UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil

Dissertação de Mestrado

Kleber Henrique de Jesus Prado

São Cristóvão – Sergipe

2020

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Kleber Henrique de Jesus Prado

Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Prof. Dr. Methanias Colaço Rodrigues Júnior

São Cristóvão – Sergipe

2020

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL
UNIVERSIDADE FEDERAL DE SERGIPE

Prado, Kleber Henrique de Jesus
P896d Data science aplicada à análise criminal baseada nos dados abertos governamentais do Brasil / Kleber Henrique de Jesus Prado ; orientador Methanias Colaço Rodrigues Júnior. - São Cristóvão, 2020.
144 f. : il.

Dissertação (mestrado em Ciência da Computação) – Universidade Federal de Sergipe, 2020.

1. Computação. 2. Crime. 3. Banco de dados. 4. Administração pública. I. Rodrigues Junior, Methanias Colaçõ orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
COORDENAÇÃO DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Ata da Sessão Solene de Defesa da Dissertação do Curso de Mestrado em Ciência da Computação-UFS.
Candidato: KLEBER HENRIQUE DE JESUS PRADO

Em 25 dias do mês de novembro do ano de dois mil e vinte, com início às 16h00min, realizou-se na Sala virtual https://meet.google.com/eej-drny-joa. A Sessão Pública de Defesa de Dissertação de Mestrado do candidato **KLEBER HENRIQUE DE JESUS PRADO**, que desenvolveu o trabalho intitulado: **"Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil"**, sob a orientação do Prof. Dr. **Methanias Colaço Rodrigues Júnior**. A Sessão foi presidida pelo Prof. Dr. **Methanias Colaço Rodrigues Júnior** (PROCC/UFS), que após a apresentação da dissertação passou a palavra aos outros membros da Banca Examinadora, Prof. Dr. **Mário André de Freitas Farias (**IFS**),** a Profª. Drª. **Leila Maciel de Almeida e Silva**(Dcomp/UFS) e, em seguida, ao Prof. Dr. **Edward David Moreno Ordonez** (PROCC/UFS). Após as discussões, a Banca Examinadora reuniu-se e considerou o mestrando aprovado. Atendidas as exigências da Instrução Normativa 01/2017/PROCC, do Regimento Interno do PROCC (Resolução 67/2014/CONEPE), Resolução nº 25/2014/CONEPE e da Portaria nº 413 de 27 de maio de 2020 (Banca por videoconferência) que regulamentam a Apresentação e Defesa de Dissertação, e nada mais havendo a tratar, a Banca Examinadora elaborou esta Ata que será assinada pelos seus membros e pelo mestrando.

Cidade Universitária "*Prof. José Aloísio de Campos",* 25 de novembro de 2020.

Prof. Dr. Methanias Colaço Rodrigues Júnior
(PROCC/UFS)
Presidente

Profª. Drª. Leila Maciel de Almeida e Silva
(DComp/UFS)
Examinadora Externa ao Programa

Prof. Dr. Edward David Moreno Ordonez
(PROCC/UFS)
Examinador Interno

Prof. Dr. Mário André Freitas Farias
(IFS)
Examinador Externo à Instituição

Kleber Henrique de Jesus Prado
Discente

À toda minha família, especialmente aos meus queridos e amados pais, irmãos, esposa, filho e enteados.

Agradecimentos

Primeiramente à Deus, dando-me sabedoria, força e serenidade para enfrentar e vencer os obstáculos que cruzam minha caminhada.

Aos meus amados pais, por todo apoio e por estarem sempre ao meu lado.

Aos meus queridos irmãos Kátia e Edu, pelo incentivo diário.

À minha amada esposa Clécia pela compreensão, apoio e paciência, durante todos esses anos.

Ao meu amado filho Lucas, que me fortalece, a cada dia, com os sorrisos e abraços.

Aos meus queridos enteados João e Laura, pela cumplicidade e pela harmoniosa convivência.

Ao meu orientador, Prof. Dr. Methanias Colaço Rodrigues Júnior, pela paciência, pelo apoio, pelo incentivo, pela sua disponibilidade total e, principalmente, pelos valiosos conselhos e orientações.

Aos amigos e companheiros de mestrado que contribuíram diretamente nesta minha formação, em especial à Anderson Barroso, Layse Souza e Thiago Oliveira.

E por fim, obrigado a todos que, de alguma maneira, contribuíram para o meu crescimento pessoal e para a realização deste trabalho.

"Deus nos concede a cada dia, uma página de vida nova no livro do tempo.
Aquilo que colocarmos nela, corre por nossa conta."
(Chico Xavier)

Resumo

Contexto: Crime é um problema social comum e complexo, que afeta a qualidade de vida, o crescimento econômico e a reputação de uma nação. Governantes e a sociedade em geral têm tido enormes problemas causados por esse fenômeno, gastando, a cada ano, milhões de dólares combatendo a violência e, consequentemente, causando grande preocupação com o seu controle para as agências de segurança pública. Portanto, novas abordagens e sistemas avançados são necessários para melhorar a análise de crimes e para proteger a sociedade. Neste contexto, a *Data Science* vem desempenhando um papel fundamental na melhoria dos resultados das investigações e detecções criminais, facilitando o registro, a análise de recuperação e o compartilhamento das informações. **Objetivo:** Aplicar fundamentos de *Data Science* e fornecer um modelo automatizado, constantemente atualizado, para analisar dados abertos governamentais relacionados aos crimes ocorridos nas Unidades Federativas (UFs) brasileiras e nos municípios de Minas Gerais. **Método:** Inicialmente, foi executada uma Revisão Sistemática (RS) quantitativa (com metanálise), como forma de identificar e sistematizar as principais abordagens, técnicas e algoritmos utilizados na análise inteligente de dados governamentais abertos relacionados a incidentes criminais. Em seguida, dois experimentos controlados foram executados para descoberta de regras de associação entre estados, municípios, crimes, Regiões Integradas de Segurança Pública (RISPs), alvos de roubo e alvos de furto. Adicionalmente, foi realizada a detecção de *outliers* em relação às taxas de criminalidade e foram desenvolvidos *rankings* que demostram os locais (estados, municípios ou RISPs) mais perigosos. **Resultados:** No contexto dos estados brasileiros, do ponto de vista geral, com ponderações para os crimes, o Paraná foi o local mais perigoso, em todos os anos avaliados. Destaque também para o Rio de Janeiro, ocupando sempre a segunda posição. Além disso, os estados de Goiás, Pernambuco e Rondônia foram classificados entre os cinco mais perigosos, em três dos cinco anos analisados. Sob a perspectiva única dos assassinatos, em 2019, os estados de Roraima, Rio Grande do Norte, Sergipe, Acre e Pernambuco foram classificados entre os dez mais violentos, sendo Pernambuco e Acre os estados mais perigosos nas duas perspectivas (média ponderada e homicídios). Em relação às regras de associação, ficou evidenciado que existem dependências entre crimes e estados. No âmbito do estado de Minas Gerais, os municípios de Belo Horizonte, Confins e Contagem estiveram, constantemente, entre os cinco mais perigosos. Além disso, ficou evidenciado que há dependências entre: crimes e municípios, crimes e RISPs, alvos de roubo e municípios, e alvos de roubo e RISPs. Por outro lado, não foram detectadas associações entre alvos de furto e municípios, e alvos de furto e RISPs. **Conclusão:** A *Data Science* possibilita a execução de diagnósticos mais precisos e mais céleres, auxiliando o planejamento estratégico e a tomada de decisão em Segurança Pública.

Palavras-chave: Regras de Associação, Análise Criminal, Ciência de Dados, Dados Abertos Governamentais e Metanálise.

Abstract

Context: Crime is a common and complex social problem that affects a nation's quality of life, economic growth and reputation. Governments and society in general have had enormous problems caused by this phenomenon. Each year, governments spend millions of dollars fighting violence and, consequently, crime prevention and control are issues of great concern to public security agencies. Therefore, new approaches and advanced systems are needed to improve crime analysis and to protect their communities. In this context, Data Science has been playing a vital role in improving the results of criminal investigations and detections, facilitating registration, recovery analysis and information sharing. **Objective:** Apply fundamentals of Data Science and provide an automated model, constantly updated, to analyze open government data related to crimes occurred in the Federative Units (FUs) and in the municipalities of Minas Gerais. **Method:** Initially, we performed a quantitative Systematic Review (SR) (with meta-analysis), as a way to identify and systematize the main approaches, techniques and algorithms used in the intelligent analysis of open government data related to criminal initiates. Then, we performed two experiments to discover rules of association between states, municipalities, crimes, Integrated Public Security Regions (IPSRs), theft targets and theft targets. Additionally, we detect outliers in relation to crime rates and developed rankings that show the most dangerous locations (states, municipalities or IPSRs). **Results:** In the context of Brazilian states, from a general point of view, with weights for crimes, Paraná was the most dangerous place in all the years evaluated. Also noteworthy for Rio de Janeiro, always occupying the second position. Besides that, the states of Goiás, Pernambuco and Rondônia were classified among the five most dangerous, in three of the five years analyzed. From the single perspective of murders, in 2019, the states of Roraima, Rio Grande do Norte, Sergipe, Acre and Pernambuco were ranked as the ten most violent ones, being Pernambuco and Acre among the most dangerous states from the two perspectives (weighted average and murders). Regarding the association rules, it became evident that there are dependencies between crimes and states. In the context of the state of Minas Gerais, Belo Horizonte, Confins and Contagem were constantly among the five most dangerous municipalities. In addition, it became evident that there are dependencies between: crimes and municipalities, crimes and IPSRs, robbery targets and municipalities, and robbery targets and IPSRs. However, no associations were detected between theft targets and municipalities, and theft targets and IPSRs. **Conclusion:** Data Science enables the execution of more accurate and faster diagnostics, helping strategic planning and decision making in Public Security.

Keywords: Association Rules, Criminal Analysis, Data Science, Open Government Data and Meta-analysis.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

1.1 Contextualização

Com o aumento da urbanização diversas transformações sociais, econômicas e ambientais têm ocorrido em toda parte do mundo. Os desafios enfrentados pelos governantes estão cada vez maiores. Áreas como mobilidade urbana, saúde e segurança pública têm recebido uma atenção especial (CATLETT et al., 2018). Nos últimos anos, o aumento da violência tem sido objeto de estudo de muitos pesquisadores (DAMASCENO; TEIXEIRA; CAMPOS, 2012). O crime é um problema social predominante e sua prevenção é uma função extremamente importante. Governantes e a sociedade, em geral, têm tido enormes problemas causados por esse fenômeno.

A cada ano, os governos gastam milhões de dólares combatendo a violência, fornecendo equipamentos, treinamento e adquirindo ferramentas para auxiliar o trabalho policial. É responsabilidade das agências de aplicação da lei monitorar e reduzir a taxa de atividades criminosas que estão acontecendo continuamente nos dias de hoje (DAMASCENO; TEIXEIRA; CAMPOS, 2012)(MARZAN et al., 2017). Portanto, a prevenção e controle do crime são questões de grande preocupação para os governos e agências de segurança pública. Tais questões, se não forem bem controladas e gerenciadas, podem afetar drasticamente a economia de um país ao longo do tempo, uma vez que mais emigração ocorrerá naturalmente (TOPPIREDDY; SAINI; MAHAJAN, 2018).

Outro desafio enfrentado por essas organizações é lidar com um grande volume de informações referentes aos crimes e criminosos. De acordo com (CATLETT et al., 2018), uma quantidade significativa de dados com informações espaciais e temporais é obtida diariamente. Consequentemente, novas abordagens e sistemas avançados são necessários para melhorar a análise de crimes e para proteger suas comunidades, permitindo uma maior compreensão da dinâmica das atividades criminosas e fornecendo respostas como “onde”, “quando” e “por que” certos crimes são prováveis de acontecer (TOPPIREDDY; SAINI; MAHAJAN, 2018) (PHILLIPS; LEE, 2011). Neste contexto, a *Data Science*, aliada aos sistemas computacionais

inteligentes, vem desempenhando um papel vital na melhoria dos resultados das investigações e detecções criminais, facilitando o registro, a análise de recuperação e o compartilhamento das informações (GUPTA; CHANDRA; GUPTA, 2014).

Por outro lado, a transparência e o acesso às informações públicas vêm se tornando pilares essenciais para administração pública moderna. Com a evolução substancial da Tecnologia da Informação e Comunicação (TIC), novas maneiras de disponibilizar informações públicas à população estão sendo criadas. Novos sistemas foram desenvolvidos, novos serviços oferecidos e integrações sistêmicas aconteceram. Todas as revoluções tecnológicas, juntamente com a popularização da internet, geraram mudanças nos processos internos e nas relações do governo com o público externo. Tais mudanças podem ser chamadas de Governo Eletrônico, ou simplesmente *e-Gov*. Segundo Janssen, Charalabidis e Zuiderwijk (2012), com o *e-Gov*, os órgãos públicos estão entre os maiores criadores e coletores de dados em muitos domínios.

Diversas iniciativas internacionais foram criadas como intuito de promover a melhoria na transparência por meio da sua comunicação interna e externa. A *Open Government Partnership* (OGP) tem a finalidade de promover a transparência, o aumento da participação cívica e aproveitar novas tecnologias para tornar governos de diversos países mais abertos, eficazes e responsáveis (FREITAS; DACORSO, 2014). A ampliação da divulgação das ações governamentais aos cidadãos fortalece o regime democrático e melhora a gestão pública. Com a publicação de dados, é possível o controle social mais efetivo, pois conhecendo a situação das contas, a sociedade terá mais condições de cobrar, exigir e fiscalizar seus governantes, proporcionando-lhes ferramentas para compreender e utilizar a informação disponível e começar a cunhar um pensamento crítico sobre informações e serviços (HARRISON et al., 2012) (CAMPOS, 2018). Em contrapartida, com participação mais ativa da população, os administradores públicos estão tendo uma boa oportunidade de entender melhor as reais necessidades da população.

Um fenômeno que vem chamando bastante atenção e tem contribuído para o acesso a informações públicas é o *Open Government Data* (OGD). Este movimento, vem sendo adotado por diversos países e as principais razões para a implementação do OGD são o aumento da responsabilidade democrática, maior transparência, a prestação de serviços públicos autoconcedidos pelos cidadãos, o estímulo ao crescimento econômico e maior eficiência e eficácia por parte dos órgãos públicos, aumentando a credibilidade das suas informações (SharePSI, 2014). De acordo com Hardy e Maurushat (2017), os governos comprometidos com o movimento dados abertos estão postando milhares de conjuntos de dados em portais *online*. O objetivo de liberar esses dados não é meramente fornecer informações ao público, e sim impulsionar a inovação por meio da análise do "grande repositório de dados". Portanto, os governos ao "abrirem" seus *datasets*, permitem que diversas empresas, pesquisadores e o público em geral possam extrair novas informações (*insight*) desse "mar de dados" e contribuir com soluções inovadoras para políticas complexas.

No Brasil, o Governo Federal, na tentativa de tornar as contas públicas mais transparentes,

tem adotado alguns instrumentos de Políticas Públicas de Informação, mecanismos legais para auxiliar na busca por dados mais transparentes. O portal da Transparência [1] e de Dados Abertos [2] do Governo Federal são ferramentas que funcionam como um grande catálogo que facilita a busca e uso de todo e qualquer tipo de dado publicado pelos órgãos do governo. Informações sobre saúde suplementar, sistema de transporte, segurança pública, indicadores de educação, gastos governamentais, processo eleitoral, programas sociais e outros podem ser facilmente encontradas. Com isso, abre-se um canal direto entre o cidadão e governo, com o intuito de melhorar a utilização dos dados, proporcionando o fortalecimento do processo democrático e melhorando a qualidade de vida da população.

Na esfera criminal, o Governo Federal instituiu, pela Lei nº 12.681, de 4 de julho de 2012, o Sistema Nacional de Informações de Segurança Pública, Prisionais e de Rastreabilidade de Armas e Munições, de Material Genético, de Digitais e de Drogas (SINESP). Este sistema é um portal de informações integradas, possibilitando consultas operacionais, investigativas e estratégicas, sobre drogas, segurança pública, justiça, sistema prisional, entre outras, implementado em parceria com os entes federados (MJSP, 2020). Em Março de 2019, o Ministério da Justiça e Segurança Pública (MJSP) e pela Secretaria de Estado de Justiça lançou uma plataforma inédita de estatísticas oficiais de segurança pública com base nos boletins de ocorrência de todos os estados e do Distrito Federal do sistema SINESP, cujo o objetivo é consolidar, de forma inédita, uma fonte oficial, pública e gratuita de dados nacionais com informações sobre segurança pública de forma célere e transparente. A plataforma torna acessíveis informações estatísticas sobre crimes como: estupro, lesão corporal seguida de morte, homicídio doloso, latrocínio, tentativa de homicídio, roubo de veículo, furto de veículo, roubo de carga e roubo à instituição financeira.

No entanto, mesmo com a publicação de informações sobre estatísticas oficiais de segurança pública, realizadas pelos governos federal e estaduais brasileiros, nenhuma dessas iniciativas parecem promover uma transparência clara e consistente à população. São informações publicadas frequentemente, sem a certeza de que sejam utilizáveis e sem uma aplicação inteligente sobre os dados avaliados. Muitas vezes, sem métricas que possibilitem ao cidadão inferir conclusões acerca da publicação.

Em razão disso, a proposta deste trabalho é conduzir um processo experimental, seguindo as diretrizes de (WOHLIN et al., 2012), para realizar uma avaliação sobre dados abertos governamentais relacionados a incidentes criminais, disponibilizados pelo Ministério da Justiça e Segurança Pública (MJSP) e Secretaria de Estado de Justiça e Segurança Pública de Minas Gerais (Sejusp/MG). Os dados do MJSP possibilitarão uma análise no nível das Unidades Federativas (UFs), enquanto que os da Sejusp/MG proporcionarão uma avaliação sobre os municípios mineiros. A escolha pelos dados de Minas Gerais é resultado de uma busca minuciosa de *datasets* estaduais disponibilizados de forma *online* na *web*, a qual constatou que este estado,

[1] Portal da Transparência - <http://www.portaltransparencia.gov.br>
[2] Portal de Dados Abertos - <http://dados.gov.br>

até o momento da pesquisa, apresentou, em relação às outras UFs, uma melhor qualidade na disponibilização e composição dos seus *datasets*, pois são publicados em formatos abertos (arquivos .cvs), padronizados, contextualizados e que podem ser lidos facilmente por máquina. Adicionalmente, ficou evidenciado que o número de atributos contidos nos *datasets* criminais de Minas Gerais [3] era maior em relação aos demais estados, o que proporcionará uma análise inteligente mais robusta. Logo, o intuito desta pesquisa é detectar padrões e anomalias, avaliando a possibilidade de criação de aplicação que promova maior transparência, visando auxiliar o processo de apoio às decisões estratégicas e operacionais dos governantes e agentes da lei, no combate efetivo da criminalidade.

Após execução dos experimentos concluímos que, do ponto de vista geral, com ponderações para os crimes, o Paraná foi o estado mais perigoso, em todos os anos avaliados, seguido sempre pelo o Rio de Janeiro. Destaque, também, para os estados de Goiás, Pernambuco e Rondônia, que foram classificados entre os cinco mais perigosos, em três dos cinco anos analisados. Sob a perspectiva única dos assassinatos, em 2019, os estados de Roraima, Rio Grande do Norte, Sergipe, Acre e Pernambuco foram classificados entre os dez mais violentos, sendo Pernambuco e Acre os estados mais perigosos nas duas perspectivas (média ponderada e homicídios). Em relação às Regras de Associação (RAs), ficou evidenciado que existe dependência entre crimes e estados. Além disso, 7 valores discrepantes (*outliers*), relacionados às taxas de criminalidade, foram detectados nas regiões Norte (2) e Nordeste (5).

No contexto dos municípios de Minas Gerais, Belo Horizonte, Confins e Contagem estiveram, constantemente, entre os cinco mais perigosos. Com relação às regras de associação, ficou evidenciado que existem dependências entre: crimes e municípios, crimes e RISPs, alvos de roubo e municípios, e alvos de roubo e RISPs. Em contrapartida, associações entre alvos de furto e municípios, e alvos de furto e RISPs não foram detectadas.

1.2 Análise do Problema

À medida que a população cresce, novos tipos de crimes são cometidos, tornando as cidades cada vez mais vulneráveis e comprometendo diretamente a segurança da população. Portanto, medidas preventivas são de extrema importância para a prevenção do crime. Os governantes e as agências de aplicação da lei precisam ter uma abordagem ágil para enfrentar os crimes que estão em constante mudança (SINGH; JOSHI, 2018).

No âmbito do Brasil, o governo vem investindo em segurança pública, inclusive nos departamentos de inteligência da polícia, para lidar com o crescente número de incidentes criminais e a redução de recursos econômicos. Tais investimentos são uma tendência mundial devido ao impacto que este fenômeno acarreta na economia (FARIAS et al., 2018). Segundo

[3] Secretaria de Estado de Justiça e Segurança Pública de Minas Gerais (Sejusp/MG) - <http://www.seguranca.mg.gov.br>

Cerqueira et al. (2018), os gastos em segurança pública no Brasil em 2018 totalizaram R$ 91,2 bilhões, o que correspondeu a 1,34% do PIB. Em relação a 2017, houve aumento real de 3,9% nas despesas empenhadas, sendo que o crescimento ocorreu de forma diferenciada entre os entes federativos. Enquanto a União aumentou os seus gastos em 12,4%, os estados e municípios majoraram seus dispêndios em 2,3% e 8,7%, respectivamente. Em 2018, o custo para manter o aparato de segurança pública no país corresponderia a um gasto, por cada brasileiro, igual a R$ 409,66. Nesse mesmo ano, as despesas per capita realizadas com a segurança variaram nas Unidades Federativas (UF) entre R$ 228,60, no Piauí, a R$ 674,08, no Acre.

Por outro lado, abordagens inteligentes e robustas ajudariam as agências policiais a manter a segurança pública e a paz nas cidades, impedindo a proliferação da prática de crimes (SINGH; JOSHI, 2018), e proporcionaria, simultaneamente, uma redução orçamentária considerável destinada a área de segurança pública. Atualmente, existem numerosos *datasets* criminais nas organizações de segurança pública e órgãos policiais. Porém, a inspeção manual, a exploração e a análise desses dados, especialmente em países em desenvolvimento são inadequadas, podendo omitir relações complexas e cruciais. Além disso, o volume de dados que pode ser processado, dentro de um prazo razoável, é limitado e essa análise humana é mais propensa a erros (ISAFIADE; BAGULA, 2013). Previsões precisas de crimes em tempo real ajudam a reduzir a taxa de criminalidade, mas continuam sendo um problema desafiador para a comunidade científica, pois as ocorrências de crimes dependem de muitos fatores complexos (TOPPIREDDY; SAINI; MAHAJAN, 2018).

Para Singh e Joshi (2018), as técnicas de *Data Mining* provaram ser eficazes na análise do conjunto de dados e na coleta de informações úteis em muitos domínios. No campo criminal, a mineração de dados está recebendo maior atenção para descobrir padrões subjacentes nos dados sobre crimes. O desejo de agir rapidamente para suprimir as atividades criminosas e descobrir relações entre vários ativos de informação ainda persiste. As soluções, em tempo real, podem economizar recursos significativos e aumentar a capacidade da aplicação da lei (BUCZAK; GIFFORD, 2010). Técnicas de mineração de dados fornecem informações e padrões, não discerníveis para as agências policiais e analistas, sobre o status atual de segurança pública e ambiental de uma comunidade. Além disso, é possível prever possíveis ocorrências futuras com o objetivo de obter reconhecimento e gerenciamento da situação (ISAFIADE; BAGULA, 2013).

Diante do exposto, a presente proposta tem em seu escopo a aplicação de *Data Science*, com o intuito de encontrar aberrações (*outliers*) e associações, sobre o rol de informações criminais, disponíveis nos bancos de dados abertos governamentais. Então, as questões de pesquisa que guiaram este trabalho foram:

- **Experimento 1**
  - **Q1**: Existem discrepâncias entre as taxas de criminalidade (taxa por cem mil habitantes) das Regiões Brasileiras?

  - **Q2**: Quais os estados vêm se destacando como os mais perigosos?
  - **Q3**: Há associações entre tipos de crime e os estados?

- **Experimento 2**

  - **Q4**: Quais os municípios vêm se destacando como os mais perigosos?
  - **Q5**: Quais as Risps mais perigosas em 2019?
  - **Q6**: Há associações entre tipos de crime e municípios?
  - **Q7**: Há associações entre tipos de crime e Risps?
  - **Q8**: Há associações entre alvos de roubo e municípios?
  - **Q9**: Há associações entre alvos de roubo e Risps?
  - **Q10**: Há associações entre alvos de furto e municípios?
  - **Q11**: Há associações entre alvos de furto e Risps?

As questões **Q1**, **Q2** e **Q3**, elencadas para o **Experimento 1**, focam na análise dos dados criminais dos estados e das regiões brasileiras. Tais questões foram respondidas usando os dados das UFs, disponibilizados pelo MJSP/Brasil. As demais questões (**Q4**, **Q5**, **Q6**, **Q7**, **Q8**, **Q9**, **Q10** e **Q11**), descritas para o **Experimento 2**, contemplam uma análise sobre os dados criminais dos municípios de Minas Gerais, disponibilizados pela Sejusp/MG. A partir dessas questões gerais, algumas hipóteses foram testadas:

- **Experimento 1**

  - **Hipótese 1 (Q3)**
    - H_0: Os tipos de crime são independentes dos estados.
    - H_1: Os tipos de crime são dependentes dos estados.

- **Experimento 2**

  - **Hipótese 2 (Q6)**
    - H_0: Os tipos de crime são independentes dos municípios.
    - H_1: Os tipos de crime são dependentes dos municípios.
  - **Hipótese 3 (Q7)**
    - H_0: Os tipos de crime são independentes das Risps.
    - H_1: Os tipos de crime são dependentes das Risps.
  - **Hipótese 4 (Q8)**
    - H_0: Os alvos de roubo são independentes dos municípios.

    - H_1: Os alvos de roubo são dependentes dos municípios.

- **Hipótese 5 (Q9)**
    - H_0: Os alvos de roubo são independentes das Risps.
    - H_1: Os alvos de roubo são dependentes das Risps.
- **Hipótese 6 (Q10)**
    - H_0: Os alvos de furto são independentes dos municípios.
    - H_1: Os alvos de furto são dependentes dos municípios.
- **Hipótese 7 (Q11)**
    - H_0: Os alvos de furto são independentes das Risps.
    - H_1: Os alvos de furto são dependentes das Risps.

1.3 Justificativa

De acordo com o "Atlas da Violência 2019" , produzido pelo Instituto de Pesquisa Econômica Aplicada (IPEA) [4] e pelo Fórum Brasileiro de Segurança Pública (FBSP) [5], a violência constitui uma das maiores questões de políticas públicas no Brasil (CERQUEIRA et al., 2018). A cada ano, os governos gastam milhões de dólares combatendo a violência, fornecendo equipamentos, treinamento e adquirindo ferramentas para auxiliar o trabalho policial (DAMASCENO; TEIXEIRA; CAMPOS, 2012).

O FBSP publicou o 13º Anuário Brasileiro de Segurança Pública, que se baseia em informações fornecidas pelas secretarias de segurança pública estaduais, pelo Tesouro Nacional (TN), pelas polícias civis, militares e federal, entre outras fontes oficiais da segurança pública. Trata-se do mais amplo retrato da segurança pública brasileira, que compila e analisa dados de registros policiais sobre violência e criminalidade, violência contra mulheres, violência sexual, informações sobre o sistema prisional e gastos com segurança pública no ano de 2018. Segundo o anuário, o Brasil gastou aproximadamente R$ 91 bilhões como a segurança pública, um aumento de 3,9% em relação ao ano anterior (LIMA; BUENO, 2019).

Os dados oficiais do Sistema de Informações sobre Mortalidade, do Ministério da Saúde (SIM/MS) mostram que, em 2017, aconteceram 65.602 homicídios no Brasil, o que equivale a uma taxa de aproximadamente 31,6 mortes para cada cem mil habitantes. Trata-se do maior nível histórico de letalidade violenta intencional no país (CERQUEIRA et al., 2018). Com relação aos feminicídios, Lima e Bueno (2019) relataram que estes casos corresponderam a 29,6% dos homicídios dolosos de mulheres em 2018. Foram 1.151 casos, em 2017, e 1.206, em 2018, um crescimento de 4% nos números absolutos. Desde que a Lei do Feminicídio entrou em vigor, em 09 de março de 2015, os casos de feminicídio subiram 62,7%.

[4] Instituto de Pesquisa Econômica Aplicada (IPEA) - <http://www.ipea.gov.br>
[5] Fórum Brasileiro de Segurança Pública (FBSP) - <http://www.forumseguranca.org.br>

No âmbito de crimes patrimoniais, foram registrados 490.956 roubos e furtos de veículos no Brasil, em 2018, correspondendo a 33,26% de todos os crimes patrimoniais registrados no período. Isso equivale a mais de mil e trezentos roubos e furtos por dia, ou seja, a mais cinquenta e sete veículos roubados e furtados por hora (quase um por minuto). Portanto, são oscilações de números que se estabeleceram em patamares muito altos há algum tempo (LIMA; BUENO, 2019).

Consequentemente, diante do que foi contextualizado, reforça-se a necessidade de implementação de meios que ajudem na superação desse fenômeno, de modo a possibilitar a realização de diagnósticos mais precisos, por meio da detecção de padrões e anomalias (*outliers*) sobre dados de incidentes criminais, e que ajudem no planejamento estratégico governamental e o processo de tomada de decisão, buscando conter o orçamento empregado na área de segurança pública, e combater e mitigar a criminalidade.

1.4 Objetivos da Pesquisa

O objetivo desta pesquisa é aplicar *Data Science*, para analisar dados abertos governamentais relacionados a incidentes criminais ocorridos no Brasil, com intuito de encontrar aberrações (*outliers*) e correlações que auxiliem o controle social e o processo de tomada de decisão.

1.4.1 Objetivos Específicos

Os objetivos norteadores do presente trabalho são expostos a seguir:

- Revisão sistemática quantitativa da literatura, com a finalidade de identificar, caracterizar e metanalisar trabalhos científicos que fizeram uso de algoritmos, técnicas e abordagens inteligentes sobre dados criminais;
- Arquitetura centralizada, com objetivo de automatizar as etapas de *download*, ETL (Extração, Transformação e Carga) e mineração dos dados;
- Dois experimentos controlados, sobre os dados criminais ocorridos no Brasil, disponibilizados pelo MJSP e pela Sesjusp/MG, para investigar a existência de correlações e anomalias que auxiliem no combate ostensivo e preventivo da criminalidade;
- Aplicações públicas que tornem mais fácil a consulta e visualização dos resultados obtidos, proporcionando uma clara percepção sobre o status da criminalidade ao longo dos anos, bem como possibilitando que governantes, agentes da lei e cidadãos infiram conclusões acerca do problema.

1.5 Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, uma Revisão Sistemática (RS) quantitativa da literatura, publicada em (PRADO et al., 2020), baseada no protocolo proposto por Kitchenham (2004), tendo por finalidade encontrar o estado da arte das pesquisas sobre análise inteligente de dados relacionados a incidentes criminais. Para operacionalizar a revisão, acessamos a base Scopus por meio do portal de periódicos da CAPES disponível em (CAPES, 2019), o qual permite fazer *download* dos artigos sem restrições. A Scopus foi escolhida por incluir buscas em diferentes bancos de dados científicos (IEEE, ACM e outros).

Ato contínuo, para realização do objetivo principal desta pesquisa, foram utilizados os dados abertos criminais disponibilizados pelo MJSP/Brasil e pela Sejusp/MG. Todavia, algumas dessas informações são fornecidas em arquivos muito grandes, com dados dispersos, limitando o entendimento do cidadão com relação ao significado ou à importância dos dados abertos. Desta forma, este trabalho permitiu e permitirá fazer a transição dos dados brutos do governo para informações estruturadas, perfazendo o *download* do(s) arquivo(s), nos formatos CSV e XLSX, bem como a leitura, interpretação do conteúdo e armazenamento numa base de dados estruturada.

Com o objetivo de facilitar a obtenção, o tratamento, a manipulação e a análise dos dados criminais foi desenvolvida uma arquitetura centralizada (unificada), tendo como ponto principal o Sistema de Gerenciamento de Bancos de Dados (SGBD) *PostgreSQL*. Dessa forma, procurou-se automatizar ao máximo os processos existentes, que vão desde a obtenção dos *datasets* até a detecção dos padrões criminais. Além disso, essa arquitetura utilizou a ferramenta *Power BI* da *Microsoft* para a apresentação gráfica dos resultados.

Na sequência, foram realizados dois experimentos controlados, seguindo as diretrizes de (WOHLIN et al., 2012), que aplicaram *Data Science* para avaliar os dados abertos governamentais relacionados a crimes. O primeiro experimento abordou os incidentes criminais ocorridos nas Unidades Federativas do Brasil, disponibilizados pelo Governo Federal por meio do MJSP. O segundo analisou os crimes ocorridos nos munícipios de Minas Gerais, disponibilizados pela Sejusp/MG. De acordo com Wohlin et al. (2012), uma experimentação não é uma tarefa simples, pois envolve preparar, conduzir e analisar experimentos corretamente. Os autores destacam como uma das principais vantagens da experimentação o controle dos sujeitos, objetos e instrumentação, o que torna possível extrair conclusões mais gerais sobre o assunto investigado. Além disto, outra vantagem inclui a habilidade de realizar análises estatísticas, utilizando métodos de teste de hipóteses e oportunidades para replicação. Do ponto de vista da classificação, em Computação, especificamente na *Data Science*, os objetos de estudo são dados e algoritmos, os quais foram analisados em laboratório, ambiente controlado, com a averiguação e o teste de hipóteses. Desta forma, este estudo pode ser classificado como experimental, pela ocorrência de testes de hipóteses (JURISTO; MORENO, 2013) (BLACKBURN, 2016), e tem características de um estudo "*in vitro*", pela experiência em laboratório.

