de desenvolvimento do trabalho contendo cada uma das etapas realizadas, havendo um maior grau de detalhamentos acerca destas no decorrer dessa subseção.

Figura 6 – Fluxograma de desenvolvimento do primeiro estudo de caso



Fonte: Elaborado pelo autor

4.2.1 Origem dos Dados

Os microdados utilizados na execução das análises desse estudo foram extraídos dos censos demográficos realizados pelo IBGE nos anos de 2000 e 2010. As informações referentes aos censos anteriores, como por exemplo 1991, foram desconsiderados na pesquisa devido aos altos índices de inflação recorrentes na época (GIAMBIAGI; MOREIRA, 1999), gerando por consequência baixa confiabilidade nos resultados gerados.

O censo demográfico brasileiro trata-se de uma pesquisa nacional realizada cada dez anos, visando o reconhecimento das condições de vida da população brasileira em todas as grandes regiões, unidades federativas, mesorregiões, microrregiões, regiões metropolitanas, municípios, distritos, subdistritos e setores censitários, tendo como objeto de coleta as pessoas residentes (na data de referência) em domicílios presentes no território nacional (IBGE, 2011). No site do censo[2] é possível encontrar os principais resultados, estatísticas, tabelas e publicações realizadas pelo IBGE. Entretanto, o instituto também fornece os microdados e documentação referentes a cada pesquisa realizada, possibilitando que a comunidade acadêmica, empresas ou até mesmo pessoas físicas realizem pesquisas acerca destes.

Os microdados disponibilizados pelas bases são divididos em 3 grupos de informações, os quais abrangem pessoas, famílias e domicílios. Essa classificação possibilita

[2] Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9662-censo-demografico> Acesso em: 20 Dez, 2019

estudos estatísticos acerca de qualquer uma dessas categorias de forma isolada, permitindo análises com alto grau de detalhamento, dependendo da dimensão da problemática em questão. O presente trabalho norteou-se a partir das classes pessoas e domicílios, com recortes espaciais para as unidades federativas e municípios brasileiros.

O censo demográfico difere-se da PNAD, principalmente, em abrangência territorial e quantidade de informações exploradas. Enquanto a primeira pesquisa realiza uma investigação profunda em todo território nacional sobre as características socioeconômicas[3] da população, a segunda propõe uma apuração domiciliar mais limitada[4], restringindo apenas ao grau de representatividade necessária para os seus resultados acerca dos diversos níveis geográficos[5] definidos na sua divulgação.

4.2.2 Pré-Processamento dos Dados

Os censos demográficos foram carregados em um Sistema Gerenciador de Banco de Dados (SGBD) objetivando facilitar as consultas acerca dos microdados em questão. Constatou-se que a base de dados do ano 2000 possuía mais de 280 variáveis relacionadas à domicílios, famílias e pessoas, e que a de 2010 ultrapassava 350 variáveis correspondentes a domicílios, pessoas, mortalidade e emigração. Dessa forma, foi realizada uma etapa preliminar de seleção de atributos com o intuito de filtrar somente as informações de interesse para as posteriores análises. A Tabela 2 exibe as 7 variáveis utilizadas no presente estudo e suas respectivas descrições.

Tabela 2 – Variáveis selecionadas dos censos demográficos e suas descrições

Código das Variáveis		**Descrição**
2000	**2010**	
v0102	v0001	Código do estado
v0103	v0002	Código do município
v7616	v6529	Rendimento mensal domiciliar
v4614	v6527	Rendimento mensal per capita
v4573	v6591	Rendimento dos benefícios de aposentadorias e pensões
-	v0656	Aposentadorias e pensões da Previdência Social
p001	v0010	Peso de cada registro

Fonte: Elaborado pelo autor a partir dos metadados dos censos demográficos de 2000 e 2010

[3] Resumidamente, o censo demográfico aborda os seguintes temas: características dos domicílios, identificação étnico-racial, nupcialidade, núcleo familiar, religião, deficiência, migração, educação, trabalho, rendimento e mortalidade.

[4] Os temas abordados na PNAD são: habitação, características gerais dos moradores, características do trabalho, rendimentos e educação.

[5] A PNAD abrange o Brasil, Grandes Regiões, Unidades da Federação, Regiões Metropolitanas que contêm Municípios das Capitais, Municípios das Capitais e Região Integrada de Desenvolvimento da Grande Teresina.

A seleção dessas variáveis norteou-se com base nas informações necessárias para mensurar o grau de desigualdade e calcular a decomposição do índice de Gini (conforme demonstrado na subseção 2.5.1) acerca das unidades territoriais de interesse. Observando a Tabela 2, nota-se a presença de um atributo extra na base do censo demográfico do ano de 2010, tal fator é explicado pela necessidade de uma variável auxiliar para designar a qual programa social cada benefício pertence. Adicionalmente foram utilizados os valores de salário mínimo registrados nos meses de abril de ambos os anos abordados, sendo 151 e 510 reais para 2000 e 2010 respectivamente.

Posteriormente, foi realizada uma etapa de tratamento de dados objetivando corrigir erros e possíveis inconsistências existentes nas bases. Devido ao fato de ser uma pesquisa com elevado grau de importância e responsabilidade, foram identificadas poucas anomalias que comprometessem as análises realizadas. No entanto, foi necessária a substituição dos dados ausentes de rendimentos de aposentadorias e pensões por valores zeros (casos específicos de pessoas que não gozavam desses benefícios) e a correção dos pesos que apresentavam escalas erradas em estados e cidades especificas.

Além disso, em decorrência do elevado dinamismo da malha territorial brasileira (IBGE, 2019), constatou-se uma divergência no número de municípios existente em cada base de dados, registrando 5.507 cidades para o ano de 2000 e 5.564 para 2010. Dessa forma, houve a necessidade de normalizar o número de cidades para ambos os anos adotando como estratégia a criação de Área Mínimas Comparáveis (AMC), as quais são definidas como sendo áreas geográficas com a máxima unidade territorial desagregada, possibilitando dessa forma a comparação entre dois pontos diferentes no tempo (EHRL, 2017; SILVA; BACHA, 2011). Sendo assim, define-se cada AMC como:

$$AMC(p, r, b, w) = \sum_{i}^{n} mun_i(p, r, b, w), \{\forall n \in N\} \tag{4.1}$$

onde, p corresponde à população, r aos rendimentos, b os valores dos benefícios de aposentadorias e pensões e w os pesos de cada município n. Sendo N cada subconjunto de municípios que foram desmembrados em relação aos originais do ano 2000.

4.2.3 Processamento

Após a realização dos estágios de seleção, tratamento e normalização dos microdados, foi possível efetuar a etapa de processamento, sendo essa responsável por mensurar os valores do coeficiente de Gini, o percentual de participação das aposentadorias e pensões e o impacto que esses benefícios causam na distribuição de renda da população acerca das unidades territoriais de interesse Para isso, foi adotada a metodologia abordada na subseção 2.5.1, utilizando como parâmetros das equações as variáveis contidas na Tabela 3.

Tabela 3 – Parâmetros de indexação e notação geral.

Variável	Significado	Valor
k	estado	São Paulo, Rio de Janeiro, Pará, etc.
i	cidade	Fortaleza, Belém, Curitiba, etc.
r	renda per capita	R$ 0, ... , R$ 1.000, ...
d	renda domiciliar	R$ 0, ... , R$ 1.000, ...
b	renda aposentadoria/pensão	R$ 0, ... , R$ 1.000, ...
s	salário mínimo	R$ 151 ou R$ 510
t	ano	2000 ou 2010
p	peso	0, ... , 0.01, ... , 1.10, ...

Fonte: Elaborado pelo autor.

Para o cálculo do índice de Gini optou-se por utilizar a renda per capita domiciliar conforme é retratado na literatura (HOFFMANN, 2009; BAPTISTELLA; SOLANGE; FERREIRA, 2011; NAKATANI-MACEDO, 2015; HOFFMANN; LEONE, 2004; THE WORLD BANK, 2019; VANDERPLAS, 2016). Dessa forma, o grau de desigualdade dos estados e municípios brasileiros ($Desigualdade$) pode ser interpretado pela Equação 4.2, sendo:

$$Desigualdade(t, k, i) = Gini(t, k, i, d, p), \{\forall d > 0\} \tag{4.2}$$

onde $Desigualdade$ corresponde ao coeficiente de Gini, dos domicílios com rendimento, para cada município i e estado k nos anos t, conforme descrito em 2.6.

Posteriormente, para mensurar o grau de participação das aposentadorias e pensões na composição de renda da população, calculou-se o quociente entre os valores desses benefícios e o total de rendimentos per capita de cada unidade territorial k e i, e ano t. Dessa forma, pode-se entender o grau de participação ($PartBeneficio$) como sendo:

$$PercBeneficios(t, k, i) = \frac{\sum\limits_{N} b(n, p, t, k, i)}{\sum\limits_{M} r(m, p, t, k, i)}, \{n \subset N \ \ \& \ \ m \subset M\} \tag{4.3}$$

sendo N e M os conjuntos universos de valores de renda de aposentadorias/pensões e rendimentos per capita, respectivamente.

Por fim, aferiu-se o impacto que esses benefícios de aposentadorias e pensões causam acerca da variável $Desigualdade$ utilizando a metodologia de decomposição do índice de Gini para cada cidade i e estado k nos anos t, semelhante a equação 4.2. Sendo assim,

estimou-se a razão concentração ($Concentracao$) desses benefícios como:

$$Concentracao(t, k, i) = C_h(t, k, i, b, p), \{\forall b > 0\} \tag{4.4}$$

onde C_h corresponde a Equação descrita em 2.10.

Dessa forma, a medida de progressividade (π) pode ser compreendida por:

$$\pi(t, k, i) = Desigualdade(t, k, i) - Concentracao(t, k, i) \tag{4.5}$$

Paralelamente, utilizando os conceitos das Equações 4.4 e 4.5, calculou-se também as medidas de progressividade para os benefícios até 1 salário mínimo e acima de 1 salário mínimo acerca de cada unidade salarial s.

Para a realização dos cálculos propostos, foi utilizada a linguagem de programação *Python 3*[6], devido a sua facilidade para trabalhar com grandes volumes de dados, e as bibliotecas de análise dados *Pandas*[7], *Numpy*[8] e *Scikit-Learn*[9]. Além disso, foi utilizado o *software* QGIS 2.18[10] para a elaboração de mapas temáticos objetivando proporcionar uma melhor visualização acerca dos resultados encontrados. O projeto está disponibilizado na plataforma do GitHub[11] sob uma licença GNU *General Public License* (GPL) v3.0, contendo todo o procedimento para a reprodução das análises realizadas, tais como documentação, *scripts*, arquivos e referências necessárias para que a sua replicação e possíveis contribuições possam ser realizadas pela comunidade científica. A Tabela 4 descreve a estrutura do projeto e os principais arquivos existentes.

[6] Disponível em: <https://www.python.org/> Acesso em: 05 Jan, 2020
[7] Disponível em: <https://www.pandas.pydata.org/> Acesso em: 05 Jan, 2020
[8] Disponível em: <https://www.numpy.org/> Acesso em: 05 Jan, 2020
[9] Disponível em: <https://www.scikit-learn.org/stable/> Acesso em: 05 Jan, 2020
[10] Disponível em: <https://www.qgis.org/pt_BR/site/> Acesso em: 05 Jan, 2020
[11] Disponível em: <https://github.com/jralbbuquerque/censo-beneficio> Acesso em: 05 Jan, 2020

Tabela 4 – Descrição dos principais arquivos do projeto no GitHub referente ao primeiro estudo de caso

requirements.txt	arquivo de texto descrevendo todas as dependências e pacotes Python necessários para a execução do projeto.
dataset	pasta na qual serão salvos todos os resultados (arquivos .xlsx) da execução do projeto.
datasrc	pasta contendo todos os dados necessários para a execução das análises.
util	pasta com os diversos módulos/métodos para os cálculos dos indicadores.
metadata	pasta contendo os dicionários e documentações das bases de dados, disponibilizados no site do IBGE.
maps	pasta com os arquivos e dados de entrada do software QGIS necessários para a elaboração dos mapas.
analisys	pasta com as diversas análises realizadas utilizando a *framework jupyter notebook*.
main.py	arquivo principal para a execução do projeto.
README.md	descrição do funcionamento do projeto.

Fonte: Elaborado pelo autor.

4.3 Resultados e Discussões

O Rendimento Domiciliar per capita (RD), utilizado para os cálculos do trabalho, é referente a renda total, soma de todas as parcelas investigadas nos censos demográficos, dos domicílios registrados nas bases de dados. Diante disso, a Tabela 5 exibe os valores de índice de Gini (G) dos estados brasileiros em 2000 e 2010, e seus respectivos valores de renda média per capita. Nesta também é possível verificar as variações ocorridas entre os anos investigados, possibilitando uma visualização nas alterações de distribuições de renda dos estados brasileiros para o intervalo analisado. Os dados equivalentes a renda mensal de 2000 foram convertidos para a mesma unidade monetária de 2010, levando em consideração o Índice Nacional de Preços ao Consumidor (INCP), deflacionando todos os valores para outubro de 2010.

Tabela 5 – Índice de Gini (G) e renda média domiciliar (RD), Brasil 2000 e 2010.

	AC	AL	AM	AP	BA	CE	DF	ES	GO
Gini (G) - 2000	0,6233	0,6615	0,6470	0,6126	0,6437	0,6537	0,6226	0,5947	0,6164
Gini (G) - 2010	0,6133	0,6087	0,6355	0,5924	0,6071	0,5996	0,6185	0,5632	0,5603
Variação Gini (Δ%)	-1,61%	-7,98%	-1,77%	-3,29%	-5,70%	-8,28%	-0,67%	-5,30%	-9,10%
RD (R$) - 2000	1521,24	1190,33	1669,78	1978,99	1291,75	1295,35	4376,14	2088,34	2003,57
RD (R$) - 2010	1875,25	1533,41	2182,59	2419,99	1626,95	1576,03	5393,59	2496,47	2456,14
Variação RD (Δ%)	23,27%	28,82%	30,71%	22,28%	25,95%	21,67%	23,25%	19,54%	22,59%
	MA	MG	MS	MT	PA	PB	PE	PI	PR
Gini (G) - 2000	0,6390	0,6035	0,6221	0,6281	0,6311	0,6285	0,6527	0,6425	0,5976
Gini (G) - 2010	0,6071	0,5565	0,5605	0,5636	0,6032	0,5982	0,6195	0,6041	0,5384
Variação Gini (Δ%)	-4,99%	-7,80%	-9,91%	-10,27%	-4,41%	-4,83%	-5,09%	-5,98%	-9,91%
RD (R$) - 2000	992,43	2027,05	2051,77	2137,66	1552,25	1196,18	1447,89	1088,33	2244,49
RD (R$) - 2010	1375,88	2337,88	2459,68	2378,52	1725,3	1588,88	1729,64	1487,87	2707,04
Variação RD (Δ%)	38,64%	15,33%	19,88%	11,27%	11,15%	32,83%	19,46%	36,71%	20,61%
	RJ	RN	RO	RR	RS	SC	SE	SP	TO
Gini (G) - 2000	0,5964	0,6393	0,6022	0,6036	0,5774	0,5539	0,6396	0,5834	0,6411
Gini (G) - 2010	0,5955	0,5931	0,5669	0,6158	0,5400	0,4918	0,6110	0,5632	0,5949
Variação Gini (Δ%)	-0,14%	-7,23%	-5,86%	2,01%	-6,49%	-11,20%	-4,47%	-3,46%	-7,21%
RD (R$) - 2000	2730,73	1428,89	1808,49	1934,10	2334,21	2438,09	1306,36	3065,76	1384,11
RD (R$) - 2010	2980,77	1849,94	2145,43	2194,02	2734,68	2984,14	1754,60	3278,10	1955,62
Variação RD (Δ%)	9,16%	29,47%	18,63%	13,44%	17,16%	22,40%	34,31%	6,93%	41,29%

Fonte: Elaborado pelo autor.

