UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Estudo do impacto econômico da pandemia de COVID-19 através de dados de Documentos Fiscais

Leonardo Alves Miguel

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

ICMC USP
SÃO CARLOS

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Leonardo Alves Miguel

Estudo do impacto econômico da pandemia de COVID-19 através de dados de Documentos Fiscais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria. *EXEMPLAR DE DEFESA*

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
Janeiro de 2021

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A474e

Alves Miguel, Leonardo
Estudo do impacto econômico da pandemia de COVID-19 através de dados de Documentos Fiscais / Leonardo Alves Miguel; orientador Francisco Aparecido Rodrigues. -- São Carlos, 2021.
134 p.

Dissertação (Mestrado - Programa de Pós-Graduação em Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2021.

1. Análise Exploratória. 2. Redes Complexas. 3. Crise econômica. 4. Aprendizado de Máquina. I. Aparecido Rodrigues, Francisco, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Leonardo Alves Miguel

Study of economic impact of the COVID-19 pandemic using Invoices data

Dissertation submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Professional Master's Program in Mathematics Statistics and Computing Applied to Industry, for the degree of Master in Science. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Francisco Aparecido Rodrigues

USP – São Carlos
January 2021

Dedico esse trabalho a todos aqueles que fazem pesquisa no Brasil, em um momento paradoxal em que precisamos da ciência e ela é cada vez mais atacada e questionada. Para que os pesquisadores brasileiros não desanimem e continuem trabalhando em prol da divulgação científica.

Dedico também a minha família, em especial aos meus pais que tudo fizeram por mim, à melhor irmã do mundo, e às minhas avós.

Por fim, dedico também a todas as famílias daqueles que se foram desnecessariamente durante esta terrível pandemia, para que encontrem conforto.

AGRADECIMENTOS

Agradeço primeiramente à Arquivei, pela cessão dos dados e colaboração através do fornecimento de infraestrutura para a realização deste trabalho. Também por me dar a oportunidade de crescer profissionalmente no dia-a-dia e em trabalhos como este. Em especial, agradeço a Bruno Oliveira e Bruno Calza pela colaboração.

Agradeço também a meu orientador, Francisco, pela ajuda e paciência em vários momentos, também motivando quando necessário, e tendo sido com certeza um dos melhores professores com quem tive a honra e o prazer de estudar.

Agradeço também ao amigo Marcus Araújo, pela colaboração com este trabalho através ideias, discussões, e ajudando na revisão. Obrigado por compartilhar das dores, desânimos, revoltas, e indecisões.

Agradeço a todos os amigos que me ajudaram a procrastinar mas no fim entregar este trabalho. Que em breve possamos presencialmente comemorar este trabalho juntos e pagar a caixa que desde já lhes devo.

Agradeço à comunidade *open source* pelo fornecimento de diversas bibliotecas e ferramentas utilizadas na execução deste trabalho, vindas de contribuições muitas vezes voluntárias e não remuneradas.

Agradeço também à Universidade de São Paulo pela oportunidade e pelo fornecimento de recursos para a realização deste trabalho.

"O que podia ser sempre vai ser melhor que é, as possibilidades vão perseguir o que a gente escolher"

(Banda Tópaz)

RESUMO

MIGUEL, L. A. **Estudo do impacto econômico da pandemia de COVID-19 através de dados de Documentos Fiscais**. 2021. 134 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

A crise mais grave em pelo menos um século trouxe uma série de consequências para a saúde, política, e economia brasileira. O termo "novo normal" é cada vez mais comum no vocabulário das pessoas, e os efeitos desse evento serão sentidos ainda por muitos anos. Neste trabalho, utilizamos dados adquiridos através do processamento de documentos fiscais para analisar o impacto econômico da crise causada pela pandemia de COVID-19 em uma amostra da economia brasileira, utilizando uma modelagem em Redes Complexas e Aprendizado de Máquina. O objetivo é utilizar dados de empresas, utilizados burocraticamente para cumprir obrigações fiscais e contábeis, para tirar conclusões sobre este evento.

Palavras-chave: Análise Exploratória, Redes Complexas, Crise Econômica, Aprendizado de Máquina, COVID-19.

ABSTRACT

MIGUEL, L. A. **Study of economic impact of the COVID-19 pandemic using Invoices data**. 2021. 134 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2021.

The most severe crisis in at least a century brought a series of consequences to brazilian health systems, politics, and economy. The expression "new normal" is increasingly common in the people's mouth and the effects of this event will be felt for many years to come. In this work, we used data acquired through processing invoices to analyze the economic impact of the COVID-19 pandemic in a sample of the brazilian economy using Complex Networks and Machine Learning concepts. We intend to use enterprise data usually processed for accounting and tax obligations to get useful insights about this event.

Keywords: Exploratory Analysis, Complex Networks, Economic Crisis, Machine Learning, COVID-19.

LISTA DE ILUSTRAÇÕES

Figura 1 – Gráfico do índice de isolamento social no Brasil 32
Figura 2 – Exemplo de DANFE através do Portal Arquivei 39
Figura 3 – Exemplo de grafo acíclico direcionado não-ponderado 42
Figura 4 – Exemplo de Curva ROC 49
Figura 5 – O dilema viés-variância 50
Figura 6 – Participação por região das empresas presentes na base de dados 60
Figura 7 – Participação por UF das empresas presentes na base de dados 61
Figura 8 – Histograma do percentual das empresas de cada município brasileiro presente na base de dados 62
Figura 9 – Participação por porte das empresas presentes na base de dados 64
Figura 10 – Participação por opção pelo MEI das empresas presentes na base de dados . 66
Figura 11 – Participação por região das empresas presentes na base de dados 67
Figura 12 – Participação por seção do CNAE das empresas presentes na base de dados . 69
Figura 13 – Histograma do percentual das empresas de cada CNAE presente na base de dados 70
Figura 14 – Valor total transacionado por mês no período analisado 73
Figura 15 – Comparação do valor total transacionado por mês entre 2019 e 2020 (Parte 4) 74
Figura 16 – Valor total transacionado por mês comparado com a previsão para o período analisado 75
Figura 17 – Valor total transacionado por região no período analisado 77
Figura 18 – Valor trimestral transacionado por região no período analisado 77
Figura 19 – Comparação do valor mensal transacionado por região entre 2019 e 2020 . . 79
Figura 20 – Valor total transacionado por UF no período analisado 80
Figura 21 – Valor trimestral transacionado por UF no período analisado 80
Figura 22 – Valor total transacionado por seção no período analisado 83
Figura 23 – Valor trimestral transacionado por seção no período analisado 84
Figura 24 – Histograma do valor total transacionado por CNAE no período analisado . . 87
Figura 25 – Diagrama de caixa dos valores mensais transacionados por CNAE no período analisado 87
Figura 26 – Histograma do grau dos vértices da rede CNAE-CNAE 94
Figura 27 – Percentual de variância explicada de cada componente principal para dados de UFs 106
Figura 28 – Visualização dos dados de UFs projetados sobre as duas componentes principais 107

Figura 29 – Matrizes de confusão para classificadores aplicados sobre dados de UFs para impacto mensal 108
Figura 30 – Matrizes de confusão para classificadores aplicados sobre dados de UFs para impacto total (Parte 1) 108
Figura 31 – Matrizes de confusão para classificadores aplicados sobre dados de UFs para impacto total (Parte 2) 109
Figura 32 – Percentual de variância explicada de cada componente principal para dados de seções 109
Figura 33 – Visualização dos dados de seções projetados sobre as duas componentes principais 110
Figura 34 – Matrizes de confusão para classificadores aplicados sobre dados de seções para impacto mensal 111
Figura 35 – Matrizes de confusão para classificadores aplicados sobre dados de seções para impacto total 112
Figura 36 – Percentual de variância explicada de cada componente principal para dados de CNAEs 113
Figura 37 – Visualização dos dados de CNAEs projetados sobre as duas componentes principais 114
Figura 38 – Resultado das tentativas de classificação para cada método aplicado sobre dados de CNAEs para impacto mensal 116
Figura 39 – Resultado das tentativas de classificação para cada método aplicado sobre dados de CNAEs para impacto total 117
Figura 40 – Interface do Google BigQuery para seleção de dados 132

LISTA DE QUADROS

Quadro 1 – Matriz de confusão 48
Quadro 2 – Campos de documentos fiscais considerados 56
Quadro 3 – Campos do Cadastro Nacional de Pessoas Jurídicas considerados 58
Quadro 4 – Códigos de Região segundo DTB 59
Quadro 5 – Denominação das Seções de CNAE 68

LISTA DE CÓDIGOS-FONTE

Código-fonte 1 – Exemplo de fragmento de NFe com dados sensíveis removidos . . . 133

LISTA DE TABELAS

Tabela 1 – Percentual das empresas presentes na base de dados por Região 60
Tabela 2 – Percentual das empresas presentes na base de dados por UF 62
Tabela 3 – Métricas sobre o percentual das empresas presentes na base de dados por município 63
Tabela 4 – Métricas sobre o percentual das empresas presentes na base de dados por porte 64
Tabela 5 – Métricas sobre o percentual das empresas presentes na base de dados pela opção pelo MEI 66
Tabela 6 – Métricas sobre o percentual das empresas presentes na base de dados pela Opção pelo Simples Nacional 66
Tabela 7 – Métricas sobre o percentual das empresas presentes na base de dados por Seção do CNAE 70
Tabela 8 – Métricas sobre o percentual das empresas presentes na base de dados por CNAE 71
Tabela 9 – Variação entre o valor transacionado para cada mês em 2020 em relação a 2019 74
Tabela 10 – Variação entre o valor transacionado para cada mês em relação ao previsto após o início da pandemia 76
Tabela 11 – Diferença entre valores transacionados previstos e obtidos por região no período da pandemia 78
Tabela 12 – Diferença entre valores mensais transacionados previstos e obtidos por UF no período da pandemia 81
Tabela 13 – Diferença entre valores totais transacionados previstos e obtidos por UF no período da pandemia 82
Tabela 14 – Quantidade de UFs impactadas negativamente pela pandemia em relação às variações mensais ou totais 83
Tabela 15 – Diferença entre valores mensais transacionados previstos e obtidos por seção no período da pandemia 85
Tabela 16 – Diferença entre valores totais transacionados previstos e obtidos por seção no período da pandemia 86
Tabela 17 – Quantidade de seções impactadas negativamente pela pandemia em relação às variações mensais ou totais 86
Tabela 18 – Quantidade de CNAEs impactadas negativamente pela pandemia em relação às variações mensais ou totais 88
Tabela 19 – Métricas do grafo de transações entre UFs 89

Tabela 20 – Métricas do grafo de transações entre UFs segmentadas por UF 90
Tabela 21 – Métricas do grafo de transações entre seções 91
Tabela 22 – Métricas do grafo de transações entre seções segmentadas por seção (Parte 1) 92
Tabela 23 – Métricas do grafo de transações entre seções segmentadas por seção (Parte 2) 93
Tabela 24 – Métricas do grafo de transações entre CNAEs 94
Tabela 25 – Métricas do grafo de transações entre CNAEs segmentadas por CNAE (Parte 1) 95
Tabela 26 – Métricas do grafo de transações entre CNAEs segmentadas por CNAE (Parte 2) 95
Tabela 27 – Variação percentual das métricas de redes mês a mês para o grafo de UFs (Parte 1) 96
Tabela 28 – Variação das métricas de redes mês a mês para o grafo de UFs (Parte 2) . . . 97
Tabela 29 – Variação das métricas de rede a cada trimestre para o grafo de UFs (Parte 1) 98
Tabela 30 – Variação das métricas de rede a cada trimestre para o grafo de UFs (Parte 2) 98
Tabela 31 – Variação percentual das métricas de redes mês a mês para o grafo de seções (Parte 1) 100
Tabela 32 – Variação das métricas de redes mês a mês para o grafo de seções (Parte 2) . 101
Tabela 33 – Variação das métricas de rede a cada trimestre para o grafo de seções (Parte 1) 101
Tabela 34 – Variação das métricas de rede a cada trimestre para o grafo de seções (Parte 2) 102
Tabela 35 – Variação percentual das métricas de redes mês a mês para o grafo de CNAEs (Parte 1) 102
Tabela 36 – Variação das métricas de redes mês a mês para o grafo de CNAEs (Parte 2) . 103
Tabela 37 – Variação das métricas de rede a cada trimestre para o grafo de CNAEs (Parte 1) 104
Tabela 38 – Variação das métricas de rede a cada trimestre para o grafo de CNAEs (Parte 2) 104
Tabela 39 – Direção de cada métrica em relação a cada uma das duas componentes principais 106
Tabela 40 – Direção de cada métrica em relação a cada uma das duas componentes principais para os dados de seções 110
Tabela 41 – Direção de cada métrica em relação a cada uma das duas componentes principais para os dados de CNAEs 113

LISTA DE ABREVIATURAS E SIGLAS

Anvisa	Agência Nacional de Vigilância Sanitária
CFOP	Código Fiscal de Operações e Prestações
CNAE	Classificação Nacional de Atividades Econômicas
CNPJ	Cadastro Nacional da Pessoa Jurídica
CONCLA	Comissão Nacional de Classificação
CTe	Conhecimento de Transporte Eletrônico
DACTE	Documento Auxiliar do Conhecimento de Transporte Eletrônico
DANFE	Documento Auxiliar da Nota Fiscal Eletrônica
DFe	Documento Fiscal Eletrônico
DTB	Divisão Territorial Brasileira
ENAT	Encontro Nacional de Administradores Tributários
EPP	Empresa de Pequeno Porte
IBGE	Instituto Brasil de Geografica e Estatística
ICMS	Imposto Sobre Operações Relativas à Circulação de Mercadorias e Prestações de Serviços de Transporte Interestadual e Intermunicipal e de Comunicação
ISS	Imposto Sobre Serviços de Qualquer Natureza
MDFe	Manifesto Econômico de Documentos Fiscais
ME	Microempresa
MEI	Microempreendedor Individual
NCM	Nomenclatura Comum do Mercosul
NFe	Nota Fiscal Eletrônica
NFSe	Nota Fiscal de Serviço Eletrônica
OLAP	Online Analytical Processing
PIB	Produto Interno Bruto
Sefaz	Secretaria de Estado da Fazenda
SPED	Sistema Público de Escrituração Digital
SQL	*Structured Query Language*
TED	Transmissão Eletrônica de Documentos
UF	Unidade Federativa
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO . . . 29
1.1 Documentos Fiscais Eletrônicos . . . 29
1.2 A Pandemia de COVID-19 . . . 30
1.3 Crise econômica . . . 33
1.4 Objetivos . . . 34

2 DOCUMENTOS FISCAIS . . . 35
2.1 Legislação Brasileira . . . 35
2.1.1 ENAT . . . 36
2.1.2 Documentos Fiscais Eletrônicos . . . 38
2.2 A Arquivei . . . 39
2.2.1 Segurança dos dados . . . 40

3 METODOLOGIA . . . 41
3.1 Redes Complexas . . . 41
3.1.1 Conceitos . . . 41
3.1.2 Representação . . . 42
3.1.3 Métricas . . . 42
3.1.3.1 Grau . . . 43
3.1.3.2 Caminhos mínimos . . . 43
3.1.3.3 Excentricidade . . . 44
3.1.3.4 Transitividade . . . 44
3.1.3.5 Assortatividade . . . 44
3.1.3.6 Centralidade . . . 45
3.2 Análise de Componentes Principais . . . 47
3.3 Classificadores . . . 47
3.3.1 Conceitos . . . 47
3.3.2 K-Nearest Neighbors . . . 51
3.3.3 Random Forest . . . 51
3.3.4 Regressão Logística . . . 52
3.3.5 SVC . . . 52

4 BASE DE DADOS . . . 55
4.1 Dados de Documentos Fiscais . . . 55

4.1.1 Aquisição *55*
4.1.2 Pré-processamento *56*
4.2 Dados Abertos de Empresas Brasileiras 57
4.2.1 Aquisição *57*
4.3 Análise Descritiva da Base de Dados 58
4.3.1 Abrangência territorial *58*
4.3.2 Abrangência econômica *62*
4.3.3 Abrangência setorial *67*
4.4 Conclusões 70

5 IMPACTO ECONÔMICO DA PANDEMIA 73
5.1 Cenário geral 73
5.2 Cenário regional 76
5.3 Cenário econômico 82
5.4 Conclusões 85

6 RESULTADOS 89
6.1 Caracterização de redes 89
6.1.1 Redes UF-UF *89*
6.1.2 Redes Seção-Seção *91*
6.1.3 Redes CNAE-CNAE *93*
6.1.4 Conclusões *96*
6.2 Alterações topológicas 96
6.2.1 Redes UF-UF *96*
6.2.2 Redes Seção-Seção *99*
6.2.3 Redes CNAE-CNAE *102*
6.2.4 Conclusões *105*
6.3 Detecção de impacto econômico 105
6.3.1 Dados de UFs *105*
6.3.2 Dados de Seções *109*
6.3.3 Dados de CNAEs *113*
6.3.4 Conclusões *115*

7 CONCLUSÃO 119
7.1 Pontos positivos e negativos 119
7.2 Próximos passos 120

REFERÊNCIAS 121

GLOSSÁRIO 127

APÊNDICE A ARMAZÉM DE DADOS 129
A.1 Arquitetura de Ingestão 129
A.2 Plataforma de DFe 130
A.3 Armazém de Dados 130

APÊNDICE B TECNOLOGIAS UTILIZADAS 131
B.1 Google BigQuery 131
B.2 Google AI Platform Notebooks 131
B.3 Bibliotecas 131

APÊNDICE C EXEMPLO DE NOTA FISCAL ELETRÔNICA 133

CAPÍTULO

1

INTRODUÇÃO

1.1 Documentos Fiscais Eletrônicos

Segundo dados da Secretaria Especial da Receita Federal do Brasil (2020), o Brasil possuía em setembro de 2020 mais de 45 milhões de empresas ativas que realizam transações todos os dias, envolvendo aquisição e transferência de mercadorias, prestação de serviços das mais diversas naturezas, transportes e devoluções. Essas empresas possuem uma série de obrigações contábeis e fiscais a cumprir todos os meses. Obrigações essas que se traduzem em documentos enviados ao governo contendo uma série de informações sobre cada uma dessas transações efetuadas.

A lei que obriga a emissão de notas fiscais é datada do ano de 1994 (BRASIL, 1994). À época, a emissão desses documentos era feita em papel, o que causava grandes transtornos às empresas tanto por conta dos próprios processos envolvidos quanto por conta da guarda desses documentos, que a depender do porte e do volume de transações das empresas exigia um grande esforço de armazenamento de milhares de documentos todos os meses. Era comum nas empresas a presença de arquivos de aço para a guarda desses documentos, ocupando considerável espaço nos escritórios dessas empresas.

A modernização da legislação que aconteceu na década seguinte possibilitou a diminuição desse problema, e através de parcerias entre os governos foi possível informatizar esses processos e o que era então papel passou a ser arquivo eletrônico armazenado em computadores ou servidores, viabilizando também a automação desses processos. Esse processo de modernização, ocorrido também em vários países da América Latina, é descrito por Mello (2014) que relata uma convergência entre vários países para o uso de tecnologias semelhantes, como a estruturação dos dados em arquivos XML, a existência de uma versão impressa, como o Documento Auxiliar da Nota Fiscal Eletrônica (DANFE) e o Documento Auxiliar do Conhecimento de Transporte Eletrônico (DACTE), o uso de assinaturas digitais para verificação de autenticidade e autoria, e o

uso de *web services* para a emissão e disponibilização desses documentos. Pasa (2001) descreve uma série de normas e recomendações que viabilizaram o uso de documentos fiscais de forma confiável e legalmente viável e fala da necessidade de "transformar o contador em um verdadeiro analista de informação".

Outro importante ganho advindo da informatização desses documentos foi justamente a facilitação do uso desses documentos para não só cumprir obrigações contábeis e fiscais, mas também em processos internos de empresas e até mesmo análises financeiras e econômicas de diversas finalidades. Os documentos fiscais eletrônicos são ricos em informações, contendo dados detalhados de forma estruturada acerca das transações efetuadas entre empresas, onde são descritos produtos, valores, impostos, e classificações que possuem uma infinidade de aplicações.

Alguns trabalhos acadêmicos já exploraram anteriormente algumas dessas aplicações através do uso de documentos fiscais para propor soluções para a sociedade. Fernandez (2012) propõe o uso de documentos fiscais, como a Nota Fiscal Eletrônica (NFe), o Conhecimento de Transporte Eletrônico (CTe), e do Manifesto Econômico de Documentos Fiscais (MDFe) para o rastreamento de cargas e veículos, os tornando importantes ferramentas de prevenção e combate à sonegação e fraudes. O trabalho de Madeira (2015) também propõe um método de mineração de texto para a detecção de fraudes e incorreções através de dados de Notas Fiscais de Serviços Eletrônicas, classificando-as quanto a essas incorreções com métodos computacionais.

Outra aplicação é apresentada no trabalho de Santos (2015), que aplica o uso de dados de NFes no estudo de alternativas para a obtenção de dados para o planejamento do transporte de cargas urbano, como forma de levantar dados para a tomada de decisões que evitem congestionamentos ou otimizem fluxos de tráfego devido a esse tipo de transporte.

Por fim, o trabalho de Junior (2019) faz o uso de MDFes para a obtenção de matrizes Origem-Destino, uma ferramenta de visualização de demanda, mostrando em seus resultados que as matrizes obtidas através de dados de documento fiscais eram semelhantes às obtidas com métodos tradicionais.

Todos esses trabalhos atestam a riqueza, viabilidade, e confiabilidade dos dados de documentos fiscais para uso em variados domínios e aplicações.

1.2 A Pandemia de COVID-19

O ano de 2020 foi marcado por uma pandemia de proporções globais poucas vezes vista na história da humanidade. Segundo Zhou *et al.* (2020), a epidemia se iniciou no dia 12 de dezembro de 2019 na cidade de Wuhan, na China, causada por infecções pelo coronavírus Sars-Cov-2 capaz de causar Síndrome Respiratória Aguda Grave, doença chamada COVID-19, cujos sintomas incluem febre, tosse seca, dificuldade respiratória, dor de cabeça, podendo evoluir para casos mais graves com pneumonia, danos pulmonares, e morte.

Em 26 de fevereiro de 2020, o Brasil confirmou seu primeiro caso de infecção pelo vírus em um paciente de 61 anos vindo da Itália (JORNAL FOLHA DE SÃO PAULO, 2020a), e sua primeira morte no dia 17 de março de 2020 na cidade de São Paulo, um homem de 62 anos (JORNAL FOLHA DE SÃO PAULO, 2020b). O preocupante espalhamento do vírus por diversos países levou a Organização Mundial da Saúde a elevar em 11 de março de 2020 a classificação para pandemia (JORNAL FOLHA DE SÃO PAULO, 2020e), após a disseminação para mais de cem países em todos os continentes.

A OMS recomenda como forma de prevenção o distanciamento social, o uso de máscaras, e a higiene contínua das mãos (WORLD HEALTH ORGANIZATION, 2021b). O alto risco de contaminação por pacientes assintomáticos e pressintomáticos infectados torna o controle da doença muito difícil, sendo os principais métodos populacionais de prevenção indicados o isolamento e a quarentena (WORLD HEALTH ORGANIZATION, 2021a). A doença ainda não possui remédios eficazes indicados pela OMS, sendo recomendados apenas a prevenção e tratamentos intensivos para pacientes já em estado grave, o que pode sobrecarregar os sistemas de saúde por todo o mundo.

O Brasil, assim como diversos países do mundo, adotou medidas para conter o avanço do vírus. Por meio da lei nº 13.979/2020 (BRASIL, 2020a), alterada posteriormente pelas leis nº 14.019/2020 (BRASIL, 2020b) e 14.035/2020 (BRASIL, 2020c), foram estabelecidas medidas para enfrentamento da pandemia, que no Artigo 3º prevê:

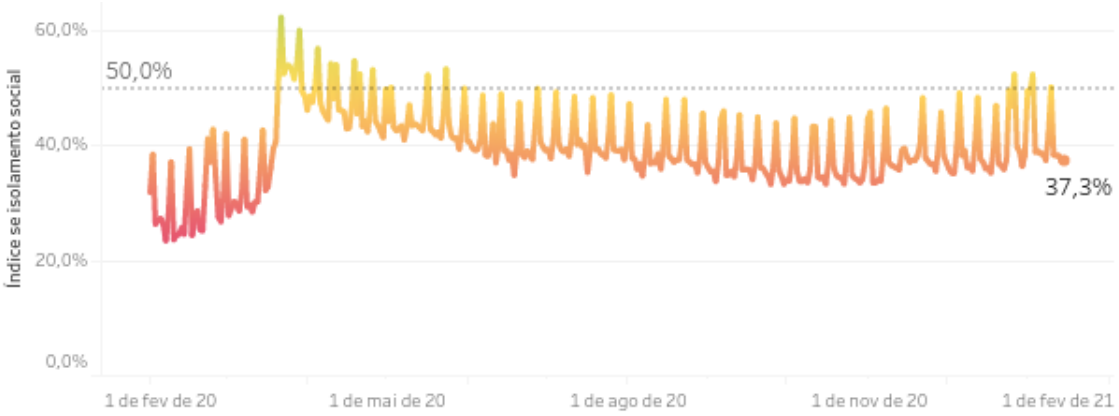
> Art. 3º Para enfrentamento da emergência de saúde pública de importância internacional de que trata esta Lei, as autoridades poderão adotar, no âmbito de suas competências, entre outras, as seguintes medidas:
>
> I - isolamento;
>
> II - quarentena;
>
> III - determinação de realização compulsória de:
>
> a) exames médicos;
>
> b) testes laboratoriais;
>
> c) coleta de amostras clínicas;
>
> d) vacinação e outras medidas profiláticas; ou
>
> e) tratamentos médicos específicos;
>
> III-A – uso obrigatório de máscaras de proteção individual;
>
> ...
>
> VI – restrição excepcional e temporária, por rodovias, portos ou aeroportos, de:
>
> a) entrada e saída do País; e
>
> b) locomoção interestadual e intermunicipal;

A adoção dessas medidas, embora não tenha ocorrido de forma ordenada em todo o país, afetou a mobilidade social. Como exemplo de tais medidas, no dia 21 de março de 2020

o governo do estado de São Paulo (JORNAL FOLHA DE SÃO PAULO, 2020d) decreta o fechamento do comércio e de todos os serviços essenciais por 15 dias, medida posteriormente prolongada cujos efeitos duraram várias semanas. Também na cidade do Rio de Janeiro, foi decretado o fechamento obrigatório do comércio no dia 22 de março de 2020, também em tentativa de conter o vírus (UOL, 2020). No dia 18 de maio de 2020, a cidade de São Paulo aprovou a antecipação de dois feriados (JORNAL FOLHA DE SÃO PAULO, 2020c) como forma de incentivar o aumento do índice de isolamento social, que indica o percentual da população que está efetivamente respeitando o isolamento social.

A Figura 1 ilustra como o índice de isolamento social variou entre fevereiro de 2020 e janeiro de 2021. O pico de isolamento ocorreu no dia 22 de março de 2020 alcançando 62.2%, mas posteriormente diminuiu estando no patamar de 37.3% em 15 de janeiro de 2021.

Figura 1 – Gráfico do índice de isolamento social no Brasil



Fonte: In Loco (2021).

Em janeiro de 2021, a OMS registra mais de 97 milhões de casos confirmados de infecção, sendo mais de 2 milhões de mortes causadas pela doença (WORLD HEALTH ORGANIZATION, 2021d), sendo que a cada dia estão ainda sendo registrados em média cerca de 100 mil novos casos e 2 mil mortes. O Brasil registra mais de 8 milhões de casos confirmados sendo 215 mil mortes causadas pela doença, sendo ainda registrados em média cerca de 51 mil casos e 1 mil mortos por dia. No momento, a OMS avalia 15 vacinas candidatas a imunizantes, algumas já com pesquisas avançadas e em uso emergencial por alguns países (WORLD HEALTH ORGANIZATION, 2021c).

O Brasil aplicou seu primeiro imunizante (JORNAL FOLHA DE SÃO PAULO, 2021) no dia 17 de janeiro de 2021, após aprovação para uso em caráter emergencial pela Agência Nacional de Vigilância Sanitária (Anvisa), em uma enfermeira de 54 anos que trabalha na UTI do Instituto de Infectologia Emílio Ribas, em São Paulo.

1.3 Crise econômica

O fechamento do comércio, as medidas de incentivo ao isolamento social, além dos altos números de mortos, são fatores que afetam economicamente diversos setores da economia.

McKibbin e Fernando (2020, tradução nossa) aponta que "em um mundo fortemente conectado e integrado, os impactos da doença (COVID-19) vão além da mortalidade (aqueles que morrem) e da morbidade (aqueles que ficam incapazes de trabalhar por um período de tempo)". O estudo, datado de março de 2020, ainda previa cenários onde a pandemia poderia se manter restrita à China, mas também a cenários que hoje sabemos mais próximos da realidade com a doença se espalhando por diversos países. Este estudo já apontava a possibilidade, hoje confirmada, de um alto número de mortes e alto impacto no Produto Interno Bruto (PIB) nos países estudados. Mesmo nos cenários mais otimistas, de menor espalhamento do vírus, o estudo já previa um desvio negativo importante para o PIB de diversos países, podendo chegar a 9.9% de desvio negativo nos piores casos. Segundo os dados do Instituto Brasil de Geografica e Estatística (IBGE), o Brasil teve no acumulado entre o último trimestre de 2019 e os três primeiros trimestres de 2020 uma queda de 3.4% no PIB Instituto Brasil de Geografia e Estatística (2021).

O trabalho de Fernandes (2020, tradução nossa) faz uma análise dos impactos econômicos da COVID-19 no cenário mundial ainda em março de 2020. O estudo aponta que "os impactos da doença vão além da mortalidade. Sendo que, governos por todo mundo tem se preparado para planos de contingência e pacotes de ajuda para sustentar suas economias', apontando também para o impacto dos *lockdowns*, uma das medidas restritivas adotadas, na redução do consumo e em interrupções na produção. O estudo também compara o evento de 2020 com crises anteriores e descreve que tanto em relação à crise econômica de 2008 quanto em relação à pandemia de SARS de 2002 há diferenças signiticativas que tornam a crise de 2020 mais grave e abrangente.

Em outro estudo, Sun, Wandelt e Zhang (2020) analisa o impacto da pandemia no transporte aéreo global e como este poderia ter contribuído para a pandemia, usando dados de radares aéreos do mundo todo. Neste trabalho, Sun, Wandelt e Zhang (2020) mostra uma diminuição da conectividade em um grafo de conexões em aeroportos de todo o mundo se comparado o mês de janeiro com o mês de maio de 2020. Ao comparar métricas de sistemas complexos neste período, o estudo também detecta uma queda de 50% na centralidade de grau, uma redução também da assortatividade de grau, um aumento significativo na centralidade de intermediação, e também na quantidade de comunidades da rede. Este estudo é um precedente importante para o nosso trabalho uma vez que aponta a alterações topológicas em sistemas complexos causadas pela pandemia.

Usando dados do mercado de ações, Aslam *et al.* (2020) faz uma análise da variação da correlação entre diversos mercado e mudanças estruturais do ponto de vista de sistemas complexos entre o período anterior e durante o espalhamento da COVID-19. Este estudo também identifica mudanças topológicas no relacionamento entre mercado de todo o mundo, relatando

um aumento do caminho mínimo médio entre eles, além de uma diminuição na centralidade de intermediação e na centralidade de proximidade. Essa análise considera apenas a árvore geradora mínima, mas já aponta mudanças diferentes das relatadas por Sun, Wandelt e Zhang (2020).

Estudos sobre o impacto econômico da crise causada pela COVID-19 ainda serão complementados por outros trabalhos, tendo em vista que ainda não alcançamos e o fim da pandemia é difícil prever quando isso acontecerá, e também por ser um evento relativamente recente e que certamente ainda será muito analisado.

1.4 Objetivos

Tendo em vista a oportunidade de utilizar valiosos dados de documentos fiscais, o impacto econômico previsto por outros trabalhos, e a importância do evento ocorrido, este trabalho se propõe a contribuir com uma análise do impacto econômico da crise provocada pela pandemia de COVID-19 através do ponto de vista de sistemas complexos, e com uma análise do impacto de mudanças topológicas de redes na magnitude e abrangência da crise para cada entidade estudada, apresentando uma categorização de impacto econômico, e verificando se há possibilidade de utilizar esses dados para categorizá-las entre entidades afetadas e não-afetadas pela crise.

No capítulo 2 será apresentado o domínio dos dados utilizados neste trabalho com uma explicação sobre documentos fiscais, a legislação brasileira envolvida, e os sistemas usados para captação dos dados. A seguir, será descrita a base de dados utilizada, com uma análise descritiva e detalhes do pré-processamento feito para a coleta dos dados no capítulo 4. O impacto econômico da crise é descrito no capítulo 5 que mostra sob diferentes visões a abrangência e magnitude da mesma. No capítulo 6 é feita uma análise descritiva com diferentes modelagens e métricas das redes utilizadas e são mostradas possíveis alterações estruturais no ponto de vista de sistemas complexos, com diferentes modelagens, e uma tentativa de identificação de características que podem levar a categorizar um agente econômico como afetado ou não pela pandemia. O capítulo 7 encerra este trabalho com as conclusões obtidas e os próximos passos propostos por este trabalho.

CAPÍTULO

2

DOCUMENTOS FISCAIS

A legislação fiscal brasileira é notoriamente conhecida por ser complexa. Nas duas últimas décadas, representantes das três esferas do governo brasileiro, municipal, estadual, e federal, se reuniram para tentar mudar essa realidade promovendo uma atuação mais integrada e modernizando o sistema. O mercado brasileiro também respondeu a essas mudanças criando oportunidades que visam facilitar o cumprimento das obrigações fiscais de empresas de todo país.

2.1 Legislação Brasileira

O Sistema Tributário Nacional foi instituído pela lei nº 5172 de 1966 (BRASIL, 1966), há mais de cinco décadas. Desde então passou por diversas reformas visando aumentar sua eficiência, modernizando-o. O artigo 37 da Constituição Federal, alterado pela Emenda Constitucional nº 42 de 19 de dezembro de 2003 (BRASIL, 2003), em seu vigésimo segundo inciso prevê:

> Art. 37. A administração pública direta e indireta de qualquer dos Poderes da União, dos Estados, do Distrito Federal e dos Municípios obedecerá aos princípios de legalidade, impessoalidade, moralidade, publicidade e eficiência e, também, ao seguinte:
>
> ...
>
> XXII - as administrações tributárias da União, dos Estados, do Distrito Federal e dos Municípios, atividades essenciais ao funcionamento do Estado, exercidas por servidores de carreiras específicas, terão recursos prioritários para a realização de suas atividades e atuarão de forma integrada, inclusive com o compartilhamento de cadastros e de informações fiscais, na forma da lei ou convênio (BRASIL, 1988).

O compartilhamento de cadastros e de informações previsto neste artigo permitiu então que as esferas do governo brasileiro tomassem atitudes no sentido de centralizar sistemas tributários, criando protocolos e sistemas para tal.

Desde então, diversos secretários da Fazenda Federal, das Fazendas Estaduais e do Distrito Federal, e de Fazendas Municipais se reúnem anualmente no Encontro Nacional de Administradores Tributários (ENAT) para discutir como modernizar e tornar mais efetivo o Sistema Tributário Brasileiro. A seguir, são descritos alguns importantes protocolos assinados nesses encontros.

2.1.1 ENAT

Em julho de 2004, em Salvador-BA, ocorre a primeira edição do ENAT. Neste encontro, foram assinados dois protocolos de cooperação técnica para a instalação do Projeto Cadastro Sincronizado e do Projeto de Escrituração Digital.