Em relação à classificação desta pesquisa, podemos citar, quanto à natureza, como sendo aplicada, pois produziu conhecimento para aplicação de seus resultados com o objetivo de contribuir para fins práticos, visando à solução imediata do problema encontrado na realidade (APPOLINÁRIO, 2007). Quanto à abordagem dos dados, foi considerada quantitativa, pois as variáveis estão associadas a valores numéricos, foram obtidas de medições objetivas e analisadas estatisticamente (ABRAMOVICI et al., 2008).

Os capítulos 3 e 4 descrevem os experimentos controlados, são autocontidos, do ponto de vista metodológico, pois descrevem, em cada seção, os passos adotados no processo experimental realizado.

1.6 Organização da Dissertação

Este documento está organizado de acordo com a Instrução Normativa Nº 02/2015/PROCC, a qual permite que a Dissertação seja "uma compilação de artigos científicos submetidos ou publicados em veículos com Qualis, desde que seja contextualizada com seções de Introdução e Conclusão, não limitada a estas". São 5 capítulos que fornecem uma base conceitual e experimental para o entendimento sistêmico. Os tópicos a seguir descrevem o conteúdo de cada um dos capítulos:

- O Capítulo 1 apresenta esta Introdução, explicando as justificativas juntamente com as hipóteses levantadas;
- O Capítulo 2 traz um artigo da Revisão Sistemática Quantitativa que foi publicado no periódico *Journal of Applied Security Research* (PRADO et al., 2020);
- O Capítulo 3 traz um artigo relativo ao primeiro experimento controlado (**Experimento 1**), publicado no periódico *Journal of Applied Security Research* (PRADO; COLAÇO JÚNIOR, 2020b);
- O Capítulo 4 traz um artigo relativo ao segundo experimento controlado (**Experimento 2**), publicado no periódico *Research, Society and Development* (PRADO; COLAÇO JÚNIOR, 2020a);
- Finalmente, no capítulo 5, é apresentado um compilado de conclusões, contribuições e sugestões de trabalhos futuros.

2

Análise Inteligente de Dados Aplicada a Dados Governamentais Relacionados a Incidentes Criminais: Uma Revisão Sistemática

Este capítulo traz um artigo da Revisão Sistemática Quantitativa que foi publicado no periódico *Journal of Applied Security Research* (PRADO et al., 2020).

2.1 Introdução

Com o aumento da urbanização diversas transformações sociais, econômicas e ambientais têm ocorrido em toda parte do mundo. Os desafios enfrentados pelos governantes estão cada vez maiores. Áreas como mobilidade urbana, saúde e segurança pública têm recebido uma atenção especial (CATLETT et al., 2018).

Nos últimos anos, o aumento da violência tem sido objeto de estudo de muitos pesquisadores (DAMASCENO; TEIXEIRA; CAMPOS, 2012). O crime é um problema social predominante e sua prevenção é uma função extremamente importante. Governantes e a sociedade, em geral, têm tido enormes problemas causados por esse fenômeno. A cada ano, os governos gastam milhões de dólares combatendo a violência, fornecendo equipamentos, treinamento e adquirindo ferramentas para auxiliar o trabalho policial. É da responsabilidade das agências de aplicação da lei monitorar e reduzir a taxa de atividades criminosas que estão acontecendo continuamente nos dias de hoje (DAMASCENO; TEIXEIRA; CAMPOS, 2012)(MARZAN et al., 2017).

No entanto, com a capacidade cada vez maior das organizações públicas e departamentos de polícia de coletar e armazenar dados detalhados de rastreamento de eventos criminais, uma quantidade significativa de dados com informações espaciais e temporais é obtida diariamente (CATLETT et al., 2018). O grande desafio enfrentado por essas organizações é lidar com um grande volume de informações referentes a crimes e criminosos. Sistemas computacionais inteligentes vêm desempenhando um papel vital na melhoria dos resultados das investigações e

detecções criminais, facilitando o registro, a análise de recuperação e o compartilhamento das informações (GUPTA; CHANDRA; GUPTA, 2014).

A análise criminal permite uma maior compreensão da dinâmica das atividades criminosas, fornecendo respostas como "para onde", "quando" e "por que" certos crimes são prováveis de acontecer. Esta análise é de grande importância para os governantes, autoridades policiais e os próprios residentes (PHILLIPS; LEE, 2011).

Este artigo apresenta uma Revisão Sistemática (RS) quantitativa que teve como objetivo identificar, caracterizar e metanalisar as abordagens, técnicas e algoritmos inteligentes utilizados sobre dados criminais, utilizando artigos de importantes bases de dados de Ciência da Computação (CC).

Após responder às questões de pesquisa, identificou-se que as principais abordagens exploradas foram: *Unsupervised Machine Learning*, com 42 estudos (38,53%), *Supervised Machine Learning*, com 33 (30,28%) e *Association Rules*, com 19 estudos (17,48%). Em relação aos algoritmos, *K-Means* foi o mais utilizado, alcançando 19 trabalhos (14,39%), seguido de *K-Nearest Neighbors (KNN)*, com 15 (11,36%), e *Apriori* com 13 (9,85%). No contexto dos tipos de estudo, o "Estudo de caso", com 73 (83,91%) publicações, superou largamente os outros tipos analisados. Em relação aos países, a Índia (22), Estados Unidos (16) e China (13) lideram o *ranking* de publicações sobre o tema. Entre os principais veículos, as conferências se destacaram com 62 pesquisas (71,26%), enquanto que os periódicos atingiram 25 (28,74%).

E, por fim, com intuito de consistir e agregar resultados, foi realizada uma metanálise com os trabalhos identificados como elegíveis para este fim. Neste contexto, após uma análise minuciosa dos artigos, três estudos foram selecionados ((ZHUANG et al., 2017),(KUO; CHANG; CHEN, 2017),(BAPPEE; JÚNIOR; MATWIN, 2018)), apresentando avaliações de algoritmos em comum em suas pesquisas. Para isto, foi coletada a quantidade de dados utilizados (amostra) e o total de erros das predições, Falsos Positivos (FP) somados aos Falsos Negativos (FN), realizadas por cada algoritmo aplicado. Dessa forma, foi possível combinar oito situações (gráficos), as quais são analisadas individualmente, na seção 2.5.

O restante deste artigo está organizado da seguinte forma. Na seção 2.2, os trabalhos relacionados sobre o tema são apresentados. Na seção 2.3, o método adotado neste trabalho é abordado. A seção 2.4 apresenta e discute os resultados alcançados. A seção 2.5 detalhada a metanálise executada sobre os estudos selecionados. Na seção 2.6, as ameaças à validade encontradas são detalhadas. E, finalmente, na seção 2.7, a conclusão é apresentada.

2.2 Trabalhos Relacionados

Alguns trabalhos secundários relacionados ao tema desta pesquisa foram encontrados. Com intuito de encontrar lacunas (*gaps*), Noor et al. (2015) apresentaram uma revisão sistemática

e abrangente sobre um *framework* de suporte à tomada de decisão para classificação na prevenção de crime. Quarenta e quatro artigos foram analisados e classificados em duas categorias de crimes (crimes violentos e crimes contra a propriedade) e seis classes de técnicas de mineração de dados (predição, classificação, visualização, regressão, clusterização e detecção de anomalias). Os autores propuseram um *framework* de classificação utilizando *data mining* e concluíram que as principais técnicas de mineração de dados utilizadas foram *Bayesian*, *Neural Network* e *Nearest Neighbor*.

Em (REWARI; SINGH, 2017), os pesquisadores visam rever consistentemente as técnicas, tipos de *datasets*, desafios e problemas encontrados na análise de dados criminais volumosos (*big data*) em estudos existentes, fornecendo aos pesquisadores um resumo do status na área e suas lacunas (*gaps*). Eles concluíram que a "revolução da *big data*" trouxe uma visão de que as armas contra os crimes, o extremismo e terrorismo não são mais balas ou bombas, e sim os grandes dados que aprimoram a análise criminal. Finalmente, Tomar e Manjhvar (2016) realizaram uma pesquisa abrangente sobre os fundamentos da clusterização e os vários aspectos que a afetam. Os autores relataram, de forma breve, as principais técnicas de agrupamento que podem ser aplicadas a dados criminais. Além disso, eles observaram que o *K-means*, com algum aprimoramento, poderá ajudar no processo de identificação de padrões criminais. Então, os pesquisadores utilizaram o algoritmo *Ant Colony Optimization* (ACO) para solucionar os problemas computacionais e produzir bons resultados através gráficos.

Este trabalho se distingue dos anteriores pelo fato de possuir uma maior abrangência. No tocante ao mapeamento, além de analisar as principais abordagens e algoritmos inteligentes sobre dados de incidentes criminais, também foram investigadas questões como os principais veículos de publicação, tipos de estudo, países pesquisadores, evolução da quantidade de artigos por ano e o status da utilização dos dados abertos governamentais por parte dos pesquisadores e cientistas. Adicionalmente, com objetivo de aplicar o correto emprego da evidência científica disponível, realizamos uma metanálise sobre os estudos selecionados. De acordo com Pereira e Galvão (2014) e Berwanger et al. (2007), uma RS relevante deve apresentar resultados consistentes (caso tenha sido realizada metanálise) ou a causa de heterogeneidade deve ter sido explorada. A maneira mais elaborada de resumir e divulgar os dados é por meio de metanálise, ou seja, uma soma estatística dos resultados de cada estudo.

2.3 Método

Alguns pesquisadores têm trabalhado para estabelecer métodos estáveis de aplicação do processo de revisão sistemática na literatura. Uma RS consiste em um protocolo sistemático de busca e seleção de estudos relevantes com o objetivo de extrair informações e mapear os resultados para uma questão específica de pesquisa. Por basear-se em um protocolo de pesquisa, pode ser reproduzida por outros pesquisadores (KITCHENHAM, 2004).

O presente estudo foi baseado no protocolo proposto por Kitchenham (2004) e seu o objetivo está limitado ao desenvolvimento de uma RS com a intenção de identificar, caracterizar e metanalisar trabalhos científicos a fim de caracterizar o uso de algoritmos, técnicas e abordagens inteligentes sobre dados de incidentes criminais. A definição das questões de pesquisa, seu escopo, estratégia de busca e os critérios de seleção serão descritos nas seções seguintes.

2.3.1 Questões de Pesquisa

As questões de pesquisa foram desenvolvidas com o propósito de apresentar uma visão geral da área, evidenciando aspectos chaves dos estudos primários (KITCHENHAM, 2004)(PETERSEN; VAKKALANKA; KUZNIARZ, 2015). Para este estudo, as questões de pesquisa tentam fornecer uma visão específica sobre os aspectos relevantes na análise inteligente de incidentes criminais. Além de incluir perguntas sobre quais são as abordagens e os algoritmos mais utilizados na análise de dados sobre crimes, também foram avaliadas as formas de obtenção dos *datasets* utilizados e os tipos de pesquisa realizados (experimento controlado, estudo de caso, prova de conceito, revisão sistemática, mapeamento sistemático e revisão quasi-sistemática).

O experimento controlado é uma forma de estudo experimental na qual o investigador tem controle sobre os principais aspectos do estudo e as variáveis independentes que estão sendo estudadas. Além disso, este tipo de estudo é caracterizado pelo controle sistemático das variáveis e do processo, tendo como objetivo confirmar teorias, conhecimento convencional, explorar relacionamentos, avaliar a predição de modelos ou validar medidas. Além disso, envolve a formulação de hipóteses, que precisarão ser verificadas em relação aos resultados obtidos (WOHLIN et al., 2012) (DELAMARO; JINO; MALDONADO, 2017). Já o estudo de caso baseia-se na utilização de um ou mais métodos qualitativos ou não segue uma linha rígida de investigação. Consiste geralmente no estudo aprofundado de um único “caso” ou de “casos relacionados”, sendo executada em condições típicas, por exemplo, a partir de alguns projetos típicos representativos (READ, 2003). **Um estudo de caso tende a ser um estudo observacional, enquanto um experimento é um estudo controlado**.

Prova de conceito é um termo utilizado para denominar um modelo prático que possa provar o conceito (teórico) estabelecido por uma pesquisa ou artigo técnico. Pode ser considerado também uma implementação, em geral resumida ou incompleta, de um método ou de uma ideia, realizada com o propósito de verificar que o conceito ou teoria em questão é suscetível de ser explorado de uma maneira útil (LINKEDIN, 2015)(FARIAS et al., 2019).

Uma revisão sistemática da literatura é um meio de identificar, avaliar e interpretar todas as pesquisas disponíveis relevantes para uma determinada questão de pesquisa, área temática ou fenômeno de interesse (KITCHENHAM, 2004). O processo de desenvolvimento desse tipo de estudo inclui caracterizar cada estudo selecionado, avaliar a qualidade deles, identificar conceitos importantes, comparar as análises estatísticas apresentadas e concluir sobre o que a literatura informa em relação a determinada intervenção, apontando ainda

problemas/questões que necessitam de novos estudos. As revisões sistemáticas são desenhadas para serem metódicas, explícitas e passíveis de reprodução (SAMPAIO; MANCINI, 2007). Em algumas áreas, a metanálise é descrita como uma revisão sistemática quantitativa, contudo, esta deve representar a combinação estatística de pelo menos dois estudos, para produzir uma estimativa única (BERWANGER et al., 2007), assim como foi feita neste trabalho. O mapeamento sistemático tem como objetivo fazer uma pesquisa em largura na literatura, e não em profundidade (KITCHENHAM, 2004) (PETERSEN et al., 2008), ou seja, são projetados para fornecer uma ampla visão geral de uma área de pesquisa, para estabelecer se a evidência de pesquisa existe sobre um tema e fornecer uma indicação da quantidade de evidência (KEELE et al., 2007). Já uma revisão quasi-sistemática tem o objetivo de caracterização, ou seja, não há necessidade de conhecimento prévio para realizar comparações sobre o objeto pesquisado (KITCHENHAM, 2004).

Portanto, as questões de pesquisa foram elaboradas e aplicadas nos estudos de controle do modelo PICO, para que fosse comprovada sua capacidade de caracterização e classificação. O PICO, um acrônimo para População, Intervenção, Controle e *Outcomes* (Resultados), tem por finalidade evidenciar os efeitos de uma **intervenção** em uma determinada **população**. Um ambiente de **controle** com artigos pré-mapeados foi definido para validar o **resultado** das questões e os termos de pesquisa (JAMES; RANDALL; HADDAWAY, 2016). A Tabela 1 ilustra o modelo PICO utilizado e a Tabela 2 descreve as questões de pesquisa deste trabalho.

2.3.2 Escopo da Pesquisa

Para a execução da RS, a base de dados da Scopus (SCOPUS, 2019) foi escolhida por incluir buscas em diferentes bancos de dados científicos (IEEE, ACM, Springer e Elsevier). Tais bases são responsáveis pela publicação dos principais periódicos da área de CC.

As fontes foram selecionadas de acordo com a disponibilidade de consulta através da internet, as quais foram indexadas nas bases citadas, podendo ser encontradas por meio da busca por palavras-chave. Apenas estudos em inglês, trabalhos relacionados à CC e artigos publicados em conferências, periódicos ou capítulos de livros foram selecionados.

Após executar a função de busca por título, resumo ou palavras-chave, foi utilizada a opção avançada de refinamento da busca, para selecionar apenas resultados pertencentes à área da CC, cujo idioma fosse o inglês. Também foram excluídos resultados que se referiam a recapitulações de conferências e notas.

O acesso a Scopus deu-se por meio do Portal da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (CAPES, 2019) com assinatura da instituição de ensino. Desta forma foi possível consultar os textos sem nenhuma restrição.

Tabela 1 – Modelo PICO para conformidade das questões de pesquisa.

Categoria	Descrição
População	Publicações de pesquisadores e desenvolvedores tendo em vista aplicações que promovem a análise de dados criminais.
Intervenção	Contexto das aplicações que utilizam abordagens, técnicas e algoritmos inteligentes para análises de dados criminais.
Controle	Aplicações com análises criminais sem o uso de inteligência. **Artigos de controle** **Artigos das aplicações que obedecem à intervenção:** - Data mining for prevention of crimes (SINGH; JOSHI, 2018); - Application of Classification Techniques for Prediction and Analysis of Crime in India (DAS; DAS, 2019); - Time series analysis and crime pattern forecasting of city crime data (MARZAN et al., 2017); **Artigos das aplicações mais comuns que não se enquadram na intervenção:** - Urban navigation beyond shortest route: The case of safe paths (GALBRUN; PELECHRINIS; TERZI, 2016); - On the measurement and analysis of safety in a large city (IBRAHIM; SHAFIQ, 2017);
Resultado	Análise inteligente para projetar taxas e tendências criminais.

Fonte: Elaborada pelo autor.

Tabela 2 – Questões de pesquisa.

Número	Descrição
Q1	Quais são as abordagens de análise inteligente de dados mais utilizadas como base para auxiliar a análise e a transparência de incidentes criminais em estudos primários?
Q2	Dentre as abordagens existentes, quais os algoritmos específicos utilizados para descoberta de padrões sobre dados de incidentes criminais em estudos primários?
Q3	Quais estudos primários utilizaram dados abertos de incidentes criminais em suas pesquisas?
Q4	Quais os tipos de estudos executados para análise inteligente de incidentes criminais?
Q5	Quais países possuem mais pesquisadores publicando sobre esse tema?
Q6	Em quais anos foram publicados mais trabalhos nessa área?
Q7	Quais os principais periódicos e conferências sobre o tema?
Q8	Quais os meios de publicação mais populares?

Fonte: Elaborada pelo autor.

2.3.3 Método de Busca de Publicações

Para realizar a pesquisa nas bases digitais, foi definida uma *string* de busca com a utilização de termos em inglês e do uso de vários sinônimos, associados ao pressuposto de que

os estudos estariam contidos nas áreas da computação que lidam com inteligência e análise de dados sobre ocorrências criminais. Tais termos foram identificados com auxílio dos artigos de controles do modelo PICO, descritos na seção 2.3.1 (Tabela 1), e posteriormente, refinados e adaptados para o maior aproveitamento da *string*. A tabela 3 mostra os termos, antes de refiná-los, que foram selecionados.

Tabela 3 – Categorias do modelo PICO e termos identificados para pesquisa bibliográfica antes de refiná-los.

Categoria	**Descrição**
População	Crime, felony, crime occurrence, criminal occurrence, felony occurrence.
Intervenção	Data analytics, data science, machine learning, data mining.
Controle	Análise estatística descritiva de dados criminais (Sem *strings*).
Resultado	Crime rate, crime trend, felony rate, felony trend, criminal rate, criminal trend.

Fonte: Elaborada pelo autor.

Após o refinamento, os termos ajustados foram utilizados para construir a *string* de busca, os quais estão descritos na Tabela 4.

Tabela 4 – Termos utilizados na string de busca.

Termos da string de busca		
Data mining	Crime data	Crime rate*
Data analytics	Crime occurrence*	Crime trend*
Data science	Felony data	Felony rate*
Machine learning	Felony occurrence*	Felony trend*
	Criminal data	Criminal rate*
	Criminal occurrence*	Criminal trend*

Fonte: Elaborada pelo autor.

A *string* de pesquisa gerada com os termos evidenciados acima foi:

(“crime rate” OR “crime data” OR “crime trend*” OR “crime occurrence*” OR “felony rate*” OR “felony data” OR “felony trend*” OR “felony occurrence*” OR “criminal rate*” OR “criminal data” OR “criminal trend*” OR “criminal occurrence*”) AND (“data mining” OR “data analytics” OR “data science” OR “machine learning”) AND (PUBYEAR > 2009) AND (LIMIT-TO (SUBJAREA ,“COMP”)) AND (LIMIT-TO (DOCTYPE,“cp”) OR LIMIT-TO (DOCTYPE,“ar”)) AND (LIMIT-TO (LANGUAGE,“English”))*

Por fim, como a pesquisa realizada durante o mês Março de 2019, a busca retornou 532 resultados. Após esta fase, a seleção dos artigos foi iniciada.

2.3.4 Critérios de Seleção

A fim de filtrar os documentos relevantes para esta Revisão Sistemática foram estabelecidos os critérios de inclusão e exclusão. Após as buscas realizadas, utilizando como base a *string* de

busca citada na seção 2.3.3, os resultados coletados foram contabilizados levando em consideração apenas os estudos selecionados para avaliação. Para a confirmação dos critérios de inclusão, analisou-se o resumo e a introdução de cada artigo. Paralelamente, os artigos foram analisados de acordo com os critérios de exclusão. Finalmente, após aplicar os critérios de seleção, os artigos selecionados foram lidos, analisados e encaminhados a etapa de extração dos resultados.

Os critérios de inclusão utilizados foram:

- Disponibilidade de consulta através da *web*, em bibliotecas digitais *online* e indexadas;
- Os tipos de publicação dos artigos deverão ser conferências ou periódicos;
- Garantia de resultados únicos através das palavras chaves e consultas personalizadas;
- Os artigos devem conter o tema deste estudo no título, resumo ou palavras-chave;
- Os artigos precisam explorar um algoritmo, técnica, mecanismo ou abordagem de análise inteligente sobre dados de incidentes criminais;
- Os *datasets* utilizados nas pesquisas devem conter dados oficiais e obtidos de algum órgão governamental (tais como delegacias, estações policiais, secretarias de segurança pública, entre outros).

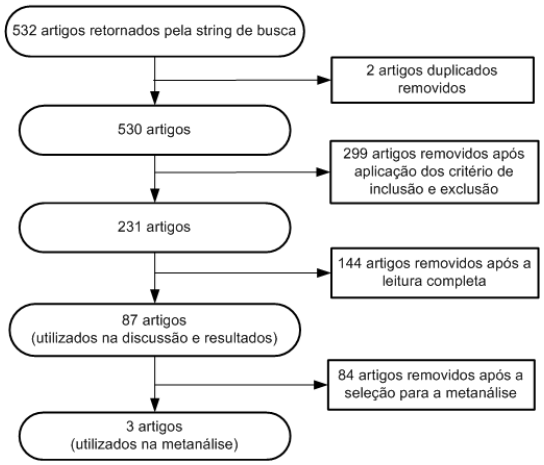
Já os critérios de exclusão utilizados foram:

- Artigos de áreas que não estejam dentro do escopo da Ciência da Computação;
- Publicações em que o idioma seja diferente do inglês;
- Estudos preliminares;
- Estudos duplicados;
- Estudos que não utilizaram *datasets* oficiais de órgãos governamentais, como: site de notícias, redes sociais, jornais *online*, entre outros;
- Trabalhos com mais de dez anos de publicação.

A busca inicial retornou um total de 532 publicações. Destas, 2 foram removidas por serem pesquisas duplicadas. Em seguida, foram aplicados os critérios de inclusão e exclusão e 299 artigos foram rejeitados, restando um total de 231 trabalhos para a leitura completa e análise. Durante esta fase percebeu-se que 144 trabalhos não atendiam as premissas propostas por esta revisão sistemática e foram descartados. Então, 87 artigos foram selecionados para a fase de extração dos dados. Finalmente, após uma análise mais profunda sobre estes trabalhos remanescentes, apenas 3 deles foram selecionados para fase de compilação dos estudos (metanálise), pois atenderam os

requisitos necessários, ou seja, apresentavam características similares (tipos de participantes, intervenções e medições de resultados). A metanálise, realizada neste trabalho, será descrita detalhadamente na seção 2.5. Na Figura 1, é possível observar de forma resumida o processo de seleção realizado para esta RS.

Figura 1 – Processo de busca e seleção dos artigos.

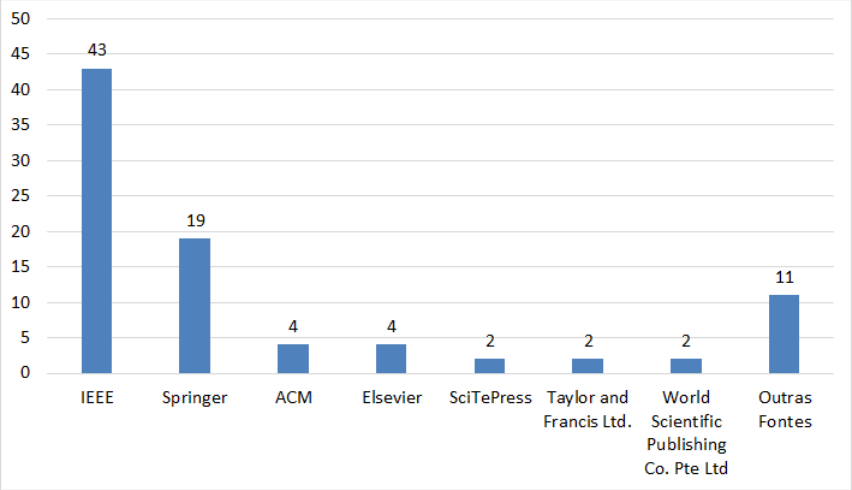


Fonte: Elaborada pelo autor.

2.4 Discussão e Resultados

Nesta seção serão apresentados os resultados dos estudos selecionados, respondendo às questões de pesquisa apresentadas anteriormente. Como foi dito, ao final do processo de seleção, 87 estudos foram selecionados e sua distribuição por repositório científico pode ser vista na Figura 2.

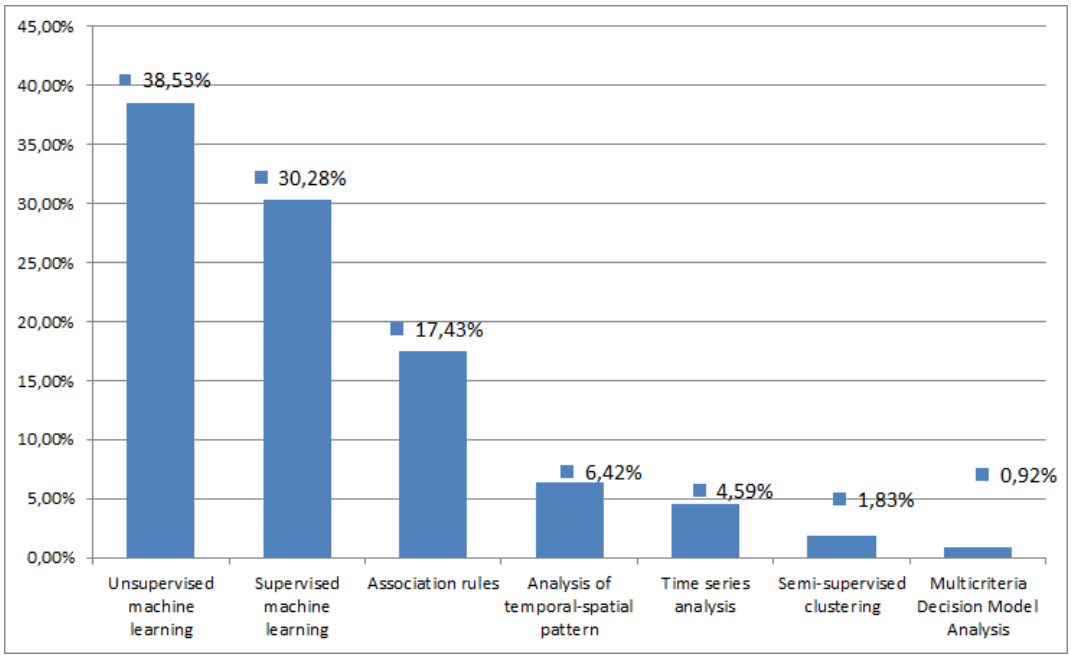
Figura 2 – Seleção de artigos por repositório científico.



Fonte: Elaborada pelo autor.

O gráfico da Figura 3 apresenta a caracterização das principais abordagens de análises inteligentes encontradas. Como resposta para a questão Q1, entre as técnicas analisadas se destacaram: *Unsupervised Machine Learning*, com 42 instâncias (38,53%) e *Supervised Machine Learning*, com 33 (30,28%). Juntas alcançaram mais de 58% das instâncias analisadas. Em seguida, estão *Association Rules*, com 19 (17,43%), *Analysis of Temporal-Spatial Pattern*, com 7 (6,42%), e *Time Series Analysis*, com 5 (4,59%). Por fim, com menos destaque, estão *Semi-Supervised Clustering*, com 2 (1,83%), e *Multicriteria Decision Model Analysis*, com 1 (0,92%). A Tabela 5 mostra, de forma mais detalhada, o mapeamento entre as abordagens encontradas e os seus respectivos trabalhos.

Figura 3 – Caracterização das abordagens de análises inteligentes.



Fonte: Elaborada pelo autor.

Tabela 5 – Referências por abordagem.

Abordagem	Referências
Unsupervised machine learning	(CATLETT et al., 2018), (GUPTA; CHANDRA; GUPTA, 2014), (SILVA et al., 2017), (FARIAS et al., 2018), (BENGTSSON; HEIN; OLSSON, 2012), (WANG et al., 2015), (OZGUL et al., 2012), (ARYAL; WANG, 2018), (TOPPIREDDY; SAINI; MAHAJAN, 2018), (BOGAHAWATTE; ADIKARI, 2013), (DUAN; XU, 2016), (BHARATHI; INDRANI; PRABAKAR, 2017), (MA; CHEN; HUANG, 2010), (R.SUJATHA, 2014), (CHAN; LEONG, 2010), (LI; KUO; TSAI, 2010), (SRIDHAR; SATHYRAJ; BALASUBRAMANIAM, 2012), (WEI et al., 2016), (BELESIOTIS; PAPADAKIS; SKOUTAS, 2018), (DAS; DAS, 2019) , (ANSARI; PRAKASH et al., 2018), (DAS; DAS, 2017), (FENG et al., 2018), (MOHAN et al., 2011), (ALKHAIBARI; CHUNG, 2017), (THOTA et al., 2017), (ZHUANG et al., 2017), (JOHANSSON; GÅHLIN; BORG, 2015), (YADAV MEET TIMBADIA; YADAV, 2017), (CAVADAS; BRANCO; PEREIRA, 2015), (SINGH; JOSHI, 2018), (LI et al., 2017), (CALVO et al., 2017), (BUCZAK; GIFFORD, 2010), (LI et al., 2018a), (BACULO et al., 2017), (TOMAR; MANJHVAR, 2018), (KELER; MAZIMPAKA, 2016), (ANDRIENKO et al., 2010), (YANG et al., 2013), (SIVARANJANI; SIVAKUMARI; AASHA, 2016) e (TAYAL et al., 2015)

Continuação na próxima página

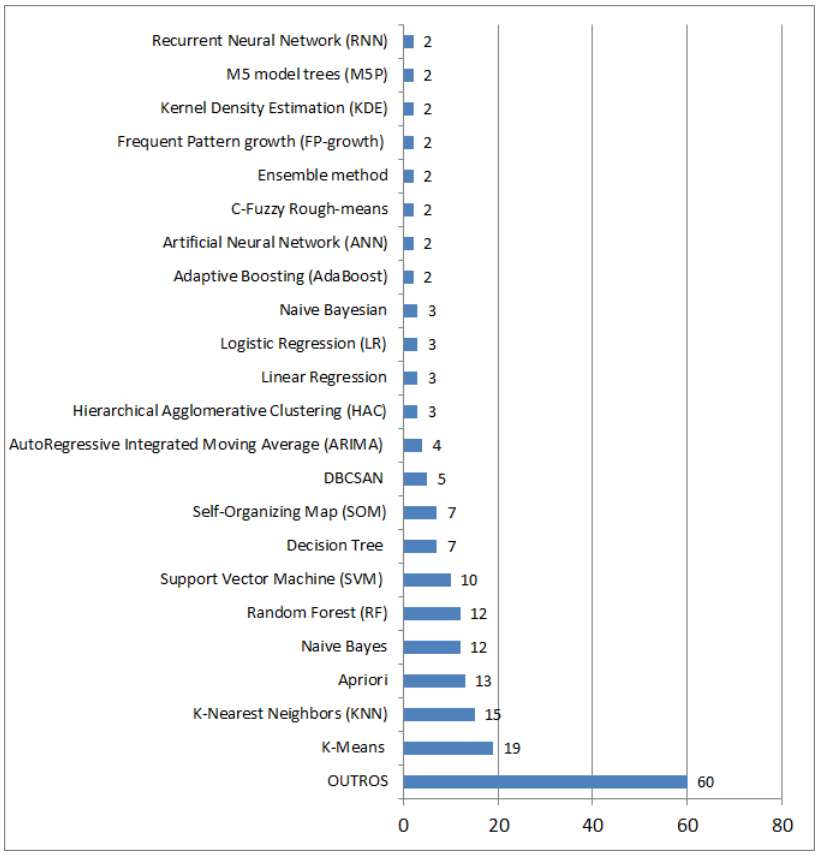
Tabela 5 – *Referências por abordagem (continuação).*

Abordagem	Referências
Supervised machine learning	(KIM et al., 2018), (CHI et al., 2017), (GUPTA; CHANDRA; GUPTA, 2014), (VINEETH; PANDEY; PRADHAN, 2016), (DAMASCENO; TEIXEIRA; CAMPOS, 2012), (TAYAL et al., 2015), (RUMI; DENG; SALIM, 2018a), (BENGTSSON; HEIN; OLSSON, 2012), (KADAR; PLETIKOSA, 2018), (HUANG; LI; JENG, 2015) (WANG et al., 2016), (AWAL et al., 2016), (RUMI; DENG; SALIM, 2018b), (AGHABABAEI; MAKREHCHI, 2015), (BONI; GERBER, 2016b), (BAPPEE; JÚNIOR; MATWIN, 2018), (KEYVANPOUR; EBRAHIMI; JAVIDEH, 2012), (SUN et al., 2014), (KUO; CHANG; CHEN, 2017), (DUAN; XU, 2016), (HUANG, 2013), (CHAN; LEONG, 2010), (SALTOS; COCEA, 2017) (DAS; DAS, 2019), (BONI; GERBER, 2016a), (FENG et al., 2018), (THOTA et al., 2017), (AGRAWAL; SEJWAR, 2017)(CAVADAS; BRANCO; PEREIRA, 2015), (LI et al., 2017) (LI et al., 2018a), (BACULO et al., 2017) e (ZHU; XIE, 2018)
Association rules	(GUPTA; CHANDRA; GUPTA, 2014), (VINEETH; PANDEY; PRADHAN, 2016), (RUIZ et al., 2014), (CHEN et al., 2015), (MARZAN et al., 2017), (SURVE; LU; DAI, 2015), (HUANG, 2013), (R.SUJATHA, 2014), (DAS; DAS, 2017), (AGRAWAL; SEJWAR, 2017) (YADAV MEET TIMBADIA; YADAV, 2017), (BALOIAN CORONEL ENRIQUE BASSALETTI, 2017), (SINGH; JOSHI, 2018), (RAJESWARI P.SURYA TEJA, 2018), (BUCZAK; GIFFORD, 2010), (CRANDELL; KORKMAZ, 2018), (SANDIG et al., 2013), (PHILLIPS; LEE, 2011) e (ISAFIADE; BAGULA, 2013)
Analysis of temporal-spatial pattern	(PHILLIPS; LEE, 2011), (KUMAR et al., 2018), (ORONG; SISON; HERNANDEZ, 2018) (YU; LAY, 2011), (PHILLIPS; LEE, 2012), (SIVARANJANI; AASHA; SIVAKUMARI, 2018) e (WANG; BROWN; GERBER, 2012)
Time series analysis	(CATLETT et al., 2018), (MARZAN et al., 2017), (LI et al., 2018b), (OLIVEIRA et al., 2018) e (CHANDRA; GUPTA, 2013)
Semi-supervised clustering	(GUPTA; CHANDRA; GUPTA, 2014) e (OZGUL et al., 2010)
Multicriteria decision model analysis	(TURET; COSTA, 2018)

Na Figura 4, os principais algoritmos encontrados nos estudos primários serão apresentados (questão Q2). Nota-se que o *K-Means* foi o mais utilizado, com 19 instâncias (14,39%). Em seguida, vieram *K-Nearest Neighbors (KNN)* e *Apriori*, com 15 (11,36%) e 13 (9,85%), respectivamente. As técnicas *Naive Bayes* e *Random Forest (RF)* foram usadas em 12 trabalhos (9,09%). Seguindo, tivemos *Support Vector Machine (SVM)*, com 10 (7,57%), *Decision Tree e Self-Organizing Map (SOM)*, com 7 (5,30%), *DBCSAN* atingiu 5 estudos (3,79%) e *AutoRegressive Integrated Moving Average (ARIMA)* foi utilizado em 4 (3,03%). Os algoritmos *Hierarchical Agglomerative Clustering (HAC)*, *Linear Regression, Logistic Regression (LR) e Naive Bayes* foram avaliados em 3 (2,27%) pesquisas. E, finalmente, na faixa de 2 artigos publicados (1,51%), *Adaptive Boosting (AdaBoost)*, *Artificial Neural Network (ANN)*, *C-Fuzzy Rough-means*, *Ensemble method*, *Frequent Pattern growth (FP-growth)*, *Kernel Density Estimation (KDE)*, *M5 model trees (M5P)* e *Recurrent Neural Network (RNN)*. Vale ressaltar que os outros 60 algoritmos foram encontrados em 60 estudos diferentes. A Tabela 28, descrita no Apêndice A, ilustra o mapeamento entre os algoritmos abordados e seus respectivos estudos.