É possível observar uma redução no índice de Gini para todos os estados, evidenciando uma diminuição na concentração de renda em todo o território brasileiro, com um decréscimo no indicador de aproximadamente -5,59% em escala nacional. Destacam-se positivamente os estados de Goiás (GO), Mato Grosso (MT), Mato Grosso do Sul (MS), Santa Catarina (SC) e Paraná (PR) com uma redução percentual acima de 9% para o índice. Além disso, houve um evidente acréscimo nas rendas médias domiciliares estaduais, com variação percentual média de aproximadamente +22,84%, fator essencial e determinante para a redução da desigualdade indicada pelo coeficiente de Gini.

Ainda que haja uma evidente melhora na redução do coeficiente de Gini para o período analisado em todo território brasileiro, é notória a heterogeneidade de concentração de renda existente entre os estados. Por exemplo, em 2010 foram reportados índices de Gini para o Brasil e Argentina de aproximadamente 0,525 e 0,430 respectivamente, contabilizando uma diferença em torno de 0,095 entre os países (THE WORLD BANK, 2019). Ao realizarmos a mesma análise nos estados do Amazonas (AM) e de Santa Catarina (SC), observamos uma diferença de 0,144 para ano de 2010 (valor acima da diferença dos índices de Gini entre Brasil e Argentina), evidenciando que a dispersão encontrada entre os estados brasileiro é equivalente a existente entre países com economias distintas (NAKATANI-MACEDO, 2015).

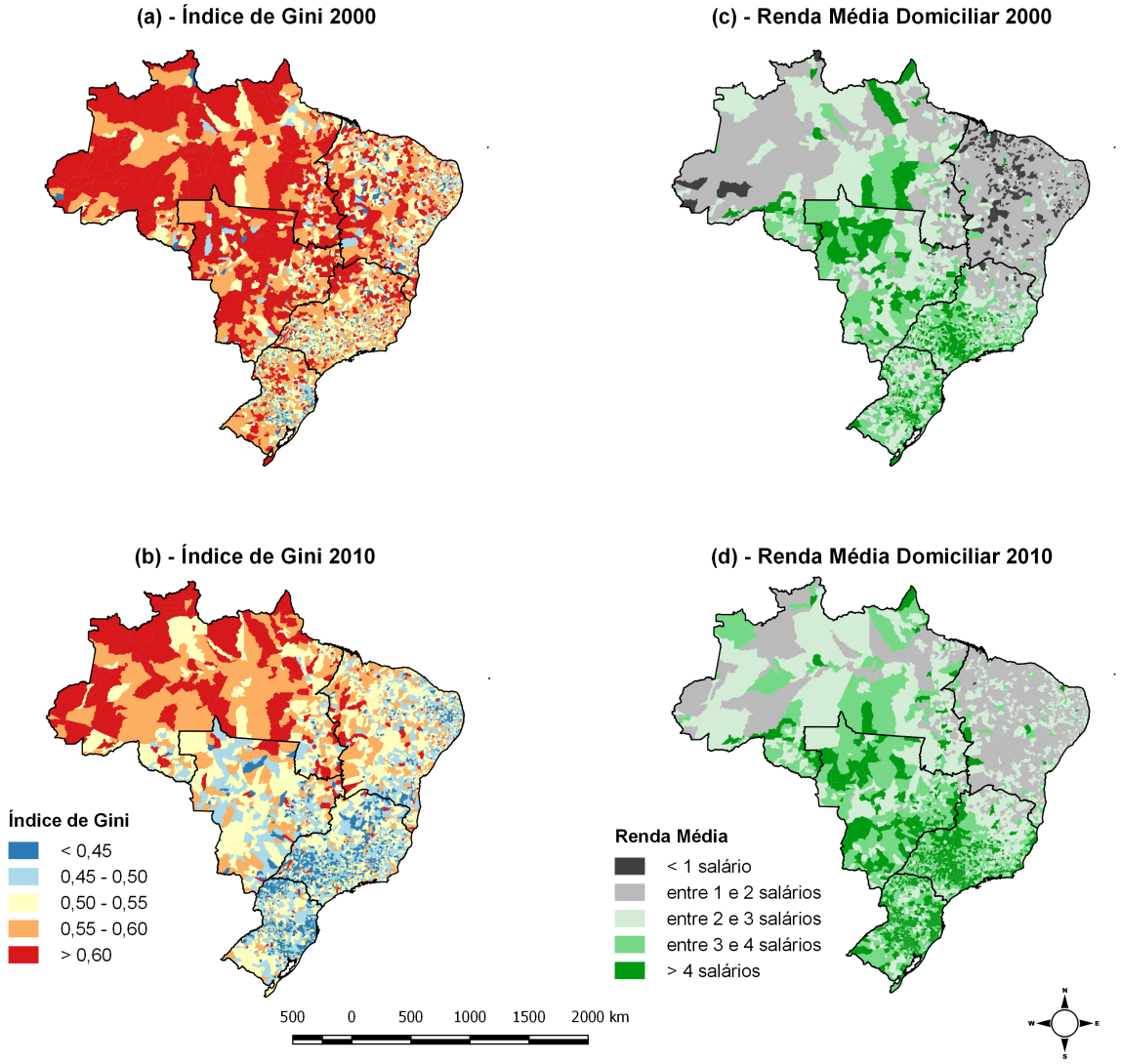
Na Figura 7 é ilustrada a situação do coeficiente de Gini no Brasil com recorte geográfico para cidades e seus rendimentos médios domiciliares. Nas subfiguras 7(a) e 7(b) são exibidas as dispersões da variável acerca dos municípios nos anos de 2000 e 2010 respectivamente, enfatizando a evidente diminuição do parâmetro. Destacam-se positivamente as regiões Centro-Oeste, Sul, Nordeste e Sudeste com consideráveis reduções municipais no indicador. Embora tenha ocorrido uma redução na concentração de renda dos estados e cidades brasileiras, é importante salientar que os valores calculados permaneceram bem acima do desejado quando comparados a maioria dos países europeus e asiáticos ($G < 0,40$) (THE WORLD BANK, 2019).

Do mesmo modo, as subfiguras 7(c) e 7(d) ilustram os rendimentos médios domiciliares (valores deflacionados para outubro de 2010 pelo INCP) dos municípios brasileiros nos anos de 2000 e 2010. Acerca desses mapas, é evidente o contraste existente no país, no qual cidades das regiões sul, sudeste e centro-oeste apresentam índices de renda média superiores aos demais municípios do norte e nordeste.

Quantificando esses resultados, aproximadamente 82% das cidades tiveram redução no coeficiente, sendo que 25% dessas tiveram uma diminuição acima de 0,1 para o indicador. Observando as subfiguras 7(a) e 7(b) ressalta-se que esses dados de redução não apresentaram tanto impacto nos municípios da região norte em relação as demais regiões. Além disso, os índices de rendimento médio domiciliar para esta região se mantiveram baixo em ambos os anos analisados, evidenciando dessa forma a necessidade de atenção especial para essas cidades no âmbito de políticas públicas.

Analisando a região nordeste, é notória a redução do indicador acerca dos seus municípios. Todavia, realizando uma avaliação comparativa entre índice de Gini e rendimento médio domiciliar, observa-se que embora a região apresente uma melhora na métrica, seus valores de renda média apresentam-se baixos, ou seja, embora a distribuição de renda apresente aparência igualitária a mesma é nivelada por valores baixos em relação a cidades das regiões Sul, Sudeste e Centro-Oeste.

Figura 7 – (a) e (b): Índice de Gini das cidades brasileiras. (c) e (d): Rendimento médio domiciliar dos municípios brasileiros

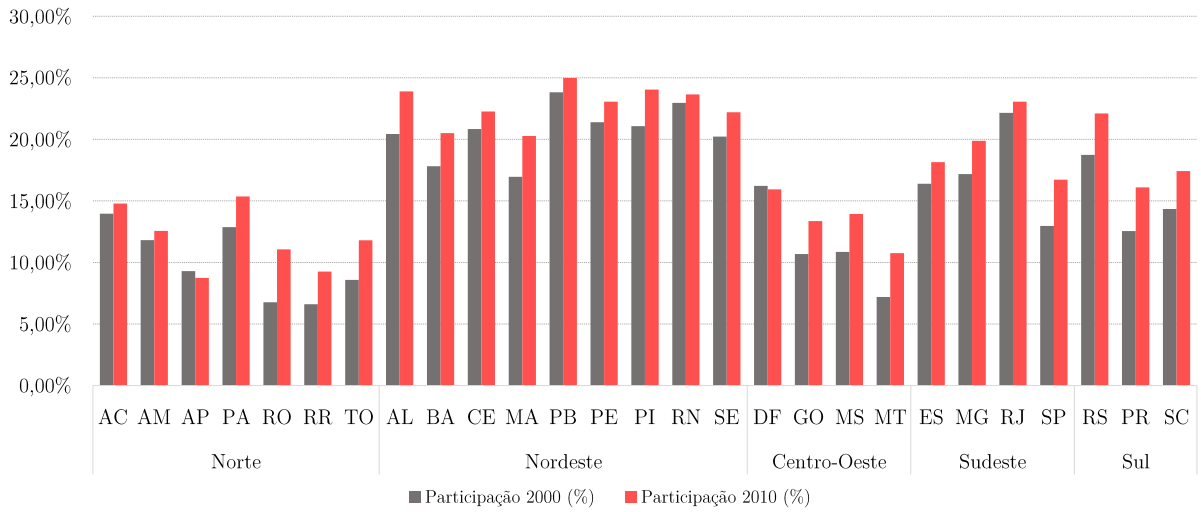


Fonte: Elaborado pelo autor

Posteriormente, foi analisado o impacto da parcela de aposentadorias e pensões acerca dos resultados encontrados para o índice de Gini nas cidades e estados brasileiros. Primeiramente, optou-se por examinar a sua participação percentual acerca do rendimento mensal per capita de cada unidade federativa do Brasil, podendo ser observado na Figura 8.

Examinando os resultados encontrados, percebe-se que essa parcela tem importante participação no rendimento mensal per capita, com médias estaduais de 15,36% e 17,63% para 2000 e 2010 respectivamente. Nota-se também que apenas as unidades federativas do Amapá e do Distrito Federal não tiveram uma variação percentual positiva no ano de 2010 em relação aos dados de 2000, evidenciando uma crescente presença desses benefícios acerca da economia nacional.

Figura 8 – Participação percentual das rendas de aposentadorias e pensões no total dos rendimentos per capita das unidades federativas brasileiras, 2000 e 2010



Fonte: Elaborado pelo autor

Em relação a evolução da participação total da parcela entre 2000 e 2010, destacam-se os estados de RO (Rondônia), RR (Roraima) e MT, com uma variação percentual positiva acima de 40%. Além disso, os estados da região nordeste ganham atenção especial nessa análise, uma vez que em todos as aposentadorias e pensões representam pelo menos 1/5 de todo o rendimento per capita da população. Em escala municipal, a participação desses benefícios no rendimento da população é ainda mais impactante, com aproximadamente 45% das cidades (2.465 municípios) apresentando pelo menos 1/3 dos rendimentos per capita provenientes dos benefícios citados no ano de 2010.

Conforme descrito na Equação 2.13 da Metodologia, a progressividade (π_h), ou razão concentração, é uma métrica utilizada para aferir a contribuição de uma parcela de rendimento no aumento ou diminuição da desigualdade de renda de uma população. Estudos realizados na literatura identificam que a parcela de aposentadorias e pensões contribui para a concentração de renda no Brasil (HOFFMANN, 2009).

Entretanto, analisando a Tabela 6 observa-se que a categoria até um salário mínimo apresenta caráter progressivo ($\pi_h > 0$), colaborando para a diminuição da concentração de renda, para todos os estados brasileiros em ambos os anos examinados. Em contrapartida, as aposentadorias e pensões acima de um salário mínimo se mostram regressivas ($\pi_h < 0$) em todos os estados, indicando uma participação dessa parcela para o aumento da concentração de renda no Brasil.

Tabela 6 – Progressividade (π_h) da parcela de rendimentos de aposentadorias e pensões por categorias para os estados brasileiros, 2000 e 2010

	AC	AL	AM	AP	BA	CE	DF	ES	GO
até um salário - 2000	0,8904	0,7139	0,9482	0,9839	0,7086	0,7085	1,2129	0,9615	0,9554
até um salário - 2010	0,7168	0,5275	0,7288	0,6726	0,5609	0,4856	1,1914	0,8210	0,7927
acima de um salário - 2000	-0,1043	-0,2503	-0,1719	-0,1718	-0,2289	-0,2524	-0,1387	-0,1734	-0,1819
acima de um salário - 2010	-0,2019	-0,2878	-0,2334	-0,1781	-0,2588	-0,2866	-0,1349	-0,2130	-0,2293
	MA	MG	MS	MT	PA	PB	PE	PI	PR
até um salário - 2000	0,5486	0,9630	0,9998	1,0264	0,8381	0,6896	0,8255	0,5637	1,0244
até um salário - 2010	0,4037	0,8138	0,8721	0,8693	0,5670	0,5435	0,6020	0,4403	0,9043
acima de um salário - 2000	-0,2847	-0,1635	-0,1530	-0,1602	-0,2124	-0,2715	-0,2097	-0,2628	-0,1048
acima de um salário - 2010	-0,2713	-0,2329	-0,2118	-0,2173	-0,2612	-0,2727	-0,2638	-0,2619	-0,1700
	RJ	RN	RO	RR	RS	SC	SE	SP	TO
até um salário - 2000	1,1402	0,8181	0,9099	0,9947	1,0494	1,0829	0,7038	1,1800	0,7877
até um salário - 2010	1,0064	0,5769	0,6994	0,7557	0,8935	0,9628	0,5785	1,0513	0,7044
acima de um salário - 2000	-0,1008	-0,2492	-0,1095	-0,1463	-0,0955	-0,0726	-0,2429	-0,0026	-0,1976
acima de um salário - 2010	-0,1711	-0,2726	-0,1775	-0,1652	-0,1675	-0,1603	-0,2710	-0,1273	-0,1873

