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

Data Science Aplicado a Dados Abertos do Governo Federal: Estudos de Caso sobre a Economia dos Municípios Brasileiros

Sandio Maciel dos Santos

DM: 13/2020

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

Sandio Maciel dos Santos

Data Science Aplicado a Dados Abertos do Governo Federal: Estudos de Caso sobre a Economia dos Municípios Brasileiros

Dissertação de Mestrado submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA como requisito 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)

S237d Santos, Sandio Maciel dos
Data Science Aplicado a Dados Abertos do Governo Federal : Estudos de Caso sobre a Economia dos Municípios Brasileiros / Sandio Maciel dos Santos. — 2020.
xiv, 55 f. : il. color.

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

1. Data Science. 2. Dados Abertos. 3. processo de KDD. 4. Dados Governamentais Brasileiros. I. Título.

CDD 006.312

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"DATA SCIENCE APLICADO A DADOS ABERTOS DO GOVERNO FEDERAL: ESTUDOS DE CASO SOBRE A ECONOMIA DOS MUNICÍPIOS BRASILEIROS"

AUTOR: **SÂNDIO MACIEL DOS SANTOS**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 13/03/2020

BANCA EXAMINADORA:

Prof. Dr. Marcelino Silva da Silva
(Orientador – PPGEE/UFPA)

Prof. Dr. Carlos Renato Lisboa Francês
(Avaliador Interno – PPGEE/UFPA)

Prof. Dr. Cláudio Alberto Castelo Branco Puty
(Avaliador Externo ao Programa – ICSA/UFPA)

Prof.ª Dr.ª Liviane Ponte Rego
(Avaliadora Externa ao Programa – ICSA/UFPA)

VISTO:

Prof.ª Dr.ª Maria Emília de Lima Tostes
(Coordenadora do PPGEE/ITEC/UFPA)

A minha Mãe
Sônia Maria Maciel dos Santos e Familiares
por sempre estarem comigo em todos os momentos.

Agradecimentos

Primeiramente, gostaria de agradecer a Deus, que durante toda a minha vida, foi o Deus com um amor incalculável e, um pai de imensurável proteção, que me proporcionou inúmeras conquistas e graças.

Gostaria também de agradece a minha mãe Sônia Maria, a qual é responsável por essa conquista, pois sempre foi um exemplo de determinação e dedicação em todas as áreas da minha vida, através de ensinamentos de conselhos que foram de suma importância para construção do caráter e valores éticos e morais.

Aos meus irmãos Celso Júnior, Sérgio Santos e a Suelane Santos, pelo apoio emocional e aos meus familiares como um todo: tios, primos, avos, que sempre contribuíram de forma direta ou indireta com essa conquista. Em especial agradeço a Adrihele Leal pela paciência que tem comigo, carinho e compreensão nas horas mais difíceis.

Aos professores do PPGEE pelos ensinamentos ao longo desses dois anos de pós-graduação, em especial ao meu orientador Prof. Dr. Marcelino Silva da Silva por ter aceitado o grande desafio de me orientar, por todo o conhecimento repassado e por toda a paciência. Também, agradeço a Prof.ª Dr.ª Liviane Ponte Rêgo por todo os ensinamentos, conselhos, puxões de orelha e pela cobrança de bons resultados.

Aos amigos dos grupos de pesquisa do LINC e do LPO, os quais tive e tenho orgulho de fazer parte, por serem companheiros nessa longa jornada e todo apoio para a conclusão deste trabalho, em especial a Ewerton Oliveira, Lucas Caldas, Francisco Eguinaldo, Daniel Victor, Igor Falcão e André Pereira, pois sem eles o trabalho dificilmente teria sido concluído.

Ao CNPq -Conselho Nacional de Desenvolvimento Científico e Tecnológico e ao CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), por financiar minha pesquisa por meio de uma bolsa de estudos.

Por fim, e não menos importante agradeço a Universidade Federal do Pará (UFPA) via ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE) pela oportunidade acadêmica e profissional de aprendizado e todo apoio.

" Ensinar não é transferir conhecimento,
mas criar as possibilidades para a sua
própria produção ou a sua construção."
(Paulo Freire)

Resumo

O processo de análise de dados nos últimos anos obteve bastante destaque no cenário brasileiro a partir da concessão da Lei 12.527/2011, a qual garante o acesso à informação pública, permitindo uma melhor transparência dos gastos públicos pela sociedade. Aliado à isso, inúmeras discussões surgiram em torno da utilização dos microdados governamentais brasileiros, entre elas destacamos as discussões sobre a reforma da previdência social e as análises voltados a saúde fiscal dos municípios brasileiros através de abordagens previdenciárias. Assim, este trabalho foca na utilização da *Data Science*, especificamente no processo de KDD para analisar microdados dos municípios brasileiros. Assim, neste trabalho são feitas duas abordagens diferentes nos a primeira realiza uma análise estatística descritiva e sem inferências, para compreender a saúde fiscal dos municípios brasileiros entre 2010 a 2017, por meio das transferências do RGPS. A segunda abordagem de análise fiscal via o modelo STVAR através das seguintes variáveis: despesa, receita e PIB do município de São Paulo. Os resultados da análise I apontam que municípios que possuem populações superior 100 habitantes não exibem défice mediante a diferença entre arrecadações municipais e transferências do RGPS. Na análise II os resultados encontrados mostram que ciclo econômico analisado ao sofrer choque exógeno (ou, impulso externo) pode gera alteração nos estados de recessão e expansão como duração média de 12 meses.

Palavras-chaves: *Data Science*. Dados Abertos. processo de KDD e Dados Governamentais Brasileiros.

Abstract

The process of analyzing open databases in recent years has gained considerable prominence in the Brazilian scenario since the granting of Law 12,527 / 2011, which guarantees access to public information, allowing for better transparency of public spending by society. Allied to this, numerous discussions arose around the use of Brazilian government microdata, among which we highlight the discussions on social security reform and the analysis of fiscal health in Brazilian municipalities through social security approaches. Thus, this work focuses on the use of Data Science, specifically in the KDD process to analyze microdata from Brazilian municipalities. Thus, in this work, two different approaches are made, the first of which performs a descriptive statistical analysis without inferences, to understand the fiscal health of Brazilian municipalities between 2010 and 2017, through transfers from the RGPS. The second approach to fiscal analysis using the STVAR model through the following variables: expenditure, revenue, and GDP of the municipality of São Paulo. The results of analysis I show that municipalities with populations greater than 100 inhabitants do not show a deficit due to the difference between municipal collections and transfers from the RGPS. In analysis II, the results found show that the economic cycle analyzed when undergoing exogenous shock (or external impulse) can generate changes in the states of recession and expansion with an average duration of 12 months.

Keywords: Data Science. Open Data. KDD process and Brazilian Government Data.

Lista de ilustrações

Figura 1 – Estimativa de crescimento de dados digitais de 2010 a 2020. 1
Figura 2 – Competência essenciais para a ciência de dados. 9
Figura 3 – Etapas operacionais do processo de KDD 10
Figura 4 – Percurso metodológico adotado. 24
Figura 5 – Percentual dos repasses previdenciários emitidos em 2010 e 2017. . . . 30
Figura 6 – Total de emissões de aposentadorias em 2017 por região. 30
Figura 7 – Percentual de aposentadorias emitidas a região norte em 2017. 31
Figura 8 – Diferença entre as transferências previdenciárias e arrecadações dos municípios com mais de 100 mil habitantes durante 2010 e 2017. 31
Figura 9 – Diferença entre o endividamento dos municípios com até 5 mil habitantes por região. 33
Figura 10 – Séries fiscais municipais observadas e ajustas via ARIMA X-13. 36
Figura 11 – Fluxograma do modelo STVAR. 37
Figura 12 – Datação dos períodos de recessão do ciclo econômico pelo CODACE. . 39
Figura 13 – Estimação dos períodos de recessão via função $F(z_t)$. 41
Figura 14 – Influência dos choques exógenos nos gasto G_t e a resposta no produto P_t. 42
Figura 15 – Influência dos choques exógenos nos gasto G_t e a resposta na receita R_t. 43
Figura 16 – Influência dos choques exógenos nos gasto G_t e a resposta nos gastos G_t. 44

Lista de tabelas

Tabela 1 – Descrição e origem dos microdados utilizados no estudo de caso I. . . . 27
Tabela 2 – Variáveis selecionadas para análise. 28
Tabela 3 – Discretização dos municípios número de habitantes. 28
Tabela 4 – O comprometimento do PIB pela diferença entre arrecadações e despesas. 32
Tabela 5 – Percentual de municípios em que FPM e menor que os benefícios pagos. 33
Tabela 6 – Comparação entre o IDH's nacionais com os IDH's das regiões brasileiras para os municípios que possuem até 5 mil habitantes. 34
Tabela 7 – Variáveis utilizadas no estudo de caso II. 35

Lista de abreviaturas e siglas

ABC	Algoritmo de Colônia Artificial de Abelhas
AG	Algoritmo Genético
ANFIP	Associação Nacional dos Auditores Fiscais da Receita Federal do Brasil
ANOVA	Análise de variância
ARIMA	Modelo Auto-Regressivo Integrado de Médias Móveis
CODACE	Comitê de Datação de Ciclos Econômico
CSV	*comma-separated-values*
EQM	Erro Médio Quadrático
FRINBRA	Finanças do Brasil
FPM	Fundo de Participação dos Município
IDH	Índices de Desenvolvimento Human
IBGE	Instituto Brasileiro de Geografia e Estatística
IPEA	Instituto de Pesquisa Econômica Aplicada
IPCA	Índice de Preços ao Consumidor Amplo
JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
LOAS	Lei Orgânica da Assistência Social
MIMO	*Multiple-Input-Multiple-Outpu*
MQ	Mínimos Quadrados
PBC	Benefício assistencial à pessoa com deficiência
PDF	*Portable Document Format*
PEC	Proposta de Emenda Constitucional
PIB	Produto Interno Bruto

PSO	Optimização por Enxame de Partícula
RGPS	Regime Geral de Previdência Social
SEADE	Fundação de Sistema Estadual de Análise de Dados
SIGEO-BI	Sistema de Informações Gerenciais da Execução Orçamentária
SISO	*Simple-Input-Simple-Output*
STN	Tesouro Nacional
STVAR	*Smooth Transition Vector Autoregressive*
Sys-GMM	*Generalized Method of Moment Estimation*
VAR	Modelo de Vetores Autorregressivos
WEB	*World Wide Web*

Lista de símbolos

Ω_E	Matrizes de variância-covariância de expansão
Ω_R	Matrizes de variância-covariância de recessão
γ	Parâmetro de ajuste
Π_E	Coeficientes de expansão
Π_R	Coeficientes de recessão
ε_t	Ruído do modelo
$F(z_t)$	Função para calcular dinâmica do modelo
x_t	Saída do modelo
G_t	Vetor de gastos
R_t	Vetor de receitas
P_t	Vetor de PIB
z_t	Média da taxa de crescimento do produto

Sumário

1 INTRODUÇÃO 1
1.1 Contextualização e Justificativa 1
1.2 Bases de Dados Abertas 2
1.3 Economia dos Municípios Brasileiros 4
1.4 Objetivo Geral 6
1.4.1 Objetivos Específicos 6
1.5 Organização do Texto 7

2 REFERENCIAL TEÓRICO 8
2.1 Ciência de dados 8
2.2 Descoberta de Conhecimento em Bases de Dados 9
2.2.1 Pré-Processamento de Dados 11
2.2.2 Mineração de Dados 11
2.2.3 Pós-Processamento de Dados 12
2.3 Análise Estatística de Dados 12
2.4 Modelos de Otimização Matemáticos 13
2.4.1 Método dos Mínimos Quadrados 13
2.4.2 Método dos Mínimos Generalizados - MIMO 15
2.5 Modelo de Estimação Não linear - STVAR 16
2.6 Considerações Finais 17

3 TRABALHOS RELACIONADOS 18
3.1 Ciência de Dados Aplicada em Dados Demográficos 18
3.2 Análises de Dados Econométricas 19
3.3 Considerações Finais 21

4 MATERIAIS E MÉTODOS 22
4.1 Ambiente Computacional 22
4.2 Metodologia Aplicada 23
4.3 Considerações Finais 24

5 RESULTADOS E DISCUSSÕES 26
5.1 Estudo de caso I 26
5.1.1 Aquisição de Dados 26
5.1.2 Seleção de Dados 27
5.1.3 Tratamento de Dados 28

5.1.4 Processamento de Dados 29
5.1.5 Resultados 29
5.2 Estudo de caso II 34
5.2.1 Aquisição de Dados 34
5.2.2 Seleção de Dados 35
5.2.3 Tratamento de Dados 35
5.2.4 Processamento de Dados 37
5.2.5 Resultados 41
5.3 Considerações Finais 45

6 CONCLUSÕES E TRABALHOS FUTUROS 46
6.1 Conclusões 46
6.2 Trabalhos Futuros 47
6.3 Dificuldades 48
6.4 Publicações 48
6.5 Publicações Adicionais 48

REFERÊNCIAS 50

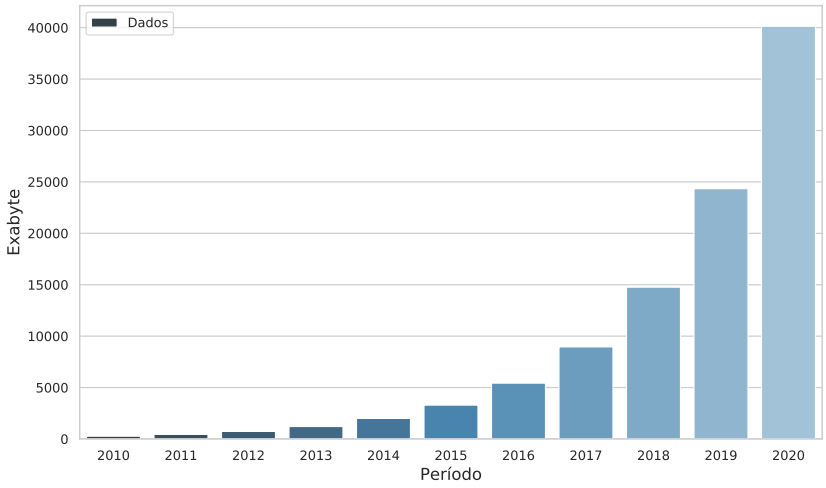
1 Introdução

1.1 Contextualização e Justificativa

Atualmente, o acúmulo de dados e de informações em diversas áreas do conhecimento, tem feito com que estudos voltados a análise de dados tenha se tornado uma das habilidades mais requeridas nos últimos anos (MUELLER, 2019). Pois, é a partir deles que alguns órgãos governamentais e outras instituições em geral realizam pesquisas que possam contribuir com a melhoria da vida social e agregar algum tipo de valor de informação (ESPÍNDOLA et al., 2018).

Em Isotani e Bittencourt (2015) é descrito que estudos realizados no período de 2006 a 2010 indicavam que a quantidade de dados digitais gerado no mundo aumentou de 166 *exbaytes* para 988 *exbaytes*, e também estimou-se que no ano 2014 estes dados já apresentavam-se em *zetabytes* (FLORIDI, 2010). No que diz respeito ao crescimento da geração de dados, prepõem-se que em 2020 o volume dados deve alcançar 40 *zetabytes* (GANTZ; REINSEL, 2012), conforme a Figura 1.

Figura 1 – Estimativa de crescimento de dados digitais de 2010 a 2020.



Fonte: Adaptado de Isotani e Bittencourt (2015)

A grande quantidade de dados gerados representa um enorme potencial de conhecimento, porém o processo compreensão dos resultados em sua maioria não é um processo trivial. Mediante isso, a sociedade acadêmica tem buscado novos mecanismos computacionais para falicitar a interpretação dos resultados, a partir dos diferentes tipos de dados, independentemente da sua área de conhecimento. Entre os diferentes ciências de análises de dados temos o *Data Science*, que nos últimos anos vem se destacando.

A Ciência de Dados (do inglês, *Data Science*) é uma ciência voltada para o estudo de dados a qual visa a extração de informações relevantes, para solucionar um determinado problema (VANDERPLAS, 2016). Trata-se de uma ciência interdisciplinar, pois é aplicada em diferentes áreas do conhecimento, como: economia, computação, estatística, meteorologia, dentre outras (CURTY; CERVANTES, 2016).

Dessa forma, esta ciência faz uso de técnicas amplamente disseminadas no campo de inteligência computacional: Aprendizado de Maquina, Redes Neurais e Mineração de dados (VIOLINO, 2018), e também de outras áreas do conhecimento, como por exemplo: econométricas, engenharia de *software* e bioinformática com finalidade de analisar relações entre os dados estudados (CRUZ, 2018).

No cenário mundial a ciência de dados vem se destacando através de estudos voltados para analises em bases de dados abertas e públicas, cujo, objetivo é extrair informações que futuramente podem ser usadas para a tomada de decisão de curto, médio e longo prazo. No Brasil, o processo de aquisição de conhecimento em bases de dados, ganha destaque na cenário atual brasileiro, quando o acesso aos dados governamentais passou a ser público a sociedade em geral, na seção seguinte é realizado um panorama sobre dados abertos.

1.2 Bases de Dados Abertas

O termo dados abertos teve início no ano 1995 através de um documento científico de uma agência norte americana. Neste documento foi debatido o livre acesso aos dados ambientais e geofísicos de diferentes países, com intuito de obter a colaboração de diferentes pesquisadores para analisar e compreender as informações referentes às características dos fenômenos naturais contidos nos dados (CHIGNARD, 2013).

A partir desse documento, alguns países que disponibilizavam o acesso a dados não governamentais passaram a impulsionar o livre acesso aos dados governamentais que colaboram positivamente na ampliação de estudos que visam melhorar transparência dos gastos públicos. Esse acesso aos dados ganhou maior estímulo após o memorando do presidente Barack Obama, que na ocasião tratava sobre a Transparência e Dados Governamentais norte-americano (OBAMA, 2008). A partir dessa perspectiva, 8 países, incluindo o governo brasileiro, formalizaram política de dados abertos em 2011 denominada de *Open Government Partnership* (MUELLER, 2019).

No entanto, em 18 novembro de 2011 vigora a Lei 12.527/2011 a qual garante o acesso à informação pública, em seu artigo 8, reconheceu a necessidade de disponibilizar dados governamentais em formato aberto. Os dados abertos governamentais fazem parte da política de acesso à informação do governo federal (INDA) (BRASIL, 2019a).

Os dados abertos têm o papel primordial na difusão de informações de uma determinada área, as quais podem ser livremente utilizadas e reutilizada inúmeras vezes por qualquer instituição e/ou pessoa física ou jurídica. Conjuntamente, ao acesso livre de dados está a divulgação da fonte de aquisição, a qual tenta evitar qualquer tipo de restrição ao seu acesso (OPEN DATA HANDBOOK, 2019). Na tentativa de evitar retenção de informação a própria definição de dados abertos sustenta três preceitos fundamentais (OPEN DATA COMMONS, 2019).

- **Disponibilidade de acesso:** Os dados devem estar disponíveis como um todo sendo possível acessá-los ou baixá-los via internet de qualquer local, seu custo não deve ser superior ao próprio custo da sua reprodução. A reprodução também deve estar disponível em formato adequado e modificável.
- **Reúso e redistribuição:** Os dados devem apresentar características legíveis que permitam a sua reutilização e redistribuição, bem como, fusão como outras bases de dados caso essa seja a vontade do seu manipulador sem que esse altere a sua essência.
- **Participação universal:** Consiste no acesso e divulgação dos dados para sociedade como um todo sem discriminação de área, ou seja, a reutilização e redistribuição dos dados deve ser garantida independentemente da sua utilização.