No protocolo de Projeto Cadastro Sincronizado, fica previsto a construção de um cadastro de contribuintes sincronizado seguindo as seguintes diretivas:

> Na construção do cadastro referido na cláusula primeira, serão observados os seguintes parâmetros, entre outros que vierem a ser definidos de comum acordo pelos partícipes:
>
> I - entrada de dados única;
>
> II - bases de dados independentes, porém sincronizadas;
>
> III - reciprocidade na aceitação da legislação de cada ente signatário;
>
> IV - adoção do número de inscrição no Cadastro Nacional da Pessoa Jurídica (CNPJ) como identificador cadastral dos contribuintes do ICMS e ISS (ENAT, 2004).

Já no protocolo de Projeto de Escrituração Digital, são criadas iniciativas para a modernização da administração tributária brasileira segundo os seguintes parâmetros:

> A administração tributária está assentada sobre três pilares básicos, a saber, o cadastro de contribuintes, que permita a perfeita identificação e individualização das pessoas, o documento básico de comércio, Nota Fiscal, que registre a atividade comercial com suas particularidades e a codificação das mercadorias e serviços (ENAT, 2004);

Mais adiante, o documento define as seguintes prioridades a serem atacadas, a destacar:

> IV – criar grupo de trabalho para elaborar proposta com vistas à adaptação da codificação da Nomenclatura Comum do Mercosul - NCM às especificidades tributárias do ICMS;
>
> IX - investir no desenvolvimento de modelo de dados único e padronizado para todos os Fiscos, relativamente às demais ferramentas de Administração Tributárias;
>
> X - investir na regulamentação do uso da certificação digital em todos os documentos fiscais, utilizado o programa de Transmissão Eletrônica de Documentos - TED , como padrão nacional;

> XI - investir na harmonização da legislação das Unidades da Federação e, na padronização da escrituração fiscal e das informações econômico-fiscais;
>
> XII – compartilhar os sistemas de auditoria fiscal existentes nas diversas Unidades Federadas e promover o desenvolvimento conjunto de novas ferramentas de apoio a ação fiscal;
>
> XIII – desenvolver e disponibilizar aos contribuintes programa que permita a captura e transmissão on-line dos documentos fiscais emitidos para as administrações fiscais;

Em agosto do ano seguinte, em São Paulo-SP, ocorre a segunda edição do ENAT, adicionando quatro importantes protocolos de cooperação técnica, das quais ressaltam-se três.

O segundo estabelece o Sistema Público de Escrituração Digital (SPED) sob os seguintes pressupostos:

> I - bases de dados compartilhadas entre as Administrações Tributárias;
>
> II - reciprocidade na aceitação da legislação de cada ente signatário, relativa aos livros contábeis e fiscais;
>
> III - validade jurídica dos livros contábeis e fiscais em meio digital, dispensando a emissão e guarda de documentos e livros em papel;
>
> IV - eliminação da redundância de informações através da padronização e racionalização das obrigações acessórias;
>
> V - preservação do sigilo fiscal, nos termos do Código Tributário Nacional (ENAT, 2005a).

O terceiro estabelece a Nota Fiscal Eletrônica, parte integrante do SPED, definindo:

> I - substituição das notas fiscais em papel por documento eletrônico;
>
> II - validade jurídica dos documentos digitais;
>
> III - padronização nacional da NF-e;
>
> IV - mínima interferência no ambiente operacional do contribuinte;
>
> V - compartilhamento da NF-e entre as administrações tributárias;
>
> VI - preservação do sigilo fiscal, nos termos do Código Tributário Nacional (ENAT, 2005b).

E por fim, a padronização e aplicação da Classificação Nacional de Atividades Econômicas (CNAE) para o Cadastro Sincronizado de Contribuintes (ENAT, 2005c).

Essas iniciativas viriam possibilitar então a criação de sistemas automatizados de controle fiscal. Em especial, a criação do SPED e da NF-e possibilitaram o uso de documentos fiscais não só com a finalidade de simplificar as obrigações fiscais das empresas brasileiras, mas também sua utilização como artefato de controle interno, devido à sua riqueza de informações estruturadas.

Na edição seguinte, em 2006, o ENAT ocorreu em Fortaleza-CE e foram estabelecidos protocolos de cooperação técnica que seguiram adicionando novos artefatos, os quais ressaltam-se a implantação da Nota Fiscal de Serviço Eletrônica (NFSe) e do Conhecimento de Transporte Eletrônico, outros dois documentos fiscais eletrônicos. O encontro segue ocorrendo anualmente e os protocolos tem se aperfeiçoado para melhorar e integrar o sistema tributário e fiscal brasileiro.

2.1.2 Documentos Fiscais Eletrônicos

Existem diversos documentos fiscais, como já mencionado neste trabalho. As empresas tem a obrigação fiscal de emitir e armazenar estes documentos. O prazo prescricionário destes documentos é apresentado nos artigos 173 e 174 do Código Tributário Nacional (BRASIL, 1966):

> Art. 173. O direito de a Fazenda Pública constituir o crédito tributário extingue-se após 5 (cinco) anos, contados:
>
> I - do primeiro dia do exercício seguinte àquele em que o lançamento poderia ter sido efetuado;
>
> II - da data em que se tornar definitiva a decisão que houver anulado, por vício formal, o lançamento anteriormente efetuado.
>
> Parágrafo único. O direito a que se refere este artigo extingue-se definitivamente com o decurso do prazo nele previsto, contado da data em que tenha sido iniciada a constituição do crédito tributário pela notificação, ao sujeito passivo, de qualquer medida preparatória indispensável ao lançamento.
>
> Art. 174. A ação para a cobrança do crédito tributário prescreve em cinco anos, contados da data da sua constituição definitiva.
>
> Parágrafo único. A prescrição se interrompe:
>
> I – pelo despacho do juiz que ordenar a citação em execução fiscal;
>
> II - pelo protesto judicial;
>
> III - por qualquer ato judicial que constitua em mora o devedor;
>
> IV - por qualquer ato inequívoco ainda que extrajudicial, que importe em reconhecimento do débito pelo devedor.

Ou seja, as empresas tem a obrigação da guarda de documentos por pelo menos cinco anos, sendo maior a depender do tipo de documento fiscal.

A Secretaria de Estado da Fazenda (Sefaz), um órgão vinculado ao Ministério da Fazenda, disponibiliza o Portal da Nota Fiscal Eletrônica (BRASIL, 2021a), onde as empresas brasileiras podem consultar os seus documentos fiscais utilizando seu certificado digital. Também são disponibilizados serviços *web* para a emissão e consulta dos documentos fiscais (BRASIL, 2021b), vinculados às Secretarias da Fazenda estaduais. Algumas unidades federativas utilizam o Ambiente Nacional, em uma proposta de unificação dos sistemas.

Podemos dizer que a Nota Fiscal Eletrônica é o principal documento dentre os documentos fiscais, uma vez que é ela quem efetivamente descreve todos as movimentações de mercadorias de uma empresa. A Nota Fiscal Eletrônica é disponibilizada no formato Extensible Markup Language (XML) (W3C, 2021), uma linguagem de marcação comum em protocolos SOAP. Um exemplo da estrutura e conteúdo de uma NFe está disponível no Apêndice C

Figura 2 – Exemplo de DANFE através do Portal Arquivei

IDENTIFICAÇÃO DO EMITENTE
LOURENCO S.CARLOS COM.LUBR.EM GERAL LTDA
R EPISCOPAL,3055,
CIDADE JARDIM - 13560580
SAO CARLOS - SP Fone/Fax: 1633613168

DANFE
Documento Auxiliar da Nota Fiscal Eletrônica
1 - Saída
Nº 33041
SÉRIE 1

CHAVE DE ACESSO 3521 0101 1801 6900 0133 5500 1000 0330 4110 2215 6312
Consulta de autenticidade no portal nacional da NF-e www.nfe.fazenda.gov.br/portal ou no site da Sefaz Autorizadora

NATUREZA DA OPERAÇÃO: VENDA
PROTOCOLO DE AUTORIZAÇÃO DE USO: 135210063446995 - 19/01/2021 10:33:24

INSCRIÇÃO ESTADUAL: 637124250112
INSCRIÇÃO ESTADUAL DO SUBST. TRIBUT.:
CNPJ: 01.180.169/0001-33

DESTINATÁRIO

NOME / RAZÃO SOCIAL	CNPJ/CPF	DATA DA EMISSÃO
C.JR.CONSTRUTORA LTDA	03.160.081/0001-85	19/01/2021

ENDEREÇO	BAIRRO / DISTRITO	CEP	DATA DA ENTRADA / SAÍDA
ROD W. LUIZ KM 240, S/N	RURAL	13560-970	19/01/2021

MUNICÍPIO	UF	FONE / FAX	INSCRIÇÃO ESTADUAL	HORA DA SAÍDA
3548906 - SAO CARLOS	SP	1633612039	637143185117	10:34:00

FATURA / DUPLICATA

Núm.	001
Venc.	16/02/2021
Valor	R$ 479,00

CÁLCULO DO IMPOSTO

BASE DE CÁLCULO DO ICMS	VALOR DO ICMS	BASE DE CÁLC. DO ICMS ST	VALOR DO ICMS ST	VALOR DO IMP. IMPORTAÇÃO	VALOR DO PIS	VALOR TOTAL DOS PRODUTOS
R$ 0,00	R$ 0,00	R$ 0,00	R$ 0,00	R$ 0,00	R$ 3,11	R$ 479,00

VALOR DO FRETE	VALOR DO SEGURO	DESCONTO	OUTRAS DESPESAS	VALOR TOTAL DO IPI	VALOR DA COFINS	VALOR TOTAL DA NOTA
R$ 0,00	R$ 0,00	R$ 0,00	R$ 0,00	R$ 0,00	R$ 14,37	R$ 479,00

TRANSPORTADORA / VOLUMES TRANSPORTADOS

NOME/RAZÃO SOCIAL	FRETE POR CONTA	CÓDIGO ANTT	PLACA DO VEÍCULO	UF	CNPJ/CPF
SEM FRETE	9				

ENDEREÇO	MUNICÍPIO	UF	INSCRIÇÃO ESTADUAL

QUANTIDADE	ESPÉCIE	MARCA	NÚMERO	PESO BRUTO	PESO LÍQUIDO
0				0,000	0,000

Fonte: Elaborada pelo autor.

A disponibilização dos documentos é feita conforme um critério determinado pela Sefaz em que a empresa tem acesso ao documento de acordo com o papel que ela exerce nele. Por exemplo, NFes são disponibilizadas ao destinatário de uma NFe mas não ao emitente. Desta forma, para obter a completude de todos os documentos fiscais de uma empresa de forma automatizada, é necessário que o sistema emitente também armazene os dados.

2.2 A Arquivei

Os dados obtidos neste trabalho foram adquiridos através da empresa parceira, Arquivei (ARQUIVEI, 2021). A Arquivei é uma empresa especializada em inteligência em documentos fiscais, fundada em 2013 na cidade São Carlos no interior de São Paulo. A empresa disponibiliza um sistema automatizado de consulta de documentos fiscais, utilizando os certificados digitais de seus clientes para obter e armazenar os dados de documentos fiscais. Através do portal, os clientes podem então não apenas cumprir com sua obrigação tributária, mas emitir relatórios e automatizar processos internos das empresas a partir desses dados.

Através dos dados dos clientes da empresa, é possível obter dados de terceiros por meio de um efeito de rede, uma vez que documentos fiscais podem citar diversas empresas. Esse efeito de rede é importante pois aumenta a abrangência dos dados analisados.

A empresa tem hoje mais de 150 funcionários, 12 mil clientes, e atende mais de 85 mil empresas de todo o Brasil. Este trabalho utiliza uma amostra da base de dados da empresa, cuja abrangência será descrita em detalhes no capítulo 4.

2.2.1 Segurança dos dados

O tratamento dos dados nos ambientes da empresa parceira foram executados apenas pelo autor deste trabalho, de maneira anonimizada, mantendo o sigilo e privacidade previstos na legislação atual e os contratos de prestação de serviço entre a empresa parceira e seus clientes, além de cumprir os critérios previstos no contrato de trabalho entre a empresa parceira e autor deste trabalho. Os dados são armazenados pela empresa parceira utilizando avançados critérios e procedimentos de segurança e com acesso restrito, não tendo tais critérios e procedimentos sido comprometidos de forma alguma na execução deste trabalho.

CAPÍTULO

3

METODOLOGIA

3.1 Redes Complexas

Nesta seção serão apresentados os métodos, conceitos, e técnicas utilizados neste trabalho para referência.

3.1.1 Conceitos

Newman (NEWMAN, 2018) define uma rede da seguinte forma: "Uma **rede** é, em sua forma mais simples, uma coleção de pontos unidos em pares por linhas. Na nomenclatura da área um ponto é chamado **nó** ou **vértice** e a linha é chamada **aresta**". Uma rede é também chamada de **grafo** na literatura matemática.

A quantidade de vértices de um grafo G é denotado por n e a quantidade de arestas é denotado por m. Uma aresta que conecta um vértice a ele mesmo é chamado **ciclo**. Um grafo pode ter mais de uma aresta entre dois vértices, neste caso estas são chamadas **multiarestas**, e o grafo é chamado **multigrafo**. Um grafo que possui ciclos é chamado **grafo cíclico**, do caso contrário **grafo acíclico**. Um grafo que não possui ciclos nem multiarestas é chamado **grafo simples**. Um grafo em que cada vértice está conectada a todos os outros vértices é chamado **grafo completo**.

As arestas de um grafo podem ou não possuir uma direção associada, indicando que cada aresta parte de um vértice em direção a outro. Um grafo que possui arestas direcionadas é chamado **grafo direcionado**, também chamado **dígrafo**, caso contrário é chamado **grafo não-direcionado**. As arestas de um grafo também podem ou não possuir valores associados, chamados de **pesos**. Um grafo que possui pesos é chamado **grafo ponderado**.

Um **subgrafo** é um conjunto de vértices e arestas que os conectam pertencentes a um grafo.

3.1.2 Representação

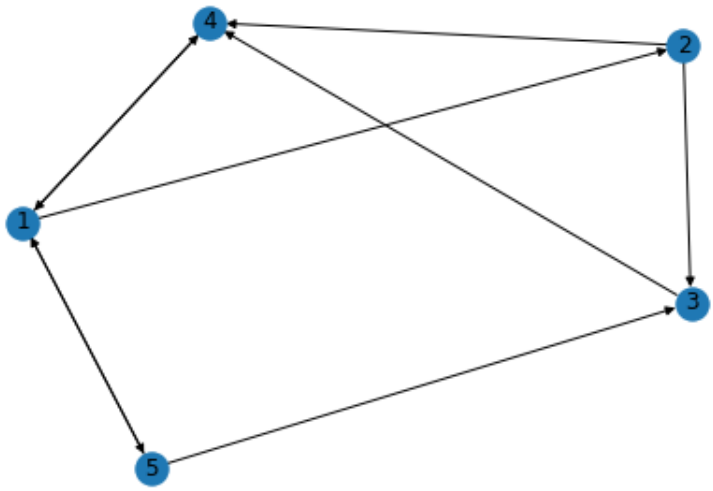
Um grafo G composto por vértices $v_0, v1, ..., v_n$ e arestas e_{ij}, sendo $0 \leq i \leq n$ e $0 \leq j \leq n$, é comumente representado por uma matriz de adjacências na forma:

$$A_{ij} = \begin{cases} 1 & \text{se existe uma aresta } e_{ij} \text{ entre os vértices i e j,} \\ 0 & \text{caso contrário} \end{cases} \tag{3.1}$$

No caso de um grafo ponderado, a matriz toma a seguinte forma:

$$A_{ij} = \begin{cases} w_{ij} & \text{se existe uma aresta } e_{ij} \text{ entre os vértices i e j de peso } w_{ij}, \\ 0 & \text{caso contrário} \end{cases} \tag{3.2}$$

Figura 3 – Exemplo de grafo acíclico direcionado não-ponderado



Fonte: Elaborada pelo autor.

A figura 3 representa um grafo que pode ser representado pela matriz:

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{3.3}$$

3.1.3 Métricas

A seguir são descritas algumas métricas de caracterização de redes e vértices usados no decorrer do trabalho.

3.1.3.1 Grau

O **grau** de um vértice, também denotado por $< k >$, é o número de arestas adjacentes a este vértice. No caso de grafos direcionados, o número de arestas que partem de um determinado vértice é chamado **grau de saída**, enquano o número de arestas que chegam a um determinado vértice é chamado **grau de entrada**. Em um grafo ponderado, pode-se calcular também o **grau ponderado** de um vértice através da soma dos pesos de cada aresta conectada aquele vértice.

O **grau médio** de uma rede é definido como:

$$< k >_G = \frac{1}{n} \sum_{i=0}^{n} < k >_i \tag{3.4}$$

Enquanto o **grau médio ponderado** é definido da mesma forma mas através da soma do grau ponderado de cada vértice.

A **densidade** de um grafo indica a razão de arestas que existem em relação a todas as conexões possíveis. Para grafos não-direcionados ela é definida como:

$$d = \frac{2m}{n(n-1)} \tag{3.5}$$

Enquanto para grafos direcionados ela é definida como:

$$d = \frac{m}{n(n-1)} \tag{3.6}$$

3.1.3.2 Caminhos mínimos

Um **caminho** é definido como uma sequência de vértices $v_i, v_{i+1}, ..., v_k$ de forma que exista uma aresta e_{ij} que conecta cada vértice v_i ao vértice v_{i+1}. A distância percorrida por um caminho é dado pela quantidade de arestas que o compõe. O **caminho mínimo** entre dois vértices é o caminho de menor quantidade de arestas que leva de um vértice a outro. A **distância** entre dois vértices é dada pela distância percorrida pelo caminho mínimo entre elas.

Em grafos ponderados, a **distância ponderada** é dada pela soma de todos os pesos w_{ij} pertencentes às arestas de um caminho. O **caminho mínimo ponderado** é definido como o caminho de menor valor que leva de um vértice a outro.

Um grafo é chamado **conexo** se existe um caminho entre qualquer par de vértices, e **desconexo** caso contrário. Uma **componente conexa** é um subgrafo conexo maximal, isto é, que um subgrafo que não está estritamente contido em outros subgrafos conexos. A **maior componente conexa** é a componente conexa de um grafo que tenha o maior número de vértices em relação às demais. Para as formulações a seguir vamos supor que as medidas são aplicadas sobre um grafo conexo.

3.1.3.3 Excentricidade

A **excentricidade** de um vértice é dado pela distância máxima entre este vértice e todos os outros. O **diâmetro** de uma rede é dado pela excentricidade máxima dentre todos os vértices da rede, e o **raio** é dado pela excentricidade mínima.

3.1.3.4 Transitividade

Essa métrica foi descrita por Luce e Perry (1949), e posteriormente nomeada e aprofundada por Wasserman, Faust *et al.* (1994), e é definida da seguinte forma: uma **tríade** é definida por duas arestas que compartilham o mesmo vértice. Um **triângulo** é definido como três vértices que são conectados entre si. A **transitividade** de um grafo é definida como:

$$T = 3\frac{\#triades}{\#triangulos} \tag{3.7}$$

onde $\#triades$ mede a quantidade de tríades em um grafo, e $\#triangulos$ a quantidade de triângulos.

A transitividade também é conhecida como **coeficiente de agrupamento global**, pois indica o quão agrupados são os vértices de uma rede. Da mesma forma, proposto por Watts e Strogatz (1998), defini-se o **coeficiente de agrupamento local** de um vértice i como:

$$c_i = \frac{2T_i}{< k >_i (< k >_i - 1)} \tag{3.8}$$

Para grafos ponderados, usa-se a média geométrica dos pesos das arestas:

$$c_i = \frac{2T_i}{< k >_i (< k >_i - 1)} \sum_{jk} (w_{ij} w_{ik} w_{jk})^{\frac{1}{3}} \tag{3.9}$$

Neste caso, é comum que seja feita a normalização dos pesos dos vértices em relação ao peso máximo da rede.

3.1.3.5 Assortatividade

A **assortatividade** é uma métrica de correlação entre as características dos vértices de uma rede. Ela é descrita por Newman (2003) e é calculada a partir de uma matriz e em que cada elemento e_{ij} representa a fração das arestas da rede que conectam um vértice do tipo i a um vértice do tipo j. Então são definidos:

$$a_i = \sum_j e_{ij}, b_j = \sum_i e_{ij} \tag{3.10}$$

O coeficiente de assortatividade é então definido da seguinte maneira:

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{3.11}$$

Quando $r = 1$, a rede é totalmente **assortativa** e seus vértices tendem a se conectar apenas a vértices semelhantes, enquanto $r = -1$ representa uma rede totalmente **disassortativa** onde os vértices se conectam apenas à vértices diferentes. Por fim, $r = 0$ indica uma rede **não-assortativa** onde os vértices não tem uma preferência definida.

A assortatividade pode ser utilizada para medir a correlação entre qualquer atributo relacionado aos vértices e de uma rede. A **assortatividade de grau**, por exemplo, é comumente usada para medir a preferência de vértices mais conectados, chamados *hubs*, de se conectar com vértices menos conectados, e vice-versa. Enquanto cada vértice pode ter um atributo relacionado, por exemplo o sexo de uma pessoa, e podemos então medir a preferência de homens e mulheres de se conectarem a pessoas de determinado sexo em uma rede social.

3.1.3.6 Centralidade

A centralidade pode ser intuitivamente definida como a medida de importância de um vértice em uma rede. Quanto mais *central* é um véritce, maior a importância dele para a rede. Existem várias formas de se calcular a centralidade, sob diversos pontos de vista, os quais apresentaremos a seguir.

A primeira forma de calcular centralidade é utilizando a **centralidade de grau**. A centralidade de grau c_i de um vértice i é calculada pela fração dos vértices de uma redes que estão conectada ao um vértice i:

$$c_i = \frac{< k >_i}{n} \tag{3.12}$$

Ou seja, essa medida é análoga ao grau de um vértice mas normalizado pela quantidade de vértices da rede.

Freeman (1979) define outra métrica de centralidade, a **centralidade de proximidade**. Ela é definida como a distância média entre um vértice e todos os outros daquela rede, da seguinte forma:

$$c_i = \frac{n-1}{\sum_{j=0}^{n} d(i,j)} \tag{3.13}$$

onde $d(i,j)$ indica a distância entre os vértices i e j.

Uma variação da centralidade de proximidade é a **centralidade de informação**, proposta por Stephenson e Zelen (1989), que ao invés de utilizar a distância como medida usa a distância resistiva, uma medida derivada de sistemas elétricos, para representar o fluxo de informação entre dois vértices.

Outra medida relacionada a caminhos mínimos, a **centralidade de intermediação** também proposta por Freeman (1977) mede a participação de um vértice em relação a todos os caminhos mínimos entre quaisquer vértices da rede. Dessa forma, definimos a centralidade de intermediação da seguinte forma:

$$c_i = \sum_{j=0,k=0}^{n} \frac{\sigma(j,k|i)}{\sigma(j,k)} \tag{3.14}$$

sendo que $\sigma(j,k)$ denota a quantidade de caminhos mínimos entre j e k, e $\sigma(j,k|i)$ denota quantos desses caminhos passam por k. Tal medida pode ou não usar a formulação ponderada para caminhos mínimos.

A **centralidade de autovalor**. proposta por Bonacich (1987), usa uma formulação matricial para a sua formulação. Através de uma formulação recursiva, podemos definir a centralidade de um vértice através da centralidade de seus vizinhos usando uma matriz de adjacências A_{ij}:

$$c_i = \frac{1}{\lambda} \sum_{j \in adj(i)} a_{ij} c_j \tag{3.15}$$

sendo λ uma constante e $adj(i)$ se refere a lista de vizinhos de i. A centralidade de um vértice é então, de forma simplificada, a soma da centralidade de seus vizinhos. Ao rearranjar essa formulação, obtemos:

$$Ax = \lambda x \tag{3.16}$$

também chamada de equação de autovalores, uma equação comum na Álgebra Linear e que possui diferentes métodos de solução através de métodos iterativos, como o Método das Potências de Mises e Pollaczek-Geiringer (1929).

Outro método baseado na centralidade de autovalores é o algoritmo **PageRank**, proposto pelos fundadores da Google Page *et al.* (1999), que utiliza uma formulação análoga mas considerando a direção das arestas, recebendo maior centralidade o vértice que possui arestas que partem de vértices também de alta centralidade.

3.2 Análise de Componentes Principais

Proposto por Pearson (1901), a Análise de Componentes Principais, também chamada PCA do inglês Principal Components Analysis, é uma ferramenta comum de análise exploratória de dados. O PCA é definido como uma transformação linear ortogonal aplicada a uma matriz para que os dados sejam projetados em um novo sistemas de coordenadas de forma que as coordenadas estejam ordenadas refletindo a variância de cada coordenada.

Intuitivamente, o objetivo com PCA é encontrar uma base que diminua a redundância entre um conjunto de variáveis analisadas. Uma medida de redundância é a correlação entre duas variáveis, se eles são totalmente correlacionadas então existe muita redundância no conjunto de variáveis e podemos reduzir a quantidade de variáveis sem perda de informação, por exemplo.

Aplicando esse método sobre uma matriz X_{nxm} com m variáveis que descrevem n experimentos, obtemos uma matriz de autovetores A_{nxm}, e um vetor de autovalores V_{mx1}. Cada componente do vetor de autovalores indicará a quantidade de variância que a aquela componente principal explica. Intuitivamente, se normalizarmos o vetor de autovalores, podemos obter o percentual de informação que cada componente contém em relação ao conjunto de dados inicial, ou seja, o primeiro componente do vetor indicará a variância explicada pela primeira componente, a segunda indica a variância da segunda componente, e assim por diante. A matriz de autovetores indica a influência de cada variável em relação a cada componente, ou seja, a primeira linha desta matriz indicará a direção da transformação aplicada sobre cada variável para a primeira componente, e quanto maior a magnitude dessa direção, maior a influência dessa variável sobre aquela componente.

A PCA pode ser utilizada para diminuição de dimensionalidade, uma vez que selecionando as duas primeiras componentes principais podemos, por exemplo, projetar os dados iniciais em um espaço de duas dimensões, e então ter uma visualização dos dados em um espaço cartesiano de forma que sabemos quanto da informação contida no conjunto de dados inicial está explicada ali. Isso é particularmente útil para observar o espalhamento de dados para problemas de classificação.

3.3 Classificadores

Nesta seção, serão apresentados alguns conceitos e métodos de classificação utilizados no decorrer desta dissertação.

3.3.1 Conceitos

Vapnik (1999) nos dá uma visão geral da Teoria do Aprendizado Estatístico, um campo introduzido em 1960, com uma formulação de como ocorre o aprendizado. Ele modela o problema do aprendizado da seguinte forma: são dados três componentes:

1. Uma distribuição desconhecida $P(x)$
2. Um supervisor capaz de produzir um vetor y a partir de uma distribuição também desconhecida $P(y|x)$
3. Uma máquina de aprendizado capaz de implementar um conjunto de funções $f(x,\alpha)$

O aprendizado é descrito como um problema de minimização do risco, definido como:

$$R(\alpha) = \int L(y, f(x,\alpha))dP(x,y) \tag{3.17}$$

Queremos escolher uma função $f(x,\alpha)$ de forma que o risco seja o menor possível. De forma intuitiva, queremos construir algoritmos que sejam capazes de rotular dados de forma eficiente.

Geralmente, se divide o problema do aprendizado em **classificação**, quando os rótulos são discretos, e **regressão**, quando são contínuos.

Uma vez aplicado um algoritmo de classificação, existem uma série de métricas para avaliarmos se houve sucesso ou não na tarefa. Para um problema binário, onde queremos rotular um conjunto de entidades em Positivos e Negativos, podemos construir uma **matriz de confusão** como ilustrado no quadro 1.

Quadro 1 – Matriz de confusão

Verdadeiros Negativos (VN)	Falsos Positivos (FN)
Falsos Negativos (FP)	Verdadeiros Positivos (VP)

Ou seja, cada linha representa o rótulo esperado e cada coluna representa o rótulo predito pelo classificador. O rótulo predito pode ser Verdadeiro ou Falso, se o classificador tiver acertado ou não.

A primeira métrica usada para a análise de classificadores é a **acurácia**, que mede quantas predições o classificador acertou em relação ao total de amostras:

$$A = \frac{VP + VN}{Total} \tag{3.18}$$

Essa métrica nos diz quanto um determinado classificador acerta em relação ao total, mas isso nem sempre é suficiente, principalmente em conjuntos de dados desbalanceados.

A **precisão** indica quantos Positivos eram de fato Positivos quando foram classificados dessa forma:

$$P = \frac{VP}{VP + FP} \tag{3.19}$$

Outras duas métricas são importantes, a **sensibilidade**, também chamada *recall*, que indica quantos de todos os casos Positivos foram efetivamente rotulados corretamente:

$$S = \frac{VP}{VP + FN} \tag{3.20}$$

E a **especificidade**, que faz o mesmo para Negativos:

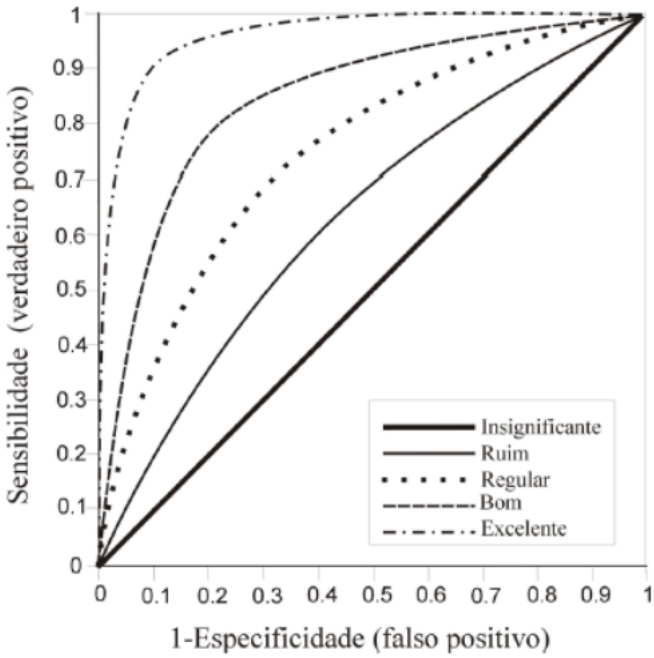
$$E = \frac{VN}{VN + FP} \tag{3.21}$$

Para ponderar especificidade e sensibilidade, existe a medida **F1-score**, que é a média harmônica entre precisão e sensibilidade:

$$F = 2\frac{P * S}{P + S} \tag{3.22}$$

Outro meio de avaliar um classificador, usando um método visual, é a curva ROC, do inglês Receiver Operating Characteristic, ou em tradução lívre Característica de Operação do Receptor. Essa curva representa um gráfico da especificidade de um algoritmo em relação a sua sensibilidade, ilustrado na Figura 4.

Figura 4 – Exemplo de Curva ROC



Fonte: Souza (2008, p.06).

Para a validação da performance de um classificador sobre um conjunto de dados são utilizadas também uma série de técnicas. A maior parte dessas técnicas utiliza uma divisão do conjunto de dados em conjunto de treino e outro conjunto de testes. O primeiro é utilizado pelo

classificador para que este ajuste seus parâmetros, e posteriormente este classificador é aplicado ao conjunto de validação e então são observados os resultados.

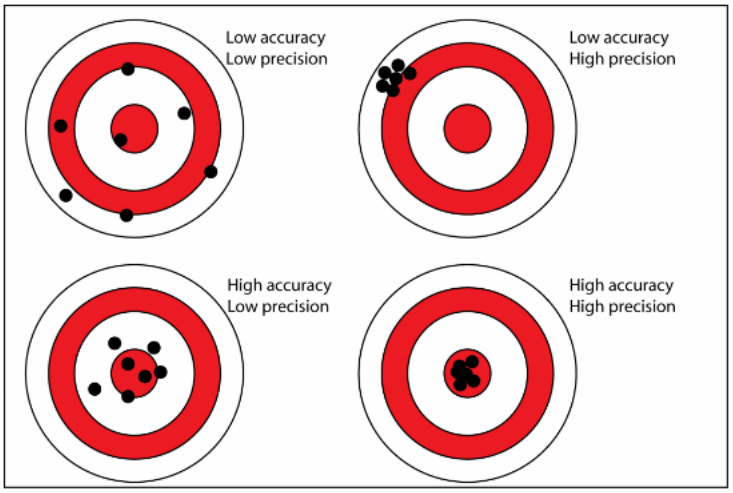
Um método comum para conjuntos de dados pequenos é o **leave-one-out**, em tradução livre *deixe um fora*. Neste método, para um conjunto de dados com N amostrar, são feitos N testes onde a cada iteração uma amostra é escolhida como teste e todo o restante do conjunto de dados é utilizado como conjunto de treino. Essa é uma das formas mais fortes de validação de um algoritmo, uma vez que toda informação é usada para classificação.

Outra técnica comumente usada, mas em conjuntos de dados maiores, chama-se **k-fold**, ou *k-dobras*. Nela, o conjunto de dados é dividido em k subconjuntos mutualmente exclusivos, e a cada iteração $k-1$ subconjuntos são utilizados para treinamento e o subconjunto restante é utilizado para validação. Essa técnica também possui sua versão estratificada que mantém a proporção de classes do conjunto inicial.

Algo importante de se observar em todo classificador é o dilema viés-variância. O **viés** de um classificador é a tendência dele de fazer suposições erradas sobre o conjunto de dados. Um alto viés ira causar **underfitting**, isto é, o classificador será incapaz de prever características nos dados necessárias para rotular corretamente os dados. Por outro lado, viés é algo necessário para que o classificador seja objetivo.

A **variância** se refere à capacidade de um algoritmo de lidar com diferentes características em um conjunto de dados. Um algoritmo com alta sensibilidade à variância irá utilizar cada pequena diferença nos dados para rotular os dados, não conseguindo alcançar o que é chamado de **generalização**. A generalização é a capacidade de um algoritmo de classificação de rotular corretamente os dados ignorando ruídos nos dados, evitando **overfitting**, isto é evitando que o classificador seja demasiadamente influenciado pelo conjunto de treino.

Figura 5 – O dilema viés-variância



Fonte: Portable Spectral Services (2019, p.01).

A seguir, utilizamos alguns classificadores utilizados neste trabalho.

3.3.2 K-Nearest Neighbors

O primeiro algoritmo aqui citado é o K-Neares Neighbors, ou em tradução livre K-Vizinhos mais próximos. Este classificador, desenvolvido por Fix (1951), é um dos mais comumente utilizados como referência dada usa simplicidade.

Este algoritmo funciona através do cálculo da distância entre pontos no espaço multidimensional. Para classificar cada amostra, o algoritmo calcula quais são as k amostras mais próximas e faz uma votação entre elas para decidir qual será o rótulo predito para a amostra sendo testada. O rótulo mais comum dentre os vizinhos de uma amostra indicará o rótulo desta amostra.

Este é um método de baixa complexidade computacional, uma vez que o parâmetro k é escolhido e funciona bem para problemas onde há uma boa separação espacial das classes.

3.3.3 Random Forest

O classificador random forest, ou florestas aleatórias, é um classificador *ensemble*, isto é, um classificador que se utiliza de diversos classificadores. O algoritmo se utiliza de árvores aleatórias de decisão, propostas por Ho (1995) a partir do trabalho de discriminzação estatística proposta por Kleinberg (1990).

Uma árvore de decisão é uma estrutura de decisão construída a partir do conjunto de treino. Nela, cada folha representará um estado, e cada vértice do topo até as folhas representará uma variável e condições relacionadas a essa variável. Para classificar um teste, percorre-se essa árvore desde o topo, e para cada nó verificam-se as condições necessárias para seguir para um dos vértices-filhos até que se chegue a uma folha.