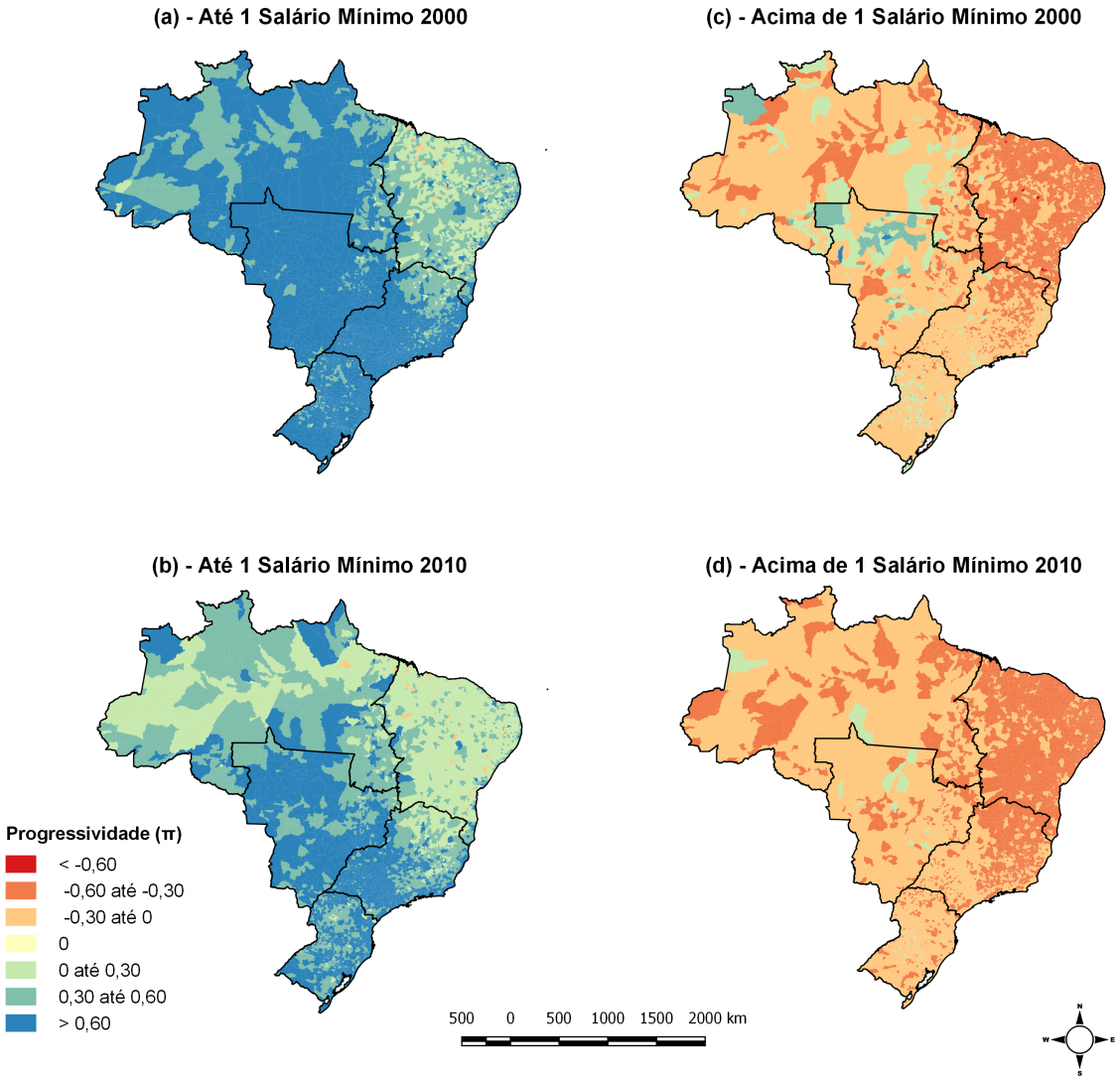
Fonte: Elaborado pelo autor.

Na Figura 9 é exibida a situação da progressividade acerca dos municípios brasileiros. As subfiguras 9(a) e 9(c), correspondentes à parcela até um salário mínimo para 2000 e 2010, respectivamente, demonstram que o padrão exibido na análise das unidades federativas brasileiras se repete em quase todo território nacional, onde grande parte dos municípios apresentam caráter progressivo ($\pi_h > 0$) para a categoria. Destacam-se as cidades das regiões Norte, Sul, Sudeste, Centro-Oeste para 2000 e Sul, Sudeste e Centro-Oeste para 2010, as quais apresentam maiores índices de progressividade.

Do mesmo modo, as subfiguras 9(b) e 9(d) ilustram as situações das progressividades para as parcelas de aposentadorias e pensões acima de um salário mínimo nos anos de 2000 e 2010 respectivamente. O caráter regressivo ($\pi_h < 0$) observado em todos as unidades federativas para essa variável se repete acerca de grande parte dos municípios brasileiros, demostrando a importante participação dessa categoria no aumento da desigualdade de rendimentos no país.

As cidades da região nordeste destacam-se em relação as demais regiões, apresentando valores ligeiramente negativos para o indicador na categoria acima de um salário mínimo, e índices baixos na parcela de até um salário em ambos os anos analisados. Tal resultado indica que embora grande parte dos rendimentos nesta região sejam proveniente dessa categoria de benefício, conforme indicado na Figura 9, possivelmente as aposentadorias e pensões acima de um salário estão concentradas na menor parcela de beneficiários, colaborando diretamente para concentração de renda existente nesses municípios.

Figura 9 – Progressividade (π_h) da parcela de rendimentos de aposentadorias e pensões por categorias para os municípios brasileiros, 2000 e 2010

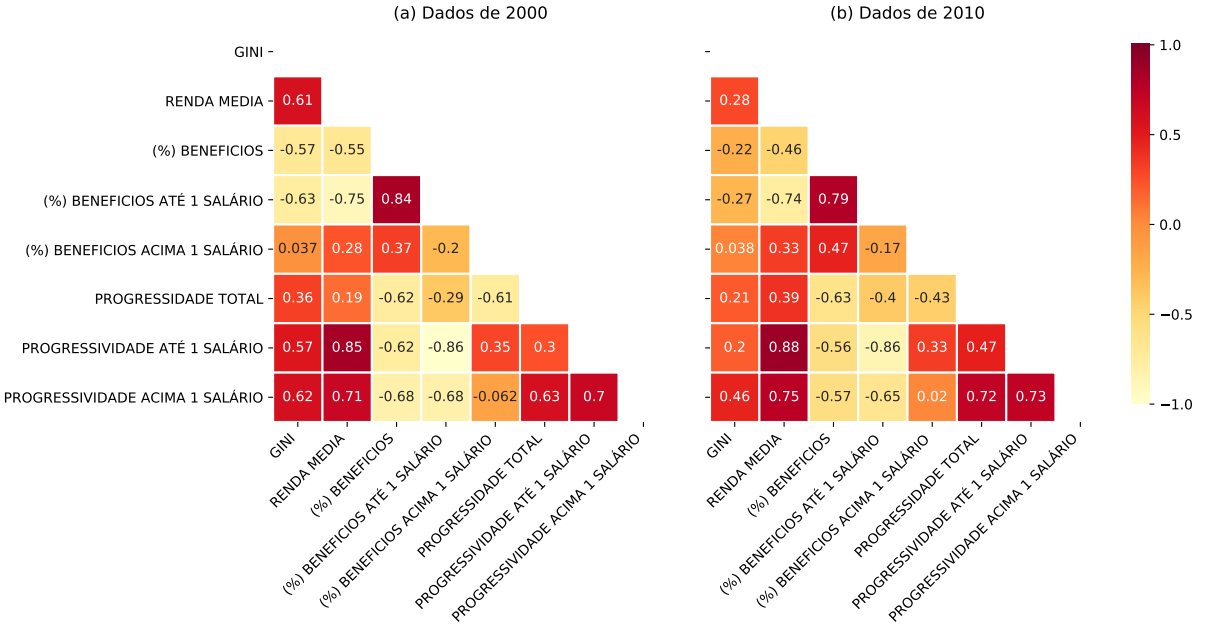


Fonte: Elaborado pelo autor

Posteriormente, foi avaliada a interdependência estatística existente entre o índice de Gini, os percentuais de participação dos benefícios acerca da composição de renda per capita e as progressividades resultantes para aposentadorias e pensões. Dessa forma, utilizando o coeficiente de Pearson, foram desenvolvidas as matrizes de correlação dos atributos analisados para os anos de 2000 e 2010 exibidas nas Figuras 10(a) e 10(b) respectivamente.

Avaliando a imagem, é possível observar uma diminuição na inter-relação existente entre a desigualdade de renda e as demais variáveis examinadas para o intervalo de tempo analisado. Destaca-se a correlação negativa existente entre o índice de Gini e o percentual de participação das aposentadorias e pensões - total e até um salário mínimo - na composição de renda da população. Tal fator indica uma tendencia inversamente proporcional entre as

Figura 10 – Matrizes de correlação das variáveis. (a) 2000 (b) 2010



Fonte: Elaborado pelo autor

variáveis, ou seja, quanto maior taxa de participação dessa categoria acerca dos municípios brasileiros, menor será a desigualdade de distribuição de riquezas.

No entanto, evidencia-se a interdependência negativa entre o percentual de participação dos benefícios e a renda média da população, pressupondo que embora essa categoria colabore para equalização de rendimentos, a mesma nivela os valores de renda por baixo. Por outro lado, a correlação existente entre as progressividades e a participação das aposentadorias e pensões demonstram, novamente, o importante papel que esses benefícios possuem na desigualdade de renda existente nos municípios. Todavia, destaca-se a importância de examinar variáveis geográficas, como a distancia das cidades em relação às metrópoles ou suas clientelas, afim de compreender melhor as interdependências existentes.

4.4 Considerações Finais

No presente estudo foi analisada a participação dos benefícios de aposentadorias e pensões na distribuição de renda per capita do Brasil (estados e municípios) nos anos de 2000 e 2010. A partir dos resultados demonstrados para o índice de Gini, conclui-se que embora haja uma heterogeneidade na sua distribuição acerca do território nacional, houve uma melhora no parâmetro, tanto a nível estadual quanto municipal, entre os anos analisados. Paralelamente, perante o cenário relatado anteriormente, o qual expõe a extensão da longevidade da população brasileira e o impacto econômico gerado pelo aumento na distribuição das aposentadorias e pensões, foi comprovado um acréscimo

na participação percentual dessa parcela no total de rendimento per capita em todos os estados brasileiros.

No entanto, quando analisadas as participações das aposentadorias e pensões, dividindo-as em categorias (até um salário e acima de um salário), os resultados demonstraram que ambas têm papel divergente em sua contribuição para a concentração de renda. Com um padrão se repetindo em todos os estados – e maioria das cidades brasileiras –, os benefícios acima de um salário apresentaram caráter regressivo ($\pi_h < 0$), colaborando para um aumento do índice de Gini, destacando-se a maior parte dos municípios do nordeste. Por outro lado, a categoria de benefícios até um salário demonstrou caráter progressivo ($\pi_h > 0$) em todo o país, colaborando para uma desconcentração de renda e decréscimo do índice de Gini.

Com isso, identifica-se que embora as aposentadorias e pensões colaborem para a concentração de renda no Brasil, conforme apresentado por (FERREIRA, 2006; HOFFMANN; LEONE, 2004; NICHOLSON, 2007) , tal fator apresenta caráter contraditório ao segmentarmos os benefícios em diferentes faixas salariais, uma vez que cada categoria colabora de forma diferente acerca da distribuição de renda. Dessa forma, correlacionando tal fator à heterogeneidade que o índice de Gini apresenta no cenário brasileiro, é possível vislumbrar discussões e análises cada vez mais complexas que objetivem equalizar as variáveis relacionadas a essa problemática.

Contudo, destaca-se que relatórios e estudos recentes apontam uma variação positiva para o coeficiente de Gini nos últimos anos, fator diretamente relacionado a crise econômica nacional a qual registrou uma queda de aproximadamente 9% no PIB entre 2014 e 2016 (NERI, 2019). Dessa forma, evidencia-se a necessidade de pesquisas com dados atualizados acerca dos municípios brasileiros a fim de mensurar a real situação de distribuição de renda do país, além do impacto que as aposentadorias e pensões causam nessa. Entretanto, os valores de rendimento domiciliar (e dos benefícios que os compõe), a nível municipal, são informações disponibilizadas apenas pelo censo demográfico, impossibilitando a realização de tais análises no presente momento.

Diante disso, como trabalhos futuros, propõem-se a replicação do atual estudo utilizando os dados futuramente disponibilizados pelo IBGE relacionados ao censo demográfico de 2020, a exploração de outros indicadores econômico-sociais (como por exemplo o índice de *Theil*), a aplicação técnicas de inteligência computacional para a extração de novos *insights* acerca da problemática e a utilização de métodos estatísticos espaciais (como o Índice de Moran (MORAN, 1950)) a fim de compreender melhor a autocorrelação existente na dispersão das variáveis entre os municípios.

5 Os Impactos da Reforma Previdenciária Acerca das Concessões de Benefícios de Aposentadoria

5.1 Considerações Iniciais

Conforme demonstrado no capítulo anterior, as aposentadorias e pensões compreendem uma importante categoria na composição de renda da população brasileira, impactando diretamente na desigualdade socioeconômica nacional. Dessa forma, acerca da magnitude social que esses benefícios proporcionam, evidencia-se que alterações nas regras de acesso ao regime previdenciário, como as previstas pela PEC 287/16[1]) ou PEC 6/2019[2], tendem a apresentar importantes consequências acerca da sociedade.

Embora diversos trabalhos abordem a sustentabilidade da Previdência Social, estimando projeções econômicas a curto e longo prazo (COSTANZI; ANSILIERO, 2017; SILVA et al., 2017), ou a confiabilidade dos resultados divulgados pelo governo para justificar as referidas reformas (BASTOS et al., 2019; SILVA, 2017), observa-se na literatura a ausência de pesquisas que investiguem os efeitos que tais alterações nas regras de acesso causariam nas concessões de aposentadorias.

Dessa forma, esse estudo de caso objetivou avaliar os impactos que a reforma prevista pela PEC 06/2019 pode causar no cenário previdenciário brasileiro acerca das concessões dos benefícios de aposentadoria. Especificamente, visou, através de uma abordagem de ciência de dados, utilizar os dados oficiais disponibilizados pela CPI da Previdência, pelo IBGE e pelo Anuário Estatístico da Previdência Social, para simular os efeitos que a PEC 06/2019 causariam no período de 1995 a 2016 se a mesma já estivesse em vigor no momento em que cada benefício foi concedido.

Para isso, esse capítulo foi dividido em 4 subseções incluindo esta breve introdução. Em seguida, na subseção 5.2, será abordada a metodologia para o desenvolvimento das análises realizadas, evidenciando a arquitetura utilizada para armazenamento dos dados, validação das informações disponibilizadas e detalhes da simulação proposta. Na subseção 5.3 serão discutidos os principais resultados encontrados, finalizando o estudo na subseção 5.4 com as considerações finais.

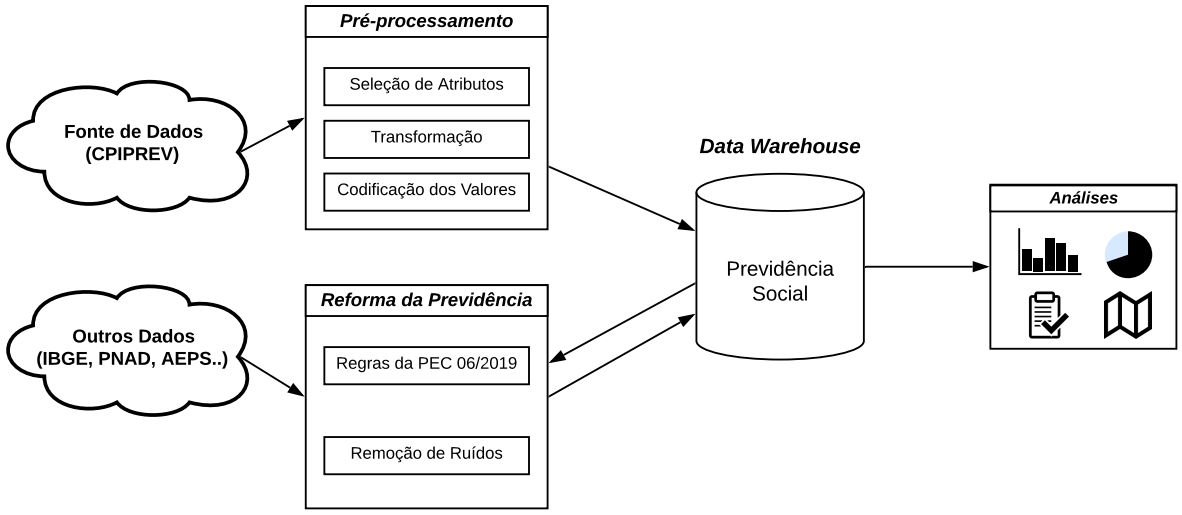
[1] Disponível em: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2119881> Acesso em: 14 Jan, 2020

[2] Disponível em: <https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2192459> Acesso em: 14 Jan, 2020

5.2 Metodologia Utilizada

Para o desenvolvimento do estudo proposto foi estabelecida uma sequência de etapas relacionadas a análise de dados, semelhante ao estudo anterior. No entanto, devido à natureza da problemática, e ao fato da pesquisa necessitar de outras fontes de informações para a abstração de melhores *insights*, foi adotada uma estratégia metodológica diferente. Sendo assim, como forma de otimizar a organização, armazenamento e consultas nos dados, foi elaborado uma estrutura de *Data Warehouse*[3] contendo todas as bases de dados inerentes ao estudo. A Figura 11 exibe o fluxo de desenvolvimento deste estudo de caso, contendo cada uma das etapas realizadas seguido por um grau de detalhamento mais aprofundado acerca dessa seção.