No cenário internacional a política de dados abertos pode ser vista como um termômetro de desenvolvimento socioeconômico da sociedade atual. Em (FLORIDI, 2010), é observado que os países membros do G7 possuem 70% do PIB dependentes de bens relacionados à informação pública.

A partir dos estímulos voltados para a abertura de dados, houve a necessidade de entender como estes são armazenados e conectados (MUELLER, 2019), pois a forma que estes são alocados pode potencializar e melhorar do aprimoramento de técnicas (de associação, agrupamento, classificação e *etc.*) usadas na análise de dados. Isso passou à favorecer o surgimento de novas demandas de exploração de dados, entre elas, podemos destacar estudos sobre microdados referentes a economia brasileira, especificamente dos municípios.

Dessa forma, a sociedade passa a perceber a importância das informações contidas nos microdados, bem como as vantagens sociais (transparência dos gastos públicos, melhora dos serviços oferecidos e controle das transferência monetárias) que esse acesso aos dados governamentais podem trazer a sociedade em geral. A seguir é feita uma contextualização sobre economia dos municípios a partir de discussões previdenciárias.

1.3 Economia dos Municípios Brasileiros

Desde a década de 90, inúmeras análises têm sido realizadas acerca de políticas públicas do governo federal brasileiro para manter o *superavit* fiscal, essa mesma proposta de observação do *superavit* fiscal também passou a ser adotada pelos estados e municípios brasileiros. De fato, a partir uma série histórica de medidas de ajuste e controle fiscal no Brasil, tem-se conseguido avançar positivamente no acompanhamento sistemático das contas governamentais (SAKURAI, 2014; FIRJAN, 2018).

A literatura econômica relata uma série de estudos empíricos que justificam o comportamento fiscal do setor público, quando este apresenta saldo positivo ou negativo do *superavit* fiscal. Essa dinâmica ondulatória da economia é definida pela literatura clássica em estados de recessão e expansão. O estado de recessão é presente na economia quando há o registro de dois trimestres consecutivos em queda do PIB, em contrapartida, o estado expansivo consiste na inclinação positiva desse saldo fiscal.

Assim, em estudos voltados aos défices públicos, a literatura neoclássica destaca o modelo apresentado por Barro (1979) o qual pode ser considerado o ponto de partida dessa temática. Assim, em seu modelo o autor busca explicar como défice pode ser percebido pelo governo e como o governo pode utilizar essas informação para minimizar as distorções que este ocasionar na área tributária.

O modelo mencionado acima, parte de uma economia fechada sem capital, na qual o agente representativo, que consome, trabalha e poupa passa a ter a sua função de utilidade potencializada pelo governo que é um planejador benevolente, cujo, objetivo é financiar o gasto público através da tributação de renda e aplicação de impostos distorcivos do trabalhador, pois a partir dessas medidas busca-se afetar diretamente a oferta de trabalho (GIUBERTI, 2005; MEDEIROS et al., 2017).

Uma aplicação prática do modelo de Barro (1979) é relatada por GIUBERTI (2005), a qual destaca que os países da OCDE e os países como economias sub-desenvolvidas, como por exemplo países da América Latina na década de 70, os quais apesentavam défice fiscal por longos períodos pós-guerra. Além disto, esse comportamento também foi observado em outros países que não fazem parte América Latina, mas que possuem economia semelhante. Porém, é importante ressaltar que tais fatos dificilmente coincidem a visão neoclássica.

Desse modo, assim como o modelo Barro (1979) outros modelos foram criados para buscar uma explicação aceitável da ocorrência de défice e/ou endividamento fiscal por meio de aspectos político-institucionais, tais como a organização do estado, para a compreensão do por quê os governos acabam cometendo a produção de défices orçamentários, que muitas das vezes são de natureza persistente.

Esses aspectos político-institucionais em instituições orçamentárias são norteados

a partir da Lei de Responsabilidade Fiscal[1], a qual visa compreender a produção do défice orçamentário e os mecanismos que possibilitam a sua redução. As instituições orçamentárias podem ser entendidas como aquelas que realizam a preparação, execução e aprovação dos orçamentos, seja este de natureza pública e/ou privada.

No Brasil é adotada uma abordagem mais acentuada acerca do federalismo fiscal com descentralização das atividades do governo, para ampliar a atuação dos estados e municípios nas atribuições fiscais, tais como à arrecadação de impostos[2], dando a estes maior controle administrativo (THOMAZINI, 2020). Além disso, alguns recurso são transferidos aos estados e municípios em particular os Fundos de Participação dos estados e municípios e transferências do Regime Geral de Previdência Social (RGPS)[3] previstos pela própria União[4].

Mediante a descentralização de atribuições orçamentárias entre repartições federativas, houve um aumento da responsabilidade fiscal acerca da prestação de serviços públicos essências, sem a contrapartida de receita própria, isto é, uma maior dependência das transferências governamentais GIUBERTI (2005). Ainda sobre a mesma perspectiva o autor relata que ano de 2002, em média 50% da receita corrente dos municípios eram provenientes de transferências da União, os quais contribuem para a descentralização fiscal (BRASIL, 2016).

Atualmente, foi publicada uma análise na Associação Nacional dos Auditores Fiscais da Receita Federal do Brasil (ANFIP)[5] que colaboram com as evidências de dependência fiscal dos municípios brasileiros pontuadas anteriormente. Em sua análise (SÓLON et al., 2019) constatou que 87.9% dos municípios brasileiros em 2017 apresentaram um desequilíbrio fiscal, pois o número de benefícios previdenciários pagos superavam os valores das suas arrecadações. É, importante frisar que nessa análise não foram incluídas algumas transferências previdenciárias, como: BPC/LOAS e os benefícios previstos em legislação especial .

Esse tipo de análise ganha bastante atenção no cenário econômico brasileiro quando são levadas em consideração rendas provenientes de aposentadorias, pensões, auxílios, entre outros benefícios. Estes benefícios juntos compõem as transferências previdenciárias do RGPS aos municípios brasileiros, os quais são garantidos pela Previdência Social[6] e têm a finalidade de assegurar os rendimentos de contribuintes quando estes não dispõem de suas capacidades laborais mínimas (MINISTÉRIO DA FAZENDA, 2018).

Diante isso, é notório que as mudanças propostas pelo novo sistema previdenciário

[1] http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp101.htm
[2] https://www.senado.leg.br/atividade/const/con1988/con1988_26.06.2019/art_153_.asp
[3] http://www.previdencia.gov.br/perguntas-frequentes/regime-geral-rgps/
[4] https://portal.tcu.gov.br/ouvidoria/duvidas-frequentes/fpe-e-fpm.htm
[5] https://www.anfip.org.br/
[6] http://www.previdencia.gov.br/

a partir da PEC 6/2019 (BRASIL, 2019b), podem ocasionar um desequilíbrio econômico acerca dos anos subsequentes após a sua implementação, pois 89.7% dos municípios brasileiros apresentam dependência quase que total dos recursos transferidos pela RGPS. Entre as alterações propostas na PEC 6/2019, o aumento de idade mínima para a concessão de aposentadorias em ambos os sexos é aquela que pode ser vista como a mudança de impacto social, pois entre todos os benefícios transferidos as aposentadorias detém a maior parcela das transferências do RGPS aos municípios brasileiros.

Portanto, as análises de dados econômicas no Brasil tem provocado debates polêmicos quando enfatizamos a análise realizada por (SÓLON et al., 2019) frente aos microdados municipais brasileiros, pois estas são capazes de apresentar o cenário e a dinâmica econômica de uma área analisada.

1.4 Objetivo Geral

O objetivo deste trabalho se foca na obtenção, processamento e análise dos dados públicos (de distintas fontes - IBGE[7], STN[8], DATAPREV[9] e *etc.*) referentes ao sistema previdenciário brasileiro, especificamente do RGPS, usando técnicas de *Data Science* para avaliar a saúde fiscal dos municípios brasileiros, tanto por uma abordagem previdenciária, quanto por uma análise fiscal.

1.4.1 Objetivos Específicos

A partir da realização do objetivo geral, temos os objetivos específicos associados a ele, os quais apresentam uma visão mais detalhada das etapas metodológicas deste estudo, sendo apresentados abaixo:

- Realizar um amplo levantamento bibliográfico dos trabalhos voltados à temática desta dissertação;
- Identificar as problemáticas acerca da coleta de dados, narradas pelos pesquisadores dentre o acervo bibliográfico utilizado;
- Obter e tratar as diferentes bases de dados para identificar quais as variáveis mais importantes;
- Modelar um ambiente de trabalho flexível e consistente de fácil manipulação durante o processo de tratamento e análise de dados;

[7] https://www.ibge.gov.br/
[8] http://sisweb.tesouro.gov.br/apex/f?p=2600:1:::NO:RP%2C1::
[9] http://www.previdencia.gov.br/

- Aplicar diversas técnicas de análise de dados, como: amostragem aleatória, modelo STVAR, aprendizado de máquina e processo (KDD) sobre os dados;
- Interpretar e validar os resultados obtidos junto aos especialistas da área, buscando obter informações capaz de auxiliar à tomada de decisão.

1.5 Organização do Texto

Esta dissertação é organizada e composta por cinco seções, incluindo esta seção introdutória. Sendo as demais seções estruturados da seguinte forma:

Capítulo 2 : apresenta conceitos referentes à ciência de dados e processo de KDD com as suas respectivas caracterizações, posteriormente são descritas técnicas de otimização matemática e de previsão de série por meio do modelo STVAR.

Capítulo 3 : são abordados as trabalhos relacionados a esta dissertação, onde são organização em duas categorias. A primeira relatar estudos com aplicação de dados sócio-demográficos e os trabalhos sobre análises temporais com estímulos exógenos.

Capítulo 4 : é apresentado o ambiente computacional empregado nesta dissertação. Além disto, é disposta a metodologia adotada para o processamento dos dados, e também o percurso metodológico.

Capítulo 5 : são apresentados dois estudos de casos. O primeiro estudo faz uma análise estatística do endividamento através de dados previdenciários e no segundo é realizada estimava via o função impulso-resposta no modelo STVAR.

Capítulo 6 : estão dispostas as conclusões de ambos os estudos de caso, bem como as dificuldades encontradas no decorrer deste trabalho, as propostas de trabalhos futuros, publicações e por fim, as referências bibliográficas utilizadas.

2 Referencial Teórico

Neste capítulo, são introduzidos conceitos teóricos e técnicos amplamente utilizados na área computacional para manipular e extrair conhecimento útil através de base de dados, algumas ciências destacam-se neste tipo de atividade, tais como a ciência de dados e a descoberta do conhecimento em base de dados (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; FRANK; WITTEN, 2000). Consequentemente, são debatidos assuntos referentes à importância do acesso aos dados público, bem como seu papel frente à sociedade em geral.

Além disso, são discutidas algumas técnicas presentes no contexto de estatística e econometria utilizadas durante o processo de análise de dados que também fazem parte do escopo da ciência de dados, por exemplo: medidas de dispersão, amostragem e o modelo auto regressivo de transição gradual suave (STVAR) e os modelos otimização matemáticos mais usuais para esta atividade.

2.1 Ciência de dados

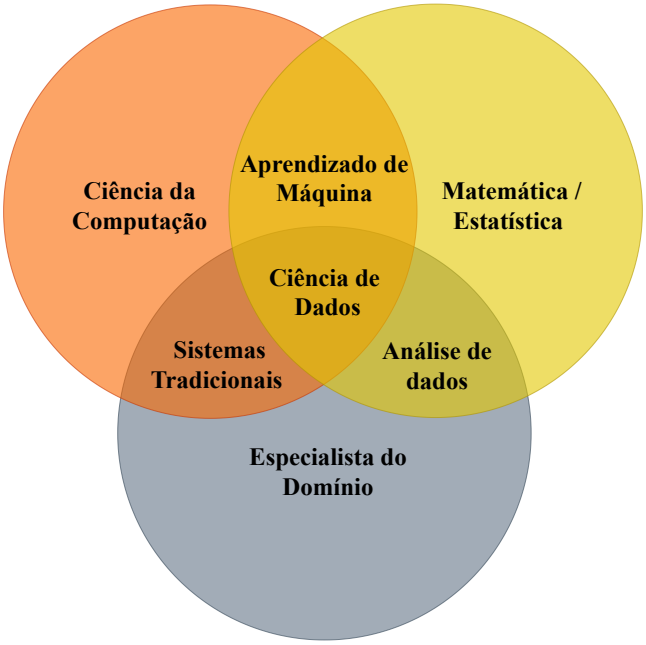
A ciência de dados é uma área fundamental para o processo de aquisição do conhecimento, a qual está direcionada na extração de informações e a formulação de novas fontes de conhecimento através de base de dados (OLIVEIRA; GUERRA; MCDONNELL, 2018). Por esse motivo, a ciência de dados vem se popularizado entre as diferentes áreas do conhecimento, por exemplo: Ciência da Computação, Biologia, Economia, Engenharias e *etc.* Apesar da ampla utilização da ciência de dados nos dias atuais, ela não é uma área nova. Porém, ela obteve maior visibilidade nos últimos anos devido ao aumento expressivo dos dados gerados e por demandas voltadas para tomada de decisão.

Para que uma análise de dados possa ser considera como ciência de dados é necessário que esta inclua em seu escopo técnicas que utilizem: álgebra linear, modelagem estatística, aprendizado de máquina, análise de gráficos, visualização e *etc.* (BOSCHETTI; MASSARON, 2015). Ou seja, a ciência de dados não se restringe unicamente ao processo de aquisição e de tratamento de dados, mais sim em todas as etapas necessárias para obtenção de conhecimento. Isto é, esta deve representar algum tipo de conhecimento relevante logo após a sua aplicação.

A ciência de dados deve conter algumas competências necessárias em seu escopo durante o processo de análise de dados, sendo elas: conhecimento computacional, conhecimento exato (matemático e estatísticos) e o conhecimento do especialista da área estudada, as quais caracterizam à sua multidisciplinaridade e à sua adequação na geração de soluções a problemas diversos (CADY, 2017; VANDERPLAS, 2016). Na Figura 2 são demonstradas

as relações e as correlações entre as diferentes áreas conhecimento.

Figura 2 – Competência essenciais para a ciência de dados.



Fonte: Elaborado pelo autor, 2019.

A utilização de ciência de dados para análises de dados pode facilmente adotar a metodologia do processo de processo de descoberta de conhecimento em bases de dados (do inglês, *Knowledge Discovery in Databases* (KDD)) (ELMASRI; NAVATHE, 2005; FRANK; WITTEN, 2000), o qual consiste basicamente na inspeção, limpeza e transformação dos dados, o qual visa a busca de padrões contidos em bases dados através de métodos analíticos computacionais que são utilizados em quase toda à sua execução (OLAVSRUD, 2018). No entanto, outras áreas de conhecimento, como: matemática convencional, estatística e economia adotam diferentes metodologias para extração de informação.

2.2 Descoberta de Conhecimento em Bases de Dados

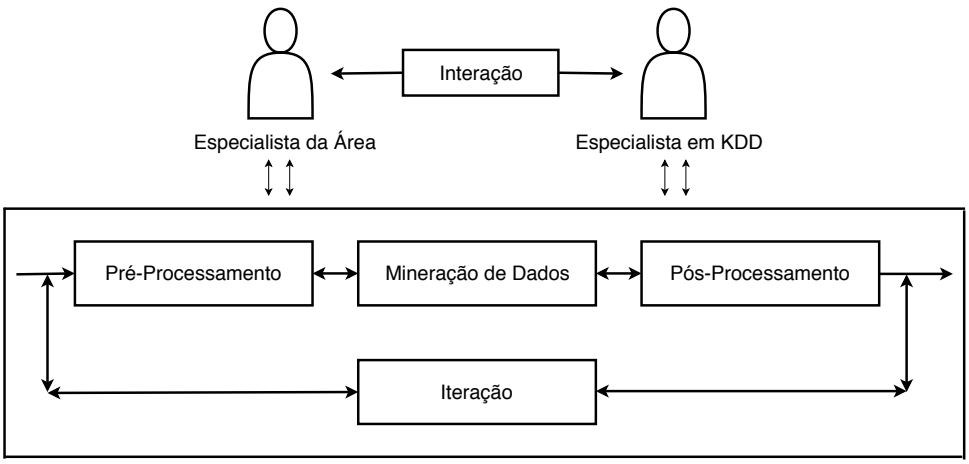
A descoberta de conhecimento em bases de dados ou o processo de KDD pode ser entendido como uma metodologia com diferentes etapas, porém as etapas são bem definidas, interligadas e interativas, cujo, objetivo final é a detecção de padrões em diferentes bases de dados via a métodos analíticos computacionais (GOLDSCHMIDT; PASSOS, 2005; FRANK; WITTEN, 2000). Além disso, o processo de KDD utiliza a mineração dados como a etapa principal de extração de conhecimento de dados, a qual normalmente faz uso de algoritmos sofisticados para tal atividade.

O processo de KDD engloba em seu escopo as seguintes fases operacionais de processamento de dados: aquisição, seleção, limpeza, transformação, mineração de dados ou inferências analíticas e interpretação dos resultados. Além disso, processo de extração de informação deve apresentar basicamente três componentes durante a sua execução (GOLDSCHMIDT; PASSOS, 2005):

- **Conjunto de dados:** Consiste na própria base de dados de forma que essa esteja armazenada em local acessível e possa ser utilizada posteriormente no processo de extração de informação.
- **Especialista da Área:** É aquele que detém conhecimento teórico e prático sobre o problema em questão.
- **Especialista em KDD:** É pessoa com conhecimento especifico e experiência a aplicação de métodos de extração de informação em base de dados

A partir das competências básicas do processo de KDD é proposto um ciclo de etapas operacionais generalizada que estabelece o fluxo de informação entre as fases de análise, conforme ilustrado na Figura 3.

Figura 3 – Etapas operacionais do processo de KDD



Fonte: Extraído de Goldschmidt, Passos e Bezerra (2015)

A Figura 3, demonstra todo o processo comunicativo entre os especialistas da área e os especialistas em KDD no decorrer do processo de análise e avaliação de dados. Segundo Goldschmidt e Passos (2005), Goldschmidt, Passos e Bezerra (2015) o processo de KDD é uma metodologia não trivial, interativa utilizada na identificação de padrões compreensíveis, válidos, potencialmente novos e úteis. Então, a complexidade do processo de KDD está essencialmente contida na percepção, observação e na interpretação adequada para cada iteração que o compõe. A seguir serão brevemente descritas as etapas de pré-processamento, mineração de dados e pós-processamento de dados.

2.2.1 Pré-Processamento de Dados

O pré-processamento compreende as atividades de seleção e tratamento de dados, as quais visam a preparação dos dados para a etapa de mineração de dados. Ou melhor, é competência do pré-processamento transformar os dados em parâmetros de entrada com formato apropriado para análises subsequentes, por exemplo: a seleção de dados está diretamente ligada a remoção de ruídos e dados duplicados (TAN; STEINBACH; KUMAR, 2009).