Uma árvore de decisão pode ser construída de várias formas, geralmente utilizando uma abordagem *top-down*, onde a cada nova amostra de treino, a árvore é percorrida e é calculado o ganho de informação com uma nova divisão da árvore. Existem diversas métricas de ganho de informação, como por exemplo o coeficiente de gini ou entropia. De forma intuitiva, uma árvore de decisão divide o espaço em diversas regiões retangulares delimitadas pelas condições para divisão de cada variável.

O algoritmo de florestas aleatórias cria então diversas árvores de decisão e ao classificar uma amostra de teste faz uma eleição dentre elas para então rotular o dado. Este algoritmo é eficiente em situações onde há um bom espalhamento dos dados e é particularmente sensível a *overfitting*, sendo computacionalmente eficiente.

3.3.4 Regressão Logística

Segundo Cramer (2002), a regressão logística é produto de diversos trabalhos nos séculos XVIII e XIX. A regressão logística é usada hoje como técnica de classificação que aplica uma regressão sobre a função logística, também chamada sigmóide (FACURE, 2017):

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{3.23}$$

O que essa técnica faz é minimizar a chamada entropia cruzada, aqui denotada por:

$$E = \begin{cases} -log(\hat{x}), se\, x = 1 - log(1-\hat{x}), se\, x = 0 \end{cases} \tag{3.24}$$

onde x é o rótulo real e $\hat{x}$ o rótulo previsto. Veja que se $x = 1$ e $\hat{x} = 0$, então E será infinito, o que também ocorre se $x = 0$ e $\hat{x} = 1$, ou seja, estamos penalizando erros de classificação com essa modelagem. Podemos reescrever o nosso objetivo então como:

$$E = \sum -xlog(\hat{x}) - (1-x)log(1-\hat{x}) \tag{3.25}$$

Minimizando essa função, obteremos uma série de parâmetros que podem ser utilizados para predizer um valor $\hat{x}$ para uma amostra, que irá variar entre 0 e 1, que são interpretadas como a chance dessa amostra pertencer a uma classe ou outra.

Esse método é particularmente eficiente em situações onde já diferenças consideráveis entre as classes, seja desbalanceamento ou diferentes variâncias, sendo bastante útil.

3.3.5 SVC

Por fim, o último método aqui utilizado é *Support Vector Classifier*, ou Classificadores de Vetores de Suporte. As Máquinas de Vetores de Suporte foram propostas a partir da Teoria do Aprendizado Estatístico por Vapnik (1963).

Intuitivamente, as máquinas de vetores de suporte encontram o superplano que melhor divide os dados de um conjunto de dados maximizando a distância entre este superplano e cada uma das classes. Ou seja, dada uma lista $(x_1, y_1), ..., (x_n, y_n)$ onde x_i representa um ponto no espaço, e cada y_i representa uma classe binária.

Usando a formulação linear, queremos encontrar o hiperplano que maximize a distância entre ele e o conjunto de pontos $x_1, ..., x_n$. O vetor normal ao hiperplano pode ser escrito como:

$$w^T x - b = 0 \tag{3.26}$$

Se o nosso conjunto de dados for linearmente separável, podemos então definir outros dois hiperplanos: $w^T x - b = 1$ e $w^T x - b = -1$ que indicam a borda de cada uma das classes no

conjunto de dados. Isto é, qualquer ponto localizado em $w^T x - b \geq 1$ terá classe 1, e qualquer ponto localizado em $w^T x - b \leq -1$ terá classe -1.

Sendo assim, esse classificador pode ser descrito pelo problema de otimização de minimização da seguinte função:

$$y(w^T x - b) \geq 1 \tag{3.27}$$

Esse classificador é bastante eficiente caso o conjunto de dados seja separável. Existem outras funções que podem ser utilizadas para separar os dados, como sigmóide e polinômios.

CAPÍTULO

4

BASE DE DADOS

Neste capítulo, serão descritos os conjuntos de dados utilizados neste trabalho, bem como os processos de extração e transformação envolvidos.

4.1 Dados de Documentos Fiscais

Nesta seção, será apresentado o tratamento dos dados obtidos através da coleta de documentos fiscais.

4.1.1 Aquisição

Como descrito na Seção 2.2, os dados utilizados neste trabalho são obtidos através do armazém de dados da empresa parceira. Estes dados passam por um complexo processo de ingestão e tratamento descrito em detalhes no Apêndice A.

A seleção dos dados foi feita usando o armazém de dados, através da linguagem *Structured Query Language* (SQL). Os critérios de seleção utilizados para obter os dados do armazém de dados foram:

- Foram utilizados apenas dados de Notas Fiscais Eletrônicas, excluindo outros documentos fiscais
- Foram utilizados apenas documentos obtidos através do *webservice* da Sefaz
- Foram utilizados apenas dados de documentos emitidos entre 01 de janeiro de 2019 e 31 de outubro de 2020
- Foram utilizados apenas documentos contendo a assinatura digital do emitente
- Foram utilizados apenas documentos contendo o protocolo de processamento do *webservice* da Sefaz

- Foram utilizados apenas documentos cuja emissão foi autorizada pelo *webservice* da Sefaz
- Foram utilizados apenas documentos referentes a transações cujo Código Fiscal de Operações e Prestações (CFOP) se refere às movimentações entre empresas envolvendo comercialização de mercadorias.

Os campos considerados são descritos no Quadro 2. Os tipos de dados e suas peculiaridades são descritos no Apêndice B.

Quadro 2 – Campos de documentos fiscais considerados

Nome do campo	**Tipo de dado**	**Descrição**
ID	STRING	Chave de acesso da NFe (anonimizado)
dhEmi	DATETIME	Data e Hora de emissão do Documento Fiscal
emit	RECORD	Identificação do Destinatário
emit.CNPJ	STRING	Número do CNPJ do emitente (anonimizado)
dest	RECORD	Identificação do Destinatário
dest.CNPJ	STRING	Número do CNPJ do destinatário (anonimizado)
det	ARRAY	Dados dos detalhes da NF-e
det.prod	RECORD	Dados dos produtos e serviços da NF-e
det.prod._nItem	INTEGER	Número do item do NF
det.prod.VProd	FLOAT	Valor bruto do produto ou serviço
det.prod.CFOP	STRING	CFOP

Cada linha do conjunto de dados possui dados de uma transação entre duas empresas, identificada por uma chave de acesso única por transação, e que contém um ou mais produtos envolvidos com seus respectivos números de item, valor bruto, e CFOP.

Todos os dados obtidos a partir do armazém de dados foram anonimizados. Dados identificadores ou sensíveis foram removidos e reindexados de forma a manter o sigilo e privacidade de tais dados respeitando tanto a legislação atual quanto os contratos de prestação de serviços da empresa parceira.

4.1.2 Pré-processamento

Para cada documento foi criada uma nova coluna com o valor bruto agregado dos produtos transacionados, respeitando os critérios de CFOP e sendo eliminados os documentos de valor zero (o que é possível devido a inconsistências durante a emissão do documento).

Para remover eventuais problemas de completude nos dados, foram considerados apenas documentos de transações entre empresas recorrentes. Chamamos de **empresas recorrentes** aquelas que possuem pelo menos um documento na base de dados emitido para cada mês entre janeiro de 2019 e outubro de 2020.

Uma vez tratados os dados, estes foram agregados em três conjuntos adicionais de dados, agregados por:

- Unidade Federativa (UF)
- Seção de CNAE
- CNAE

Cada um dos três conjuntos de dados agrega informações das transações efetuadas durante um mês, de janeiro de 2019 a outubro de 2020, sendo agregada a informação do valor total bruto dos produtos transacionados.

Uma nova coluna foi criada contendo a quantidade de transações mensal entre cada entidade. Foi considerado que cada NF-e documenta uma única transação independente da quantidade de produtos envolvida.

4.2 Dados Abertos de Empresas Brasileiras

Nesta seção, será apresentado o tratamento dos dados obtidos através da base de dados abertos da Receita Federal.

4.2.1 Aquisição

Além dos dados descritos na Seção 4.1, foram utilizados também dados coletados das bases de dados abertas sobre o Cadastro Nacional de Pessoas Jurídicas da Receita Federal (SECRETARIA ESPECIAL DA RECEITA FEDERAL DO BRASIL, 2020).

Tais dados são divulgados conforme previsto no Artigo 5º da Constituição Federal (BRASIL, 1988):

> Art. 5º Todos são iguais perante a lei, sem distinção de qualquer natureza, garantindo-se aos brasileiros e aos estrangeiros residentes no País a inviolabilidade do direito à vida, à liberdade, à igualdade, à segurança e à propriedade, nos termos seguintes:
>
> ...
>
> XXXIII - todos têm direito a receber dos órgãos públicos informações de seu interesse particular, ou de interesse coletivo ou geral, que serão prestadas no prazo da lei, sob pena de responsabilidade, ressalvadas aquelas cujo sigilo seja imprescindível à segurança da sociedade e do Estado;
>
> ...

Regulado pela Lei de Acesso a Informação (BRASIL, 2011):

> Art. 8º É dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no

> âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas.
>
> ...
>
> § 3º Os sítios de que trata o § 2º deverão, na forma de regulamento, atender, entre outros, aos seguintes requisitos:
>
> ...
>
> III - possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina;
>
> ...

Esses dados são também tratados e estão disponíveis no armazém de dados da empresa parceira através do processo descrito no Apêndice A. De lá foram coletadas as seguintes informações a partir da base de dados completa, cujos campos são descritos no Quadro 2.

Quadro 3 – Campos do Cadastro Nacional de Pessoas Jurídicas considerados

Nome do campo	**Tipo de dado**	**Descrição**
CNPJ	STRING	Número de Inscrição no CNPJ
SituacaoCadastral	STRING	Código da Situação Cadastral
DataInicioAtividade	STRING	Data de Início da Atividade
CnaeFiscal	STRING	Código da Atividade Econômica
UF	STRING	Sigla da UF em que se encontra o estabelecimento
Municipio	STRING	Código do município de jurisdição onde se encontra o estabelecimento
Porte	STRING	Código do porte da empresa
OpcaoSimples	STRING	Indicador da existência da opção pelo simples
OpcaoMei	STRING	Indicador da existência da opção pelo MEI

Através do número de inscrição das empresas foi possível cruzar os dados obtidos nesta base de dados com a base de dados descrita na Seção 4.1, o que foi feito antes de todo o processo de anonimização citado nessa seção.

4.3 Análise Descritiva da Base de Dados

Nesta seção, será apresentada uma análise descritiva dos dados utilizados neste trabalho, descrevendo os dados coletados de documentos fiscais e comparando sua abrangência com os dados obtidos da base de dados aberta da Receita Federal.

4.3.1 Abrangência territorial

Nesta subseção, será apresentada a análise descritiva do ponto de vista territorial. Será descrita a participação das empresas citadas na base de dados em relação aos totais de cada divisão territorial brasileira a partir dos dados descritos na Seção 4.2.

O IBGE criou a Divisão Territorial Brasileira (DTB), uma codificação hierárquica de sete dígitos do território brasileiro onde:

- O primeiro dígito indica a região
- O segundo dígito indica a UF
- O terceiro e quarto dígito indicam a mesorregião
- O sexto dígito indica a microrregião
- O sétimo dígito indica o município

O Quadro 4 descreve as regiões brasileiras segundo esta divisão, que foi usada no estudo de abrangência territorial dos dados.

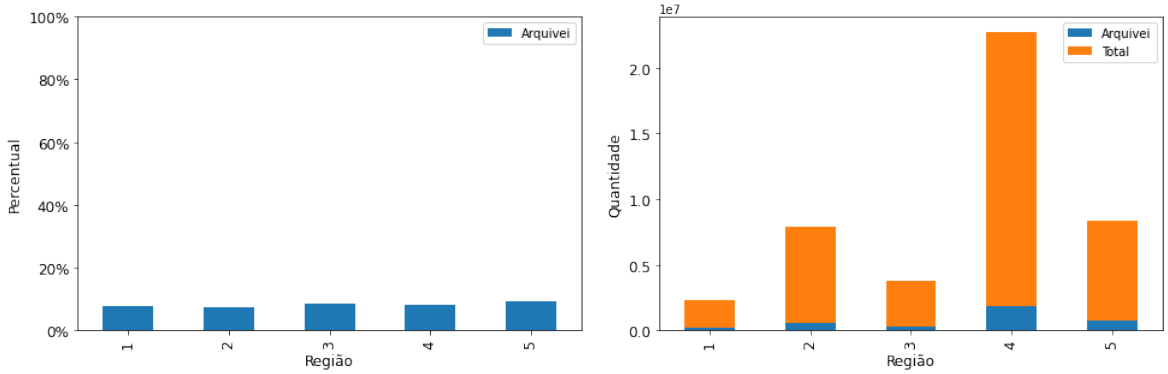
Quadro 4 – Códigos de Região segundo DTB

Código da Região	**Região**	**Unidade Federativa**
1	Região Norte	Acre Amapá Amazonas Pará Rondônia Roraima Tocantins
2	Região Nordeste	Alagoas Bahia Ceará Maranhão Paraíba Pernambuco Piauí Rio Grande do Norte Sergipe
3	Região Centro-Oeste	Distrito Federal Goiás Mato Grosso Mato Grosso do Sul
4	Região Sudeste	Espírito Santo Minas Gerais São Paulo Rio de Janeiro
5	Região Sul	Paraná Rio Grande do Sul Santa Catarina

Na Figura 6a é descrito o percentual do total das empresas de cada região que foi encontrada na base de dados deste trabalho. Os dados são descritos em detalhe na Tabela 1. A Figura 6b descreve as quantidades de empresas presentes na base de dados em relação ao total de cada região brasileira.

Com esses dados é possível observar que a participação em termos regionais é significativa e que há um razoável balanceamento regional, com uma participação regional que varia entre 7 e 10%, mas sem um grande viés para uma das regiões. A quantidade de empresas da Região 4 é significativamente maior em comparação à outras regiões, mas reflete proporcionalmente a realidade se verificarmos os valores percentuais.

Figura 6 – Participação por região das empresas presentes na base de dados



(a) Percentual das empresas de cada região brasileira presente na base de dados

(b) Quantidade de empresas de cada região brasileira presente na base de dados relativa ao total

Fonte: Dados da pesquisa.

Tabela 1 – Percentual das empresas presentes na base de dados por Região

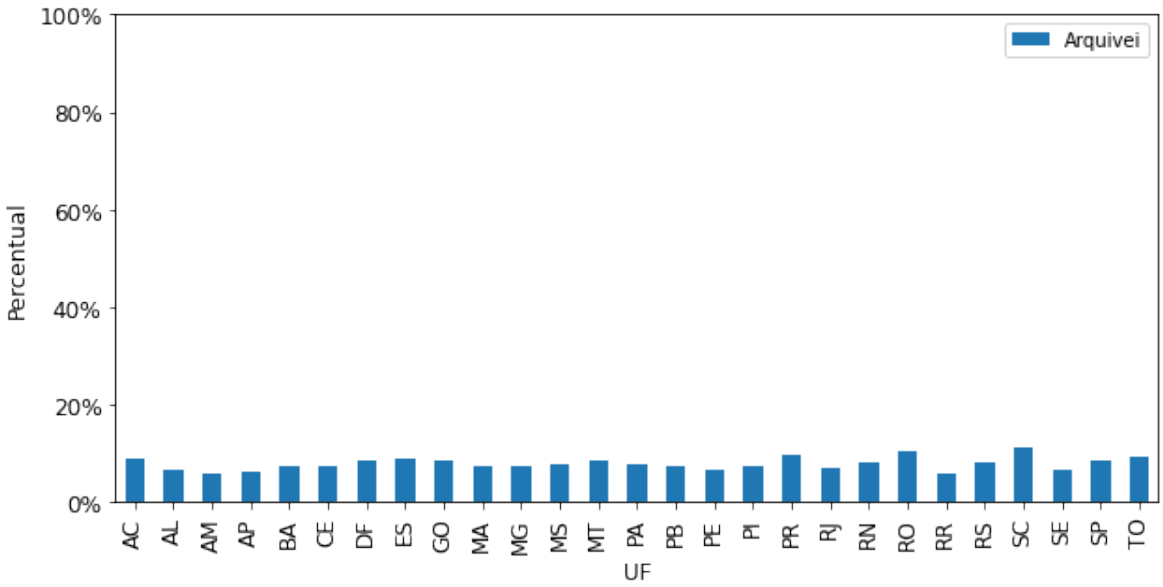
Região	Participação
1	7.88%
2	7.26%
3	8.41%
4	8.00%
5	9.39%

Fonte: Dados da pesquisa.

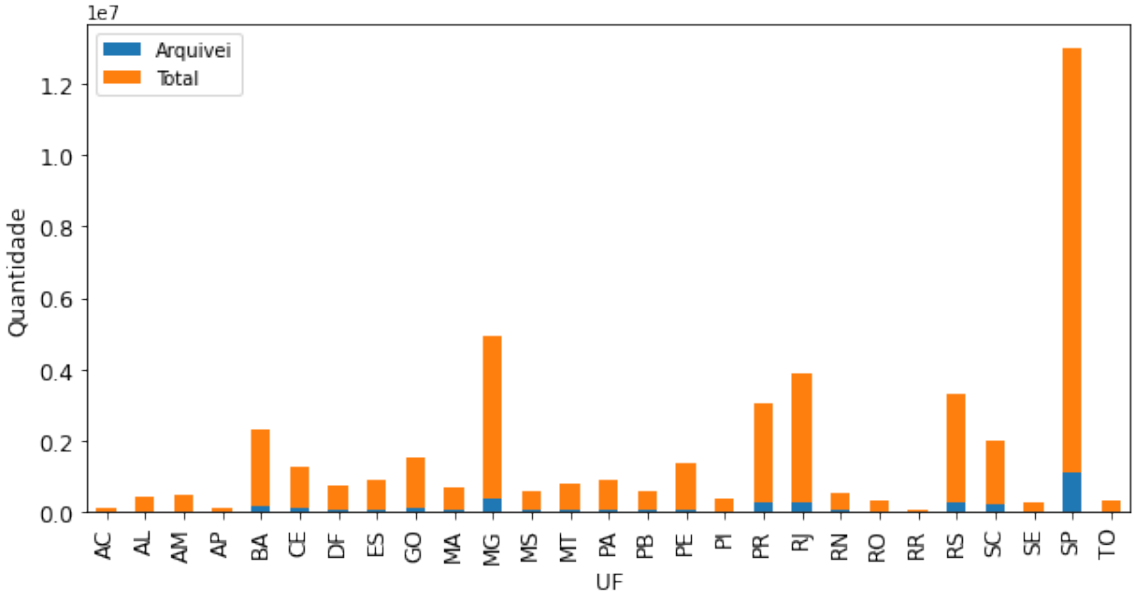
A seguir, repetimos a análise agora para as Unidades Federativas do Brasil, ilustrada em participação percentual na Figura 7a e na Tabela 2, e em quantidades em relação à quantidade total de cada UF na Figura 7b.

Da mesma forma, é possível observar que a participação por unidade federativa é significativa e que embora haja um desbalanceamento um pouco maior, com uma participação por UF que varia entre 6 e 12%, também não há um grande viés percentual para uma das UFs, refletindo razoavelmente a realidade.

Figura 7 – Participação por UF das empresas presentes na base de dados



(a) Percentual das empresas das cada UF presente na base de dados



(b) Quantidade de empresas de cada UF presente na base de dados relativa ao total

Fonte: Dados da pesquisa.

Por fim, a Figura 8 mostra um histograma relacionando o percentual do total de empresas presentes na base de dados em relação aos municípios brasileiros. A Tabela 3 mostra algumas métricas sobre esta análise por municípios, mostrando que existe participação em todos os municípios brasileiros com uma média e mediana próximos de 7.5% e pico de 21%.

Assim, pode-se observar uma forte participação territorial dos dados tanto observando do ponto de vista regional como também estadual e municipal, estando presentes nas mais diversas

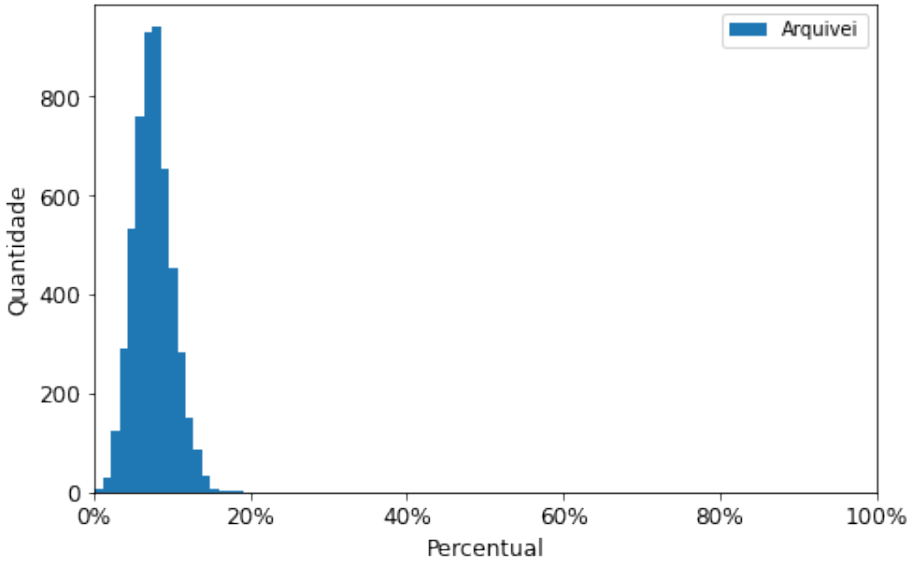
Tabela 2 – Percentual das empresas presentes na base de dados por UF

UF	Participação
AC	8.93%
AL	6.75%
AM	5.76%
AP	6.25%
BA	7.34%
CE	7.46%
DF	8.48%
ES	8.95%
GO	8.64%
MA	7.56%
MG	7.41%
MS	7.67%
MT	8.44%
PA	7.91%

UF	Participação
PB	7.43%
PE	6.68%
PI	7.49%
PR	9.58%
RJ	7.14%
RN	8.03%
RO	0.34%
RR	6.04%
RS	8.16%
SC	1.15%
SE	6.45%
SP	8.41%
TO	9.13%

Fonte: Dados da pesquisa.

Figura 8 – Histograma do percentual das empresas de cada município brasileiro presente na base de dados



Fonte: Dados da pesquisa.

realidades brasileiras.

4.3.2 Abrangência econômica

Os dados descritos na Seção 4.2 apresentam uma série de informações e classificações das empresas que podem ser utilizadas para ter uma melhor noção do porte e dos tipos de empresas envolvidas. Nesta subseção, será apresentada a análise descritiva do ponto de vista

Tabela 3 – Métricas sobre o percentual das empresas presentes na base de dados por município

Métrica	Valor
Média	7.522%
Desvio Padrão	2.462%
Mínimo	0.002%
25º percentil	5.834%
50º percentil	7.419%
75º percentil	9.065%
Máximo	21.230%

Fonte: Dados da pesquisa.

econômico em relação aos totais coletados daquela base.

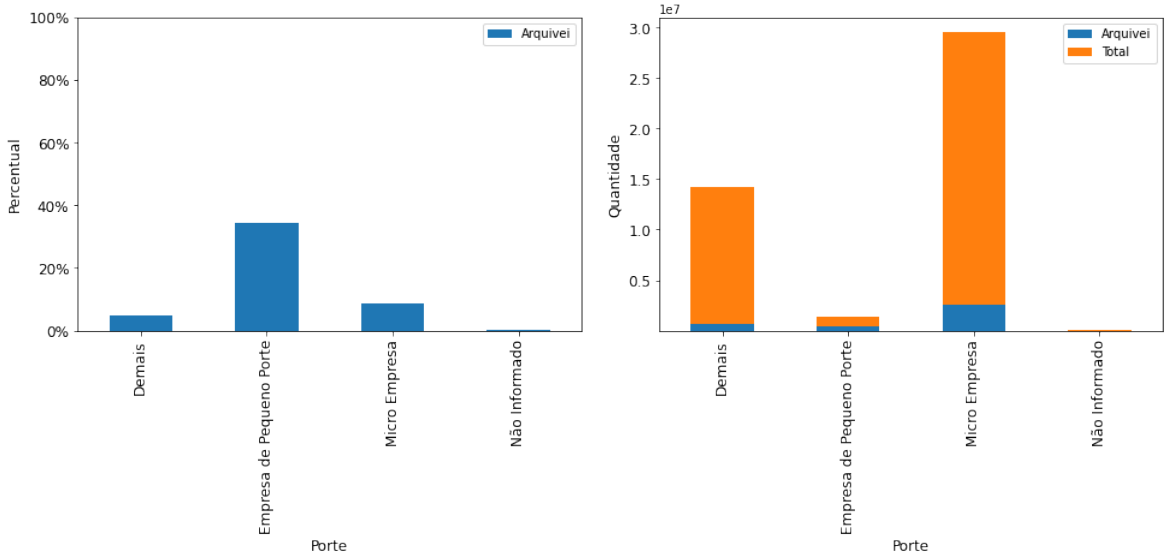
A primeira classificação que vamos considerar para esta análise é o **porte** de uma empresa. Regulada pela Lei nº 5172/1966 (BRASIL, 1966), que em seu Artigo 3º define:

> Art. 3º Para os efeitos desta Lei Complementar, consideram-se **microempresas** ou **empresas de pequeno porte**, a sociedade empresária, a sociedade simples, a empresa individual de responsabilidade limitada e o empresário a que se refere o art. 966 da Lei no 10.406, de 10 de janeiro de 2002 (Código Civil), devidamente registrados no Registro de Empresas Mercantis ou no Registro Civil de Pessoas Jurídicas, conforme o caso, desde que:
>
> I - no caso da microempresa, aufira, em cada ano-calendário, receita bruta igual ou inferior a R$ 360.000,00 (trezentos e sessenta mil reais); e
>
> II - no caso de empresa de pequeno porte, aufira, em cada ano-calendário, receita bruta superior a R$ 360.000,00 (trezentos e sessenta mil reais) e igual ou inferior a R$ 4.800.000,00 (quatro milhões e oitocentos mil reais).
>
> § 1º Considera-se receita bruta, para fins do disposto no caput deste artigo, o produto da venda de bens e serviços nas operações de conta própria, o preço dos serviços prestados e o resultado nas operações em conta alheia, não incluídas as vendas canceladas e os descontos incondicionais concedidos.
>
> (...)

As empresas declaram seu porte no momento da sua inscrição no Cadastro Nacional de Pessoas Jurídicas, e esta declaração é revisada pela Receita Federal a partir de suas declarações fiscais anuais. Empresas cuja receita bruta não se enquadra nas categorias Microempresa (ME) e Empresa de Pequeno Porte (EPP) são enquadradas na categoria **Demais**. Em alguns casos específicos, as empresas não declaram sua receita bruta no ato da inscrição no cadastro, em sua maioria empresas estrangeiras presentes no cadastro, essas são enquadradas na categoria **Não Informado**.

A partir dos dados das Figuras 9a e 9b é possível verificar a participação das empresas presentes na base de dados em relação aos totais. A Tabela 4 descreve os valores em detalhes.

Figura 9 – Participação por porte das empresas presentes na base de dados



(a) Percentual das empresas presente na base de dados categorizadas por porte

(b) Quantidade de empresas categorizadas por porte presente na base de dados relativa ao total

Fonte: Dados da pesquisa.

Tabela 4 – Métricas sobre o percentual das empresas presentes na base de dados por porte

Porte	Participação
Demais	4.76%
Empresa de Pequeno Porte	34.46%
Micro Empresa	8.59%
Não Informado	0.05%

Fonte: Dados da pesquisa.

É possível observar que há uma participação significativa desproporcional maior de EPPs nas empresas presentes da base de dados em relação à base de comparação. A participação é significativa também nos segmentos de Micro Empresas e Demais. Dentre as empresas que não informaram o porte a participação é proporcionalmente bastante baixa e suas quantidades totais na base de comparação são também muito pequenas.

A legislação brasileira define também a figura do Microempreendedor Individual (MEI) através do Código Civil Brasileiro (BRASIL, 2002) que define a figura do empresário como:

> Art. 966. Considera-se empresário quem exerce profissionalmente atividade econômica organizada para a produção ou a circulação de bens ou de serviços.

> Parágrafo único. Não se considera empresário quem exerce profissão intelectual, de natureza científica, literária ou artística, ainda com o concurso de auxiliares ou colaboradores, salvo se o exercício da profissão constituir elemento de empresa.

E da Lei nº 128/2008 (BRASIL, 2008) que estende essa definição para a figura do MEI:

> Art. 18-A. O **Microempreendedor Individual - MEI** poderá optar pelo recolhimento dos impostos e contribuições abrangidos pelo Simples Nacional em valores fixos mensais, independentemente da receita bruta por ele auferida no mês, na forma prevista neste artigo.
>
> § 1o Para os efeitos desta Lei, considera-se MEI o empresário individual a que se refere o art. 966 da Lei no 10.406, de 10 de janeiro de 2002 – Código Civil, que tenha auferido receita bruta, no ano-calendário anterior, de até R$ 36.000,00 (trinta e seis mil reais), optante pelo Simples Nacional e que não esteja impedido de optar pela sistemática prevista neste artigo.
>
> § 2o No caso de início de atividades, o limite de que trata o § 1o deste artigo será de R$ 3.000,00 (três mil reais) multiplicados pelo número de meses compreendido entre o início da atividade e o final do respectivo ano-calendário, consideradas as frações de meses como um mês inteiro.

Empresários que optam pelo regime MEI tem uma série de facilidades fiscais que visam incentivar pequenos empresários, o que pode ser um fator importante para a sobrevivência do mesmo durante uma crise.

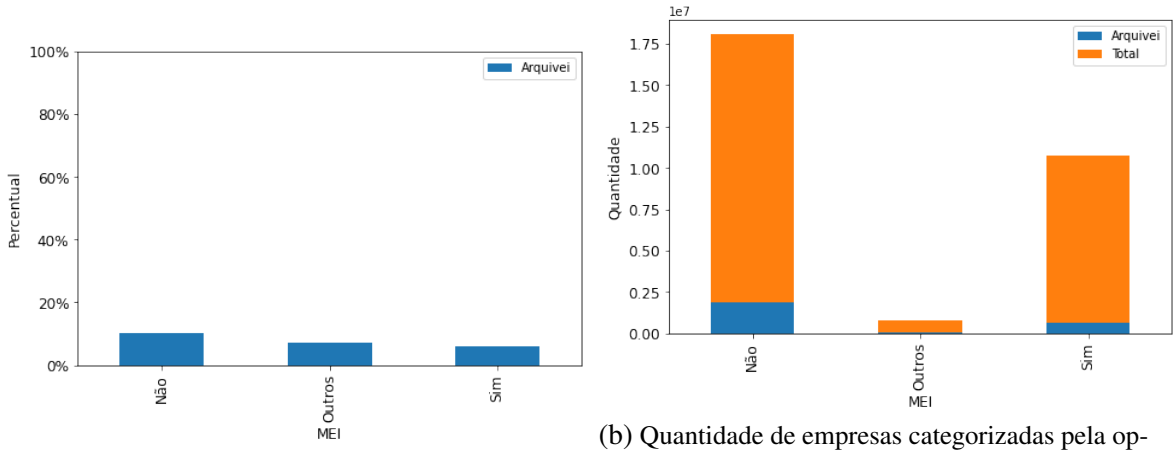
A partir dos dados sobre a participação por porte e da Opção pelo MEI, podemos verificar qual o percentual de participação nesta faixa econômica. Nesta análise, foram consideradas apenas as empresas que já declaram ter porte de Micro Empresa, um requisito devido à receita bruta máxima prevista na legislação, e então verificamos quais os percentuais presentes na base de dados.

A Figura 10a e 10b, cujos dados são detalhados na Tabela 5 mostram os dados de participação em relação à Opção pelo MEI. A microempresa tem a opção de optar ou não por este enquadramento no momento de sua abertura, mas existem casos em que não foi feita a opção, seja por omissão no momento do cadastro ou por ser anterior à lei que define o MEI. Casos onde não foi feita a opção são descritos nos dados como **Outros**.

É possível observar que nos dados coletados da empresa parceira existe algum viés para microempreendedores que não optam pelo MEI, seja explicitamente ou devido a outros fatores.

A Lei nº 123/2006 (BRASIL, 2006a) institui o Simples Nacional, um regime tributário que facilita o recolhimentos de impostos para MEIs e EPPs. A base de dados abertos da Receita Federal também possui a informação da opção ou não pelo regime Simples Nacional. As empresas podem solicitar a exclusão do regime Simples Nacional mediante apresentação de

Figura 10 – Participação por opção pelo MEI das empresas presentes na base de dados



(a) Percentual das empresas presente na base de dados categorizadas pela opção por MEI

(b) Quantidade de empresas categorizadas pela opção por MEI presente na base de dados relativa ao total

Fonte: Dados da pesquisa.

Tabela 5 – Métricas sobre o percentual das empresas presentes na base de dados pela opção pelo MEI

Opção pelo MEI	Participação
Não	10.20%
Outros	7.03%
Sim	5.98%

Fonte: Dados da pesquisa.

ofício, ou em alguns casos previstos na lei onde a empresa deixa de cumprir alguns requisitos ou obrigações, nestes casos a empresa consta na base de dados como Excluído.

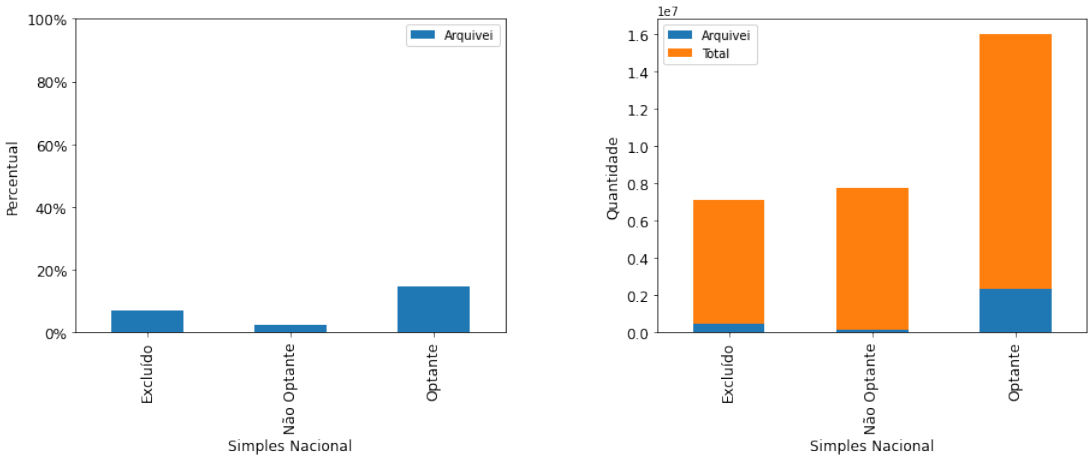
As Figuras 11 e 11b, detalhados na Tabela 6, mostram a relação entre os dados presentes na base de dados em relação aos totais coletados da Receita Federal. É possível verificar que na base de dados é mais comum a presença de empresas optantes pelo regime Simples Nacional, e assim como em relação à opção ou não pelo MEI isso pode ser um fator de resiliência ou não à crises.

Tabela 6 – Métricas sobre o percentual das empresas presentes na base de dados pela Opção pelo Simples Nacional

Opção pelo Simples Nacional	Participação
Excluído	0.0693
Não Optante	0.0229
Optante	0.1450

Fonte: Dados da pesquisa.