Figura 11 – Fluxograma de desenvolvimento do segundo estudo de caso



Fonte: Elaborado pelo autor

5.2.1 Origem dos Dados e Elaboração do *Data Warehouse*

O estágio inicial do procedimento metodológico deste estudo norteou-se na aquisição dos microdados previdenciários necessários para a reprodução da PEC 06/2019. Sendo assim, após a divulgação do trabalho de (SILVA et al., 2017), foram solicitadas as informações ausentes nos anexos das LDOs, necessárias para a reprodução dos seus cálculos. Dessa forma, o Ministério da Fazenda disponibilizou os dados referentes a aposentadorias

[3] Data warehouse compreende uma organização para banco de dados comumente utilizada em inteligência de negócios. Sua principal característica esta relacionada ao armazenamento de grandes quantidade dados, de diferentes fontes, em uma arquitetura que possibilite a realização de consultas complexas com maior grau de facilidade.

e pensões no site[4] CPI da Previdência de 2017, enumerados pelos documentos DOC090 e DOC097.

Esses documentos compreendem séries históricas de dados relacionados a concessão de benefícios de diversas espécies, sendo as principais: aposentadorias por tempo de contribuição (espécie 42), por tempo de contribuição especial (espécies 46 e 57), por idade (espécie 41), por invalidez (espécie 32), além dos benefícios de pensão por morte (espécie 21) e auxílio doença (espécie 31). A Tabela 7 exibe um breve resumo acerca dos dados disponibilizados pela CPI da Previdência de 2017.

Tabela 7 – Informações sobre os microdados disponibilizados pela CPI da Previdência de 2017.

Origem	**Arquivos**	**Nº de Registros**	**Descrição**
DOC090	APOSENTADORIAS.csv	46.545.767	Benefícios de Aposentadorias (42, 46, 57, 41 e 32) e Auxílio 31 concedidos entre 2000 e 2015
	PENSAO.csv	5.647.457	Benefícios de Pensão da espécie 21 concedidas entre 2000 e 2015
DOC097	SEN_MICRODADOS01.APO.AUX.csv	57.955.872	Benefícios de Aposentadorias (42, 46, 57, 41, 32 e 92) e Auxílios (91 e 31) concedidos entre 1995 e 2016
	SEN_MICRODADOS02_PENS.csv	7,964,185	Benefícios de Pensão por Morte (21, 93) concedidos entre 1995 e 2016
		118.113.281	

Fonte: Elaborado pelo autor.

Para a realização das análises e simulações propostas, optou-se por utilizar o arquivo SEN_MICRODADOS01.APO.AUX.csv devido ao fato desse agregar informações relacionadas às principais espécies de aposentadorias em um maior intervalo de tempo. Assim sendo, a base de dados foi inserida na estrutura de *Data Warehouse* dividindo a base de dados em diferentes tabelas do tipo FATO e DIMENSAO, sendo as primeiras responsáveis pelo conteúdo bruto dos dados (registros de aposentadorias) e as segundas por dados complementares (formato de datas, espécies de aposentadorias, tipo de clientela, etc) utilizadas principalmente em consultas mais específicas e complexas.

Para a elaboração da estrutura da *Data Warehouse* e bancos de dados, além da criação das tabelas FATO e DIMESAO e conversão dos dados contidos no arquivo .csv (*comma separated values*) em registros, foi utilizado o *software* de código aberto *Pentaho Data Integration* (PDI)[5]. Devido ao fato dessa ferramenta ser fortemente utilizada na área de inteligência de negócios, a mesma promove um conjunto de funcionalidades relacionadas a análise de dados tais como suporte à linguagem Structured Query Language (SQL) e comunicação com os principais sistemas gerenciadores de bancos de dados.

[4] Disponível em: <https://legis.senado.leg.br/comissoes/docsRecCPI?codcol=2093> Acesso em: 17 Fev, 2020

[5] Disponível em: <https://community.hitachivantara.com/s/article/data-integration-kettle> Acesso em: 18 Fev, 2020

Diante disso, foram utilizadas funcionalidades do PDI e da linguagem de programação SQL para realizar uma etapa de pré-processamento dos dados. Essa etapa compreendeu a seleção dos atributos relevantes para o estudo e transformações acerca dos microdados, como por exemplo codificação de valores textuais em numéricos para diminuir a demanda computacional necessária. Por fim, os microdados resultantes, correspondentes a série histórica de benefícios de aposentadorias/auxílios concedidos entre os anos de 1995 e 2016, foram armazenados em SGBD PostgreSQL. O Apêndice A apresenta, resumidamente, os principais metadados presentes nesta base de dados, podendo ser encontrado ao final deste documento.

Diante disso, a partir da utilização de funcionalidades do PDI e da linguagem de programação SQL, foi realizada uma etapa de pré-processamento dos dados, a qual compreendeu a seleção dos atributos relevantes para o estudo e transformações acerca dos microdados, como a codificação dos valores textuais em numéricos. Posteriormente, os dados resultantes, correspondentes a série histórica de benefícios de aposentadorias/auxílios concedidos entre os anos de 1995 e 2016, foram armazenados em um SGBD PostgreSQL.

Destaque-se que a utilização de uma estrutura de *Data Warehouse* não possibilita a inserção ou remoção de registros no banco de dados de maneira automática e dinâmica. Para tal funcionalidade, o administrador da arquitetura – utilizando técnicas de *Extract, Transfom and Load* (ETL) responsáveis por extrair, transformar/pré-processar e carregar dados de diferentes origens em bancos de dados – promove a atualização do *Data Warehouse* de acordo com a demanda necessária (KIMBALL; CASERTA, 2011; ELSAPPAGH; HENDAWI; BASTAWISSY, 2011). Todavia, devido à natureza da aplicação realizada nesse estudo não promover uma atualização dinâmica nos dados, fator justificado pela dificuldade de acesso a microdados recentes de concessões de aposentadorias, essa ferramenta apresentou-se promissora uma vez que possibilita consultas mais complexas e com menor grau de dificuldade.

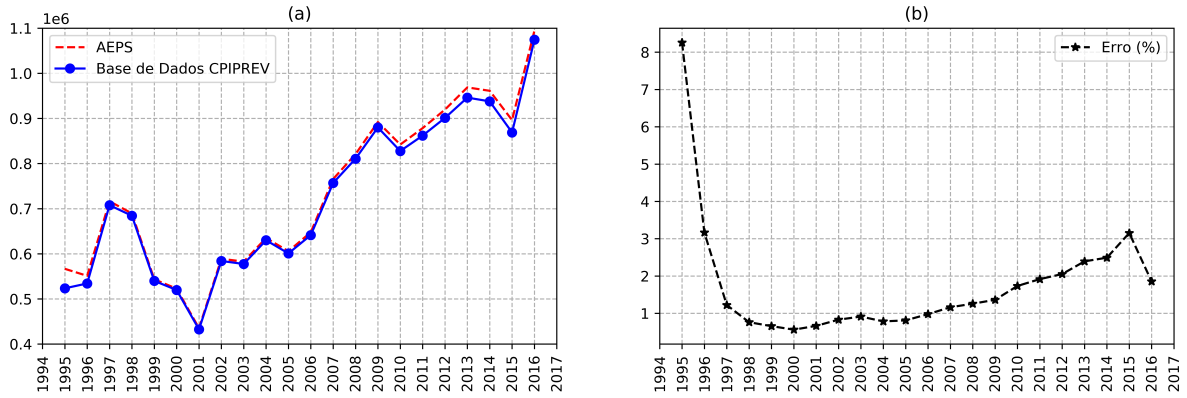
5.2.2 Análise Exploratória

Uma vez que o desenvolvimento do *Data Warehouse* responsável pelo armazenamento dos microdados previdenciários foi realizado, constatou-se a necessidade de efetuar uma análise exploratória acerca dos dados, objetivando, principalmente, compreender o seu conteúdo e avaliar a confiabilidade das informações disponibilizadas. Dessa forma, foram utilizados os microdados fornecidos pelo DATAPREV para recalcular os resultados divulgados pelo Anuário Estatístico da Previdência Social (AEPS)[6]. Inicialmente foi elaborada uma série histórica comparativa referentes a concessão dos benefícios de aposentadorias. A Figura 12(a) ilustra o comportamento de ambas as fontes de dados para o intervalo de

[6] O AEPS foi criado em 1992 e é responsável por disponibilizar para a populações dados estatísticos e básicos referentes a situação dos benefícios mantidos pelo INSS e sobre suas arrecadações de contribuições previdenciárias (IBGE, 2017)

1995-2016. Por sua vez, a Figura 12(b) exibe o erro percentual existente entre as duas, enfatizando o ano de 1995, o que apresentou uma diferença acima de 8% entre dados disponibilizados pela CPI e o AEPS daquele ano.

Figura 12 – (a) Comparação entre os dados disponibilizados pela CPI da Previdência e as informações contidas no APES. (b) Erro percentual.



Fonte: Elaborado pelo autor

Similarmente, foi reproduzido um dos resultados contidos no AEPS de 2016, intitulado "A.1 - Quantidade de benefícios concedidos, por clientela, segundo os grupos de espécies - 2014/2016", contido na página 21[7], utilizando os microdados disponibilizados pelo DATAPREV. Com isso, conforme exibido na Tabela 8, notou-se uma pequena variação entre os resultados, proveniente, possivelmente, de ações jurídicas relacionadas a concessões de benefícios, ou seja, contribuintes que deram entrada em benefícios e tiveram retorno após a data de disponibilização dos dados pela CPI da previdência.

Tabela 8 – Comparação entre os dados disponibilizados pelo DATAPREV e as informações contidas no APES: Quantidade de benefícios concedidos por grupos de espécies entre 2014/2016

Espécie	**AEPS**			**DATAPREV**		
	2014	**2015**	**2016**	**2014**	**2015**	**2016**
Aposentadorias	1.150.880	1.058.151	1.263.974	1.115.107	1.058.379	1.290.267
- Tempo de Contribuição	315.542	320.460	432.033	315.483	320.414	439.850
- Idade	645.687	575.841	662.366	645.744	575.886	674.338
- Invalidez	189.651	161.850	169.575	189.880	162.079	176.079

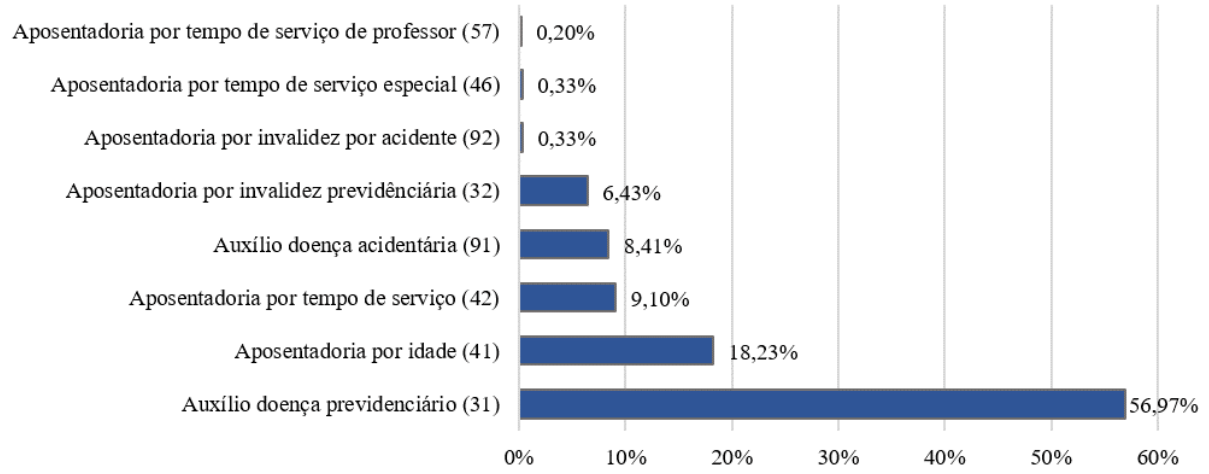
Fonte: Elaborado pelo autor

Posteriormente, após validar a confiabilidade dos dados utilizados, foi estimada a quantidade de registros existentes na base de dados classificando-os por espécies de benefícios. Dessa forma, a Figura 13 exibe um gráfico de barras contendo os percentuais de participação de cada categoria na base de dados, enfatizando as aposentadorias das

[7] Disponível em: <http://sa.previdencia.gov.br/site/2018/08/aeps2016.pdf> Acesso em: 5 de Mar, 2020

espécies 42, 46, 57, 41, 32 e 92 (34,62% do total) correspondentes aos benefícios diretamente atingidos pela reforma previdenciária prevista na PEC 06/2019.

Figura 13 – Percentual de participação dos benefícios da base dados classificados por espécie.



Fonte: Elaborado pelo autor

Em contrapartida, quando analisada a proporção de aposentadorias concedidas classificando-as por sexo, identificou-se relativo grau de equilíbrio com 55,06% dos segurados correspondendo ao sexo masculino, 44,93% ao feminino e 0,01% com o sexo ignorado/indefinido. No entanto, devido a natureza das simulações propostas, a última categoria foi desconsiderada visto que as regras previstas na PEC 06/2019 não incorporavam diretrizes relacionadas a sexo indefinido.

Após a realização da análise exploratória e validação das informações existentes no banco de dados, contatou-se a possibilidade de simular a reforma da previdenciária conforme previsto nos objetivos deste estudo. Além disso, outras análises estatísticas e quantitativas mais detalhadas foram realizadas a fim de compreender melhor o conteúdo desses microdados. Foram disponibilizadas as informações referentes a essas análises complementares, a documentação necessária para sua replicação os códigos fontes utilizados para cada cálculo e consulta realizados acerca da base de dados no repositório do projeto disponível na plataforma GitHub[8].

5.2.3 Simulação da Reforma Previdenciária (PEC06/2019)

Após a realização das etapas anteriores, e utilizando dados públicos disponibilizados pelo IBGE e AEPS, foi elaborada e incorporada uma nova base de dados ao *Data Warehouse* referente a uma simulação das regras previstas pela reforma da previdência prevista na

[8] Disponível em: <https://github.com/andrespp/ds-prev> Acesso em: 3 Mai, 2020

PEC 06/2019[9]. A Tabela 9 exibe de forma resumida as diretrizes previstas pela reforma para as principais espécies de benefícios em relação a idade mínima e tempo de contribuição em anos.

Tabela 9 – Regras para obtenção de aposentadorias propostas pela PEC 06/2019.