A seleção de dados é considerada como a etapa responsável pela redução de dados, pois nela é feita a seleção dos atributos essenciais que são utilizados para solucionar um dado problema de extração de informação em base de dados. É, importantes frisar que essa filtragem de dados é realizada junto com os especialistas da área (GOLDSCHMIDT; PASSOS; BEZERRA, 2015), visto que os atributos selecionados nessa etapa serão utilizados durante todo o processo de análise.

Além disso, a seleção de dados também objetiva verificar se há problemas na forma em que os dados estão alocados. Enquanto, a transformação de dados preocupa-se em modificar o formato em que os dados encontram-se, em formatos compreensíveis pelos algoritmos de análise de dados. Ou melhor, cabe a etapa de transformação de dados adequar os dados de maneira que os algoritmos possam ser facilmente utilizados como entrada na fase de mineração de dados (TAN; STEINBACH; KUMAR, 2009).

2.2.2 Mineração de Dados

A mineração de dados é a etapa fundamental do processo de descoberta de conhecimento em bancos de dados, a qual consiste no processo analítico capaz de explorar gigantescas bases de dados de diferentes domínios de aplicação na busca por padrões consistentes que sirvam como informações úteis e valiosas para à tomada de decisão (TAN; STEINBACH; KUMAR, 2009). Desse modo, a mineração de dados é considerada como a etapa de maior relevância no processo de descoberta de conhecimento, já que ela é a responsável pelo processamento dos dados em informações indispensáveis para a tomada de decisão (FRANK; WITTEN, 2000; HAND; MANNILA; SMYTH, 2001). Com isso a mineração de dados pode ser entendida como uma atividade multi-aplicável, porque, sua aplicação varia de acordo com a área de atuação.

Na fase da aplicação de mineração de dados, é necessário que o especialista em KDD identifique quais técnicas e algoritmos terão melhor desempenho na atividade de descoberta de conhecimento. Ou seja, as técnicas e os algoritmos escolhidos devem ser precedentemente definidas levando em consideração as características descritas pelos especialistas da área. A seguir serão apresentadas breves contextualizações das tarefas mais usuais na etapa de mineração de dados:

- **Associação:** a tarefa de associação tem como objetivo buscar correlação entre atributos em uma base de dados, em outras palavras, busca encontrar itens que impliquem a presença de outros itens na mesma transação.
- **Classificação:** pode ser definida como um processo supervisionado que é um aprendizado indutivo, ou seja, um aprendizado capaz de relacionar eventuais atributos existentes entre uma base de dados.
- **Agrupamento:** Para Elmasri e Navathe (2005) "o objetivo do agrupamento é colocar os registros em grupos, de tal forma que os registros de um grupo sejam similares aos demais do mesmo grupo e diferentes daqueles dos demais grupos".
- **Regressões:** esta tarefa é constituída a partir de uma modelagem preditiva onde sua variável alvo é compreendida como continua, ou seja, para que um problema X qualquer possa ser resolvido é necessário usar outros indicadores que se relacionam a este problema, para que ele seja solucionado.

2.2.3 Pós-Processamento de Dados

O pós-processamento de dados é a etapa responsável pela decodificação das saídas da etapa de mineração de dados, onde as respostas da mineração são convertidas em formatos compreensíveis pelos especialistas em KDD, com o objetivo de facilitar a análise de interpretação dos resultados obtidos. Aliado à isso, o processo de conversão de dados pode ser feito por meio de algumas técnicas computacionais bastantes utilizadas no âmbito acadêmico em geral, como a visualização de dados e a representação gráfica (STEELE, 2010).

Contudo, para que possa havê uma análise de dados computacional eficiente e bem estruturada é necessário que o processo de extração de conhecimento em bases de dados, utilize métodos analíticos bem solidificados na areá de domínio na qual pretende-se aplicá-lo. Dessa forma, buscou-se na literatura empírica técnicas amplamente disseminadas na área econômica para solucionar a problemática de visualização e regressão não-linear de dados via técnicas estatísticas (tendencia, média e modelo econométrico) e otimização matemática.

2.3 Análise Estatística de Dados

A estatística é uma ciência exata que estuda as formas de coleta, organização, análise e interpretação de dados através de amostras de uma dada população. Além disso, os métodos estatísticos são indispensáveis para produção de conhecimento e também são utilizados para obtenção de novas percepções comportamentais do mundo (DEVORE, 2015; MEDEIROS, 2007).

Assim, a ampla utilização dos métodos estatísticos descritivos ganha bastante atenção no cenário computacional uma vez que estes estão diretamente interligados com a análise de dados de forma geral (SANTOS, 2018), pois essa abordagem normalmente faz uso de representações de natureza gráficas (histogramas, gráficos *boxplot* e gráficos de dispersão) para exploração dos dados. Além disso, a estatística também engloba em seu escopo métodos para organizar e reduzir o universo de dados pesquisado.

A metodologia de sintetização de dados é recorrente em medidas de amostragens a partir de uma população de dados, as quais baseiam-se em uma determinada amostra de dados a partir de universo de observação acerca da relação entre as unidades e a precisão mínima pretendida (REIS et al., 2007). Portanto, as amostragens estatísticas podem ser utilizadas por meio de métodos não probabilísticos empíricos (conveniência, julgamento e quotas) e métodos probabilísticos aleatórios (simples, sistemática, estratificada e conglomerada) (SANTOS, 2018).

Dentre as inferências estatísticas, algumas como: intervalo de confiança, Análise de Variância - (ANOVA) e Erro Quadrático Médio - (EQM), têm apresentando inúmeras aplicações no âmbito computacional para considerar análises aferidas sobre diferentes temáticas e também em diferentes áreas do conhecimento, por exemplo: Ciências Humanas, Ciências Agrarias e Ciência da Saúde (BATTISTI; BATTIST, 2008).

2.4 Modelos de Otimização Matemáticos

A econometria é uma subárea da economia que se propõe a observar, estimar, testar e avaliar hipóteses que tendem a implementação de novas políticas que estão voltadas ao ramo de negócios (WOOLDRIDGE, 2010; WOOLDRIDGE, 2017). A aplicabilidade mais comum da econometria está direcionada a previsão de variáveis macroeconômicas citando caso análogo taxa de inflação, taxa de juros e o PIB, buscando inter-relação apriorística econômicas através de funções matemáticas (WOOLDRIDGE, 2010; GUJARATI, 2011; WOOLDRIDGE, 2017).

As funções matemáticas mais usuais para estimação de coeficientes adotada na econometria são os métodos de otimização clássicos de mínimos quadrados e suas variações (AGUIRRE, 2007), que são amplamente utilizados para estimar parâmetros através de regressões lineares e não lineares, as quais objetivam identificar parâmetros de modelos matemáticos capazes de simular comportamento de um cenário específico (DANTAS, 2013).

2.4.1 Método dos Mínimos Quadrados

O método dos mínimos quadrados foi proposto por Gauss em 1795, o qual é um dos métodos mais conhecidos e usuais em diferentes áreas do conhecimento (AGUIRRE,

2007). A partir deste método é possível estimar parâmetros $\hat{\theta}$ que representam a dinâmica do modelo analisado. Onde $\hat{\theta}$ é vetor de parâmetros apresentado por:

$$\hat{\theta} = \left[\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n\right]$$

Para realizar a estimativa $\hat{\theta}$ é necessário utilizar regressores $y(t)$ que são atrasos ou observações passadas da própria dinâmica do modelo, isto é, os regressores podem ser entendidos como um conjunto de informações organizadas de forma cronológica em um dado período de tempo, conforme a ilustração vetorial de $y(t)$ apresentada abaixo.

$$y(t) = \left[y(0), y(1), \ldots, y(n-1)\right]$$

Para calcular a previsão do sistema $\hat{y}(t)$ é utilizada a Equação 2.1, ψ representa as entradas do modelo de forma vetorial semelhante ao vetor $y(t)$ ilustrado acima. De acordo com Gauss, os parâmetros estimados devem se apresentar de tal forma que a soma dos quadrados da diferença entre os valores previstos seja mínima, conforme a Equação 2.2.

$$\hat{y}(t) = \psi\hat{\theta} + \xi \tag{2.1}$$

$$\varepsilon = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{2.2}$$

Onde, N é o número de obervações do modelo, ε é o erro de previsão o qual representa a diferença entre $\hat{y}$ amostras reais e y respostas estimadas, ξ é o erro do modelo. Assim, através do método de mínimos quadrados pretende-se determinar o valor $\hat{\theta}$ que minimize o valor ε. Para (DANTAS, 2013) quando os métodos polinomiais são lineares em seus parâmetros usualmente utiliza-se mínimos quadrados para estimar os seus parâmetros, a Equação 2.3 demonstra a sua formulação matemática.

$$y(k) = \sum_{i=1}^{n=\theta} \psi^T(k)\hat{\theta}_{n_\theta} + \xi(k) \tag{2.3}$$

Onde, n_θ é o números de termos do modelo, $\psi(k)$ consistem nos regressores do sistema, $\xi(k)$ é ruído do modelo. Ao observar a Equação 2.4 de forma matricial, tem-se:

$$Y = \Psi\Theta + E \tag{2.4}$$

Podemos admitir que sejam feitas N medidas necessárias para definir os parâmetros, Y e E são respectivamente vetores de saída e de erro do modelo, assim temos as seguintes resoluções vetoriais:

$$Y = \Big[y(1), y(2), \ldots, y(N-1)\Big]^T$$

$$E = \Big[\xi(1), \xi(2), \ldots, \xi(N-1)\Big]^T$$

$$\Psi = \begin{bmatrix} \psi(1) & \psi(1) & \cdots & \psi_{n_\theta}(1) \\ \psi(2) & \psi(2) & \cdots & \psi_{n_\theta}(2) \\ \psi(3) & \psi(3) & \cdots & \psi_{n_\theta}(3) \\ \vdots & \vdots & \ddots & \vdots \\ \psi(N-1) & \psi(N-1) & ... & \psi_{n_\theta}(N-1) \end{bmatrix}$$

Onde, T indica a matriz transposta ou o vetor inverso, Ψ a matriz de regressores do modelo e θ corresponde ao vetor de parâmetros estimados por mínimos quadrados.

O método de mínimos quadrados Equação 2.5 é uma transformação linear X em relação a Y denominada de estimador linear que minimiza o custo de função, conforme ilustrado na Equação 2.1 (COELHO; COELHO, 2004; COELHO, 2016).

$$\hat{\theta} = \Big[X^T X\Big]^{-1} XY \tag{2.5}$$

Para (DANTAS, 2013) a matriz $X^T X$ é conhecida como matriz de informação ou matriz de covariância dos regressores. A matriz é simétrica e definida positiva quando X tem característica pleno de colunas e semi-definida positiva em caso contrário.

2.4.2 Método dos Mínimos Generalizados - MIMO

O método dos mínimos quadrados generalizados foi introduzido por Alexander Aitken em 1934 (AITKEN, 1936) o qual apresenta inúmeras aplicações no âmbito acadêmico, dentre elas temos o método dos mínimos quadrados MIMO (*Multiple-Input-Multiple-Outpu*) que é amplamente utilizado na área telecomunicações e econometria (SIBILLE; OESTGES; ZANELLA, 2010; WOOLDRIDGE, 2017), conforme a Equação 2.6.

$$Y_{i,n} = \mu_{i,n} + C_{i,m} X_{m,i} + \xi_{i,n} \tag{2.6}$$

Onde, n é número de colunas de Y e X que são multi-saídas observadas; μ e C são o coeficientes a serem estimados pelo modelo e ξ é a matriz de resíduos. A Equação 2.6 pode se observada de maneira matricial, conforme ilustra a Equação 2.7:

$$\begin{bmatrix} Y_{i,n} \\ Y_{n,i} \end{bmatrix} [Y] = \begin{bmatrix} \mu_{i,n} \\ \mu_{n,i} \end{bmatrix} [\mu] + \begin{bmatrix} C_{i,m} \\ C_{m,i} \end{bmatrix} [C] \begin{bmatrix} X_{m,i} \\ X_{i,m} \end{bmatrix} [X] + \begin{bmatrix} \xi_{i,n} \\ \xi_{n,i} \end{bmatrix} [\xi] \tag{2.7}$$

$$\hat{\theta} = YX^T[XX^T]^{-1} \tag{2.8}$$

Portanto, para estimar os coeficientes a partir dos mínimos quadrados MIMO é proposta a Equação 2.8 (AITKEN, 1936), que é tida como uma solução sub-ótima para o problema de estimação matricial dos coeficiente μ e C.

2.5 Modelo de Estimação Não linear - STVAR

Atualmente a literatura empírica responsável pela avaliação dos feitos de choques da política fiscal tem declinado nas investigações acerca da dependência entre os multiplicadores fiscais e o ciclo econômico (ALVES, 2017), pois estes modelos englobam em escopo características, como a recessão e expansão econômica que são utilizados para simular comportamento do ciclo econômico estudado (OLIVEIRA, 2018).

Vários trabalhos são propostos para analisar ciclos econômicos de diferentes locais do mundo, entre essas análises temos o popular modelo VAR - (*Vector Autoregression*) o qual é capaz de transitar de maneira gradual entre regimes (expansão e recessão) (DUTRA, 2018; SANTOS, 2009). Tais características são aplicadas no modelo *Smooth Transition Vector Autoregression* - STVAR (DUTRA, 2018; OLIVEIRA, 2018; SANTOS, 2009). Na área de econometria o modelo STVAR obteve ampla utilização a partir da codificação de (AUERBACH; GORODNICHENKO, 2012a; AUERBACH; GORODNICHENKO, 2012b), cuja, especificação econométrica segue o sistema de equações:

$$x_t = \Big[1 - F(z_{t-1})\Big]\ \Pi_R(L)\ x_{t-1} + F(z_{t-1})\Pi_E(L)\ x_{t-1} + \varepsilon_t \tag{2.9}$$

$$\varepsilon_t \sim N(0, \Omega_t) \tag{2.10}$$

$$\Omega_t = \Big[1 - F(z_{t-1})\Big]\ \Omega_R + F(z_{t-1})\Omega_E \tag{2.11}$$

$$F(z_t) = \frac{exp(-\gamma z_t)}{1 + exp(-\gamma z_t)},\ \gamma > 0 \tag{2.12}$$

$$var(z_t) = 1,\ E(z_t) = 0 \tag{2.13}$$

Onde, $x_t = [G_t, R_t, P_t]$ é um vetor composto pelas variáveis: G_t são os gastos públicos, R_t representa as receitas e P_t é o PIB - (Produto Interno Bruto) (ORAIR; SIQUEIRA; GOBETTI, 2016); A variável z_t ilustrada na Equação 2.12 é um indicador capaz de captura a transição entre as etapas do ciclo econômico tendo valores positivos para

períodos de expansão e negativos nas recessões, normalizado com média zero e variância igual a 1 Equação 2.13.

As matrizes $\Pi_i(L)$ a $\Omega_i(L)$ das Equações 2.9 e 2.11 representam respectivamente os coeficientes estimados pelo modelo e a matriz covariância, i representa os regimes de expansão ($i = E$) e recessão ($i = R$) do ciclo econômico; a função $F(z)$ Equação 2.12 consiste na interpretação probabilista de está ou não em recessão (AUERBACH; GORODNICHENKO, 2012b) e a variável γ deve ser a justa de tal forma que períodos recessivos estimados coincidam com períodos reais do modelo (ORAIR; SIQUEIRA; GOBETTI, 2016).

2.6 Considerações Finais

Nesse capítulo, foi apresentada a fundamentação teóricas acerca do desenvolvimento desta dissertação, a qual iniciou a partir da contextualização a respeito da ciência de dados (OLIVEIRA; GUERRA; MCDONNELL, 2018), bem como a sua interdisciplinaridade voltado para a busca de informação em base de dados. A partir da perspectiva de detecção de conhecimento úteis em base de dados, foi utilizada a técnica de descoberta de conhecimento em base de dados (ELMASRI; NAVATHE, 2005) e suas etapas operacionais.

Neste contexto, é descrita a conjuntora de base de dados abertas é a sua importância para o desenvolvimento social e o quão disseminada a prática de abertura de dados está presente nas maiores economias do mundo (MUELLER, 2019). Também, foram apresentadas métricas de avaliação estatísticas para mensurá as analises realizadas (DEVORE, 2015). Ademais, são contextualizados os modelos de estimação econométricas mais utilizados, bem como as técnicas de otimização matemática mais usuais para o problema, tais como: mínimos quadrados ordinários e mínimos quadrados MIMO (AGUIRRE, 2007; COELHO; COELHO, 2004; COELHO, 2016).

3 Trabalhos Relacionados

Devido à grande abrangência da área de análises de dados, nesta seção serão abordados trabalhos cujo os seus objetivos foram analisar dados abertos e públicos voltadas especificamente para a área econômica. Dessa forma, neste capítulo apresenta-se um panorama dos trabalhos relacionados ao tema, ressaltando trabalhos que utilizam ciência de dados em suas análises sobre recenseamento do tema referido.

3.1 Ciência de Dados Aplicada em Dados Demográficos

Atualmente, o processo de aquisição de conhecimento em bases de dados é bastante utilizado em diferentes áreas do conhecimento, como: computação, estatística, engenharias e *etc.* (TAN; STEINBACH; KUMAR, 2009; MUELLER, 2019), com o propósito de se obter informações úteis que sirvam com base para a tomada de decisão (GOLDSCHMIDT; PASSOS; BEZERRA, 2015; LIMA et al., 2017). Porém, a literatura sobre à temática inclui em seu escopo diferentes metodologias (Análise de Dados, Ciência de Dados, processo de KDD e *etc.*) de como lidar com a problemática (QIANG et al., 2019). No entanto, à ciência de dados recentemente tem se mostrado cada vez mais predominante entre as pesquisas acadêmicas.

O trabalho prático inicial desta pesquisa é narrado por (BAERLOCHER; PARENTE; RIOS-NETO, 2019) o qual utiliza a técnica de *Sys-GMM*[10] sobre dados referentes às diferenças entre as microrregiões brasileiras durante o período de 1997 a 2000 visando as delimitações demográficas contábeis entre elas. Algumas variáveis econômicas utilizadas no estudo: população em idade ativa, número de cidadãos entre 15 e 64 anos, PIB per capita, anos de escolaridade e a média de anos de escolaridade da população acima de 25 anos e *etc.*

A partir da análise aferida foi constatado que o dividendo das microrregiões apresenta indícios de que mudança de faixa etária implica diretamente no faturamento dessas regiões e que o grau de escolaridade é fator com maior impacto no crescimento econômico de cada área estudada e que o efeito contábil é responsável por menos de 10% da diferença de renda entre as regiões mais pobres e mais ricas do Brasil.

Em (SÓLON et al., 2019) é empregado um estudo por meio de métodos estatísticos os quais são utilizados para estimar o quão dependente os municípios brasileiros estão dos repasses previdenciários referentes aos dados de 2017. Assim, os resultados mostram que 87.9% dos municípios brasileiros apresentam défice positivo em relação às arrecadações

[10] https://rdrr.io/rforge/gmm/man/sysGmm.html

municipais e os benefícios pagos. Isso reforça a importância da manutenção e do crescimento linear do salário mínimo, pois este é um valioso instrumento de redistribuição de renda pela Previdência Social.