Figura 11 – Participação por região das empresas presentes na base de dados



(a) Percentual das empresas presente na base de dados categorizadas pela opção pelo Simples Nacional

(b) Quantidade de empresas categorizadas pela opção pelo Simples Nacional presente na base de dados relativa ao total

Fonte: Dados da pesquisa.

4.3.3 Abrangência setorial

Nesta Seção, será feita a análise comparativa dos dados presentes na base de dados de documentos fiscais em relação aos dados abertos descritos na Seção 4.2 visando descrever a participação em relação aos setores econômicos da economia brasileira.

Como citado no Capítulo 2, através dos protocolos do ENAT foi feito o compromisso para a unificação do Cadastro Nacional de Atividades Econômicas (CNAE). Esse cadastro foi implementado então pela Resolução IBGE/CONCLA nº 01 de 04 de setembro de 2006 (BRASIL, 2006b) e Resolução IBGE/CONCLA nº nº 02, de 15 de dezembro de 2006 (BRASIL, 2006c). Como descrito no Art. 1º da Resolução nº 01, essa classificação é hierárquica:

> O PRESIDENTE DA COMISSÃO NACIONAL DE CLASSIFICAÇÃO – CONCLA, no uso de suas atribuições, resolve:
>
> Art. 1 o Aprovar e divulgar a estrutura completa da Classificação Nacional de Atividades Econômicas – CNAE – versão 2.0, organizada em cinco níveis hierárquicos: seções, divisões, grupos, classes e subclasses, sendo o detalhamento das subclasses destinado ao uso da Administração Pública Brasileira.
>
> (...)

O anexo de tal resolução contém uma tabela com o detalhamento dessa classificação, posteriormente retificada.

As empresas informam o código de sua CNAE Fiscal no momento da sua inscrição no Cadastro Nacional de Pessoas Jurídicas, e este dado fica disponível para consulta pública através

de sua base de dados aberta. São informadas também os códigos de suas CNAEs secundários referentes a atividades secundárias das empresas, estes não foram utilizados nesta análise.

O Quadro 5 descreve a que se refere cada seção segundo o IBGE/CONCLA.

Quadro 5 – Denominação das Seções de CNAE

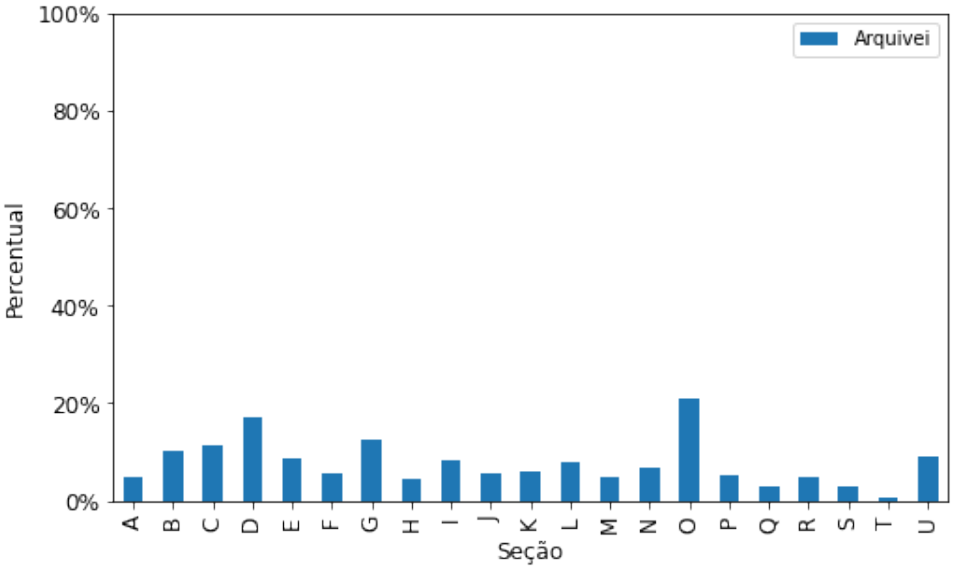
Seção	Faixa de Divisões	Denominação
A	01 a 03	Agricultura, pecuária, produção florestal, pesca e aquicultura
B	05 a 09	Indústrias extrativas
C	10 a 33	Indústrias de transformação
D	35	Eletricidade e gás
E	36 a 39	Água, esgoto, atividades de gestão de resíduos e descontaminação
F	41 a 43	Construção
G	45 a 47	Comércio; Reparação de veículos automotores e motocicletas
H	49 a 53	Transporte, armazenagem e correio
I	55 a 56	Alojamento e alimentação
J	58 a 63	Informação e comunicação
K	64 a 66	Atividades financeiras, de seguros e serviços relacionados
L	68	Atividades imobiliárias
M	69 a 75	Atividades profissionais, científicas e técnicas
N	77 a 82	Atividades administrativas e serviços complementares
O	84	Administração pública, defesa e seguridade social
P	85	Educação
Q	86 a 88	Saúde humana e serviços sociais
R	90 a 93	Artes, cultura e serviços sociais
S	94 a 96	Outras atividades de serviços
T	97	Serviços Domésticos
U	99	Organismos internacionais e outras instituições extraterritoriais

As Figuras 12a e 12b descrevem a participação percentual por seção econômica e suas quantidades em relação aos totais. A Tabela 7 descreve a participação percentual em detalhes.

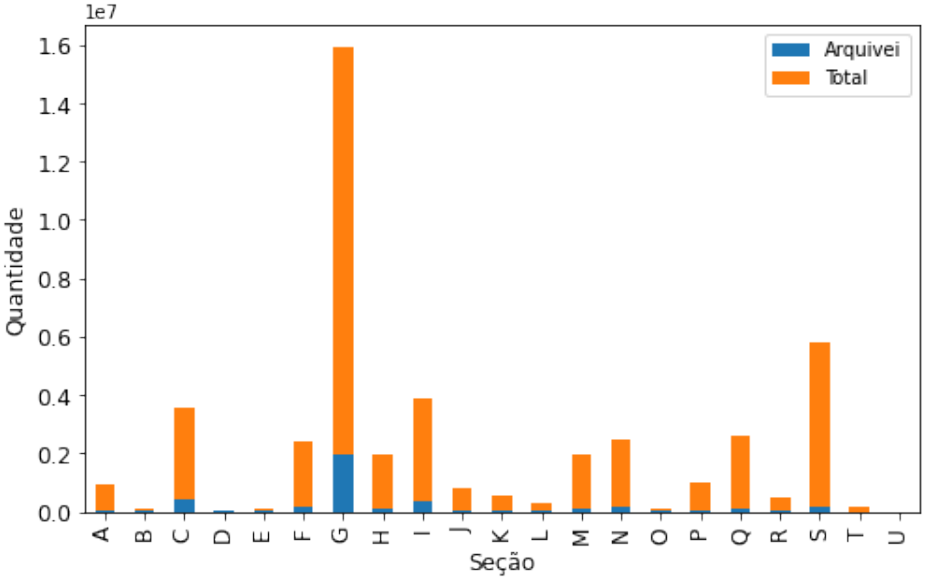
Através destes dados é possível notar que existe uma participação significativa em praticamente todas as seções, com exceção da seção T, de serviços domésticos. Os dados também possuem razoável viés para alguns setores que possuem altos percentuais, particularmente D, G, e O, em oposição aos outros. A quantidade de empresas de cada seção também varia significativamente, particularmente a seção G que possui um número bastante alto de empresas, o que não necessariamente implica maior quantidade de documentos fiscais ou valor transacionado, análise que será feita posteriormente.

A Figura 13 ilustra a participação percentual por CNAE, e a Tabela 8 nos dá algumas métricas acerca da participação por CNAE. Dos 1342 CNAEs existentes na base de dados de comparação, 1327 são citados na base de dados de documentos fsicais (98.88%), com uma média de cerca de 15% de participação e desvio padrão de cerca de 12%.

Figura 12 – Participação por seção do CNAE das empresas presentes na base de dados



(a) Percentual das empresas presente na base de dados categorizadas pela seção do CNAE



(b) Quantidade de empresas categorizadas seção do CNAE presente na base de dados relativa ao total

Fonte: Dados da pesquisa.

É possível observar então que que há uma grande abrangência setorial, uma vez que a base de dados possui exemples da grande maioria das seções, mas o histograma mostra que existe uma grande quantidade de CNAEs com baixo percentual de participação.

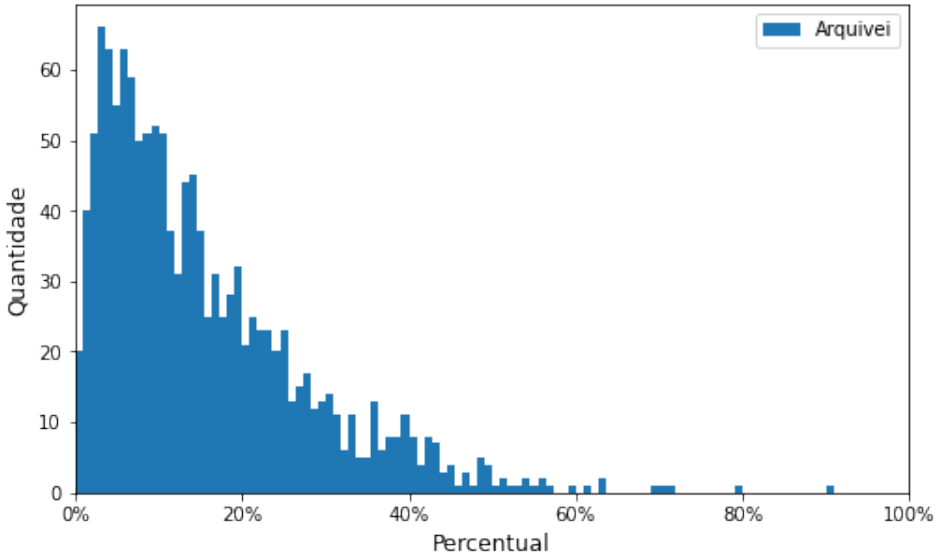
Tabela 7 – Métricas sobre o percentual das empresas presentes na base de dados por Seção do CNAE

Seção	Participação
A	4.74%
B	10.31%
C	11.30%
D	17.17%
E	8.69%
F	5.66%
G	12.34%
H	4.36%
I	8.40%
J	5.41%
K	5.91%

Seção	Participação
L	7.87%
M	4.71%
N	6.66%
O	20.72%
P	5.22%
Q	3.01%
R	4.75%
S	2.82%
T	0.71%
U	9.07%

Fonte: Dados da pesquisa.

Figura 13 – Histograma do percentual das empresas de cada CNAE presente na base de dados



Fonte: Dados da pesquisa.

4.4 Conclusões

Neste capítulo, pudemos observar como os dados da base de dados de documentos fiscais utilizada neste trabalho se comparam à base de dados aberta de empresas brasileiras. Do ponto de vista territorial, pudemos observar que há uma distribuição consideravelmente balanceada nos dados, não havendo grande preferência regional. Na análise de abrangência econômica, foi possível ver que existe certo viés para empresas de menor porte, com uma participação maior de MEs e EPPs em relação às Demais, o que reflete também a participação no Simples Nacional, uma tipificação típica de empresas de menor porte. Por fim, analisando a participação da base de dados em relação às seções da economia brasileira, pudemos observar que existe alguma

Tabela 8 – Métricas sobre o percentual das empresas presentes na base de dados por CNAE

Métrica	Valor
Média	15.371%
Desvio Padrão	12.452%
Mínimo	0.004%
25º percentil	5.984%
50º percentil	12.056%
75º percentil	21.570%
Máximo	90.909%

Fonte: Dados da pesquisa.

variância entre as diversas seções de CNAE, e também entre todos os CNAEs.

Com isso, concluímos que os dados possuem algum viés e não são perfeitamente balanceados conforme a totalidade dos dados. Além disso, esta análise possui suas limitações, uma vez que não sabemos exatamente o volume de documentos nem de valores transacionados para cada tipo de empresa. O objetivo dessa análise foi tornar mais explícito a perspectiva que os dados apresentam acerca da totalidade dos dados da economia brasileira. Sob essa perspectiva faremos às análises seguintes, que são aplicadas sobre uma amostra dos dados totais, que pudemos ver que é significativa e razoavelmente balanceada segundo as segmentações que utilizamos.

CAPÍTULO

5

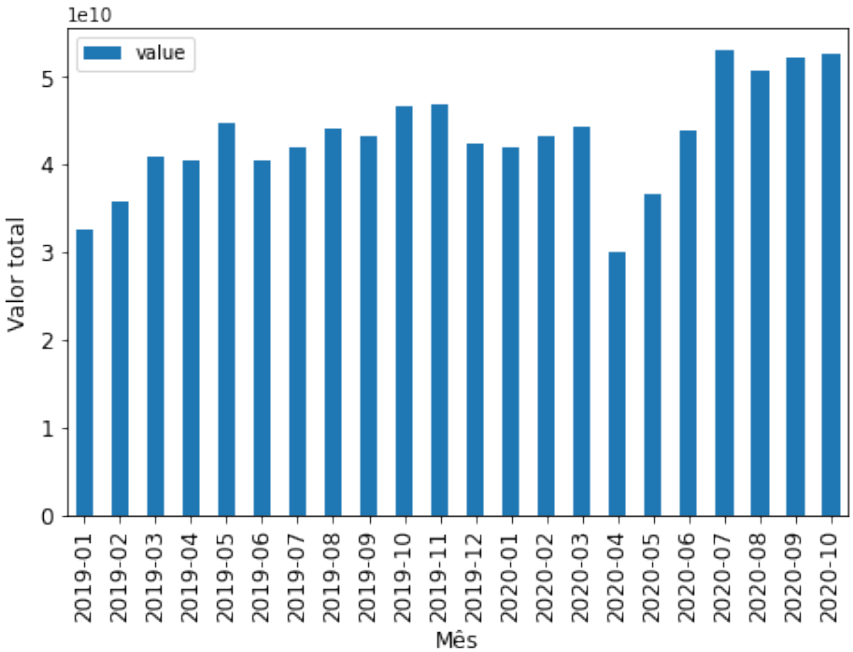
IMPACTO ECONÔMICO DA PANDEMIA

Neste seção, será descrito o possível impacto econômico da Pandemia de COVID-19, contextualizada na Seção 1.2, verificando mensalmente se o volume transacionado foi próximo ou não no volume esperado.

5.1 Cenário geral

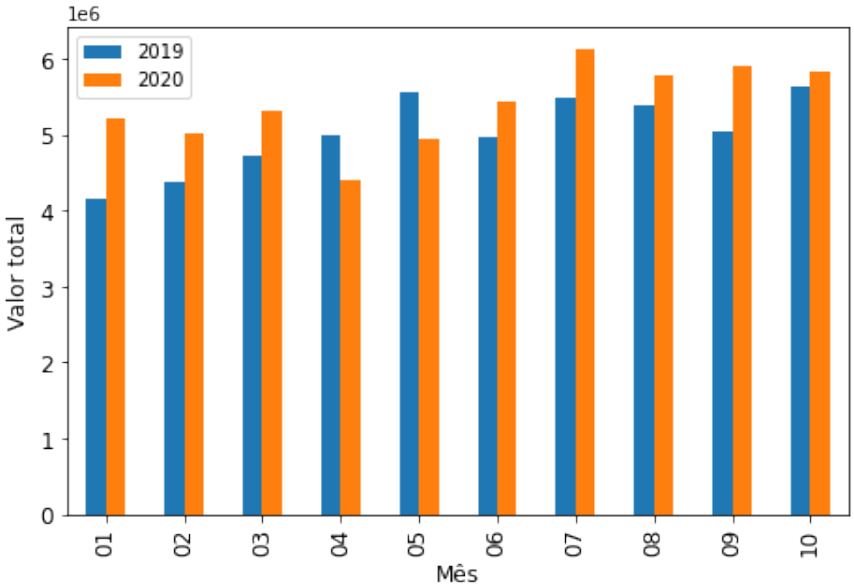
Nessa seção, descrevemos os dados da base totalizados mensalmente de forma agregada.

Figura 14 – Valor total transacionado por mês no período analisado



Fonte: Dados da pesquisa.

Figura 15 – Comparação do valor total transacionado por mês entre 2019 e 2020 (Parte 4)



Fonte: Dados da pesquisa.

A Figura 14 mostra o valor total transacionado em cada mês entre janeiro de 2019 e outubro de 2020, onde é visível um menor valor transacionado principalmente em abril e maio de 2020.

Tabela 9 – Variação entre o valor transacionado para cada mês em 2020 em relação a 2019

Mês	Variação
Janeiro	25.49%
Fevereiro	14.58%
Março	12.71%
Abril	-12.24%
Maio	-11.24%
Junho	9.20%
Julho	11.64%
Agosto	7.52%
Setembro	17.02%
Outubro	3.33%

Fonte: Dados da pesquisa.

A Figura 15 mostra uma comparação dos valores mensais transacionados em 2020 em relação aos valores mensais de 2019. Verifica-se então que os meses de abril e maio tiveram um valor total transacionado menor no ano de 2020, cuja variação por UF está descritos em detalhes na Tabela 9.

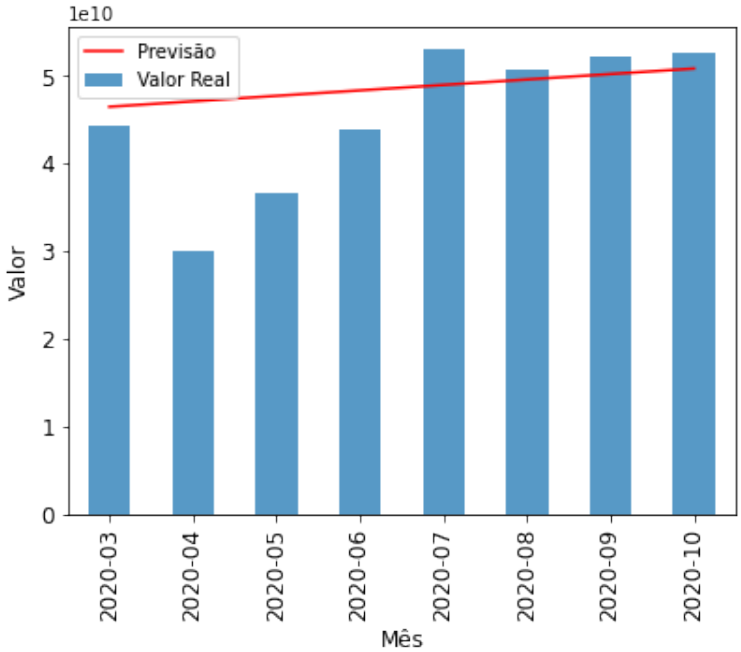
Como descrito na Seção 1.2, os meses de março e abril marcam exatamente o início do isolamento e das quarentenas no Brasil, e a aparentemente o impacto econômico nesses

momentos foi mais importante.

Aplicando uma regressão linear a partir dos dados entre janeiro de 2019 e fevereiro de 2020 é possível obter uma da previsão de valor transacionado para os meses seguintes e então comparar com o valor obtido nos meses seguintes. Uma regressão linear não necessariamente é a melhor forma de prever o valor a ser transacionado para determinado mês, dados que os dados podem ter sazonalidades, porém se observarmos os dados da Figura 14, podemos verificar uma tendência de crescimento linear mês a mês para os meses do ano de 2019. Isso não necessariamente ocorre porque houve um aumento real de transações das empresas envolvidas, mas porque a base de dados da empresa parceira tem crescido expressivamente mês a mês, com isso mesmo considerando o critério de recorrência, citado na seção 4.1.2, a base de dados passa a receber mais dados de determinadas empresas, notoriamente aquelas que não são clientes da empresa parceira.

A Figura 16 mostra os valores obtidos em relação aos valores previstos a partir da regressão linear aplicada sobre os dados, cuja variação é detalhada na na Tabela 10.

Figura 16 – Valor total transacionado por mês comparado com a previsão para o período analisado



Fonte: Dados da pesquisa.

A partir desses dados, é possível verificar que no período entre março e junho de 2020 o valor transacionado total foi menor que o previsto, verificando uma retomada posterior. Notoriamente, no mês de abril houve uma diferença de cerca de 36% entre o valor previsto e o valor real. Dada essa leitura, vamos considerar posteriormente para efeitos de comparação que o segundo trimestre de 2020 foi o mais afetado economicamente pela pandemia, mesmo que no mês de março haja também impacto não o consideraremos pois as medidas de isolamento

Tabela 10 – Variação entre o valor transacionado para cada mês em relação ao previsto após o início da pandemia

Mês	Variação
Março	-4.46%
Abril	-36.22%
Maio	-23.11%
Junho	-9.15%
Julho	8.28%
Agosto	2.09%
Setembro	4.18%
Outubro	3.46%
Total	-6.58%

Fonte: Dados da pesquisa.

e distanciamento foram aplicadas apenas ao fim deste mês e o impacto não foi suficiente para alterar significativamente os totais do primeiro trimestre de 2020.

A partir de julho de 2020, o valor real transacionado foi maior que o valor previsto para esses meses. Isso pode indicar um menor impacto da pandemia para esses meses, e podemos observar na seção 1.2 a partir da análise do índice de isolamento social que o isolamento social apesar de uma alta inicial em março e abril de 2020, começou a regredir a partir de julho de 2020, se mantendo estável após isso. Com isso, os níveis de consumo e produção possivelmente foram impactados positivamente com uma retomada, uma hipótese que ainda precisa ser mais lapidada uma vez que estamos analisando um período ainda limitado.

5.2 Cenário regional

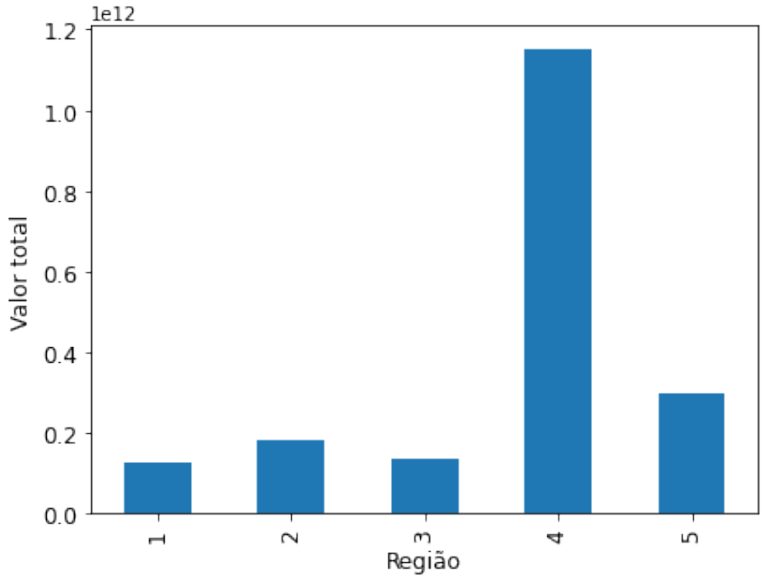
Nesta seção, descrevemos o impacto econômico da pandemia do ponto de vista regional. Usamos aqui a definição de região utilizada na seção 4.3.1. A Figura 17 mostra o valor total transacionado por região entre janeiro de 2019 e outubro de 2020 na base de dados utilizada neste trabalho.

A região 4 (sudeste) possui um valor transacionado muito maior em comparação às outras regiões. Segundo comparação feita na seção 4.3.1, não há uma preferência territorial clara para empresas da região sudeste possuindo um percentual de participação semelhante à de outras regiões. Esta região é conhecida por sua alta participação no cenário econômico brasileiro.

Na Figura 18 é feita uma comparação trimestral entre o valor transacionado em cada um dos três primeiros trimestres de 2020 para cada região.

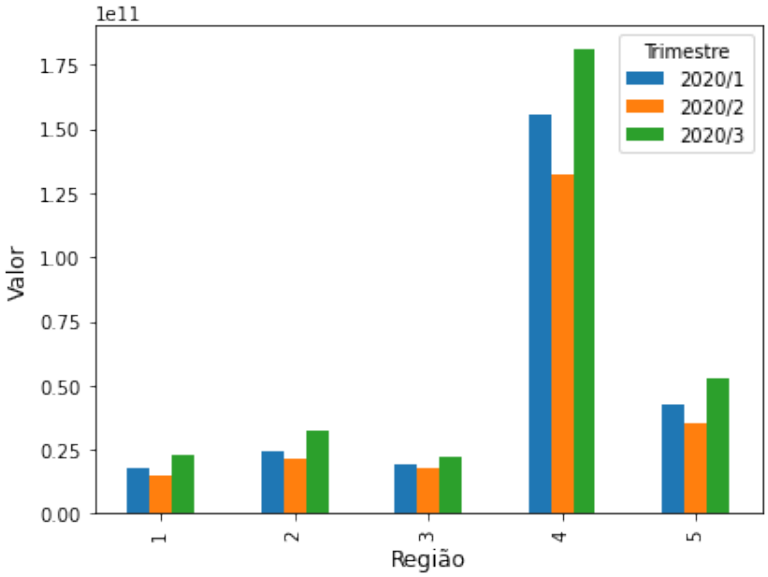
Pode-se observar que todas as regiões tiveram um valor transacionado menor no segundo trimestre de 2020 em comparação aos outros dois trimestres analisados, o que indica um impacto

Figura 17 – Valor total transacionado por região no período analisado



Fonte: Dados da pesquisa.

Figura 18 – Valor trimestral transacionado por região no período analisado



Fonte: Dados da pesquisa.

econômico devido à crise causada pela pandemia, uma vez que os dados possuem uma tendência de crescimento no valor total transacionado, como argumentamos na seção 5.1.

A Tabela 11 descreve a diferença entre os valores previstos a partir de uma regressão linear aplicada sobre os dados de janeiro de 2019 a fevereiro de 2020 e àqueles obtidos para cada região entre março e outubro de 2020, da mesma forma como descrito na seção 5.1. Também é apresentada na mesma tabela a variação entre o valor total previto e obtido entre os meses de março a outubro de 2020.

Tabela 11 – Diferença entre valores transacionados previstos e obtidos por região no período da pandemia

Região	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro
1	-13.4%	-48.8%	-26.1%	-9.9%	8.9%	1.6%	0.7%	6.4%
2	-9.2%	-39.6%	-22.3%	-5.3%	15.2%	10.1%	12.9%	18.9%
3	8.2%	-27.7%	-10.7%	3.5%	9.6%	6.5%	8.4%	1.9%
4	-3.0%	-34.9%	-23.8%	-10.2%	3.0%	1.5%	3.4%	2.5%
5	-8.5%	-37.3%	-24.9%	-12.6%	22.5%	-2.3%	1.6%	-2.9%

(a) Diferença entre valor mensal previsto e obtido

Região	Variação total
1	-9.4%
2	-1.9%
3	0.1%
4	-7.5%
5	-7.7%

(b) Diferença entre valor total previsto e obtido

Fonte: Dados da pesquisa.

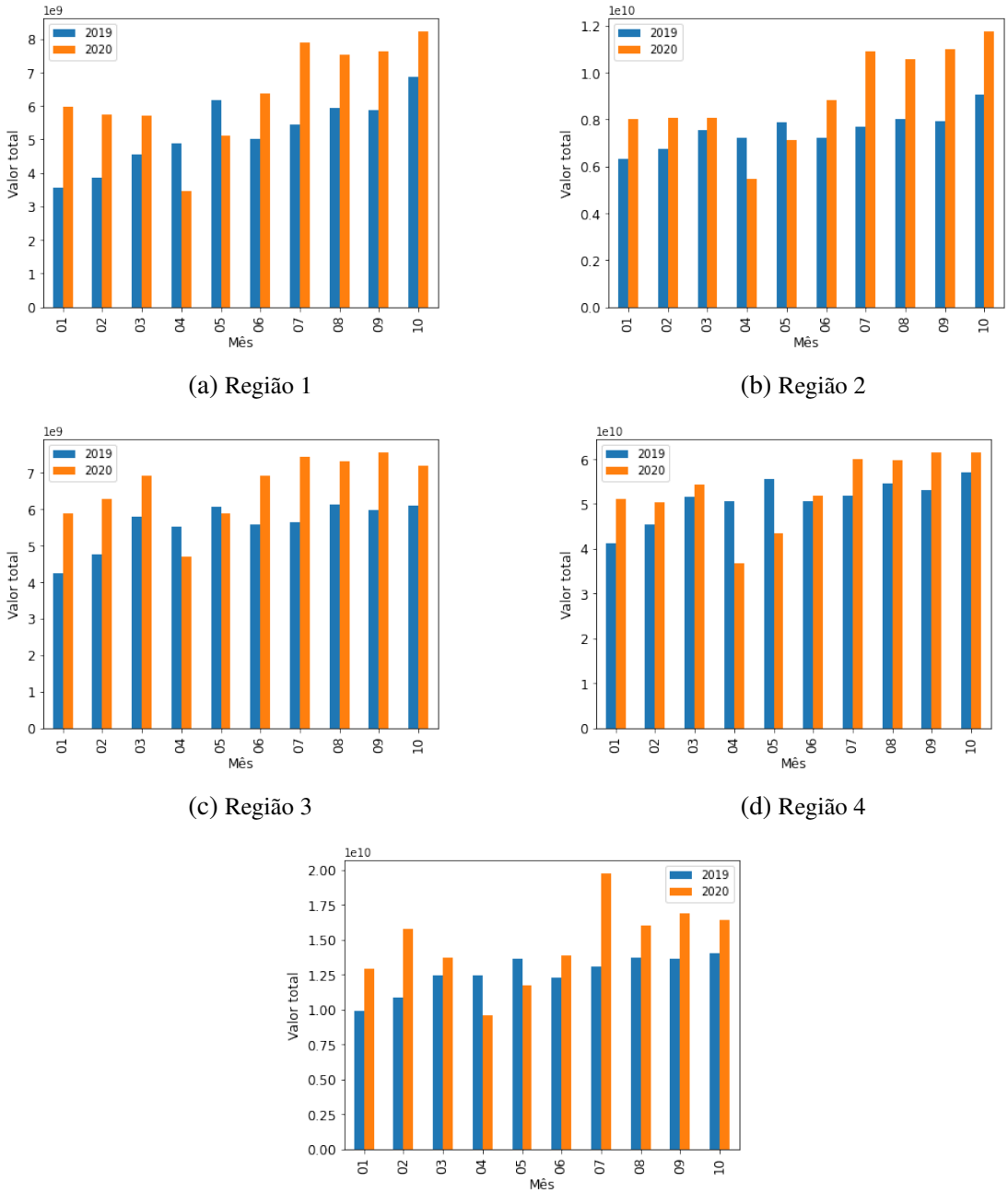
Verfificamos novamente que os meses mais impactados foram aqueles relativos ao segundo trimestre de 2020, de abril a junho. A diferença para o mês de abril chegou a ser de cerca de 48% para a região 1 (norte), o que representa um impacto bastante significativo. É possível observar também cerca resiliência da região 3 (centro-oeste), sendo que essa região não apontou impacto no mês de março e foi capaz de se recuperar mais rapidamente já apresentando resultados positivos para o mês de junho. Com exceção da região 3, todas as regiões tiveram uma diferença entre o valor total previsto e o obtido para o período de março a outubro de 2020.

A Figura 19 apresenta mês a mês o valor total transacionado em 2019 e 2020 para os meses de janeiro a outubro. Apesar da tendência de crescimento do valor transacionado mês a mês, o valor transacionado em abril e maio de 2020 foi menor que o valor transacionado em abril e maio de 2019 para todas as regiões.

Na Figura 20 vemos uma análise do valor total transacionado por UF no período de janeiro de 2019 e outubro de 2020. É possível observar um valor total transacionado consideravelmente maior no estado de São Paulo que, segundo descrição da seção 5.2, tem uma participação de 8% do total das empresas deste estado, um percentual não muito diferente de outros estados citados. Outras UFs tem um valor total transacionado consideravelmente maior também, como Minas Gerais e Rio de Janeiro que reforçam a participação maior da região 4, e os estados da região 5 (sul): Paraná, Santa Catarina, e Rio Grande do Sul.

A seguir, a Figura 21 apresenta graficamente uma comparação trimestral sobre o valor total transacionado em cada UF. Embora o valor total transacionado tenha grandes discrepâncias entre os estados, o que dificulta a identificação visual do impacto trimestral, é possível verificar

Figura 19 – Comparação do valor mensal transacionado por região entre 2019 e 2020



(a) Região 1

(b) Região 2

(c) Região 3

(d) Região 4

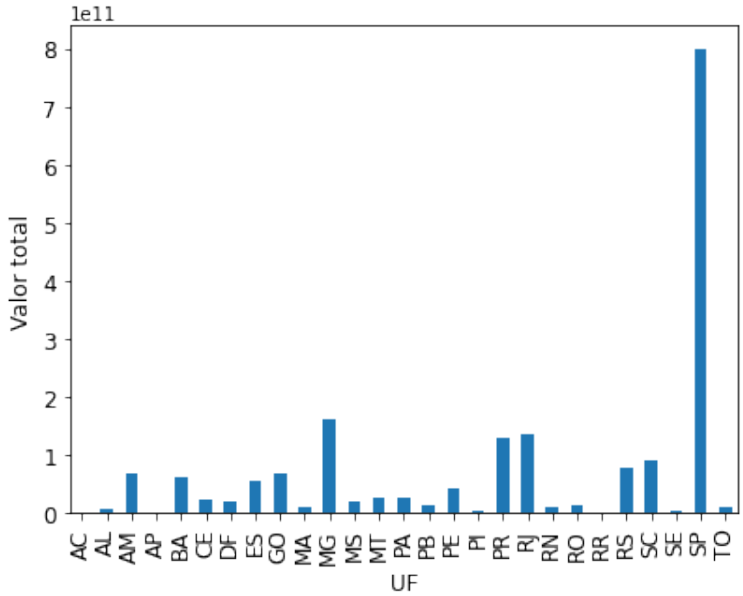
(e) Região 5

Fonte: Dados da pesquisa.

que na maioria das UFs o valor total transacionado no segundo trimestre de 2020 é menor que no primeiro e no terceiro trimestre, o que indica um impacto maior devido à crise causada pela pandemia de COVID-19 nesse trimestre.

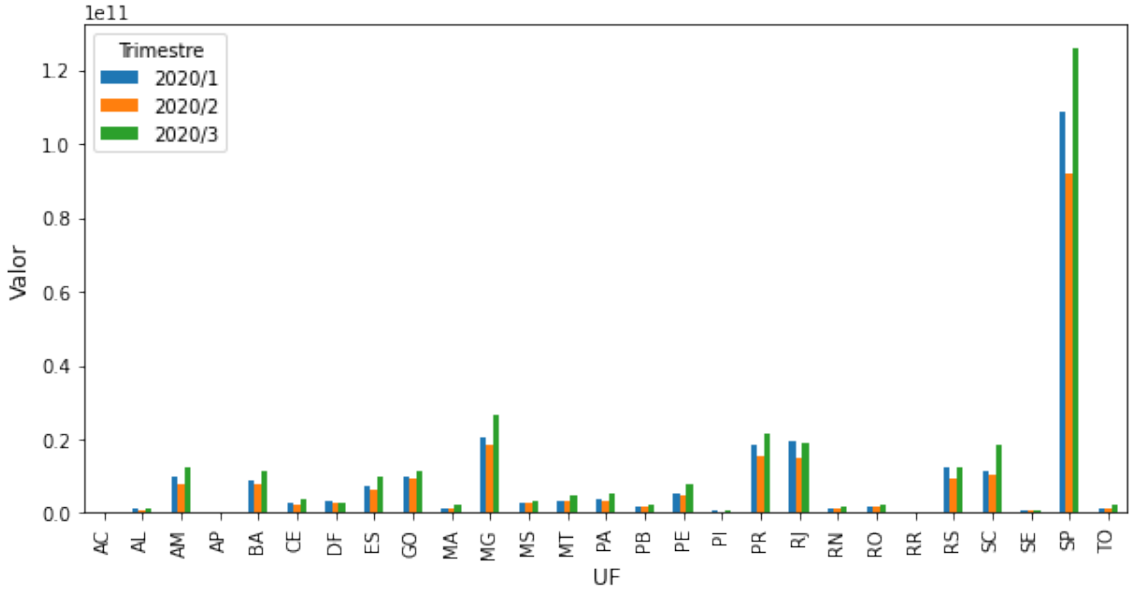
Os dados são então detalhados na Tabela 12, onde é apresentada a diferença entre o valor mensal transacionado previsto para cada UF utilizando uma regressão linear sobre os dados entre janeiro de 2019 e fevereiro de 2020 e o valor real obtido para cada mês seguinte, de março a

Figura 20 – Valor total transacionado por UF no período analisado



Fonte: Dados da pesquisa.