Espécie	**Idade**		**Tempo de Contribuição**	
	Mulher	**Homem**	**Mulher**	**Homem**
Urbana (41/42)	62	65	15	20
Rural (41/42)	55	60	15	
Professor (57)	57	60	25	
Especial (46)	55		15	
Especial (46)	58		20	
Especial (46)	60		25	

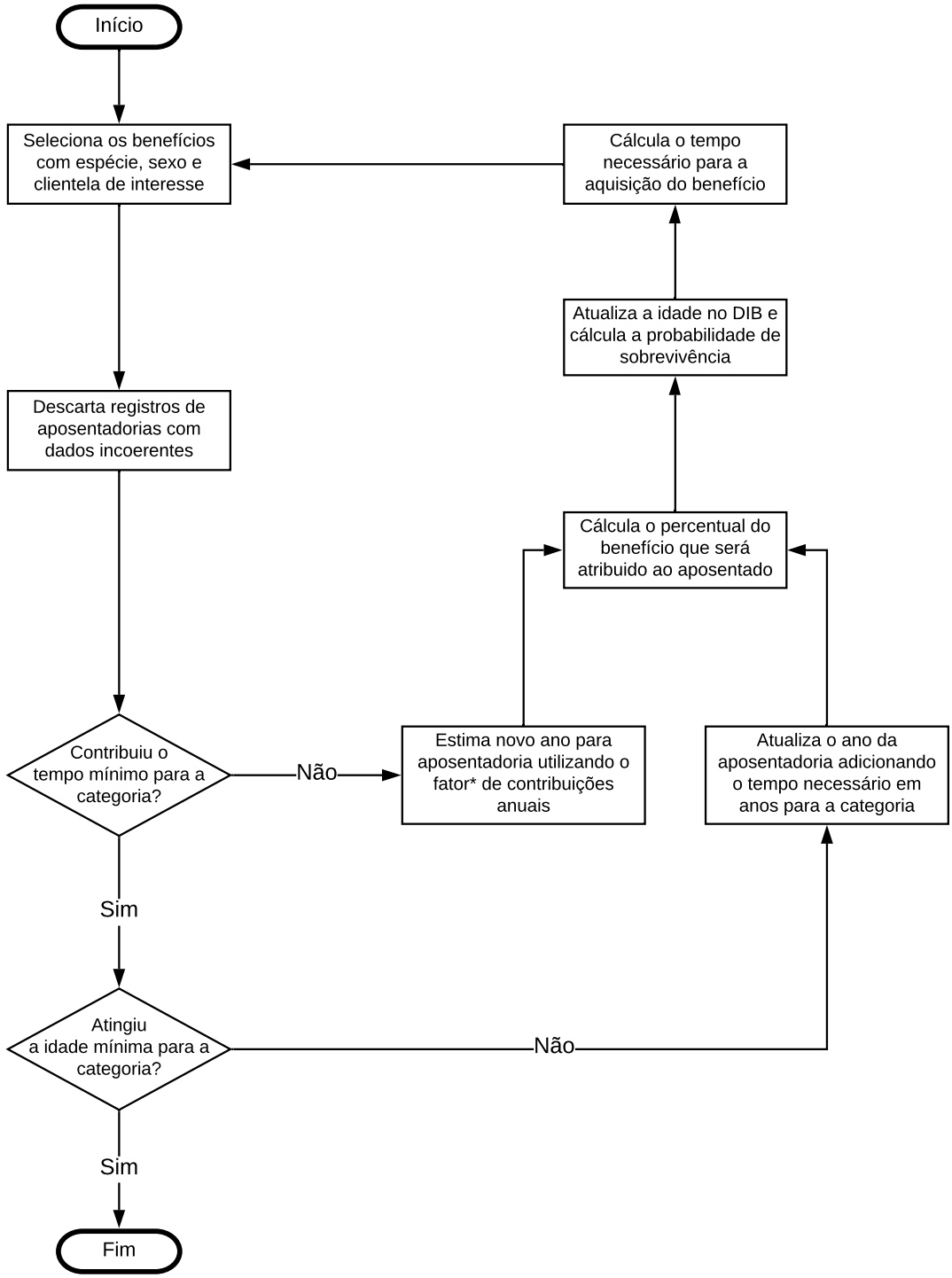
Fonte: Elaborado pelo autor

Essa nova base de dados, além de compreender o universo de registros existentes no banco de dados da previdência, agrega novos atributos referentes a projeções de como tais benefícios se comportariam caso a reforma já estivesse em vigor no momento em que cada aposentadoria foi concedida[10]. Na Figura 14 é possível visualizar, de forma resumida, a lógica utilizada para a realização da simulação proposta através de um fluxograma, o qual terá um maior grau de detalhamento das tarefas realizadas no decorrer desta subseção.

[9] As regras utilizadas são referentes à proposta em vigor no dia 15/06/2019, podendo eventualmente ter sofrido alterações após a publicação do trabalho.

[10] A simulação não considera regras de transição, mas sim as regras finais para concessão de cada benefício.

Figura 14 – Fluxograma da simulação da reforma previdenciária PEC 06/2016



Fonte: Elaborado pelo autor

* O fator de contribuições anuais, destacado no fluxograma, corresponde ao número médio de contribuições anuais, em meses, realizado por cada beneficiário.

Inicialmente, houve a leitura e seleção dos registros inerentes a pesquisa, correspondendo aos benefícios das espécies 41, 42, 46, 57, 32 e 92 – aposentadorias por idade, tempo de contribuição, tempo de contribuição especial, tempo de contribuição professor, invalidez previdenciária e invalidez acidentária, respectivamente – para os gêneros masculino e feminino, com clientela urbana e rural, os quais corresponderam a 19.775.373 registros da base de dados. Entretanto, devido a existência de anomalias na idade e tempo de contribuição dos beneficiários, houve uma etapa de triagem nos microdados, reduzindo o universo de interesse em aproximadamente 12,84% em relação a quantidade anterior, resultando em 17.236.900 registros válidos.

Após a seleção e remoção de ruídos nos dados de interesse, foram realizadas as simulações propostas. Diante disso, primeiramente, foi verificado se o tempo de contribuição de cada beneficiário correspondia ao tempo mínimo em anos necessário para que a sua categoria obtivesse acesso ao benefício. Caso sim, o registro era selecionado para a próxima etapa, no entanto, caso não, seria estimado e agregado ao beneficiário o tempo em anos necessário para que o mesmo atingisse a regra proposta. Para tal estimativa, levou-se em consideração a dificuldade dos trabalhadores do setor privado se manterem em seus respectivos trabalhos devido, principalmente, as altas taxas de desemprego e informalidade existentes no país (PNAD, 2019). Em seguida, verificou-se se cada beneficiário atingia a idade mínima da sua categoria para a concessão da aposentadoria, acrescentando o tempo a mais que o mesmo haveria de esperar para ter acesso a seu benefício quando necessário.

Dessa forma, após estimar as novas idades e tempo a mais de contribuição, foram calculadas as probabilidades dos beneficiários que não atendiam as regras previstas na PEC 06/2019 estarem vivos na futura concessão dos seus benefícios. Para isso, optou-se por utilizar a projeção populacional realizada pelo IBGE, a qual corresponde a dados de estimativas demográficas a nível nacional classificadas por sexo, idade e clientela, de 2010 a 2060. Assim sendo, a probabilidade de sobrevivência de cada beneficiário pode ser calculada através da Equação 5.1 abaixo.

$$prob_sobrev(i_{proj}, a_{proj}, s) = \frac{POP_proj(i_{proj}, a_{proj}, s)}{POP_atual(i_{atual}, a_{atual}, s)}, \{a \epsilon \mathbb{I} | 2010 \leq a \leq 2060\} \quad (5.1)$$

sendo s o sexo do beneficiário e, i_{proj}, i_{atual}, a_{proj} e i_{atual} a idade na projeção, idade atual, ano na projeção e ano atual, receptivamente. Além disso, as variáveis POP_atual e $POP_projetada$ correspondem as populações atuais e projetadas pelo IBGE para cada idade, sexo e ano de interesse.

Por fim, foi calculado o percentual a ser aplicado na média dos salários (contribuições) de cada beneficiário baseando-se nas regras previstas pela reforma da previdência, a qual estipula uma mensalidade 60% do valor médio, possibilitando um acréscimo de 2%

para cada ano que exceder o tempo mínimo de contribuição . Dessa forma, foram criados sete novos atributos correspondentes a simulação da PEC 06/2019, sendo:

- `PEC6_GAP_IDADE`: quantidade de anos a mais que o beneficiário deveria esperar para obter o benefício considerando as novas regras;
- `PEC6_GAP_CONTRIB`: quantidade de anos a mais que o beneficiário deveria trabalhar para obter o benefício considerando as novas regras;
- `PEC6_GAP`: valor máximo entre o `PEC6_GAP_IDADE` e `PEC6_GAP_CONTRIB`;
- `PEC6_ANO_DIB`: ano mínimo em que o beneficiário terá acesso ao benefício considerando as novas regras;
- `PEC6_IDADE_DIB`: idade do beneficiário na nova data de início do benefício de aposentadoria considerando as novas regras;
- `PEC6_PROB`: probabilidade de o beneficiário estar vivo no ano em que ficaria elegível para obtenção do benefício;
- `PEC6_PERCENT`: percentual a ser aplicado sobre a média das contribuições para fins de cálculo do valor do benefício.

Para o desenvolvimento dessa simulação foi utilizada a linguagem de programação *Python 3*, os *frameworks* de análise de dados *Pandas* e *Numpy* e biblioteca *Psycopg2* responsável por estabelecer uma conexão com o banco de dados PostgreSQL. No repositório do projeto, disponibilizado através da plataforma GitHub[11], pode-se encontrar os *scripts* utilizados em toda simulação com toda a documentação necessária para sua reprodução, apresentando dessa forma um maior grau de detalhamento acerca dos cálculos realizado para cada novo atributo. A Tabela 10 descreve a estrutura do projeto e os principais arquivos existentes.

[11] Disponível em: <https://github.com/andrespp/ds-prev> Acesso em: 3 de Mai, 2020

Tabela 10 – Descrição dos principais arquivos do projeto no GitHub referente ao segundo estudo de caso

requirements.txt	arquivo de texto descrevendo todas as dependências e pacotes Python necessários para a execução do projeto.
dataset	pasta contendo todos os dados necessários para a execução das análises.
doc	pasta contendo documentos inerentes a simulação, tais como APES, Ofícios da CPI da Previdência, Taxas, dentre outros.
compose	pasta contendo os arquivos necessários para levantar a infraestrutura do projeto utilizando a aplicação Docker.
metadata	pasta contendo os dicionários e documentações das bases de dados.
dw	pasta contendo os arquivos necessários para a criação do data warehouse e realização da simulação.
analisys	pasta com as diversas análises realizadas utilizando a *framework jupyter notebook*.
README.md	descrição das características e funcionamento do projeto.

Fonte: Elaborado pelo autor.

5.3 Resultados e Discussão

A partir da estruturação e implementação da arquitetura do *Data Warehouse* para o armazenamento dos microdados previdenciários, e aplicando técnicas/estratégias de ciência de dados para a extração de conhecimento válido acerca da problemática abordada, foi possível realizar a simulação proposta nos objetivos desse estudo. Em vista disso, avaliando as atuais diretrizes do sistema previdenciário nacional, observa-se que as duas principais informações utilizadas para avaliar a viabilidade de um contribuinte se aposentar, tanto no regime atual quanto nas regras da PEC 06/2019, consistem na idade da pessoa e do seu tempo de contribuição para o RGPS. A partir dessas informações, e dependendo das características do trabalho exercido pela contribuinte em questão (trabalhador rural ou urbano, do sexo feminino ou masculino, do tipo professor ou não, dentre outros), este pode passar a ter direito ao benefício da sua aposentadoria.

Dessa forma, como ponto de partida para as análises realizadas, e utilizando os dados reais de aposentadorias concedidas no período analisado , calculou-se a idade média de entrada nos benefícios e o tempo médio em anos de contribuições dos aposentados para o intervalo de tempo analisado, exibindo os resultados nas Tabelas 11 e 12. Assim, destaca-se que, de forma geral, os homens se aposentavam em média com 58,24 anos de idade contribuindo com 26,77 anos de trabalho, enquanto as mulheres aposentavam-se com 57,23 anos de idade com uma média de 20,57 anos de trabalho contribuídos para a previdência social.

Tabela 11 – Idade média de entrada em benefício pelas regras atuais do RGPS.

	Idade	**T.C.**	**T.C. Esp.**	**Professor**	**Rural**
Homens	65,81	53,53	48,84	56,00	60,84
Mulheres	61,58	51,37	49,16	50,66	56,99

Elaborado pelo autor.

Tabela 12 – Tempo (em anos) médio de contribuição pelas regras atuais do RGPS.

	Idade	**T.C.**	**T.C. Esp.**	**Professor**	**Rural**
Homens	20,35	34,52	26,02	30,90	18,39
Mulheres	17,27	29.37	25,86	26,55	17,82

Elaborado pelo autor.

Avaliando as Tabelas 11 e 12, observa-se que, na média, os homens e mulheres que deram entrada em seus benefícios atendiam as regras expostas na Tabela 9. No entanto, é importante evidenciar que nem todos os trabalhadores que se aposentaram por idade, por

exemplo, possuíam o tempo mínimo de contribuição previsto nas regas da PEC 06/2019. Dessa forma, destaca-se que esse período faltando não compreende necessariamente o tempo de trabalho adicional que, na prática, esses contribuintes teriam que cumprir para se aposentarem. Tal fator é justificado pelas elevadas taxas de desemprego e informalidade que impactam diretamente na estabilidade dos trabalhadores do setor privado, dificultando que os mesmos trabalhem formalmente de forma ininterrupta, e consequentemente implicando na idade que estes iriam se aposentar.

Sendo assim, estima-se que os homens aposentados por idade pelas regras atuais necessitariam trabalhar um tempo adicional maior que o previsto, aplicando a mesma analogia para as mulheres e para os demais tipos de aposentadorias consideradas nessa pesquisa. Com isso, utilizando as informações disponibilizadas pelas Tabelas 11 e 12, e partindo do princípio de que todos começaram a trabalhar com 18 anos, foi calculado o número médio de contribuições[12] por ano para cada espécie de aposentadoria avaliada, tendo os resultados exibidos na Tabela 13.

Tabela 13 – Número médio de contribuições por ano.

	Idade	**T.C.**	**T.C. Esp.**	**Professor**	**Rural**
Homens	5.11	11,66	10,13	9,76	5,15
Mulheres	4,76	11,57	9,96	9,76	5.49

Fonte: Elaborado pelo autor.

Tais resultados demonstram que, pelas regras previstas na PEC 06/2019, em algumas situações, o tempo de contribuição necessário para se aposentar, em média, chegaria a equivaler ao dobro do tempo de contribuição restante. Por exemplo, um homem com 65 anos de idade e 15 anos de contribuição, que nas regras atuais consegue a concessão de aposentadoria por idade mas não atende as regras da PEC 06/2019 pela necessidade de 5 anos extras de contribuição, necessitaria trabalhar em média 11,74 anos a mais, visto que a cada ano o mesmo só conseguiria contribuir com 5,11 parcelas, resultando em uma idade de aproximadamente 77 anos ao se aposentar.

Além disso, é importante esclarecer que o número médio de contribuições anuais pode variar de acordo com o cenário econômico vivenciado pelo trabalhador. Crises econômicas e elevações nas taxas de desemprego e informalidade tendem a diminuir as contribuições realizadas, gerando, consequentemente, aumento no tempo adicional de trabalho para as concessões dos benefícios de acordo com as regras da PEC 06/2019.