Além disso, os resultados obtidos à partir do estudo sinalizam que os municípios que apresentam populações entre 10 mil a 20 mil habitantes, ostentam o maior percentual de pagamento de benefícios que superam as arrecadações. Em escala monetária nacional isto representa aproximadamente R$ 25 bilhões de reais, e quando comparado ao PIB equivale à 7.5% para o conjunto de municípios enquadrado na faixa de habitantes.

Mediante ao crescente número de análises elaboradas sobre dados demográficos, os autores (KREYENFELD; WILLEKENS, 2015) fazem uma análise dos tipos de dados demográficos existentes , que podem ser facilmente utilizados para medir as correlações entre a construção de indicadores capazes de mensurar a migração social urbana e as taxas responsáveis pela construção de famílias, como: casamentos e divórcios e morte.

No entanto, em sua análise o autor acentua que os dados demográficos abertos em sua maioria não apresentam informações suficientes, o que pode delimitar o entendimento de alguns comportamentos sociais. Além disso, são explicitadas algumas vantagens da utilização destes dados, bem como as armadilhas que estes podem apresentam durante o processo de análise.

Assim, o diferencial deste trabalho quando comparado aos demais descritos nesta seção, está relacionado ao uso de técnicas da ciência de dados para o entendimento e análise sobre a dependência contábil dos municípios brasileiros através dos repasses previdenciários solvidos entre o período de 2010 a 2017. Isto é, foi realizada uma análise a partir das transferências da RGPS anuais aos municípios brasileiros, para compreender o comportamento do seu ciclo administrativo. Tal estudo refere-se ao primeiro estudo de caso apresentado nesta dissertação.

3.2 Análises de Dados Econométricas

Em econometria inúmeros estudos são focados na análises de dados, cujo, o seu objetivo é modelar a dinâmica de microeconomias e macroeconomias de diferentes localidades do mundo. Então, relacionado a isso várias pesquisas têm direcionado sua atenção para a estimação de séries temporais, que apresentam características não-lineares, para compreender de forma consistente da relação de alguns variáveis presentes modelo econômico, como por exemplo: gasto, receita e produto, as quis são capazes de demonstrar os impactos de uma caso haja variação na modelo econômico estuda via a aplicação de multiplicadores fiscal.

Assim, no Brasil alguns autores (ORAIR; SIQUEIRA; GOBETTI, 2016) em suas

pesquisas buscaram compreender de forma mais consistente essa previsão, a partir do modelo proposto por Auerbach e Gorodnichenko (2012a), o qual considera três variáveis de entrada (despesa, receitas e PIB) que permitem investigar os diferentes impactos sobre as atividades econômicas.

Dessa forma, em sua análise os autores (CERQUEIRA; RIBEIRO; MARTINEZ, 2014; ORAIR; SIQUEIRA; GOBETTI, 2016; ORAIR; SIQUEIRA, 2018) inicialmente destacam que no Brasil, os dados disponíveis apresentam-se em formatos inadequados e em séries curtas para aplicação no modelo de estimação. Por isso, os autores construíram suas próprias séries com periodicidade mensal entre o período de 2002-1 a 2016-4, com base nos dados anuais brasileiros disponíveis IBGE e STN para validar o modelo.

Então, os resultados obtidos demonstram que há uma forte depressão econômica para o período analisado, a partir do influência de alguns gasto público sobretudo os de investimentos. Por outro lado, quando são avaliados benefícios sociais e gastos pessoais a duração do impacto é sensivelmente maior. Tal análise baseia-se a partir de resposta a impulso inferido no modelo, para uma melhor visualização do efeito dos multiplicadores fiscais na economia analisada.

Em (CERQUEIRA; RIBEIRO; MARTINEZ, 2014; ORAIR; SIQUEIRA, 2018) é proposta uma análise empírica a respeito dos choques monetários assimétricos na economia brasileira. Como instrumento de análise utiliza-se modelo STVAR para variáveis relacionadas ao produto, às taxas de inflação e de câmbio e ao indicador de política monetária. Para avaliar o modelo, é utilizada a função resposta a impulso, a qual é capaz de mensurá o impacto dos choques monetários expansionistas e contracionistas, que apresentam efeitos assimétricos sobre o crescimento do produto e a inflação.

No estudo de (GRUDTNER; ARAGON, 2017) é utilizado o modelo de estimação vetorial autorregressivo de transição não-linear com o objetivo de avaliar os impactos da aplicação dos multiplicadores de gasto públicos brasileiros entre o período trimestral de 1999-1 ao quarto período trimestral 2015-4, mediante as variáveis de consumo, investimento e salários pagos aos servidores públicos.

Apesar disso, os resultados obtidos evidenciam que os multiplicadores dos gastos do governo apresentam comportamento equivalentes a partir dos períodos de recessão e expansão. Além disto, os choques aplicados ao modelo apresentam período de expansão alto e recessão profunda acerca dos gastos do governo referente aos multiplicadores fiscais de investimento, via a razão dívida/PIB e ao grau de abertura econômica e taxa de câmbio.

Portanto, o segundo estudo de caso presente nesta dissertação realiza uma análise sobre os dados municipais de são paulo, a partir do uso de técnicas de ciência de dados para tratar e obter as bases de dados alocadas em distintas fontes (STN, Fundação de Sistema

Estadual de Análise de Dados (SEADE)[14], IPEADATA[11] e *etc.*) E, posteriormente aplicar os dados no modelo vetorial auto-regressivo de transição suave - STVAR, para para medir comportamento econômico do ciclo e compreender a sua dinâmica via a multiplicador fiscal.

3.3 Considerações Finais

Em relação aos novos estudos voltados para ciência de dados, a área de análise de dados demográficos vêm despertando bastante interesse entre diversos pesquisadores de diferentes domínios do conhecimento (MAGALHÃES, 2015; BRITO; KERSTENETZKY, 2019). No que diz respeito ao setor da economia a análise de dados é capaz de gerar grandes impactos sociais que implicam diretamente na qualidade de vida (SILVA et al., 2011; BRITO; KERSTENETZKY, 2019).

[14] https://www.seade.gov.br/produtos/pib-mensal/
[11] http://www.ipeadata.gov.br/Default.aspx

4 Materiais e Métodos

Neste capítulo são apresentadas as ferramentas computacionais utilizadas no desenvolvimento desta dissertação, as quais têm sido amplamente aplicadas para customizar o tempo de processamento de atividades que envolvem algum tipo de análise de dados. Estas também apresentam artifícios que facilitam tanto a visualização quanto a manipulação de dados de forma mais intuitiva. Além disto, também é descrita a metodologia empregada, a começar pelo pré-processamento até o pós-processamento dos dados.

4.1 Ambiente Computacional

As inovações tecnológicas têm gerado inúmeras melhorias quando voltamos nossa atenção para a análise de dados, principalmente na redução do tempo de pré-processamento, o qual é a atividade de maior esforço. Assim, algumas ferramentas têm se destacado neste cenário, como exemplo: Miniconda, Python, Pandas, Jupyter Lab, Matplotlib, Docker e *etc.* A seguir será realizada uma breve descrição do ambiente computacional adotado nesta dissertação.

- O Miniconda é uma distribuição gratuita minimizada do anaconda que inclui apenas o Python e o conda, o que torna essa distribuição mais leve e dá ao usuário maior controle das suas aplicações. Nela o usuário apenas faz uso dos pacotes necessários para a sua análise. Além disso, ela possui um fácil gerenciamento de pacotes através do gerenciador nativo conda (CONTINUUM ANALYTICS, 2020).

- O Python é uma linguagem de programação interpretada, imperativa, orientada a objetos, fracamente tipada e de acesso livre. Foi lançada pelo matemático Guido van Rossum em 1991. Além disso, o Python é largamente utilizado em atividades que envolvem processamento de textos, dados científicos e para a criação de CGIs utilizados em páginas dinâmicas *web* (REITZ, 2018). Partindo dessa ótica, o Python foi a linguagem base para desenvolvimento deste trabalho devido a sua simplicidade e a sua facilidade de codificação.

- Pandas é uma biblioteca do Python que fornece estruturas de dados rápidas, flexíveis e expressivas, projetadas para facilitar intuitivamente o trabalho com dados "relacionais"ou "rotulados". O seu objetivo fundamental é ser o bloco de construção de alto nível voltado a análise prática de dados. Além de que, ela busca torna-se a ferramenta mais poderosa de manipulação de dados e de código aberto, flexível e disponível entre as demais (MCKINNEY, 2019).

- Jupyter Lab consiste em uma plataforma de desenvolvimento que permite ao usuário uma maior interação e flexibilidade, além de proporcionar uma melhor organização de arquivos a partir de um ambiente unificado de visualização e manipulação de dados. Outra característica do Jupyter Lab, corresponde a sua compreensão mediante aos diferentes tipos de arquivo (*csv, json, markdown, pdf, vega, vega-lite e etc.*) (PROJECT JUPYTER, 2019).

- O Matplotlib é uma biblioteca de plotagem 2D do Python capaz de produzir figuras com alta qualidade e em vários formatos, através de um ambiente interativo que independem da plataforma de desenvolvimento. Ela pode ser usada em *scripts* Python, Shell Python, IPython, Jupyter Notebook e suas variações e em servidores de aplicativos da *web* (HUNTER et al., 2019).

- O Docker pode ser entendido como uma plataforma *open source* que facilita a criação e administração de ambientes isolados. Ele permite encapsular um ambiente de trabalho dentro de um *container*, sendo este portátil para qualquer outro *host* que contenha o *docker* instalado (GUEDES, 2018). Ou melhor, a ideia por trás do *docker* é utilizar apenas uma máquina ao invés de várias, para rodar diferentes aplicações sem ocorrer nenhum tipo de conflitos entre elas.

4.2 Metodologia Aplicada

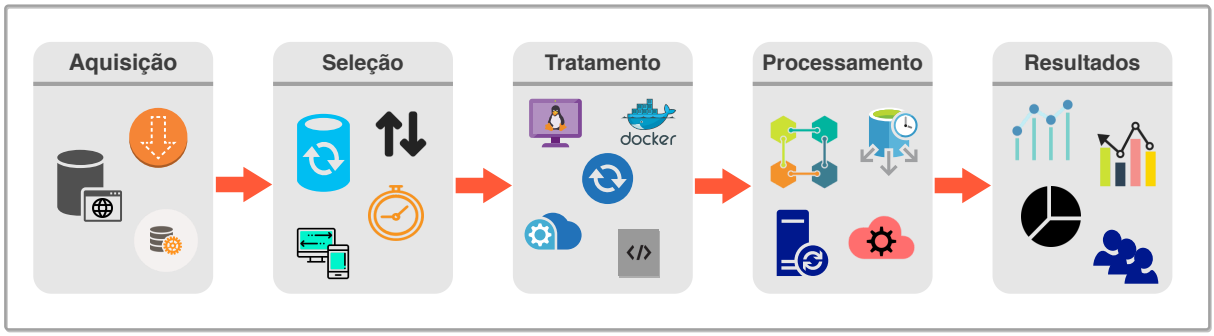
A metodologia adotada nesta dissertação engloba a aplicação do processo de KDD como processo principal de extração do conhecimento em base de dados, que tem como um de seus objetivos a utilização das informações adquiridas para a tomada de decisão. Assim, os dados utilizados como objeto de estudo desta pesquisa são oriundos de instituições governamentais brasileiras (IBGE, STN e Previdência Social) que apresentam dados de acesso aberto e público.

A implementação desta metodologia visa inicialmente compreender as relações empíricas entre dados previdenciários (aposentadorias em geral, auxílios, pensões e *etc.*), fundo participação dos municípios e PIB dos municípios brasileiros entre o período de 2010 a 2017 através de uma abordagem previdencial, como é disposto na seção 2.5. Além disso, é realizada outra abordagem por meio análise fiscal, cujo, objetivo é prever a dinâmica do ciclo econômico do município de São Paulo através do modelo STVAR adaptado por Auerbach e Gorodnichenko (2012a), mediante aplicação de choques exógenos (BEZERRA; SILVA; LIMA, 2014) na variável de despesa do município na etapa de teste 2013-7 e 2017-12.

Apesar das técnicas de análise apresentarem abordagens diferentes, ambas inicialmente percorreram todas as etapas contidas na fase de pré-processamento, como exemplo:

aquisição, seleção e tratamento de dados, para facilitar os demais processos de análise. Posteriormente, são aplicadas as fases que processam as informações (dados) e por fim à fase de pós-processamento de dados. Para uma melhor visualização do processo metodológico adotado.

Figura 4 – Percurso metodológico adotado.



Fonte: Elaborado pelo autor, (2020).

Na Figura 4 são brevemente ilustradas as etapas de análise de dados aplicadas nesta trabalho dissertação, para uma compreensão do fluxo de execução.

- Na etapa de aquisição é realizada a coleta de dados, em diferentes bases de dados. Pois, estes são alocados de forma descentralizada.
- Na etapa de seleção é realizada a triagem dos dados necessários em qualquer pesquisa a partir de base de dados já prontas.
- Na etapa de tratamento consistiu em verificar se os dados apresentam alguma anomalia, tais como: ruídos, dados em branco e dados com formatos inadequados, sem comprometer a integridade dos dados.
- Na etapa de processamento são aplicadas técnicas, como: processo de KDD, análise estatísticas e Modelo de estimação não-lineares na busca por conhecimento úteis e novos.
- Na etapa dos resultados são analisadas as saídas obtidas pelos algoritmos e as demais técnicas de análise, bem como a interpretação dos resultados obtidos junto ao especialista da área.

4.3 Considerações Finais

Neste capítulo foi apresentado o ambiente computacional utilizado para aplicação da análise de dados desta dissertação, por meio de ferramentas *open source* (Python, Miniconda, Pandas e *etc.*) que visam contribuir para uma melhor produtividade das

analises. Buscou-se também identificar quais as vantagens computacionais a partir da sua utilização. E, por fim, é apresentado o percurso metodológico empregado em ambos os estudos de caso.

5 Resultados e Discussões

Neste capítulo são apresentados dois estudos de caso a partir da aplicação do processo de KDD sobre as bases de dados dados abertas e públicas referentes aos microdados da economia dos municípios brasileiros, tanto por uma abordagem previdenciária, quanto por uma análise fiscal. E, por fim são apresentados os resultados obtidos através da aplicação do percurso metodológico adotado nesta dissertação conforme apresentado no capítulo anterior.

5.1 Estudo de caso I

No primeiro estudo de caso é realizada uma análise através de uma abordagem previdenciária, cujo, objetivo é identificar quais as transferências do RGPS (aposentadorias, pensões, auxílios, benefícios de legislação específica e *etc*) aos municipais são mais relevantes para manter a saúde fiscal destas repartições estáveis. Por outro lado, é feita uma análise a partir das emissões do Fundo de Participação dos Municípios (FPM) para mensurar o impacto desta transferência nas contas municipais.

5.1.1 Aquisição de Dados

O processo de aquisição de dados é compreendido como a etapa inicial de qualquer tipo de metodologia voltada para a análise de dados, caso essa seja uma nova área este processo pode ser demorado, pois dependendo da área tida como objeto de estudo, esta etapa pode demandar anos até que todos os dados estejam disponíveis (ORAIR; SIQUEIRA; GOBETTI, 2016).

Dessa forma, os dados necessários para a análise podem estar armazenados de forma descentralizada, não conectada e também podem não pertencer a uma única instituição em geral, podendo tornar a análise insustentável (TAN; STEINBACH; KUMAR, 2009). Outro ponto interessante sobre aquisição de dados é a busca por informações junto ao especialista da área para determinar se a pesquisa tem cunho inédito ou se esta já apresenta algum tipo de análise inicial, pois assim diminui-se o esforço durante o processo de aquisição de dados.

Assim, os microdados utilizados neste estudo de caso estão disponíveis no site do DATAPREV de forma descentralizada e não conectada e, estes são organizados em três arquivos diferentes para cada ano analisado. Cada arquivos contém as seguintes informações, os valores e as quantidades das transferências previdenciárias, bem como as arrecada-

ção municipais. Outros microdados são obtidos no site Secretaria do Tesouro Nacional, especificamente o FPM dos municípios que transferências da União aos municípios.

Além destes, também foram selecionados arquivos que contém dados referentes aos PIB's e as projeções populacionais, estes dados podem ser encontrados a partir do site IBGE. Vale ressaltar que todas bases de dados mencionadas anteriormente estão compreendidas entre 2010 e 2017, pois são estas apesentam maior quantidade de não nulos. A Tabela 1 abaixo ilustra os locais, aquivos, extensão, tamanho, quantidade e período referente ao ano em que os microdados foram armazenados.

Tabela 1 – Descrição e origem dos microdados utilizados no estudo de caso I.

Locais	**Arquivo**	**Extensão**	**Tamanho**	**Quantidade**	**Período**
DATAPREV	QTE	*.xlsx*	2.5 megabytes	8	2010 a 2017
DATAPREV	VTE	*.xlsx*	2.2 megabytes	8	2010 a 2017
DATAPREV	VAM	*.xlsx*	2.2 megabytes	8	2010 a 2017
IBGE	PPM	*.csv*	4.0 megabytes	8	2010 a 2016
IBGE	EPM	*.csv*	3.3 megabytes	8	2010 a 2017
STN	FPM	*.csv*	70 megabytes	1	2010 a 2017

Fonte: Elaborado pelo autor.

Na Tabela 1 são descritos arquivos utilizados, sendo eles: Quantidade de transferências Emitidas (QTE), Valores das Transferências Emitidas (VTE), Valores Arrecadados por Município (OBE), Projeção da População por Município (PPM), Estimativa do PIB Municipal (EPM)). Todos arquivos utilizados nessa análise são de livre acesso a comunidade em geral.

5.1.2 Seleção de Dados

Na etapa de seleção foram identificadas as variáveis cruciais para o processo de análises de dados, ou seja, aquelas variáveis julgadas como menos relevantes foram descartadas, como por exemplo: Renda per capita, Participação Percentual (Acumulada e Relativa) e a quantidade anual de benefícios transferidos em dezembro por município. Após, o processo de triagem de dados, foi necessário compactar os arquivos em apenas um para facilitar as manipulações de dados subsequentes, o qual resultou em arquivo com 17 variáveis e 46.160 instâncias. Na Tabela 2 são ilustradas as variáveis mais relevantes para análise de dados.

Tabela 2 – Variáveis selecionadas para análise.

Variáveis	**Descrição**
ApoPI_Mun_VA	Valor anual das aposentadorias por idade municipais
ApoTC_Mun_VA	Valor das aposentadorias por tempo contribuição municipais
Arr_Mun_VA	Valor anual das arrecadações municipais
Pib_Mun_VA	Valor anual do PIB dos municipais
Pen_Mun_VA	Valor anual dos pensões municipais
Aux_Mun_VA	Valor anual dos auxílios municipais
Apo_Mun_VT	Valor anual total de aposentadorias
Ben_Mun_VT	Valor anual total de benefícios
FPM_Mun_VA	Valor anual do Fundo de Participação do Municípios
Ano	Ano de referência das coletas

Fonte: Elaborado pelo autor.

Em seguida, foram identificados alguns valores de arrecadação ausentes na base de dados o que poderia limitar ou guiar as análises a resultados tendenciosos. Mediante a essa problemática optou-se por realizar uma limpeza nos dados faltosos para manter a veracidade dos mesmos, o qual resultou em 28.896 instâncias como informações completas da série de dados obtida.