Figura 21 – Valor trimestral transacionado por UF no período analisado



Fonte: Dados da pesquisa.

outubro de 2020.

É possível observar que apenas o estado de Roraima não foi afetado pela crise segundo essa análise, mas como indicado na Tabela 12 os dados desse estado não possuem um ruído considerável. Todos as outras UFs apresentaram pelo menos um mês cujo valor transacionado previsto e real indicam impacto negativo da crise. Os meses que concentram maior impacto novamente são os meses de março a junho, novamente indicando que o segundo de trimestre de

Tabela 12 – Diferença entre valores mensais transacionados previstos e obtidos por UF no período da pandemia

UF	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro
AC	-7.3%	-33.9%	-5.1%	9.2%	27.3%	20.5%	20.6%	15.4%
AL	-11.0%	-33.6%	-26.5%	-12.3%	6.9%	3.9%	15.4%	9.0%
AM	-15.8%	-62.5%	-31.5%	-14.3%	0.5%	-4.5%	-8.4%	4.1%
AP	-15.6%	-29.1%	-17.5%	-8.1%	17.4%	23.9%	22.1%	12.9%
BA	-0.4%	-36.0%	-21.7%	-3.3%	11.0%	4.9%	12.7%	13.5%
CE	-15.6%	-45.1%	-26.3%	-7.7%	10.8%	7.6%	7.4%	19.1%
DF	32.3%	-18.6%	-19.0%	16.4%	9.8%	2.8%	10.7%	6.6%
ES	-6.5%	-40.1%	-24.2%	-5.8%	6.1%	5.6%	16.2%	6.4%
GO	10.8%	-29.5%	-7.9%	4.0%	9.5%	1.3%	7.3%	0.4%
MA	-20.9%	-31.2%	-24.8%	0.2%	22.3%	21.2%	6.8%	13.7%
MG	-9.0%	-35.1%	-20.6%	-6.7%	12.2%	9.1%	12.0%	11.1%
MS	3.4%	-16.0%	-9.1%	-1.6%	12.8%	5.8%	-3.5%	-15.3%
MT	-13.3%	-38.9%	-13.4%	-3.2%	7.0%	24.5%	19.8%	17.0%
PA	-13.6%	-33.3%	-25.8%	-8.8%	18.5%	5.9%	9.7%	6.1%
PB	-8.7%	-38.5%	-11.6%	1.7%	36.1%	15.8%	20.2%	30.7%
PE	-14.6%	-43.4%	-21.0%	-7.9%	17.5%	15.8%	11.8%	23.9%
PI	-10.6%	-47.1%	-24.5%	-7.0%	25.2%	13.4%	24.3%	15.1%
PR	-2.2%	-37.2%	-22.8%	-11.5%	1.5%	-0.7%	1.4%	-2.0%
RJ	-6.7%	-43.6%	-25.6%	-17.2%	-8.6%	-10.4%	-17.7%	-17.7%
RN	-14.9%	-47.3%	-39.7%	-23.3%	10.6%	8.1%	24.7%	31.9%
RO	-3.0%	-25.2%	-10.0%	-4.7%	17.1%	2.1%	-0.1%	-4.8%
RR	82.5%	90.6%	189.3%	370.4%	1480.1%	-	-	-
RS	-18.4%	-42.0%	-33.7%	-20.1%	-7.9%	-17.9%	-15.0%	-23.9%
SC	-8.2%	-32.5%	-18.8%	-6.3%	89.5%	12.3%	20.3%	19.7%
SE	-4.2%	-38.7%	6.2%	23.9%	16.0%	13.6%	7.4%	26.6%
SP	-0.9%	-32.9%	-24.1%	-9.9%	3.0%	1.9%	4.6%	4.1%
TO	-13.4%	-32.1%	-18.9%	2.4%	18.4%	13.9%	23.7%	19.7%

Nota – Os dados de Roraima foram insuficientes para uma previsão correta e registraram valores com bastante ruído, fazendo com que a previsão de valores utilizando a regressão linear indicasse valores bastante pessimistas para os meses da pandemia. Foram mantidos os valores de março a julho e removidos os valores posteriores devido ao ruído. Da mesma forma, consideramos que o estado de Roraima não foi impactado negativamente pela pandemia

Fonte: Dados da pesquisa.

2020 foi o mais impactado pela pandemia.

Na Tabela 13 apresentamos a diferença entre o valor total previsto e obtido para o período de março a outubro de 2020.

É possível verificar que dezoito UFs tiveram impacto no valor total transacionado em relação ao previsto, sendo os mais afetados o Rio Grande do Sul, o Rio de Janeiro, e o Amazonas. Nove UFs tiveram um valor previsto menor que o obtido, o que segundo esta análise significa

Tabela 13 – Diferença entre valores totais transacionados previstos e obtidos por UF no período da pandemia

UF	Variação total
AC	6.2%
AL	-5.5%
AM	-15.7%
AP	1.3%
BA	-1.8%
CE	-5.7%
DF	5.1%
ES	-4.9%
GO	-0.4%
MA	-0.7%
MG	-3.1%
MS	-2.9%
MT	0.4%
PA	-4.6%
PB	5.9%
PE	-1.7%
PI	-0.7%
PR	-8.9%
RJ	-18.4%
RN	-6.1%
RO	-3.4%
RR	484.7%
RS	-22.1%
SC	10.0%
SE	6.6%
SP	-6.6%
TO	2.5%

Fonte: Dados da pesquisa.

que a retomada para estes estados pode ter compensado as perdas dos meses mais críticos.

Na Tabela 14 é descrita a quantidade de UFs impactada mensalmente e totalmente no período analisado. Consideramos aqui que o impacto mensal significa que pelo menos um mês teve um valor transacionado obtido 25% menor que o valor previsto, e impacto total indica que o valor total obtido foi menor que o previsto.

5.3 Cenário econômico

Nesta seção, serão apresentados os resultados da análise de impacto da pandemia de COVID-19 na economia do ponto de vista setorial. Para isso, vamos usar o CNAE para segmentar a economia, como descrito na seção 4.3.2.

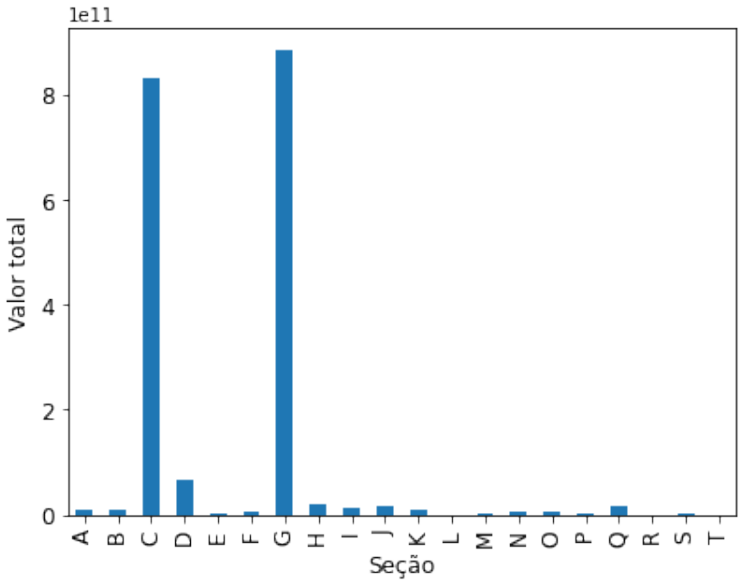
Tabela 14 – Quantidade de UFs impactadas negativamente pela pandemia em relação às variações mensais ou totais

Impacto	Mensal	Total
Sim	24	18
Não	3	9

Fonte: Dados da pesquisa.

Na Figura 22 é apresentado o volume transacionado por seção de CNAE considerando o total entre janeiro de 2019 e outubro de 2020.

Figura 22 – Valor total transacionado por seção no período analisado

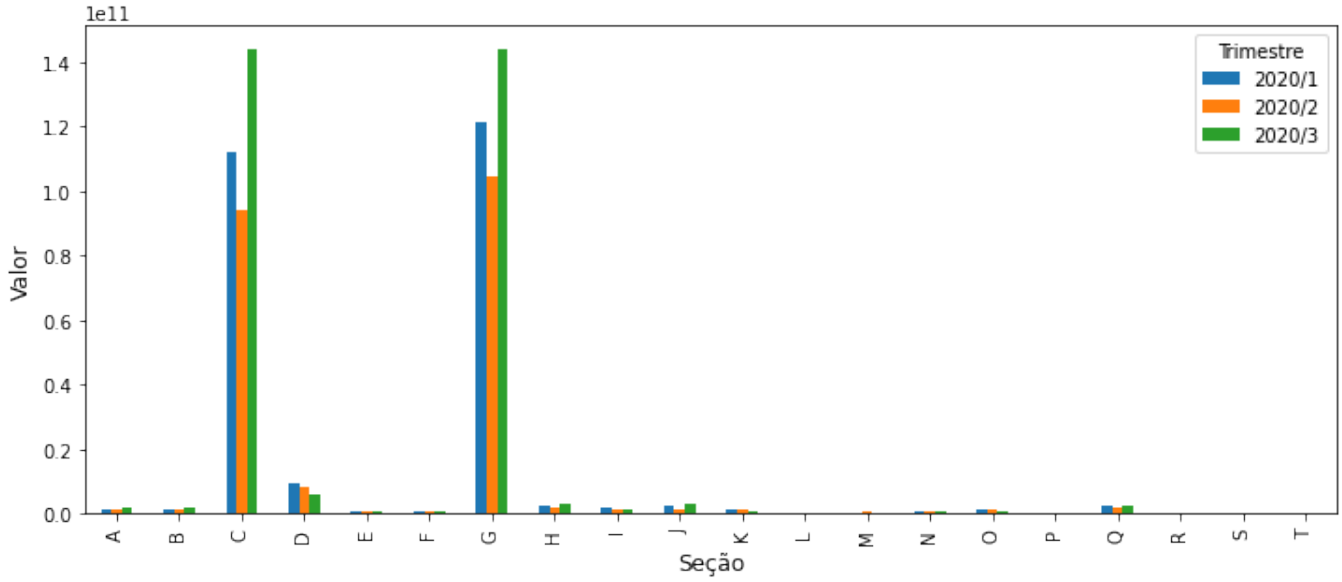


Fonte: Dados da pesquisa.

A partir desses dados, podemos ver um volume muito maior transacionado nas seções C e G do CNAE, respectivamente as seções de "Indústria de Transformação"e "Comércio; Reparação de veículos automotores e motocicletas". Como pudemos ver na seção 4.3.2, as seções C e G apresentam uma participação um pouco maior que as demais na base de dados, sendo que a seção G possui uma quantidade bastante grande de empresas também na base de referência em relação às demais, fato correspondido também aqui.

Na Figura 23 repetimos a análise trimestral agora por seção, ilustrando os valores transacionados nos três primeiros trimestres de 2020 para cada seção. Como as seções C e G possuem volumes muito maiores, é difícil visualizar o impacto para todas as seções, mas podemos verificar a mesma tendência encontrada até aqui: o segundo trimestre de 2020 apresentou um impacto negativo em relação ao primeiro e ao terceiro trimestres, com acentuada queda o volume transacionado total.

Figura 23 – Valor trimestral transacionado por seção no período analisado



Fonte: Dados da pesquisa.

Na Tabela 15 é descrito em detalhes a diferença entre o valor previsto via regressão linear aplicada aos dados de valor mensal transacionado entre janeiro de 2019 a fevereiro de 2020 para os meses de março a outubro de 2020 e o valor obtido nesses meses.

Podemos verificar que a maior parte do impacto na maioria das seções da economia se dá entre março e junho de 2020, com perdas significativas registradas neste período para várias seções. Ao contrário de outras análises até aqui, podemos verificar que algumas seções da economia foram pouco impactadas e apresentaram resultados positivos no período estudado. A seção B, de indústrias extrativas, apresentou resultados positivos em todo o período estudado. Outras seções apresentaram um impacto negativo estendido para todos os meses, sendo estas as seções D, F, H, L, P, R, e T.

A Tabela 16 apresenta a diferença entre o valor total previsto e obtido para o período estudado. As seçõs B e O, respectivamente referentes a Indústrias Extrativas e Administração pública, foram aquelas em que o impacto foi positivo, obtendo um valor maior que o previsto, enquanto outras seções foram severamente impactadas, como L, R, T, respectivamente referentes a Atividades Imobiliárias, Artes, Cultura e Esporte, e Serviços Domésticos.

A Tabela 17 apresenta a consolidação da categorização de seções quanto ao impacto mensal e total. Novamente, impacto mensal se refere às seções cujo o valor obtido foi 25% menor que o previsto para pelo menos um mês da pandemia, e impacto total se o valor total obtido para o período foi menor que o previsto.

A Figura 24 estende a análise para um estudo do impacto por CNAE. Ela apresenta um diargama de caixas dos valores mensais transacionados por CNAE entre janeiro de 2019 e

Tabela 15 – Diferença entre valores mensais transacionados previstos e obtidos por seção no período da pandemia

Seção	Março	Abril	Maio	Junho	Julho	Agosto	Setembro	Outubro
A	-15.4%	-24.4%	-14.2%	-10.8%	6.9%	8.2%	32.9%	28.6%
B	2.8%	12.8%	19.1%	32.7%	50.6%	51.0%	58.5%	64.6%
C	-4.2%	-39.0%	-23.6%	-5.8%	16.0%	11.7%	15.5%	16.2%
D	-1.4%	-6.1%	-16.7%	-14.4%	-12.8%	-39.3%	-62.2%	-64.5%
E	10.8%	-32.0%	-17.3%	17.1%	65.5%	125.2%	-1.3%	15.3%
F	-4.7%	-30.0%	-26.8%	-20.0%	-8.1%	-13.8%	-10.0%	-15.0%
G	-5.9%	-36.8%	-23.0%	-11.6%	4.0%	-2.3%	-0.2%	-2.3%
H	-16.2%	-40.0%	-39.2%	-24.5%	-17.0%	-15.5%	-7.0%	-9.7%
I	-25.8%	-53.7%	-51.7%	-42.2%	-33.1%	-33.4%	-27.9%	-25.3%
J	24.3%	-61.6%	-47.4%	-32.3%	-6.4%	-5.3%	-19.3%	5.4%
K	4.6%	15.1%	12.6%	13.2%	14.2%	-72.6%	-64.8%	-66.1%
L	-28.3%	-34.7%	-56.8%	-51.3%	-49.2%	-29.1%	-34.2%	-57.0%
M	-12.2%	-12.2%	57.9%	71.4%	-33.2%	-32.2%	-16.3%	-12.8%
N	-1.6%	-52.1%	-42.2%	-10.0%	28.9%	-4.7%	5.7%	1.9%
O	53.8%	20.7%	2.8%	77.7%	11.1%	-4.6%	27.2%	-28.1%
P	-12.0%	-56.8%	-44.6%	-33.5%	-22.1%	-34.0%	-31.3%	-39.5%
Q	33.1%	-10.5%	-6.4%	-7.2%	6.7%	-3.9%	2.7%	-11.2%
R	-32.2%	-68.9%	-69.6%	-50.6%	-42.7%	-41.5%	-18.5%	-18.3%
S	-5.4%	-32.9%	-24.7%	-22.5%	-10.4%	-19.5%	2.0%	-0.2%
T	-40.5%	-38.2%	-61.9%	-48.5%	-47.5%	-54.7%	-54.2%	-27.6%

Fonte: Dados da pesquisa.

outubro de 2020.

Os dados de valores transacionados por CNAE tem grande variância, o que dificulta uma análise, mas é possível ver através do histograma que novamente os meses de abril e maio possuem uma distribuição de valores ligeiramente menor que o restante, mantendo o padrão anterior.

Sobre os dados de CNAE também foi aplicada uma regressão linear sobre os dados obtidos nos meses de janeiro de 2019 a fevereiro de 2020 para a previsão dos valores para os meses de março de a outubro de 2020. Da mesma forma, a Tabela 18 mostra uma categorização de impacto mensal e total para CNAEs usando os mesmos critérios apresentados anteriormente: é considerado impacto mensal aquele onde para um determinado mês o valor obtido é 25% menor que o valor previsto, e impacto total se o valor total obtido é menor que o previsto.

5.4 Conclusões

Finalizamos então a análise de impacto para diferentes segmentações, e pudemos verificar que em todos os cenários houve algum impacto entre os valores previstos e obtidos. Foi

Tabela 16 – Diferença entre valores totais transacionados previstos e obtidos por seção no período da pandemia

Seção	Variação total
A	2.4%
B	36.0%
C	-1.3%
D	-27.0%
E	23.5%
F	-16.0%
G	-9.5%
H	-20.8%
I	-36.5%
J	-17.4%
K	-17.7%
L	-42.8%
M	0.5%
N	-9.1%
O	20.8%
P	-34.2%
Q	0.3%
R	-42.7%
S	-14.0%
T	-46.6%

Fonte: Dados da pesquisa.

Tabela 17 – Quantidade de seções impactadas negativamente pela pandemia em relação às variações mensais ou totais

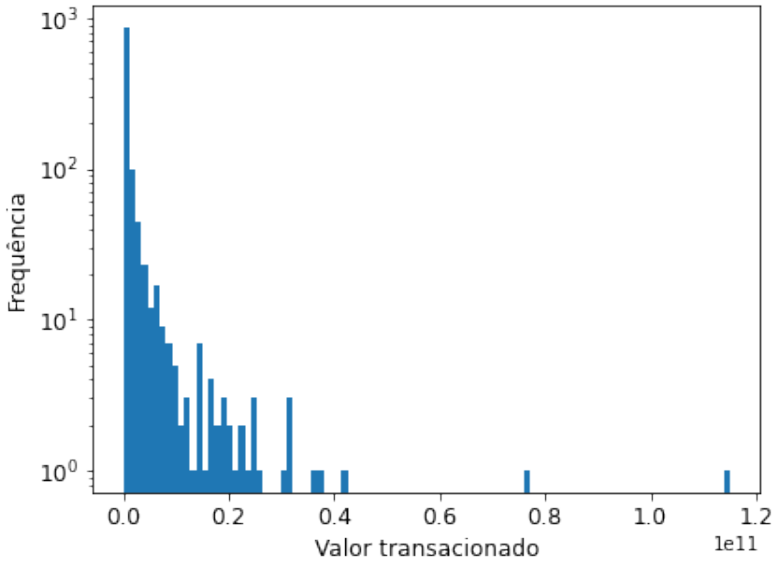
Impacto	Mensal	Total
Sim	17	14
Não	3	6

Fonte: Dados da pesquisa.

apresentada uma análise comparativa entre 2019 e 2020, e uma análise usando uma previsão de valor transacionado e valor real obtido. O impacto foi mais localizado entre os meses de março e junho, tornando o segundo trimestre de 2020 o mais afetado.

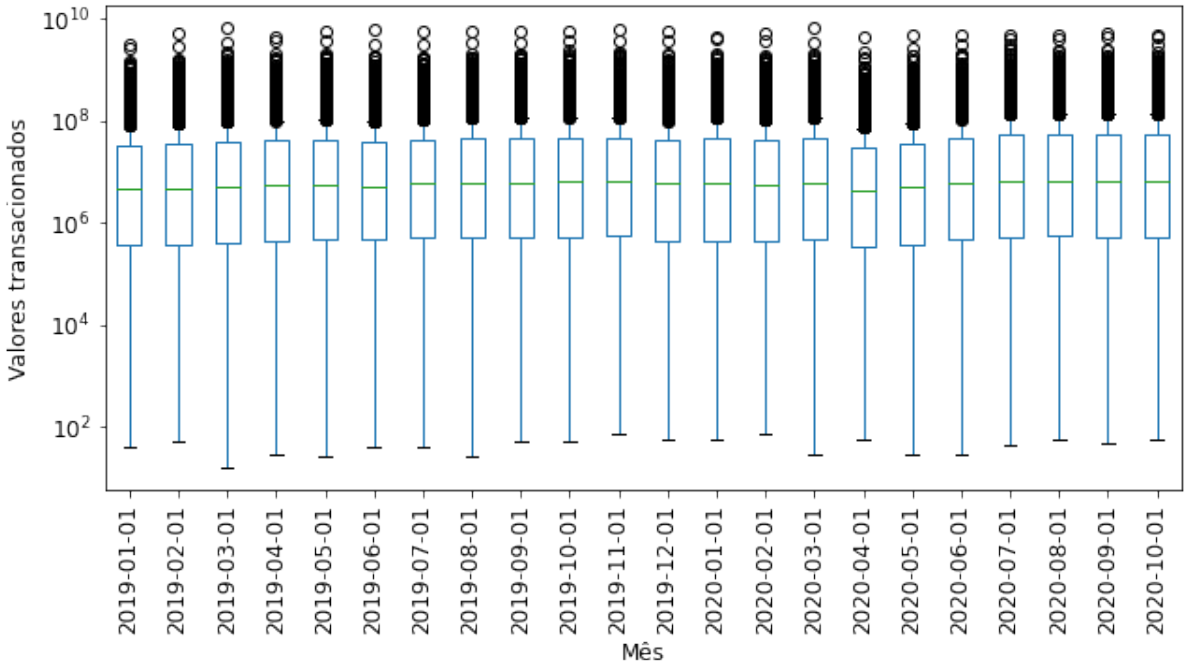
Cada cenário possui proporções diferentes de entidades afetadas e não afetadas pela pandemia, com um desbalanceamento maior para entidades afetadas. O objetivo, portanto, foi apresentar uma categorização das entidades ao mesmo tempo que aprofundamos a análise do Capítulo 4 para mostrar sob o ponto de vista de valores transacionados uma nova perspectiva dos dados aqui apresentados.

Figura 24 – Histograma do valor total transacionado por CNAE no período analisado



Fonte: Dados da pesquisa.

Figura 25 – Diagrama de caixa dos valores mensais transacionados por CNAE no período analisado



Fonte: Dados da pesquisa.

Tabela 18 – Quantidade de CNAEs impactadas negativamente pela pandemia em relação às variações mensais ou totais

Impacto	Mensal	Total
Sim	852	600
Não	273	525

Fonte: Dados da pesquisa.

CAPÍTULO

6

RESULTADOS

Neste capítulo, serão apresentados os resultados obtidos a partir da aplicação do que foi proposto na seção 1.4, seguido de uma análise dos dados.

6.1 Caracterização de redes

Neste capítulo, faremos uma caracterização dos dados a partir de diferentes modelagens utilizando redes complexas.

6.1.1 Redes UF-UF

Nesta seção, será apresentada a caracterização de uma rede entre unidades federativas brasileiras.

Tabela 19 – Métricas do grafo de transações entre UFs

Métrica	Valor
Quantidade de nós	27
Quantidade de arestas	377
Diâmetro	2
Raio	1
Grau médio	26.9259
Densidade	1.0740
Transitividade	0.9971
Assortatividade de grau	-0.0356
Assortatividade de região	-0.0025

Fonte: Dados da pesquisa.

Essa rede foi montada utilizando os dados de transações entre janeiro de 2019 e dezembro de 2019. Cada vértice representa uma UF e cada aresta representa o valor total transacionado

no período entre duas UFs. Nesta modelagem consideramos um grafo não-direcionado cíclico. Cada vértice possui um atributo chamado **região** indicando a região a que pertence aquela UF. Cada aresta além do atributo contendo o valor total transacional, e um atributo chamado **distância** correspondente ao inverso do valor total transacionado. Este atributo reflete a distância econômica entre dois vértices, uma vez que vértices que tenham um alto valor transacionado entre si terão uma relação mais próxima, o que pode ser útil na formulação de alguns algoritmos. A Tabela 19 apresenta as métricas do grafo para esta modelagem.

Tabela 20 – Métricas do grafo de transações entre UFs segmentadas por UF

UF	Grau	Coeficiente de agrupamento	Caminho mínimo médio poderado	Centralidade de informação	Centralidade de autovalor	*PageRank*
AC	26	0.0003	$2.61e^{-09}$	$4.51e^{+07}$	0.0016	0.0077
AL	27	0.0008	$1.07e^{-09}$	$1.15e^{+08}$	0.0057	0.0112
AM	27	0.0045	$5.06e^{-10}$	$2.33e^{+08}$	0.0714	0.0468
AP	26	0.0005	$2.66e^{-09}$	$5.57e^{+07}$	0.0016	0.0077
BA	27	0.0030	$5.27e^{-10}$	$2.16e^{+08}$	0.0538	0.0348
CE	27	0.0019	$6.39e^{-10}$	$1.81e^{+08}$	0.0199	0.0185
DF	27	0.0015	$5.74e^{-10}$	$1.83e^{+08}$	0.0290	0.0162
ES	27	0.0033	$5.03e^{-10}$	$2.24e^{+08}$	0.0801	0.0334
GO	27	0.0037	$5.17e^{-10}$	$2.25e^{+08}$	0.0634	0.0401
MA	27	0.0012	$8.34e^{-10}$	$1.46e^{+08}$	0.0093	0.0131
MG	27	0.0054	$4.82e^{-10}$	$2.39e^{+08}$	0.1696	0.0744
MS	27	0.0015	$5.96e^{-10}$	$1.75e^{+08}$	0.0249	0.0153
MT	27	0.0019	$6.21e^{-10}$	$1.85e^{+08}$	0.0219	0.0204
PA	27	0.0021	$5.89e^{-10}$	$1.90e^{+08}$	0.0261	0.0216
PB	27	0.0014	$7.46e^{-10}$	$1.58e^{+08}$	0.0124	0.0140
PE	27	0.0029	$5.59e^{-10}$	$2.14e^{+08}$	0.0339	0.0325
PI	27	0.0007	$1.37e^{-09}$	$9.40e^{+07}$	0.0039	0.0084
PR	27	0.0049	$4.85e^{-10}$	$2.37e^{+08}$	0.1416	0.0630
RJ	27	0.0051	$4.82e^{-10}$	$2.38e^{+08}$	0.1596	0.0659
RN	27	0.0011	$8.10e^{-10}$	$1.33e^{+08}$	0.0097	0.0125
RO	27	0.0010	$8.09e^{-10}$	$1.45e^{+08}$	0.0100	0.0138
RR	27	0.0003	$1.83e^{-09}$	$4.51e^{+07}$	0.0009	0.0067
RS	27	0.0038	$5.16e^{-10}$	$2.26e^{+08}$	0.0680	0.0416
SC	27	0.0042	$4.97e^{-10}$	$2.31e^{+08}$	0.0971	0.0474
SE	27	0.0006	$1.64e^{-09}$	$9.23e^{+07}$	0.0025	0.0091
SP	27	0.0119	$4.59e^{-10}$	$2.52e^{+08}$	0.9426	0.3101
TO	27	0.0011	$8.96e^{-10}$	$1.42e^{+08}$	0.0085	0.0127

Fonte: Dados da pesquisa.

É possível notar que o grafo entre UFs é quase completo, com alto grau médio, refletindo o baixo diâmetro e raio da rede. Dado que estamos usando uma modelagem cíclica, aceitando arestas que conectam um vértice a ele mesmo, a densidade desta rede é maior que 1.0. Temos também alta transitividade, e baixa assortatividade de grau e de região. Por ser um grafo bastante

completo, é difícil tomar qualquer conclusão a partir destas métricas uma vez que elas irão refletir a simplicidade da topologia desta rede.

A Tabela 25 nos dá uma ideia melhor da topologia do grafo. Para essas métricas agora consideramos a formulação usando arestas ponderadas pelo valor total transacionado pelas UFs. Para os caminhos mínimos, foi utilizada a distância entre UFs, como definido anteriormente.

Todos as UFs estão conectadas entre si com uma exceção: o estado do Acre não possui conexão com o estado do Amapá. O estado de Sâo Paulo se sobressai com maiores valores de coeficiente de agrupamento, centralidade de autovalor, e pagerank por ter um maior valor transacionado entre este estado e todos os outros se comparado aos demais, algo que já foi descrito na seção 5. Observando outros estados da região sudeste, como Rio de Janeiro de Minas Gerais, estes também tem valores importantes para essas métricas mostrando sua importância, novamente refletindo a tendência dessa região de se sobressair quanto ao valor transacionado e influenciando positivamente nestas métricas. Os estados da região norte, como Acre, Amapá, e Roraima apresentaram baixos valores para estas métricas refletindo os baixos valores transacionados por estas UFs com as demais. Sendo assim, apesar de uma topologia simples, ela se mostra razoavelmente eficaz em mostrar a importância de cada estado em relação às relações econômicas entre UFs brasileiras, também obedecendo às limitações e ao viés desta base de dados.

Nesta análise, não consideramos as métricas de centralidade de intermediação, caminho mínimo médio, e excentricidade, uma vez que o grafo é quase completo e essas medidas refletiriam apenas o grau de cada vértice, apresentando pouca variância e sendo redundante com o restando dos dados obtidos.

6.1.2 Redes Seção-Seção

Nesta seção, será apresentada a caracterização de uma rede entre seções de CNAE.

Tabela 21 – Métricas do grafo de transações entre seções

Métrica	Valor
Quantidade de nós	20
Quantidade de arestas	186
Diâmetro	2
Raio	1
Grau médio	17.6000
Densidade	0.9789
Transitividade	0.9618
Assortatividade de grau	-0.1433

Fonte: Dados da pesquisa.

Essa rede foi montada utilizando os dados de transações entre janeiro de 2019 e dezembro de 2019, como anteriormente. Cada vértice representa uma seção e cada aresta representa o valor

total transacionado no período entre duas seções. Nesta modelagem consideramos novamente um grafo não-direcionado cíclico. Cada aresta possui também um atributo chamado **distância** correspondente ao inverso do valor total transacionado. A Tabela 21 apresenta as métricas do grafo para esta modelagem.

Podemos observar que este grafo não é tão completo quanto o o grafo UF-UF, mas também é bastante completo com um alto grau médio e densidade, encurtando os valores de raio e diâmetro. A assortatividade de grau indica leve tendência à assortatividade negativa, indicando alguma tendência de vértices menos conectados se conectarem à vértices mais conectados e vice-versa, uma característica que reflete também a própria topologia dessa rede, com pouca variância entre o grau de seus vértice uma vez que é uma rede bastante conectada.

As Tabelas 22 e 23 detalham as métricas agora por seção, sendo novamente considerada a formulação usando arestas ponderadas e utilizando a distância entre seções para a formulação de caminhos mínimos ponderados.

Tabela 22 – Métricas do grafo de transações entre seções segmentadas por seção (Parte 1)

Seção	Grau	Coeficiente de agrupamento	Caminho mínimo médio	Caminho mínimo médio ponderado
A	19	0.0006	1	$1.635e^{-06}$
B	19	0.0007	1	$1.635e^{-06}$
C	20	0.0047	0.95	$1.634e^{-06}$
D	19	0.0010	1	$1.634e^{-06}$
E	19	0.0003	1	$1.635e^{-06}$
F	19	0.0005	1	$1.635e^{-06}$
G	20	0.0048	0.95	$1.634e^{-06}$
H	19	0.0010	1	$1.634e^{-06}$
I	19	0.0008	1	$1.634e^{-06}$
J	19	0.0009	1	$1.634e^{-06}$
K	17	0.0005	1.1	$1.635e^{-06}$
L	15	>0.0001	1.2	$1.648e^{-06}$
M	19	0.0002	1	$1.635e^{-06}$
N	19	0.0005	1	$1.635e^{-06}$
O	18	0.0005	1.05	$1.634e^{-06}$
P	18	0.0001	1.05	$1.637e^{-06}$
Q	19	0.0008	1	$1.634e^{-06}$
R	15	>0.0001	1.1	$1.646e^{-06}$
S	19	0.0002	1	$1.635e^{-06}$
T	1	>0.0001	1.8	$3.102e^{-05}$

Fonte: Dados da pesquisa.

Podemos observar que assim como anteriormente, as seções C e G se sobressaem com altos valores de agrupamento e centralidade, indicando influencia dos altos valores transacionados por essas seções com outras regiões. A seção T se sobressai com uma baixa conectividade com

Tabela 23 – Métricas do grafo de transações entre seções segmentadas por seção (Parte 2)

Seção	Centralidade de informação	Centralidade de autovalor	Centralidade de intermediação	Centralidade de proximidade	PageRank
A	$3.907e^{+04}$	0.0116	0.0020	0.95	0.0131
B	$3.907e^{+04}$	0.0090	0.0020	0.95	0.0132
C	$3.907e^{+04}$	0.6961	0.0517	1	0.3633
D	$3.907e^{+04}$	0.0139	0.0020	0.95	0.0457
E	$3.906e^{+04}$	0.0038	0.0020	0.95	0.0093
F	$3.906e^{+04}$	0.0067	0.0020	0.95	0.0110
G	$3.907e^{+04}$	0.7167	0.0517	1	0.3954
H	$3.907e^{+04}$	0.0213	0.0020	0.95	0.0171
I	$3.907e^{+04}$	0.0123	0.0020	0.95	0.0143
J	$3.907e^{+04}$	0.0173	0.0020	0.95	0.0169
K	$3.907e^{+04}$	0.0032	0	0.86	0.0147
L	$3.887e^{+04}$	0.0001	0	0.79	0.0075
M	$3.905e^{+04}$	0.0020	0.0020	0.95	0.0090
N	$3.906e^{+04}$	0.0066	0.0020	0.95	0.0105
O	$3.907e^{+04}$	0.0085	0.0003	0.90	0.0113
P	$3.903e^{+04}$	0.0006	0.0003	0.90	0.0078
Q	$3.907e^{+04}$	0.0177	0.0020	0.95	0.0156
R	$3.881e^{+04}$	0.0001	0	0.86	0.0075
S	$3.905e^{+04}$	0.0016	0.0020	0.95	0.0084
T	$2.058e^{+03}$	4.826e	0	0.53	0.0075

Fonte: Dados da pesquisa.

outras seções, e as seções L e R são bastante influenciadas pelos baixos valores transacionados. O caminho mínimo médio, a centralidade de proximidade, e a centralidade de intermediação refletem naturalmente o grau de cada vértice, e observamos pouca variância ao analisarmos o caminho mínimo médio ponderado, com exceção da seção T, o que também acontece para a centralidade de informação. A centralidade de autovalor e o pagerank refletem novamente a magnitude dos valores transacionados por cada seção, com destaque positivo para as seções C e G, e negativo para as seções L, R, e T. No geral, da mesma forma como a rede UF-UF, a importância de cada seção reflete os valores totais transacionados de acordo com o viés e as limitações da base de dados utilizada.

6.1.3 Redes CNAE-CNAE

Nesta seção, vamos repetir mais uma vez a análise agora para a caracterização de uma rede entre CNAEs.