Paralelamente, uma vez que há modificações nas regras para admissão dos beneficiários, evidencia-se alterações no cenário previdenciário em relação a concessões de

[12] Idealmente, o número médio de contribuições por ano deveria ser igual a 12, correspondente a 12 meses de trabalho.

aposentadorias. Em geral, as diretrizes atuais especificam que os trabalhadores devem possuir uma idade mínima ou ter um tempo de contribuição mínimo para que o benefício possa ser obtido. Uma vez que as regras propostas pela reforma conciliam ambos os critérios como pré-requisitos para obtenção da aposentadoria, estima-se que uma grande parcela dessas pessoas, possivelmente, não conseguiria mais se aposentar na mesma época em que se aposentou pelas regras atuais.

Diante disso, utilizando os microdados previdenciários e estratégias computacionais para análises em grandes quantidades de dados[13], foi estimado o percentual de aposentados que teriam acesso aos seus benefícios perante as novas regras propostas. Dessa forma, constatou-se que aproximadamente 83,28% dos trabalhadores que tiveram suas aposentadorias concedidas durante o intervalo de tempo analisado não teriam acesso a esses benefícios caso a reforma estivesse em vigor desde 1995. A Tabela 14 apresenta a proporção de aposentadorias concedidas entre 1995 e 2016 que não atendem as regras pela PEC 06/2019, classificando-as de acordo com seus respectivos sexos, clientelas e espécies do benefício obtido.

Tabela 14 – Proporção de pessoas que não atendem as regras da PEC na mesma data que deram entrada nos benefícios com as regras atuais.

	Idade	**T.C.**	**T.C. Esp.**	**Professor**	**Rural**
Homens	51,29%	98,57%	96,88%	77,89%	55,77%
Mulheres	84,27%	98,90%	97,91%	86,18%	87,13%

Fonte: Elaborado pelo autor.

Assim, dos homens que se aposentaram por idade, 51,29% não teriam conseguido o benefício, enquanto para as mulheres, das que se aposentaram por idade, 84,27% não teriam conseguido se aposentar. A mesma leitura é aplicada para as pessoas que se aposentaram nas espécies de Aposentadoria por Tempo de Contribuição, Aposentadoria por Tempo de Contribuição Especial, Aposentadoria de Professor e Aposentadoria Rural. Além disso, analisando os resultados, constata-se que as mulheres são as mais impactadas em todas as categorias. Tal fator ocorre, principalmente, devido ao acréscimo na idade mínima – de 60 para 62 anos – necessária para que as pessoas do sexo feminino possam ter acesso aos seus benefícios.

Paralelamente, avaliando os resultados da simulação relacionados ao tempo médio adicional de trabalho, ou espera, necessários para a obtenção da aposentadoria daqueles

[13] Devido ao fato da base de dados possuir um volume de informações maior que a quantidade de memória tipicamente disponível em um computador doméstico, optou-se por um processamento em lotes, onde subconjutos dos dados de tamanhos compatíveis com os recursos computacionais disponíveis foram processados sequencialmente. Tal estratégia foi adotada pela natureza dos dados, na qual cada lote é independente do anterior possibilitando dessa forma processamentos isolados.

que não cumpriram as regras previstas na PEC, considerando todo o período de 1995 a 2016, observa-se a existência de duas possíveis situações: (i) pessoas que não atingiram a idade mínima, mas já cumpriram o tempo mínimo de contribuição; (ii) pessoas que atingiram a idade mínima, mas não cumpriram o tempo mínimo de contribuição.

Na primeira situação, temos os trabalhadores que pelas regras atuais obtiveram o benefício da aposentadoria por tempo de contribuição, incluindo os professores que se aposentaram com idade inferior à proposta pela PEC 06/2019. Neste caso, estes trabalhadores necessitariam apenas esperar a idade mínima necessária sem a exigência de continuar contribuindo para a previdência social. Com isso, estimou-se que os homens que se aposentaram por tempo de contribuição deveriam em média esperar aproximadamente mais 11,18 anos até alcançarem a idade mínima, enquanto as mulheres mais 10,31 anos. No caso dos professores e professoras, haveria a necessidade de aguardar em média 5,15 a 7,19 anos, respectivamente. Embora não seja obrigatoriamente exigido que essas pessoas continuassem trabalhando e contribuindo para alcançarem a aposentadoria, provavelmente as mesmas iriam, em detrimento da necessidade de arcar com sua subsistência.

Por outro lado, na segunda categoria encontram-se, no geral, as pessoas que, pelas regras atuais, aposentam-se por idade, incluindo praticamente 100% dos trabalhadores rurais. Nessa situação específica, é obrigatório que os trabalhadores continuem contribuindo para o sistema previdenciário até cumprirem os requisitos exigidos pela PEC. Desse modo, observa-se que os homens urbanos e rurais que alcançaram os requisitos mínimos para a aposentadoria por idade pelas regras atuais, necessitariam trabalhar, em média, aproximadamente 4,80 e 6,20 anos a mais, respectivamente. Já para as mulheres, haveria a necessidade de acréscimo médio de 3,31 anos no setor urbano e 4,55 anos no rural. Acerca disso, a Tabela 15 apresenta uma síntese da idade média estimada para as aquisições dos benefícios de aposentadoria pelas regras previstas na PEC 06/2019.

Tabela 15 – Idade média estimada para a concessão das aposentadorias de acordo as diretrizes previstas pela PEC 06/2019.

	Idade + T.C.		**Professor**	**Rural**
Homens	68,24	$\frac{\text{75,30 – A. Idade}}{\text{65,03 – A. T.C.}}$	60,51	76,76
Mulheres	65,34	$\frac{\text{67,45 – A. Idade}}{\text{62,03 – A. TC}}$	57,35	67,14

Fonte: Elaborado pelo autor.

Adicionalmente, diante do que foi exposto, é importante ponderar a possibilidade do contribuinte estar vivo ou não até que o mesmo atinja os requisitos mínimos para a concessão do seu benefício. Diante desse cenário, considerando as datas de cessação

de benefícios pela causa de morte, notou-se que aproximadamente 6,97% (994.173) dos aposentados pelas regras atuais teriam falecido antes de dar entrada no benefício pelas regras previstas na PEC 06/2019. Além disso, utilizando a Equação 5.1 descrita na metodologia desse estudo de caso, estima-se que dos demais, 509.763 homens e 78.453 mulheres teriam 30% ou menos de chance de estarem vivos antes da sua elegibilidade para a aposentadoria.

Destaca-se também que a atual reforma previdenciária proposta pelo governo, além de dificultar o acesso dos contribuintes as suas aposentadorias, aplica modificações diretas nos valores de mensalidade repassadas pelo instituto responsável. Nesse contexto, as novas diretrizes preveem um valor de benefício correspondente a 60% da média total dos salários do trabalhador, com a possibilidade de acréscimo de 2% a mais para cada ano trabalhado além do mínimo, conforme relatado anteriormente. Dessa forma, observando os aposentados inclusos no espaço amostral analisado, estima-se que apenas 3% desses receberiam a média integral, ou mais, das suas contribuições. A Tabela 16 exibe, com maior grau de detalhamento, o percentual médio a ser aplicado para o cálculo do valor dos benefícios, classificando-os por tipo de aposentadoria e sexo.

Tabela 16 – Percentual médio a ser aplicado para cálculo do valor do benefício.

	Idade	**T.C.**	**T.C. Esp.**	**Professor**	**Rural**
Homens	0,66	0,89	0,72	0,81	0,66
Mulheres	0,66	0,88	0,82	0,83	0,69

Fonte: Elaborado pelo autor.

Além disso, evidencia-se que a parcela de segurados com possibilidade de receber o benefícios integral corresponde, em sua grande maioria, a aposentados por tempo de contribuição. No entanto, a quantidade de concessões de aposentadorias por tempo de contribuição representa apenas 29,24% do total de concessões. Isto indica que, com a aprovação da PEC, a grande maioria dos aposentados, tenderiam a receber em média um valor abaixo de 70% da média de todos os salários de contribuição.

É importante destacar que todos os resultados descritos foram validados e discutidos por especialistas da Associação Nacional dos Auditores Fiscais da Receita Federal (ANFIP), os quais divulgaram um relatório relacionado a temática abordada disponível em (GENTIL, 2019).

Por outro lado, avaliando a questão computacional, constatou-se que a abordagem de ciência de dados, aliada a utilização de uma estrutura de *Data Warehouse*, simplificou a execução das consultas, análises e simulação acerca dos microdados previdenciários. Tal fator justifica-se pela natureza dessa arquitetura, a qual objetiva, sobretudo, proporcionar

aos usuários e administradores do banco de dados a agregação de diversas dimensões de dados para a abstração de *insights* complexos.

Além disso, a aplicação de processamento em lotes para a simulação da reforma da previdência - conforme descrito na metodologia deste estudo - possibilitou a efetuação dos cálculos propostos acerca do grande volume de dados utilizados. Tal estratégia, embora não seja uma técnica de paralelização, viabiliza a execução da análise em computadores com recursos limitados, uma vez que executa as tarefas de leitura, processamento e escrita dos resultados acerca de parcelas de dados, descartando, dessa forma, a necessidade de memórias principais com elevadas capacidade de armazenamento, ou utilização do disco rígido como memória auxiliar.

5.4 Considerações Finais

No presente estudo foram realizadas análises acerca dos impactos que a reforma da previdência prevista pela PEC 06/2019 poderia causar no cenário previdenciário nacional, avaliando a série histórica de concessões de aposentadorias entre os anos 1995 e 2016. Nesse contexto, os resultados demonstraram que aproximadamente 78,66% das aposentadorias concedidas no período analisado não atendem as regras previstas na PEC, fazendo-se necessário um acréscimo em anos de contribuição, ou espera, para que os beneficiários tenham acesso a essas aposentadorias. Paralelamente, foi relatada a dificuldade que os trabalhadores, possivelmente, enfrentariam perante ao novo sistema proposto, uma vez que haverá a necessidade desses manterem-se formais em um mercado de trabalho que não lhe garante total estabilidade.

Adicionalmente, considerando a sobrevida dos aposentados, estimou-se que aproximadamente 6,49% dos trabalhadores teriam falecido antes de ter acesso ao seu benefício e que 509.763 homens e 78.453 mulheres teriam probabilidade abaixo de 30% de estarem vivos. Além disso, foi calculado o percentual de repasse que os beneficiários receberiam considerando a regra dos 60% + 2% por ano trabalhado, observando que apenas 3% dos trabalhadores teriam acesso a média integral de todos seus salários, demonstrando a clara dificuldade que as novas regras implicarão para os trabalhadores gozarem desse benefício.

É evidente que caso as regras previstas pela PEC 06/2019 já estivessem em vigor desde o primeiro ano analisado, o cenário resultante seria diferente do estipulado pelos resultados, afinal, tais regras estimulariam os trabalhadores a terem um perfil diferente no mercado de trabalho. No entanto, é valido considerar os principais desafios que a proposta apresentará para a população, como forma de refletir se essa é melhor alternativa para solucionar o atual problema existente no regime previdenciário nacional.

6 Considerações Finais

A população brasileira está diante de uma possível reforma previdenciária que causará impactos diretos na vida de todos os trabalhadores. Logo, é de extrema importância que a sociedade compreenda as razões que motivaram o governo propor tal medida e os efeitos que tais alterações tendem a causar no cenário social. Paralelamente, é possível observar que o advento da ciência de dados aliada à disseminação de dados abertos governamentais, apresentam-se como importantes ferramentas para a realização de análises cada vez mais complexas e precisas acerca da problemática descrita.

Embora tal abordagem apresente-se em evidência na literatura, destaca-se a ausência de trabalhos que analisem os impactos socioeconômicos relacionados a temática abordada. Isto é, investigando o estado da arte, observou-se uma gama de pesquisas que avaliaram as participações de aposentadorias e pensões na distribuição de renda em um caráter nacional - impossibilitando interpretações em unidades territoriais menores -, e estudos que investigaram os impactos macroeconômicos que a reforma da previdência poderia causar.

Diante disso, esta dissertação visou explorar a atual situação do sistema previdenciário brasileiro a partir da aplicação de técnicas e estratégias de Ciência de Dados em Dados Abertos Governamentais da Previdência Social. Para isso, foram apresentados dois estudos de caso, objetivando, de forma geral, compreender o impacto socioeconômico que o atual cenário previdenciário nacional causa na população, e simular os efeitos que a reforma prevista pela PEC 06/2019 pode provocar na concessão de benefícios. Com isso, a primeira pesquisa utilizou a metodologia de decomposição do índice de Gini e os microdados disponibilizados pelos censos demográficos de 2000 e 2010 para aferir o impacto que as aposentadorias e pensões causam na distribuição de renda dos estados e municípios brasileiros.

Avaliando os resultados, nota-se que houve um aumento da participação dessa parcela de benefícios na composição de rendimentos da população, e uma diminuição no coeficiente de desigualdade econômica. Todavia, embora as aposentadorias e pensões apresentassem caráter regressivo na distribuição de riquezas, observou-se que os benefícios até um salário mínimo colaboravam para a diminuição da concentração de renda. Tal fator destacou-se nos municípios da região nordeste, que, sobretudo, apresentou os menores índices de médias salariais do Brasil. Além disso, destacam-se as interdependências existentes entre as variáveis analisadas, evidenciando a correlação negativa entre a participação dos benefícios avaliados com o índice de Gini, indicando que ao passo que um aumenta o outro diminui.

Por outro lado, o segundo estudo utilizou os microdados de aposentadorias disponibilizados pela CPI da Previdência Social para simular os efeitos que as novas diretrizes, previstas na PEC 06/2019, causariam nos benefícios de 1995 a 2016 se as mesmas já estivessem em vigor na data em que cada aposentadoria foi concedida. Para isso, foram aplicadas estratégias de ciência de dados para facilitar o armazenamento dos dados, consultas e realização de futuras análises.

Dessa forma, após validar as informações disponibilizadas, observou-se que 83,28% das aposentadorias concedidas no período analisado não atendiam às novas regras previstas na reforma. Evidenciou-se, então, que a PEC 06/2019, aliada aos elevados índices de informalidade existentes no Brasil, dificultaria o acesso a tais benefícios, dado a dificuldade que trabalhadores da RGPS tem em manterem-se estáveis em seus empregos. Além disso, estimou-se que aproximadamente 6,48% dos beneficiários teriam falecido antes de dar entrada nas suas aposentadorias, e que somente 3% do total receberiam a média integral, ou mais, das suas contribuições como mensalidade.

Embora não tenham sido aplicadas técnicas de inteligência computacional nas análises realizadas, destaca-se a eficiência de utilizar ciência de dados em bases de dados abertas governamentais, uma vez que tal estratégia possibilitou a extração de conhecimento válido acerca de dados sem representatividade, podendo, dessa forma, servir como ferramenta para auxiliar possíveis tomadas de decisão em gestão pública.

Com objetivo de promover um trabalho colaborativo, foram disponibilizados todos os *scripts*, ferramentas e dados utilizados nessa dissertação na plataforma GitHub, possibilitando esforços da comunidade cientifica para o enriquecimento nas análises realizadas. Com isso, devido ao fato de compreender um campo interdisciplinar, a Previdência Social possibilita que pesquisadores de diversas áreas desenvolvam alternativas para o debate com diferentes óticas, possibilitando, dessa forma, a elaboração políticas públicas cada vez mais justas e eficientes diante do cenário brasileiro.

6.1 Dificuldades Encontradas

Durante o desenvolvimento do trabalho foram encontradas uma série de dificuldades relacionadas, principalmente, a dificuldade de acesso a informações e granularidade dos dados disponibilizados.

Dentre as principais dificuldades encontradas durante a realização deste trabalho, destacam-se:

- Ausência de microdados atualizados dos rendimentos de aposentadorias e pensões acerca dos municípios brasileiros, impossibilitando a realização de investigações mais precisas da atual participação que esses benefícios tem na distribuição de renda.