5.1.3 Tratamento de Dados

Partindo do processo de seleção e limpeza de dados realizado na base, foi realizada uma categorização a partir do porte dos municípios conforme o processo de classificação por habitantes aplicada em (SÓLON et al., 2019). Posteriormente, os dados foram divididos em seis subconjuntos de dados, conforme ilustrado na Tabela 3. Após a classificação dos dados, foram retiradas amostras aleatórias de cada classe de município, as quais correspondem 80% de instâncias por agrupamento pelo porte para as análises.

Tabela 3 – Discretização dos municípios número de habitantes.

Classes	**Porte do Município**
FH1	[0 a 5000 habitantes]
FH2	[5000 a 10000 habitantes]
FH3	[10000 a 20000 habitantes]
FH4	[20000 a 50000 habitantes]
FH5	[50000 a 100000 habitantes]
FH6	[Acima de 100000 habitantes]

Fonte: Elaborado pelo autor.

Anteriormente ao processo de discretização de dados, foi realizada uma projeção dos PIB's municipais referente ao ano de 2017 mediante ao método dos mínimos quadrados

ordinários (MQO), porque durante o processo de analise foi identificado que o órgão responsável por essa estimativa ainda não havia feito a sua projeção. Isso fez como houvesse a necessidade de utilizar ferramentas computacionais para ajudar no tempo de estimação desta variável, pois este processo de estimação deveria ser aplicado a para cada município presente na analise.

5.1.4 Processamento de Dados

Etapa de processamento de dados aplicado neste estudo de caso, foi realizada mediante o uso métodos estatísticos descritivos (Tabela, Gráficos, *Histograma* e *etc*), pois os métodos estatísticos descritivos possibilitam uma análise visual de fácil compreensão, sem a contrapartida de qualquer inferência estatística. Todavia, essa visualização simplesmente nortear o especialista em análise de dados (ou, em KDD) a escolher técnicas analíticas de maneira mais sistemática ao problema em questão.

Assim, as inferências estatísticas aplicadas sobre microdados dos municípios brasileiros têm o proposito de comparar o relação entre a diferença de transferências previdenciárias ou do repasses da União o FPM com os valores arrecadados pelos próprios municípios, para mensurá a saúde fiscal destes, além de identificar quais os possíveis municípios não conseguiriam manter as contas equilibradas caso os repasses da União e da previdência Social não fossem mantidos.

5.1.5 Resultados

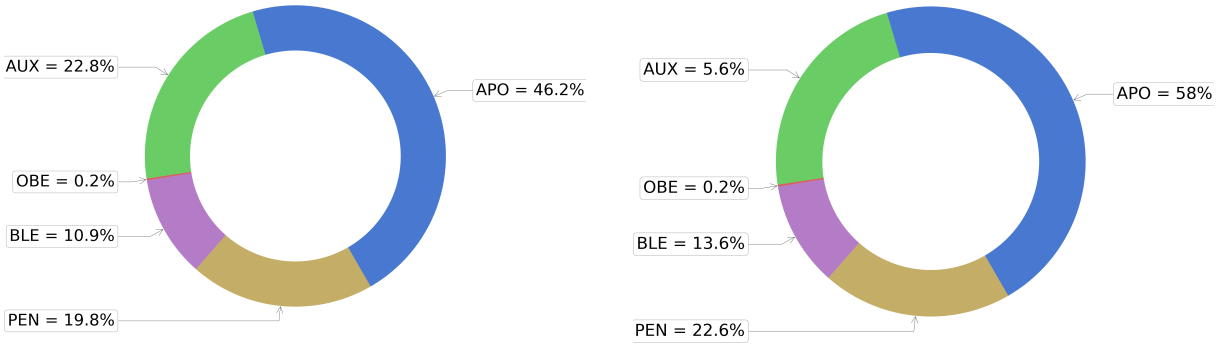
Inicialmente, foi realizado uma análise para quantificar quais transferências previdenciárias municipais (Aposentadorias - (APO), Auxílios - (AUX), Outros Benefícios - (OBE), Benefícios de Legislação Específica - (BLE) e Pensões - (PEN)) apresentavam o maior valor dos emissões do RGPS aos municípios brasileiros, conforme é ilustrado na Figura 5.

Na Figura 5 é evidenciado que entre todos os benefícios previdenciários é importante salientar que as aposentadorias em geral são os benefícios que se apresentam em maiores quantidades e estes também são responsáveis por mais de 50% das transferências previdenciárias emitidas no cenário atual. Sendo possível observar que entre 2010 a 2017 houve um aumento 11.8% na quantidade de aposentadorias, as quais correspondem a R$: 181.038 bilhões de reais das despesas públicas.

Partindo da observação dos repasses de aposentadorias emitidas durante o ano 2017. Foi realizada uma análise para identificar qual o tipo de aposentadoria que apresentava o maior impacto nas contas de uma dada região, ou seja, aquela que representa a maior parcela de valores transferidos ao ano. Então, a partir da análise aferida é possível observar que as regiões norte, nordeste e centro oeste possuem o maior quantitativo de municípios

Figura 5 – Percentual dos repasses previdenciários emitidos em 2010 e 2017.

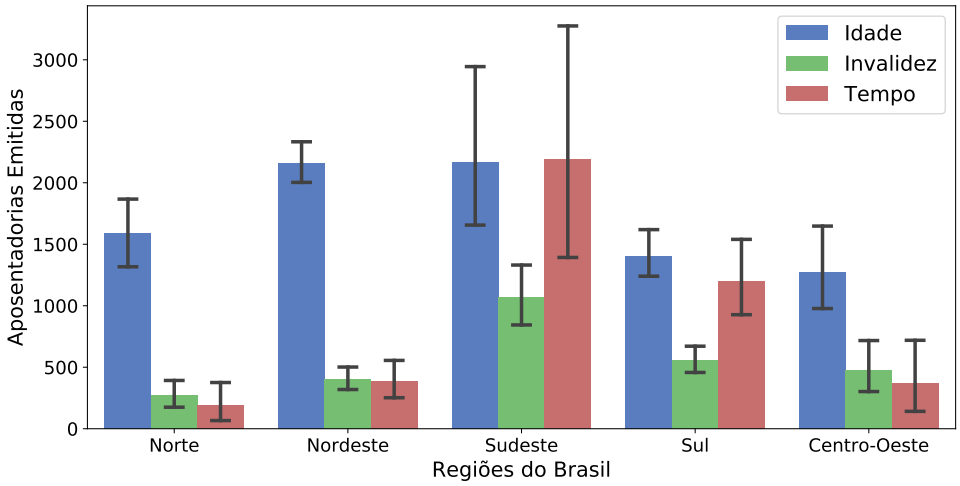
(a) Repasses emitidos em 2010. (b) Repasses emitidos em 2017.



Fonte: Elaborado pelo autor.

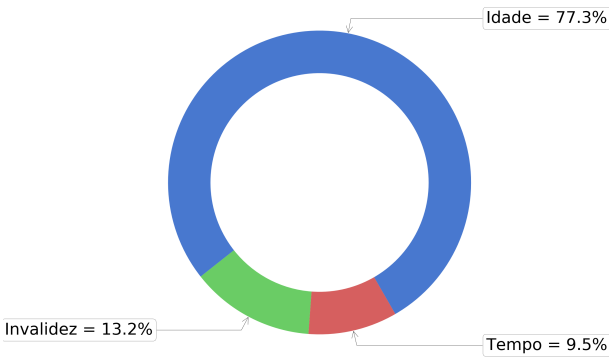
que a aposentadoria por idade é maior que 59% quando comparadas às demais que apresentam uma distribuição mais uniforme destes repasses, conforme as Figuras 6 e 7

Figura 6 – Total de emissões de aposentadorias em 2017 por região.



Fonte: Elaborado pelo autor.

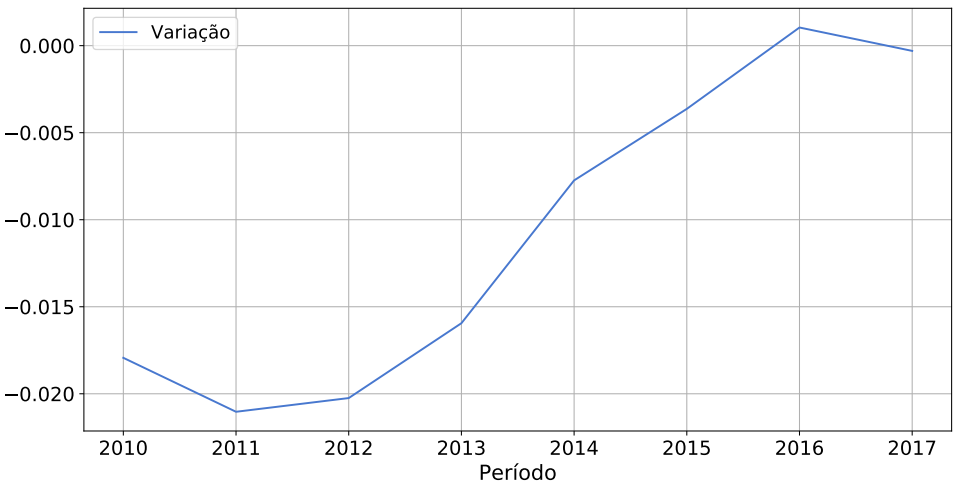
Figura 7 – Percentual de aposentadorias emitidas a região norte em 2017.



Fonte: Elaborado pelo autor.

A segunda análise aplicada neste trabalho foi para constatar que os municípios que possuem população superior a 100.000 mil habitantes apresentam um *superavit* fiscal econômico. Essa afirmação é embasada na diferença entre o montante arrecadado pelos municípios e as transferências do RGPS realizados pela União (SÓLON et al., 2019). Ou seja, os municípios que detém o número de indivíduos menor que 100.000 mil não conseguem pagar suas contas unicamente com base nas suas próprias receitas, como disposto na Figura 8.

Figura 8 – Diferença entre as transferências previdenciárias e arrecadações dos municípios com mais de 100 mil habitantes durante 2010 e 2017.



Fonte: Elaborado pelo autor.

Esta inferência foi feita entre os anos de 2010 a 2017 para avaliar se esta ocorrência era um fenômeno isolado ou se estes vinham ser propagando ao longo dos anos na economia dos municípios. Por outro lado, apenas no ano de 2016 os municípios que apresentavam suas populações superiores a 100.000 habitantes mostraram um saldo positivo

essa ocorrência pode estar relacionada ao fato do Brasil ter sofrido um rebaixamento do grau de investimento para grau especulativo[12].

Na Tabela 4 é possível observar que os municípios no ano 2010 que estavam inseridos na faixa de até 5 mil habitantes possuíam o maior desequilíbrio fiscal, onde o volume de benefícios pagos supera os de arrecadação gerando uma despesa de 936 milhões de reais ao ano, a qual impacta diretamente no PIB em 6,2% do conjunto de municípios estudado. No entanto, no ano 2017, os municípios com até 5 mil habitantes foram superados pelos que tinham populações entre 10 mil e 20 mil habitantes, os quais apresentam um défice de 84 milhões de reais, o que representa 6,7% do PIB.

Essa alteração está relacionada ao fator do crescimento populacional, o que faz com que municípios troquem de faixas e está troca de faixa condicionada ao aumento da expectativa de vida dos indivíduos, porque quanto maior for a expectativa de vida maior, concomitantemente será o número de benefícios previdenciários emitidos. Porém, os demais anos analisados sobre essa classificação não indicavam carência para manter as suas contas em dia, como é ilustrado na Tabela 4.

Tabela 4 – O comprometimento do PIB pela diferença entre arrecadações e despesas.

Faixas	**2010**			**2017**		
	DS	**PS**	Δ(%)	**DS**	**PS**	Δ(%)
FH1	0.936	15.047	6,2	1.829	28.356	6,4
FH2	1.575	28.220	5,6	3.186	56.837	6,0
FH3	2.784	48.422	5,8	5.675	84.515	6,7
FH4	5.213	118.311	4,4	11.226	188.947	5,9
FH5	9.135	235.642	3,9	22.209	437.779	5,0
FH6	-39.539	2.204.673	-1,8	-7.499	2.487.596	-0,3

Legenda: Δ impacto percentual no PIB; DS: somatório dos défices (arrecadação - despesa)e PS: é o somatório dos PIB municipais.

Fonte: Elaborado pelo autor.

Outra análise interessante é a relação entre os municípios em que os seus fundos de participação são superados pelos benefícios previdenciários pagos - BPG. Assim, a partir da Tabela 5 é possível observar que entre as cinco grandes regiões brasileiras a região sul se destaca por apresentar a maior taxa (76.17%) de municípios em que transferências do FPM é insuficiente para sanar as suas despesas (SÓLON et al., 2019). Em contrapartida a região norte é a que apresenta o maior percentual (48.99 %) de municípios em que repasses do FPM são suficientes para pagar os benefícios previdenciários.

[12] https://www.parmais.com.br/blog/economia-brasileira-em-2016/

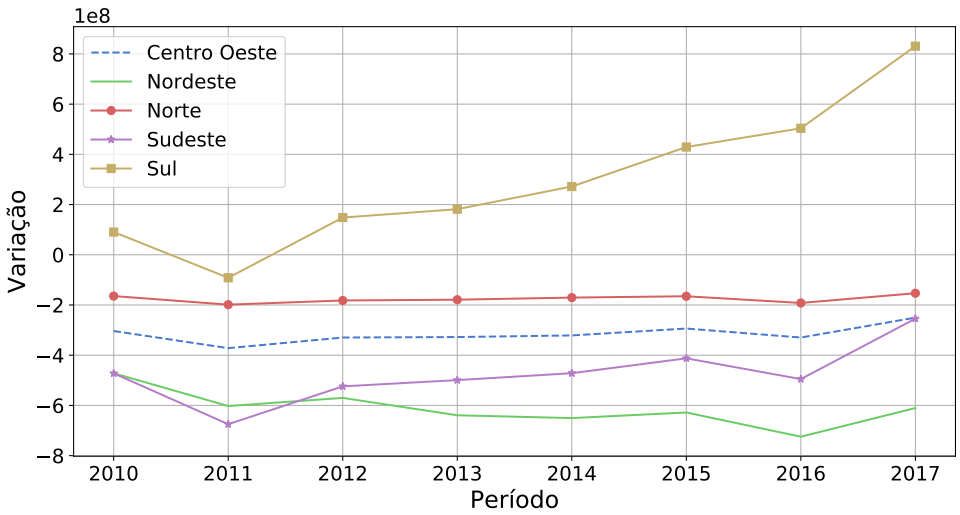
Tabela 5 – Percentual de municípios em que FPM e menor que os benefícios pagos.

Região	**2010** **BPG > FPM**	(%)	**2017** **BPG > FPM**	(%)	**Municípios**
Centro Oeste	275	59.01	329	70.44	467
Nordeste	1207	67.31	1341	74.79	1794
Norte	220	48.99	298	66.22	450
Sudeste	1255	75.23	1351	80.99	1668
Sul	905	76.17	982	82.45	1191
Total	**3862**	-	**4301**	-	**5570**

Fonte: Elaborado pelo autor.

Os municípios enquadrados nas distribuições populacionais estabelecidas na metodologia desta dissertação, foi observado que os municípios da região sul que tinham suas população com até 5 mil habitantes, demostravam um crescimento positivo o que indica que o endividamento é cada vez mais menor em relação ao comprometimento do PIB. Por outro lado, as demais regiões brasileiras apresentam uma de dinâmica do crescimento inverso ao da região sul em relação ao comprometimento PIB através da mesma perspectiva de classificação populacional de 5 mil habitantes por município, conforme o ilustrado na Figura 9.

Figura 9 – Diferença entre o endividamento dos municípios com até 5 mil habitantes por região.



Fonte: Elaborado pelo autor.

Vale ressaltar que essa diminuição do défice da região sul nos últimos anos também pode ser vista a partir de outro ângulo, isto é, quando analisamos relação dessa diminuição do endividamento com os Índices de Desenviolamento Humanos (IDH's) nacionais. Onde é possível analisar que dentre todas as regiões brasileiras a região sul é a que aproxima-se mais dos valores médios nacionais, conforme é ilustrado na Tabela 6.

Tabela 6 – Comparação entre o IDH's nacionais com os IDH's das regiões brasileiras para os municípios que possuem até 5 mil habitantes.

IDH's	**(M)**	**(E)**	**(L)**	**(R)**
Nacional**	0.7270	0.6370	0.8160	0.7390
Centro Oeste	0.6870	0.5825	0.8190	0.6700
Nordeste	0.5810	0.4770	0.7520	0.5530
Norte	0.6355	0.5305	0.7920	0.5980
Sudeste	0.6860	0.5910	0.8210	0.6670
Sul*	0.7100	0.6075	0.8350	0.7040

Fonte: Elaborado pelo autor.

Na Tabela 6 são ilustrados os quatro IDH's mais revelantes no mundo, sendo eles: municipal (M), educacional (E), longevidade (L) e renda - (R). Ao analisa-los, é possível notar que a região nordeste é aquela tem o menor valor médio dos IDH's e que o seu IDH de renda apresenta uma diferença de 0.1857 em relação a média nacional, enquanto a região sul exibe a valor de 0.035 de diferença.

Assim, algumas hipóteses são levantadas a respeito dessa diferença de IDH's entre as regiões brasileiras, uma delas está relacionada ao fato da região nordeste apresentar a segunda maior população do Brasil e uma grande parte da sua população morando na zona rural o que corresponde a metade da população brasileira (IBGE, 2010), sendo que o sua maior fonte de renda é provida das atividades exercidas na zona rural.

5.2 Estudo de caso II

No segundo estudo caso é feita uma análise fiscal sobre os microdados da cidade de São Paulo, cujo, objetivo foca na utilização do modelo STVAR para correlacionar os parâmetros fiscais de despesas, receitas e PIB, para mensurar o impacto que cada variável econômica pode ocasionar no comportamento do ciclo econômico analisado. Esse impacto é medido via a função (Resposta a Impulso), ou seja, é injetado um choque exógeno no modelo com a finalidade de medir o tempo de duração do mesmo no ciclo econômico fiscal.

5.2.1 Aquisição de Dados

Os microdados utilizados no estudo de caso II são disponibilizados pelo IBGE, Sistema de Informações Gerenciais da Execução Orçamentária (SIGEO-BI)[13] e o SEADE. O SIGEO-BI e o SEADE disponibilizam dados da economia fiscal do estado de São Paulo, pois mediante a busca por microdados foi evidenciado que São Paulo disponibiliza dados com periodicidade mensal, de livre acesso e, além disso, São Paulo apresenta um comportamento econômico semelhante ao ciclo econômico nacional brasileiro.

[13] https://portal.fazenda.sp.gov.br/servicos/sigeo-bi

Assim, o processo de aquisição de dados foi realizado a partir de minuciosas coletas individuas dos arquivos de despesas e receitas, as quais estão disponíveis no site SIGEO-BI e só podem ser coletadas uma a uma em arquivos com formato: *csv*, *xlsx*, PDF, *xml* e *etc.* Por outro lado, as variáveis de taxa de crescimento do produto e PIB foram catalogadas via o site SEADE que também dá acesso livre aos microdados e, por fim, o Índice de Preços ao Consumidor Amplo (IPCA).