Essa rede foi também montada utilizando os dados de transações entre janeiro de 2019 e dezembro de 2019. Cada vértice representa um CNAE e cada aresta representa o valor total transacionado no período entre dois CNAEs. Nesta modelagem consideramos um grafo não-

Tabela 24 – Métricas do grafo de transações entre CNAEs

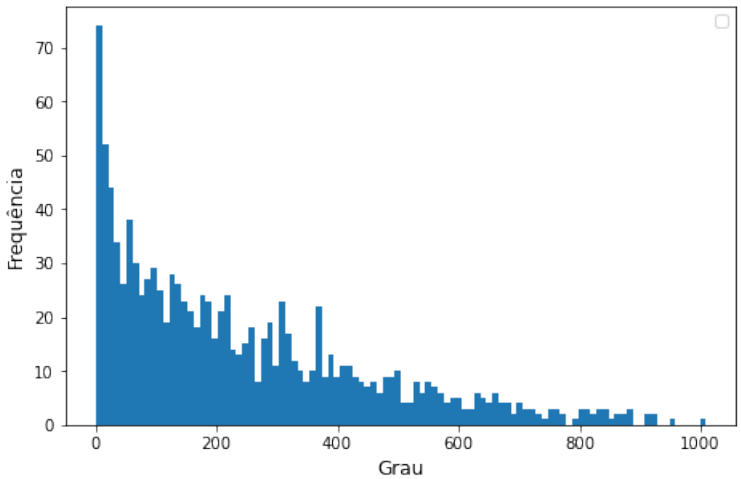
Métrica	Valor
Quantidade de nós	1125
Quantidade de arestas	135607
Diâmetro	4
Raio	2
Grau médio	241.0791
Densidade	0.2145
Transitividade	0.5293
Assortatividade de grau	-0.3449
Assortatividade de seção	0.0015

Fonte: Dados da pesquisa.

direcionado cíclico. Cada vértice possui um atributo chamado **seção** indicando a seção a que pertence aquele CNAE. Cada aresta possui também um atributo chamado **distância** correspondente ao inverso do valor total transacionado. A Tabela 24 apresenta as métricas do grafo para esta modelagem.

Através da Tabela 24 podemos observar que essa rede possui uma topologia mais complexo que as anteriores. Temos agora uma rede de menor grau médio e densidade. Podemos fazer então a análise da distribuição de grau através do histograma com o grau dos vértices apresentado na Figura 26. É possível observar a ocorrência de *hubs*, vértices de alto grau, e existe uma variância significativa entre o grau dos vértices. O diâmetro e o raio indicam também que é uma rede curta.

Figura 26 – Histograma do grau dos vértices da rede CNAE-CNAE



Fonte: Dados da pesquisa.

A rede também apresenta uma transitividade significativa que reflete a conectividade da rede e uma tendência leve à desassortatividade de grau. A assortatividade de seção não indicou

nenhuma tendência, indicando que não existe preferência dos vértices de se conectarem ou não à vértices da mesma seção.

As Tabelas 25 e 26 apresentam a distribuição das outras métricas estudadas. Novamente, foram usadas as formulações utilizando arestas ponderadas para o estudo.

Tabela 25 – Métricas do grafo de transações entre CNAEs segmentadas por CNAE (Parte 1)

Métrica	Grau	Coeficiente de agrupamento	Excentricidade	Caminho mínimo médio	Caminho mínimo médio ponderado
Média	241.1	$1.471e^{-05}$	2.917	1.8160	$1.140e^{-05}$
Desvio Padrão	213.7	$1.946e^{-05}$	0.309	0.2289	$6.641e^{-05}$
Mínimo	1	0	2	1.1030	$5.708e^{-06}$
25º percentil	66	$4.441e^{-06}$	3	1.6800	$5.722e^{-06}$
50º percentil	183	$9.605e^{-06}$	3	1.8470	$5.776e^{-06}$
75º percentil	365	$1.790e^{-05}$	3	1.9640	$6.351e^{-06}$
Máximo	1009	0.0003	4	2.7970	0.0019

Fonte: Dados da pesquisa.

Tabela 26 – Métricas do grafo de transações entre CNAEs segmentadas por CNAE (Parte 2)

Métrica	Centralidade de informação	Centralidade de autovalor	Centralidade de intermediação	Centralidade de proximidade	PageRank
Média	225.9	0.0020	0.0007	0.5598	0.0008
Desvio Padrão	47.49	0.0298	0.0019	0.0769	0.0021
Mínimo	0.4702	$3.198e^{-12}$	0	0.3572	0.0001
25º percentil	235.5	$3.452e^{-07}$	$3.047e^{-06}$	0.5086	0.0001
50º percentil	245.4	$5.856e^{-06}$	$5.415e^{-05}$	0.5409	0.0002
75º percentil	246.3	$7.721e^{-05}$	0.0005	0.5947	0.0007
Máximo	246.4	0.8161	0.0287	0.9057	0.0316

Fonte: Dados da pesquisa.

Analisando a distribuição do coeficiente de agrupamento, é possível verificar que existe uma certa variância que indicam a presença de alguns vértices com agrupamento local significativo, embora a magnitude da métrica seja bastante influenciada pelo fato de usar a formulação para arestas ponderadas. Por outro lado, a excentricidade dos vértices apresenta pouca variância o que reflete uma rede curta, de baixo diâmetro e raio, da mesma forma que a distribuição do caminho mínimo médio partindo dos vértices também possui uma variância limitada e fortemente correlacionada ao grau.

As métricas de centralidade de informação, de proximidade e de proximidade são novamente influenciadas pelo volume transacionado por cada CNAE e sofrem certo efeito também do baixo comprimento da rede, apresentando uma variância também baixa. Já a centralidade de autovalor e o pagerank apresentam variância significativa possivelmente refletindo o valor total transacionado pelos vértices e não sendo tão influenciados pela topologia da rede.

6.1.4 Conclusões

Assim, fica claro a partir da caracterização das redes com dados de 2019 que são redes com topologias não tão complexas e que possuem uma forte correlação com os valores transacionados por cada entidade. Apesar disso, este estudo pode adicionar informações importantes sobre a topologia dessas redes que serão usadas no estudo do impacto da pandemia do ponto de vista topológico e podem ser usados para a detecção ou decisão de entidades que tiveram impacto da pandemia.

6.2 Alterações topológicas

Nesta seção, vamos analisar mudanças que ocorreram na estrutura das redes para o período da pandemia em comparação com a topologia vista na caracterização anterior.

6.2.1 Redes UF-UF

Tabela 27 – Variação percentual das métricas de redes mês a mês para o grafo de UFs (Parte 1)

Mês	Densidade	Transitividade	Grau médio	Grau médio ponderado
01/2019	-	-	-	-
02/2019	+0.54%	+0.55%	+0.54%	+9.88%
03/2019	+0.00%	+0.17%	+0.00%	+25.54%
04/2019	+0.54%	+0.57%	+0.54%	+23.86%
05/2019	+0.81%	+0.81%	+0.81%	+36.91%
06/2019	+0.27%	+0.32%	+0.27%	+23.88%
07/2019	+0.54%	+0.57%	+0.54%	+28.36%
08/2019	+0.54%	+0.57%	+0.54%	+35.37%
09/2019	+0.27%	+0.36%	+0.27%	+32.62%
10/2019	+0.81%	+0.82%	+0.81%	+42.89%
11/2019	+0.27%	+0.34%	+0.27%	+43.93%
12/2019	+0.27%	+0.36%	+0.27%	+29.66%
01/2020	+0.81%	+0.83%	+0.81%	+28.84%
02/2020	+0.27%	+0.44%	+0.27%	+32.22%
03/2020	+0.54%	+0.59%	+0.54%	+36.09%
04/2020	-0.27%	-0.07%	-0.27%	-7.91%
05/2020	-0.27%	-0.18%	-0.27%	+12.47%
06/2020	+0.00%	+0.11%	+0.00%	+34.61%
07/2020	+0.54%	+0.57%	+0.54%	+62.51%
08/2020	+0.54%	+0.59%	+0.54%	+55.16%
09/2020	+1.08%	+1.09%	+1.08%	+60.32%
10/2020	+0.00%	+0.17%	+0.00%	+61.19%

Fonte: Dados da pesquisa.

Primeiramente, vamos analisar alterações topológicas na estrutura das redes UF-UF no período que compreende janeiro de 2019 e outubro de 2020. Para esta análise, os dados foram dividido mês a mês e um grafo foi criado para cada mês do período estudado. Cada vértice destes grafos representa uma UF e cada aresta representa o valor total transacionado entre duas UFs naquele mês. Usamos novamente a modelagem de grafo cíclico ponderado, sendo que cada aresta possui novamente um atributo indicando a distância entre os vértices igual ao inverso do valor transacionado. Cada vértice possui também um atributo indicando a região a qual pertence aquela UF, usado em uma das métricas de assortatividade.

A Tabelas 27 e 28 mostram a evolução de algumas métricas gerais das redes UF-UF para cada mês do período analisado. Cada linha da tabela representa o valor da métrica analisada em relação ao mês de janeiro de 2019, mostrando assim a evolução percentual em relação a esse mês-base.

Tabela 28 – Variação das métricas de redes mês a mês para o grafo de UFs (Parte 2)

Mês	Assortatividade de grau	Assortatividade de região	Caminho mínimo médio	Caminho mínimo médio ponderado
01/2019	-	-	-	-
02/2019	-15.60%	-50.13%	-0.55%	-8.47%
03/2019	+0.05%	-0.00%	+0.00%	-25.12%
04/2019	+7.91%	-15.39%	-0.55%	-22.23%
05/2019	-31.31%	-85.53%	-0.83%	-25.07%
06/2019	+21.99%	+13.35%	-0.27%	+2.44%
07/2019	-17.07%	-57.33%	-0.55%	-1.69%
08/2019	+7.91%	-57.33%	-0.55%	-11.02%
09/2019	+4.73%	-28.82%	-0.27%	-2.55%
10/2019	-6.87%	-43.83%	-0.83%	-11.15%
11/2019	+8.56%	+13.35%	-0.27%	-14.48%
12/2019	+4.73%	-70.86%	-0.27%	+2.50%
01/2020	-25.43%	-43.83%	-0.83%	+1.69%
02/2020	-18.97%	-28.82%	-0.27%	-5.58%
03/2020	-7.22%	-15.39%	-0.55%	-4.69%
04/2020	+1.98%	-13.38%	+0.27%	+36.70%
05/2020	+37.79%	+29.14%	+0.27%	+2.72%
06/2020	+19.66%	-42.28%	+0.00%	-15.66%
07/2020	+7.91%	-15.39%	-0.55%	-32.97%
08/2020	-7.22%	-15.39%	-0.55%	-35.56%
09/2020	-16.88%	-71.96%	-1.11%	-37.64%
10/2020	+0.05%	-42.28%	0.00%	-30.06%

Fonte: Dados da pesquisa.

A partir desses dados, é possível ver que as oscilações mês a mês nas métricas são pequenas, muitas vezes menor que 1%. Quando analisamos as densidade, transitividade, e grau médio, podemos ver que nos meses de abril e maio de 2020 há uma queda pequena mas

consistente em todas essas métricas. Isso mostra uma pequena perda de de conectividade geral da rede, que se recupera e retoma os níveis anteriores a partir de agosto de 2020. Analisando o grau médio ponderado, uma métrica que envolve valores transacionados, podemos ver que a variância nos meses anteriores à crise é maior, em patamares de dezenas de pontos percentuais, e da mesma forma o impacto da crise nos meses de abril e maio é mais significativo, chegando a ser 7% menor que o mês-base.

Analisando a assortatividade de grau e de região, podemos notar também uma variância mais significativa a partir dos dados nos meses que precedem a pandemia. Não há uma tendência clara em nenhuma dessas métricas, sendo que elas oscilam positivamente e negativamente no decorrer do tempo e não podemos deduzir que a variação para os meses de maior impacto da pandemia tem relação com a crise. No mês de maio, há um aumento pontual de ambas assortatividades, mas logo retornando às variações e aos patamares anteriores nos meses seguintes.

Ao observar o caminho mínimo médio verificamos uma variância pequena nos meses que precedem a pandemia, tendendo a encurtar os caminhos mínimos em relação ao mês-base, e nos meses do segundo trimestre de 2020 há um aumento tímido mas consistente desta métrica, voltando aos patamares anteriores após esse período. Utilizando a análise de caminhos mínimos ponderados observamos uma variação maior dos dados, novamente com tendência de diminuição da métrica nos meses que precedem a pandemia em relação ao mês base, sendo que nos meses de abril e maio a métrica tem um aumento consideravelmente consistente e depois alcança um patamar de alta diminuição dos caminhos mínimos em relação ao mês-base.

Tabela 29 – Variação das métricas de rede a cada trimestre para o grafo de UFs (Parte 1)

Mês	Densidade	Transitividade	Grau médio	Grau médio ponderado
2020/1	-	-	-	-
2020/2	-0.80%	-0.78%	-0.80%	-14.60%
2020/3	+0.27%	+0.22%	+0.27%	+20.35%

Fonte: Dados da pesquisa.

Tabela 30 – Variação das métricas de rede a cada trimestre para o grafo de UFs (Parte 2)

Mês	Assortatividade de grau	Assortatividade de região	Caminho mínimo médio	Caminho mínimo médio ponderado
2020/1	-	-	-	-
2020/2	+77.29%	+1.84%	+0.85%	+6.63%
2020/3	-34.35%	-83.36%	-0.28%	-33.74%

Fonte: Dados da pesquisa.

É possível traçar uma relação entre a diminuição da conectividade e o aumento dos

caminhos mínimos, uma vez que tais métricas são naturalmente inversamente proporcionais: quanto mais conectado é uma rede, menor são os caminhos mínimos necessários para percorrê-la. Isso não significa necessariamente uma relação causal, uma vez que a perda de conectividade da rede foi aparentemente bastante pequena, apesar de consistente.

Por fim, analisamos as Tabelas 29 e 30 que repetem a mesma análise, mas agora com grafos utilizando dados trimestrais ao invés de mensais. São mantidas as mesmas regras de formação dos grafos conforme a descrição anterior. A análise agora passa a ser em relação ao primeiro trimestre de 2020, que tomamos como trimestre-base.

Novamente, é possível observar perda de conectividade na rede em pequenos mas consistentes percentuais, afetando as métricas densidade, transitividade, grau médio, e grau médio ponderado. Essa conectividade é aparentemente recuperada no trimestre posterior. Por outro lado, as redes apresentam um ganho significativo de assortatividade de grau, indicando preferência dos vértices de se conectar a vértices semelhantes. A assortatividade de região e as métricas de caminhos mínimos médios apresentaram tímido aumento percentual. Essas outras métricas tiveram queda no trimestre seguinte, em sentido contrário ao apresentado.

6.2.2 Redes Seção-Seção

A seguir, serão analisadas alterações topológicas na estrutura das redes Seção-Seção no período que compreende janeiro de 2019 e outubro de 2020. Nesta análise, os dados foram mais uma vez dividido mês a mês e um grafo foi criado para cada mês do período estudado. Cada vértice destes grafos representa uma seção e cada aresta representa o valor total transacionado entre duas seções naquele mês. Usamos novamente a modelagem de grafo cíclico ponderado, sendo que cada aresta possui novamente um atributo indicando a distância entre os vértices igual ao inverso do valor transacionado.

As Tabelas 31 e 32 apresentam a evolução das métricas das redes Seção-Seção para cada mês. Cada linha mostra portanto a variação do valor das métricas do grafo montado para aquele mês em relação ao mês de janeiro de 2019, usado novamente como mês-base.

Assim como nas redes UF-UF, temos aqui uma baixa variação nos dados das métricas de densidade, transitividade, e grau médio, sendo que a magnitude da variação dessas métricas é menor que 2.5%, o que dificulta afirmações mais assertivas. Para essas três métricas, não existe uma tendência clara de crescimento ou redução de cada métrica no decorrer do tempo, sendo que comparando os meses imediatamente anteriores à crise com os meses mais críticos (abril e maio de 2020), há uma aparente queda de densidade e transitividade indicando uma leve redução na conectividade da rede, que apresenta evolução nos meses seguintes. Quando observamos o grau médio ponderado, uma métrica influenciada pelo volume mensal de transações, podemos então ver que nos meses mais críticos da crise o grau médio diminuiu, indicando um menor fluxo de transações entre os vértices das redes.

Tabela 31 – Variação percentual das métricas de redes mês a mês para o grafo de seções (Parte 1)

Mês	Densidade	Transitividade	Grau médio	Grau médio ponderado
01/2019	-	-	-	-
02/2019	-0.57%	-0.55%	-0.57%	+9.89%
03/2019	+0.57%	+0.31%	+0.57%	+25.54%
04/2019	-1.15%	-1.20%	-1.15%	+23.86%
05/2019	+1.15%	+0.15%	+1.15%	+36.92%
06/2019	-1.72%	-1.33%	-1.72%	+23.89%
07/2019	-0.57%	-1.00%	-0.57%	+28.37%
08/2019	-0.57%	-0.63%	-0.57%	+35.37%
09/2019	+0.00%	+0.18%	+0.00%	+32.63%
10/2019	-0.57%	+0.13%	-0.57%	+42.89%
11/2019	-1.72%	-1.25%	-1.72%	+43.94%
12/2019	+1.15%	+0.85%	+1.15%	+29.66%
01/2020	+2.30%	+1.87%	+2.30%	+28.85%
02/2020	+1.15%	+0.46%	+1.15%	+32.22%
03/2020	+1.72%	+0.98%	+1.72%	+36.10%
04/2020	-0.57%	-0.38%	-0.57%	-7.92%
05/2020	-0.57%	-0.47%	-0.57%	+12.47%
06/2020	-1.72%	-0.72%	-1.72%	+34.62%
07/2020	-1.72%	-1.97%	-1.72%	+62.52%
08/2020	+0.00%	-0.15%	+0.00%	+55.16%
09/2020	+1.72%	+1.24%	+1.72%	+60.33%
10/2020	+0.00%	-0.60%	+0.00%	+61.20%

Fonte: Dados da pesquisa.

Ao analisar a assortatividade de grau podemos ver uma variação nessas métricas sem tendência clara de crescimento ou redução para todo o período analisado. A métrica caminho mínimo médio também apresenta uma variação sem tendência clara e com baixa magnitude, com uma leve alta nos meses mais críticos da crise em relação aos meses imediatamente anteriores, mas não sendo o suficiente para indicar uma tendência. Quando analisamos o grau médio ponderado por outro lado, métrica influenciada pelo volume de transações, podemos verificar que nos meses de abril e maio de 2020 há uma variação significativa que indica aumento desta métrica em relação aos meses imediatamente anteriores e posteriores.

As Tabelas 33 e 34 mostram a mesma análise mas do ponto de vista trimestral em relação ao primeiro trimestre de 2020, tomado como trimestre-base.

Ao observarmos agora de forma agregada os dados trimestrais de 2020, podemos ver pela densidade, transitividade, e grau médio que há uma leve redução na conectividade da rede nos trimestres após o começo da pandemia em relação ao trimestre-base. Quando incluímos o grau médio ponderado, vemos que há uma queda significativa no segundo trimestre em relação ao trimestre-base e que há recuperação no trimestre seguinte.

Tabela 32 – Variação das métricas de redes mês a mês para o grafo de seções (Parte 2)

Mês	Assortatividade de grau	Caminho mínimo médio	Caminho mínimo médio ponderado
01/2019	-	-	-
02/2019	+2.76%	+0.45%	+29.22%
03/2019	-9.42%	-0.45%	-11.35%
04/2019	+4.86%	+0.90%	+2.19%
05/2019	+1.47%	-0.45%	+3.75%
06/2019	+16.41%	+0.90%	+5.20%
07/2019	+1.48%	+0.45%	-15.91%
08/2019	+6.16%	+0.45%	-13.54%
09/2019	-5.13%	-0.45%	-10.76%
10/2019	-1.58%	+0.00%	-15.12%
11/2019	10.43%	+0.90%	-38.71%
12/2019	-3.53%	-0.90%	-41.20%
01/2020	-14.48%	-1.80%	-25.39%
02/2020	-1.24%	-0.45%	-36.64%
03/2020	-7.40%	-0.90%	+22.58%
04/2020	-4.91%	+0.45%	+28.02%
05/2020	-8.57%	+0.45%	+114.33%
06/2020	-3.99%	+1.35%	+72.81%
07/2020	-4.68%	+1.80%	+65.47%
08/2020	-11.69%	+0.45%	+60.28%
09/2020	-10.17%	-0.90%	+56.97%
10/2020	-6.61%	+0.45%	-33.58%

Fonte: Dados da pesquisa.

A seguir, verificamos uma tendência de alta da assortatividade de grau durante o trimestre mais crítico da crise, retornando ao patamar anterior no trimestre seguinte. E ao analisar o caminho mínimo médio podemos ver que não há variação significativa em relação ao trimestre-base, mas quando fazemos a análise da métrica ponderada é possível verificar um aumento considerável no distância entre os vértices da rede.

Tabela 33 – Variação das métricas de rede a cada trimestre para o grafo de seções (Parte 1)

Mês	Densidade	Transitividade	Grau médio	Grau médio ponderado
2020/1	-	-	-	-
2020/2	-1.66%	-0.92%	-1.66%	-14.60%
2020/3	-1.10%	-0.59%	-1.10%	+20.35%

Fonte: Dados da pesquisa.

Tabela 34 – Variação das métricas de rede a cada trimestre para o grafo de seções (Parte 2)

Mês	Assortatividade de grau	Caminho mínimo médio	Caminho mínimo médio ponderado
2020/1	-	-	- -
2020/2	+11.16%	+0.93%	+122.98%
2020/3	-2.87%	+0.93%	+100.19%

Fonte: Dados da pesquisa.

6.2.3 Redes CNAE-CNAE

Por fim, serão apresentadas as alterações topológicas na estrutura das redes CNAE-CNAE no período que compreende janeiro de 2019 e outubro de 2020.

Tabela 35 – Variação percentual das métricas de redes mês a mês para o grafo de CNAEs (Parte 1)

Mês	Densidade	Transitividade	Grau médio	Grau médio ponderado
01/2019	-	-	-	-
02/2019	+4.35%	+1.75%	+4.45%	+9.79%
03/2019	+5.96%	+2.29%	+6.15%	+25.32%
04/2019	+9.60%	+3.67%	+9.70%	+23.75%
05/2019	+11.46%	+4.22%	+11.66%	+36.67%
06/2019	+7.55%	+2.71%	+7.55%	+23.89%
07/2019	+12.16%	+4.16%	+12.06%	+28.48%
08/2019	+12.70%	+4.64%	+12.80%	+35.25%
09/2019	+11.38%	+4.45%	+11.48%	+32.51%
10/2019	+13.87%	+5.36%	+14.27%	+42.38%
11/2019	+12.57%	+4.84%	+12.67%	+43.81%
12/2019	+6.15%	+2.35%	+6.62%	+29.09%
01/2020	+9.87%	+3.64%	+10.27%	+28.39%
02/2020	+8.58%	+3.01%	+8.87%	+31.87%
03/2020	+9.29%	+3.80%	+9.68%	+35.62%
04/2020	-2.40%	-0.24%	-2.40%	-7.92%
05/2020	+1.98%	+1.67%	+2.07%	+12.37%
06/2020	+7.94%	+3.99%	+8.13%	+34.38%
07/2020	+11.86%	+5.21%	+12.46%	+61.65%
08/2020	+10.99%	+4.48%	+10.89%	+55.30%
09/2020	+9.68%	+4.69%	+10.26%	+59.47%
10/2020	+7.63%	+3.96%	+8.02%	+60.62%

Fonte: Dados da pesquisa.

Da mesma forma, os dados foram dividido mês a mês e um grafo foi criado para cada mês do período estudado. Cada vértice destes grafos representa um CNAE e cada aresta representa o valor total transacionado entre dois CNAEs naquele mês. Usamos novamente a modelagem de grafo cíclico ponderado, sendo que cada aresta possui novamente um atributo

indicando a distância entre os vértices igual ao inverso do valor transacionado. Cada vértice possui também um atributo indicando a seção a que pertence aquele CNAE para uso na métrica de assortatividade.

As Tabelas 35 e 36 apresentam as métricas desta análise, apresentando a variação percentual de cada métrica para cada mês em relação ao mês-base de janeiro de 2019.

Tabela 36 – Variação das métricas de redes mês a mês para o grafo de CNAEs (Parte 2)

Mês	Assortatividade de grau	Assortatividade de seção	Caminho mínimo médio	Caminho mínimo médio ponderado
01/2019	-	-	-	-
02/2019	+0.38%	-6.76%	-0.72%	-9.37%
03/2019	+1.52%	-22.04%	-0.93%	+35.22%
04/2019	+1.80%	-8.60%	-1.49%	+33.50%
05/2019	+2.69%	-23.49%	-1.83%	-13.22%
06/2019	+1.02%	-24.50%	-1.15%	-15.84%
07/2019	+1.76%	+4.38%	-1.77%	-17.73%
08/2019	+1.77%	-16.28%	-1.64%	+3.81%
09/2019	+0.86%	+3.95%	-1.65%	-29.89%
10/2019	+1.55%	-4.69%	-1.83%	-35.86%
11/2019	+1.02%	-8.01%	-1.83%	-36.74%
12/2019	+0.52%	-10.46%	-1.08%	-30.56%
01/2020	+2.06%	-6.52%	-1.50%	-28.02%
02/2020	+1.27%	+9.35%	-1.30%	-23.95%
03/2020	+1.61%	+1.73%	-1.40%	-23.14%
04/2020	-0.02%	+8.10%	+0.58%	-8.54%
05/2020	+1.14%	+16.10%	-0.27%	+9.73%
06/2020	+2.35%	+0.88%	-1.13%	+16.51%
07/2020	+3.60%	+15.25%	-1.66%	+1.86%
08/2020	+1.83%	+12.58%	-1.57%	-26.79%
09/2020	+2.24%	+16.47%	-1.21%	+5.56%
10/2020	+1.30%	-4.01%	-0.98%	-17.64%

Fonte: Dados da pesquisa.

Mais uma vez, podemos verificar a redução da conectividade das redes através da densidade, da transitividade e do grau médio da rede em relação ao mês-base, com as três métricas apresentando redução em relação aos meses imediatamente anteriores ao início da crise e se recuperando nos meses posteriores aos meses mais críticos. Essa métricas agora apresentam uma variação de magnitude significativa, sendo mais importante que nas redes analisadas anteriormente. Se considerarmos o grau médio ponderado essa diferença se torna ainda mais significativa, também com tendência de redução durante os meses mais críticos e posterior recuperação.

Analisando a assortatividade de grau das redes, é possível ver que existe pouca variação em relação ao mês-base, e que no mês de abril de 2020 há uma variação diferente para leve disas-

sortatividade, voltando para os patamares anteriores nos meses seguintes. Já quando observamos a assortatividade de seção, vemos uma variação de magnitude mais significativa, e que apresenta tendência positiva nos meses da pandemia que se mantém até setembro de 2020. Essa tendência de aumento na assortatividade de seção é interessante pois indicaria que na rede de CNAEs as seções tem alguma preferência de se conectar com seções semelhantes.

Por fim, analisando o caminho mínimo médio, é possível verificar certa estabilidade no período analisado em relação ao mês-base, com exceção dos meses de abril de maio de 2020 quando aparentemente há um aumento da distância entre os vértices das redes. Quando incluímos ponderação pelo valor transacionado na análise, essa diferença se torna ainda mais evidente, com uma alta que impacta também meses posteriores aos meses mais críticos.

Tabela 37 – Variação das métricas de rede a cada trimestre para o grafo de CNAEs (Parte 1)

Mês	Densidade	Transitividade	Grau médio	Grau médio ponderado
2020/1	-	-	-	-
2020/2	-4.91%	-1.09%	-4.91%	-14.60%
2020/3	+0.82%	+0.98%	+0.91%	+20.25%

Fonte: Dados da pesquisa.

Tabela 38 – Variação das métricas de rede a cada trimestre para o grafo de CNAEs (Parte 2)

Mês	Assortatividade de grau	Assortatividade de seção	Caminho mínimo médio	Caminho mínimo médio ponderado
2020/1	-	-	-	-
2020/2	-0.21%	+8.07%	+0.98%	+71.35%
2020/3	+0.66%	+3.81%	+0.05%	+26.02%

Fonte: Dados da pesquisa.

As Tabelas 37 e 38 complementam a análise anterior fazendo novamente a análise agora agregando os dados em trimestres e mostrando a variação em relação ao primeiro trimestre de 2020, o trimestre-base.

Novamente, podemos verificar a perda de conectividade da rede no trimestre mais crítico, o segundo de 2020, com uma redução na densidade, transitividade, e grau médio da rede, apresentando uma retomada no trimestre seguinte. Incluindo o grau médio ponderado, novamente essa variação se mostra mais evidente.

Por fim, analisando a variação das métricas de assortatividade, a assortatividade de grau apresenta leve redução no trimestre mais crítico seguida de leve retomada no trimestre seguinte. A assortatividade de seção apresenta alta no trimestre mais crítico e regularização no trimestre seguinte. A distância entre os vértices medida através do caminho mínimo médio apresenta leve

tendência de alta nos trimestres da pandemia e se acentuam quando usamos a análise da métrica ponderada.

6.2.4 Conclusões

Nesta seção, fizemos uma descrição da oscilação de diversas métricas de redes através e uma modelagem temporal, comparando com as redes se comportam através do tempo e eventuais tendências.

Para o estudo das redes de topologia mais simples, UF-UF e Seção-Seção, a variação das métricas apresentou uma variação de baixa magnitude, muitas vezes menor que 1%, mas puderam apresentar algumas tendências consistentes mesmo assim. Pudemos ver que mesmo nessas redes há uma perda de conectividade das redes, principalmente nos meses e trimestres mais críticos, evidenciada quando levamos em consideração as formulações que levam em conta os valores transacionados.

As redes CNAE-CNAE que tem uma topologia mais complexa apresentaram a mesma tendência mas de forma mais significativa, sendo que pudemos observar uma magnitude maior nas variações de cada métricas, refletindo da mesma a perda de conectividade da rede.

A perda de conectividade é um sintoma citado nos trabalhos descritos na seção 1.3, que analisam dados de áreas significativamente diferentes, como o transporte aéreo global. Assim como o trabalho citado sobre a correlação do impacto da pandemia no mercado de ações apontava para um aumento do caminho mínimo médio, uma tendência também encontrada neste trabalho.

6.3 Detecção de impacto econômico

Nesta seção, nas últimas seções pudemos verificar como a base de dados analisada se comportou durante a pandemia, sob diferentes perspectivas. Agora vamos apresentar os métodos tentados para a detecção e classificação de impacto econômico, visando tentar identificar de forma automatizada entidades afetadas pela crise sob diferentes modelagens. Todos os conjuntos de dados a seguir foram normalizados utilizando norma euclideana, ou norma L2.

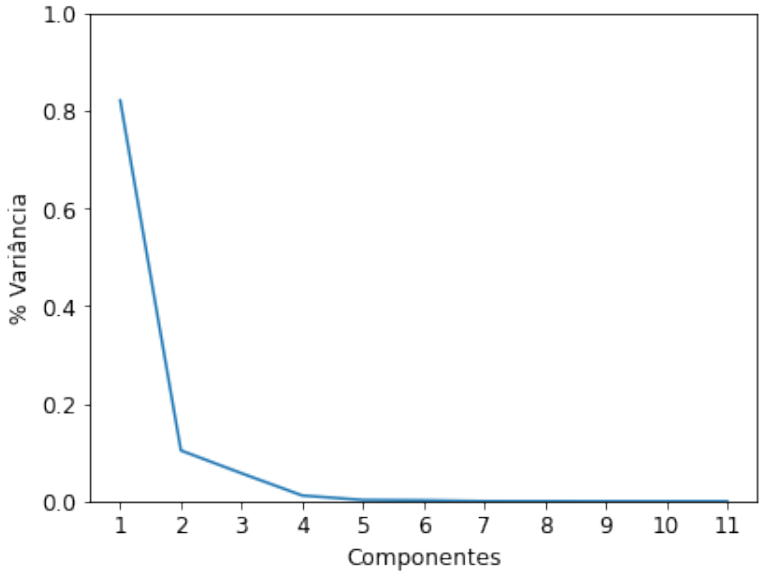
6.3.1 Dados de UFs

Para esta análise, criamos uma matriz foram reunidos os dados descritos na seção 6.1.1, as métricas de caracterização da rede UF-UF, com algumas métricas relacionadas ao volume transacionado por cada UF: volume total, quantidade de transações, e valor médio de cada transação. Assim, foi criada uma matriz onde cada linha representa uma UF e cada coluna o valor de uma das métricas de redes ou de transações, totalizando onze variáveis.

Também foi uma criada uma segunda tabela com o impacto mensal e total para cada UF, conforme descrito na seção 5.2. O objetivo aqui é observar o comportamento de classificadores

ao tentar decidir a partir dos dados das métricas obtidas se é possível classificar uma UF como impactada ou não.

Figura 27 – Percentual de variância explicada de cada componente principal para dados de UFs



Fonte: Dados da pesquisa.

Foi feita uma análise de componentes principais a partir dos dados. A Figura 27 ilustra o percentual da variância explicada pelo conjunto de dados explicada por cada componente principal. As duas primeiras componentes principais concentram a maior parte da variância explicada pelo conjunto de dados.

Tabela 39 – Direção de cada métrica em relação a cada uma das duas componentes principais

Componente	1	2
Valor total	0.4746	0.1108
Quantidade de transações	0.4708	0.1544
Valor médio por transação	-0.0077	-0.3831
Grau	0.0009	-0.0101
Coeficiente de agrupamento	0.3311	-0.2535
Excentricidade	-0.0226	0.2543
Caminho mínimo médio	-0.0010	0.0108
Caminho mínimo médio ponderado	-0.1081	0.7227
Centralidade de informação	0.0976	-0.3850
Centralidade de autovalores	0.4783	0.1268
PageRank	0.4392	0.0377

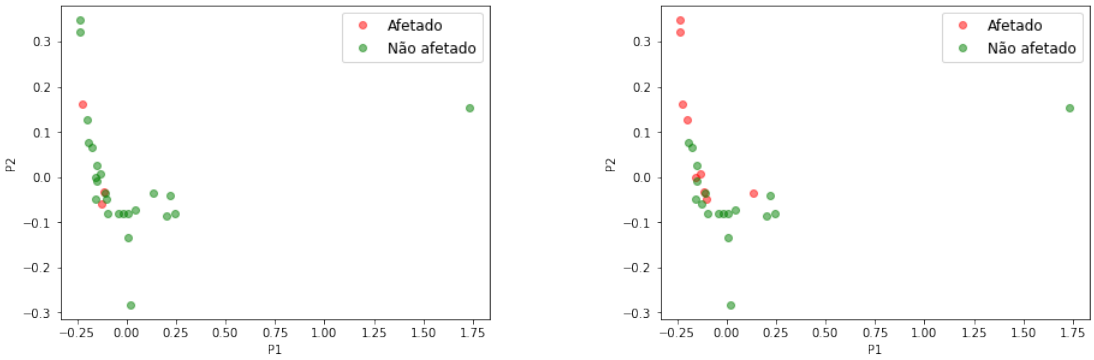
Fonte: Dados da pesquisa.

Na Tabela 39 são descritas as direções de cada variável em relação à cada uma das duas componentes principais. Para a primeira componente se destacam a magnitude do valor total e

da quantidade de transações, além da centralidade de autovalores e de informação, enquanto para a segunda componente se destacam o valor médio por transação e novamente a centralidade de informação. As variáveis com maior magnitude terão maior impacto decisivo nos classificadores que usaremos posteriormente.

As Figuras 28a e 28b ilustram a projeção dos dados em relação às duas componentes principais para dar uma noção do espalhamento dos dados, em relação respectivamente ao impacto mensal e impacto total.

Figura 28 – Visualização dos dados de UFs projetados sobre as duas componentes principais



(a) Impacto mensal

(b) Impacto total

Fonte: Dados da pesquisa.