Com relação à questão **Q3**, a Figura 5 mostra que a maioria dos estudos primários

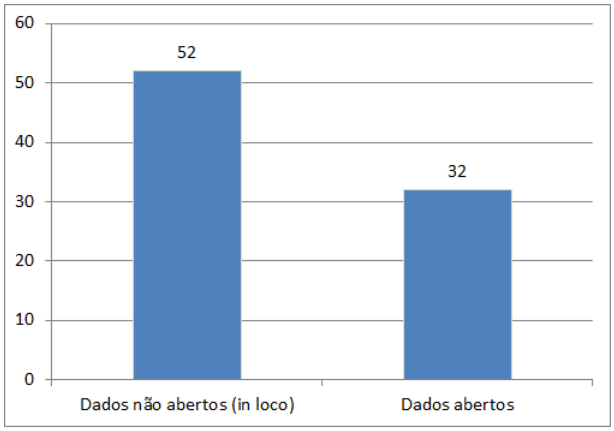
Figura 4 – Quantidade de artigos por algoritmo.



Fonte: Elaborada pelo autor.

analisados utilizaram *datasets* que não estavam disponíveis em plataformas abertas. O número de artigos contabilizados para esse grupo foi 52 (61,90%), enquanto que o outro grupo, i. e., dos que utilizaram dados abertos, somou 32 trabalhos (38,10%).

Figura 5 – Quantidade de trabalhos que utilizaram dados abertos ou não.

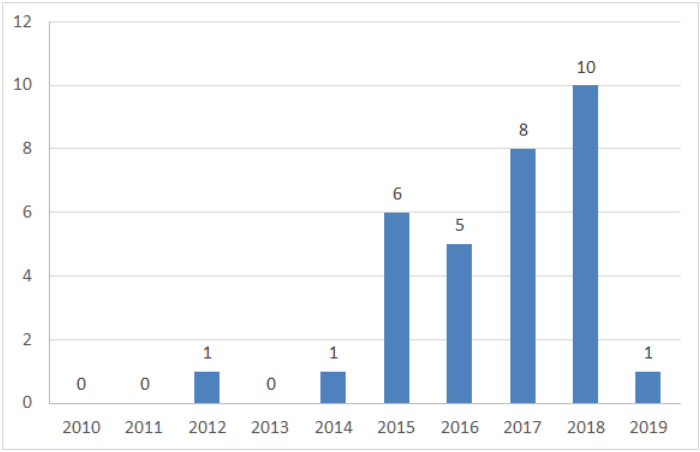


Fonte: Elaborada pelo autor.

Todavia, analisando especificamente o grupo de artigos que utilizaram fontes abertas,

notamos que nos últimos anos houve um aumento do número de pesquisas publicadas, como mostra a Figura 6. O valor de 1 artigo para o ano de 2019 é justificado pelo fato de que esta pesquisa teve início neste ano. Vale ressaltar que este trabalho apenas verificou se o conjunto de dados utilizados pelos trabalhos estavam disponíveis de forma *online*, sem se preocupar se os mesmos estavam de acordo com os princípios de dados abertos.

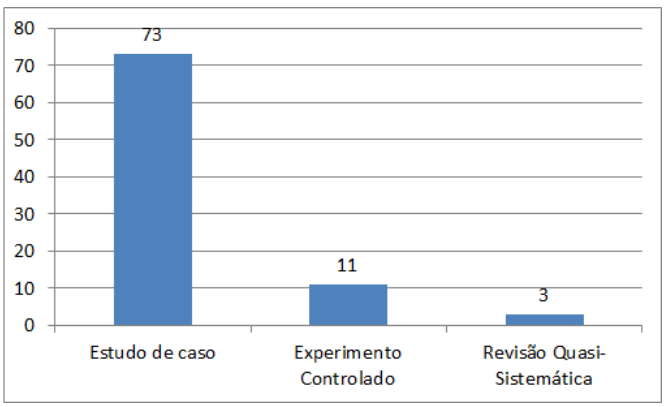
Figura 6 – Quantidade de pesquisas publicadas por ano que utilizaram dados abertos.



Fonte: Elaborada pelo autor.

Analisando a Figura 7, entre os tipos de estudos encontrados (questão Q4), o tipo "Estudo de caso" superou largamente os outros com 73 (83,91%) publicações. Em seguida, vieram o "Experimento Controlado", com 11 (12,64%) e, por fim, a "Revisão Quasi-Sistemática", com 3 (3,45%). Estudos do tipo "Mapeamento Sistemático", "Revisão Sistemática" e "Prova de Conceito" não foram encontrados. A ausência deste último tipo é justificada pelos critérios de inclusão e exclusão, em que somente trabalhos que utilizaram *datasets* oficiais foram selecionados, ou seja, conjuntos de dados com informações reais, os quais, em sua grande maioria, foram analisados em ambientes reais de produção.

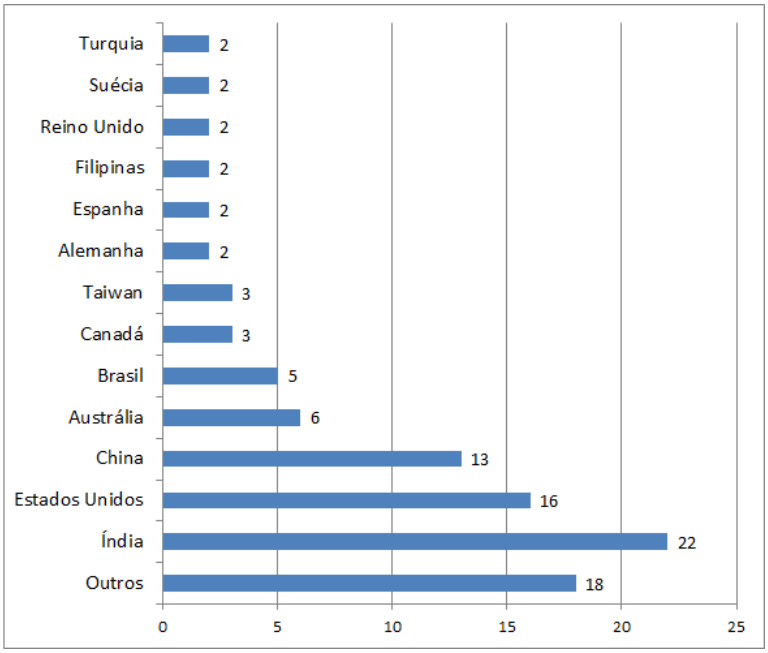
Figura 7 – Artigos por tipo de estudo.



Fonte: Elaborada pelo autor.

Na Figura 8, a qual responde a questão **Q5**, vemos que a Índia é o país com o maior número de publicações (22), seguido de Estados Unidos (16) e China (13). Posteriormente temos Austrália (6), Brasil (5) e Taiwan e Canadá com 3 trabalhos. Alemanha, Espanha, Filipinas, Reino Unido, Suécia e Turquia publicaram 2 artigos. E, por fim, um grupo (outros) composto por 18 países, no qual cada um realizou apenas uma publicação.

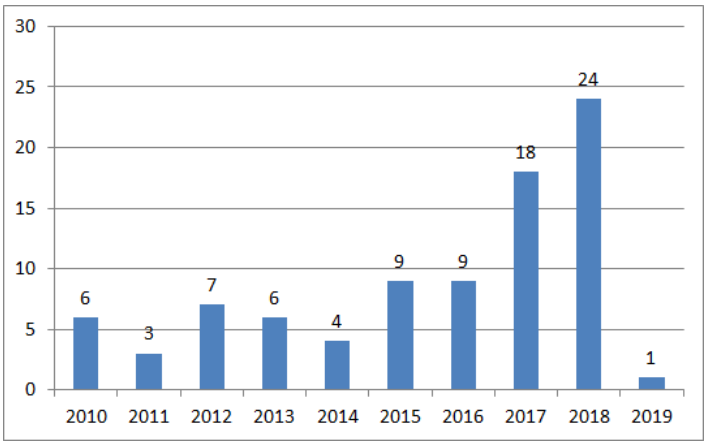
Figura 8 – Artigos por país.



Fonte: Elaborada pelo autor.

Como podemos ver, as publicações ocorreram em diversas partes do mundo caracterizando homogeneidade de publicações. Com isso, verifica-se que a área criminal é um problema global que está sendo estudada por muitos países.

Figura 9 – Publicações por ano.



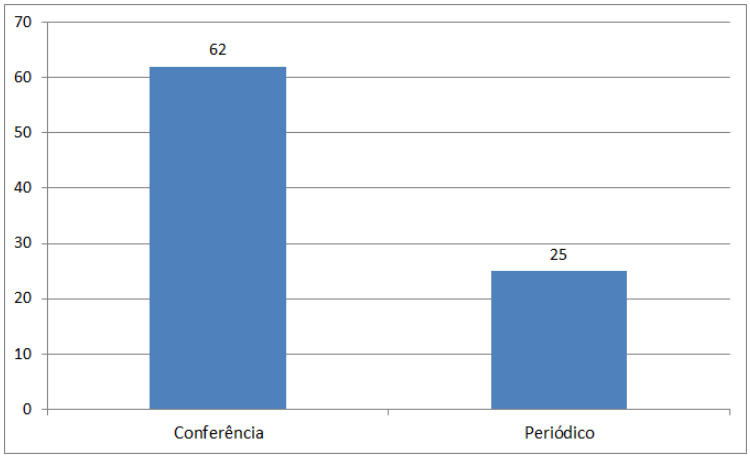
Fonte: Elaborada pelo autor.

A resposta para questão **Q6** é apresentada na Figura 9, na qual pode-se observar uma oscilação na quantidade de publicações entres os anos de 2010 e 2014. Inicialmente, no ano de 2010, a quantidade de pesquisas foram 6. Em seguida, ocorreu um decréscimo em 2011 (3), voltando a subir em 2012 (7) e tendo uma nova queda nos dois anos seguintes 2013 (6) e 2014 (4). A partir do ano de 2015 o número de trabalhos voltou a crescer consideravelmente. Em 2015 e 2016 foram 9 artigos, no ano seguinte chegou a 18, e no ano de 2018 chegou ao seu ápice, atingindo 24 trabalhos. O valor de 1 artigo para o ano de 2019 é justificado pelo fato de que esta pesquisa teve início neste ano.

O aumento no número de publicações poderá estar vinculado ao fenômeno *Open Data* ocorrido em diversos países nos últimos anos. Como demonstrado anteriormente, e pode ser visto na Figura 6, houve um aumento significante do número de pesquisas utilizando *datasets* abertos. O acesso aberto a diversas bases governamentais tem facilitado e conquistado, cada vez mais, pesquisadores ao redor do mundo.

Já a questão **Q7** trata sobre quais são as principais conferências e periódicos sobre o tema. A "*International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*" e "*International Conference on Systems, Man, and Cybernetics (SMC)*" foram as principais conferências, sendo cada uma delas a fonte de 2 publicações. Em relação aos periódicos, destacaram-se a "*EPJ Data Science*", com 3 publicações, e a "*AI and Society*", com 2. Todos os outros veículos (conferências ou periódicos) publicaram apenas um artigo cada.

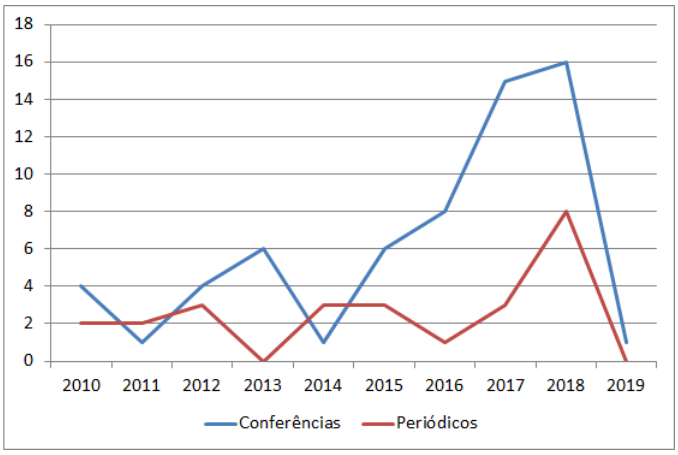
Figura 10 – Distribuição das publicações.



Fonte: Elaborada pelo autor.

E, finalmente, a Figura 10 apresenta o resultado para a questão **Q8**. Vemos que o meio mais popular para publicação dos trabalhos foi a conferência, que obteve 71,26% (62) das publicações, enquanto os periódicos tiverem 28,74% (25). Esse padrão não surpreende, visto que as conferências são conhecidamente os meios mais acessíveis para publicações científicas. A base Scopus, por exemplo, indexa mais de 120 mil eventos de conferências em todo o mundo, enquanto o número de *journals* indexados é de quase 23 mil (SCOPUS, 2019).

Figura 11 – Distribuição das publicações por ano.



Fonte: Elaborada pelo autor.

A Figura 11 representa um refinamento do gráfico apresentado na Figura 10, em que observamos a distribuição de publicações por ano, divididas pelos respectivos meios de publicação. Nota-se que nos dez anos avaliados, em apenas dois (2011 e 2014) o índice de publicações de periódicos superou, suavemente, o de conferências.

Como exposto anteriormente, com a abertura dos dados governamentais ocorrida em diversos países, houve um aumento significante do número de pesquisas publicadas. Estas publicações ocorreram em diversas partes do mundo, caracterizando homogeneidade de publicações. Com isso, verifica-se que a área criminal é um problema global que está sendo estudado por diversos pesquisadores. Além disso, foi visto que diversas abordagens/algoritmos estão sendo utilizados para auxiliar a análise inteligente de dados relacionados a incidentes criminais. Porém, uma abordagem ainda pouco explorada nesta área vem ganhando força nos últimos anos, a *Deep Learning* (DL). A DL é um subgrupo específico de técnicas de *Machine Learning*, as quais utilizam redes neurais profundas e dependem de muitos dados para o treinamento, obtendo resultados mais satisfatórios em relação às outras abordagens.

E, por fim, outro ponto a ser tocado, principalmente em relação à área de Ciência da Computação, é a carência de pesquisas que realizem replicações para consolidação e validação de trabalhos, como também, de estudos experimentais com protocolos mais rigorosos que permitam estas replicações. Consequentemente, metanálises maiores e melhores são impedidas de serem realizadas.

A seguir, será apresentada uma rara metánalise nesta área, para a qual houve a dificuldade de seleção de trabalhos aptos. Mesmo para os trabalhos selecionados, algumas informações precisaram ser inferidas, pela ausência de dados completos dos trabalhos, bem como apenas partes de cada trabalho possuíam interseções passíveis de agregação de resultados.

2.5 Metanálise para Revisão Sistemática

A metanálise é uma análise estatística utilizada para interpretar os resultados de diferentes estudos individuais e independentes, geralmente extraídos de trabalhos publicados, com a aplicação de uma ou mais técnicas estatísticas, e com o objetivo de integrá-los, combinando e resumindo seus resultados, para sintetizar as suas conclusões ou mesmo extrair uma nova conclusão, além de possibilitar a inclusão de futuros estudos que venham a ser publicados (MONTEIRO, 2010).

Assim, pode-se afirmar que a metanálise é um estudo observacional da evidência e que se baseia na aplicação do método estatístico a um estudo de revisão sistemática, integrando dois ou mais estudos primários (SANTOS; CUNHA, 2013). Uma de suas vantagens é elevar a objetividade das revisões sistemáticas de literatura, minimizando possíveis vieses e aumentando a quantidade de estudos analisados (FILHO et al., 2014). Portanto, a metanálise possibilita uma estimativa imparcial com aumento da precisão.

Os resultados de uma metanálise são representados por um gráfico do tipo *Forest plot*. A vantagem dos *Forest plots* é sumarizar todas as informações sobre a eficácia dos métodos avaliados (intervenções estudadas), acurácia dos algoritmos, no nosso caso, e a contribuição de cada estudo para a análise (BERWANGER et al., 2007).

Nesse tipo de gráfico, a linha vertical central estabelece a ausência de efeito (Risco Relativo ou RR = 1,0), indicando que não há diferenças significantes entre os grupos, referentes ao efeito das intervenções estudadas. Por outro lado, a linha horizontal estabelece os valores do risco relativo. Valores deslocados à esquerda da vertical indicam que os achados favorecem a intervenção de comparação (controle), os valores à direita favorecem o tratamento em estudo. O status de controle ou tratamento é relativo ao estudo, pois o gráfico também pode ser utilizado com dois tratamentos, nosso caso, e o deslocamento, à esquerda ou à direita, vai indicar a intervenção (algoritmo) favorecida. Para cada estudo, o ponto central representa a estimação pontual do efeito e o comprimento da linha, representa o seu intervalo de confiança. Este corresponde ao intervalo de valores dentro do qual se assume que, com 95% de significância, encontra-se o verdadeiro valor do efeito, permitindo estabelecer também direção e precisão. As linhas mais curtas dos intervalos de confiança indicam maior precisão dos resultados, melhorando a sua validade (MEDINA; PAILAQUILÉN, 2010). Quando esta linha toca o eixo central, temos a indicação de que são necessários mais estudos e mais amostras populacionais para que a significância estatística da possível diferença entre as intervenções seja alcançada.

Em metanálises com desfechos binários (por exemplo, doença/ausência de doença), os achados individuais do estudo são exibidos como “n/N”, sendo que “n” é o número de participantes com o desfecho, ou seja, com efeitos adversos, e “N” é número total de participantes da intervenção (RIED, 2006). Nos gráficos apresentados por este estudo, as varáveis “n” e “N” são representados, receptivamente, pelas variáveis *Events* e *Total*.

Os dados provenientes de vários estudos somente podem ser combinados na metanálise, se possuírem características similares (tipos de participantes, intervenções e medições de resultados). É essencial que as diferenças estatísticas dos resultados dos estudos sejam pesquisadas, desta forma, é possível avaliar a variabilidade nos resultados estimados. A conclusão mais importante de uma metanálise é o resumo quantitativo de resultados, simbolizado por um diamante (MEDINA; PAILAQUILÉN, 2010).

Os estudos tendem a ser diferentes em relação aos tipos de intervenção utilizados, amostras e definição do desfecho. Tal diferença é denominada heterogeneidade, a qual é a medida mais utilizada para avaliação da diversidade, sendo muito importante no resultado de uma metanálise (BERWANGER et al., 2007). As maneiras mais usuais de se verificar a existência de heterogeneidade em metanálises são pelo teste Q, de Cochran, ou pela estatística I^2, de Higgins e Thompson. Nos dois casos, a ideia principal é definir que a heterogeneidade das medidas de efeito é constituída de duas fontes de variação: a verdadeira heterogeneidade e o erro aleatório (RODRIGUES, 2010).

Desta forma, a heterogeneidade indica o percentual da variação do resultado entre os estudos que ultrapassam o efeito da variação aleatória (acaso), i.e., a variação entre os estudos que decorre de diferenças reais. Uma desvantagem importante do I^2 é que não há pontos de corte desenvolvidos empiricamente para determinar quando há heterogeneidade excessiva para fazer uma metanálise. Higgins et al. (2003) sugeriram interpretações de regra geral, tais como 25% representa baixa heterogeneidade, 50% representa heterogeneidade média e 75% representa alta heterogeneidade. A identificação da heterogeneidade estatística também pode ser realizada pela aplicação do teste estatístico χ^2 ou qui-quadrado, que tem como objetivo avaliar se as diferenças observadas nos resultados são compatíveis com o acaso. O valor do *p-value*, apresentado pelo χ^2, é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra sob a hipótese nula (as variáveis são independentes). Se o $p\text{-}value < 0,05$, há evidências que as variáveis são fortemente dependentes dos estudos, i.e., alta heterogeneidade (SANTOS; CUNHA, 2013).

Além disso, na avaliação da heterogeneidade, podemos observar outra variável chamada de τ de *Kendall*, ou coeficiente de correlação de *Kendall*, representado por τ^2. Seu objetivo é verificar se existe correlação não-paramétrica entre duas variáveis, sendo interpretado como uma medida de concordância entre dois conjuntos de classificações relativas a um conjunto de objetos de estudo, i.e., uma estimativa amostral de uma medida da magnitude em que os efeitos de tratamento variam entre os estudos. É adequado quando as amostras têm tamanhos reduzidos (DERSIMONIAN; LAIRD, 1986). Se $\tau^2 = 0$, assume-se que os estudos envolvidos na metanálise são independentes, $\tau^2 = 1$, os estudos são iguais, e $\tau^2 = -1$ os estudos são diferentes. Sua vantagem é a possibilidade de estimarmos medidas metanalíticas sem a necessidade de pressupormos que os estudos que compõem a metanálise são homogêneos (DERSIMONIAN; LAIRD, 1986).

Neste mesmo contexto, estudos diferentes nunca terão resultados idênticos, sempre haverá

alguma diferença. Estas diferenças resultam do acaso somado às diferenças verdadeiras. Se as diferenças são só pelo acaso, então estas vão apenas até um determinado limite. Além deste limite, o que há de diferença pode ser devido à discordância real entre os estudos. Vale ressaltar que quanto maior for o tamanho amostral dos estudos, mais fácil será para detectar a heterogeneidade, pois esses estudos serão mais precisos, o efeito do acaso se reduz e eventuais diferenças tendem a ser mais reais.

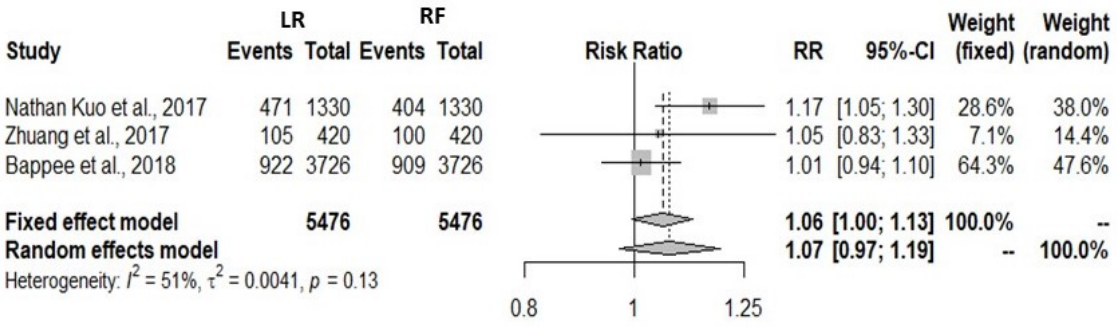
Os gráficos a seguir, elaborados com o software R (TEAM, 2014), mostram a metanálise de três estudos selecionados a partir desta RS, considerando sempre a avaliação de heterogeneidade a partir de três métodos estatísticos, e Intervalo de Confiança (IC) de 95%. Como elucidado anteriormente, vale a pena ressaltar que a variável *p*, ilustrada nos gráficos, representa o valor do *p-value* resultante da aplicação do teste do qui-quadrado (χ^2), para a avaliação da heterogeneidade.

De cada estudo foram coletados a quantidade de dados (amostra) analisados (representada no gráfico pela variável *Total*), como também o total de erros das predições realizadas por cada algoritmo aplicado (representados nos gráficos pela variável *Events*). Como dito anteriormente, a variável *Events* contabiliza os desfechos com efeitos adversos, logo, para este trabalho, representa o número de predições erradas realizadas por cada algoritmo avaliado. Tais erros podem ser computados por meio da soma da quantidade dos Falsos Positivos (FP) com a quantidade dos Falsos Negativos (FN), coletados durante a etapa de classificação. Uma outra alternativa para se chegar aos erros cometidos pelos algoritmos é por meio do valor da acurácia. A acurácia representa o percentual de acertos de um classificador ([verdadeiro positivo + verdadeiro negativo] / total de predições). Se um classificador apresenta uma acurácia de 80%, a diferença para 100%, ou seja, 20%, representa seu percentual de erros. Então, como os estudos selecionados, por não possuírem protocolo experimental rigoroso e não considerarem a possibilidade de estudos secundários a partir dos seus resultados, apenas informaram as acurácias dos seus algoritmos avaliados, foi utilizada esta última estratégia para a obtenção desses valores. Desta forma, os totais de erros de cada classificador foram obtidos utilizando a seguinte fórmula: *(1-acuracia)* * *Total*, onde, $0 <= acuracia <= 1$.

Após uma análise minuciosa dos artigos, três estudos foram selecionados para este propósito, (Zhuang et al. 2017 (ZHUANG et al., 2017), Nathan Kuo et al. 2017 (KUO; CHANG; CHEN, 2017) e Bappee et al. 2018 (BAPPEE; JÚNIOR; MATWIN, 2018)), os quais apresentaram avaliações de algoritmos em comum em suas pesquisas. Estes trabalhos tentam predizer áreas com altos índices de criminalidade (*hotspots*), explorando, principalmente, a relação entre os dados históricos dos crimes e os diversos fatores geográficos, demográficos e sociais.

Na Figura 12 é demonstrada a metanálise **(M1)** realizada entre os três estudos, observando os algoritmos em comum *Logistic Regression (LR)* e *Random Forest (RF)*. O resultado mostra que $I^2 = 51\%$, $\tau^2 = 0{,}0041$ e $p = 0{,}13$, ou seja, os estudos apresentaram uma heterogeneidade média. Os trabalhos de Bappee et al. (2018) e Nathan Kuo et al. (2017) analisaram amostras com uma quantidade maior de dados, consequentemente, obtiveram os maiores pesos no resultado

Figura 12 – Logistic Regression (LR) x Random Forest (RF)



Fonte: Elaborada pelo autor.

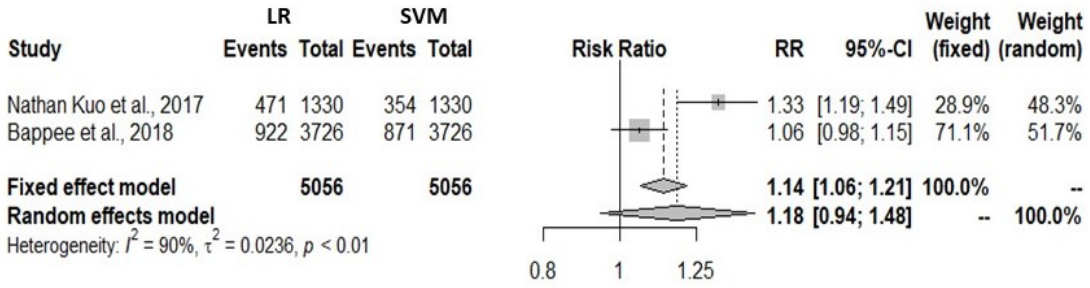
Nota: Gráfico gerado pelo Software R.

da metanálise. Ainda em (BAPPEE; JÚNIOR; MATWIN, 2018), percebe-se que a linha que representa o Intervalo de Confiança (IC) toca no eixo do Risco Relativo (RR), indicando que não há diferença estatística significativa entre os dois algoritmos avaliados, diferentemente de Nathan Kuo et al. (2017), para o qual a linha do Intervalo de Confiança (IC) não tocou o eixo do Risco Relativo (RR), revelando que há diferença estatística significativa favorecendo o *Random Forest*. Devido ao uso de uma amostra menor, Zhuang et al. (2017) apresentaram um grande Intervalo de Confiança (IC), perfazendo uma maior imprecisão. Consequentemente, este fato pode indicar que os efeitos encontrados por este estudo são atribuídos ao acaso. Em contrapartida, por ter uma composição amostral relevante, o estudo de Bappee et al. (2018) teve uma maior influência nos resultados desta metanálise, pois seus pesos (aleatório = 47,6% e fixo = 64,3%) superaram os valores dos pesos dos trabalhos de Nathan Kuo et al. (2017) (aleatório = 38% e fixo = 28,6%) e Zhuang et al. (2017) (aleatório = 14,4% e fixo = 7,1%).

Apesar de cada um dos três estudos demostrar que *Random Forest (RF)* é melhor do que *Logistic Regression (LR)*, o resultado final (metanálise) indica que não há diferença estatística significante, i.e., favorecimento de um dos dois algoritmos, considerando tanto o efeito fixo, quando as diferenças são ignoradas e é mais fácil encontrar evidências, quanto ao efeito aleatório, quando as diferenças dos estudos são levadas em consideração. Ambos os diamantes tocam ou ultrapassam o eixo central.

A Figura 13 apresenta a metanálise **(M2)** realizada entre os estudos Bappee et al. (2018) e Nathan Kuo et al. (2017), observando os algoritmos *Logistic Regression (LR)* e *Support Vector Machine (SVM)*. Nota-se que I^2 = 90%, τ^2 = 0,0236 e $p < 0,01$, ou seja, os estudos apresentaram uma heterogeneidade alta. Como pode ser visto, em Bappee et al. (2018), o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR), indicando que não existe diferença estatística significativa entres os métodos avaliados, diferentemente de Nathan Kuo et al. (2017), para o qual a linha do Intervalo de Confiança (IC) não ultrapassou o eixo do Risco Relativo (RR),

Figura 13 – Logistic Regression (LR) x Support Vector Machine (SVM)



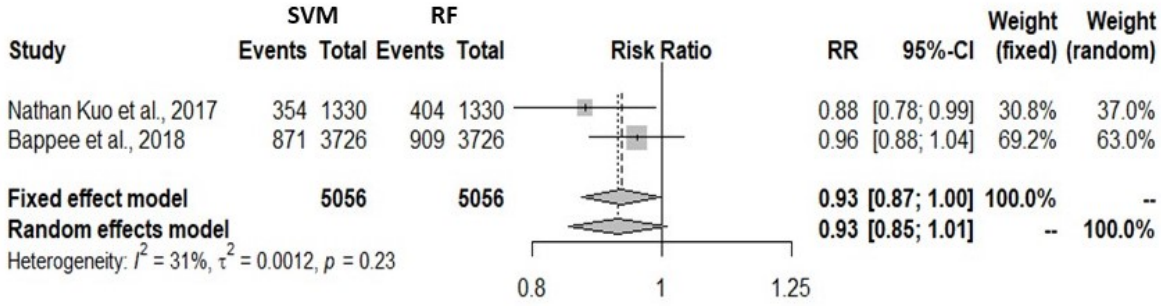
Fonte: Elaborada pelo autor.