- Ausência das identificações de estados e municípios nos microdados da CPI da Previdência, impossibilitando análises territoriais mais precisas dos impactos que a PEC 06/2019 pode causar.
- Ausência das datas de cessação dos benefícios contidos na base da CPI da Previdência, impossibilitando uma avaliação histórica de benefícios ativos.
- A granularidade das informações disponibilizadas pelo AEPS, divulgando informações agregadas que impossibilitam análises mais detalhadas e complexas.

6.2 Trabalhos Futuros

Esta pesquisa encontra-se em desenvolvimento, possibilitando a realização de novas análises e discussões acerca da temática abordada. Dessa forma, os seguintes itens estão em processo de aperfeiçoamento e serão publicados em trabalhos futuros:

- A exploração complementar da metodologia de decomposição do índice de Gini, efeitos composição e concentração, em outras subcategorias de salários (2, 3 e 4 salários mínimos, por exemplo), para avaliar de forma mais precisão a participação que as aposentadorias e pensões possuem na distribuição de riquezas da população.
- A aplicação de métodos estatísticos espaciais, como a varredura espacial de Kulldorff, afim de detectar regiões de *clusters* acerca da das variáveis analisadas.
- A aplicação de técnicas de inteligência computacional, como Redes Bayesianas, para identificar as principais variáveis que influenciam na distribuição de renda da população.
- Utilizar as taxas de informalidade e desemprego para aferir, de forma mais precisa, o comportamento que as concessões de aposentadorias e pensões teriam se as regras previstas pela PEC 06/2019 já estivessem em vigor.
- Avaliar os impactos que a reforma da previdência causaria na distribuição de renda da população brasileira utilizando o índice de Gini e outras fontes de dados, como por exemplo a PNAD.

6.3 Publicações

Publicações realizadas durante o desenvolvimento desta dissertação:

(i) **FÉLIX JUNIOR, F. E. A.** ; PEREIRA, A. A. S. ; SANTOS, S. M. ; PUTY, C. A. C. B. ; REGO, L. P. ; SILVA, M. S. Impactos da reforma previdenciária: um estudo

acerca das concessões de benefícios. VIII Congresso Internacional Interdisciplinar em Sociais e Humanidades (CONINTER), 2019, UNIT/AL. Anais do(a) Anais do VIII CONINTER. Recife: Even3, 2019. v. 8. p. 1-15.

(ii) **FÉLIX JUNIOR, F. E. A.** ; SANTOS, S. M. ; FRANCES, C. R. L. ; MONTEIRO, M. A. ; SILVA, M. S. . Análise da participação das aposentadorias e pensões na distribuição de renda per capita do Brasil - 2000 e 2010. VIII Congresso Internacional Interdisciplinar em Sociais e Humanidades (CONINTER), 2019, UNIT/AL. Anais do(a) Anais do VIII CONINTER. Recife: Even3, 2019. v. 8. p. 1-15.

(iii) SANTOS, S. M. ; **FÉLIX JUNIOR, F. E. A.** ; FRANCES, C. R. L. ; REGO, L. P. ; SILVA, M. S. Ciência de dados aplicada em base de dados abertas: uma análise da economia dos municípios brasileiros - 2010 a 2017. VIII Congresso Internacional Interdisciplinar em Sociais e Humanidades (CONINTER), 2019, UNIT/AL. Anais do(a) Anais do VIII CONINTER. Recife: Even3, 2019. v. 8. p. 1-15.

(iv) OLIVEIRA, E. C. L. ; SANTOS, S. M. ; **FÉLIX JUNIOR, F. E. A.** ; SILVA, M. S. ; ARAUJO, J. P. L. Identificação paramétrica de moto CC utilizando algoritmos metaheurísticos e mínimos quadrado. XIV Congresso Brasileiro de Inteligência Computacional (CBIC), 2019, Belém. Anais do XIV CBIC, 2019.

(v) **FÉLIX JUNIOR, F. E. A.**; SANTOS, S. M. ; SILVA, M. S. ; CARVALHO, J. C. C. . Utilização de técnicas de machine learning para classificação de sinais de eletrocardiograma. XIV Congresso Brasileiro de Inteligência Computacional (CBIC), 2019, Belém. Anais do XIV CBIC, 2019.

Referências

BAESENS, B. *Analytics in a Big Data World: The Essential Guide to Data Science and Its Applications*. 1st. ed. [S.l.]: Wiley Publishing, 2014. ISBN 978-1-118-89274-9. Citado 4 vezes nas páginas 5, 6, 8 e 17.

BAPTISTELLA, J. C. F.; SOLANGE, I. de Souza de C.; FERREIRA, C. R. Concentração de Renda nas Macrorregiões Brasileiras: Estudo no Período 2003-2006. In: I Circuito de Debates Acadêmicos. Brasília - DF: IPEA, 2011. Citado 2 vezes nas páginas 24 e 28.

BARBOSA, C. de Freitas e R. A previdência social e as distorções na distribuição de renda. *Pesquisa & Debate*, v. 26, n. 1(47), 2015. ISSN 1806-9029. Citado na página 19.

BASTOS, P. P. Z. et al. *A Falsificação nas Contas Oficiais da Reforma da Previdência: O Caso do Regime Geral de Previdência Social*. Campinas, 2019. 17 p. Acessado em: 21 de Janeiro de 2020. Disponível em: <https://www.eco.unicamp.br/images/destaque/A-Falsificacao-nas-Contas-Oficiais-da-Reforma-da-Previdencia-Nota-CECON8.pdf>. Citado 2 vezes nas páginas 20 e 39.

BENESTY, J. et al. Pearson correlation coefficient. In: _____. *Noise Reduction in Speech Processing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN 978-3-642-00296-0. Citado na página 15.

Brasil. *Lei de Acesso a Informação - LAI*. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acessado em: 6 de Novembro de 2019. Citado na página 11.

Brasil. *Portal Brasileiro de Dados Abertos*. 2019. Disponível em: <http://dados.gov.br/>. Acessado em: 7 de Novembro de 2019. Citado na página 11.

BUSSAB, P. A. M. W. D. O. *Estatística Básica*. [S.l.]: Saraiva, 2017. 568 p. ISBN 9788547220228. Citado 3 vezes nas páginas 15, 16 e 17.

BUZETI, L. C. *Aplicação de Metodologia de Descoberta de Conhecimento em Dados Públicos: Explorando os Sistemas DATASUS e ALICEWEB*. Dissertação (Mestrado) — PPGGCO - UNICESUMAR, 2017. Citado na página 18.

CAMARANO, A. A.; FERNANDES, D. A previdência social brasileira. Instituto de Pesquisa Econômica Aplicada (Ipea), 2016. Citado na página 1.

CAO, L. Data science and analytics: A new era. *Springer*, International Journal of Data Science and Analytics, 2016. Citado na página 8.

CAO, L. Data science: A comprehensive overview. *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA, v. 50, n. 3, jun. 2017. ISSN 0360-0300. Citado na página 8.

CARDOSO, P. H. *Ciência de dados aplicada a dados governamentais abertos sob a ótica da Ciência da Informação*. Tese (Doutorado) — Universidade Federal do Rio de Janeiro - Escola de Comunicação, Rio de Janeiro, RJ, Brasil, 2019. Citado na página 2.

CHOO, C. The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions. *International Journal of Information Management*, v. 16, n. 5, p. 329 – 340, 1996. ISSN 0268-4012. Citado 2 vezes nas páginas 7 e 16.

COSTANZI, R. N. et al. *Reforma da Previdência Social*. [S.l.]: Instituto de Pesquisa Econômica Aplicada, 2018. (Texto para Discussão). ISBN 1415-4765. Citado na página 20.

COSTANZI, R. N.; ANSILIERO, G. *Impacto Fiscal da Demografia na Previdência Social*. [S.l.]: Instituto de Pesquisa Econômica Aplicada, 2017. (Texto para Discussão). Capítulo 20. ISBN 1415-4765. Citado 3 vezes nas páginas 1, 20 e 39.

DATAPREV. *DATAPREV: Quem Somos*. 2020. Disponível em: <https://portal.dataprev.gov.br/dataprev/quem-somos>. Acessado em: 19 de Fevereiro de 2020. Citado na página 1.

EHRL, P. Minimum comparable areas for the period 1872-2010: an aggregation of Brazilian municipalities. *Estudos Econômicos (São Paulo)*, scielo, v. 47, p. 215 – 229, 03 2017. ISSN 0101-4161. Citado na página 27.

EL-SAPPAGH, S. H. A.; HENDAWI, A. M. A.; BASTAWISSY, A. H. E. A proposed model for data warehouse etl processes. *Journal of King Saud University - Computer and Information Sciences*, v. 23, n. 2, p. 91 – 104, 2011. ISSN 1319-1578. Citado na página 42.

EMBARAK, O. *Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems*. [S.l.]: Apress, 2018. ISBN 978-1-48424-109-7. Citado na página 8.

EMC. *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. 1st. ed. [S.l.]: Wiley Publishing, 2015. ISBN 978-1-118-87613-8. Citado 3 vezes nas páginas 5, 16 e 17.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *Advances in Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1996. 1–34 p. ISBN 0-262-56097-6. Citado 3 vezes nas páginas 6, 7 e 16.

FERREIRA, C. R. Aposentadorias e distribuição da renda no Brasil: uma nota sobre o período 1981 a 2001. *Revista Brasileira de Economia*, v. 6, n. 3, p. 247–260, 2006. Citado 2 vezes nas páginas 24 e 38.

GENTIL, D. L. *Nota Técnica: Uma contribuição à avaliação das incosistências do modelo de projeção atuarial do governo federal e dos impactos de exclusão social da PEC 06/2019*. [S.l.], 2019. 32 p. Disponível em: <https://www.anfip.org.br/publicacoes/documentos/uma-contribuicao-a-avaliacao-das-inconsistencias-do-modelo-de-projecao-atuarial-do-governo-federal-e-dos-impactos-de-exclusao-social-da-pec-06-2019/>. Citado na página 54.

GHANI, N. A. et al. Social media big data analytics: A survey. *Computers in Human Behavior*, v. 101, p. 417 – 428, 2019. ISSN 0747-5632. Citado na página 5.

GIAMBIAGI, F.; MOREIRA, M. M. *A Economia Brasileira nos Anos 90*. [S.l.]: Banco Nacional de Desenvolvimento Econômico e Social, 1999. 496 p. ISBN 8587545019. Citado na página 25.

GIAMBIAGI, F.; PINTO, F.; ROTHMULLER, L. *Reforma previdenciária em 2019: elementos para uma tomada de decisão*. [S.l.]: Banco Nacional de Desenvolvimento Econômico e Social, 2018. 48 p. (Texto para Discussão). Citado na página 20.

GINI, C. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. [S.l.]: Tipogr. di P. Cuppini, 1912. (Studi economico-giuridici pubblicati per cura della facoltà di Giurisprudenza della R. Università di Cagliari, pt. 1). Citado 2 vezes nas páginas 12 e 17.

GORUNESCU, F. *Data Mining: Concepts, Models and Techniques*. [S.l.]: Springer, 2011. ISBN 978-3-642-19720-8. Citado na página 6.

GRUS, J. *Data Science do Zero: Primeiras Regras com Python*. 1st. ed. [S.l.]: Alta Books, 2016. ISBN 978-85-508-0387-6. Citado 3 vezes nas páginas 8, 16 e 17.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. [S.l.]: Morgan Kaufmann Publishers Inc., 2011. Citado 4 vezes nas páginas 5, 6, 7 e 16.

HOFFMANN, R. Queda da desigualdade de renda no brasil de 1995 a 2005, e delimitação dos relativamente ricos em 2005. *Revista Ipea*, Brasília, v. 1, 2006. Citado na página 12.

HOFFMANN, R. Desigualdade da distribuição da renda no brasil: a contribuição de aposentadorias e pensões e de outras parcelas do rendimento domiciliar per capita. *Economia e Sociedade*, Campinas, v. 18, n. 1, p. 213–231, 2009. Citado 5 vezes nas páginas 14, 18, 24, 28 e 34.

HOFFMANN, R.; LEONE, E. T. Participação da mulher no mercado de trabalho e desigualdade da renda domiciliar per capita no Brasil: 1981-2002. *Revista Brasileira de Economia*, Belo Horizonte, v. 14, n. 2, p. 35–38, 2004. Citado 3 vezes nas páginas 24, 28 e 38.

IBGE. *Sinopse do Censo Demográfico : 2010*. [S.l.]: Instituto Brasileiro de Geografia e Estatística, 2011. Citado na página 25.

IBGE. *Anuário Estatístico da PRevidência Social*. [S.l.]: DATAPREV, 2017. v. 24. 908 p. ISSN 0104-8139. Citado na página 42.

IBGE. *Projeções da População: Brasil e Unidades da Federação*. 2th. ed. [S.l.]: Série Relatórios Metodológicos, 2018. Citado na página 1.

IBGE. *Evolução da Malha Territorial*. 2019. Disponível em: <https://atlasescolar.ibge.gov.br/images/atlas/mapas_brasil/brasil_evolucao_malha_municipal.pdf>. Acessado em: 6 de Dezembro de 2019. Citado na página 27.

INSS. *Instituto Nacional do Seguro Social*. 2020. Disponível em: <https://www.inss.gov.br/acesso-a-informacao/institucional/>. Acessado em: 19 de Fevereiro de 2020. Citado 2 vezes nas páginas 1 e 2.

IPEA. *Ipea Data*. 2019. Disponível em: <http://www.ipeadata.gov.br/Default.aspx>. Acessado em: 22 de Novembro de 2019. Citado na página 23.

ISOTANI, S.; BITTENCOURT, I. *Dados Abertos Conectados: Em busca da Web do Conhecimento*. [S.l.]: Novatec, 2015. ISBN 9788575224496. Citado 2 vezes nas páginas 9 e 10.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, Taylor & Francis, v. 29, n. 4, p. 258–268, 2012. Citado 2 vezes nas páginas 10 e 11.

JOURNAL OF DATA SCIENCE. *About JDS*. 2019. Disponível em: <http://www.jds-online.com/about>. Acessado em: 21 de Novembro de 2019. Citado na página 8.

KENDALL, M. G. A New Measure of Rank Correlation. *Biometrika*, v. 30, n. 1-2, p. 81–93, 06 1938. ISSN 0006-3444. Citado na página 16.

KIMBALL, R.; CASERTA, J. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. [S.l.]: Wiley, 2011. ISBN 9781118079683. Citado na página 42.

KLEMETTINEN, M. A knowledge discovery methodology for telecommunication network alarm databases. 05 1999. Citado na página 6.

MAHESHWARI, A. *Business Intelligence and Data Mining*. [S.l.]: Business Expert Press, 2015. 162 p. ISBN 978-1-63157-121-3. Citado na página 7.

MANYIKA, J. et al. Open data: Unlocking innovation and performance with liquid information. *McKinsey Global Institute*, v. 21, 2013. Citado na página 10.

MINISTÉRIO DA FAZENDA. Evolução da Proteção Previdenciária no Brasil - 2017. *Informe de Previdência Social*, Brasília, 2018. ISSN 2318-5759. Citado 2 vezes nas páginas 1 e 24.

MORAN, P. A. P. Notes on continuous stochastic phenomena. *Biometrika*, [Oxford University Press, Biometrika Trust], v. 37, n. 1/2, p. 17–23, 1950. Citado na página 38.

MUKAKA, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi medical journal : the journal of Medical Association of Malawi*, v. 24, n. 3, p. 69–71, 2012. Citado na página 15.