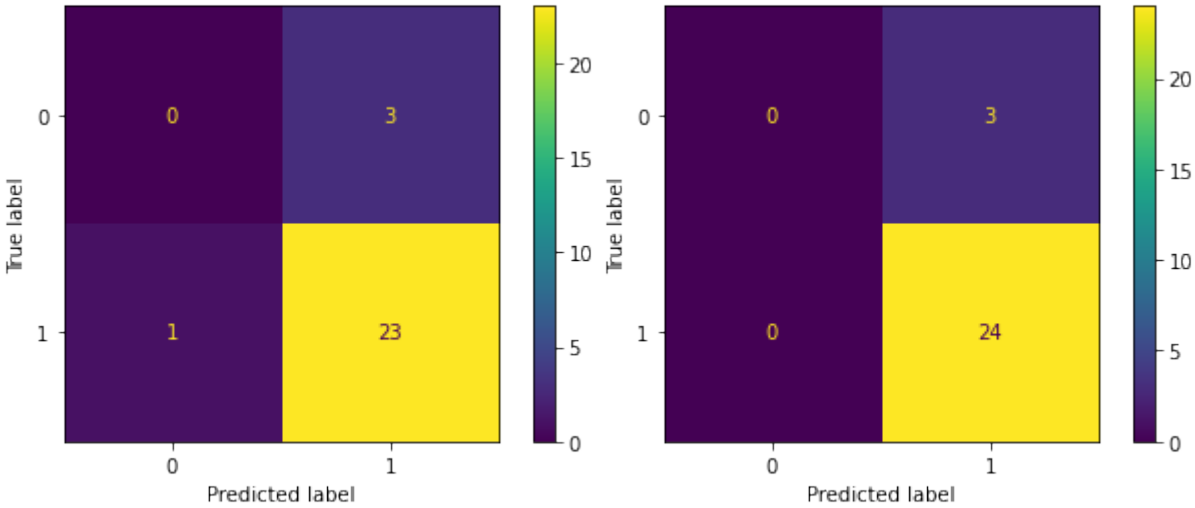
O conjunto de dados possui algum espalhamento nos dados, e parece ter algumas UFs um pouco mais distantes do restante. Porém, não existem regiões onde é clara a distinção de afetados e não afetados tanto para o impacto mensal quanto total. No caso do impacto total, existem UFs afetadas que parecem ocupar uma região próxima onde essas poderiam ser separadas linearmente, enquanto outras permanecem em uma região pouco divisível. Também é importante notar que ambos conjuntos de dados são desbalanceados, o que é um desafio para classificadores.

Ao aplicarmos classificadores sobre os dados foram obtidos os resultados ilustrados nas Figuras 29, 30 e 31. A primeira se refere aos dados de impacto mensal e as seguintes ao impacto total, conforme descrito anteriormente. Para validação da técnica, foi utilizado o método *leave-one-out*.

Para a tentativa de classificação dos dados de UFs para impacto mensal, três classificadores tiveram o mesmo resultado: Random Forest, Regressão Logística, e SVC, onde todas as UFs foram classificadas como positivos. Já utilizando o classificador KNN, o resultado foi semelhante mas uma das UFs que deveria ser classificada como positivo foi agora classificada como negativo. Apesar de uma alta acurácia, o resultado não é interessante uma vez que temos uma grande quantidade de falsos positivos para todos os métodos, faltando generalização.

Para os dados de impacto total, o resultado foi um pouco mais positivo. As maiores acurácias foram apresentadas pelos classificadores Regressão Logística e SVC, com cerca de

Figura 29 – Matrizes de confusão para classificadores aplicados sobre dados de UFs para impacto mensal

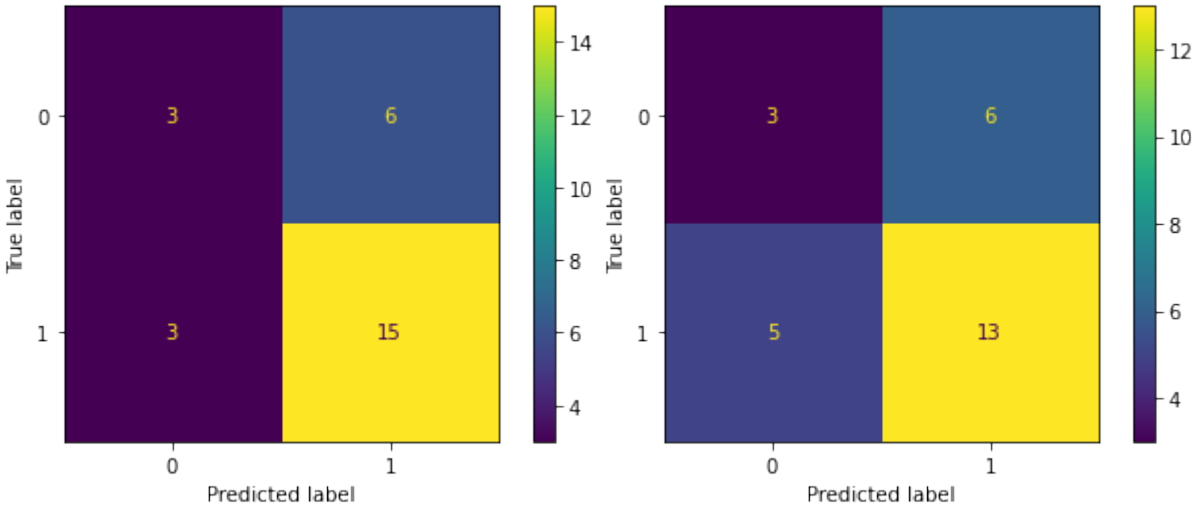


(a) KNN (b) Random Forest, Regressão Logística, e SVC

Fonte: Dados da pesquisa.

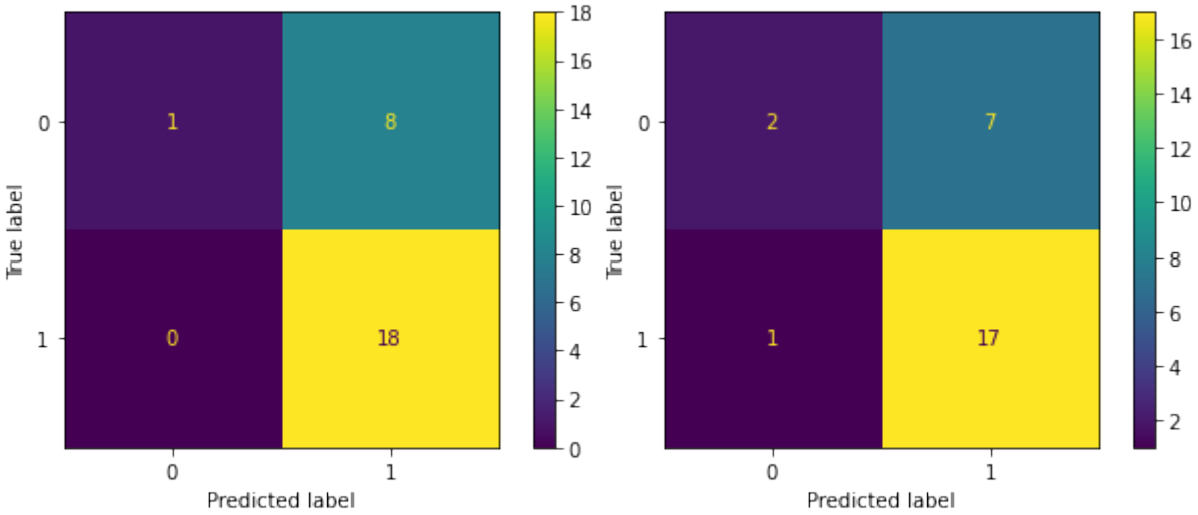
70%, seguido de KNN com 66% e Regressão Logística com 59%. Evidentemente, a acurácia para este conjunto de dados desbalanceado esconde um grande problema dessa classificação, que é conseguir uma boa especificidade para este conjuntos de dados. Todos os classificadores tiveram problemas com a quantidade de falsos positivos e falsos negativos. Nenhum dos classificadores portanto conseguiu efetivamente alcançar uma generalização aceitável.

Figura 30 – Matrizes de confusão para classificadores aplicados sobre dados de UFs para impacto total (Parte 1)



(a) KNN (b) Random Forest

Figura 31 – Matrizes de confusão para classificadores aplicados sobre dados de UFs para impacto total (Parte 2)



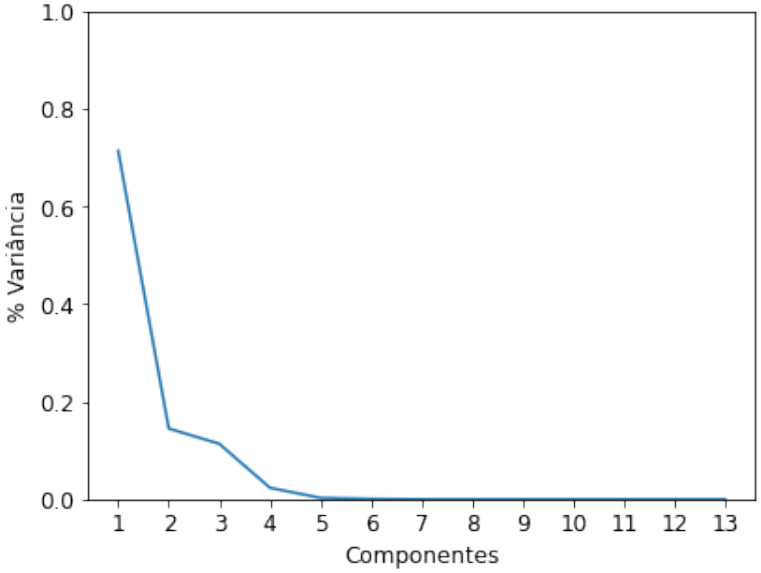
(a) Regressão Logística (b) SVC

Fonte: Dados da pesquisa.

6.3.2 Dados de Seções

Repetimos então a análise, construindo uma matriz agora usando os dados de seções, com os dados descritos na seção 6.1.2 junto das métricas de volume total, quantidade de transações, e valor médio de transações.

Figura 32 – Percentual de variância explicada de cada componente principal para dados de seções



Fonte: Dados da pesquisa.

A matriz resultante possui uma linha para cada seção analisada e uma coluna para cada

métrica, totalizando treze variáveis. Utilizamos também os dados de impacto mensal e total por seção descritos na seção 5.3.

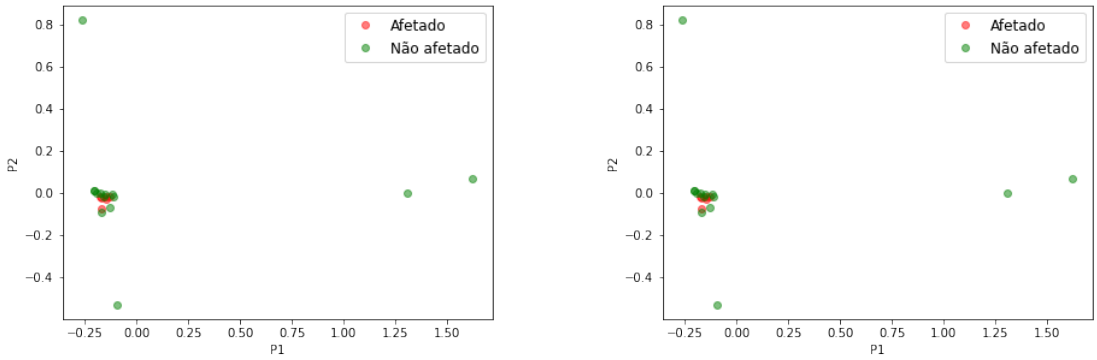
Tabela 40 – Direção de cada métrica em relação a cada uma das duas componentes principais para os dados de seções

Componente	1	2
Valor total	0.4238	0.0153
Quantidade de transações	0.3972	0.0648
Valor médio por transação	-0.0133	-0.5534
Grau	0.0243	-0.1738
Coeficiente de agrupamento	0.3633	-0.0632
Excentricidade	-0.0702	-0.0082
Caminho mínimo médio	-0.0195	0.1362
Caminho mínimo médio ponderado	-0.0504	0.7709
Centralidade de proximidade	0.0171	-0.0829
Centralidade de informação	0.0120	-0.1816
Centralidade de autovalores	0.4237	0.0396
PageRank	0.4121	0.0084
Centralidade de intermediação	0.4151	0.0285

Fonte: Dados da pesquisa.

Assim como anteriormente, foi aplicada a análise das componentes principais dessa matriz. A Figura 32 mostra o percentual da variância total explicada pelo conjunto de dados para cada componente. Novamente, as duas primeiras componentes explicam a maior parte da variância do conjunto de dados.

Figura 33 – Visualização dos dados de seções projetados sobre as duas componentes principais



(a) Impacto mensal

(b) Impacto total

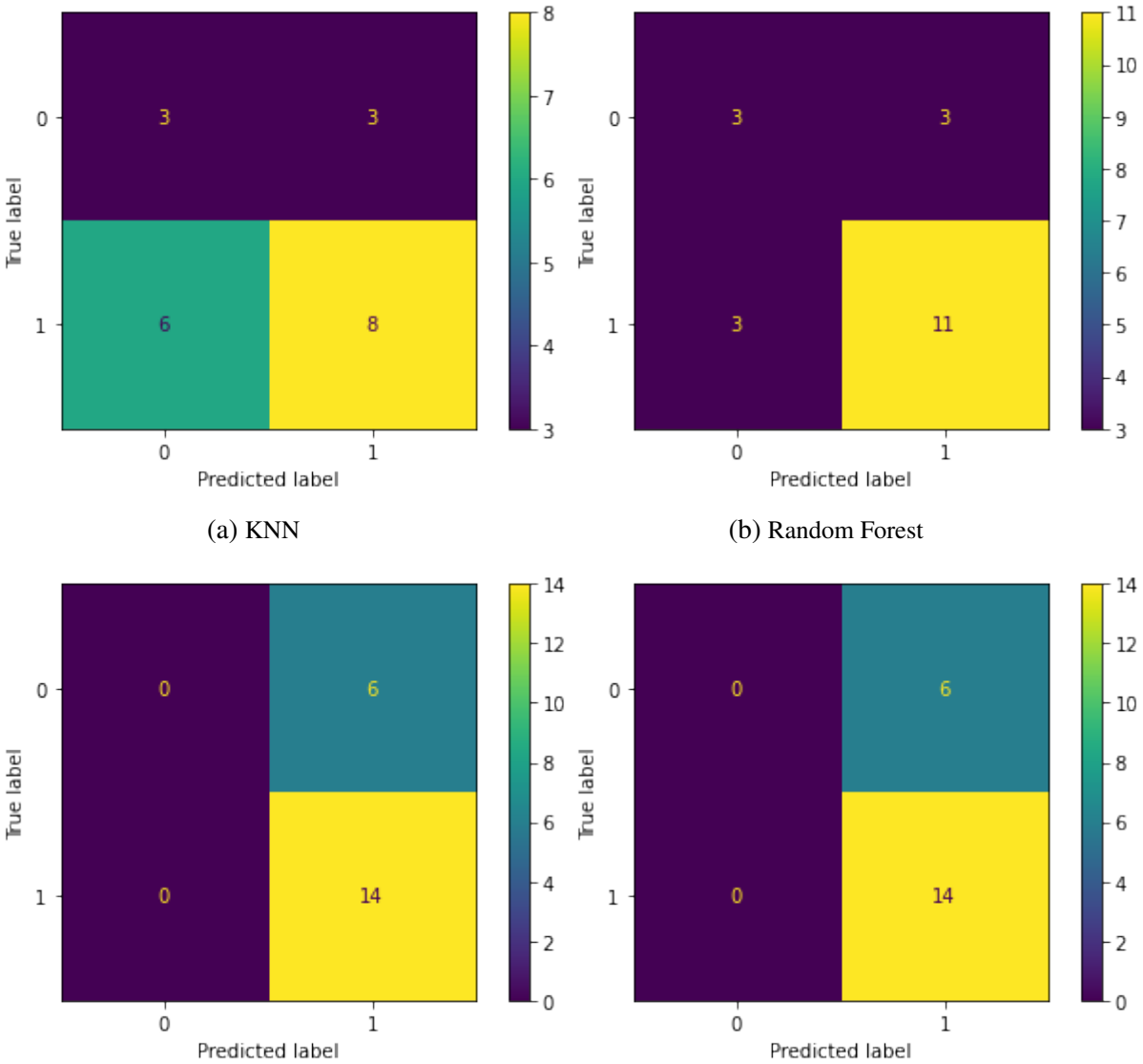
Fonte: Dados da pesquisa.

A seguir, na Tabela 40, é descrita a direção de cara variável em relação às duas principais componentes. O valor total novamente tem uma participação importante, mas as métricas de centralidade de autovalores, *pagerank*, e centralidade de intermediação apresentam também

magnitude significativa em relação à primeira componente, enquanto as maiores magnitudes para a segundo componentes pertencem às váriáveis valor médio por transação e caminho mínimo médio ponderado.

A Figura 33 apresenta o espalhamento dos dados projetos sobre as duas componentes principais representando o impacto mensal e total. É possível ver que os dados são pouco espalhados e se concentram em uma região principal do espaço de variáveis, além de possuir quatro seções distantes com um aparente comportamento de *outliers*. Novamente, os dados são bastante desbalanceados.

Figura 34 – Matrizes de confusão para classificadores aplicados sobre dados de seções para impacto mensal



(a) KNN

(b) Random Forest

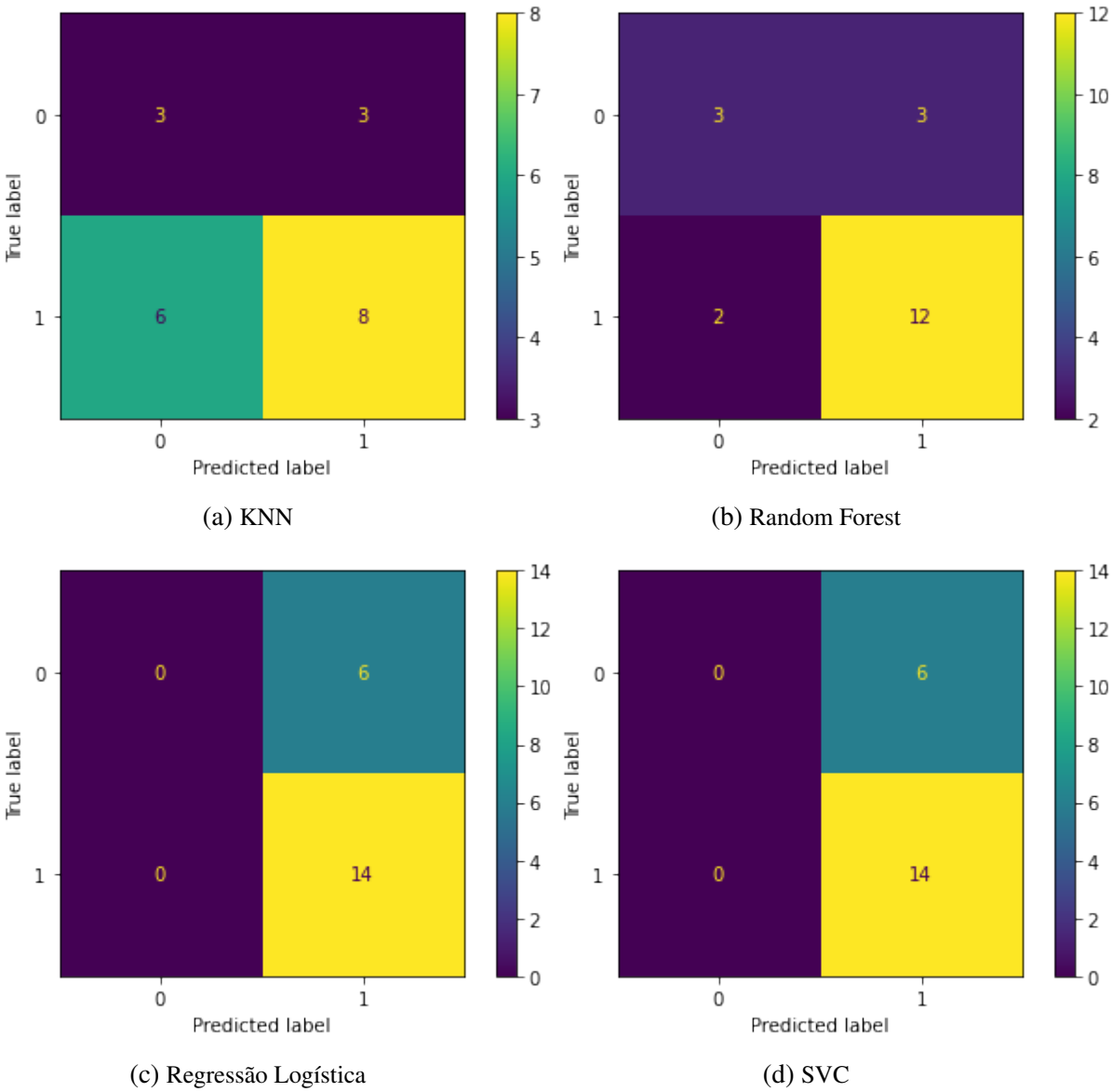
(c) Regressão Logística

(d) SVC

Fonte: Dados da pesquisa.

Foram aplicados também classificadores sobre os dados, que são ilustrados nas Figuras 34

Figura 35 – Matrizes de confusão para classificadores aplicados sobre dados de seções para impacto total



(a) KNN

(b) Random Forest

(c) Regressão Logística

(d) SVC

Fonte: Dados da pesquisa.

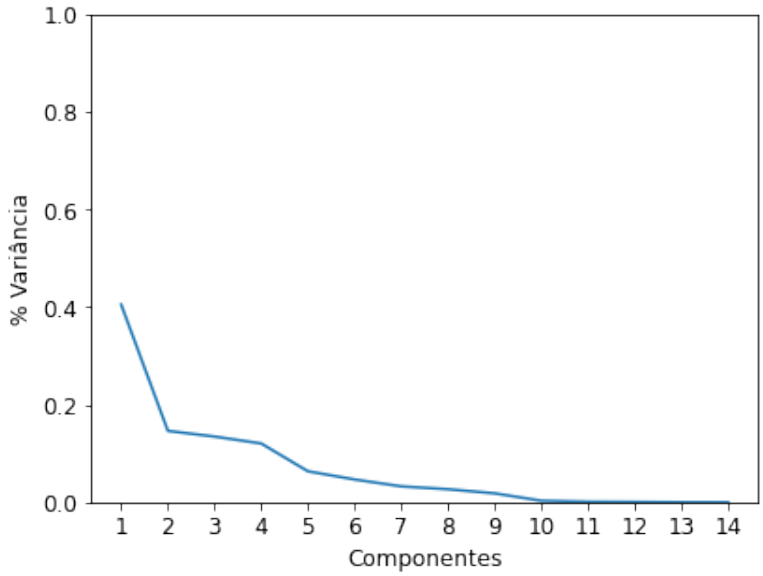
e 35, referentes ao impacto mensal e total respectivamente. Para a validação foi novamente utilizada a técnica *leave-one-out*.

Novamente, os classificadores não foram capazes de apresentar boa generalização sobre os dados, sendo que dois deles classificaram todas as seções como afetadas pela crise. O classificador KNN apresentou baixa precisão, com uma alta taxa de falsos positivos. E o classificador Random Forest apresentou um resultado razoável, com uma acurácia de 70%, precisão de 78% e especificidade de 50%.

6.3.3 Dados de CNAEs

Por fim, aplicamos a mesma técnica anterior sobre os dados de CNAEs.

Figura 36 – Percentual de variância explicada de cada componente principal para dados de CNAEs



Fonte: Dados da pesquisa.

Montamos uma matriz novamente agregando os dados de volume total transacionado, quantidade de transações, valor médio de transações, seção do CNAE (discretizada), e as métricas de redes da seção 6.1.3, totalizando catorze variáveis.

Tabela 41 – Direção de cada métrica em relação a cada uma das duas componentes principais para os dados de CNAEs

Componente	1	2	3	4
Seção	0.5035	-0.0108	0.0966	0.0441
Valor total	0.4563	0.3150	-0.0079	-0.1132
Quantidade de transações	0.0612	-0.4170	0.6541	0.1840
Valor médio por transação	0.1895	-0.2848	-0.3259	0.0631
Grau	0.1765	-0.3230	0.3309	0.1040
Coeficiente de agrupamento	-0.0217	0.0357	0.0442	-0.0124
Excentricidade	-0.0328	0.0529	0.0559	-0.0012
Caminho mínimo médio	-0.0295	0.3470	-0.0146	0.9292
Caminho mínimo médio ponderado	0.0388	-0.0607	-0.0697	0.0092
Centralidade de proximidade	0.0207	-0.0689	-0.0039	-0.0521
Centralidade de informação	0.4090	0.4709	0.1726	-0.1805
Centralidade de autovalores	0.4696	-0.1466	0.0038	0.0944
PageRank	0.2737	-0.3887	-0.5499	0.1673
Centralidade de intermediação	-0.0225	0.1185	-0.0681	0.0283

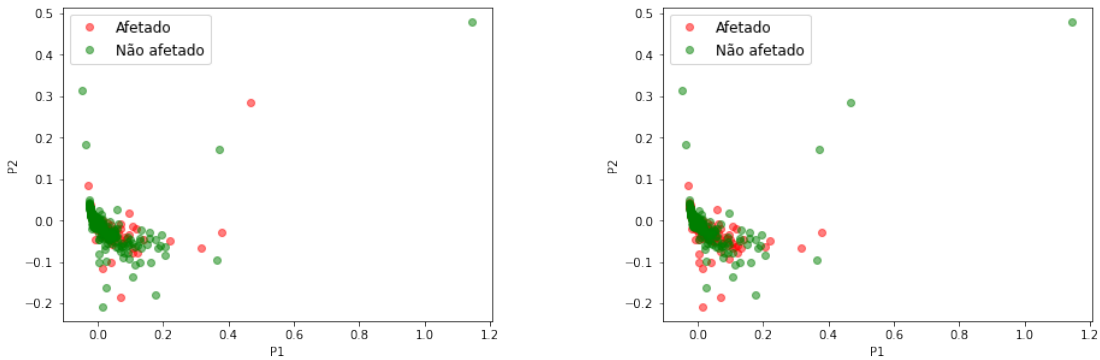
Fonte: Dados da pesquisa.

A Figura 36 apresenta o percentual da variância explicada pelo conjunto de dados para cada componente principal dessa matriz. Ao contrário dos conjuntos anteriores, há maior distribuição da variância explicada por entre as componentes. São necessárias agora quatro componentes principais para totalizar 80.6% da variância total do conjunto de dados.

A Tabela 41 mostra a direção de cada variável em relação a cada uma das quatro componentes principais. Algumas métricas já citadas nas análises anteriores novamente se sobressaem: valor total, quantidade de transações, centralidade de informação, centralidade de autovalores, *pagerank*. A variável seção, adicionada para melhorar o espalhamento dos dados, também tem participação importante principalmente na primeira componente principal.

A Figura 36 mostra o espalhamento dos dados projetados sobre as duas componentes principais em relação ao impacto mensal e total. É possível novamente ver que não há um espalhamento significativo, com exceção de alguns *outliers*, e as duas classes se sobrepõem uma a outra sem regiões bem definidas. O conjunto considerando impacto mensal é bastante desbalanceado, enquanto o conjunto que considera impacto total é menos desbalanceado.

Figura 37 – Visualização dos dados de CNAEs projetados sobre as duas componentes principais



(a) Impacto mensal

(b) Impacto total

Fonte: Dados da pesquisa.

Mais uma vez aplicamos classificadores sobre os dados para identificação do impacto mensal e total. Desta vez, dado o tamanho do dataset utilizados a técnica k-fold estratificado para validarmos os dados obtidos através dos classificadores. As Figuras 38 e 39 mostram os resultados obtidos, através das matrizes de confusão e curva ROC.

Para o conjunto de dados de impacto mensal, novamente os classificadores falharam em identificar verdadeiros negativos. O classificador que melhor conseguiu generalizar os dados foi novamente Random Forest, com uma precisão de 95% mas uma especificidade de apenas 12%. Também para o conjunto de dados de impacto total, significativamente mais balanceado, os resultados apresentaram pouca generalização. Todos os métodos apresentaram uma alta taxa de falsos positivos e negativos.

6.3.4 Conclusões

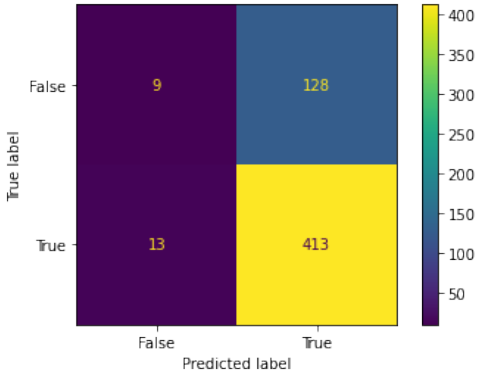
Nesta seção, montamos uma tentativa de classificação de entidades impactadas e não impactadas pela crise causada pela pandemia. Foram utilizadas três abordagens, por UF, por seção, e por CNAE, e fizemos um estudo da variância e de como os dados se espalham.

Já na análise de componentes principais, foi possível perceber para os três conjuntos de dados que eles não possuem um espalhamento adequado para que um classificador esteja apto a separar os conjuntos de dados. Sendo que os classificadores escolhidos foram utilizados justamente para tentar utilizar diferentes abordagens de classificação.

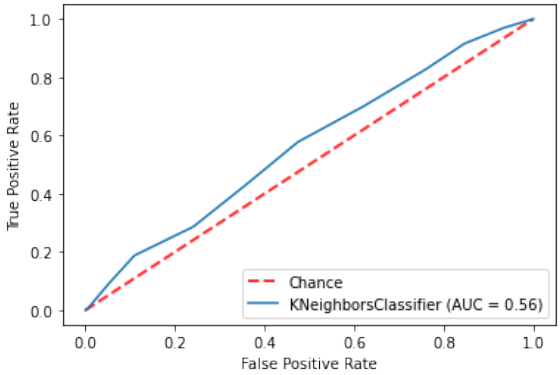
Foram utilizadas também algumas técnicas para tentar reduzir o *overfitting* nos dados: utilizamos a análise de componentes principais para selecionar apenas dimensões de maior taxa de explicabilidade, utilizamos técnicas de discretização dos dados para tentar uma melhor separação dos dados, e algumas variáveis não apresentadas aqui também foram inseridas no conjunto de dados e depois removidas ou por serem redundante com alguma outra variável, ou por serem insuficientes para um melhor espalhamento dos dados.

O resultado aqui obtido não é o ideal, mas de forma alguma eles significam que seja impossível fazer um classificador que esteja apto a esses conjuntos de dados. Outros classificadores não utilizados aqui podem ter a capacidade de obter melhores resultados, e até mesmo os classificadores que foram utilizados podem apresentar resultados melhores com outros tratamentos aplicados aos dados ou com a adição de novas variáveis.

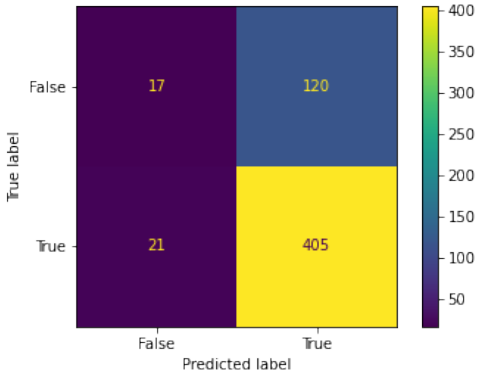
Figura 38 – Resultado das tentativas de classificação para cada método aplicado sobre dados de CNAEs para impacto mensal



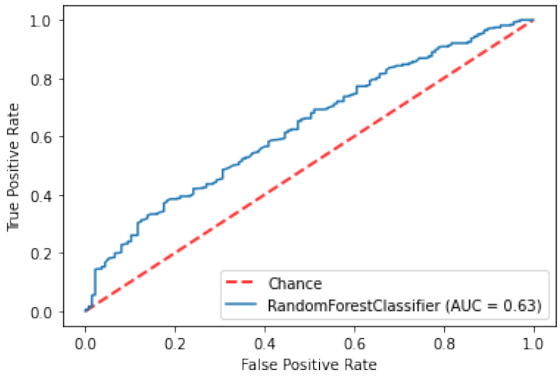
(a) Matriz de confusão para KNN



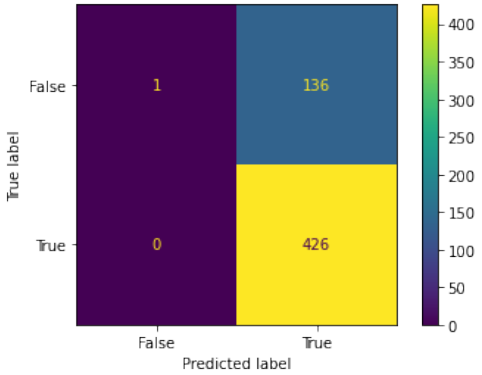
(b) Curva ROC para KNN



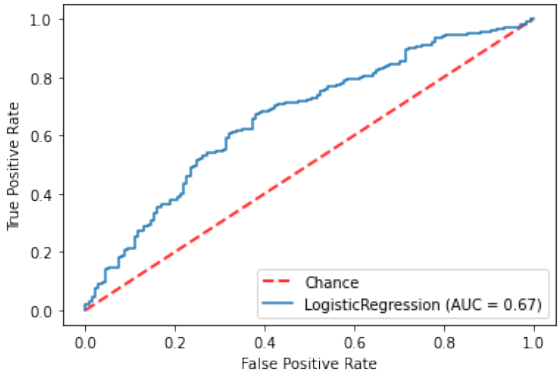
(c) Matriz de confusão para Random Forest



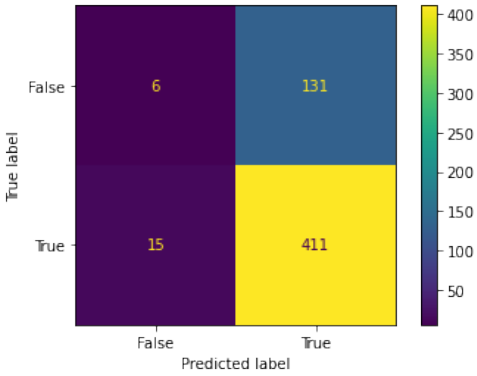
(d) Curva ROC para Random Forest



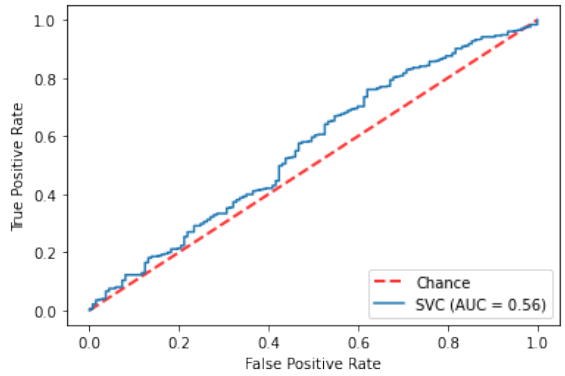
(e) Matriz de confusão para Regressão logística



(f) Curva ROC para Regressão Logística



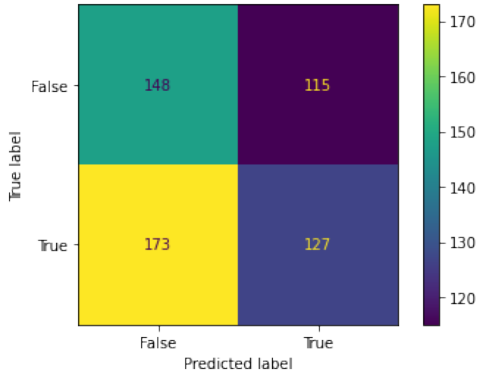
(g) Matriz de confusão para SVC



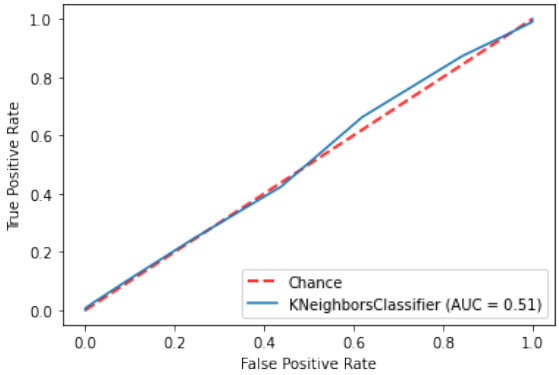
(h) Curva ROC para SVC

Fonte: Dados da pesquisa.