Nota: Gráfico gerado pelo Software R.

mostrando que há diferença estatística significativa.

Os estudos demonstraram, individualmente, que *Support Vector Machine (SVM)* é melhor do que *Logistic Regression (LR)*. Contudo, o resultado da metanálise indica que, considerando o efeito fixo, é estatisticamente significante a vantagem do *Support Vector Machine (SVM)* sobre o *Logistic Regression (LR)*, na análise de dados sobre incidentes criminais. Em relação ao efeito aleatório, não há diferença estatística significativa entre os algoritmos avaliados, isto pode ser explicado pela alta heterogeneidade dos estudos e pode ser visto pela projeção do diamante, ultrapassando o eixo central. Com relação aos pesos, percebe-se que, para qualquer um dos efeitos, Bappee et al. (2018) obtiveram valores (aleatório = 51,7% e fixo = 71,1%) superiores aos da outra pesquisa (aleatório = 48,3% e fixo = 28,9%), produzindo uma influência maior deste trabalho no resultado da metanálise.

Figura 14 – Support Vector Machine (SVM) x Random Forest (RF)



Fonte: Elaborada pelo autor.

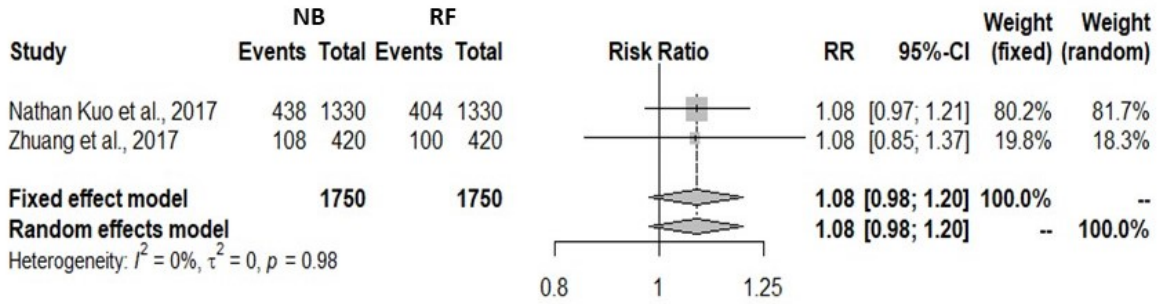
Nota: Gráfico gerado pelo Software R.

A Figura 14 apresenta a metanálise **(M3)** realizada entre os estudos Bappee et al. (2018) e Nathan Kuo et al. (2017), observando os algoritmos *Support Vector Machine (SVM)* e *Random*

Forest (RF). Como pode ser visto, I^2 = 31%, τ^2 = 0,0012 e p = 0,23, ou seja, os estudos apresentaram uma heterogeneidade média. Nota-se que, em Bappee et al. (2018), o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR), indicando que não existe diferença estatística significativa entres os algoritmos avaliados, diferentemente de Nathan Kuo et al. (2017), para o qual a linha do Intervalo de Confiança (IC) não ultrapassou o eixo do Risco Relativo (RR), mostrando que há diferença estatística significativa. Além disso, devido ao uso de uma amostra menor, a pesquisa realizada por Nathan Kuo et al. (2017) apresentou um Intervalo de Confiança (IC) mais aberto, perfazendo uma maior imprecisão. Consequentemente, este fato pode indicar que os efeitos encontrados por este estudo são atribuídos ao acaso. Como no caso anterior, os pesos obtidos por Bappee et al. (2018) (aleatório = 63,0% e fixo = 69,2%) superaram os da outra pesquisa (aleatório = 37,0% e fixo = 30,8%), provocando uma maior influência deste trabalho no resultado da metanálise.

Individualmente, os estudos mostraram que o *Support Vector Machine (SVM)* é melhor que *Random Forest (RF)*, todavia, o resultado final indica que não há diferença estatística significante entre os algoritmos, considerando o efeito fixo ou aleatório.

Figura 15 – Naive Bayes (NB) x Random Forest (RF)



Fonte: Elaborada pelo autor.

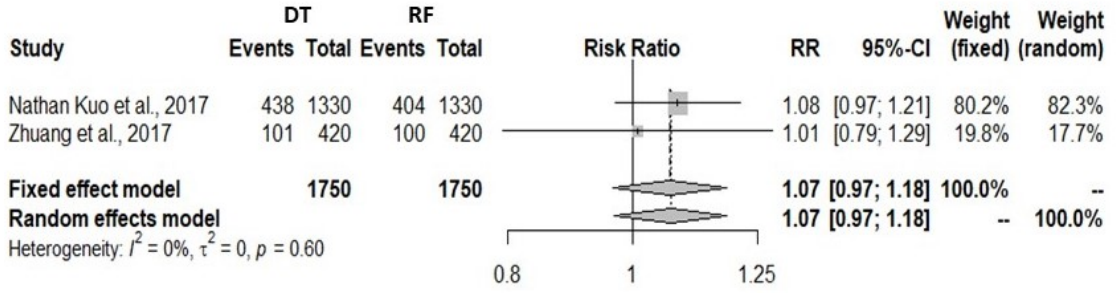
Nota: Gráfico gerado pelo Software R.

A Figura 15 apresenta a metanálise **(M4)** realizada entre os estudos Zhuang et al. (2017) e Nathan Kuo et al. (2017), observando os algoritmos *Naive Bayes (NB)* e *Random Forest (RF)*. O gráfico mostra que I^2 = 0%, τ^2 = 0 e p = 0,98, ou seja, os estudos apresentaram uma heterogeneidade baixa. Como pode ser visto, nos dois estudos, o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR), constatando que não houve diferença significante entre os algoritmos. Devido ao tamanho de sua amostra, percebe-se que o trabalho de Zhuang et al. (2017) apresentou um Intervalo de Confiança (IC) maior, o que pode indicar que os efeitos encontrados por este estudo são atribuídos ao acaso. Em relação aos pesos, para qualquer um dos efeitos, Nathan Kuo et al. (2017) obtiveram valores superiores (aleatório = 81,7% e fixo = 80,2%) aos da outra pesquisa (aleatório = 18,3% e fixo = 19,8%), provocando uma influência

maior deste estudo no resultado da metanálise.

Apesar dos estudos demonstrarem que *Random Forest (RF)* é melhor do que *Naive Bayes (NB)*, para este caso, o resultado final indica que não há diferença estatística significante entre os algoritmos, considerando o efeito fixo ou aleatório.

Figura 16 – Decision Tree (DT) x Random Forest (RF)



Fonte: Elaborada pelo autor.

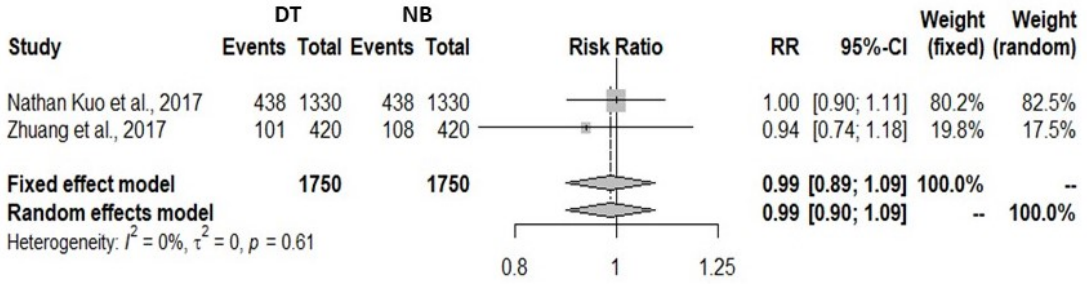
Nota: Gráfico gerado pelo Software R.

A Figura 16 exibe a metanálise **(M5)** realizada entre os estudos Zhuang et al. (2017) e Nathan Kuo et al. (2017), observando os algoritmos *Decision Tree (DT)* e *Random Forest (RF)*. O resultado mostra que o I^2 = 0%, τ^2 = 0 e p = 0,60, ou seja, os estudos apresentaram uma heterogeneidade baixa. Como na metanálise anterior, em ambos os casos, o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR), constatando que não houve diferença significante entre os algoritmos. Além disso, o trabalho de Zhuang et al. (2017) apresentou, mais uma vez, um Intervalo de Confiança (IC) mais aberto em relação a outra pesquisa, causada, como já vimos, pela sua composição amostral não expressiva. Quanto aos pesos, o estudo de Nathan Kuo et al. (2017), por ter uma amostra mais relevante, ocasionou uma maior influência nos resultados desta metanálise, uma vez que seus valores (aleatório = 82,3% e fixo = 80,2%) superaram os de Zhuang et al. (2017) (aleatório = 17,7% e fixo = 19,8%).

Apesar de os estudos, individualmente, apontarem que *Random Forest (RF)* é melhor que *Decision Tree (DT)*, a metanálise sinaliza que não há diferença estatisticamente significativa entre os métodos avaliados, considerando qualquer efeito (fixo ou aleatório).

A Figura 17 apresenta a metanálise **(M6)** realizada entre os estudos Zhuang et al. (2017) e Nathan Kuo et al. (2017), observando os algoritmos *Decision Tree (DT)* e *Naive Bayes (NB)*. Nota-se que I^2 = 0%, τ^2 = 0 e p = 0,61, ou seja, os estudos apresentaram uma heterogeneidade baixa. Nas duas pesquisas, o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR), constatando que não houve diferença significante entre os algoritmos. No tocante aos pesos, para qualquer um dos efeitos, Nathan Kuo et al. (2017) obteviveram valores superiores (aleatório = 82,5% e fixo = 80,2%) os pesos de Zhuang et al. (2017) (aleatório = 17,5% e fixo

Figura 17 – Decision Tree (DT) x Naive Bayes (NB)



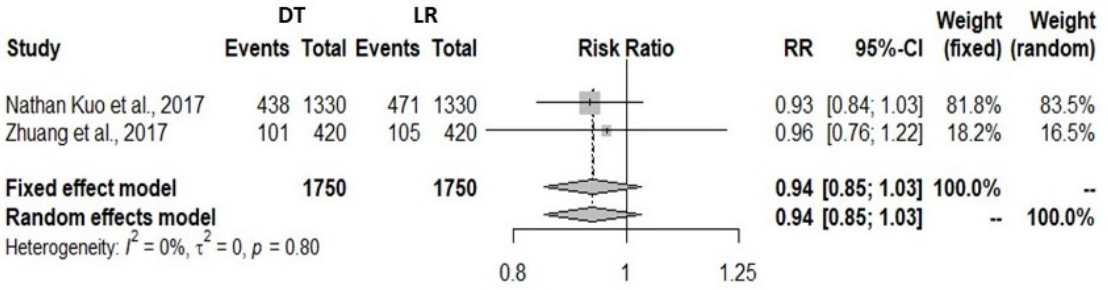
Fonte: Elaborada pelo autor.

Nota: Gráfico gerado pelo Software R.

= 19,8%). Adicionalmente, percebe-se que, o estudo de Zhuang et al. (2017) apresentou um Intervalo de Confiança (IC) mais amplo, perfazendo uma maior imprecisão. Consequentemente, este fato pode indicar que os efeitos encontrados por este estudo são atribuídos ao acaso.

Individualmente, os estudos demonstraram que o *Decision Tree (DT)* é levemente superior ao *Naive Bayes (NB)*. Todavia, o resultado da compilação mostra que não há diferença estatística significante entre os algoritmos, considerando o efeito fixo ou aleatório.

Figura 18 – Decision Tree (DT) x Logistic Regression (LR)



Fonte: Elaborada pelo autor.

Nota: Gráfico gerado pelo Software R.

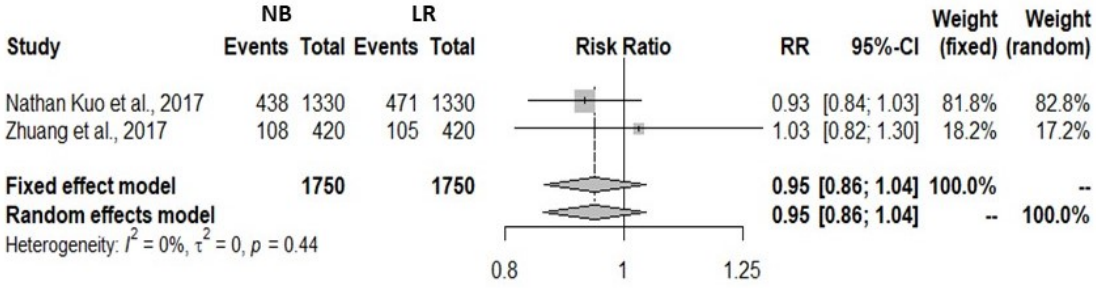
A Figura 18 apresenta a metanálise **(M7)** realizada entre os estudos Zhuang et al. (2017) e Nathan Kuo et al. (2017), observando aos algoritmos *Decision Tree (DT)* e *Logistic Regression (LR)*. Nesta metanálise, o I^2 = 0%, τ^2 = 0, e p = 0,80, ou seja, os estudos apresentam uma heterogeneidade baixa. Como pode ser visto, o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR) em ambos os casos, constatando que não houve diferença significante entre os algoritmos. Como no caso anterior, Nathan Kuo et al. (2017) obtiveram valores superiores (aleatório = 83,5% e fixo = 81,8%) aos da outra pesquisa (aleatório = 16,5% e

fixo = 18,2%), provocando uma influência maior deste estudo no resultado da metanálise.

Por utilizar uma amostra pequena em seus estudos e, consequentemente, perfazendo uma maior imprecisão, Zhuang et al. (2017) apresentaram um Intervalo de Confiança (IC) mais aberto, o que pode significar que os efeitos encontrados por este estudo são atribuídos ao acaso.

Adicionalmente, os estudos demostraram que o algoritmo *Decision Tree (DT)* é superior ao *Logistic Regression (LR)*. Entretanto, a metanálise sinaliza que não há diferença estatisticamente significativa entre os métodos avaliados, considerando qualquer efeito (fixo ou aleatório).

Figura 19 – Naive Bayes (NB) x Logistic Regression (LR)



Fonte: Elaborada pelo autor.

Nota: Gráfico gerado pelo Software R.

E, por fim, a Figura 19 apresenta a metanálise **(M8)** realizada entre os estudos Zhuang et al. (2017) e Nathan Kuo et al. (2017), avaliando-se os algoritmos *Naive Bayes (NB)* e *Logistic Regression (LR)*. Percebe-se que $I^2 = 0\%$, $\tau^2 = 0$ e $p = 0{,}44$, ou seja, os estudos apresentaram uma heterogeneidade baixa. Novamente, em ambos os casos, o Intervalo de Confiança (IC) ultrapassou o eixo (central) do Risco Relativo (RR), constatando que não houve diferença significante entre os algoritmos. Com relação aos pesos, os valores atingidos por Nathan Kuo et al. (2017) (aleatório = 82,8% e fixo = 81,8%) foram superiores aos alcançados pelo outro trabalho (aleatório = 17,2% e fixo = 18,2%). Como nas análises anteriores e justificado pela sua composição amostral, o estudo de Zhuang et al. (2017) apresentou um Intervalo de Confiança (IC) mais amplo, o que pode indicar que os efeitos encontrados por este estudo são atribuídos ao acaso, produzindo uma maior imprecisão.

Apesar de os estudos mostrarem resultados opostos, ou seja, o *Naive Bayes (NB)* foi melhor em Nathan Kuo et al. (2017) e *Logistic Regression (LR)* em Zhuang et al. (2017), a metanálise demonstra que não há diferença estatisticamente significante entre os algoritmos analisados.

De forma resumida, a Tabela 6 mostra os resultados das metanálises aplicadas por este trabalho. Do total de oito compilações, tivemos 5 metanálises com heterogeneidade baixa (**M4, M5, M6, M7 e M8**), 2 médias (**M1 e M3**) e 1 alta (**M2**). Nota-se, também, que a metanálise M2,

realizada entre os trabalhos (KUO; CHANG; CHEN, 2017) e (BAPPEE; JÚNIOR; MATWIN, 2018) foi a única que apresentou diferença estatística significativa entre os algoritmos avaliados (*Logistic Regression* e *Support Vector Machine*), considerando o efeito fixo. Neste caso, o SVM superou o outro classificador.

Tabela 6 – Resumo da metanálise.

Metanálise	**Artigos Avaliados**	**Algoritmos Avaliados**	**Heterogeneidade (I^2 %)**	χ^2 ***p-value***	**Algoritmo vencedor (com diferença estatística significante)**
M1	(ZHUANG et al., 2017), (KUO; CHANG; CHEN, 2017) e (BAPPEE; JÚNIOR; MATWIN, 2018)	LR e RF	Média (51%)	p = 0,13	Não Houve
M2	(KUO; CHANG; CHEN, 2017) e (BAPPEE; JÚNIOR; MATWIN, 2018)	LR e SVM	Alta (90%)	p <0,01	SVM (efeito fixo)
M3	(KUO; CHANG; CHEN, 2017) e (BAPPEE; JÚNIOR; MATWIN, 2018)	SVM e RF	Média (31%)	p = 0,23	Não Houve
M4	(ZHUANG et al., 2017) e (KUO; CHANG; CHEN, 2017)	NB e RF	Baixa (0%)	p = 0,98	Não Houve
M5	(ZHUANG et al., 2017) e (KUO; CHANG; CHEN, 2017)	DT e RF	Baixa (0%)	p = 0,60	Não Houve
M6	(ZHUANG et al., 2017) e (KUO; CHANG; CHEN, 2017)	DT e NB	Baixa (0%)	p = 0,61	Não Houve
M7	(ZHUANG et al., 2017) e (KUO; CHANG; CHEN, 2017)	DT e LR	Baixa (0%)	p = 0,80	Não Houve
M8	(ZHUANG et al., 2017) e (KUO; CHANG; CHEN, 2017)	NB e LR	Baixa (0%)	p = 0,44	Não Houve

Fonte: Elaborada pelo autor.

2.6 Ameaças à Validade

As ameaças à validade podem limitar a habilidade de interpretar e/ou descrever resultados dos dados obtidos. Portanto, não há como desconsiderar as seguintes ameaças encontradas nesse estudo.

- **Validade de Construção:** A *string* de busca e as questões de pesquisa utilizadas podem não cobrir a área de análise inteligente de dados relacionados às ocorrências criminais. Para mitigar essa ameaça, tentou-se elaborar uma *string* mais abrangente possível, quanto aos termos que pudessem ser usados na área, utilizando vários sinônimos. Tais termos foram identificados e refinados, com auxílio dos artigos de controles norteados pelo modelo PICO, utilizando trabalhos que interessavam à pesquisa (intervenção) e falsos positivos, com o objetivo de calibrar a *string* de busca. Além disso, foram consideradas as opiniões de três pesquisadores.

- **Validade Interna: (Extração de dados):** Três pesquisadores foram responsáveis por extrair e classificar os dados de cada publicação. Logo, vieses ou problemas na extração dos dados podem ameaçar a validade da caracterização dos dados. **(Viés de Seleção):** Inicialmente, os artigos foram incluídos ou excluídos de acordo com julgamento dos próprios pesquisadores. Consequentemente, alguns estudos podem ter sido categorizados incorretamente. Para mitigar estas ameaças, as revisões da seleção e extração foram feitas por todos os pesquisadores envolvidos e as discordâncias encontradas foram resolvidas em uma votação final. **(Viés de Classificação):** Alguns artigos selecionados não deixaram claro de que forma eles obtiveram o conjunto de dados utilizados em seus trabalhos, ou seja, se os dados utilizados são abertos ou não. Para mitigar esses vieses, os sites referenciados pelos trabalhos foram acessados e avaliados na busca de tais informações. Caso não fossem encontrados, estes dados foram classificados como não abertos. Por fim, a interpretação do tamanho da amostra utilizado para o cálculo da acurácia pode ter sido influenciada pelas faltas de detalhe e clareza de algum artigo.

- **Validade Externa:** Apesar da Scopus ser a maior base de literatura científica, com mais de 21.950 *journals* e 120 mil conferências (SCOPUS, 2019), não é possível afirmar que os resultados dessa revisão sistemática abrangeram toda a área da Ciência da Computação. No entanto, este trabalho apresentou evidências das principais técnicas utilizadas, identificando lacunas a serem exploradas e servindo como guia para futuros trabalhos nesta linha.

2.7 Conclusão

Neste trabalho, foi realizada uma revisão sistemática quantitativa, visando analisar artigos científicos para identificar, caracterizar e metanalisar as abordagens, técnicas e algoritmos pertinentes a análise inteligente de dados aplicada a dados relacionados à incidentes criminais, seguindo o protocolo de pesquisa e seleção de estudos apresentados na seção 2.3. Com este método, os dados de 87 trabalhos, sendo 73 estudos de caso, 11 experimentos controlados e 3 revisões quasi-sistemáticas foram extraídos e analisadas, identificando tendências nesta área.

Como resultados, as principais abordagem exploradas foram *Unsupervised machine learning* com 42 estudos (38,53%), *Supervised machine learning*, com 33 (30,28%), e *Association rules*, com 19 (17,48%). Em relação aos algoritmos, os mais utilizados foram K-Means, com 19 trabalhos (14,39%), seguido de *K-Nearest Neighbors (KNN)*, com 15 (11,36%), e *Apriori*, com 13 (9,85%). No contexto dos tipos de estudos, destacaram-se o "Estudo de caso", com 73 (83,91%) publicações, o "Experimento Controlado", com 11 (12,64%), e, por fim, a "Revisão Quasi-Sistemática", com 3 (3,45%). Entre os países, a Índia (22), os Estados Unidos (16) e a China (13) lideraram o *ranking* de publicações sobre o tema. Além disso, as conferências "*International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*" e "*International Conference on Systems, Man, and Cybernetics (SMC)*", e os periódicos "*EPJ Data*

Science" e "*AI and Society*", destacaram-se como maiores publicadores.

Na metanálise, foram analisados três estudos (Zhuang et al. (ZHUANG et al., 2017), Nathan Kuo et al. (KUO; CHANG; CHEN, 2017) e Bappee et al. (BAPPEE; JÚNIOR; MATWIN, 2018)) observando os algoritmos *Decision Tree, Logistic Regression, Naive Bayes, Random Forest* e *Support Vector Machine*. Dentre as oito metanálises avaliadas, tivemos 5 com heterogeneidade baixa (**M4, M5, M6, M7 e M8**), 2 médias (**M1 e M3**) e 1 alta (**M2**). Além disso, apenas uma compilação (**M2**), realizada entre os trabalhos Nathan Kuo et al. (KUO; CHANG; CHEN, 2017) e Bappee et al. (BAPPEE; JÚNIOR; MATWIN, 2018), apresentou diferença estatística significativa entre os algoritmos avaliados (*Logistic Regression* e *Support Vector Machine*), considerando o efeito fixo, elegendo o SVM vencedor.

E, finalmente, um ponto a ser ressaltado, principalmente em relação à área de Ciência da Computação, é a carência de pesquisas que realizem replicações para consolidação e validação de trabalhos, bem como, a escassez de estudos experimentais com protocolos mais rigorosos que permitam estas replicações. Consequentemente, metanálises maiores e melhores são impedidas de serem realizadas. Este estudo descreve um caminho para realização de metanálises com trabalhos de avaliação de algoritmos e similares, assim como também evidencia que abordagens específicas de agregação de estudos necessitam ser padronizadas, criadas e/ou adaptadas para área de Ciência da Computação.

Acreditamos que este trabalho é relevante para a academia, governantes, autoridades policiais e a comunidade em geral, apresentando-lhes tendências por meio da análise inteligente sobre dados criminais. Além disso, pode oferecer ainda outra abordagem na busca pelas melhores soluções para o cenário atual e para combater a ameaça de crimes contra a sociedade.

3

Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais do Brasil

Este capítulo traz um artigo relativo ao primeiro experimento controlado, publicado no periódico *Journal of Applied Security Research* (PRADO; COLAÇO JÚNIOR, 2020b).

3.1 Introdução

Crime é um problema social comum e complexo, que afeta a qualidade de vida, o crescimento econômico e a reputação de uma nação. Nos últimos anos, o aumento da violência tem sido objeto de estudo de muitos pesquisadores (DAMASCENO; TEIXEIRA; CAMPOS, 2012), uma vez que, com o aumento da urbanização, diversas transformações sociais, econômicas e ambientais têm ocorrido em todas as partes do mundo, com desafios cada vez maiores enfrentados pelos governantes. Áreas como mobilidade urbana, saúde e segurança pública têm recebido uma atenção especial (CATLETT et al., 2018).

Governantes e a sociedade, em geral, têm tido enormes problemas causados por esse fenômeno. A cada ano, os governos gastam milhões de dólares combatendo a violência, fornecendo equipamentos, treinamento e adquirindo ferramentas para auxiliar o trabalho policial. É responsabilidade das agências de aplicação da lei monitorar e reduzir a taxa de atividades criminosas que estão acontecendo continuamente nos dias de hoje (DAMASCENO; TEIXEIRA; CAMPOS, 2012)(MARZAN et al., 2017), consequentemente, a prevenção e controle do crime são questões de grande preocupação para os governos e agências de segurança pública. Tais questões, se não forem bem controladas e gerenciadas, podem afetar drasticamente a economia de um país ao longo do tempo, uma vez que mais emigração ocorrerá naturalmente (TOPPIREDDY; SAINI; MAHAJAN, 2018).

Com a capacidade cada vez maior das organizações públicas e departamentos de polícia

de coletar e armazenar dados detalhados de rastreamento de eventos criminais, uma quantidade significativa de dados com informações espaciais e temporais é obtida diariamente (CATLETT et al., 2018). O grande desafio enfrentado por essas organizações é lidar com um grande volume de informações referentes aos crimes e criminosos. Consequentemente, novas abordagens e sistemas avançados são necessários para melhorar a análise de crimes e para proteger suas comunidades (TOPPIREDDY; SAINI; MAHAJAN, 2018). Neste contexto, a *Data Science*, aliada aos sistemas computacionais inteligentes, vem desempenhando um papel vital na melhoria dos resultados das investigações e detecções criminais, facilitando o registro, a análise de recuperação e o compartilhamento das informações (GUPTA; CHANDRA; GUPTA, 2014).

Por outro lado, a transparência e o acesso às informações públicas vêm se tornando pilares essenciais para administração pública moderna. Com a evolução substancial da Tecnologia da Informação e Comunicação (TIC), novas maneiras de disponibilizar informações públicas à população estão sendo criadas. Novos sistemas foram desenvolvidos, novos serviços oferecidos e integrações sistêmicas aconteceram. Todas as revoluções tecnológicas, juntamente com a popularização da internet, geraram mudanças nos processos internos e nas relações do governo com o público externo. Tais mudanças podem ser chamadas de Governo Eletrônico, ou simplesmente *e-Gov*. Segundo Janssen, Charalabidis e Zuiderwijk (2012), com o *e-Gov*, os órgãos públicos estão entre os maiores criadores e coletores de dados em muitos domínios. Estes domínios de dados variam entre tráfego, clima, informações geográficas, turísticas, segurança pública, estatísticas, negócios, orçamentação do setor público e vários outros.

Entretanto, apesar das ações realizadas pelos governos federais e estaduais brasileiros para publicitar informações sobre estatísticas oficiais de segurança pública, nenhuma dessas iniciativas que promovem a transparência parece ser suficiente para produzir informações claras, consistentes e transparentes ao grande público. São milhares de fluxos de dados publicados frequentemente, sem a certeza de que sejam utilizáveis e sem uma aplicação inteligente sobre os dados avaliados. Muitas vezes, sem métricas que possibilitem ao cidadão inferir conclusões acerca da publicação.

Em razão disso, a proposta deste artigo é aplicar *Data Science* e conduzir um processo experimental, seguindo as diretrizes de (WOHLIN et al., 2012), para realizar uma avaliação sobre dados abertos governamentais relacionados a incidentes criminais, das Unidades Federativas (UF) do Brasil, disponibilizados pelo Governo Federal por meio do Ministério da Justiça e Segurança Pública (MJSP). Logo, esta pesquisa objetiva detectar padrões e anomalias, bem como promover maior transparência, visando auxiliar o processo de apoio às tomadas de decisões estratégicas e operacionais dos governantes e agentes da lei, no combate efetivo da criminalidade.

O restante deste artigo está organizado da seguinte forma. Na seção 3.2, os trabalhos relacionados sobre o tema são apresentados. Na seção 3.3, a metodologia adotada é abordada. A seção 3.4 descreve alguns conceitos básicos necessários para o entendimento deste trabalho. Na seção 3.5, a definição e o planejamento do experimento controlado são apresentados. A seção

3.6, detalha a operação do experimento, desde a preparação até a coleta dos dados. Na seção 3.7, os resultados são analisados. E, finalmente, na seção 3.8, a conclusão e os trabalhos futuros são apresentados.

3.2 Trabalhos Relacionados

As técnicas de *Data Mining* têm se mostrado eficazes na análise de conjuntos de dados e na coleta de informações úteis em muitos domínios. No campo criminal, a mineração de dados está recebendo maior atenção para descobrir padrões subjacentes nos dados sobre crimes (SINGH; JOSHI, 2018). A seguir, são descritos alguns trabalhos que utilizaram bases de dados que, minimamente, contêm informações semelhantes às disponibilizadas pelo governo brasileiro. Na maioria dos casos, em uma situação diferente do Brasil, o maior detalhamento dos dados fornecido pelos governos permite o uso de mais opções de algoritmos, não explorados neste artigo pela incompletude dos dados disponíveis.

Em (KIM et al., 2018), os autores analisaram os dados de crimes do departamento de polícia de Vancouver dos últimos 15 anos. Utilizando os algoritmos *K-Nearest Neighbour* e *Boosted Decision Tree*, os pesquisadores criaram um modelo que prevê a ocorrência de um crime. A pesquisa constatou que a complexidade e o tempo de treinamento dos classificadores foram ligeiramente diferentes e que a acurácia, dos métodos avaliados, ficou na faixa de 39% e 44%. Apesar da baixa precisão dos classificadores, os autores ressaltaram que o modelo pode fornecer uma estrutura preliminar para análises adicionais. Além disso, para uma melhor compreensão, o modelo fornece gráficos envolvendo análises geográficas (*hotspots*), estatísticas (*boxplot*) e tendências criminais. (BAPPEE; JÚNIOR; MATWIN, 2018) desenvolveram um modelo de aprendizado de máquina para prever crimes usando recursos geoespaciais e dados criminais da província de Nova Escócia, Canadá. Foram utilizados dois recursos espaciais: a geocodificação e a menor distância para um *hotspot*, os quais embasaram um experimento controlado para avaliar os classificadores *Logistic Regression, Support Vector Machine, Random Forest* e o método *Ensemble*. Os resultados apontaram que a inclusão dos novos recursos espaciais aumentou o desempenho, em termos de acurácia e área sob a curva ROC (*Receiver Operating Characteristic*).

O trabalho de Sivaranjani, Sivakumari e Aasha (2016) utilizou várias abordagens de clusterização para analisar os dados criminais com objetivo auxiliar as agências policiais de Tamilnadu a prever e detectar crimes com maior precisão. Os algoritmos de agrupamentos *K-Means, Agglomerative* e *Density Based Spatial Clustering with Noise* (DBSCAN) foram utilizados para agrupar atividades criminais com base em alguns casos predefinidos. Além disso, a pesquisa utilizou o classificador *K-Nearest Neighbour* (KNN) na previsão de crimes, cujos resultados podem ser visualizados de forma fácil e interativa utilizando a ferramenta *Google Map*. Em relação ao agrupamento, o *cluster* DBSCAN apresentou melhor desempenho entre métricas avaliadas, além de um melhor agrupamento em relação aos outros algoritmos.

Aryal e Wang (2018) desenvolveram e implementaram o *Spark-based Shared Nearest Neighbor* (SparkSNN), um algoritmo de *cluster* baseado em densidade eficiente, utilizando uma plataforma poderosa chamada *Spark*. O SparkSNN aprimorou o desempenho do algoritmo *Shared Nearest Neighbor* (SNN) tradicional para análise de *big data* em *clusters* de computação distribuídos. Em contraste com os algoritmos existentes, essa nova abordagem pode encontrar *clusters* de diferentes tamanhos, formas e densidades em dados em grande escala. A pesquisa avaliou minuciosamente as propriedades de desempenho e escalabilidade do algoritmo proposto, sobre o conjunto de dados criminal do estado americano de Maryland. Os experimentos demostraram que o SparkSNN aprimorou a eficiência e o desempenho do algoritmo SNN tradicional para análise de *big data* em *clusters* de computação distribuídos. Em (FARIAS et al., 2018), os autores desenvolveram modelos que são baseados em algoritmos de clusterização e na análise de técnicas de conceito formal, utilizando os dados de registros de crimes da cidade de Mossoró-RN, Brasil. Com o uso do algoritmo de agrupamento *Fuzzy K-Means*, identificaram locais com alta concentração de crimes (*hotspots*). Além disso, utilizaram Análise Formal de Conceitos (AFC), do inglês *Formal Concept Analysis*, para processar dados criminais, permitindo a extração de padrões de crimes por períodos específicos do dia e tipos de crimes. Quatro tipos de crimes básicos (furto, roubo, homicídio e tráfico de drogas) foram analisados, usando a teoria da AFC para gerar redes de acordo com a concentração de tais crimes ao longo de períodos específicos do dia. E, por fim, um *ranking* dos bairros de Mossoró foi produzido de acordo com seu nível de perigosidade .