5.2.2 Seleção de Dados

Na etapa de seleção de dados iniciou a partir da identificação das variáveis mais relevantes para o processo de análise de dados, ou seja, para cada aquivo baixado foi eleita e selecionada as informações de despesa e receita total para compor a base de dados. Este mesmo processo seleção de variáveis foi aplicado nos aquivos com informações do PIB, taxa de crescimento do produto e IPCA, conforme a Tabela 7.

Tabela 7 – Variáveis utilizadas no estudo de caso II.

Período	**Despesa**	**Receita**	**Taxa**	**IPCA**	**PIB**
01-01-2002	3.568.584.139,17	5.858.718.804,02	0	0.52	73,4
01-02-2002	3.586.141.273,12	3.544.351.995,94	0.2	0.36	73,5
01-03-2002	3.835.692.682,06	4.645.578.841,50	-0.8	0.6	72,8
⋮	⋮	⋮	⋮	⋮	⋮
01-12-2017	34.643.300.073,18	22.029.972.465,90	0.5	0.44	101,1

Fonte: Elaborado pelo autor.

Assim, a base de dados utilizada na segunda análise é constituída por cinco variáveis, as quais estão associadas as despesas, IPCA, receitas, taxa de crescimento do produto e o PIB. Estas variáveis apresentam periodicidade mensal e cada uma delas contém 180 instâncias ao total.

5.2.3 Tratamento de Dados

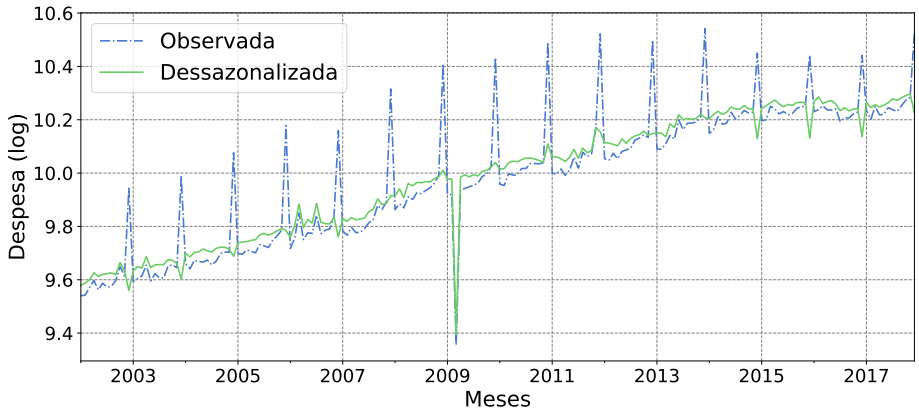
As séries utilizadas neste trabalho cobrem o período de janeiro de 2002 a dezembro de 2017, estas foram convertidas para milhões de reais e deflacionadas a partir do IPCA de dezembro de 2017. Em seguida, as séries foram convertidas valores logarítmicos, pois este procedimento é frequentemente usado na economia para reduzi escala. Além disso, as séries de despesa, receita e PIB foram dessazonalizadas via o sistema ARIMA X-13[16].

O ARIMA X-13 é *software* amplamente utilizado para realizar a ajustes sazonais do *Census Method II*. Após os procedimentos de conversão, redução e ajuste sazonal na base de dados temos representação das três principais variáveis da análise, conforme as ilustrações das Figuras 10a, 10b e 10c.

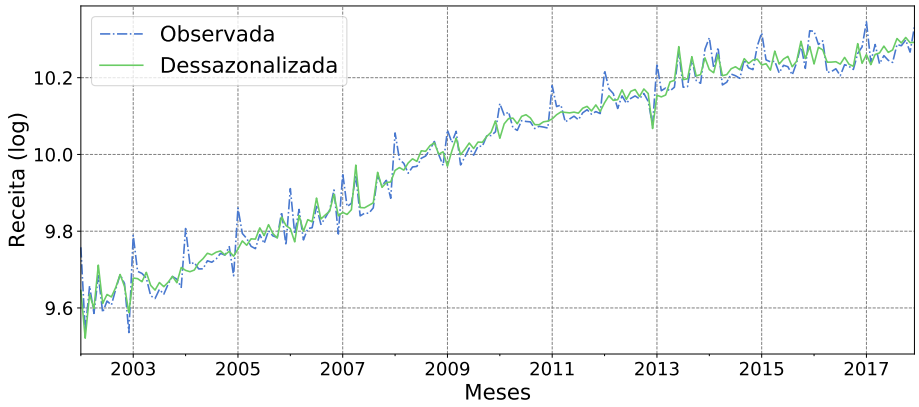
[16] http://www.statsmodels.org/0.6.1/_modules/statsmodels/tsa/x13.html

Figura 10 – Séries fiscais municipais observadas e ajustas via ARIMA X-13.

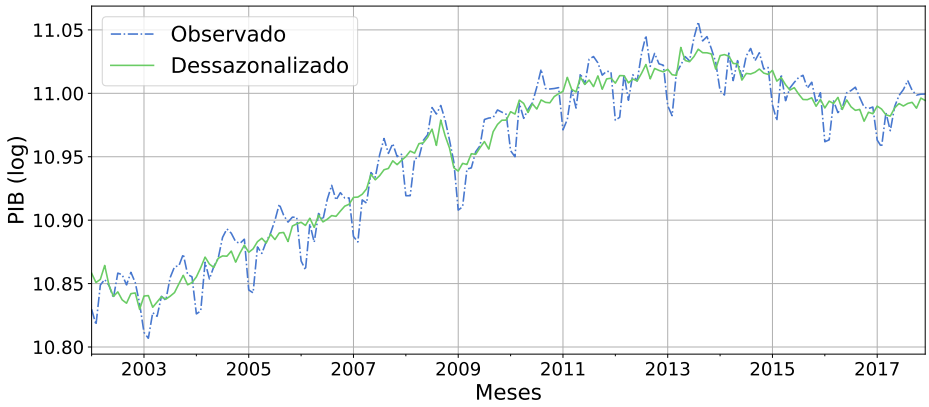
(a) Série mensal de Despesa.



(b) Série mensal de Receita.



(c) Série mensal do PIB.



Fonte: Elaborado pelo autor.

As Figuras 10a, 10b e 10c acima referem-se ao processo de análise sazonal das séries de despesa, receita e PIB. Assim, a linhas azuis demonstram o comportamento real das variáveis fiscais da cidade de São Paulo, enquanto as linhas vermelhas indicam a retirada dos períodos sazonais encontrados em cada parâmetro fiscal. É importante mencionar que

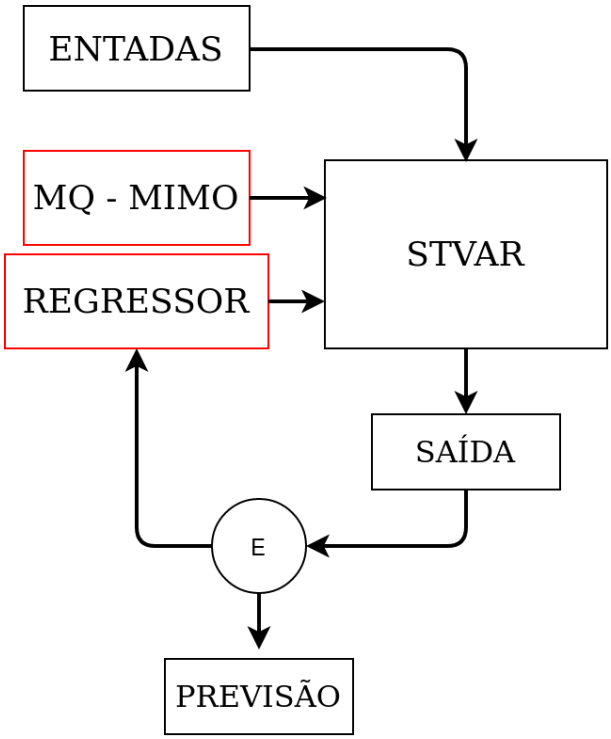
séries observadas estão em escala logarítmica.

Por conseguinte, foi realizado um processo de triagem nos dados para que estes reportem o histórico de recessão e expansão da economia, os quais são usados para calcular as matriz de variância-covariância Ω_R e Ω_E, conforme a Equação 2.11. Em seguida, é aplicada a decomposição de Cholesky sobre as matriz de variância-covariância obtidas, para estimar os resídios do modelo, como disposto na Equação 2.10.

5.2.4 Processamento de Dados

Após a construção e o tratamento da base de dados foi utilizada a abordagem do modelo STVAR adaptada pelos autores Auerbach e Gorodnichenko (2012a), Auerbach e Gorodnichenko (2012b) conforme os sistemas de equações apresentado na seção 2.5 desta dissertação, cuja, finalidade é a obtenção de multiplicadores fiscais capazes de simular o impacto fiscal aplicado a partir dos estados de expansão e recessão do ciclo econômico analisado. Na Figura 11 é representado o fluxograma de execução do modelo STVAR aplicado.

Figura 11 – Fluxograma do modelo STVAR.



Fonte: Elaborado pelo autor.

O fluxograma exposto na Figura 11 apresenta as seguintes etapas: a) MQ-MIMO: consiste na estimação de coeficientes matriciais via ao mínimos quadrados mimo, para trabalho usado 70% dos dados para essa tarefa; b) Entradas: são os próprios dados de teste utilizados do modelo; c) Regressor: representa a quantidade de informações defasadas

utilizadas no modelo para uma projeção; d) Modelo STVAR: é processamento e estimação a partir das informações de entrada; e) Saída: corresponde a projeção do modelo e f) Projeção: consiste no melhor resultado encontrado.

Assim, alguns autores como (ORAIR; SIQUEIRA; GOBETTI, 2016) ressaltam que uma das maiores vantagens em utilizar modelo STVAR está relacionada a sua capacidade de estimar os regimes por meio da variação do grau (ou probabilidade) em que se encontra um regime em particular. Porém, uma das maiores desvantagens na utilização do modelo STVAR é a sua alta sensibilidade na adequação da variável z_t. Pois, na literatura não há uma especificação da melhor forma de calibragem de z_t.

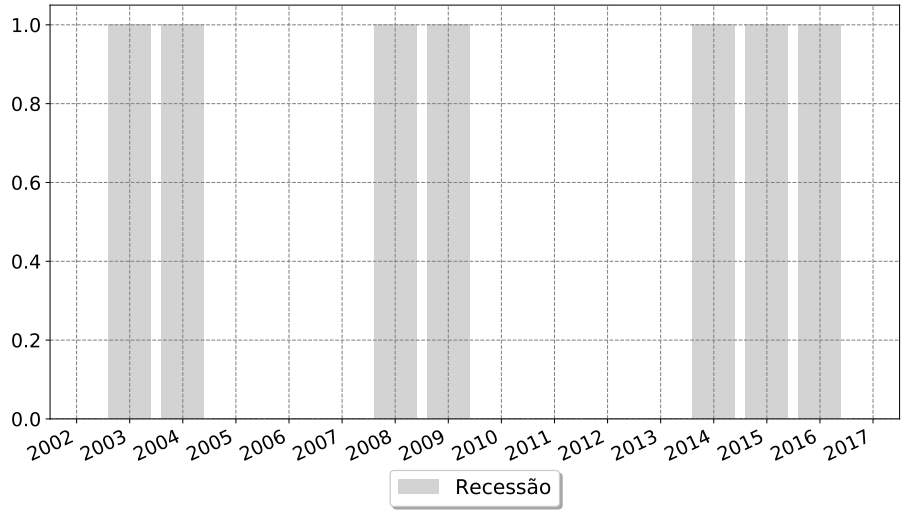
Apesar de haver inúmeras discussões voltadas ao ajuste de z_t, neste estudo é utiliza a metodologia empregada nos trabalhos (ORAIR; SIQUEIRA; GOBETTI, 2016; AUERBACH; GORODNICHENKO, 2012a), os quais utilizam médias moveis das taxas de crescimento do produto para calcular o valor de z_t sob a justificativa pragmática da sua fácil aplicabilidade.

Assim, no trabalho de Auerbach e Gorodnichenko (2012a) o calculo de z_t é realizado a partir da taxa de crescimento de seis meses consecutivos do PIB, neste estudo é adotada a mesma metodologia. Pois, uma recessão econômica pode ser compreendida de forma literal como uma contração no ciclo econômico a partir de dois trimestres ou seis meses contínuos da taxa de crescimento do produto. Além disso, foi realizada uma normalização e uma transformação nos dados. Pois, é a partir destas especificações que podemos obter os valores de entrada necessários para calculo de $F(z_t)$.

Em seguida, é realizada a calibragem do parâmetro $\gamma = 1,88$ para que este apresente a quantidade períodos recessivos via $F(z_t)$, para que equipare com datação do Comitê de Datação de Ciclos Econômicos (CODACE)[17] do Instituto Brasileiro de Economia da Fundação Getúlio Vargas (IBRE/FGV)[18], onde é mostrado que a economia brasileira esteve por cerca de 22.05% aproximadamente em período de recessão entre 2002 janeiro e 2017 dezembro, conforme é ilustrado na Figura 12.

[17] https://portalibre.fgv.br/estudos-e-pesquisas/codace/
[18] https://portal.fgv.br/fgv-ibre

Figura 12 – Datação dos períodos de recessão do ciclo econômico pelo CODACE.



Fonte: Elaborado pelo autor.

Portanto, a ultima etapa realizada foi estimação dos coeficientes matriciais de correlação dos estados de recessão (Π_R) e expensão (Π_E) via o método de mínimos quadrados MIMO, conforme a Equação 2.8. Assim, este processo de estimação dos coeficientes matriciais são demonstrados passo a passo a partir das manipulações vetoriais ilustradas abaixo. Onde $\bar{Y}$ representa vetores de gastos ($G(t)$), receita ($R(t)$) e PIB ($P(t)$) que são os parâmetros de entrada do MQ-MIMO.

$$\bar{Y} = \begin{bmatrix} G(1) & G(2) & \cdots & G(N) \\ R(1) & R(2) & \cdots & R(N) \\ P(1) & P(2) & \cdots & P(N) \end{bmatrix}$$

Para determinar $\bar{X}$ as informações defasadas de $G(t)$, $R(t)$ e $P(t)$ são relacionadas com o grau de recessão, o qual é obtido pelo Função $F(z_t)$, conforme a representação vetorial de $\Phi^T(t)$.

$$\bar{X} = \begin{bmatrix} \Phi^T(1) & \Phi^T(2) & \cdots & \Phi^T(N) \end{bmatrix}$$

$$\Phi^T(t) = \begin{bmatrix} [1 - F(z_{t-1})]G(t-1) \\ [1 - F(z_{t-1})]R(t-1) \\ [1 - F(z_{t-1})]P(t-1) \\ F(z_{t-1})G(t-1) \\ F(z_{t-1})R(t-1) \\ F(z_{t-1})P(t-1) \end{bmatrix}$$

$$\bar{X} = \begin{bmatrix} [1-F(z_0)]G(0) & [1-F(z_1)]G(1) & [1-F(z_2)]G(2) & \cdots & [1-F(z_{N-1})]G(N-1) \\ [1-F(z_0)]R(0) & [1-F(z_1)]R(1) & [1-F(z_2)]R(2) & \cdots & [1-F(z_{N-1})]R(N-1) \\ [1-F(z_0)]P(0) & [1-F(z_1)]P(1) & [1-F(z_2)]P(2) & \cdots & [1-F(z_{N-1})]P(N-1) \\ F(z_0)G(0) & F(z_1)G(1) & F(z_2)G(2) & \cdots & F(z_{N-1})G(N-1) \\ F(z_0)R(0) & F(z_1)R(1) & F(z_2)R(2) & \cdots & F(z_{N-1})R(N-1) \\ F(z_0)P(0) & F(z_1)P(1) & F(z_2)P(2) & \cdots & F(z_{N-1})P(N-1) \end{bmatrix}$$

Dessa forma, podemos aplicar os vetores $\bar{Y}$ e $\bar{X}$ a partir da equação 2.8, a qual é utilizada para estimação os coeficientes matriciais de $\hat{\Pi}_R$ e $\hat{\Pi}_E$, conforme os sistemas de equações ilustrados e as representações vetoriais abaixo.

$$\hat{\Pi} = \bar{Y}\bar{X}^T(\bar{X}\bar{X}^T)^{-1} \tag{5.1}$$

$$\hat{\Pi} = \bar{Y}_{(3xN)}\ \bar{X}^T_{(Nx6)}\ (\bar{X}_{(6xN)}\ \bar{X}^T_{(Nx6)})^{-1} \tag{5.2}$$

$$\hat{\Pi} = \bar{Y}_{(3xN)}\ \bar{X}^T_{(Nx6)} = Q_{1(3x6)} \tag{5.3}$$

$$\hat{\Pi} = (\bar{X}_{(6xN)}\ \bar{X}^T_{(Nx6)})^{-1} = (Q_{2(6x6)})^{-1} \tag{5.4}$$

$$\hat{\Pi} = Q_{1(3x6)} * (Q_{2(6x6)})^{-1} \tag{5.5}$$

$$\hat{\Pi} = Q_{1(3x6)} * Q_{3(6x6)} \tag{5.6}$$

$$\hat{\Pi}_{(6x6)} = \left[\begin{array}{ccc:ccc} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} & b_{(1,1)} & b_{(1,2)} & b_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} & b_{(2,1)} & b_{(2,2)} & b_{(2,3)} \\ a_{(3,1)} & a_{(3,2)} & a_{(3,3)} & b_{(3,1)} & b_{(3,2)} & b_{(,3)} \end{array}\right]$$

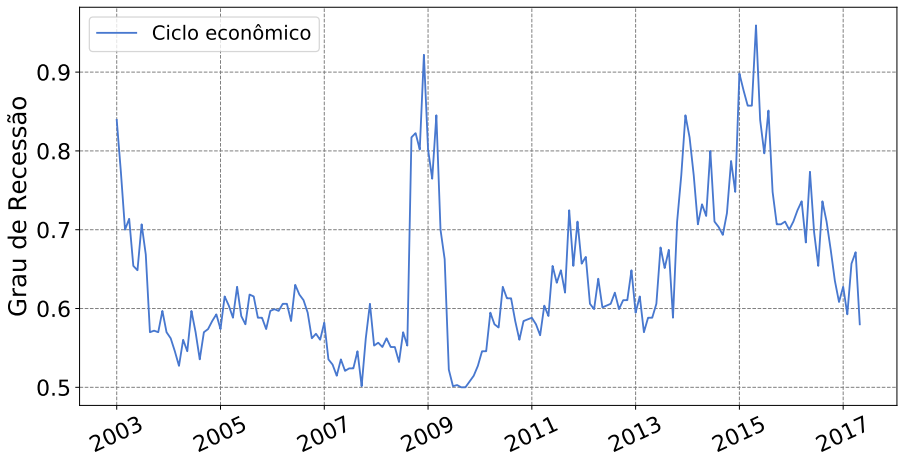
$$\hat{\Pi}_{R_{(3x3)}} = \begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} \\ a_{(3,1)} & a_{(3,2)} & a_{(3,3)} \end{bmatrix}$$

$$\hat{\Pi}_{E_{(3x3)}} = \begin{bmatrix} b_{(1,1)} & b_{(1,2)} & b_{(1,3)} \\ b_{(2,1)} & b_{(2,2)} & b_{(2,3)} \\ b_{(3,1)} & b_{(3,2)} & b_{(3,3)} \end{bmatrix}$$

5.2.5 Resultados

Para o modelo proposto nesta dissertação foi utilizada o probabilidade de $F(z_t) > 0,8$ de esta em período de recessão e seu complemento para o período de expansão. Assim, na Figura 13 são apresentados os valores estimados pela função $F(z_t)$ para modelar a dinâmica do ciclo econômico observado. É possível notar, que os resultados obtidos através da estimativa de $F(z_t)$ foram acetáveis, pois estes apresentam dinâmica similar aos dos períodos recessivos em particular datados pelo CODACE, o que viabiliza o prosseguimento desta análise.