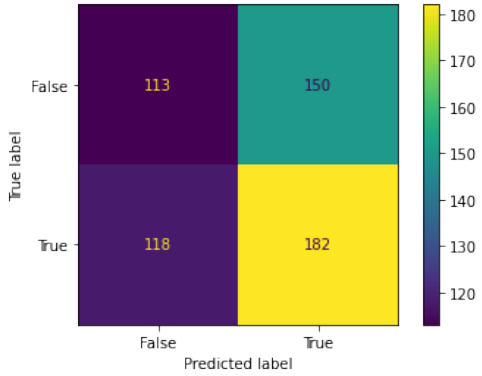
Figura 39 – Resultado das tentativas de classificação para cada método aplicado sobre dados de CNAEs para impacto total



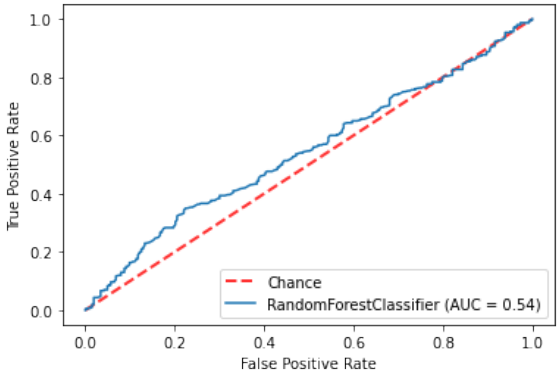
(a) Matriz de confusão para KNN



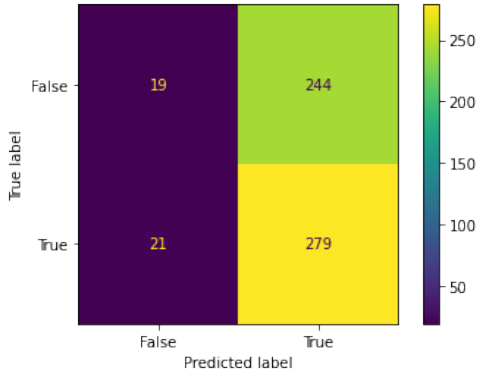
(b) Curva ROC para KNN



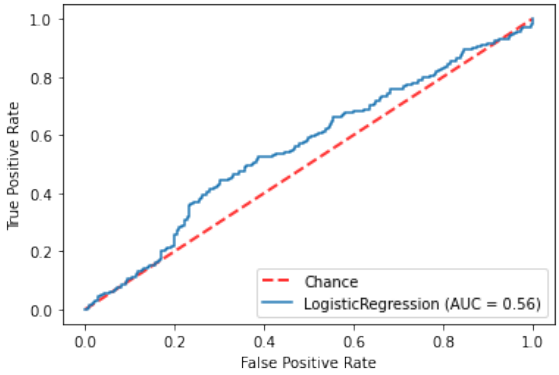
(c) Matriz de confusão para Random Forest



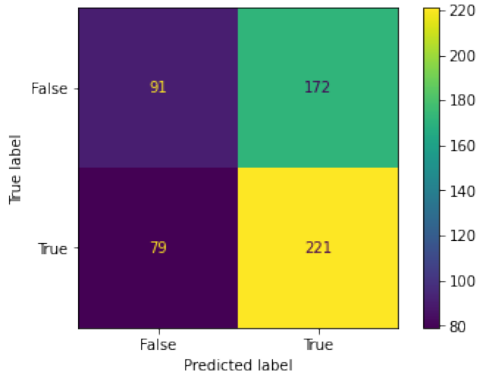
(d) Curva ROC para Random Forest



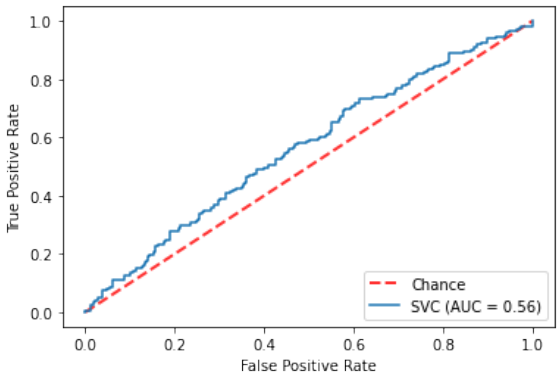
(e) Matriz de confusão para Regressão logística



(f) Curva ROC para Regressão Logística



(g) Matriz de confusão para SVC



(h) Curva ROC para SVC

Fonte: Dados da pesquisa.

CAPÍTULO

7

CONCLUSÃO

Este capítulo encerra esse trabalho com as conclusões obtidas no decorrer do mesmo.

7.1 Pontos positivos e negativos

Este trabalho obteve sucesso em demonstrar o valor dos dados de documentos fiscais, obtendo análises significativas a partir de dados que muitas vezes são vistos apenas como uma burocracia pelas empresas.

A análise descritiva dos dados e do impacto econômico da pandemia apresentam também uma nova perspectiva do impacto de um evento tão importante, mesmo que estejamos lidando neste trabalho com uma amostra de dados em relação à economia brasileira.

A análise do impacto da pandemia do ponto de vista topológico usando redes complexas colabora com outras análises semelhantes, sendo que conseguimos alcançar resultados parecidos com trabalhos prévios, como a perda de conectividade para grafos de diferentes complexidades e topologias.

O processamento de documentos fiscais possui também um desafio recorrente em trabalhos analíticos sobre eles que é o tratamento de inconsistências e incorreções nos dados. Apesar de ser um documento de alta criticidade jurídica, as informações presentes nos documentos nem sempre estão estruturadas de forma adequada e também podem apresentar problemas, o que pode causar ruído nas análises. Neste trabalho, apesar dos esforços de pré-processamento, foram encontrados alguns problemas do tipo e um importante desafio a ser ainda lapidado são novas formas de pré-processar esses documentos para melhor confiabilidade dos dados.

O estudo não conseguiu apresentar um classificador eficaz para identificar entidades afetadas pela crise, e também não apresentou análises mais granulares, como por exemplo sobre empresas e produtos.

7.2 Próximos passos

Como sugestão para trabalhos futuros, existem outros documentos fiscais com valiosos dados que não foram utilizados aqui. Por exemplo, o Conhecimento de Transporte Eletrônico é um documento de grande valor para análises de cadeias de transporte, como trabalhos anteriores já o fizeram. Outros documentos fiscais são portanto uma fonte rica de dados para futuros trabalhos.

Neste trabalho não foram utilizadas ferramentas de processamento distribuído que podem auxiliar em análises de maiores quantidades de dados, e poderiam auxiliar por exemplo em uma análise de impacto com melhor granularidade.

Outro desafio interessante seria transformar as informações aqui encontradas, de métricas e topologias de redes complexas, em produtos a serem utilizados pelas empresas brasileiras, transformando burocracia em oportunidade.

O autor deste trabalho finaliza esperando que as análises e esforços aqui feitas sejam úteis no futuro em outros trabalhos.

REFERÊNCIAS

ARQUIVEI. 2021. Disponível em: <https://arquivei.com.br/>. Acesso em: 25/01/2021. Citado na página 39.

ASLAM, F.; MOHMAND, Y. T.; FERREIRA, P.; MEMON, B. A.; KHAN, M.; KHAN, M. Network analysis of global stock markets at the beginning of the coronavirus disease (covid-19) outbreak. **Borsa Istanbul Review**, Elsevier, 2020. Citado na página 33.

BONACICH, P. Power and centrality: A family of measures. **American journal of sociology**, University of Chicago Press, v. 92, n. 5, p. 1170–1182, 1987. Citado na página 46.

BRASIL. **Lei nº 5172/1966**. 1966. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l5172.htm>. Acesso em: 25/01/2021. Citado nas páginas 35, 38 e 63.

____. **Constituição da República Federativa do Brasil**. 1988. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm>. Acesso em: 25/01/2021. Citado nas páginas 35 e 57.

____. **Lei nº 8846/1994**. 1994. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/L8846.htm>. Acesso em: 25/01/2021. Citado na página 29.

____. **Lei nº 10406/2002**. 2002. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/2002/l10406compilada.htm>. Acesso em: 25/01/2021. Citado na página 64.

____. **Emenda Constitucional nº42/2003**. 2003. Disponível em: <http://www.planalto.gov.br/ccivil_03/constituicao/Emendas/Emc/emc42.htm>. Acesso em: 25/01/2021. Citado na página 35.

____. **Lei nº 123/2006**. 2006. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp123.htm>. Acesso em: 25/01/2021. Citado na página 65.

____. **Resolução CONCLA nº 01/2006**. 2006. Disponível em: <https://concla.ibge.gov.br/images/concla/documentacao/ResConcla1-2006.pdf>. Acesso em: 25/01/2021. Citado na página 67.

____. **Resolução CONCLA nº 02/2006**. 2006. Disponível em: <https://concla.ibge.gov.br/images/concla/documentacao/ResConcla2-2006.pdf>. Acesso em: 25/01/2021. Citado na página 67.

____. **Lei nº 128/2008**. 2008. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp128.htm>. Acesso em: 25/01/2021. Citado na página 65.

____. **Lei nº 12527/2011**. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>. Acesso em: 25/01/2021. Citado na página 57.

____. **Lei nº 13979/2020**. 2020. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l13979.htm>. Acesso em: 25/01/2021. Citado na página 31.

____. **Lei nº 14019/2020**. 2020. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l13979.htm>. Acesso em: 25/01/2021. Citado na página 31.

____. **Lei nº 14035/2020**. 2020. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l13979.htm>. Acesso em: 25/01/2021. Citado na página 31.

____. **Portal da Nota Fiscal Eletrônica**. 2021. Disponível em: <https://www.nfe.fazenda.gov.br/portal/principal.aspx>. Acesso em: 25/01/2021. Citado na página 38.

____. **Portal da Nota Fiscal Eletrônica - Relação de Serviços Web**. 2021. Disponível em: <https://www.nfe.fazenda.gov.br/portal/webServices.aspx?tipoConteudo=Wak0FwB7dKs=>. Acesso em: 25/01/2021. Citado na página 38.

CRAMER, J. S. The origins of logistic regression. Tinbergen Institute Working Paper, 2002. Citado na página 52.

ENAT. **Protocolo 01/2004 - Cadastro Sincronizado**. 2004. Disponível em: <http://www.enat.receita.fazenda.gov.br/pt-br/area_nacional/areas_interesse/foruns/historico/i-enat/i-enat-2004-protocolos/protocolo-01-2004-cooperacao-praticas-i-enat/at_download/file>. Citado na página 36.

____. **Protocolo 02/2005 - SPED**. 2005. Disponível em: <http://www.enat.receita.fazenda.gov.br/pt-br/area_nacional/areas_interesse/foruns/historico/ii-enat/ii-enat-2005-protocolos/protocolo-02-2005-sped-ii-enat>. Citado na página 37.

____. **Protocolo 03/2005 - NF-e**. 2005. Disponível em: <http://www.enat.receita.fazenda.gov.br/pt-br/area_nacional/areas_interesse/foruns/historico/ii-enat/ii-enat-2005-protocolos/protocolo-03-2005-nf-e-ii-enat>. Citado na página 37.

____. **Protocolo 04/2005 - CNAE Fiscal**. 2005. Disponível em: <http://www.enat.receita.fazenda.gov.br/pt-br/area_nacional/areas_interesse/foruns/historico/ii-enat/ii-enat-2005-protocolos/protocolo-04-2005-cnae-fiscal-ii-enat>. Citado na página 37.

FACURE, M. **Regressão Logística**. 2017. Disponível em: <https://matheusfacure.github.io/2017/02/25/regr-log/>. Acesso em: 25/01/2021. Citado na página 52.

FERNANDES, N. Economic effects of coronavirus outbreak (covid-19) on the world economy. **Available at SSRN 3557504**, 2020. Citado na página 33.

FERNANDEZ, M. L. A. **Avaliação da utilização de documentos fiscais eletrônicos na rastreabilidade de cargas.** Tese (Doutorado) — Universidade de São Paulo, 2012. Citado na página 30.

FIX, E. **Discriminatory analysis: nonparametric discrimination, consistency properties**. [S.l.]: USAF school of Aviation Medicine, 1951. Citado na página 51.

FREEMAN, L. Centrality in networks: I. conceptual clarifications. social networks. **Social Network**, 1979. Citado na página 45.

FREEMAN, L. C. A set of measures of centrality based on betweenness. **Sociometry**, JSTOR, p. 35–41, 1977. Citado na página 46.

GOOGLE CLOUD PLATFORM. **Google AI Platform Notebooks**. 2021. Disponível em: <https://cloud.google.com/ai-platform-notebooks>. Acesso em: 25/01/2021. Citado na página 131.

____. **Google BigQuery**. 2021. Disponível em: <https://cloud.google.com/bigquery>. Acesso em: 25/01/2021. Citado na página 131.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. [S.l.], 1995. v. 1, p. 278–282. Citado na página 51.

IN LOCO. **Mapa brasileiro da COVID-19**. 2021. Disponível em: <https://extranet.who.int/pqweb/sites/default/files/documents/Status_COVID_VAX_Dec2020.pdf>. Acesso em: 25/01/2021. Citado na página 32.

INSTITUTO BRASIL DE GEOGRAFIA E ESTATíSTICA. **Produto Interno Bruto - PIB**. 2021. Disponível em: <https://www.ibge.gov.br/explica/pib.php>. Acesso em: 25/01/2021. Citado na página 33.

JORNAL FOLHA DE SÃO PAULO. **Brasil confirma primeiro caso do novo coronavírus**. 2020. Disponível em: <https://www1.folha.uol.com.br/equilibrioesaude/2020/02/brasil-confirma-primeiro-caso-do-novo-coronavirus.shtml>. Acesso em: 25/01/2021. Citado na página 31.

____. **Brasil tem primeira morte pelo novo coronavírus; homem tinha diabetes e hipertensão**. 2020. Disponível em: <https://www1.folha.uol.com.br/equilibrioesaude/2020/03/brasil-tem-primeira-morte-pelo-novo-coronavirus.shtml>. Acesso em: 25/01/2021. Citado na página 31.

____. **Câmara aprova antecipação de dois feriados na capital paulista**. 2020. Disponível em: <https://www1.folha.uol.com.br/equilibrioesaude/2020/05/sp-deve-ter-feriadao-de-6-dias-a-partir-de-quarta-para-aumentar-isolamento.shtml>. Acesso em: 25/01/2021. Citado na página 32.

____. **Governo de SP decreta fechamento total de serviços não essenciais por 15 dias**. 2020. Disponível em: <https://www1.folha.uol.com.br/cotidiano/2020/03/governo-de-sao-paulo-decreta-fechamento-total-de-servicos-nao-essenciais-por-15-dias.shtml>. Acesso em: 25/01/2021. Citado na página 32.

____. **OMS declara pandemia do novo coronavírus Sars-Cov-2**. 2020. Disponível em: <https://www1.folha.uol.com.br/equilibrioesaude/2020/03/oms-declara-pandemia-do-novo-coronavirus.shtml>. Acesso em: 25/01/2021. Citado na página 31.

____. **Após aval da Anvisa, enfermeira do Emilio Ribas é a primeira a ser vacinada contra a Covid-19 no Brasil**. 2021. Disponível em: <https://www1.folha.uol.com.br/equilibrioesaude/2021/01/apos-aval-da-anvisa-enfermeira-do-emilio-ribas-e-a-primeira-a-ser-vacinada-contra-a-covid-19-no-brasil.shtml>. Acesso em: 25/01/2021. Citado na página 32.

JUNIOR, N. F. R. C. Geração da matriz origem-destino para o transporte rodoviário de carga usando o manifesto eletrônico de documentos fiscais. 2019. Citado na página 30.

KLEINBERG, E. Stochastic discrimination. **Annals of Mathematics and Artificial intelligence**, Springer, v. 1, n. 1-4, p. 207–239, 1990. Citado na página 51.

LUCE, R. D.; PERRY, A. D. A method of matrix analysis of group structure. **Psychometrika**, Springer, v. 14, n. 2, p. 95–116, 1949. Citado na página 44.

MADEIRA, R. d. O. C. **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais**. Tese (Doutorado) — Fundação Getúlio Vargas, 2015. Citado na página 30.

MATPLOTLIB: Visualization with Python. 2021. Disponível em: <https://matplotlib.org/>. Acesso em: 25/01/2021. Citado na página 131.

MCKIBBIN, W.; FERNANDO, R. The economic impact of covid-19. **Economics in the Time of COVID-19**, v. 45, 2020. Citado na página 33.

MELLO, N. O. d. **Os documentos fiscais eletrônicos no Brasil e na América Latina e seus impactos na automação de processos e no intercâmbio eletrônico de informações na cadeia logística.** Tese (Doutorado) — Universidade de São Paulo, 2014. Citado na página 29.

MISES, R.; POLLACZEK-GEIRINGER, H. Praktische verfahren der gleichungsauflösung. **ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik**, Wiley Online Library, v. 9, n. 1, p. 58–77, 1929. Citado na página 46.

NETWORKX: Network Analysis in Python. 2021. Disponível em: <https://networkx.org/>. Acesso em: 25/01/2021. Citado na página 132.

NEWMAN, M. **Networks**. [S.l.]: Oxford University Press, 2018. Citado na página 41.

NEWMAN, M. E. Mixing patterns in networks. **Physical review E**, APS, v. 67, n. 2, p. 026126, 2003. Citado na página 44.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. **The PageRank citation ranking: Bringing order to the web.** [S.l.], 1999. Citado na página 46.

PANDARAL·LEL. 2021. Disponível em: <https://github.com/nalepae/pandarallel>. Acesso em: 25/01/2021. Citado na página 132.

PASA, E. C. O uso de documentos eletrônicos na contabilidade. **Revista Contabilidade & Finanças**, SciELO Brasil, v. 12, n. 25, p. 72–83, 2001. Citado na página 30.

PEARSON, K. Liii. on lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901. Citado na página 47.

PORTABLE SPECTRAL SERVICES. **Precision and Accuracy**. 2019. Disponível em: <https://www.portaspecs.com/precision-and-accuracy/>. Acesso em: 25/01/2021. Citado na página 50.

SANTOS, E. M. d. Uso de dados de documentos fiscais eletrônicos para o planejamento do transporte urbano de cargas. 2015. Citado na página 30.

SCIKIT-LEARN: Machine Learning in Python. 2021. Disponível em: <https://scikit-learn.org/stable/>. Acesso em: 25/01/2021. Citado na página 132.

SCIPY: a Python-based ecosystem of open-source software for mathematics, science, and engineering. 2021. Disponível em: <https://www.scipy.org/>. Acesso em: 25/01/2021. Citado na página 132.

SECRETARIA ESPECIAL DA RECEITA FEDERAL DO BRASIL. **Dados públicos CNPJ - Receita Federal**. 2020. Disponível em: <https://receita.economia.gov.br/orientacao/tributaria/cadastros/cadastro-nacional-de-pessoas-juridicas-cnpj/dados-publicos-cnpj>. Acesso em: 30/09/2020. Citado nas páginas 29 e 57.

SOUZA, P. R. S. de. Uma análise em possíveis casos de patologias médicas utilizando a curva roc em lógica paraconsistente anotada para apoio a decisão médica em busca de melhor precisão de resposta na web. Congresso Internacional ABED de Educação a Distância, 2008. Citado na página 49.

STEPHENSON, K.; ZELEN, M. Rethinking centrality: Methods and examples. **Social networks**, Elsevier, v. 11, n. 1, p. 1–37, 1989. Citado na página 46.

SUN, X.; WANDELT, S.; ZHANG, A. How did covid-19 impact air transportation? a first peek through the lens of complex networks. **Journal of Air Transport Management**, Elsevier, v. 89, p. 101928, 2020. Citado nas páginas 33 e 34.

UOL. **Crivella determina fechamento de comércio no Rio a partir de terça-feira**. 2020. Disponível em: <https://noticias.uol.com.br/cotidiano/ultimas-noticias/2020/03/22/crivella-determina-fechamento-de-comercio-no-rio-a-partie-de-terca-feira.htm>. Acesso em: 25/01/2021. Citado na página 32.

VAPNIK, V. Pattern recognition using generalized portrait method. **Automation and remote control**, v. 24, p. 774–780, 1963. Citado na página 52.

VAPNIK, V. N. An overview of statistical learning theory. **IEEE transactions on neural networks**, IEEE, v. 10, n. 5, p. 988–999, 1999. Citado na página 47.

W3C. **World Wide Web Consortium**. 2021. Disponível em: <https://www.w3schools.com/xml/xml_whatis.asp>. Acesso em: 25/01/2021. Citado na página 39.

WASSERMAN, S.; FAUST, K. *et al.* Social network analysis: Methods and applications. Cambridge university press, 1994. Citado na página 44.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **nature**, Nature Publishing Group, v. 393, n. 6684, p. 440–442, 1998. Citado na página 44.

WORLD HEALTH ORGANIZATION. **Coronavirus disease 2019 Q&As**. 2021. Disponível em: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19>. Acesso em: 25/01/2021. Citado na página 31.

____. **Coronavirus disease (COVID-19) advice for the public**. 2021. Disponível em: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public>. Acesso em: 25/01/2021. Citado na página 31.

____. **Status of COVID-19 Vaccines within WHO EUL/PQ evaluation process**. 2021. Disponível em: <https://extranet.who.int/pqweb/sites/default/files/documents/Status_COVID_VAX_Dec2020.pdf>. Acesso em: 25/01/2021. Citado na página 32.

____. **WHO Coronavirus Disease (COVID-19) Dashboard**. 2021. Disponível em: <https://covid19.who.int/>. Acesso em: 25/01/2021. Citado na página 32.

ZHOU, P.; YANG, X.-L.; WANG, X.-G.; HU, B.; ZHANG, L.; ZHANG, W.; SI, H.-R.; ZHU, Y.; LI, B.; HUANG, C.-L. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. **Nature**, Nature Publishing Group, v. 579, n. 7798, p. 270–273, 2020. Citado na página 30.

GLOSSÁRIO

***Application Programming Interface*:** ou *API*, em tradução livre significa Interface Programática de Aplicação, é um conjunto de ferramentas que permite a integração entre dois sistemas através de um protocolo de comunicação. Nesta integração são usadas definições comuns a ambos sistemas de modo a estabelecer um padrão de comunicação. APIs geralmente são construídas usando um protocolo de transferência de dados, como HTTP, REST, ou SOAP, e encapsulam o acesso a um banco de dados interno.

***Enterprise Resource Planning*:** ou *ERP*, em tradução livre significa Planejamento de Recursos Empresariais, é um tipo de sistema para o gerenciamento de processos de negócio que podem atender diversas áreas, como financeiro, recursos humanos, compras e vendas, etc. No contexto de documentos fiscais, muitos desses sistemas possuem emissores capazes de gerenciar algumas das obrigações contábeis de forma automatizada.

Armazém de Dados: ou *Data Warehouse*, é um sistema de armazenamento estruturado de informações usado para a consolidação de informações e conceitos, geralmente envolvendo uma grande quantidade de dados. Esses dados são comumente armazenados em bancos de dados colunares pois envolvem o cálculo de agregações, e usam técnicas de Processamento Analítico em Tempo Real, também chamado Online Analytical Processing (OLAP), para a coleta e tratamento de tais dados. Os dados são então usados em processos corporativos para a tomada de decisão, desenvolvimento e aperfeiçoamento de novos produtos, e até mesmo reaproveitados em aplicações.

Fonte Única de Verdade: é uma prática adotada na construção de sistemas onde todos os dados conceitualmente relacionados a uma entidade importante para uma empresa ou organização são centralizados em uma única base de dados que passa a se tornar um banco de referência a outros sistemas, aumentando a consistência dos dados e diminuindo a carga cognitiva necessária para times indiretamente ligados a essa entidade.

APÊNDICE

A

ARMAZÉM DE DADOS

Neste apêndice, será descrito parte da arquitetura de ingestão de documentos fiscais utilizado pela empresa parceira.

A.1 Arquitetura de Ingestão

A Arquivei, empresa parceira deste trabalho, possui em sua plataforma diversos fluxos de ingestão de documentos fiscais. O usuário pode fazer o upload de documentos por diversos meios, sendo eles:

- Upload via plataforma *web*
- Upload de arquivos zip via plataforma *web*
- Envio por email
- Através do Sincroniza Notas, um produto instalável que faz integrações com ERPs e outros sistemas contábeis
- Através de consulta automatizada utilizando certificado digital A1
- Através do Modulo de Segurança da Arquivei que utiliza uma integração *web* para fazer a consulta via certificado digital A3

O fluxo de consulta automatizada é especialmente importante porque é por este fluxo que são recebidos os documentos vindos da Secretaria da Fazenda. O sistema de consulta lida de forma eficiente com os *web services* de forma a garantir que o cliente tenha acesso a todas as informações disponíveis, da forma mais consistente possível. Estes documentos possuem um grau de confiabilidade maior em relação aos demais, uma vez que os documentos recebidos

através dos clientes podem ter inconsistências ou incompletudes por problemas de comunicação e integração. Essas inconsistências podem apresentar ruídos em análises de dados, uma vez que nem todas as informações disponíveis no documento fiscal são checadas pela Secretaria da Fazenda e as alterações dos mesmos ocorrem através de documentos específicos, chamadas Cartas de Correção, que são de difícil processamento.

Todos esses fluxos passam por um fluxo centralizado de tratamento de dados, que garante certa consistência e integridade. Os documentos coletados são então enviados a um sistema chamado Plataforma de Documentos Fiscais, descrito a seguir.

A.2 Plataforma de DFe

A Plataforma de Documentos Fiscais é o sistema responsável por armazenar os documentos de forma segura, consistente, e disponível. Esta plataforma usa o conceito de Fonte Única de Verdade para construir um conjunto de sistemas responsável por tratar os dados de documentos fiscais de forma centralizada e consistente.

A plataforma disponibiliza uma série de APIs para ingestão síncrona, ingestão assíncrona, e disponibilização dos dados. Os clientes da plataforma são times internos da Arquivei, responsáveis por construir diferentes produtos, que se usam das funcionalidades da mesma para agilizar e melhorar suas entregas.

A.3 Armazém de Dados

Uma das interfaces da Plataforma de DFe é um protocolo de replicação de dados chamado Change Data Capture (CDC), em tradução livre Captura de Modificações de Dados, que é um protocolo capaz de transmitir mudanças feitas em um banco de dados para sistemas externos de forma a possibilitar a geração de visualizações.

Visualizações são bancos de dados de propósito específico criados a partir de uma Fonte Única de Verdade, visando atender demandas específicas. Uma empresa pode escolher ter visualizações analíticas, visualizações de índice-invertido, ou visualizações transacionais, por exemplo. Cada visualização irá disponibilizar os dados através de um novo protocolo ou de uma nova modelagem, conforme as demandas da empresa ou organização.

Uma visualização utilizada são armazéns de dados. O armazém de dados da Arquivei é construído a partir do CDC visando disponibilizar os dados da base de dados para análise. Através do armazém de dados de documentos fiscais foi possível obter os dados para este trabalho. O armazém também armazena dados coletados de base de dados externas que tenham alguma interação com os dados transacionais ou de documentos fiscais.

APÊNDICE

B

TECNOLOGIAS UTILIZADAS

Serão descritas aqui algumas das tecnologias utilizadas na execução deste trabalho.

B.1 Google BigQuery

O *Google BigQuery* é um produto disponível na plataforma de nuvem *Google Cloud Platform* (GOOGLE CLOUD PLATFORM, 2021b), sendo uma tecnologia de armazém de dados gerenciada e altamente escalável. A interface de interação com os dados é a linguagem SQL. Se trata de um banco colunar capaz de lidar com dados da ordem de *petabytes* de dados, e conta com ferramentas de segurança e criptografia nativas.

O *Google BigQuery* é a tecnologia usada no armazenamento de dados do armazém de dados da empresa parceira.

B.2 Google AI Platform Notebooks

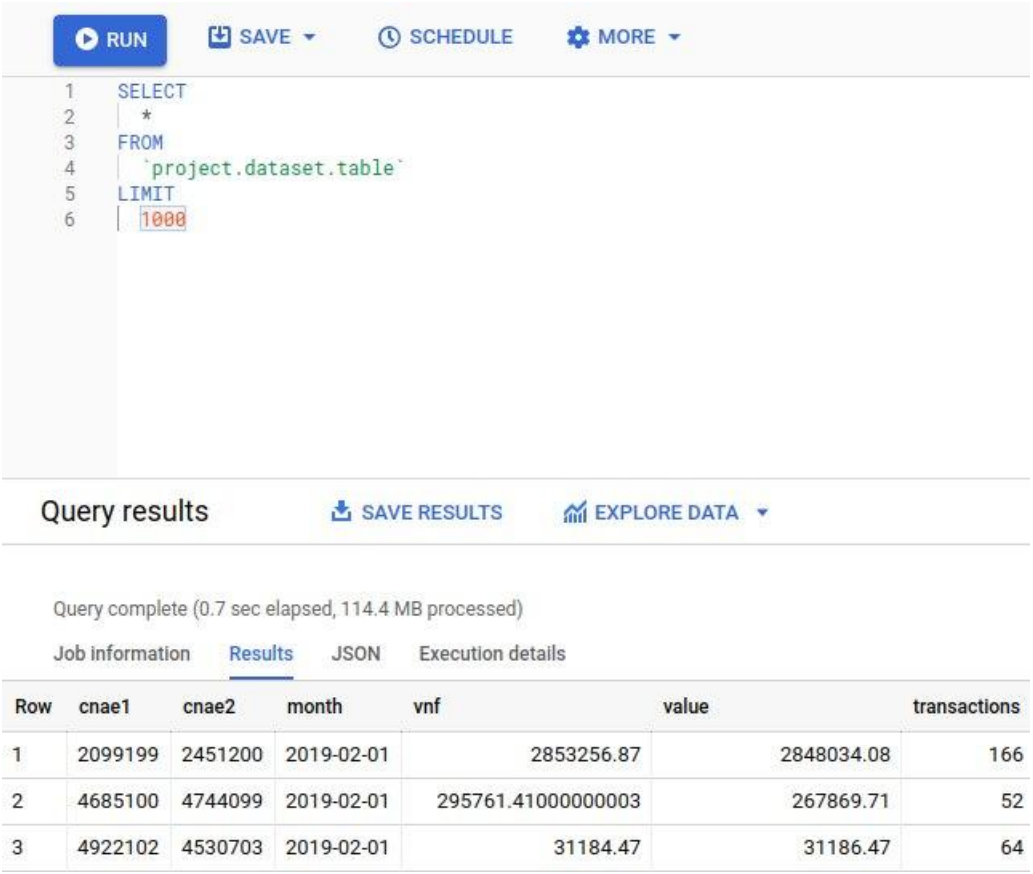
Outro produto disponível na *Google Cloud Platform*, o *Google AI Platform Notebooks* (GOOGLE CLOUD PLATFORM, 2021a), fornece um serviço gerenciado de integração segura com o ambiente JupyterLab, uma plataforma *web* para aplicações usando Jupyter Notebook, uma aplicação *web* para a criação e compartilhamento de códigos Python.

B.3 Bibliotecas

Neste trabalho foi utilizada a linguagem Python, com o auxílio das bibliotecas:

- *matplotlib* (MATPLOTLIB..., 2021): para a visualização e criação de imagens.

Figura 40 – Interface do Google BigQuery para seleção de dados



Fonte: Elaborada pelo autor.

- *sklearn* (SCIKIT-LEARN..., 2021): para pré-processamento e execução de algoritmos de análise de dados e aprendizado de máquina.
- *scipy* (SCIPY..., 2021): para algumas análises estatísticas e pré-processamentos.
- *networkx* (NETWORKX..., 2021): para execução de algoritmos de processamento de redes.
- *pandarallel* (PANDARAL·LEL, 2021): para paralelização usando processamento *multithread* para melhor escalabilidade.

APÊNDICE

C

EXEMPLO DE NOTA FISCAL ELETRÔNICA

Neste apêndice, apresentamos um fragmento de uma Nota Fiscal Eletrônica para referência com dados sensíveis removidos no Código-fonte 1.

Código-fonte 1 – Exemplo de fragmento de NFe com dados sensíveis removidos

```
<emit>
  <CNPJ>***</CNPJ>
  <xNome>LOURENCO S.CARLOS COM.LUBR.EM GERAL LTDA</xNome>
  <enderEmit>
    <xLgr>R EPISCOPAL</xLgr>
    <nro>3055</nro>
    <xBairro>CIDADE JARDIM</xBairro>
    <cMun>3548906</cMun>
    <xMun>SAO CARLOS</xMun>
    <UF>SP</UF>
    <CEP>13560580</CEP>
    <cPais>1058</cPais>
    <xPais>Brasil</xPais>
    <fone>1633613168</fone>
  </enderEmit>
  <IE>637124250112</IE>
  <CRT>3</CRT>
</emit>
<dest>
  <CNPJ>***</CNPJ>
  <xNome>C.JR.CONSTRUTORA LTDA</xNome>
  <enderDest>
    <xLgr>ROD W. LUIZ KM 240</xLgr>
```

```
24:     <nro>S/N</nro>
25:     <xBairro>RURAL</xBairro>
26:     <cMun>3548906</cMun>
27:     <xMun>SAO CARLOS</xMun>
28:     <UF>SP</UF>
29:     <CEP>13560970</CEP>
30:     <cPais>1058</cPais>
31:     <xPais>BRASIL</xPais>
32:     <fone>1633612039</fone>
33:   </enderDest>
34:   <indIEDest>9</indIEDest>
35:   <IE>637143185117</IE>
36:   <email>***</email>
37: </dest>
38: <det nItem="1">
39:   <prod>
40:     <cProd>005449</cProd>
41:     <cEAN>SEM GTIN</cEAN>
42:     <xProd>TOTAL RUBIA TIR 7900 FE SM 10W30 CJ4 L</xProd>
43:     <NCM>27101932</NCM>
44:     <CEST>0103700</CEST>
45:     <CFOP>5656</CFOP>
46:     <uCom>L</uCom>
47:     <qCom>8.0000</qCom>
48:     <vUnCom>43.000000</vUnCom>
49:     <vProd>344.00</vProd>
50:     <cEANTrib>SEM GTIN</cEANTrib>
51:     <uTrib>L</uTrib>
52:     <qTrib>8.0000</qTrib>
53:     <vUnTrib>43.000000</vUnTrib>
54:     <indTot>1</indTot>
55:     <comb>
56:       <cProdANP>620501002</cProdANP>
57:       <descANP>CICLO DIESEL</descANP>
58:       <UFCons>SP</UFCons>
59:     </comb>
60:   </prod>
61: <det>
```

ICMC USP
SÃO CARLOS