Em (AGRAWAL; SEJWAR, 2017), os pesquisadores utilizaram dois algoritmos com o objetivo de encontrar padrões ocultos em bancos de dados criminais. Inicialmente, os padrões foram extraídos por meio da execução do algoritmo *FP-Growth*, o qual tentou constatar os padrões frequentes dos tipos de crimes em relação às cidades (localidade). Em seguida, os padrões foram otimizados utilizando o algoritmo *Multi Objective Particle Swarm Optimization* (MOPSO). Segundo os autores, os resultados indicaram que a abordagem proposta é promissora, indicando com que frequência um padrão de crime aparece no banco de dados e auxiliando os analistas de crimes a obter conhecimento sobre a ocorrência de crimes em locais específicos, em menos tempo de execução. Em (RUIZ et al., 2014), os autores propuseram uma nova técnica para fundir Regras de Associação (RA) obtidas de vários bancos de dados, por meio do que chamamos de regras de meta-associação. A principal vantagem desse método é que ele pode extrair informações que não são obtidas por regras regulares simples. De acordo com os pesquisadores, os resultados obtidos são promissores e sugerem que esse novo método pode ser aplicado em diversas áreas. Ainda sobre Regras, (CHEN et al., 2015) usaram uma técnica de descoberta de RAs para desenvolver um sistema *web* integrado, com intuito de identificar padrões e extrair informações úteis. Por meio desta aplicação, os dados poderão ser visualizados em um mapa, ajudando pesquisadores, policiais e o público a explorar o conjunto de dados e recuperar informações valiosas, auxiliando-os a tomar decisões sobre crimes e segurança.

Os pesquisadores de (MARZAN et al., 2017) tiveram como objetivo identificar áreas com maior ocorrência de crimes (*hotspots*) da cidade de Manila, Filipinas. Adicionalmente,

usaram o algoritmo *Apriori*, na descoberta de padrões frequentes, para ajudar os policiais a formar uma ação preventiva. Este trabalho também avaliou vários métodos de previsão de séries temporais, como regressão linear, processos gaussianos, multicamada *Perceptron* e *SMOreg* para prever tendências futuras do crime. Como resultado, o multicamada *Perceptron* foi capaz de prever o número de crimes na maioria dos locais em Manila, com mais precisão do que as outras técnicas. De forma semelhante, Yadav Meet Timbadia e Yadav (2017) utilizaram os algoritmos *Apriori*, *K-Means*, *Naive Bayes*, bem como análise de correlação e regressão para tentar ajudar os especialistas criminais a descobrir padrões, tendências, realizar previsões, encontrar relacionamentos e possíveis explicações, mapear redes criminosas e identificar possíveis suspeitos. De acordo com os autores, o modelo desenvolvido reduzirá os crimes e auxiliará o processo de detecção de crimes de várias maneiras. Por fim, o trabalho apresentado em (SINGH; JOSHI, 2018) utilizou as técnicas de mineração de dados supervisionadas e não supervisionadas para analisar dados de crimes. Foram implementados os algoritmos de Regressão Linear Múltipla, *Apriori* e *K-Means*, para conduzir um estudo comparativo de vários padrões sobre um conjunto de dados criminais do Estado de Gujarat, na Índia. Como resultado, as RAs sugeriram os tipos de crime associados com base nos casos registrados e o modelo de Regressão Linear Múltipla ajudou a desenvolver um modelo preditivo com precisão de 85,50%. Além disso, o *K-Means* demostrou, dentre as cidades do Estado de Gujarat, quais são as mais perigosas.

3.3 Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, uma Revisão Sistemática (RS) quantitativa da literatura (PRADO et al., 2020), tendo por finalidade encontrar o estado da arte das pesquisas sobre análise inteligente de dados abertos governamentais relacionados a incidentes criminais. Para operacionalizar a revisão, acessamos a base Scopus por meio do portal de periódicos da CAPES disponível em (CAPES, 2019), o qual permite fazer *download* dos artigos sem restrições. A Scopus foi escolhida por incluir buscas em diferentes bancos de dados científicos (IEEE, ACM e outros).

Ato contínuo, para realização do objetivo principal desta pesquisa, foram utilizados os dados abertos criminais disponibilizados pelo MJSP-Brasil. Todavia, algumas dessas informações são fornecidas em arquivos muito grandes, com dados dispersos, limitando o entendimento do cidadão com relação ao significado ou à importância dos dados abertos. Desta forma, este trabalho permitiu e permitirá fazer a transição dos dados brutos do governo para informações estruturadas, perfazendo o *download* do(s) arquivo(s), nos formatos XLSX, bem como a leitura, interpretação do conteúdo e armazenamento numa base de dados estruturada. Na sequência, foi realizado um experimento controlado, com os dados estruturados, para produção da pesquisa e coleta de dados dependentes. A seção 3.4.1 ilustra uma visão geral da arquitetura utilizada para realizar as etapas que vão desde o *download* do *dataset* até a detecção dos padrões.

De acordo com Wohlin et al. (2012), uma experimentação não é uma tarefa simples, pois envolve preparar, conduzir e analisar experimentos corretamente. Os autores destacam como uma das principais vantagens da experimentação o controle dos sujeitos, objetos e instrumentação, o que torna possível extrair conclusões mais gerais sobre o assunto investigado. Além disto, outra vantagem inclui a habilidade de realizar análises estatísticas, utilizando métodos de teste de hipóteses e oportunidades para replicação. Juristo e Moreno (2013) também afirmam que a pesquisa científica, como a deste trabalho em tela, não pode ser baseada em opiniões ou interesses comerciais. Investigações científicas são representadas por estudos baseados em observação e/ou experimentação acerca do mundo real e seus comportamentos mensuráveis.

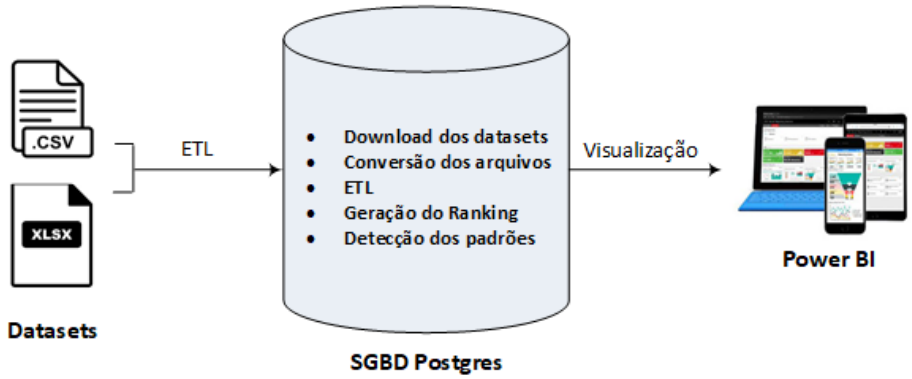
Sumarizando, o experimento teve 3 etapas macro: (1) identificação e *download* dos arquivos nos formatos XLSX; (2) construção de programas ETL (*Extract, Transform, Load*) para as cargas dos dados e da base de dados para tratamento, estruturação e armazenamento das informações contidas nos arquivos; (3) seleção, exploração, análise e validação das informações oriundas dos dados estruturados.

3.4 Base Conceitual

3.4.1 Visão Geral da Arquitetura

Com o objetivo de facilitar a obtenção, o tratamento, a manipulação e a análise dos dados criminais foi desenvolvida uma arquitetura centralizada (unificada), tendo como ponto principal o Sistema de Gerenciamento de Bancos de Dados (SGBD) *PostgreSQL*. Dessa forma, procurou-se automatizar ao máximo os processos existentes, que vão desde a obtenção dos *datasets* até a detecção dos padrões criminais. Vale ressaltar, que a arquitetura desenvolvida utilizou a ferramenta *Power BI* da *Microsoft* para a apresentação gráfica dos resultados.

Figura 20 – Visão geral da arquitetura.

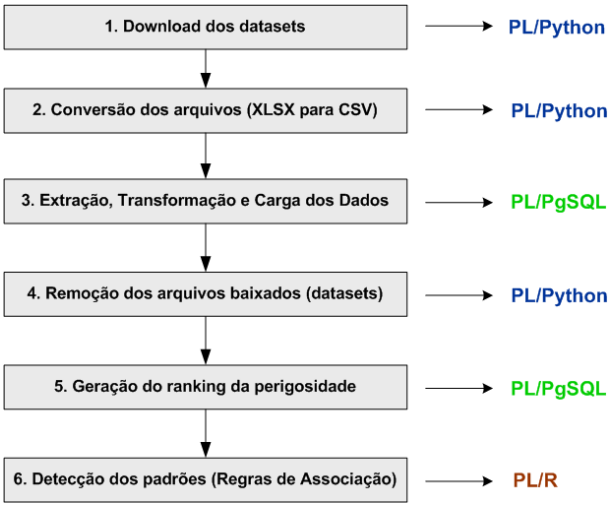


Fonte: Elaborada pelo autor.

O SGBD *PostgreSQL* permite que as funções definidas pelo usuário sejam escritas em outras linguagens além de SQL e C. Estas linguagens são chamadas genericamente de Linguagens Procedurais (LPs). No caso de uma função escrita em uma Linguagem Procedural (LP), o servidor de banco de dados não possui nenhum conhecimento interno sobre como interpretar o texto do código fonte da função. Em vez disso, a tarefa é passada para um tratador especial que conhece os detalhes da linguagem. O próprio tratador pode fazer todo o trabalho (análise gramatical e sintática, execução, etc) ou pode servir como um "elo de ligação" entre o *PostgreSQL* e a implementação existente de uma linguagem de programação (POSTGRESQL, 2019). A Figura 20 exibe uma visão geral da arquitetura desenvolvida.

Segundo (POSTGRESQL, 2019), atualmente existem quatro linguagens procedurais disponíveis na distribuição padrão *PostgreSQL*: *PL/pgSQL*, *PL/Tcl*, *PL/Perl* e *PL/Python*. Tais LPs são úteis para executar código nas linguagens *pgSQL* (linguagem nativa do *PostgresSQL*), *Tcl*, *Perl* e *Python*. Entretanto, existem várias linguagens procedurais desenvolvidas e mantidas fora da distribuição principal, como, por exemplo: *PL/Java*, *PL/Sh* e *PL/R*.

Figura 21 – Linguagem procedural utilizada em cada fase.



Fonte: Elaborada pelo autor.

A Figura 21 detalha a sequência de etapas desenvolvidas dentro da arquitetura, as quais vão desde o *download* dos *datasets* até a mineração dos dados, especificando qual linguagem procedural foi utilizada em cada fase desse processo. Em nosso trabalho, três LPs foram utilizadas, são elas:

- ***PL/pgSQL*** **-** É uma linguagem procedural, nativa, desenvolvida para ser usada dentro sistema de banco de dados. Esta LP foi utilizada, principalmente, nos procedimentos de Extração, Transformação e Carga (ETL) dos dados, geração do *ranking* de perigosidade, entre outros processos.

- ***PL/Python –*** É uma linguagem procedural que permite a implementação de funções utilizando a linguagem *Python*, as quais poderão ser executadas dentro do *PostreSQL*. Os códigos escritos em *PL/Python* foram usados para realizar os *downloads* dos *datasets*, conversão, remoção e descompactação dos arquivos.

- ***PL/R –*** Útil para realizar mineração dos dados, detectando os padrões. É uma LP que permite implementação de funções utilizando a linguagem R, oferecendo a maioria (se não todos) dos recursos que um escritor de funções possui na linguagem R.

3.4.2 Detecção de *Outliers*

Os *outliers*, também chamado de valores atípicos, aberrantes ou discrepantes, são valores que se distanciam muito dos demais, fugindo do padrão. Quando encontrados em uma amostra, podem representar um novo evento a ser investigado (TAN; STEINBACH; KUMAR, 2016). De acordo com (CONFORTI; ROSA; HOFSTEDE, 2015), *outliers* são baseados em modelos estatísticos, dos quais se espera que os dados estejam em uma distribuição normal e as observações improváveis, com base na média e desvio padrão, são consideradas como valores anormais.

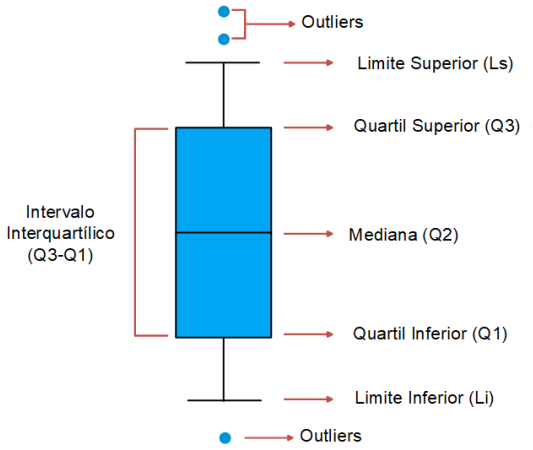
BoxPlot, no português "gráfico de caixas", é, provavelmente, a técnica estatística mais simples para se detectar *outliers* e representá-los graficamente. Autores e leitores podem usá-lo para resumir e interpretar rapidamente os dados tabulares. O *boxplot* é um recurso gráfico aperfeiçoado que cumpre com a análise exploratória e até mesmo inferencial dos dados, podendo substituir o uso de tabelas em casos específicos. Esta representação gráfica faz parte de uma família diversificada de técnicas estatísticas usada para identificar visualmente padrões, os quais poderiam estar ocultos em um conjunto de dados. Além disso, este tipo de gráfico é um recurso visual que resume os dados para exibir a mediana, quartis e os valores pontuais máximos e mínimos. Portanto, apresenta valores de tendência central, dispersão e simetria dos dados agrupados. Dentre suas aplicações incluem análise exploratória dos dados, detecção de *outliers* e comparação entre grupos (equivalência) (NETO et al., 2017).

Segundo (NETO et al., 2017), ao analisar este tipo de gráfico (ver Figura 22), verifica-se os seguintes itens:

- **Primeiro Quartil (Q1) ou Quartil Inferior**: Representa 25% dos menores valores do conjunto de dados avaliado. Representado pela linha limite inferior da caixa;

- **Segundo Quartil ou Mediana (Q2)**: É o centro da distribuição dos dados, representada pela linha dentro da caixa. O Q2 indica o 50º percentil. Uma distribuição simétrica teria a mediana no centro do retângulo. Se a mediana é próxima de Q1, então, os dados são positivamente assimétricos. Se a mediana é próxima de Q3, os dados são negativamente assimétricos;

Figura 22 – Gráfico *BoxPlot*.



Fonte: Elaborada pelo autor.

- **Terceiro Quartil (Q3) ou Quartil Superior**: Representa 75% da amostra. Indicado pela linha limite superior da caixa;
- **A dispersão dos dados**: Representada pelo intervalo interquartílico (IIQ), ou seja, o tamanho da caixa, que é a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1). O IIQ representa o intervalo de 50% dos dados. Embora a amplitude seja de fácil entendimento, o intervalo interquartílico é uma estatística mais robusta para medir variabilidade, uma vez que não sofre influência de *outliers*;
- **Limites inferior (Li) e superior (Ls)**: São as linhas que vão do retângulo até o maior (Ls) ou menor (Li) valor considerado não discrepante. Laurikkala et al. (2000) sugerem a utilização de uma heurística (1,5 x IIQ). Assim, o limite inferior é calculado por Q1- (1,5 x IIQ) e o limite superior é dado por Q3 + (1,5 x IIQ). As anomalias são representadas pelos pontos encontrados além dos valores Li e Ls. O valor 1,5, utilizado para o cálculo do intervalo de confiança (Li e Ls), é utilizado para representar 95% dos dados e os valores razoáveis extremos. Utilizando o valor 3,0, conseguimos uma cobertura de 99% dos dados;
- ***Outliers***: Indicam possíveis valores aberrantes ou atípicos. No *boxplot*, as observações são consideradas discrepantes quando estão abaixo de Li ou acima de Ls.

3.4.3 Regras de Associação

A mineração de regras de associação (RA) é uma abordagem que trabalha na descoberta de relacionamentos interessantes ocultos em grandes conjuntos de dados. Conhecida como Análise de Cesta de Mercado (*Market Basket Analysis*), as Regras de Associação (RA) são técnicas de mineração de dados que representam combinações de itens que ocorrem com determinada frequência em uma base de dados. Através de um estudo, o qual procurava encontrar

relacionamentos entre os itens nas compras dos clientes numa visita ao supermercado, Agrawal, Imieliński e Swami (1993) descobriram coocorrências entre conjuntos de itens a partir de um conjunto de dados estruturado como transações, nas quais cada transação contém um subconjunto de itens. RA é uma técnica de aprendizado não supervisionado usada para extrapolar as regras associadas entre si (SINGH; JOSHI, 2018).

A aplicabilidade das regras de associação perpassa por diversas áreas distintas como análise epidemiológica, genética, recomendação de filmes, previsão de doenças cardíacas, ontologia, *Big Data*, computação ubíqua e Internet das Coisas. Segundo (MARZAN et al., 2017), uma das abordagens de mineração de dados mais conhecidas para encontrar conjuntos de itens frequentes e gerar regras de associação é o algoritmo *Apriori*. Isso abriu o caminho para os pesquisadores resolverem problemas sobre como encontrar padrões ocultos em análises criminais.

De maneira formal, Agrawal, Imieliński e Swami (1993) definiram as regras de associação da seguinte forma: Sejam $I = i_1, i_2, ..., i_m$ um conjunto de m itens distintos e D uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens (*itemset*), tal que $T \subseteq I$. Uma regra de associação é uma expressão na forma $A \Rightarrow B$, onde $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ e $A \cap B = \emptyset$. A é denominado antecedente e B denominado consequente da regra. Tanto o antecedente quanto o consequente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens. A quantidade de itens pertencentes a um conjunto de itens é chamada de comprimento do conjunto. Um conjunto de itens de comprimento k costuma ser referenciado como um *k –itemset* (GONSALVES, 2004) (CAMPOS, 2018). Como exemplo, A e B podem ser produtos ou eventos, então, poderíamos ter a regra, "quem compra o produto A também compra o produto B", ou a regra, "quando ocorre um roubo (A) também ocorre um assassinato (B)".

Existem várias medidas de interesse que avaliam regras de acordo com as restrições iniciais do usuário. Diversos pesquisadores propõem medidas que objetivam extrair um padrão específico dos dados (CAMPOS, 2018). Em nossa pesquisa, utilizamos as seguintes medidas de interesse:

- **Suporte:** O suporte de um conjunto de itens Z, *Sup(Z)*, representa o percentual de transações da base de dados que contêm os itens do conjunto Z. Caso o limite mínimo de suporte (suporte mínimo) escolhido seja alto, alguns padrões interessantes podem ser ignorados e, consequentemente, comprometer a tomada de decisão efetiva. Em contrapartida, se o valor mínimo de limite escolhido for baixo, um grande número de regras será gerado, tornando a tarefa complexa e demorada. O suporte de uma regra de associação $A \Rightarrow B$, *Sup* $(A \Rightarrow B)$, é dado por *Sup* $(A \cup B)$, e pode ser visto na Equação 3.1.

$$Sup(A \Rightarrow B) = \frac{Transações\ \ que\ \ contêm\ \ A\ \ e\ \ B}{Total\ \ \ de\ \ Transações} \tag{3.1}$$

- **Confiança:** A confiança da regra $A \Rightarrow B$, *Conf* $(A \Rightarrow B)$, representa, dentre as transações que contêm A, a porcentagem de transações que também contêm B, é dada por *Conf* $(A \Rightarrow B) =$ *Sup* $(A \cup B) \div$ *Sup* (A), e pode ser vista na Equação 3.2.

$$Conf(A \Rightarrow B) = \frac{Número\ total\ de\ transações\ que\ contêm\ A\ e\ B}{Total\ de\ transações\ que\ contêm\ A} \tag{3.2}$$

- **Lift:** A medida de interesse *Lift*, também conhecida como *Interest*, é utilizada para avaliar as dependências entre o conjunto de itens do antecedente e o conjunto de itens do consequente de uma RA. O valor do *Lift* de uma regra de associação $A \Rightarrow B$, obtida a partir de uma base de dados de transações, indica o quanto mais frequente torna-se B quando ocorre em conjunto com A. O *Lift* de uma regra de associação $A \Rightarrow B$ é dado pela Equação 3.3. Quando *Lift*$(A \Rightarrow B) = 1$, significa que A e B são independentes, ou seja, não existe associação entre eles. Se *Lift*$(A \Rightarrow B) > 1$, então A e B são positivamente dependentes. Se *Lift*$(A \Rightarrow B) < 1$, então A e B são negativamente dependentes.

$$Lift(A \Rightarrow B) = \frac{Conf(A \Rightarrow B)}{Sup(B)} \tag{3.3}$$

- **Count:** Representa a frequência da ocorrência de um determinado conjunto de itens.

- **Coeficiente de correlação de *Pearson (r)*:** É uma medida de associação bivariada (força) do grau de relacionamento entre duas variáveis. O coeficiente de correlação *Pearson (r)* varia de -1 a 1. O sinal indica a direção da correlação (negativa ou positiva), enquanto que o valor indica a magnitude. Uma correlação positiva ($r=1$) indica que quando X aumenta, Y também aumenta, ou seja, valores altos de X estão associados a valores altos de Y. Por outro lado, uma correlação negativa ($r=-1$) indica que quando X aumenta, Y diminui, ou seja, valores altos de X estão associados a valores baixos de Y. Em particular, uma correlação com $r=0$ significa que as variáveis são ortogonais entre si (ausência de correlação) (PARANHOS et al., 2014).

- **Qui-quadrado (χ^2):** É um teste não paramétrico utilizado para avaliar a correlação entre vários itens. O princípio básico deste método é comparar proporções, isto é, as possíveis divergências entre as frequências observadas e esperadas para um certo evento. Quanto mais próximas as frequências observadas estiverem das frequências esperadas, maior o peso da evidência em favor da independência. Nas regras de associação, é utilizado para testar a independência entre os itens das regras (CAMPOS, 2018). O valor crítico da distribuição qui-quadrado com 1 grau de liberdade (tabela de contingência 2x2) em α = 0,05 é 3,84; quanto maior o valor do qui-quadrado, mais provável é a correlação das variáveis. O nível de significância é indicado pelo valor do *p-value*, fornecendo uma medida da força dos resultados de um teste, em contraste a uma simples rejeição ou não rejeição (WU et al., 2016).

3.4.4 Definição do *Ranking* de Perigosidade

Um grande problema encontrado na análise criminal é quantificar quanto um determinado local é mais perigoso que outro. Isso torna-se ainda mais problemático quando envolvemos diversos tipos de crimes dentro da mesma análise. Diariamente, diferentes regiões são alvos de diversos tipos de crimes e em várias proporções. Então, como é possível determinar que uma determinada região "*A*" é mais perigosa que uma região "*B*"?.

Neste contexto, (FARIAS et al., 2018) criaram um *ranking* de perigosidade para os bairros de uma cidade brasileira chamada Mossoró, localizada no estado do Rio Grande do Norte. Com o uso dos dados criminais combinado com uma fórmula pré-definida, eles definiram um grau de perigosidade para cada bairro da cidade. Para a definição dessa fórmula, os autores ponderaram cada tipo de crime, a fim de permitir uma comparação entre os diferentes bairros. Esta ponderação foi definida usando o bom senso e a opinião de especialistas. Por exemplo, é senso comum que um homicídio é um crime mais grave que um roubo. A Tabela 7 mostra os pesos que foram definidos para cada crime.

Tabela 7 – Pesos dos crimes.

Tipo de crime	**Peso**
Furto	1,0
Roubo	2,0
Homicídio	3,0
Tráfico de drogas	4,0

Fonte: (FARIAS et al., 2018).

Baseada nos pesos acima, o Índice de Criminalidade (*IC*) para cada bairro foi calculado por meio de uma média ponderada, ou seja, o somatório do produto entre o peso e o número de incidentes de cada tipo de crime, divido pela soma dos pesos. A Equação 3.4 mostra a fórmula utilizada, onde n representa o número de tipos de crime, P_i o valor do peso e N_i o total de ocorrências daquele tipo de crime.

$$IC = \frac{\sum_{i=1}^{n} P_i \times N_i}{\sum_{i=1}^{n} P_i} \tag{3.4}$$

Após o computo do *IC*, o Índice de Criminalidade Normalizado (*ICN*) foi obtido, utilizando o processo de normalização no intervalo [0,1] (LIMA; VIGNATTI; SILVA, 2020). Então, os bairros foram classificados do menos para o mais perigoso, onde *ICN=0* significa o menos perigoso e *ICN=1* representa o mais perigoso. Porém, a fim de permitir uma classificação mais interpretável e clara, os autores transformaram o *ICN*, que são representados por valores contínuos, em níveis de perigosidade representados por valores categóricos. A definição desses níveis foi realizada da seguinte forma:

- Não Perigoso: *0,00 <= ICN < 0,15*;

- Pouco Perigoso: *0,15 <= ICN < 0,30*;
- Perigoso: *0,30 <= ICN < 0,50*;
- Muito Perigoso: *0,50 <= ICN <= 1,00*.

Em nosso trabalho, foi utilizada a mesma estratégia adotada por (FARIAS et al., 2018) para encontrar os níveis de perigosidade. Entretanto, a fim de obter resultados mais precisos, alguns ajustes pontuais no processo foram realizados. Primeiramente, percebe-se que a Equação 3.4 utiliza o número de incidentes criminais. De certa forma, isso favorece regiões (bairros) menos populosas, pois locais com maior concentração populacional podem ter mais crimes. Logo, locais mais populosos poderiam alcançar os maiores *IC* por este fator e não necessariamente pelos níveis de segurança e criminalidade. Para mitigar essa ameaça, foi considerada a estimativa populacional anual no cálculo do *IC*. Ao invés de utilizar diretamente o número de ocorrências criminais, foi utilizada a taxa de crimes por 100.000 habitantes. Este procedimento atenuará as discrepâncias existentes, trazendo todas as regiões analisadas para o mesmo patamar, ao diminuir o potencial de confundimento com a população (fator de confusão). A Equação 3.5 representa o cálculo para encontrar a taxa de crimes por 100.000 habitantes (*TC_HAB*). A estimativa populacional anual de cada estado foi obtida no portal do Instituto Brasileiro de Geografia e Estatística (IBGE), localizado no site https://ibge.gov.br.

$$TC\_HAB = \frac{Número\ \ de\ \ Crimes}{Estimativa\ Populacional} \times 100.000 \tag{3.5}$$

A Equação 3.6 mostra a Equação 3.4 adaptada, utilizando a taxa de crimes por 100.000 habitantes em seu computo.

$$IC_{adaptado} = \frac{\sum_{i=1}^{n} P_i \times TC\_HAB_i}{\sum_{i=1}^{n} P_i} \tag{3.6}$$

Um outro ponto ajustado foram as faixas dos níveis de criminalidade. Diferente do trabalho de (FARIAS et al., 2018), foram adotadas cinco faixas de perigosidade (Baixíssimo, Baixo, Intermediário, Alto ou Altíssimo). Neste contexto, é importante ressaltar que o *ICN* sempre indicará o nível de perigosidade relativo ao Brasil e ao ano analisado, ou seja, é um *ranking* de perigosidade do Brasil, pois, ainda que os estados obtivessem níveis de criminalidade desejáveis e compatíveis com os lugares mais seguros do mundo, sempre haverá um estado ocupando a posição Altíssimo, representando a pior posição em relação ao resto do país. Então, os níveis adotados por este artigo foram:

- Baixíssimo: *0,00 <= ICN <= 0,15*;
- Baixo: *0,15 < ICN <= 0,40*;

- Intermediário: *0,40 < ICN <= 0,60*;
- Alto: *0,60 < ICN <= 0,85*;
- Altíssimo: *0,85 < ICN <= 1,00*.

Com intuito de otimizar o *ranking* aqui desenvolvido, vale ressaltar que quatro especialistas em segurança pública foram consultados, para ponderar os novos tipos de crime e calibrar os pesos existentes descritos na Tabela 7. Os especialistas trouxeram as visões da Polícia, do Ministério Público, do Judiciário e dos advogados, com as perspectivas dos crimes de um Coordenador de um Grupo de Combate ao Crime Organizado, de um Promotor, de um Juiz e de um Advogado Criminalista. O maior ponto de discussão residiu nos crimes como o de Lesão Corporal Seguida de Morte, pois apesar de ter uma pena menor, lida com a subtração do maior bem do ser humano, assim como o homicídio. Os desempates do debate consideraram sempre a pena e suas variações. Neste caso específico, observou-se o fato de o julgamento já ter considerado o caráter preterdoloso deste crime. Vale ressaltar que este é um debate amplo e profundo, o qual não tem e talvez nunca terá uma conclusão simples, uma vez que as penas nem sempre refletirão a costumaz gravidade de um crime e as opiniões divergem veementemente quando os crimes lidam com vítimas fatais.

Em resumo, foram consideradas as penas dos crimes, uniformidade de distâncias entre os pesos (considerando uma aproximação proporcional às distâncias entre as penas), clamor social, bem como o empirismo da cultura e eficácia das leis brasileiras. Infelizmente, como a tipificação fornecida pelo governo não detalha totalmente os crimes pela gravidade, as distâncias dos pesos tiveram que ser aproximadas e se basear nas médias das penas e suas variações. Após essas validações, os novos tipos de crime e os novos pesos adotados estão descritos na Tabela 8.

Tabela 8 – Pesos dos crimes utilizados neste trabalho.

Tipo de crime	**Peso**
Furto de veículo	1,0
Roubo de veículo	2,0
Roubo de carga	2,0
Tentativa de homicídio	2,0
Lesão corporal seguida de morte	2,5
Roubo a instituição financeira	2,5
Estupro	3,0
Homicídio doloso	3,5
Roubo seguido de morte (latrocínio)	5,0

Fonte: Elaborada pelo autor.

3.5 Definição e Planejamento do Experimento

Nesta e nas duas próximas seções, este trabalho é apresentado como um processo experimental. O mesmo segue as diretrizes de Wohlin (WOHLIN et al., 2012) e processos

experimentais com publicações recentes (SANTOS; COLAÇO JÚNIOR; SOUZA, 2018) (OLIVEIRA; COLAÇO JÚNIOR, 2018). Esta seção focará na definição do objetivo e planejamento do experimento.

3.5.1 Definição dos Objetivos

O objetivo principal deste experimento é analisar as possíveis associações entre os estados do Brasil e os tipos de crimes, utilizando os dados disponibilizados pelo Governo Federal, através do Ministério da Justiça e Segurança Pública (MJSP). Para atingi-lo, conduziu-se um experimento, *in vitro*, em ambiente controlado, no qual foram verificadas regras de associação, utilizando o algoritmo *Apriori*, para determinação das combinações de itens que ocorrem com determinada frequência, bem como para a geração das medidas de interesse que, estatisticamente, puderam servir de base para medir a força de tais regras. Adicionalmente, *rankings* de perigosidade dos estados foram desenvolvidos, bem como uma análise de *outliers* entre as taxas de criminalidade (proporcional à população) das regiões brasileiras foi realizada.

Baseado no modelo GQM (*Goal Question Metric*) apresentado em (BASILI et al., 2014)(BASILI; WEISS, 1984), segue a formalização do objetivo desse trabalho: **Analisar** as ocorrências criminais nos estados brasileiros, **com o propósito** de avaliá-las, **com respeito às** detecção de padrões criminais e anomalias, **do ponto de vista de** cientistas de dados, analistas criminais e cidadãos, **no contexto** de dados abertos do MJSP-Brasil.

3.5.2 Planejamento

Formulação das Hipóteses - Não foram encontrados estudos experimentais que analisaram as possíveis associações entre os estados brasileiros e os tipos de crime. Além disso, também não foram encontradas pesquisas científicas que descrevessem os níveis de perigosidade dos estados ou que realizaram alguma análise de discrepância entre as taxas de criminalidade entre as Regiões Brasileiras. Baseadas nestas premissas, três questões de pesquisa foram formalizadas para esse trabalho. Para as duas primeiras questões, **Q1** e **Q2**, serão realizadas análises estatísticas descritivas, enquanto que, para a questão **Q3**, será aplicado um teste de significância da hipótese. As questões de pesquisa são:

- **Q1**: Existem discrepâncias entre as taxas de criminalidade (taxa por cem mil habitantes) das Regiões Brasileiras?
- **Q2**: Quais os estados vêm se destacando como os mais perigosos?
- **Q3**: Há associações entre tipos de crime e os estados?

Para responder à questão **Q3**, analisaremos as regras de associação geradas para: Estado $\Rightarrow$ Tipo de crime. Portanto, a seguinte hipótese será testada:

- **Hipótese 1 (Q3)**
  - H_0: Os tipos de crime são independentes dos estados.
  - H_1: Os tipos de crime são dependentes dos estados.

Seleção do Contexto - Para a realização do experimento, foram utilizados os dados criminais dos 26 estados e do Distrito Federal, totalizando 3.443.750 incidentes.

Seleção dos Participantes e Objetos - Segundo Travassos e Barros (2003), os participantes são os indivíduos selecionados da população com a finalidade de conduzir o experimento. Logo, o conjunto de participantes deve ser representativo o bastante para que se possam generalizar os resultados do experimento a uma população desejada. Eles são os responsáveis por informar parâmetros para o experimento, tal como o valor das variáveis.

Foram selecionados todos dados de ocorrências criminais estaduais disponibilizados até o início dessa pesquisa. Tais dados englobam as informações do período de Janeiro de 2015 a Dezembro de 2019, os quais foram obtidos no portal de dados abertos do Ministério da Justiça e Segurança Pública, localizado em (MJSP, 2020).

Variáveis Dependentes - As variáveis dependentes abordadas no experimento, para validação da hipótese, foram as frequências dos conjuntos de itens analisados e as regras geradas, com seus Suportes e Confianças, das quais podem ser derivadas outras medidas de interesse objetivas para auxiliar na identificação das forças destas regras de associação: *Lift*, *r* (Coeficiente de correlação de *Pearson*) e *Qui Quadrado* (χ^2), com seu nível de significância (*p-value*).