Figura 13 – Estimação dos períodos de recessão via função $F(z_t)$.

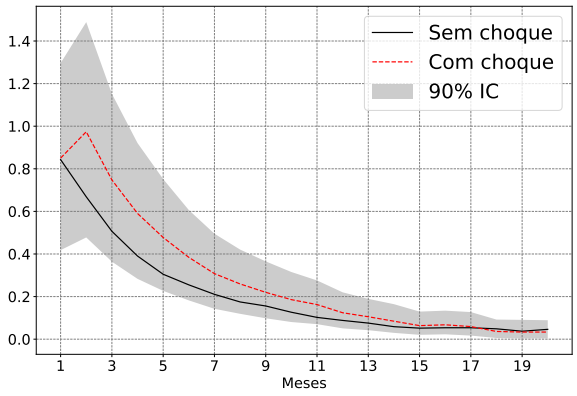


Fonte: Elaborado pelo autor.

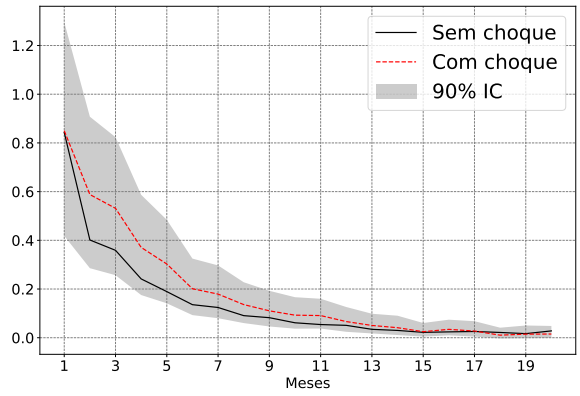
A seguir, são apresentados os resultados alcançados a partir dos experimentos econométricos de estimação via ao modelo STVAR. Apriori, é realizada uma análise para identificar o comportamento do ciclo econômico da cidade da São Paulo por meio das despesas agregadas, receitas e o PIB. Para essa atividade de observação da dinâmica econômica é aplicada um choque exógeno impulsivo (Função Resposta ao Impulso) de um real nas despesas agregadas durante 20 meses após o choque, o qual é utilizado para mensurar o comportamento de cada variáveis do ciclo econômico, conforme as Figuras 14a, 14b e 14c.

Figura 14 – Influência dos choques exógenos nos gasto G_t e a resposta no produto P_t.

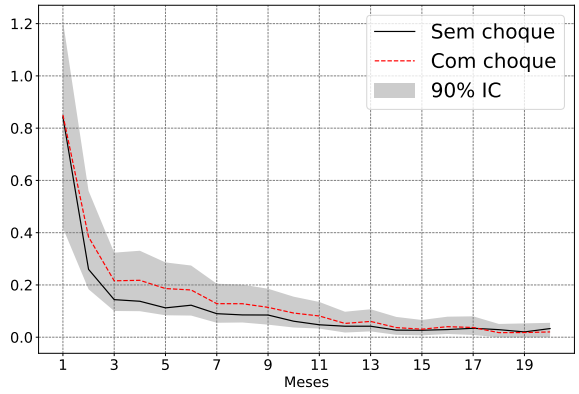
(a) Modelo Linear.



(b) Modelo Expansivo.



(c) Modelo Recessivo.



Notas: A linha preta se refere ao multiplicador de produto sem impulso exógeno, sendo que a área sombreada é o seu intervalo de confiança de 90%. A linha vermelha se refere ao multiplicador produto com impulso exógeno.

Fonte: Elaborado pelo autor.

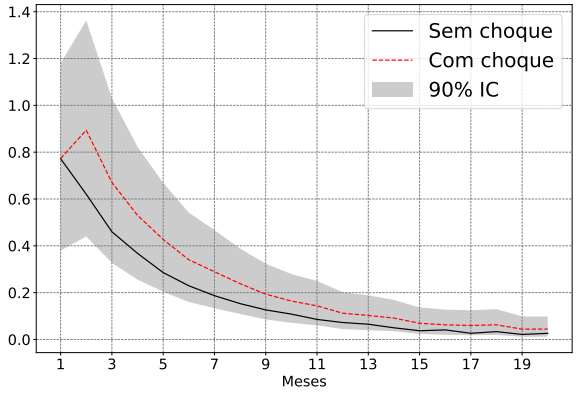
Ao se analisar as respostas do produto após aplicação de um adição monetária nas despesas do governo local, ou melhor, um impulso exógeno em G_t implicando da dinâmica de P_t, a qual exibe diferentes comportamentos levando em considerando os regimes de expansão e recessão analisados. Após a aplicação desse impulso na variável G_t é possível observar que no modelo linear o multiplicador P_t, atinge o seu ponto máximo no segundo período com o valor de 0.3111 de diferença entre P_t sem influência externas e com influências externas, a qual gera um alteração na economia durante 16 meses até o modelo estabilizar em 0.0560 a longo prazo, conforme a Figura 14a.

No regime de expansão, o P_t apresenta o seu pico em 0.1894, o qual ocorre no terceiro período analisado e permanece por 12 meses na economia até atingir o valor de 0.0320, que o seu ponto de estabilidade. Já no regime de recessão, o P_t apresenta o valor máximo de 0.0596, o qual pode ser observado no quarto mês estimado pelo modelo e este

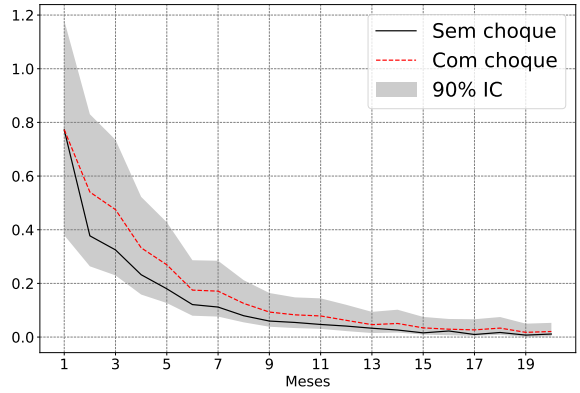
apresenta duração de 14 meses na economia conforme é disposto nas Figuras 14b e 14c.

Figura 15 – Influência dos choques exógenos nos gasto G_t e a resposta na receita R_t.

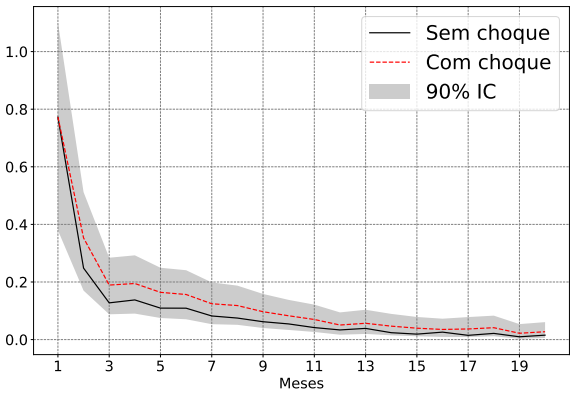
(a) Modelo Linear



(b) Modelo Expansão



(c) Modelo recessão



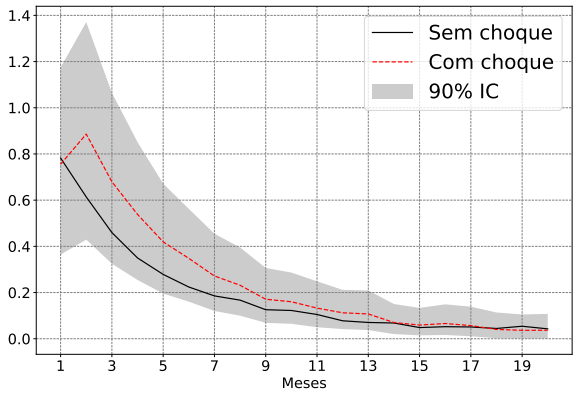
Notas: A linha preta se refere ao multiplicador de receita sem impulso exógeno, sendo que a área sombreada é o seu intervalo de confiança de 90%. A linha vermelha se refere ao multiplicador receita com impulso exógeno.

Fonte: Elaborado pelo autor.

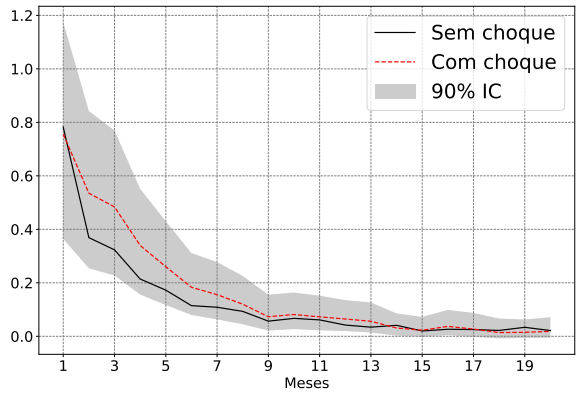
Observa-se que os multiplicador de R_t apresentam os menores valores no impacto do ciclo econômico analisado, quanto estes são comparados aos impactos da resposta ao impulso em G_t. Assim, para os modelos linear, expansivo, recessivo o multiplicado de G_t apresenta uma resposta de diferença máxima a partir segundo mês para modelo linear e expansivo e terceiro mês para modelo recessivo, com os respectivamente valores 0.2736, 0.1500 e 0.1047, sendo que ambos apresentam duração média de 15 meses para cada regime analisado, e estes passam convergir quando atingem os respectivamente valores 0.0625, 0.0506 e 0.0352, conforme nas Figuras 15a, 15b e 15c.

Figura 16 – Influência dos choques exógenos nos gasto G_t e a resposta nos gastos G_t.

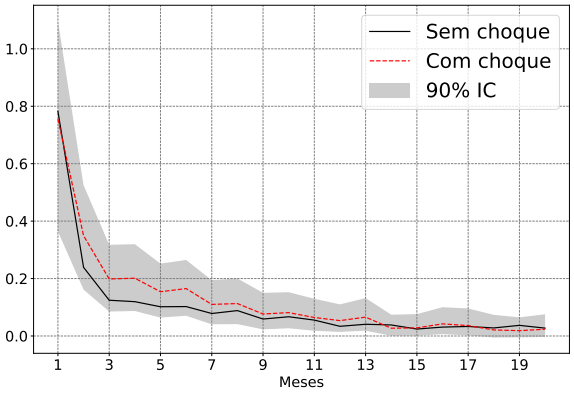
(a) Modelo Linear



(b) Modelo Expansão



(c) Modelo Recessão



Notas: A linha preta se refere ao multiplicador de gastos sem impulso exógeno, sendo que a área sombreada é o seu intervalo de confiança de 90%. A linha vermelha se refere ao multiplicador gastos com impulso exógeno.

Fonte: Elaborado pelo autor.

Ao se comparar o multiplicador de gastos, é notório que o choque exógeno aferido em G_t apresenta o maior impacto no ciclo econômico estudado. É importante frisar que essa análise se dá através de um choque impulsivo em G_t, implicando no próprio G_t. Assim, o modelo expansivo o qual exibe o valor 0.1438 com maior diferença no período analisado, esse impulso tem duração de 12 meses no ciclo econômico. No modelo linear os valores máximos encontrados assumem características mais acentuadas que o do modelo recessivo. Porém, o tempo de influência na economia do regime recessivo é mais elevada apresentado uma duração de dez meses enquanto o outro apresenta período de ocorrência de 12 meses, como ilustrado nas Figuras 16a, 16b e 16c.

Contudo, nota-se que a diferença entre os regimes é mais acentuada para o multiplicador de gastos do que para o de receita. Vale ressaltar, que assim como os multiplicadores de gastos e receita também apresentam comportamento de convergirem rapidamente para

valores estacionários em médio e longo prazo. Isto denota que os efeitos marginais ocorrem de forma mais acentuada nos primeiros 14 meses inicias do período estimado.

5.3 Considerações Finais

Neste capítulo, foram descritos dois estudos de casos voltados a análises de dados municipais econômicos do governo federal brasileiro. Inicialmente fez-se uma análise para constata a dependência municipal acerca dos repasses previdenciários (aposentadorias, pensões, auxílios e *etc.*) a partir de uma observação temporal de 16 anos. Posteriormente, é utilizado o modelo de estimação - STVAR para verificar os impactos e a duração dos choques exógenos nos multiplicadores fiscais (gastos, receitas e PIB) em economias locais.

6 Conclusões e Trabalhos Futuros

Neste capítulo, é apresentada as conclusões desta dissertação, assim como os trabalhos futuros que podem ser desenvolvidos a partir desta pesquisa. Também, são dispostas as dificuldades encontradas para ambos os estudos de casos e as publicações aceitas durante o período de mestrado, as quais imprimem a relevância do desenvolvimento deste estudo.

6.1 Conclusões

O presente trabalho apresentou dois estudos de caso voltados para análises de dados públicos do governo federal brasileiro. No primeiro estudo de caso, foi realizada uma análise frente aos dados econômicos municipais brasileiros durante o período de 2010 a 2017, por meio do porte do município conforme é apresentado na metodologia adotada deste trabalho em todo o território nacional.

Assim, os resultado obtidos demonstraram que houve um aumento da carência municipal quando confrontamos a quantidade de benefícios previdenciários emitidos com as arrecadados municipais durante o período de 2010 a 2017. Relacionado à isso, é identificado que alguns municípios apresentavam 6% de comprometimento do seu PIB o que indica um desequilíbrio fiscal entre de gastos e receitas.

Em adição, verificou-se que os municípios que possuíam população menor que 100 mil habitantes apresentam os maiores índices de comprometimento do PIB, com média 5,8 percentual de endividamento. Por outro lado, os benefícios previdenciários transferidos a essas unidades autônomas são responsáveis por manter-las com capital mínimo de giro, o qual garante a sobrevivência administrativa fiscal.

Assim, vinculado a redistribuição de renda ocasionada pelos benefícios emitidos, é importante ressaltar a permanência do aumento progressivo do salário mínimo, porque grande parte dos benefícios transferidos, como por exemplo: aposentadorias, pensões e auxílios, é realizada sob essa perspectiva salarial para a população brasileira de forma geral. O que pode ser considerado como um ponto positivo para manter essas economias ativas. Porém, isso não garante que a economia local seja autossustentável.

Dessa forma, a partir da primeira análise foi possível identificar que os municípios brasileiros apresentam uma grande dependência da redistribuição federal, para manter suas contas equilibras. Também, foi possível analisar que os municípios com população menor que 100 mil habitantes exibem maior sensibilidade a mudanças que a constituição previdenciária possa sofrer.

No segundo estudo de caso, é realizada uma análise econométrica via o modelo de estimação STVAR, o qual visa entender o comportamento do ciclo econômico fiscal local estudado, através de choques exógenos inferidos nos gastos públicos. Essa análise, considera a datação da dinâmica do ciclo econômico brasileiro entre 2002 e 2017, para que os resultados obtidos tenham maior consistência e apresentem dinâmica similar a do governo federal.

Assim, as estimativas obtidas das funções impulso-resposta via o modelo STVAR indicam diferenças consideráveis as respostas do produto e receita dependendo do tipo de regime econômico analisado. Assim, o multiplicador associado aos gastos agregados apresentam a maior diferença durante os estado de expansão. Por outro lado, as receitas em todos os estados do ciclo econômico exibem as menores respostas impulsivas encontradas.

Apesar disso, é importante destacar que as estimações não-lineares obtidas reportam resultados capazes de apresentar estados de expansão e recessão econômica separados, assim dando uma melhor avaliação do modelo. Também, vale ressaltar que os gastos agregados apresentam respostas médias e podem esconder de certa forma o real impacto dos multiplicadores de gasto. Porém, mesmo como essa curta disponibilidade de dados os resultas obtidos apresentam aceitáveis ao contexto.

Contudo, em ambos os estudos de casos os resultados apresentados indicam pontos relevantes que podem ser utilizados para o entendimento do ciclo econômico municipal, destacando que as transferências de gastos públicos em geral, as quais são responsáveis pela manutenção da atividade econômica local e mostram como esses influenciam externas alteram a dinâmica do ciclo econômico.

6.2 Trabalhos Futuros

Para trabalhos futuros, propõem-se a utilização de técnicas de inteligencia computacional para agrupar os municípios por características comuns e não só pelo porte de habitantes, como por exemplo: microrregião, tipo de área e região. Abaixo, são listados alguns possíveis trabalhos que envolvem dados no cenário econômico brasileiro:

- Utilizar técnicas de inteligência computacional para estimar o PIB dos municípios e/ou estados brasileiros que apresentam periodicidade mensal e/ou trimestral, a partir dos valores de ICMS estimados pelo Governo Federal.
- Realizar uma análise mais detalhada nos municípios que apresentaram o maior défice de endividamento do seu PIB.
- Comparar técnicas otimização matemática por exemplo MQO e variações com técnicas computacionais, como por exemplo: Redes Neurais (RN), Algoritmos Genéticos

(AG), Optimização por enxame de partículas (PSO), Algoritmo de Colônia Artificial de Abelhas (ABC) e *etc.*, para determinar os coeficientes matriciais de Π_E e Π_R do modelo STVAR.

- Utilizar mais multiplicadores fiscais de gastos para avaliar a dinâmica econômica municipal de forma mais detalhada, medindo o impacto que cada estado ou regime (expansão e recessão) apresenta ao longo prazo.

6.3 Dificuldades

Vale destacar que uma das principais dificuldades encontradas durante a realização deste trabalho, foi entorno da disponibilidade de dados abertos que apresentassem cortes geográficos locais e sub-globais acerca da economia brasileira em períodos longos. Além desta, também é pertinente a falta de dados locais que apresentem periodicidade mensal e/ou trimestral dos gastos, receitas e produto interno bruto (PIB) dos municípios brasileiros.

Outra dificuldade encontrada está relacionada aos poucos trabalhos da literatura que fazem uso de dados reais que utilizem modelo STVAR para realizar análises temporais a partir de dados municipais. Além disso, a escassez de trabalhos desta natureza limitam inferências e as comparações sob as análises de dados municipais brasileiros.

6.4 Publicações

Durante o processo de pesquisa acadêmica alguns trabalhos voltados a aplicação de *Data Science* sobre dados abertos foram desenvolvidos e publicados. Entre estes trabalhos obtivemos a seguinte publicação do estudo de caso I desta dissertação:

- **Santos**, S. M.; Felix 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 entre 2010 a 2017. In: VIII CONINTER - Congresso Internacional Interdisciplinar em Sociais e Humanidades. Maceió. 2019.

6.5 Publicações Adicionais

Além do trabalho mencionado acima também obtemos algumas publicações adicionais de diferentes contextos de análise de dados, como por exemplo:

- Felix Junior, F. E. A.; Santos, S. M.; Frances, C. R. L.; Silva, M. S. Análise da Participação das Aposentadorias e Pensões na Distribuição de Renda Per Capita do

Brasil - 2000 e 2010. In: VIII CONINTER - Congresso Internacional Interdisciplinar em Sociais e Humanidades. Maceió. 2019.