NAKATANI-MACEDO, C. D. *O Comportamento dos Rendimentos das Aposentadorias e Pensões na Renda Domiciliar Per Capita - Brasil e Suas Regiões - 2004 a 2012*. Dissertação (Mestrado) — Universidade Estadual de Londrina, 2015. Citado 4 vezes nas páginas 19, 24, 28 e 32.

NAKATANI-MACEDO, C. D. et al. Population aging in the state of Parana and impact of pensions and retirements in income. *Economia e Região*, v. 3, n. 1, p. 27–38, 2015. Citado na página 19.

NERI, M. C. *A Escala de Desigualdade: Qual foi o Impacto da Crise sobre a Distribuição de Renda a Pobreza*. 2019. 34 p. Rio de Janeiro, RJ - Fundação Getulio Vargas (FGV). Citado na página 38.

NICHOLSON, B. *A previdência injusta: como o fim dos privilégios pode mudar o Brasil*. São Paulo: Geração Editorial, 2007. Citado 2 vezes nas páginas 24 e 38.

OBAMA, B. *Transparency and Open Government*. 2008. Washington, United States of America: The White House. Citado na página 10.

OKF. *Defining Open Data – Open Knowledge Foundation*. 2020. Disponível em: <https://blog.okfn.org/2013/10/03/defining-open-data/>. Acessado em: 20 de Fevereiro de 2020. Citado 2 vezes nas páginas 10 e 11.

OPEN DATA HANDBOOK. *The Open Data Handbook*. 2019. Disponível em: <http://opendatahandbook.org/guide/en/>. Acessado em: 10 de Novembro 2019. Citado na página 9.

OPEN DEFINITION. *The Open Definition*. 2019. Disponível em: <http://opendefinition.org/>. Acessado em: 6 de Novembro de 2019. Citado 2 vezes nas páginas 9 e 17.

OPG. *Open Government Partnership*. 2019. Disponível em: <https://www.opengovpartnership.org/>. Acessado em: 4 de Novembro de 2019. Citado na página 11.

PAL, S. K.; MEHER, S. K.; SKOWRON, A. Data science, big data and granular mining. *Pattern Recognition Letters*, v. 67, p. 109 – 112, 2015. ISSN 0167-8655. Granular Mining and Knowledge Discovery. Citado na página 5.

PLAMONDON, P. et al. *Prática Atuarial na Previdência Social*. [S.l.]: Ministério da Previdência Social, 2002. v. 33. (Coleção Previdência Social, v. 33). ISBN 978-85-88219-39-7. Citado na página 2.

PNAD. *Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua*. 2019. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/17270-pnad-continua.html?=&t=downloads>. Acessado em: 22 de Novembro de 2019. Citado 2 vezes nas páginas 24 e 47.

PROVOST, F.; FAWCETT, T. Data science and its relationship to big data and data-driven decision making. *Big Data*, v. 1, p. 51–59, 03 2013. Citado 5 vezes nas páginas 5, 8, 9, 16 e 17.

REZENDE, S. O. *Sistemas inteligentes: Fundamentos e Aplicações*. 1rd. ed. [S.l.]: Manole Ltda, 2003. Citado na página 6.

ROGATI, M. *The AI Hierarchy of Needs*. 2020. Disponível em: <https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007>. Acessado em: 20 de Fevereiro de 2020. Citado na página 8.

ROSSI, P.; MELLO, G. *Choque recessivo e a maior crise da história: A economia brasileira em marcha à ré*. Campinas, 2017. 5 p. Disponível em: <http://www3.eco.unicamp.br/images/arquivos/NotaCecon1_Choque_recessivo_2.pdf>. Citado na página 2.

RUD, O. P. *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. [S.l.]: John Wiley & Sons, 2009. 256 p. ISBN 0-470-48840-9. Citado 2 vezes nas páginas 7 e 16.

SANTOS, A. M. T. B. dos. *Mineração de Dados Educacionais: Um estudo sobre os Dados Socioeconômicos na Educação na Base de Dados do INEP*. Dissertação (Mestrado) — Universidade Federal do Pará - Programa de Pós Graduação em Engenharia Elétrica, 2019. Citado na página 18.

SECRETÁRIA DE POLÍTICA ECONÔMICA. *Relatório de Análise Econômica dos Gastos Públicos Federais Evolução dos Gastos Públicos Federais no Brasil: Uma análise para o período 2006-15*. 2016. 18 p. Disponível em: <http://www.fazenda.gov.br/centrais-de-conteudos/publicacoes/relatorio-de-analise-economica-dos-gastos-publicos-federais/arquivos/2016/relatorio_gasto_publico_federal_site.pdf>. Acessado em: 20 de Janeiro de 2020. Citado na página 2.

SILVA, C. P. A. *Uma Metodologia para Aferição da Acurácia de Modelos de Projeção de Longo Prazo para a Previdência Social no Brasil*. Tese (Doutorado) — Universidade Federal do Pará - Programa de Pós-Graduação em Engenharia Elétrica, Belém, Pará, Brasil, 2017. Citado 2 vezes nas páginas 1 e 39.

SILVA, C. P. A. et al. Financial Forecasts Accuracy in Brazil's Social Security System. *PLoS ONE*, v. 12, n. 8, p. 1–20, 2017. Citado 3 vezes nas páginas 20, 39 e 40.

SILVA, R. R. d.; BACHA, C. J. C. Polígonos de Voronoi como alternativa aos problemas das áreas mínimas comparáveis: uma análise das mudanças populacionais na região norte do Brasil. scielo, v. 28, p. 133 – 151, 06 2011. ISSN 0102-3098. Citado na página 27.

SOLÓN, A. *A Previdência Social e a Economia dos Municípios*. 7. ed. São Paulo: ANFIP, 2019. ISBN 978-85-62102-32-5. Citado 2 vezes nas páginas 19 e 24.

SOUZA, P. H. G. F. de; MEDEIROS, M. The concentration of income at the top in Brazil, 2006-2014. *International Policy Centre for Inclusive Growth*, n. 163, 2017. Citado na página 23.

THE WORLD BANK. *GINI index (World Bank estimate)*. 2019. Disponível em: <https://data.worldbank.org/indicator/SI.POV.GINI>. Acessado em: 21 de Novembro de 2019. Citado 3 vezes nas páginas 23, 28 e 32.

THOMAS, V.; WANG, Y.; FAN, X. Measuring education inequality: Gini coefficients of education. *World Bank Policy Research Working*, 2001. Disponível em: <https://ssrn.com/abstract=258182>. Citado na página 12.

UNITED NATIONS. *Human Development Report 2015: Work for Human Development*. [S.l.]: UNDP, 2015. ISBN 978-92-1-126398-5. Citado na página 23.

UNITED NATIONS. *World Population Prospects: The 2017 Revision*. 2017. Disponível em: <https://population.un.org/wpp/Publications/Files/WPP2017_Wallchart.pdf>. Acessado em: 21 de Janeiro de 2020. Citado na página 1.

USA. *The home of the U.S. Government's open data*. 2019. Disponível em: <https://www.data.gov/>. Acessado em: 7 de Novembro de 2019. Citado na página 10.

VANDERPLAS, J. *Python Data Science Handbook: Essential Tools for Working with Data*. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2016. ISBN 978-1-491-91205-8. Citado 4 vezes nas páginas 8, 9, 17 e 28.

WU, X. et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, v. 14, n. 1, p. 1–37, 2008. ISSN 0219-3116. Citado na página 7.

YITZHAKI, S.; SCHECHTMAN, E. *The Gini Methodology: A Primer on a Statistical Methodology*. [S.l.]: Springer New York, 2012. (Springer Series in Statistics). ISBN 9781461447207. Citado 2 vezes nas páginas 12 e 13.

Apêndices

APÊNDICE A – Atributos dos Microdados Disponibilizados pela CPI da Previdência

Campo	Descrição Campo	Código	Descrição Código
ESPECIE	Espécie do Benefício	31	Auxilio doença previdenciário
		32	Aposentadoria por invalidez previdenciária
		41	Aposentadoria por idade
		42	Aposentadoria por tempo de serviço previdenciária
		46	Aposentadoria por tempo de serviço especial
		57	Aposentadoria por tempo de serviço de professor
		91	Auxílio doença acidentário
		92	Aposentadoria por invalidez por acidente do trabalho
DIB	Data de Início do Benefício		
DDB	Data de Despacho do Benefício		
MOT_CESSACAO	Motivo de Cessação *		
ULT_COMPET_MR	Competência da Cálculo da Última Mensalidade Reajustada		
VL_MR	Valor da Última Mensalidade Reajustada		
DT_NASC	Data de Nascimento do Titular		
VL_RMI	Valor da Renda Mensal Inicial		
CLIENTELA	Clientela	1	Masculino
		3	Feminino
		9	Ignorado
SITUACAO	Situação do Benefício	0	Ativo
		2	Cessado
		3	Suspenso
		4	Suspenso por Marca de Erro
		5	Cessado por Cess do Titular Proprio (Pa)
		7	Bloqueado Pelo Controle de Pagamento
		8	Cessado Pelo Sistema de Obitos da Dtp
		10	Recebendo Mensalid de Recuper 6 Meses
		11	Recebendo Mensalid de Recuper 18 Meses
		12	Suspenso Pelo Sistema de Revisao Rur/Urb
		16	Beneficio Suspenso Pela Auditoria
		17	Beneficio Cessado Pela Inspetoria
		18	Beneficio Cessado Pela Auditoria
		22	Cessado Pela Revisao Rural/95
		23	Susp. Pelo Sistema de Obitos da Dataprev
		24	Cancelado Pela Auditoria
DT_OBITO	Data de Óbito		
TEMPO_CONTRIB	Tempo de Contribuição, em Anos, do Titular		
IDADE_DIB	Idade na DIB do Titular		

* Códigos com os motivos de cessação listados a seguir.

Código	Descrição	Código	Descrição
2	NAO COMPROVACAO DE FE DE VIDA	51	CESSACAO POR REVISAO RURAL/95
3	CESSACAO POR SUSPEITA DE OBITO/SIM	52	ERRO ADM. INFORMADO PELA AUDITORIA
4	REVBPC NAO LOCALIZ. DO BENEFICIARIO	53	FRAUDE INFORMADA PELA AUDITORIA
5	NAO APRES. DE DOCS. OBRIGATORIA	54	LIMITE MEDICO INFORMADO P/ PERICIA
6	NAO ATENDIMENTO A CONVOC.POSTO	55	IRREGULARID./ERRO MEDICO-PERICIAL
7	BPC>2 ANOS - APRENDIZ C/ DEFIC.	56	RECUP.LESAO C/ DESCARAC. DO BENEF.
8	BENEFICIO CONCEDIDO COM CESSACAO DA DIB	57	CESS. DE AUXILIO RECLUSAO POR FUGA
9	DCA ACP2005.33.00.020219-8	58	BENEFICIO C/DCI C/ MAIS DE 60 DIAS
10	CESSACAO POR SUSPEITA DE OBITO	59	CESSADO, DIVERGENCIA DADOS CNIS
11	CESSACAO POR NAO COMPARECIMENTO AO CENSO	60	CESS.DE BENEFICIO FORA DE CADASTRO
12	LIMITE MEDICO	61	RECUSA AO PROGR. REABILIT. PROFIS.
13	OBITO DO TITULAR DO BENEFICIO	62	REVBPC N+DEF LON PZ/R>1/4S
14	CESSACAO P/ACAO REVISIONAL COMPARTILHADA	63	CESSACAO P/ FALTA DE DIGITALIZACAO
15	OPCAO CONCESSAO BENEFICIO PREVIDENCIARIO	64	OBITO SEGURADO INSTIT.AUX.RECLUSAO
16	CASAMENTO	65	BENEF. SUSPENSO POR MAIS DE 6 MESES
17	REAPARECIMENTO DO INSTITUIDOR	66	VOLTA TRABALHO-ATIV.INSALUBRE(B46)
18	DESISTENCIA BPC-LOAS PELO TITULAR	67	CESSACAO POR CONCESSAO DE B80
19	CESS. PA DEVIDO CESS. BENEF. INST.	68	REMUNERACAO APOS A DIB
20	DESISTENCIA ESCRITA TITULAR DO BENEFICIO	69	ALTA MEDICA
21	TRANSFORMACAO B87 EM B88	70	RETORNO VOLUNTARIO AO TRABALHO
22	PRORROGACAO DE BENEFICIO ANTERIOR	71	ERRO TECNICO
23	CONCES. APOSENT. P/RECLUSO	72	NAO COMPARECIMENTO
24	SUSPENSAO 64 P/MAIS 6MESES	73	RET./PERMANENC.NA ATIV.COND.ESPEC.
25	NB TRANSITADO JULG/REV.ADM	74	CANCELAMENTO POR FRAUDE/AUDITORIA
26	RENDA PER CAPITA - REVBPC	75	CANCELAMENTO P/IRREGULARIDADE/AUDITORIA
27	PM DENEGATORIA - REVBPC	76	CANCELAMENTO P/ERRO ADM.MEDICO/AUDITORIA
28	TRANSFORMACAO PARA OUTRA ESPECIE	77	IRREGULARIDADE DETECTADA PELA AUDITORIA
29	CONCESSAO DE OUTRO BENEFICIO	78	CESS. DE B80 C/CONTRATO TEMPORARIO
30	CONSTATACAO DE FRAUDE	79	CESSACAO DE B80 (120/134 DIAS)
31	CONSTATACAO IRREGULAR./ERRO ADM.	80	ABORTO NAO CRIMINOSO
32	DECISAO DE CESSACAO POR RECURSO	81	OBITO INFORMADO PELA REVBPC
33	DECISAO JUDICIAL	82	CESSACAO DE B80 (60 DIAS)
34	VOLTA AO TRABALHO	83	CESSACAO DE B80 (30 DIAS)
35	BENEFICIO SEM DEPENDENTE VALIDO	84	CESSACAO - OPCAO RECEB. PELO M. EXERCITO
36	ACUMULACAO INDEVIDA DE BENEFICIOS	85	BENEFICIO COM NIT ERRADO
37	SUSP SISOBI MAIS 6 MESES	86	CESSACAO B94/95 PARECER MEDICO CONTRARIO
38	CESSACAO ABONO P/ CONCES. APOSENT.	87	CESSACAO ACUMULACAO INDEVIDA REV. 2003
39	NAO ATENDIMENTO CONVOCACAO INSPET.	88	REVBPC N HA+DEF LONG PZ
40	CESS.P/ RECUP.TOTAL DENTRO 5 ANOS	89	CESSACAO AMPARO LEI 10559/02.
41	CESS.P/ RECUP. PARCIAL APOS 5 ANOS	90	CESSACAO DE PA DEVIDO DATA LIMITE
42	CESSADO P/ SIST. DE OBITOS(SISOBI)	91	REVISAO DE ACORDAO
43	AUX.RECL.-CUMPR.PENA,CONDIC.,ALBERG.	92	LIMITE INDEFINIDO S/CONCESSAO DE B32/92
44	AUX.DOENCA NAO COMPAREC. A PERICIA	93	CESSACAO BATIMENTO FUNASA
45	CESS. BENEFICIO - RECIPROCA	94	ALTA VOLUNTARIA
46	ESTATUTARIO TRANSF. P/ORGAO ORIGEM	95	NAO COMPARECIMENTO A REABILITACAO PROF.
47	CESS. DEVIDO PERDA QUALID. DEPEND.	96	TRANSFERENCIA P/ORGAO ORIGEM LEI 8878/94
48	BENEF. CESSADO NO SISTEMA ANTIGO	97	COMPROVADA MA FE DO BENEFICIARIO
49	OBITO INFORMADO PELA AUDITORIA	98	CESSACAO POR LEI
50	CONCESSAO DE AUXILIO FUNERAL (PAB)	99	OBITO DEPENDENTE INFORM. P/CENSO(HIPNET)