Variáveis Independentes - As variáveis independentes referem-se à entrada do processo de experimentação, ou seja, representa a causa que afeta o resultado do experimento (TRAVASSOS; BARROS, 2003). Para este trabalho, foram consideradas, como variáveis independentes, o conjunto de registros compilados e disponibilizados em arquivos, contendo os dados dos incidentes criminais ocorridos nos estados brasileiros, o algoritmo *Apriori* utilizado, bem como Suporte e Confiança mínimos, intervalos aceitáveis de *Lift* e *p-value* máximo.

Instrumentação - O processo de instrumentação teve início com a configuração do ambiente para o experimento, planejamento de coleta de dados, construção de ETL (programa de Extração, Transformação e Carga) e o desenvolvimento e execução dos algoritmos necessários. Os materiais/recursos utilizados foram:

- Arquivo com os incidentes criminais, disponibilizado pelo MJSP;
- Arquivos com as estimativas populacionais, disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE);
- Arquivos com o *script* de criação do projeto de banco de dados;

- Banco de Dados *PostgreSQL*, versão 11.6-3, para armazenar os dados e realizar o processo de ETL, com o uso da linguagem nativa *PL/pgSQL*;
- *Python*, versão 3.7.6;
- Software livre *R*, versão 3.6.0;
- Ferramenta *Power BI*.

3.6 Operação do Experimento

3.6.1 Preparação

Para a realização deste trabalho, foi preparado o ambiente adequado para armazenar os dados utilizados no experimento. Como descrito, para facilitar esse processo, foi desenvolvida e utilizada a arquitetura descrita na seção 3.4.1. Inicialmente, foi realizada a instalação do *PostgreSQL*, dos ambientes de execução *Python* e *R*, e, por fim, da ferramenta de análise de dados *Power BI*. Posteriormente, configurou-se para que o banco de dados pudesse executar código *Python* e *R*, por meio das suas respectivas LPs. Em seguida, foram executados os *scripts* de criação das tabelas e os demais objetos do banco de dados, a fim de deixar o ambiente pronto para receber os dados. Por fim, realizou-se o *download* e a carga das estimativas populacionais anuais, fornecidas pelo IBGE, disponibilizadas no site (IBGE, 2020).

3.6.2 Execução

Após a preparação do ambiente e posse de todos os materiais/recursos descritos anteriormente, e com o intuito de tornar o processo mais automatizado possível, facilitando dessa forma sua manipulação e manutenção, foram desenvolvidos procedimentos que vão desde o *download* dos *datasets* até a mineração dos dados.

Primeiramente, um código escrito em *PL/Python* foi utilizado para realizar o *download* dos *datasets* criminais e realizar a conversão dos arquivos Excel (.XLSX) em arquivos .CSV, já que este último é o formato lido (aceito) pelo *PostgreSQL*.

Na etapa seguinte, utilizou-se *PL/pgSQL* para realizar o processo de Extração, Transformação e Carga das informações. Nesta fase, os dados foram tratados, eliminando as possíveis inconsistências, e, posteriormente, armazenados em tabelas no banco de dados. Vale ressaltar que os dados disponibilizados pelo Governo Federal estão sumarizados, ou seja, cada registro informa a quantidade de incidentes criminais ocorridos em um determinado mês, estado e tipo de crime. Dessa forma, foi necessário realizar o desmembramento dessas em informações, para que cada registro represente apenas um crime, ou seja, passando a configurar uma transação, a qual será avaliada pelo algoritmo de mineração. Então, o *dataset* que, inicialmente, possuía 14.489 registros, após a etapa de desmembramento, passou a ter 3.443.750 transações. Posteriormente, ainda

fazendo o uso da linguagem nativa *PL/pgSQL*, foram construídos os *rankings* de perigosidade dos estados brasileiros.

Em sequência, foi implementada uma rotina, utilizando a Linguagem Procedural *PL/R*, para a detecção das associações, por meio do algoritmo *Apriori*. Todas essas etapas estão conectadas sequencialmente, podendo ser atualizadas, automaticamente, por meio de um procedimento agendado no banco de dados.

3.6.2.1 Coleta dos Dados

Após a execução, com o intuito de facilitar a coleta e análise dos dados, as informações foram publicadas em uma ferramenta gráfica (*Power BI*). Utilizando os diversos recursos neste *framework*, gráficos e tabelas foram construídos a fim de elucidar as questões de pesquisa proposta por este trabalho.

Com relação às RAs, para cada associação, foi gerado um conjunto de medidas de interesse. Por se tratar de mineração em uma base de dados real, o número de regras geradas foi relativamente alto, como corroborado pelos ensaios apresentados em (ZHENG; KOHAVI; MASON, 2001). Além disso, grande parte destes resultados minerados costuma ser composta por regras óbvias, redundantes ou, até mesmo, contraditórias. Para filtrar as regras interessantes para o estudo, o suporte mínimo foi estabelecido em 0,14% (0,0014) e a confiança mínima em 60% (0,60). Como visto na Equação 3.1, o valor do suporte de uma RA é dado pela relação entre a quantidade de transações, nas quais aparecem os itens A e B (Estado e Tipo de Crime), e o total de transações. Para o cálculo do suporte mínimo, adotamos o valor aproximado (para cima) de uma amostra com população infinita, considerando uma margem de erro de 3,5% e 95% de confiabilidade, o que totaliza 784 ocorrências. Logo, a quantidade mínima de transações, nas quais aparecem os itens A e B, deve ser de 800 (arredondamento para cima), para o ano em que houve menos transações (2019). Desta forma, o suporte mínimo foi obtido por meio da divisão, 800/569.194, resultando, aproximadamente, no valor 0,0014 (0,14%). Nos outros anos, com mais transações, a margem de erro será ainda menor, uma vez que o percentual aplicado foi o mesmo. Além disso, o valor do *Lift* deverá estar nas seguintes faixas: $Lift > 1$ (significando que existe uma dependência positiva) ou $0 < Lift < 0{,}1$ (indicando que existe uma dependência negativa).

Em adição, para cada regra de associação, foram gerados, também, os valores para o coeficiente de correlação de *Pearson* (r), o qui quadrado (χ^2) e o seu nível de significância (*p-value*).

3.6.2.2 Validação dos Dados

Para assegurar a análise, interpretação e validação dos resultados sobre as regras de associação, foi aplicado o teste de significância da hipótese, utilizando o teste do qui quadrado (χ^2). Além disso, também foi calculado o coeficiente de correlação de *Pearson (r)*.

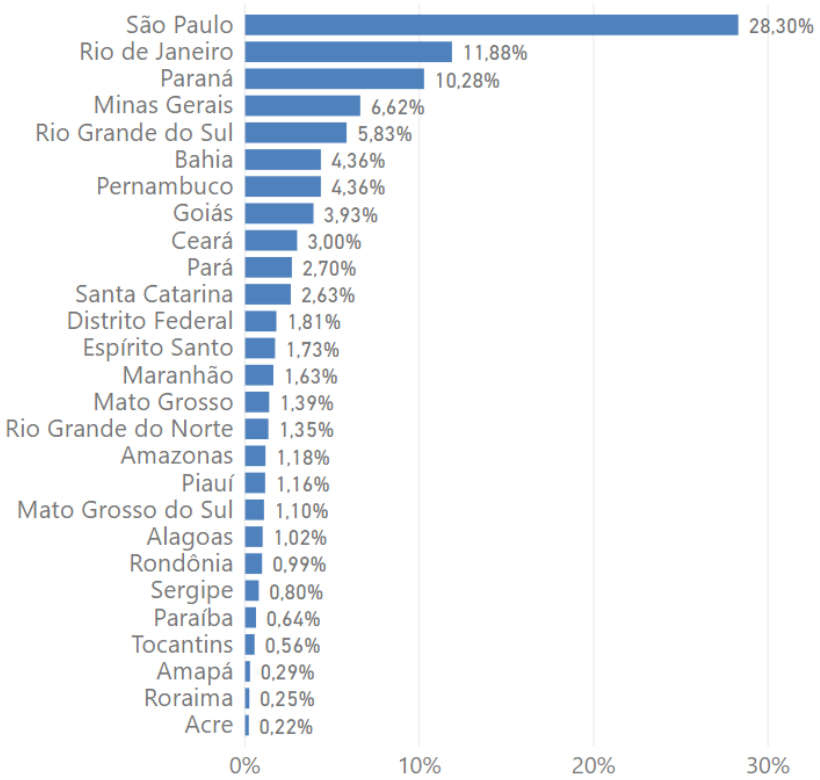
3.7 Análise dos Resultados

Nesta seção, será feita uma análise e interpretação dos resultados obtidos, a fim de responder às questões de pesquisa apresentadas anteriormente. Em seguida, serão apresentadas as ameaças à validade encontradas neste trabalho.

3.7.1 Resultados Brutos

Como exposto, o *dataset* utilizado nesta avaliação contempla os incidentes criminais ocorridos em todas as UFs do Brasil. Inicialmente, o gráfico da Figura 23 apresenta a porcentagem do número de crimes que aconteceram desde Janeiro de 2015 até Dezembro de 2019. Nota-se, na parte superior, que o estado de São Paulo supera, largamente, as demais UFs, contendo 974.465 incidentes criminais (28,30%). Em segundo lugar, está o Rio de Janeiro, com 408.987 casos (11,88%), seguido pelo Paraná, com 353.890 (10,28%). Em contrapartida, na outra extremidade, as UFs do Amapá, com 10.137 (0,29%), Roraima, com 8.748 (0,25%), e Acre, com 7.691 (0,22%), foram as regiões menos afetadas criminalmente.

Figura 23 – Porcentagem de ocorrências criminais por estado (2015-2019).



Fonte: Elaborada pelo autor.

A Tabela 9 apresenta, detalhadamente, o número e porcentagem anuais das ocorrências criminais em cada um dos estados. Como no caso anterior, Figura 23, observa-se que os estados

Tabela 9 – Número de ocorrências e porcentagem anual em cada estado brasileiro.

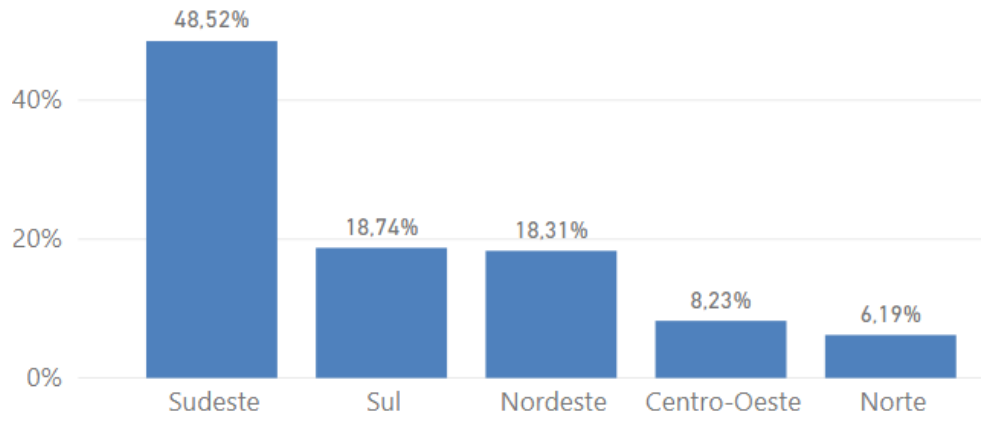
	2015	**2016**	**2017**	**2018**	**2019**	
Estado	**Total (%)**	**Total (%)**	**Total (%)**	**Total (%)**	**Total (%)**	**Variação (2018-2019)**
São Paulo	215.832 **(31,05%)**	217.018 **(28,74%)**	202.053 **(27,01%)**	183.532 **(27,13%)**	156.030 **(27,41%)**	-14,98 %
Rio de Janeiro	67.759 **(9,75%)**	81.590 **(10,81%)**	94.046 **(12,57%)**	90.592 **(13,39%)**	75.000 **(13,18%)**	-17,21 %
Paraná	65.628 **(9,44%)**	79.742 **(10,56%)**	78.026 **(10,43%)**	69.046 **(10,21%)**	61448 **(10,8%)**	-11,00 %
Minas Gerais	50871 **(7,32%)**	54.498 **(7,22%)**	49.319 **(6,59%)**	39.617 **(5,86%)**	33.539 **(5,89%)**	-15,34 %
Pernambuco	23.311 **(3,35%)**	30.189 **(4,00%)**	37.195 **(4,97%)**	30.544 **(4,52%)**	28.845 **(5,07%)**	-5,56 %
Rio Grande do Sul	46.957 **(6,76%)**	45.582 **(6,04%)**	43.159 **(5,77%)**	36.803 **(5,44%)**	28.336 **(4,98%)**	-23,01 %
Bahia	31.573 **(4,54%)**	33.213 **(4,40%)**	31.480 **(4,21%)**	29.478 **(4,36%)**	24.359 **(4,28%)**	-17,37 %
Goiás	27.984 **(4,03%)**	33.658 **(4,46%)**	29627 **(3,96%)**	26.649 **(3,94%)**	17.587 **(3,09%)**	-34,01 %
Pará	16.285 **(2,34%)**	19.108 **(2,53%)**	22.304 **(2,98%)**	19.804 **(2,93%)**	15.324 **(2,69%)**	-22,62 %
Ceará	21.486 **(3,09%)**	21.529 **(2,85%)**	24.268 **(3,24%)**	20.955 **(3,10%)**	14.971 **(2,63%)**	-28,56 %
Santa Catarina	20.982 **(3,02%)**	21.512 **(2,85%)**	18.989 **(2,54%)**	14.957 **(2,21%)**	14.231 **(2,50%)**	-4,85 %
Distrito Federal	13312 **(1,92%)**	14791 **(1,96%)**	12808 **(1,71%)**	10862 **(1,61%)**	10449 **(1,84%)**	-3,80 %
Espírito Santo	10.557 **(1,52%)**	10.843 **(1,44%)**	15.368 **(2,05%)**	12.712 **(1,88%)**	10.163 **(1,79%)**	-20,05 %
Maranhão	10.844 **(1,56%)**	12.659 **(1,68%)**	11.959 **(1,60%)**	11.095 **(1,64%)**	9.745 **(1,71%)**	-12,17 %
Piauí	6.364 **(0,92%)**	7.831 **(1,04%)**	7.761 **(1,04%)**	8.975 **(1,33%)**	9.106 **(1,60%)**	1,46 %
Mato Grosso	11.405 **(1,64%)**	10.830 **(1,43%)**	9.503 **(1,27%)**	8.243 **(1,22%)**	7.992 **(1,40%)**	-3,05 %
Rio Grande do Norte	7.721 **(1,11%)**	11.040 **(1,46%)**	10.767 **(1,44%)**	10.017 **(1,48%)**	6.855 **(1,20%)**	-31,57 %
Mato Grosso do Sul	7.839 **(1,13%)**	7.905 **(1,05%)**	7.740 **(1,03%)**	7.489 **(1,11%)**	6775 **(1,19%)**	-9,53 %
Amazonas	7.240 **(1,04%)**	8.787 **(1,16%)**	10.292 **(1,38%)**	7.783 **(1,15%)**	6.632 **(1,17%)**	-14,79 %
Rondônia	7.234 **(1,04%)**	7.716 **(1,02%)**	6.415 **(0,86%)**	6.420 **(0,95%)**	6.191 **(1,09%)**	-3,57 %
Alagoas	6.939 **(1%)**	7.802 **(1,03%)**	7.428 **(0,99%)**	6.997 **(1,03%)**	5.947 **(1,04%)**	-15,01 %
Sergipe	4.606 **(0,66%)**	6.176 **(0,82%)**	5.990 **(0,80%)**	5.540 **(0,82%)**	5.102 **(0,90%)**	-7,91 %
Paraíba	5.213 **(0,75%)**	2.666 **(0,35%)**	2.335 **(0,31%)**	7.152 **(1,06%)**	4.634 **(0,81%)**	-35,21 %
Tocantins	3.291 **(0,47%)**	3.706 **(0,49%)**	3.639 **(0,49%)**	4.720 **(0,70%)**	3.836 **(0,67%)**	-18,73 %
Acre	260 **(0,04%)**	487 **(0,06%)**	1.796 **(0,24%)**	2.550 **(0,38%)**	2.598 **(0,46%)**	1,88 %
Amapá	2.125 **(0,31%)**	2.042 **(0,27%)**	1.996 **(0,27%)**	2.025 **(0,30%)**	1949 **(0,34%)**	-3,75 %
Roraima	1.457 **(0,21%)**	2.064 **(0,27%)**	1.782 **(0,24%)**	1.895 **(0,28%)**	1.550 **(0,27%)**	-18,21 %
Total	695.075 **(100,00%)**	754.984 **(100,00%)**	748.045 **(100,00%)**	676.452 **(100,00%)**	569.194 **(100,00%)**	

Fonte: Elaborada pelo autor.

de São Paulo, Rio de Janeiro e Paraná dominaram, ao longo dos anos, o número de casos entre todas as UFs. Já em relação aos locais com o menor número de incidentes, mantiveram-se as UFs do Acre, Amapá e Roraima. Excetuando Piauí e Acre, é notória a queda do número de crimes em 2019, relativa ao ano anterior, seguindo uma tendência geral de queda em 2018, em menores proporções que em 2019. Neste contexto, os destaques vão para Paraíba, Goiás, Rio Grande do Norte e Ceará. No caso da Paraíba, o número de crimes cresceu em 2018 e aumentou a proeminência da queda da criminalidade. Nos outros estados citados, houve uma manutenção da tendência de queda, contudo, mais significativa em 2019. Apesar da queda geral, uma análise detalhada destes resultados é interessante, considerando os tipos de crimes, se houve ações que foram efetivadas, bem como, se estas ações realmente trouxeram efeitos positivos mais significativos para estes estados. O raciocínio inverso vale para os estados do Piauí e Acre. No Acre, os índices sobem desde 2015, no Piauí, a situação é semelhante. Isto poderá ser refletido também na análise de *outliers* posterior e deve ser uma ação contínua do governo para guiar suas políticas públicas de forma seletiva.

Na Figura 24, são apresentados os dados agrupados por região brasileira. Percebe-se que a maioria das ocorrências aconteceram na região Sudeste. Com 1.670.939 casos, atingindo 48,52% do total, esta localidade superou largamente as demais. Na sequência, vieram o Sul, com 645.398 incidentes (18,74%), e o Nordeste, com 630.662 (18,31%). As regiões Centro-Oeste e Norte obtiveram os menores números, com 283.448 (8,23%) e 213.303 (6,19%), respectivamente. A Tabela 10 mostra, detalhadamente, o número de ocorrências anuais de cada região e suas respectivas porcentagens. Nota-se que o padrão anual se manteve ao longo do tempo e coincide com a avaliação feita anteriormente (Figura 24). A região Sudeste obteve os maiores números entre os locais avaliados. Na sequência, em 2019, por exemplo, vieram o Nordeste, o Sul, o Centro-Oeste, e, finalmente, o Norte.

Figura 24 – Porcentagem de ocorrências criminais por região brasileira.



Fonte: Elaborada pelo autor.

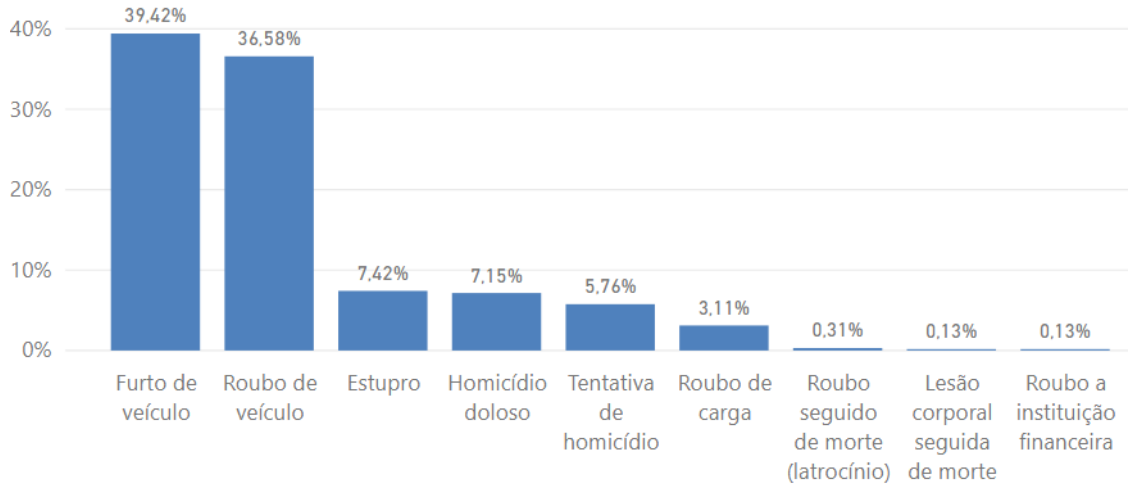
Tabela 10 – Número de ocorrências e porcentagem anual em cada região brasileira.

	2015	**2016**	**2017**	**2018**	**2019**	
Região	**Total (%)**	**Total (%)**	**Total (%)**	**Total (%)**	**Total (%)**	**Variação (2018-2019)**
Sudeste	345.019 **(49,64%)**	363.949 **(48,21%)**	360.786 **(48,23%)**	326.453 **(48,26%)**	274.732 **(48,27%)**	-15,84 %
Nordeste	118.057 **(16,98%)**	133.105 **(17,63%)**	139.183 **(18,61%)**	130.753 **(19,33%)**	109.564 **(19,25%)**	-16,21 %
Sul	133.567 **(19,22%)**	146.836 **(19,45%)**	140.174 **(18,74%)**	120.806 **(17,86%)**	104.015 **(18,27%)**	-13,9 %
Centro-Oeste	60.540 **(8,71%)**	67.184 **(8,90%)**	59.678 **(7,98%)**	53.243 **(7,87%)**	42.803 **(7,52%)**	-19,61 %
Norte	37.892 **(5,45%)**	43.910 **(5,82%)**	48.224 **(6,45%)**	45.197 **(6,68%)**	38.080 **(6,69%)**	-15,75 %
Total	695.075 **(100,00%)**	754.984 **(100,00%)**	748.045 **(100,00%)**	676.452 **(100,00%)**	569.194 **(100,00%)**	

Fonte: Elaborada pelo autor.

O gráfico da Figura 25 apresenta uma análise no contexto dos tipos de crime. Percebe-se que os crimes que mais ocorreram foram “Furto de veículo” e “Roubo de veículo”, que somados representaram mais de 75% dos casos. “Furto de veículos”, com 1.357.383 registros (39,42%), superou levemente “Roubo de veículo”, o qual alcançou 1.259.796 ocorrências (36,58%). Seguindo, tivemos “Estupro”, com 255.366 (7,42%), “Homicídio doloso”, com 246.113 (7,15%), “Tentativa de homicídio”, com 198.350 (5,76%), e “Roubo de carga”, com 107.155 (3,11%). E, por fim, alcançando valores abaixo de 1%, os crimes “Roubo seguido de morte (latrocínio)”, com 10.613 incidentes (0,31%), “Roubo a instituição financeira”, com 4.488 (0,13%), e “Lesão corporal seguida de morte”, com 4.486 (0,13%).

Figura 25 – Porcentagem de ocorrências criminais por tipo de crime.



Fonte: Elaborada pelo autor.

A Tabela 11 os números de ocorrências e porcentagens dos tipos de crime em cada ano.

Aqui, também se pode notar que "Furto de veículo" e "Roubo de veículo" lideraram as ocorrências criminais em qualquer um dos períodos analisados, coincidindo com a avaliação feita na Figura 25. Além disso, os crimes de "Roubo seguido de morte (latrocínio)", "Roubo a instituição financeira" e "Lesão corporal seguida de morte" mantiveram o padrão, com suas taxas anuais abaixo de 1%. Apesar das sucessivas quedas ao longo dos anos, "Roubo a instituição financeira" obteve destaque ainda maior em 2019, chegando a uma redução de 42,60%. Adicionalmente, "Homicídio doloso", com 17,61%, "Roubo de carga", com 18,61%, "Roubo seguido de morte (latrocínio)", com 19,31%, e "Roubo de veículo", com 25,21%, alcançaram reduções consideráveis.

Tabela 11 – Número de ocorrências e porcentagem anual por tipo de crime.

	2015	**2016**	**2017**	**2018**	**2019**	
Tipo de Crime	**Total (%)**	**Total (%)**	**Total (%)**	**Total (%)**	**Total (%)**	**Variação (2018-2019)**
Furto de veículo	285.824 **(41,12%)**	299.037 **(39,61%)**	282.345 **(37,74%)**	258.299 **(38,18%)**	231.878 **(40,74%)**	-10,23 %
Roubo de veículo	246.424 **(35,45%)**	284.185 **(37,64%)**	287.945 **(38,49%)**	252.436 **(37,32%)**	188.806 **(33,17%)**	-25,21 %
Estupro	45.267 **(6,51%)**	48.402 **(6,41%)**	52.405 **(7,01%)**	55.409 **(8,19%)**	53.883 **(9,47%)**	-2,75 %
Homicídio doloso	50.833 **(7,31%)**	53.035 **(7,02%)**	55.406 **(7,41%)**	47.612 **(7,04%)**	39.227 **(6,89%)**	-17,61 %
Tentativa de homicídio	43.770 **(6,30%)**	42.678 **(5,65%)**	40.822 **(5,46%)**	36.726 **(5,43%)**	34.354 **(6,04%)**	-6,46 %
Roubo de carga	18.612 **(2,68%)**	23.271 **(3,08%)**	24.797 **(3,31%)**	22.314 **(3,30%)**	18.161 **(3,19%)**	-18,61 %
Roubo seguido de morte (latrocínio)	2.246 **(0,32%)**	2.470 **(0,33%)**	2.435 **(0,33%)**	1.916 **(0,28%)**	1.546 **(0,27%)**	-19,31 %
Lesão corporal seguida de morte	768 **(0,11%)**	830 **(0,11%)**	1.056 **(0,14%)**	949 **(0,14%)**	885 **(0,16%)**	-6,74 %
Roubo a instituição financeira	1.331 **(0,19%)**	1.076 **(0,14%)**	834 **(0,11%)**	791 **(0,12%)**	454 **(0,08%)**	-42,60 %
Total	695.075 **(100,00%)**	754.984 **(100,00%)**	748.045 **(100,00%)**	676.452 **(100,00%)**	569.194 **(100,00%)**	

Fonte: Elaborada pelo autor.

3.7.2 Análise e Interpretação dos *Outliers*

Com o intuito de responder mais facilmente à questão de pesquisa **Q1**, gráficos *boxplot* foram utilizados na detecção dos pontos "fora da curva". Então, para realizar essa tarefa, foram selecionadas as seguintes variáveis: Ano, Região Brasileira, UF (Unidade Federativa – Estado) e a Taxa de criminalidade. Como foi descrito, o valor desta taxa criminal foi calculado, utilizando a estimativa populacional de cada localidade, a fim de ser proporcional a 100.000 habitantes. Adicionalmente, para facilitar a avaliação periódica do padrão e acompanhar a evolução na linha do tempo, as análises de *outliers* foram feitas de forma anual. Logo, cinco cenários anuais foram avaliados (2015, 2016, 2017, 2018 e 2019). A Tabela 12 e a Tabela 13 mostram as taxas proporcionais anuais dos estados e dos Tipos de crime, respectivamente. Tais tabelas estão ordenadas, de forma decrescente, pelos valores das taxas de 2019. Vale lembrar que as taxas foram calculadas utilizando a Equação 3.5. Mesmo considerando as proporcionalidades das

populações estaduais e população nacional (Tabela 13) em cada ano, os destaques para os estados e tipos de crimes continuaram os mesmos, com um acréscimo nos percentuais das diferenças relativas a 2018. Por exemplo, Ceará teve uma queda de 29% e Homicídio Doloso uma queda de 18,26%.

Tabela 12 – Taxa de criminalidade proporcional a 100.000 habitantes dos estados.

Estado	**2015**	**2016**	**2017**	**2018**	**2019**	**Variação (2018-2019)**
Paraná	583,74	709,28	689,22	608,39	537,42	-11,67 %
Rio de Janeiro	407,30	490,44	562,51	527,93	434,41	-17,71 %
Rondônia	404,75	431,72	355,25	365,27	348,35	-4,63 %
Distrito Federal	447,13	496,81	421,39	365,15	346,54	-5,10 %
São Paulo	482,31	484,96	448,06	403,02	339,79	-15,69 %
Pernambuco	247,72	320,81	392,63	321,64	301,82	-6,16 %
Acre	31,84	59,63	216,48	293,35	294,58	0,42 %
Piauí	198,12	243,79	241,08	274,92	278,20	1,19 %
Roraima	283,34	401,38	340,96	328,67	255,88	-22,15 %
Espírito Santo	265,67	272,87	382,64	320,01	252,90	-20,97 %
Goiás	417,93	502,67	437,06	385,04	250,59	-34,92 %
Rio Grande do Sul	416,05	403,86	381,17	324,84	249,06	-23,33 %
Tocantins	214,69	241,76	234,74	303,49	243,89	-19,64 %
Mato Grosso do Sul	292,24	294,70	285,28	272,52	243,79	-10,54 %
Amapá	271,64	261,03	250,21	244,12	230,45	-5,60 %
Mato Grosso	345,03	327,63	284,13	239,48	229,36	-4,23 %
Sergipe	203,29	272,58	261,79	243,16	221,95	-8,72 %
Santa Catarina	303,62	311,29	271,23	211,39	198,62	-6,04 %
Rio Grande do Norte	222,19	317,70	307,01	287,93	195,47	-32,11 %
Alagoas	206,58	232,27	220,04	210,57	178,19	-15,38 %
Pará	196,85	230,98	266,58	232,62	178,13	-23,42 %
Ceará	239,70	240,18	269,03	230,89	163,94	-29,00 %
Bahia	206,68	217,41	205,16	199,01	163,78	-17,70 %
Amazonas	180,92	219,58	253,27	190,73	160,02	-16,10 %
Minas Gerais	242,27	259,54	233,52	188,29	158,44	-15,85 %
Maranhão	155,94	182,04	170,84	157,71	137,73	-12,67 %
Paraíba	130,34	66,66	58,00	178,96	115,33	-35,56 %

Fonte: Elaborada pelo autor.

Tabela 13 – Taxa de criminalidade proporcional a 100.000 habitantes dos tipos de crime.

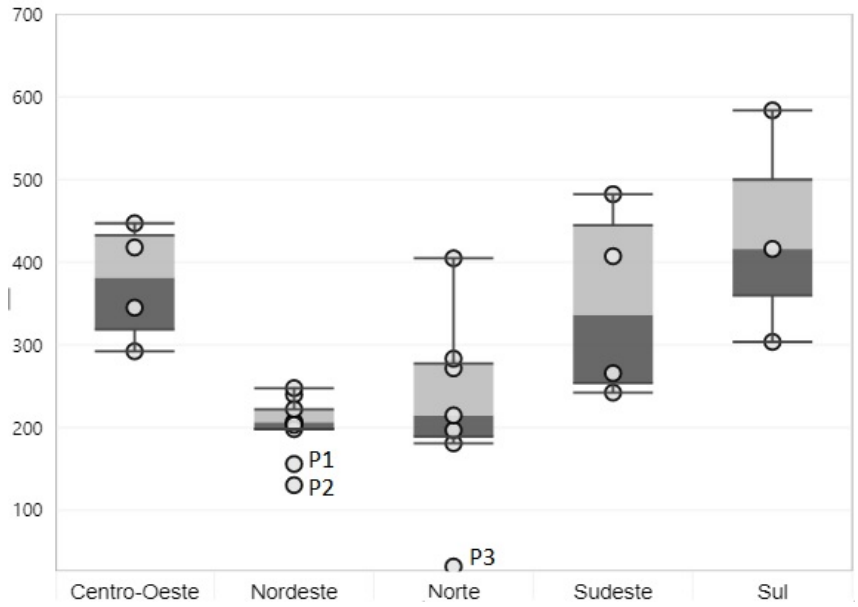
Tipo de crime	**2015**	**2016**	**2017**	**2018**	**2019**	**Variação (2018-2019)**
Furto de veículo	138,69	145,11	135,96	123,89	110,34	-10,94 %
Roubo de veículo	119,58	137,90	138,66	121,08	89,84	-25,80 %
Estupro	21,97	23,49	25,24	26,58	25,64	-3,54 %
Homicídio doloso	24,67	25,73	26,68	22,84	18,67	-18,26 %
Tentativa de homicídio	21,24	20,71	19,66	17,61	16,35	-7,16 %
Roubo de carga	9,03	11,29	11,94	10,70	8,64	-19,25 %
Roubo seguido de morte (latrocínio)	1,09	1,20	1,17	0,92	0,74	-19,57 %
Lesão corporal seguida de morte	0,37	0,40	0,51	0,46	0,42	-8,70 %
Roubo a instituição financeira	0,65	0,52	0,40	0,38	0,22	-42,11 %

Fonte: Elaborada pelo autor.

A Figura 26 representa os gráficos *boxplot* das Regiões Brasileiras para o ano de 2015. Neste exemplo, nota-se que as regiões Centro-Oeste, Sudeste e Sul concentraram as maiores taxa criminais em relação às demais, com índices acima de 242 ocorrências por cem mil habitantes. Na região Centro-Oeste, 50% dos valores das taxas estão entre 318,63 e 432,53, valores que

representam, respectivamente, o primeiro (Q1) e o terceiro (Q3) quartis. Na região Sul, essa porcentagem está entre 359,83 (Q1) e 499,89 (Q3). E, no Sudeste, metade das taxas obtiveram valores entre 253,97 (Q1) e 444,81 (Q3).