- Felix Junior F. E. A.; Pereira, A. A. S.; Santos, S. M.; Silva, M. S. Impactos da Reforma Previdenciária: Um Estudo Acerca das Concessões de Benefícios,. In: VIII CONINTER - Congresso Internacional Interdisciplinar em Sociais e Humanidades. Maceió, 2019.

Referências

AGUIRRE, L. *Introdução à Identificação de Sistemas – Técnicas Lineares e Não-Lineares Aplicadas a Sistemas Reais*. Editora UFMG, 2007. ISBN 9788570415844. Disponível em: <https://books.google.com.br/books?id=f9IwE7Ph0fYC>. Citado 3 vezes nas páginas 13, 14 e 17.

AITKEN, A. C. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh*, Royal Society of Edinburgh Scotland Foundation, v. 55, p. 42–48, 1936. Citado 2 vezes nas páginas 15 e 16.

ALVES, R. S. *O impacto da Política Fiscal Sobre a Atividade Econômica ao Longo do Ciclo Econômico: Evidências Para o Brasil*. 2017. Dissertação, USP - (Universidade de São Paulo), São Paulo, Brasil. Citado na página 16.

AUERBACH, A. J.; GORODNICHENKO, Y. Fiscal multipliers in recession and expansion. National Bureau of Economic Research, Inc, p. 63–98, September 2012. Disponível em: <https://ideas.repec.org/h/nbr/nberch/12634.html>. Citado 5 vezes nas páginas 16, 20, 23, 37 e 38.

AUERBACH, A. J.; GORODNICHENKO, Y. Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, v. 4, n. 2, p. 1–27, May 2012. Disponível em: <https://ideas.repec.org/a/aea/aejpol/v4y2012i2p1-27.html>. Citado 3 vezes nas páginas 16, 17 e 37.

BAERLOCHER, D.; PARENTE, S. L.; RIOS-NETO, E. Economic effects of demographic dividend in brazilian regions. *The Journal of the Economics of Ageing*, v. 14, p. 100198, 2019. ISSN 2212-828X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2212828X18301075>. Citado na página 18.

BARRO, R. J. On the determination of the public debt. *Journal of Political Economy*, v. 87, n. 5, Part 1, p. 940–971, 1979. Disponível em: <https://doi.org/10.1086/260807>. Citado na página 4.

BATTISTI, I. D. E.; BATTIST, G. *Métodos Estatísticos*. Ijuí, Rio Grande do Sul, Brasil: Unijuí, 2008. ISBN 978-85-7269-442-1. Citado na página 13.

BEZERRA, J. F.; SILVA, I. Ézio M.; LIMA, R. C. Os efeitos da política monetária sobre o produto no brasil: evidência empírica usando restrição de sinais. *Revista de Economia Contemporânea*, scielo, v. 18, p. 296 – 316, 08 2014. ISSN 1415-9848. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-98482014000200296&nrm=iso>. Citado na página 23.

BOSCHETTI, A.; MASSARON, L. *Python Data Science Essentials - Learn the fundamentals of Data Science with Python*. Birmingham: Packt Publishing, 2015. ISBN 978-1-78528-042-9. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=e880b26d877ff1fa8f20a6f0bb5b89bd>. Citado na página 8.

BRASIL. *Estatísticas de Finanças Públicas eConta Intermediária de Governo*. [S.l.]: Nacionais Contas, 2016. ISBN 9788524043574. Citado na página 5.

BRASIL. Lei de acesso a informação - lai. 2019. Acessado em: 11 de Novembro 2019. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Citado na página 2.

BRASIL. Proposta de emenda à constituição n^o 6, de 2019 - reforma da previdência. 2019. Acessado em: 02 de Dezembro 2020. Disponível em: <https://www25.senado.leg.br/web/atividade/materias/-/materia/137999>. Citado na página 6.

BRITO, A. S.; KERSTENETZKY, C. L. Has the minimum wage policy been important for reducing poverty in brazil? a decomposition analysis for the period from 2002 to 2013. *EconomiA*, v. 20, n. 1, p. 27 – 43, 2019. ISSN 1517-7580. Disponível em: <http://www.sciencedirect.com/science/article/pii/S151775801830081X>. Citado na página 21.

CADY, F. *The Data Science Handbook*. Wiley, 2017. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=b9f1781011afa3e617f771337919156d>. Citado na página 8.

CERQUEIRA, V. dos S.; RIBEIRO, M. B.; MARTINEZ, T. S. Propagação assimétrica de choques monetários na economia brasileira:evidências com base em um modelo vetorial não-linear de transição suave. *Revista Brasileira de Economia*, scielo, v. 68, p. 19 – 47, 03 2014. ISSN 0034-7140. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-71402014000100003&nrm=iso>. Citado na página 20.

CHIGNARD, S. *A brief history of Open Data*. 2013. Disponível em: <http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data>. Acessado em: 10 de Novembro 2019. Citado na página 2.

COELHO, A. A. R. *Identificação de Sistemas Dinâmicos Lineares*. 2. ed. [S.l.]: Editora UFSC, 2016. ISBN 8532807305, 978-8532807304. Citado 2 vezes nas páginas 15 e 17.

COELHO, A. A. R.; COELHO, L. dos S. *Identificação de Sistemas Dinâmicos Lineares*. [S.l.]: Editora UFSC, 2004. ISBN 85.328.0280-X. Citado 2 vezes nas páginas 15 e 17.

CONTINUUM ANALYTICS. *conda Documentation, Release 4.8.1.post748+271ca227*. [s.n.], 2020. Disponível em: <https://readthedocs.com/projects/continuumio-conda/downloads/pdf/latest/>. Citado na página 22.

CRUZ, L. C. da. *Data Science: Desenvolvimento de aplicação para análise de dados*. São Paulo, Brasil: [s.n.], 2018. Monografia (Ciência da Computação), FEMA (Fundação Educacional do Município de Assis). Citado na página 2.

CURTY, R. G.; CERVANTES, B. M. N. Data science: Ciência orientada a dados. *Informação & Informação*, v. 21, n. 2, p. 1, 2016. Citado na página 2.

DANTAS, A. D. O. da S. *Identificação de Modelos Polinomiais NARX Utilizando Algoritmos Combinados de Detecção de Estrutura e Estimação de Parâmetros com Aplicações Práticas*. 2013. Dissertação, UFJF (Universidade Federal do Rio Grande do Norte), Natal, Brasil. Citado 3 vezes nas páginas 13, 14 e 15.

DEVORE, J. L. *Probabilidade e Estatística para Engenharia e Ciências*. 8. ed. [S.l.]: Cengage Learning, 2015. ISBN 978-85-221-1183-1. Citado 2 vezes nas páginas 12 e 17.

DUTRA, F. N. *Multiplicadores Fiscais no Brasil: Estimativa A partir de modelos STVAR*. 2018. Monografia (Bacharel em Ciências Econômicas), UFGRS - (Universidade Federal do Rio Grande Sul), Porto Alegre, Brasil. Citado na página 16.

ELMASRI, R.; NAVATHE, S. B. *Sistemas de Banco de Dados*. 4. ed. Addison Wesley, 2005. ISBN 9788588639171. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=EE3B00130FFFB1D0F5774913401461F1>. Citado 3 vezes nas páginas 9, 12 e 17.

ESPÍNDOLA, P. L. et al. Governança de dados aplicada à ciência da informação: análise de um sistema de dados científicos para a área da saúde. *RDBCI: Revista Digital de Biblioteconomia e Ciência da Informação*, v. 16, n. 3, p. 274–298, 2018. Citado na página 1.

FIRJAN. *IFDM 2018 Índice Firjan Desenvolvimento Municipal*. IFGF, 2018. Disponível em: <https://www.firjan.com.br/ifgf/consulta-ao-indice/>. Citado na página 4.

FLORIDI, L. *Information: A Very Short Introduction*. Oxford University Press, USA, 2010. ISBN 0199551375,9780199551378. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=B30B6B125B4E8662742B466161F6B567>. Citado 2 vezes nas páginas 1 e 3.

FRANK, E.; WITTEN, I. *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations*. [s.n.], 2000. Chapter 8. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=787a145b668449b75f4e6f38d13ad06e>. Citado 3 vezes nas páginas 8, 9 e 11.

GANTZ, J.; REINSEL, D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *EMC*, 2012. Citado na página 1.

GIUBERTI, A. C. Lei de responsabilidade fiscal: Efeitos sobre o gasto com pessoal dos municípios brasileiros. XXXIII ENCONTRO NACIONAL DE ECONOMIA ANPEC, v. 33, 2005. Disponível em: <http://www.anpec.org.br/encontro2005/artigos/A05A048.pdf>. Citado 2 vezes nas páginas 4 e 5.

GOLDSCHMIDT, R.; PASSOS, E. *Data Mining: um Guia Prático*. 1. ed. Campus, 2005. ISBN 9788535218770. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=a3b3949d218211d74a13a153fa77928c>. Citado 2 vezes nas páginas 9 e 10.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining: conceito, técnicas, algoritmos orientações e aplicações*. 2. ed. Brasil: GEN LTC, 2015. ISBN 978-8535278224. Citado 4 vezes nas páginas 8, 10, 11 e 18.

GRUDTNER, V.; ARAGON, E. K. D. S. B. Multiplicador dos gastos do governo em períodos de expansão e recessão: Evidências empíricas para o brasil. *Revista Brasileira de Economia*, scielo, v. 71, p. 321 – 345, 09 2017. ISSN 0034-7140. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0034-71402017000300321&nrm=iso>. Citado na página 20.

GUEDES, M. *No final das contas: o que é o Docker e como ele funciona?* [s.n.], 2018. Disponível em: <https://www.treinaweb.com.br/blog/no-final-das-contas-o-que-e-o-docker-e-como-ele-funciona/>. Citado na página 23.

GUJARATI, D. *Econometria Básica*. 5. ed. Mc Graw Hill, 2011. ISBN 8563308327,9788563308320. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=74d18c9da279057ed7d6da0fa4cce3b8>. Citado na página 13.

HAND, D. J.; MANNILA, H.; SMYTH, P. *Principles of data mining*. MIT Press, 2001. (Adaptive computation and machine learning). ISBN 9781423731320. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=0A2AB73C0E70AC3ED2F74F59CB348D68>. Citado na página 11.

HUNTER, J. et al. *Matplotlib, Release 3.1.1*. [S.l.: s.n.], 2019. Citado na página 23.

IBGE. Instituto Brasileiro de Geografia e Estatística. Censo Demográfico. Tabela 1.8 - População nos Censos Demográficos, segundo as Grandes Regiões, as Unidades da Federação e a situação do domicílio - 1960/2010. IBGE, 08 2010. Disponível em: <https://www.overleaf.com/project/5dc6ec3652bc020001f402f1>. Citado na página 34.

ISOTANI, S.; BITTENCOURT, I. *Dados Abertos Conectados: em Busca da Web do Conhecimento*. [S.l.: s.n.], 2015. ISBN 978-85-7522-449-6. Citado na página 1.

KREYENFELD, M.; WILLEKENS, F. Data bases and statistical systems: Demography. In: WRIGHT, J. D. (Ed.). *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Second edition. Oxford: Elsevier, 2015. p. 735 – 741. ISBN 978-0-08-097087-5. Disponível em: <http://www.sciencedirect.com/science/article/pii/B9780080970868410135>. Citado na página 19.

LIMA, T. G. et al. Kdd processes in non-relational data: The case of the mineramongo tool. In: *2017 12th Iberian Conference on Information Systems and Technologies (CISTI)*. [S.l.: s.n.], 2017. p. 1–6. ISSN null. Citado na página 18.

MAGALHÃES, L. F. A. Fontes de dados demográficos e estudos de população em santa catarina. NECAT, v. 4, p. 23 – 37, 06 2015. ISSN 2317-8523. Disponível em: <http://incubadora.periodicos.ufsc.br/index.php/necat/article/view/3624>. Citado na página 21.

MCKINNEY, W. *pandas: powerful Python data analysistoolkit, Release 0.25.3*. [S.l.: s.n.], 2019. Citado na página 22.

MEDEIROS, C. A. de. *Estatística Aplicada à educação*. Brasília: Universidade de Brasília, 2007. ISBN 978-85-230-0990-8. Citado na página 12.

MEDEIROS, K. R. d. et al. Lei de responsabilidade fiscal e as despesas com pessoal da saúde: uma análise da condição dos municípios brasileiros no período de 2004 a 2009. *Ciência & Saúde Coletiva*, scielo, v. 22, p. 1759 – 1769, 06 2017. ISSN 1413-8123. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232017002601759&nrm=iso>. Citado na página 4.

MINISTÉRIO DA FAZENDA. *Evolução da proteção previdenciária no Brasil - 2017. informe de previdência social*. [S.l.: s.n.], 2018. Citado na página 5.

MUELLER, L. M. J. P. *Python for Data Science, 2nd Edition*. 2. ed. Wiley, 2019. (For Dummies). ISBN 9781119547624, 9781119547662. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=77e0c44a6390aa0c01c1159861eec3d8>. Citado 5 vezes nas páginas 1, 2, 3, 17 e 18.

OBAMA, B. *Transparency and Open Government*. 2008. Washington, United States of America: The White House. Citado na página 2.

OLAVSRUD, T. Afinal, o que é ciência de dados e o que isso tem a ver com as profissões do futuro? 2018. Disponível em: <https://itmidia.com/afinal-o-que-e-ciencia-de-dados-e-o-que-isso-tem-a-ver-com-as-profissoes-do-futuro/>. Citado na página 9.

OLIVEIRA, P. F. d.; GUERRA, S.; MCDONNELL, R. *Ciência de Dados com R: Introdução*. Brasil: IBDPA, 2018. ISBN 978-85-54230-00-5. Citado 2 vezes nas páginas 8 e 17.

OLIVEIRA, V. K. de. *Multiplicadores Fiscais de Gastos e Tributos: Um abordagem DSGE para Economia Brasileira*. 2018. Dissertação, USP - (Universidade de São Paulo), São Paulo, Brazil. Citado na página 16.

OPEN DATA COMMONS. *Open Data Commons*. 2019. Disponível em: <https://opendatacommons.org/licenses/>. Acessado em: 10 de Novembro 2019. Citado na página 3.

OPEN DATA HANDBOOK. *The Open Data Handbook*. 2019. Disponível em: <http://opendatahandbook.org/guide/en/>. Acessado em: 10 de Novembro 2019. Citado na página 3.

ORAIR, R. O.; SIQUEIRA, F. D. F.; GOBETTI, S. W. *Política Fiscal e Ciclo Econômico: uma análise baseada em multiplicadores do gasto público*. 2016. Monografia , XXI Prêmio Tesouro Nacional. Citado 6 vezes nas páginas 16, 17, 19, 20, 26 e 38.

ORAIR, R. O.; SIQUEIRA, F. de F. Investimento público no brasil e suas relações com ciclo econômico e regime fiscal. *Economia e Sociedade*, scielo, v. 27, p. 939 – 969, 12 2018. ISSN 0104-0618. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-06182018000300939&nrm=iso>. Citado na página 20.

PROJECT JUPYTER. *JupyterLab Documentation, Release 1.2.4*. [S.l.: s.n.], 2019. Citado na página 23.

QIANG, Z. et al. Research on the course system of data science and engineering major. In: *2019 IEEE International Conference on Computer Science and Educational Informatization (CSEI)*. [S.l.: s.n.], 2019. p. 90–93. ISSN null. Citado na página 18.

REIS, E. et al. *Estatística Aplicada*. 6. ed. Manchester: Edições Sílabo, 2007. v. 1. ISBN 978-972-618-819-3. Citado na página 13.

REITZ, K. *Python Guide Documentation, Release 0.0.1*. [s.n.], 2018. Disponível em: <https://buildmedia.readthedocs.org/media/pdf/python-guide/latest/python-guide.pdf>. Citado na página 22.

SAKURAI, S. N. Superávit e déficit fiscal dos municípios brasileiros: uma aplicação do modelo de viés de seleção em painel. *Nova Economia*, scielo, v. 24, p. 517 – 540, 12 2014. ISSN 0103-6351. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-63512014000300517&nrm=iso>. Citado na página 4.

SANTOS, A. J. dos. *Modelos Vetoriais Auto-Regressivos com Transição Suave Estruturados por Árvores – STVAR-Tree*. 2009. Dissertação (Mestrado), PUC - (Pontifícia Universidade Católica do Rio de Janeiro), Rio de Janeiro, Brasil. Citado na página 16.

SANTOS, C. M. L. da Silva Afonso dos. *Estatística Descritiva - Manual de Auto-Aprendizagem*. 3. ed. [S.l.]: Edições Sílabo, 2018. ISBN 9789726189688. Citado na página 13.

SIBILLE, A.; OESTGES, C.; ZANELLA, A. *MIMO: From Theory to Implementation*. 1. ed. Academic Press, 2010. ISBN 0123821940,9780123821942. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=F3CC360C32C7B1AFE931F93D245FB4CC>. Citado na página 15.

SILVA, D. A. S. et al. Associação do sobrepeso com variáveis sócio-demográficas e estilo de vida em universitários. *Ciência & Saúde Coletiva*, scielo, v. 16, p. 4473 – 4479, 11 2011. ISSN 1413-8123. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232011001200020&nrm=iso>. Citado na página 21.

STEELE, N. I. J. *Beautiful Visualization: Looking at Data through the Eyes of Experts*. 1. ed. O'Reilly Media, 2010. (Theory in Practice). ISBN 1449379869,9781449379865. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=a328e65a24a0dc295e4b0bdd71706394>. Citado na página 12.

SÓLON, A. de F. et al. *A Previdência Social e a Economia dos Municípios*. Brasília: ANFIP, 2019. ISBN 978-85-62102-32-5. Citado 6 vezes nas páginas 5, 6, 18, 28, 31 e 32.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining. Mineração de Dados*. 1. ed. [S.l.]: Ciência Moderna, 2009. ISBN 978-8573937619. Citado 3 vezes nas páginas 11, 18 e 26.

THOMAZINI, B. S. Federalismo brasileiro: Origem e evolução histórica de seus reflexos na atualidade. 2020. Disponível em: <https://ambitojuridico.com.br/cadernos/direito-constitucional/federalismo-brasileiro-origem-e-evolucao-historica-de-seus-reflexos-na-atualidade/>. Citado na página 5.

VANDERPLAS, J. *Python Data Science Handbook: Essential Tools for Working with Data*. [S.l.]: O'Reilly Media, Inc., 2016. ISBN 978-1-491-91205-8. Citado 2 vezes nas páginas 2 e 8.

VIOLINO, B. *8 habilidades essenciais para cientistas de dados de alto desempenho*. 2018. <https://cio.com.br/retrospectiva-2018-10-artigos-mais-lidos-sobre-ciencia-de-dados/>. Citado na página 2.

WOOLDRIDGE, J. *Introdução à econometria: Uma abordagem moderna*. 3. ed. [S.l.]: Cengage Learning, 2017. ISBN 8522125643,9788522125647. Citado 2 vezes nas páginas 13 e 15.

WOOLDRIDGE, J. M. *Introdução à Econometria. Uma Abordagem Moderna*. 1. ed. São Paulo: Cengage CTP, 2010. ISBN 978-8522104468. Citado na página 13.