Figura 26 – Análise de *outliers* para o ano de 2015.



Fonte: Elaborada pelo autor.

Adicionalmente, percebe-se que para as regiões Nordeste, Norte e Sul, o valor da mediana (Q2) está próximo ao primeiro quartil (Q1), corroborando com a evidência de que os dados, predominantemente, possuem mais valores das taxas acima da mediana (assimetria positiva). Por outro lado, os gráficos do Centro-Oeste e do Sudeste demostram que os valores das medianas estão próximos ao centro, mostrando um equilíbrio na distribuição normal das taxas.

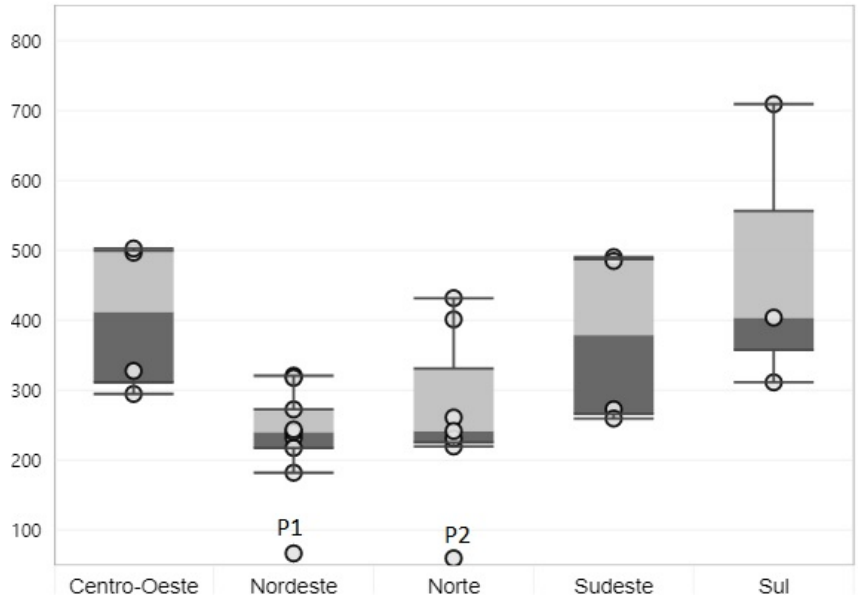
Com relação aos *outliers*, duas regiões apresentaram UFs com taxas criminais abaixo do limite mínimo (*min*), totalizando 3 valores aberrantes. Na região Nordeste, os pontos P1 e P2 esboçados no gráfico representam, respectivamente, os estados do Maranhão e da Paraíba. O primeiro obteve uma taxa criminal de 155,94, enquanto o segundo de 130,34. Já na região Norte, o estado do Acre, simbolizado pelo ponto P3, atingiu a taxa de 31,84. Considerando que se trata do limite inferior, este é um ponto positivo para esses estados.

Analisando a Tabela 12, para o ano 2015, dentre as três UFs com maiores taxas, destacaram-se: Paraná, com 583,74, São Paulo, com 482,31, e Distrito Federal, com 447,13. Por outro lado, os estados do Maranhão, Paraíba e Acre obtiveram os menores índices criminais, com os valores 155,94, 130,34 e 31,84, respectivamente

O gráfico esboçado na Figura 27 demostra o segundo cenário analisado, o ano de 2016. Aqui, verifica-se, também, que as regiões Centro-Oeste, Sudeste e Sul dominaram as maiores taxa

criminais, com valores acima de 259 ocorrências por cem mil habitantes. Na região Centro-Oeste, metade dos seus estados obtiveram taxas que estão entre 311,17 e 499,74, valores que representam, respectivamente, o primeiro (Q1) e o terceiro (Q3) quartis. Já na região Sul, 50% da UFs atingiram índices entre 357,58 (Q1) e 556,57 (Q3). E, no Sudeste, metade das taxas obtiveram valores entre 266,21 (Q1) e 487,70 (Q3).

Figura 27 – Análise de *outliers* para o ano de 2016.



Fonte: Elaborada pelo autor.

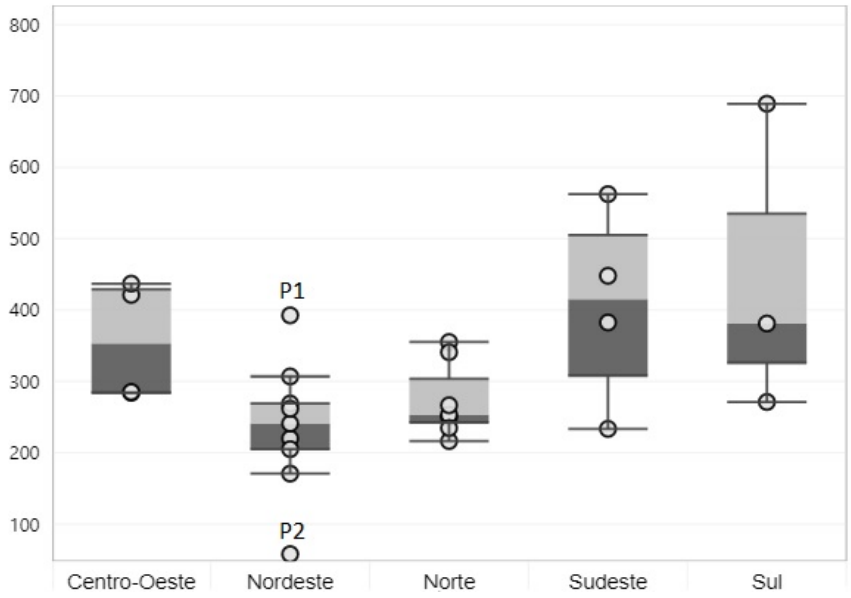
Além disso, nota-se que para as regiões Nordeste, Norte e Sul, o valor da mediana (Q2) está próxima ao primeiro quartil (Q1), ratificando a evidência de que os dados, predominantemente, possuem mais valores das taxas acima da mediana (assimetria positiva). Por outro lado, os gráficos do Centro-Oeste e Sudeste demostram que os valores das medianas estão próximos ao centro, mostrando um equilíbrio na distribuição normal dos valores.

Apenas dois *outliers* foram detectados neste ano. O primeiro, localizado na região Nordeste e simbolizado pelo ponto P1, representa o estado do Paraíba, o qual obteve o valor de 66,66. Já o segundo estado, localizado na região Norte e sinalizado pelo ponto P2, foi o Acre, com 59,63. P1 e P2 são *outliers* que se encontram abaixo do limite mínimo (*min*), o que representa mais um ano positivo para esses estados.

E, por fim, com as maiores taxas em 2016, ver Tabela 12, destacaram-se Paraná, com 709,28, Goiás, com 502,67, e Distrito Federal, com 496,81. Em contrapartida, os estados do Maranhão, Paraíba e Acre atingiram os menores índices criminais com os valores 182,04, 66,66 e 59,63, respectivamente. Apesar de não ser um ponto fora da curva, o Maranhão voltou a ter um bom desempenho.

A Figura 28 representa as taxas de criminalidade para o ano de 2017. Como nos cenários anteriores, nota-se que as regiões Centro-Oeste, Sudeste e Sul concentraram as maiores taxas criminais em relação às demais, com valores acima de 233 ocorrências por cem mil habitantes. Na região Centro-Oeste, 50% dos valores das taxas estão entre 284,71 e 429,22, valores que representam, respectivamente, o primeiro (Q1) e o terceiro (Q3) quartis. Na região Sul, essa porcentagem está entre 326,20 (Q1) e 535,19 (Q3). E, no Sudeste, metade das taxas obtiveram valores entre 308,08 (Q1) e 505,29 (Q3).

Figura 28 – Análise de *outliers* para o ano de 2017.



Fonte: Elaborada pelo autor.

Adicionalmente, verifica-se que para as regiões Norte e Sul, o valor da mediana (Q2) está próximo ao primeiro quartil (Q1), corroborando a evidência de que os dados, predominantemente, possuem mais valores das taxas acima da mediana (assimetria positiva). E, finalmente, os gráficos do Centro-Oeste, Nordeste e Sudeste indicam que os valores das medianas estão próximos ao centro, perfazendo um equilíbrio na distribuição normal das taxas.

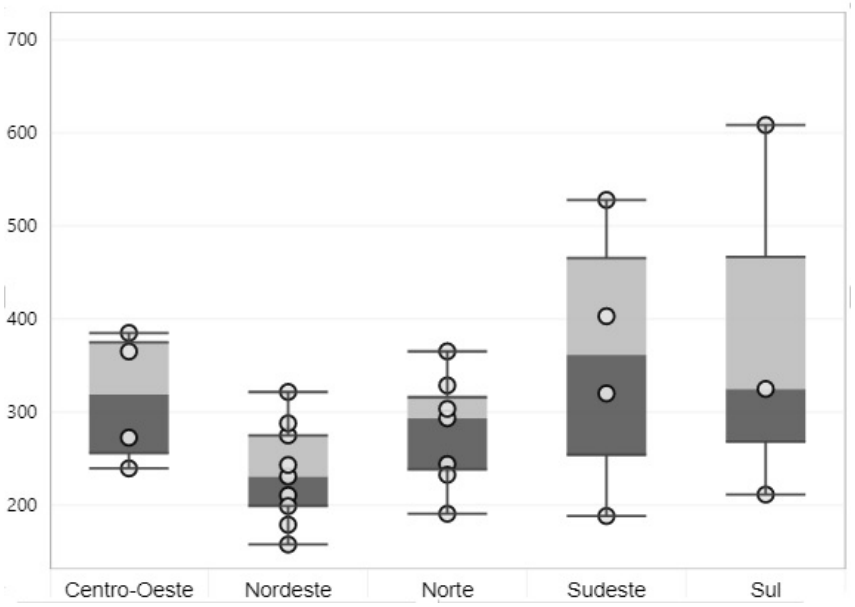
No contexto dos valores discrepantes, somente a região Nordeste apresentou dois *outliers*. Os pontos P1 e P2 esboçados no gráfico representam, respectivamente, os estados do Pernambuco e da Paraíba. O primeiro, com uma taxa criminal de 392,63, excedeu limite máximo (*max*). Já o segundo ultrapassou o limite mínimo (*min*), com o índice igual 58,00. Novamente, um destaque positivo para a Paraíba.

Em 2017, ver a Tabela 12, os estados da Bahia, Maranhão e Paraíba foram os locais que atingiram os menores índices criminais, com os valores 205,16, 170,84 e 58,00, respectivamente. Em contrapartida, as maiores taxas foram do Paraná, com 689,22, Rio de Janeiro, com 562,51, e

São Paulo, com 448,06. Maranhão e Paraíba mais uma vez positivamente entre os últimos, com a Bahia aparecendo pela primeira vez.

O gráfico esboçado na Figura 29 representa os cenários do ano de 2018. Verifica-se que as regiões Centro-Oeste e Norte concentraram taxas criminais que estão entre 190 e 386 ocorrências por cem mil habitantes, aproximadamente. Na região Centro-Oeste, metade dos seus estados obtiveram taxas que estão entre 256,00 e 375,09, valores que representam, respectivamente, o primeiro (Q1) e o terceiro (Q3) quartis. Já na região Norte, 50% da UFs atingiram índices entre 238,37 (Q1) e 316,08 (Q3).

Figura 29 – Análise de *outliers* para o ano de 2018.



Fonte: Elaborada pelo autor.

Um outro fato a destacar é o estado do Paraná, o qual está localizado no limite máximo (*max*) no gráfico da região Sul, obtendo uma taxa igual a 608,39 incidentes por cem mil habitantes. Tal valor foi muito acima dos outros apresentados pelas demais UFs.

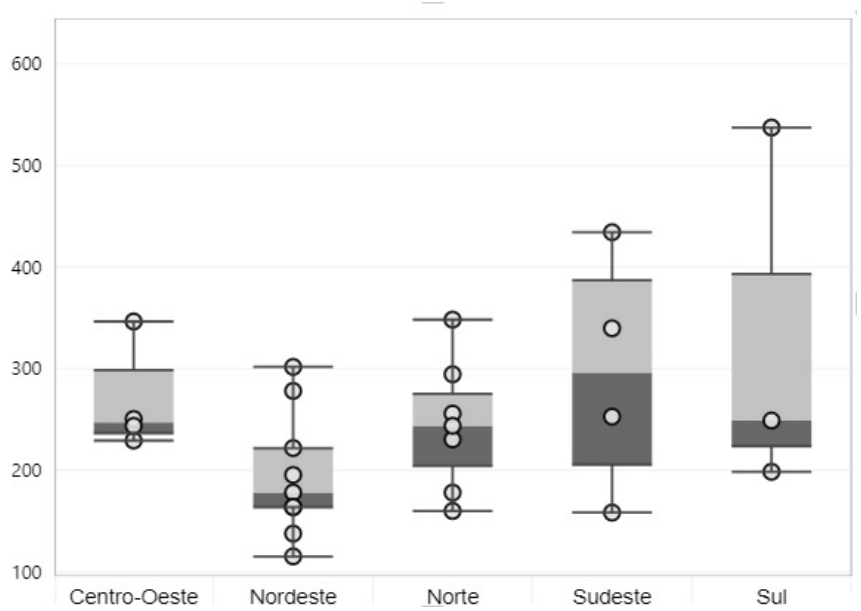
Analisando a região Norte, o valor da mediana (Q2) está próximo ao terceiro quartil (Q3), ratificando a evidência de que os dados, predominantemente, possuem mais valores das taxas abaixo da mediana (assimetria negativa). Já nas regiões Nordeste e Sul, a mediana (Q2) está mais próxima do primeiro quartil (Q1), denotando que a maior parte das taxas possui valores acima da mediana (assimetria positiva). Por outro lado, os gráficos do Centro-Oeste e do Sudeste demostram que os valores das medianas estão próximos ao centro, mostrando um equilíbrio na distribuição normal dos valores.

Baseando-se pela Tabela 12, as maiores taxas em 2018 foram do Paraná, com 608,39, Rio de Janeiro, com 527,93 e São Paulo, com 403,02. Em contrapartida, os estados de Minas Gerais,

Paraíba e Maranhão atingiram os menores índices criminais, com os valores 188,29, 178,96 e 157,71, respectivamente. E, por fim, vale ressaltar que, neste ano, não foram encontrados pontos aberrantes (*outliers*). Paraíba e Maranhão se mantiveram entre os estados com menores índices, com Minas Gerais aparecendo pela primeira vez entre os últimos.

A Figura 30 representa a análise de *outliers* para o ano de 2019, o quinto e último cenário. Avaliando as regiões Centro-Oeste, Nordeste e Sul, o valor da mediana (Q2) está próximo ao primeiro quartil (Q1), corroborando a evidência de que os dados, predominantemente, possuem mais valores das taxas acima da mediana (assimetria positiva). Já os gráficos do Sudeste e Norte indicam que o valores das medianas estão mais próximos aos centros, confirmando um equilíbrio maior na distribuição normal das taxas.

Figura 30 – Análise de *outliers* para o ano de 2019.



Fonte: Elaborada pelo autor.

Com o apoio da Tabela 12, para o ano de 2019, os estados de Minas Gerais, Maranhão e Paraíba, novamente, atingiram os menores índices criminais, com os valores 158,44, 137,73 e 115,33, respectivamente. Além disso, houve uma redução dos valores apresentados por estes estados, em relação a 2018. Por outro lado, as maiores taxas foram alcançadas pelo Paraná, com 537,42, Rio de Janeiro, com 434,41, e Rondônia, com 348,35. Verifica-se, também, uma redução de destaque nos índices de criminalidade dos estados de Goiás, Rio Grande do Norte e Ceará, confirmando os números absolutos e alcançando o menor índice em todos os anos. Além disso, vale destacar que, como no ano anterior, não foram encontrados pontos aberrantes (*outliers*) para o ano de 2019.

De forma resumida, a Tabela 14 mostra as informações dos *outliers* que foram detectados. Nota-se que somente as regiões Norte (2) e Nordeste (5) apresentaram taxas anormais, totalizando

7 pontos discrepantes, localizados acima do Limite Superior (LS) ou abaixo do Limite Inferior (LI), distribuídos entre quatro dos cinco anos avaliados. No ano de 2015, 3 *outliers*, localizados abaixo do LI, foram detectados. Em 2016, tivemos 2, abaixo do LI. E, no ano de 2017, foram 2, 1 acima do LS e 1 abaixo do LI.

Tabela 14 – Resumo dos *outliers* encontrados por ano vs. região brasileira.

Ano	Quantidade	Ponto	Região	Estado	Taxa
2015	3	P1 (abaixo do LI)	Nordeste	Maranhão	155,94
		P2 (abaixo do LI)	Nordeste	Paraíba	130,34
		P3 (abaixo do LI)	Norte	Acre	31,84
2016	2	P1 (abaixo do LI)	Nordeste	Paraíba	66,66
		P2 (abaixo do LI)	Norte	Acre	59,63
2017	2	P1 (acima do LS)	Nordeste	Pernambuco	392,63
		P2 (abaixo do LI)	Nordeste	Paraíba	58,00

Fonte: Elaborada pelo autor.

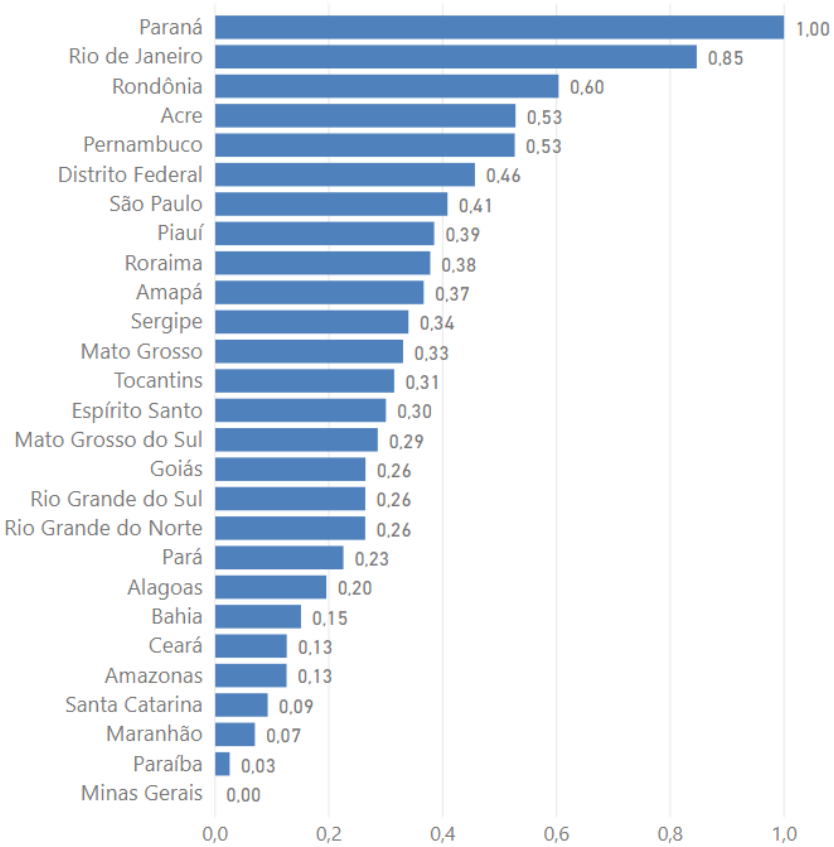
3.7.3 Análise e Interpretação do *Ranking* de Perigosidade

Com relatado na seção 3.4.4, um grande problema encontrado na análise criminal é quantificar quanto um determinado local é mais perigoso que outro. Isso torna-se ainda mais problemático quando envolvemos diversos tipos de crimes dentro da mesma análise. Para responder quais os estados vêm se destacando como os mais perigosos (questão **Q2**), um Índice de Criminalidade (*IC*) foi criado, baseado em uma média ponderada, calculada a partir da taxa da criminalidade (proporcional à população) e nos pesos adotados para os diversos tipos de crimes. Posteriormente, para permitir uma classificação mais interpretável e clara, o *IC* foi normalizado e reenquadrado em classes categóricas.

A Figura 31 mostra os Índices Normalizados dos estados, no ano mais recente (2019). Ao confrontarmos esses dados com a quantidade anual de incidentes ocorridos nos estados, ver Tabela 9, notam-se algumas alterações, evidenciando que nem sempre os locais que possuem altos números de casos criminais apresentaram, segundo os pesos adotados, altos níveis de perigosidade. Por esta nova análise, a Unidade Federativa de Minas Gerais, quarto maior local em incidentes criminais em 2019, foi classificado como o menos perigoso e seu nível obtido foi “Baixíssimo”. Por outro lado, o Paraná, o terceiro maior estado em ocorrências criminais, superou os dois primeiros, São Paulo e Rio de Janeiro, tornando-se o mais perigoso, com o grau “Altíssimo”. Vale ressaltar que São Paulo, mesmo sendo o local com maior quantidade de casos, atingiu o nível “Intermediário”.

O *ICN* indica um *ranking*, uma ordem de perigosidade, relativa ao Brasil e ao ano analisado (vide seção 3.4.4), podendo indicar, por exemplo, que as políticas públicas do governo federal devem considerar uma atenção maior para o Paraná (Altíssima), Rio de Janeiro e Rondônia, cumulativas às desenvolvidas para 2019. Um estado ser classificado em um ano como

Figura 31 – Índice de criminalidade normalizado dos estados em 2019.



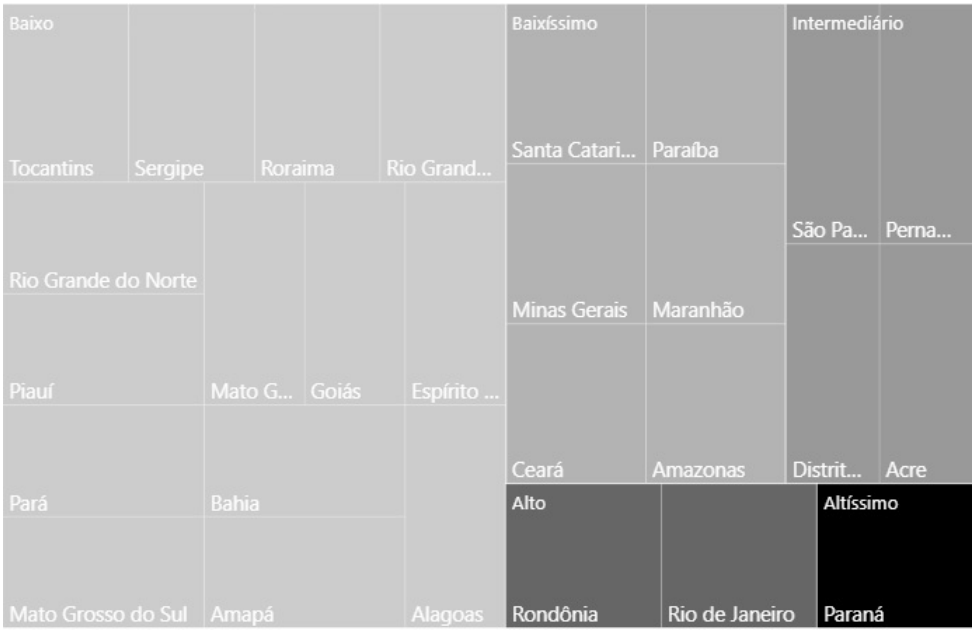
Fonte: Elaborada pelo autor.

intermediário e em outro ano como alto (vide Tabela 15) não implica um aumento no nível de perigosidade absoluto. Esta troca de posição pode indicar, apenas, uma menor distância para o primeiro colocado daquele ano.

Um gráfico *Treemap*, exibido na Figura 32, demostra a classificação dos estados com relação aos níveis de perigosidade. Gráficos *Treemap* utilizam um método para exibir dados hierárquicos, usando retângulos aninhados. Cada retângulo é preenchido com retângulos menores lado a lado, os quais representam subcategorias e possuem uma área proporcional a uma dimensão especificada nos dados. Em nossos exemplos a seguir, cada grau de perigosidade é representado por um retângulo, com uma certa tonalidade cinza. Estes retângulos são descompostos em vários outros menores, que representam os estados. Nota-se que apenas 1 (uma) Unidade Federativa, mais especificamente o Paraná, foi classificada com o grau "Altíssimo". Com o nível "Alto", tivemos 2 estados (Rio de Janeiro e Rondônia). Já como "Intermediário", foram 4 (Acre, Distrito Federal, Pernambuco e São Paulo). Com o nível "Baixo", enquadraram-se 14 estados (Alagoas, Amapá, Bahia, Espírito Santo, Goiás, Mato Grosso, Mato Grosso do Sul, Pará, Piauí, Rio Grande do Norte, Rio Grande do Sul, Roraima, Sergipe e Tocantins). E, finalmente, 6 UFs foram classificadas como as menos perigosas, isto é, com o grau "Baixíssimo" (Amazonas, Ceará,

Maranhão, Minas Gerais, Paraíba e Santa Catarina).

Figura 32 – Estados por nível de perigosidade em 2019.



Fonte: Elaborada pelo autor.

A Tabela 15 detalha, para cada ano, a quantidade de estados por nível de perigosidade e as cinco UFs mais perigosas, com seus respectivos índices de criminalidade normalizados. É importante destacar que Paraná foi o local mais perigoso em todos os períodos avaliados, sendo sempre classificado no grau "Altíssimo". Destaque também para Rio de Janeiro, presente em todos os anos, sempre na segunda colocação. Além disso, os estados de Goiás, Pernambuco e Rondônia foram classificados entre os cinco mais perigosos, em três dos cinco anos.

Nos últimos anos, o Atlas da Violência (CERQUEIRA et al., 2018) tem demonstrado que as regiões Nordeste e Norte vêm se revezando como regiões mais violentas, se considerarmos apenas o número de assassinatos por 100.000 habitantes. Isto pode ser explicado pelo crescimento econômico desordenado do Nordeste e a falta de estruturas melhores para o sistema carcerário, bem como para segurança pública como um todo, ou seja, o crescimento econômico não foi acompanhado por investimentos no treinamento e fortalecimento das polícias, bem como por melhorias no sistema prisional. Além disso, facções criminosas hegemônicas no controle do tráfico no Rio de Janeiro e em São Paulo expandiram seus negócios para o Norte e Nordeste. Esse histórico permite uma atenção maior ao ano de 2019, no qual, apesar de Rondônia, Acre e Pernambuco aparecerem entre os cincos mais perigosos, corroborando o histórico do Atlas da Violência, houve uma queda geral nos índices das duas regiões, excetuando os estados do Acre e do Piauí. Neste contexto, o destaque vai para os estados do Rio Grande do Norte, Ceará e Paraíba, no Nordeste, e Pará e Roraima, no Norte, indicando que as ações federais e estaduais devem ser analisadas e podem ser eleitas como replicáveis.

Tabela 15 – Detalhamento anual dos estados mais perigosos (Top 5).

Ano	Total de UFs por Nível	Estado - *ICN* (Nível)
2015	Altíssimo : 1 Alto : 7 Intermediário : 10 Baixo : 8 Baixíssimo : 1	Paraná - 1,00 (Altíssimo) Rio de Janeiro - 0,77 (Alto) São Paulo - 0,74 (Alto) Distrito Federal - 0,72 (Alto) Goiás - 0,72 (Alto)
2016	Altíssimo : 1 Alto : 3 Intermediário : 8 Baixo : 13 Baixíssimo : 2	Paraná - 1,00 (Altíssimo) Rio de Janeiro - 0,77 (Alto) Goiás - 0,71 (Alto) Distrito Federal - 0,63 (Alto) Rondônia - 0,58 (Intermediário)
2017	Altíssimo : 2 Alto : 1 Intermediário : 12 Baixo : 11 Baixíssimo : 1	Paraná - 1,00 (Altíssimo) Rio de Janeiro - 0,92 (Altíssimo) Pernambuco - 0,64 (Alto) Goiás - 0,60 (Intermediário) Espirito Santo - 0,53 (Intermediário)
2018	Altíssimo : 2 Alto : 0 Intermediário : 8 Baixo : 12 Baixíssimo : 5	Paraná - 1,00 (Altíssimo) Rio de Janeiro - 0,97 (Altíssimo) Rondônia - 0,50 (Intermediário) Roraima - 0,50 (Intermediário) Pernambuco - 0,49 (Intermediário)
2019	Altíssimo : 1 Alto : 2 Intermediário : 4 Baixo : 14 Baixíssimo : 6	Paraná - 1,00 (Altíssimo) Rio de Janeiro - 0,85 (Alto) Rondônia - 0,60 (Alto) Acre - 0,53 (Intermediário) Pernambuco - 0,53 (Intermediário)

Fonte: Elaborada pelo autor.

Embora a queda nos índices seja importante e indique uma tendência positiva para os estados do Norte e Nordeste destacados, se considerarmos apenas os assassinatos, isto não implica que os índices destas regiões alcançaram um patamar baixo em relação ao país. A Tabela 16 detalha o ano de 2019, por estado e por crime, considerando também o número de ocorrências proporcional a 100.000 habitantes. Na coluna 10, foram somados os crimes de "Lesão corporal seguida de morte", "Homicídio doloso" e "Roubo seguido de morte (latrocínio)", para que fossem totalizadas todas as mortes disponíveis, por estado, e para que os estados fossem reordenados por esta coluna. Sob este ponto de vista, o qual gera discussões entre juristas e especialistas em segurança pública, do primeiro ao décimo colocado, figuraram estados do Norte e Nordeste, destacando-se Roraima, Rio Grande do Norte, Sergipe, Acre e Pernambuco, com estes dois últimos configurando entre os mais perigosos, sob as duas perspectivas (geral ponderada e assassinatos).

A discussão sobre a perspectiva única dos assassinatos leva em consideração a ausência, por exemplo, do registro da barbárie e da natureza hedionda do estupro. Neste sentido, chama a atenção os números do Paraná, Rondônia, Mato Grosso do Sul, Amapá e Mato Grosso. Isto explica a posição de Rondônia no *ranking* de perigosidade, levando-se em consideração ainda que um estupro pode resultar em morte e que, infelizmente, os dados disponíveis não detalham os subtipos do estupro. De qualquer forma, ainda que não seja letal, o estupro pode mudar uma

Tabela 16 – Taxa dos tipos de crime, por estado, no ano de 2019, proporcional a 100.000 habitantes.

	Crimes									
Estado	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Roraima	101,86	52,00	0,17	21,96	1,98	0,00	42,92	32,03	2,97	36,98
Rio Grande do Norte	19,82	128,69	0,51	2,94	5,47	0,17	6,70	29,29	1,88	36,64
Sergipe	42,11	91,10	0,30	25,45	0,17	0,30	27,49	33,85	1,17	35,19
Acre	64,86	141,28	0,45	15,19	0,23	0,45	37,42	33,11	1,59	34,93
Pernambuco	64,21	137,35	5,80	37,69	0,19	0,09	21,74	33,36	1,38	34,93
Bahia	33,27	74,17	1,45	16,16	0,46	0,03	5,40	31,92	0,93	33,31
Amapá	62,19	52,85	1,77	25,07	2,72	0,95	55,69	27,43	1,77	31,92
Alagoas	41,23	72,30	1,17	12,88	0,18	0,12	19,00	30,80	0,51	31,49
Paraíba	21,95	48,85	0,32	6,69	0,00	0,55	5,57	30,93	0,45	31,38
Pará	41,71	45,29	2,09	17,73	0,37	0,29	39,70	29,54	1,42	31,33
Paraná	279,22	101,26	10,93	7,43	0,87	0,30	109,38	26,99	1,03	28,89
Amazonas	47,53	59,02	0,22	6,13	0,55	1,35	20,36	23,74	1,11	25,40
Espírito Santo	100,46	72,46	0,37	41,88	0,20	0,62	12,04	24,31	0,55	25,06
Goiás	117,93	68,28	2,15	27,58	0,77	0,30	10,13	22,44	1,00	24,21
Mato Grosso	67,93	52,69	1,09	34,38	0,66	0,06	49,28	22,10	1,18	23,94
Ceará	50,39	56,15	0,99	12,86	0,33	0,15	19,82	22,85	0,39	23,57
Rondônia	123,17	102,18	0,11	38,82	0,39	0,06	61,44	21,27	0,90	22,56
Rio de Janeiro	90,32	230,24	43,18	19,73	0,26	0,33	28,5	21,21	0,63	22,10
Maranhão	35,89	45,96	0,59	12,97	0,18	0,55	20,78	19,75	1,06	20,99
Tocantins	99,75	44,57	0,06	32,93	0,45	0,19	45,65	19,14	1,14	20,73
Piauí	89,91	131,67	0,58	17,05	0,46	0,12	20,62	16,74	1,04	18,24
Mato Grosso do Sul	124,65	19,11	0,04	25,40	0,25	0,25	58,80	14,79	0,50	15,54
Rio Grande do Sul	101,49	96,24	2,53	23,00	0,18	0,29	10,75	14,00	0,56	14,74
Distrito Federal	169,97	113,65	1,13	26,43	0,13	0,00	22,09	12,34	0,80	13,27
Minas Gerais	96,64	28,83	1,68	13,49	0,26	0,14	5,81	11,24	0,35	11,85
Santa Catarina	123,72	23,96	0,54	19,64	0,21	0,27	20,74	9,18	0,36	9,75
São Paulo	183,15	101,29	15,95	7,38	0,22	0,05	25,29	6,05	0,42	6,69

Legenda:
(1) Furto de veículos
(2) Roubo de veículos
(3) Roubo de carga
(4) Tentativa de homicídio
(5) Lesão corporal seguida de morte
(6) Roubo a instituição financeira
(7) Estupro
(8) Homicídio doloso
(9) Roubo seguido de morte (latrocínio)
(10) (5)+(8)+(9)

Fonte: Elaborada pelo autor.

vida ou, metaforicamente, ceifá-la para sempre.

Vale ressaltar que, no caso do Paraná, seu grande número de Estupros e de Furtos de Veículos o colocou em primeiro lugar no *ranking* de perigosidade. No Rio de Janeiro, o segundo lugar é justificado pelo grande número de roubos de veículos, pode-se dizer que está muito perigoso locomover-se com veículo automotor neste estado. Do ponto de vista positivo, Minas Gerais e Santa Catarina aparecem entre os melhores resultados, nas duas perspectivas (geral ponderada e assassinatos).

3.7.4 Análise e Interpretação das Regras de Associação

Para realizar esta análise, foram utilizadas 3.443.750 transações para a geração das regras de associação. Em nosso trabalho, as transações foram separadas por ano, resultando em cincos subgrupos (2015, 2016, 2017, 2018 e 2019), e, posteriormente, analisadas separadamente pelo algoritmo de mineração de dados *Apriori*. A Tabela 17 mostra algumas informações obtidas após o processamento. Entre elas estão: o número de transações analisadas, os cinco itens mais frequentes e a quantidade de regras de associação encontradas em cada ano.

Tabela 17 – Quantitativos de transações analisadas para cada ano.

Ano	Transações	Frequentes	Regras
2015	695.075	Furto de veículo: 285.824 Roubo de veículo: 246.424 São Paulo: 215.832 Rio de Janeiro: 67.759 Paraná: 65.628	1
2016	754.984	Furto de veículo: 299.037 Roubo de veículo: 284.185 São Paulo: 217.018 Rio de Janeiro: 81.590 Paraná: 79.742	2
2017	748.045	Roubo de veículo: 287.945 Furto de veículo: 282.345 São Paulo: 202.053 Rio de Janeiro: 94.046 Paraná: 78.026	2
2018	676.452	Furto de veículo: 258.299 Roubo de veículo: 252.436 São Paulo: 183.532 Rio de Janeiro: 90.592 Paraná: 69.046	2
2019	569.194	Furto de Veículo: 231.878 Roubo de Veículo: 188.806 São Paulo: 156.030 Rio de Janeiro: 75.000 Paraná: 61.448	3

Fonte: Elaborada pelo autor.

Como pode ser visto na Tabela 17, no total, 10 associações foram encontradas. Totalizando 3 associações, o ano de 2019 foi o período com a maior quantidade detectada. Na sequência, com 2, vieram os anos de 2016, 2017 e 2018. E, por fim, com apenas uma associação, o ano de 2015. Na Tabela 18, as informações encontradas pelas regras são detalhadas, juntamente com os valores de suas medidas de interesse: suporte (*supp*), confiança (*conf*), *lift*, quantidade (*count*), coeficiente de correlação de *Pearson* (r), qui quadrado (χ^2) e o seu *p-value*. Entretanto, é preciso que analisemos esta tabela sob a luz da estatística, para encontrarmos evidências que ratifiquem, ou não, as regras de associação encontradas. Para tal análise, adotamos um nível de confiança de 95% (α=0,05) para o experimento.

Analisando a Tabela 18, percebe-se que todas as associações atenderam aos requisitos

Tabela 18 – Regras encontradas para associações entre tipos de crime e estados.

Ano	**Regra**	***supp***	***conf***	***lift***	***count***	***r***	χ^2	***p-value* do χ^2**
2015	Santa Catarina ⇒ Furto de veículo	0,0202	0,67	1,63	14.068	0,09	6.006,56	0,00
2016	Rio Grande do Norte ⇒ Roubo de veículo	0,0093	0,64	1,69	7.024	0,07	3.222,18	0,00
2016	Santa Catarina ⇒ Furto de veículo	0,0189	0,66	1,68	14.291	0,09	6.660,82	0,00
2017	Rio Grande do Norte ⇒ Roubo de veículo	0,0093	0,65	1,69	6.992	0,07	3.227,08	0,00
2017	Santa Catarina ⇒ Furto de veículo	0,0160	0,63	1,67	11.987	0,08	5.341,67	0,00
2018	Rio Grande do Norte ⇒ Roubo de veículo	0,0103	0,70	1,87	6.982	0,08	4.558,42	0,00
2018	Santa Catarina ⇒ Furto de veículo	0,0135	0,61	1,60	9.133	0,07	3.391,42	0,00
2019	Minas Gerais ⇒ Furto de veículo	0,0359	0,61	1,50	20.457	0,10	6.057,41	0,00
2019	Rio Grande do Norte ⇒ Roubo de veículo	0,0079	0,66	1,98	4.513	0,08	3.339,62	0,00
2019	Santa Catarina ⇒ Furto de veículo	0,0156	0,62	1,53	8.864	0,07	2.807,33	0,00

Fonte: Elaborada pelo autor.

mínimos exigidos pelo experimento. Como relação à medida suporte, os valores alcançados superaram o suporte mínimo de 0,14% (0,0014). O mesmo aconteceu com a confiança, ou seja, os resultados foram superiores à confiança mínima estipulada (0,60 ou 60%). Em relação ao *lift*, nota-se que os valores atingidos pelas associações foram maiores que 1, indicando que existe uma dependência positiva entre o estado e o tipo de crime. Além disso, os coeficientes da correlação de *Pearson* resultaram em valores superiores a 0 (zero), demonstrando que há uma correlação positiva, ou seja, quando a frequência do "Antecedente" aumentar, a frequência do "Consequente" aumentará, na proporção indicada por *r*.

Vale apena destacar que a associação "Rio Grande do Norte ⇒ Roubo de veículo", presente em 4 anos. O poder público deve investigar as causas e ações necessárias para coibir um crime no estado que tem representado mais de 64% das ocorrências, aproximadamente o dobro dos percentuais do país nos últimos anos. Também é importante investigar a associação deste tipo de crime com o grande número de assassinatos do estado, ou seja, uma ação pode baixar dois índices em paralelo.

Apesar de não implicar causa, as regras de 2019 para Minas Gerais e Santa Catarina coadunam com seus baixíssimos níveis de perigosidade em relação ao resto do país, uma vez que furto de veículo possui o menor peso para perigosidade. Isto não implica negligenciar este fato, mas também pode nortear as políticas públicas para estes estados. No caso específico de Minas Gerais, este é um dos estados que melhor disponibiliza dados de segurança pública, detalhando melhor as localidades e tipos de crimes.

Finalmente, para responder à questão **Q3**, verifica-se que os *p-value* foram abaixo do nível de significância adotado (*p-value*$< 0,05$). Dessa forma, para essas regras, a hipótese H_0 pode ser rejeitada, indicando que há uma dependência entre o estado e o tipo de crime.

3.7.5 Ameaças à Validade

Uma das principais questões de um experimento é quão válidos são os resultados obtidos por ele. Tais resultados devem ser válidos para a população da qual o conjunto de participantes foi selecionado. Com isso, é interessante generalizar os resultados para uma população mais ampla (TRAVASSOS; BARROS, 2003). Ameaças à validade podem limitar a habilidade de interpretar e/ou descrever resultados dos dados obtidos em um experimento (CHAPETTA, 2006). Portanto, não há como desconsiderar as seguintes ameaças encontradas durante a experimentação.

- **Ameaças às validades de construção e interna -** Considerando que os dados foram obtidos, por meio de *download*, tratados e analisados pelos autores, existem ameaças a serem consideradas. Para mitigar possíveis erros, todos os artefatos de software construídos para o tratamento dos dados e os resultados por eles gerados foram homologados e revisados por mais de um pesquisador, considerando amostras de cálculos feitos pelos artefatos, contra amostras de cálculos replicadas em planilhas, manualmente e diretamente no banco de dados. Tais testes foram feitos na fase de construção (validade de construção) dos artefatos e na fase de execução (validade interna).

- **Ameaças à validade de conclusão -** Os pesos considerados para os crimes podem não refletir a cultura, o nível de violência e o clamor social de algum lugar específico. Esta ameaça foi mitigada com a consideração da opinião de especialistas, das penas dos crimes e das demais variáveis citadas, dentro de uma visão geral do contexto brasileiro. Além disso, as distâncias do pesos são aproximadamente proporcionais às distâncias das penas. Vale ressaltar que os pesos podem variar de acordo com o objetivo do estudo. Neste trabalho, consideramos o direcionamento de políticas públicas na área de segurança.

- **Ameaças à validade externa - (1)** A falta de correção dos dados criminais fornecidos pelo governo brasileiro e das estimativas populacionais fornecidas pelo IBGE (Instituto Brasileiro de Geografia e Estatística), bem como possíveis subnotificações por parte dos estados poderão influenciar diretamente o resultado do experimento. Os estados também podem enviar correções atrasadas dos dados, o que denuncia como o país ainda precisa evoluir em termos de política e arquitetura para dados abertos, padronizando estruturas e protocolos, impondo limites temporais e punindo não conformidades. Isto foi mitigado com a disponibilidade de dados "vivos" em nosso portal, os quais são atualizados mensalmente (www.transparenciatraduzida.com.br). **(2)** Considerando que as avaliações das regras de associação realizadas no experimento foram feitas anualmente, os resultados obtidos dizem respeito apenas às associações existentes entre os tipos de crimes e os estados brasileiros neste escopo. No entanto, o resultado poderá sofrer alterações caso a mesma avaliação seja aplicada em um espaço temporal diferente, como, por exemplo, uma análise semestral ou sobre toda a base de dados. **(3)** A agregação, imposta pelo governo, de diversos tipos de crime, sem o detalhamento do nível de gravidade, também pode influenciar os resultados,

uma vez que não é possível pesar e ponderar com maior precisão pela gravidade. Por fim, os *rankings* se limitam aos tipos de crimes disponibilizados.

3.8 Conclusão e Trabalhos Futuros

Este trabalho teve o intuito de detectar padrões e anomalias, bem como promover maior transparência, visando auxiliar o processo de apoio às decisões estratégicas e operacionais dos governantes e agentes da lei, no combate efetivo da criminalidade. Foram utilizados dados das 27 Unidades Federativas, relacionados a incidentes criminais, disponibilizados, de forma aberta, pelo Governo Federal, através do Ministério da Justiça e Segurança Pública (MJSP).

Para auxiliar na identificação das forças das regras de associação geradas pelo experimento, foram utilizadas as medidas de interesse: Suporte, Confiança, *Lift*, *r* (Coeficiente de correlação de *Pearson*) e *Qui Quadrado* (χ^2), com seu nível de significância (*p-value*).

Como resultado da resposta à questão **Q1**, 7 *outliers* foram detectados nas regiões Norte (2) e Nordeste (5), distribuídos nos anos 2015 (3), 2016 (2) e 2017 (2). Não foram detectados valores aberrantes para os dois últimos anos (2018 e 2019). Na resposta para a questão **Q2**, ficou constatado que, em três dos 5 anos avaliados, os estados de Goiás, Pernambuco e Rondônia estiveram entre os cinco mais perigosos. Nesta mesma perspectiva, o Paraná foi considerado o local mais perigoso em todos os períodos avaliados, seguido sempre pelo Rio de Janeiro.

O ano de 2019 deve ser observado com atenção, tanto pela queda geral dos índices de criminalidade em relação a 2018 quanto pelas quedas específicas de estados tais como Paraíba, Goiás, Rio Grande do Norte, Ceará, Pará e Roraima, ou pela considerável queda no índice ou pela localização na região Norte ou Nordeste, as quais vêm configurando entre as mais violentas nos últimos Atlas da Violência. Essas evidências indicam que as ações federais e estaduais devem ser analisadas e podem ser eleitas como replicáveis em outros estados.

Sob a perspectiva única dos assassinatos, em 2019, do primeiro ao décimo colocado, figuraram como mais violentos estados do Norte e Nordeste, destacando-se, na ordem, Roraima, Rio Grande do Norte, Sergipe, Acre e Pernambuco, com estes dois últimos configurando entre os mais perigosos, sob as duas perspectivas (geral ponderada e assassinatos). Também sob as duas perspectivas, os destaques positivos são para Minas Gerais e Santa Catarina. No caso específico de Minas Gerais, este é um dos estados que melhor disponibiliza dados de segurança pública, detalhando melhor as localidades e tipos de crimes.

No âmbito das regras de associação (questão **Q3**), considerando a limitação imposta pelos dados disponibilizados, o destaque ficou por conta da associação "Rio Grande do Norte $\Rightarrow$ Roubo de veículo", presente em 4 anos. O poder público deve investigar as causas e ações necessárias para coibir um crime no estado que tem representado mais de 64% das ocorrências, aproximadamente o dobro dos percentuais do país nos últimos anos. Também é importante

investigar a associação deste tipo de crime com o grande número de assassinatos do estado, ou seja, uma ação pode baixar dois índices em paralelo.

Além da pesquisa desenvolvida neste trabalho e como trabalhos futuros, outros possíveis desdobramentos poderão ser analisados. Dados criminais levando em consideração regiões menores, como, por exemplo, os municípios ou as regiões metropolitanas, poderão ser investigados, desde que estados ou federação os disponibilizem. Nesta linha, podem ser encontradas as associações entre crimes, considerando a localidade e os espaços temporais disponíveis, bem como permitindo a prioridade na inibição de crimes que são antecedentes de outras ocorrências dentro de um mesmo estado ou cidade. Sobre pesos de crimes, sugerimos a realização de um *Survey* com juristas, criminalistas, profissionais e especialistas em segurança pública em todo o Brasil, coletando as medianas de pesos atribuídos e ampliando a discussão sobre o que pode ser considerada uma localidade mais perigosa e sobre as penas dos crimes.

À guisa de conclusão, destacamos a relevância desta pesquisa e dos órgãos aqui envolvidos e citados, os quais têm a responsabilidade social de apresentar resultados que servem como direcionamento para as decisões sobre segurança pública em todo o país. As proeminências destes órgãos e de análises como as descritas neste artigo devem servir de alerta e direcionamento para os nossos governantes, auxiliando o planejamento estratégico e o processo de apoio à decisão, buscando conter o orçamento empregado na área de segurança pública. Os estados podem enviar correções atrasadas dos dados, o que denuncia como o país ainda precisa evoluir em termos de política e arquitetura para dados abertos, padronizando estruturas e protocolos, impondo limites temporais e punindo não conformidades. Os dados aqui apresentados poderão ser consultados, nessas e em outras perspectivas, no site do projeto Transparência Traduzida, em (www.transparenciatraduzida.ufs.br).

4

Data Science Aplicada à Análise Criminal Baseada nos Dados Abertos Governamentais dos Municípios de Minas Gerais

Este capítulo traz um artigo relativo ao segundo experimento controlado, publicado no periódico *Research, Society and Development* (PRADO; COLAÇO JÚNIOR, 2020a).

4.1 Introdução

Com o aumento da urbanização, diversas transformações sociais, econômicas e ambientais têm ocorrido em todas as partes do mundo, com desafios cada vez maiores enfrentados pelos governantes. Áreas como mobilidade urbana, saúde e segurança pública têm recebido uma atenção especial (CATLETT et al., 2018).

Nos últimos anos, a criminalidade vem se destacando como um problema social predominante e sua mitigação é extremamente importante. Segundo (TOPPIREDDY; SAINI; MAHAJAN, 2018), a prevenção e controle do crime são questões de grande preocupação para os governos e agências de segurança pública, uma vez que, se não forem bem controladas e gerenciadas, podem afetar drasticamente a economia de um país ao longo do tempo, uma vez que mais emigração ocorrerá naturalmente. A cada ano, os governos gastam milhões de dólares combatendo a violência, fornecendo equipamentos, treinamento e adquirindo ferramentas para auxiliar o trabalho policial. Consequentemente, governantes e a sociedade, em geral, têm tido enormes problemas causados por esse fenômeno. É responsabilidade das agências de aplicação da lei monitorar e reduzir a taxa de atividades criminosas que estão acontecendo continuamente nos dias de hoje (DAMASCENO; TEIXEIRA; CAMPOS, 2012)(MARZAN et al., 2017).

Um outro grande desafio enfrentado por essas organizações é lidar com um grande volume de informações referentes aos crimes e criminosos. De acordo com (CATLETT et al., 2018), uma quantidade significativa de dados com informações espaciais e temporais é obtida diariamente.

Consequentemente, novas abordagens e sistemas avançados são necessários para melhorar a análise de crimes e para proteger suas comunidades, permitindo uma maior compreensão da dinâmica das atividades criminosas e fornecendo respostas como "onde", "quando" e "por que" certos crimes são prováveis de acontecer (TOPPIREDDY; SAINI; MAHAJAN, 2018) (PHILLIPS; LEE, 2011). Neste contexto, a *Data Science*, aliada aos sistemas computacionais inteligentes, vem desempenhando um papel vital na melhoria dos resultados das investigações e detecções criminais, facilitando o registro, a análise de recuperação e o compartilhamento das informações (GUPTA; CHANDRA; GUPTA, 2014).

Essa evolução da *Data Science* e da Tecnologia da Informação e Comunicação (TIC) em geral produziu novas maneiras de disponibilizar informações públicas à população. Novos sistemas foram desenvolvidos, novos serviços oferecidos e integrações sistêmicas aconteceram, gerando mudanças nos processos internos e nas relações do governo com o público externo. Por meio do Governo Eletrônico, ou simplesmente *e-Gov*, os órgãos públicos passaram a ser os maiores criadores e coletores de dados em muitos domínios. Estes domínios de dados variam entre tráfego, clima, informações geográficas, turísticas, segurança pública, estatísticas, negócios, orçamentação do setor público e vários outros (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012).

Entretanto, apesar das ações realizadas pelos governos federais e estaduais brasileiros para publicitar informações sobre estatísticas oficiais de segurança pública, nenhuma dessas iniciativas que promovem a transparência parece ser suficiente para produzir informações claras, consistentes e transparentes ao grande público. São milhares de fluxos de dados publicados frequentemente, sem a certeza de que sejam utilizáveis e sem uma aplicação inteligente sobre os dados avaliados. Muitas vezes, sem métricas que possibilitem ao cidadão inferir conclusões acerca da publicação.

Em razão disso, a proposta deste artigo é aplicar *Data Science* apoiada em um processo experimental, para realizar uma avaliação sobre dados abertos governamentais relacionados a incidentes criminais, dos munícipios de Minas Gerais (MG), disponibilizados pela Secretaria de Estado de Justiça e Segurança Pública (Sejusp), que, apesar da limitação ainda presente, é uma das secretarias estaduais que melhor detalha os eventos criminais. Logo, a escolha do estado se deu pelo seu porte e pela disponibilidade das informações, objetivando detectar associações entre cidades e crimes, *rankings* de perigosidade, padrões e modelos replicáveis em outros estados, bem como promover maior transparência, visando auxiliar o processo de apoio às tomadas de decisões estratégicas e operacionais dos governantes e agentes da lei, no combate efetivo da criminalidade. Ainda neste contexto, este trabalho contribui para a produção e disponibilidade de bases de dados abertas e consistentes pelas Secretarias Estaduais, inclusive para o melhoramento dos dados da Sejusp/MG, bem como para que outras organizações, cidadãos e pesquisadores produzam e possuam modelos para automatização da extração, produção de conhecimento e publicação atualizada constantemente sobre dados abertos criminais, intensificando as suas

atuações como controladores sociais.

Para alcançar esse fim, o restante deste artigo está organizado da seguinte forma. Na seção 4.2, os trabalhos relacionados sobre o tema são apresentados. Na seção 4.3, a metodologia adotada é abordada. A seção 4.4 descreve alguns conceitos básicos necessários para o entendimento deste trabalho. Na seção 4.5, a definição e o planejamento do experimento são apresentados. A seção 4.6 detalha a operação do experimento. Na seção 4.7, os resultados são analisados. E, finalmente, na seção 4.8, a conclusão e os trabalhos futuros são apresentados.

4.2 Trabalhos Relacionados

De acordo com (SINGH; JOSHI, 2018), as técnicas de *Data Mining* têm se mostrado eficazes na análise de conjuntos de dados e na coleta de informações úteis em muitos domínios. No campo criminal, a mineração de dados está recebendo maior atenção para descobrir padrões subjacentes nos dados sobre crimes. A seguir, são descritos alguns trabalhos que utilizaram bases de dados que, minimamente, contêm informações semelhantes às disponibilizadas pelo governo de Minas Gerais. Na maioria dos casos, em uma situação diferente do Brasil, o maior detalhamento dos dados fornecido pelos governos permite o uso de mais opções de algoritmos, não explorados neste artigo pela incompletude dos dados disponíveis.

Neste contexto, Agrawal e Sejwar (2017) utilizaram o algoritmo chamado *Multi Objective Particle Swarm Optimization* (MOPSO). Essa nova proposta gera padrões otimizados, utilizando o algoritmo para descoberta de associações *FP-Growth* em bancos de dados criminais. Primeiramente, os padrões foram extraídos por meio da execução do algoritmo *FP-Growth*, o qual tentou constatar os padrões frequentes dos tipos de crimes em relação às cidades (localidade). Em seguida, os padrões otimizados foram gerados, utilizando o MOPSO. Segundo os autores, os resultados evidenciaram que a abordagem proposta é promissora, indicando com que frequência um padrão de crime aparece no banco de dados e auxiliando os analistas de crimes a obter conhecimento sobre a ocorrência de crimes em locais específicos, em menos tempo de execução.

Ainda sobre combate à criminalidade e Regras de Associação (RAs), os pesquisadores de (MARZAN et al., 2017) tiveram como objetivo identificar áreas com maior ocorrência de crimes (*hotspots*) da cidade de Manila, Filipinas. Adicionalmente, usaram o algoritmo *Apriori* para descoberta de padrões frequentes, ajudando os policiais a formar uma ação preventiva. Este trabalho também avaliou vários métodos de previsão de séries temporais, tais como regressão linear, processos gaussianos, multicamada *Perceptron* e *SMOreg*, para prever tendências futuras do crime. Como resultado, o algoritmo multicamada *Perceptron* foi capaz de prever o número de crimes na maioria dos locais em Manila, com mais precisão do que as outras técnicas. De forma semelhante, Yadav Meet Timbadia e Yadav (2017) utilizaram os algoritmos *Apriori*, *K-Means* e *Naive Bayes*, bem como análise de correlação e regressão, para tentar ajudar os especialistas criminais a descobrir padrões, tendências, realizar previsões, encontrar relacionamentos e

possíveis explicações, mapear redes criminosas e identificar suspeitos. De acordo com os autores, o modelo desenvolvido reduzirá os crimes e auxiliará o processo de detecção de crimes de várias maneiras. O trabalho aqui apresentado, dentro das possibilidades que os dados abertos proporcionam, também utilizou o algoritmo *Apriori* para descoberta de associações e produziu um *ranking* de perigosidade para as cidades de Minas Gerais.

Como este trabalho, apresentou dados sobre diversos furtos que ocorrem em Minas Gerais, em (CARAZZA; NETO; EMANUEL, 2020), os pesquisadores investigaram os efeitos da adoção de uma política de toque de recolher juvenil, implantada por diversas cidades do estado de São Paulo. Foi utilizada a estimativa "*difference-in-differences*", para comparar seu impacto em relação aos municípios que não adotaram tal medida. Os resultados indicaram uma redução de 17,9% da taxa de furto, nas cidades que implementaram a política.

No que diz respeito a trabalhos nacionais sobre criminalidade e regionais sobre dados criminais de Minas Gerais, independente dos procedimentos usados terem sidos totalmente automatizados, é preciso ficar atento às bases científicas locais e a alguns trabalhos mais antigos. Por exemplo, na nossa Revisão Sistemática da literatura publicada, a faixa utilizada foi dos últimos dez anos e a prioridade foram as maiores bases de dados de pesquisa do mundo, tal como a Scopus, a qual não indexa muitos dos trabalhos nacionais sobre crimes. Isto pode ser um indicativo da necessidade de maior renovação dos trabalhos ou de mais publicações em outras línguas, nesta área específica de automação da produção de conhecimento. Além disso, o foco desta pesquisa foi a busca por trabalhos que utilizaram dados abertos criminais governamentais e que, de alguma forma, automatizaram a coleta, extração, transformação, geração e publicação de conhecimento sobre estes dados, dentro da limitação do que é disponibilizado.

Não obstante não terem sido encontrados trabalhos relacionados dentro do escopo da Revisão Sistemática efetuada, fonte de entrada para esta pesquisa, em outros indexadores, tais como Scielo e DOAJ, é possível encontrar trabalhos que possuem uma relação indireta com a proposta deste trabalho, mas são de grande relevância para o relacionamento dos resultados aqui expostos com fatores econômicos, sociais e demográficos. Neste contexto, mais recentemente, destacam-se dois trabalhos. O trabalho de Barros et al. (2019) analisou a relação entre a taxa de homicídios e o nível de desenvolvimento econômico dos municípios brasileiros, observando-se que regiões, com alto nível de desenvolvimento, tendem a ser cercadas por municípios com baixo índice de criminalidade. No entanto, para alguns municípios, o nível de desenvolvimento econômico não é capaz de barrar o avanço do crime.

Com uma relação mais direta com este trabalho, do ponto de vista da localidade, Ervilha e Lima (2019) verificaram em que magnitude, e relevância, variáveis socioeconômicas e demográficas específicas afetaram as taxas de criminalidade e suas variações delituosas, nos municípios do estado de Minas Gerais, no período de 2000 a 2014. As análises foram realizadas por meio da modelagem econométrica de dados em painel e evidenciaram que políticas de combate à criminalidade devem ser conjugadas com outras políticas públicas relacionadas à

educação e assistência social, considerando a faixa etária e a vulnerabilidade socioeconômica da população. Estes trabalhos podem ser usados como base para inferências sobre os níveis de perigosidade e associações encontradas.

4.3 Metodologia

A metodologia adotada para o trabalho envolveu, inicialmente, uma Revisão Sistemática (RS) quantitativa da literatura, publicada em (PRADO et al., 2020), tendo por finalidade encontrar o estado da arte das pesquisas sobre análise inteligente de dados abertos governamentais relacionados a incidentes criminais. Para operacionalizar a revisão, acessamos a base Scopus, a qual inclui buscas em diferentes bancos de dados científicos (IEEE, ACM, Elsevier e outros), considerando publicações dos últimos dez anos.

Ato contínuo, para realização do objetivo principal desta pesquisa, foram utilizados os dados abertos criminais disponibilizados pela Sejusp/MG. Todavia, algumas dessas informações são fornecidas em arquivos muito grandes, com dados dispersos, limitando o entendimento do cidadão com relação ao significado ou à importância dos dados abertos. Desta forma, este trabalho permitiu e permitirá fazer a transição dos dados brutos do governo para informações estruturadas, perfazendo o *download* do(s) arquivo(s), nos formatos CSV, bem como a leitura, interpretação do conteúdo e armazenamento numa base de dados estruturada. A seção 4.4.5 ilustra uma visão geral da arquitetura utilizada para realizar as etapas, que vão desde o *download* do *dataset* até a detecção dos padrões.

Do ponto de vista da classificação, em Computação, especificamente na *Data Science*, os objetos de estudo são dados e algoritmos, os quais foram analisados em laboratório, ambiente controlado, com a averiguação e o teste de hipóteses, antes da publicação da aplicação. Desta forma, este estudo pode ser classificado como experimental, pela ocorrência de testes de hipóteses (JURISTO; MORENO, 2013) (BLACKBURN, 2016), e tem características de um estudo "*in vitro*", pela experiência em laboratório, do ponto de vista da Computação. Por outro lado, não pode ser classificado como "*in virtuo*" ou "*in silico*", pois não há simulação computacional de pessoas e/ou de outros aspectos do mundo real (TRAVASSOS; BARROS, 2003) (WOHLIN et al., 2012).

Neste contexto, este trabalho conduziu um experimento, o qual tem o seu método descrito de forma autocontida, no seu planejamento, detalhado na seção 4.5, com 3 etapas macro: (1) identificação e *download* dos arquivos nos formatos CSV; (2) construção de programas ETL (*Extract, Transform, Load*) para as cargas dos dados e da base de dados para tratamento, estruturação e armazenamento das informações contidas nos arquivos; (3) seleção, exploração, análise, testes de hipóteses e validação das informações oriundas dos dados estruturados.

4.4 Base Conceitual

4.4.1 Transparência Pública e Dados Abertos

A transparência na gestão pública é um mecanismo de controle social, associado ao princípio constitucional da publicidade, que prevê obrigatoriedade na divulgação das contas públicas em governos democráticos e estabelece um conjunto de aspectos que sugerem a existência de políticas, procedimentos e tecnologias que proporcionem acesso, uso, qualidade, compreensão e auditabilidade de processos e informações (ALBUQUERQUE et al., 2016)(NASCIMENTO, 2019).

Segundo (FRAGA et al., 2019), o acesso dos cidadãos e órgãos fiscalizadores às atividades desenvolvidas pela administração pública é de grande relevância, pois em algumas situações, os gestores tomam decisões em benefício próprio, em detrimento de toda a sociedade, o que acaba por influenciar nos resultados econômicos e sociais. Dessa forma, torna-se importante que ocorra transparência de informações entre as partes, para que os cidadãos possam visualizar as ações realizadas pelos governantes e tenham informações suficientes sobre a aplicação de suas contribuições.

Neste mesmo contexto, o conceito de dados abertos baseia-se no fato de que existem certas informações cujo acesso deve estar acima de *copyright*, patentes, censura ou qualquer outro acesso privado (BERTOT et al., 2014). De acordo com Hardy e Maurushat (2017), dados abertos são dados acessíveis gratuitamente ou a um custo mínimo, por qualquer pessoa, sendo reutilizados para qualquer finalidade. Os governos comprometidos com o movimento de dados abertos estão postando milhares de conjuntos de dados em portais *online*. O objetivo de liberar esses dados não é meramente fornecer informações ao público, mas impulsionar a inovação por meio da análise destes "grandes repositórios de dados". Portanto, os governos, ao "abrirem" seus *datasets*, permitem que diversas empresas, pesquisadores e o público em geral possam extrair novas informações (*insight*) desse "mar de dados" e contribuir com soluções inovadoras para políticas complexas.

No contexto brasileiro, o portal da Transparência (TRANSPARÊNCIA, 2020) e o portal de Dados Abertos (DADOS, 2020) do Governo Federal são ferramentas que funcionam como um grande catálogo que facilita a busca e o uso de todo e qualquer tipo de dado publicado pelos órgãos do governo. Informações sobre saúde suplementar, sistema de transporte, segurança pública, indicadores de educação, gastos governamentais, processo eleitoral, programas sociais e outros podem ser facilmente encontradas. Com isso, abre-se um canal direto entre o cidadão e governo, com o intuito de melhorar a utilização dos dados, proporcionando o fortalecimento do processo democrático e melhorando a qualidade de vida da população.

Na esfera criminal, o Governo Federal instituiu, pela Lei nº 12.681, de 4 de Julho de 2012, o Sistema Nacional de Informações de Segurança Pública, Prisionais e de Rastreabilidade de Armas e Munições, de Material Genético, de Digitais e de Drogas (SINESP). Este sistema é um portal de

informações integradas, possibilitando consultas operacionais, investigativas e estratégicas, sobre drogas, segurança pública, justiça, sistema prisional, entre outras, implementado em parceria com os entes federados (DADOS, 2020). Considerando que os dados do SINESP são muito gerais, nesta mesma linha, alguns estados passaram a disponibilizar, normalmente, por meio das suas secretarias de segurança pública, informações mais detalhadas, relacionadas aos incidentes criminais ocorridos em seus territórios. Dessa forma, utilizamos para este experimento, os dados criminais de Minas Gerais. A escolha pelo conjuntos de dados desse estado é resultado de uma busca minuciosa de *datasets* estaduais disponibilizados de forma *online* na *web*, a qual constatou que esta Unidade Federativa, até o momento da pesquisa, apresentou, em relação aos outros estados, uma melhor qualidade na disponibilização e composição dos seus *datasets*, os quais são publicados em formatos abertos (arquivos .csv), padronizados, contextualizados e podem ser lidos facilmente por máquinas. Adicionalmente, ficou constatado que o número de *datasets* (anos) disponíveis, relacionados a dados criminais, era maior em relação aos demais estados, o que nos possibilitou realizar uma análise mais robusta.

4.4.2 Região Integrada de Segurança Pública (Risp)

A Região Integrada de Segurança Pública (Risp) é a instância responsável pelo planejamento estratégico das áreas de coordenação integrada da segurança pública, que estabelece as diretrizes de enfrentamento da criminalidade, a partir da promoção da articulação entre as polícias civil e militar (SEJUSP, 2020). Para Maciel et al. (2019), Risp é uma divisão geográfica que permite a articulação e integração regional, no nível tático e operacional, dos órgãos competentes da pasta de segurança pública, para realizar o planejamento, controle, supervisão, avaliação e monitoramento corretivo das atividades de segurança pública. Normalmente, uma Rip é constituída por um ou mais municípios.

Atualmente, as Risps encontram-se constituídas em diversos estados brasileiros. No contexto de Minas Gerais, foco do nosso trabalho, o governo estadual assinou uma resolução conjunta, no dia 15 de fevereiro de 2008, que instituiu a integração geográfica entre as polícias civil e militar. Na época, tal resolução definiu que a correspondência circunscricional das instituições ficava estruturada em 16 Risps, em todo território estadual (SEJUSP, 2020). Com o passar dos anos, mais 3 regiões foram criadas, totalizando, atualmente, 19 regiões integradas. A Figura 33 mostra um mapa territorial de Minas Gerais que detalha a distribuição atual das Regiões Integrada de Segurança Pública.

4.4.3 Regras de Associação

Conhecida como Análise de Cesta de Mercado (*Market Basket Analysis*), as Regras de Associação (RAs) são técnicas de mineração de dados que representam combinações de itens que ocorrem com determinada frequência em uma base de dados. Segundo (MARZAN et al., 2017), uma das abordagens de mineração de dados mais conhecidas para encontrar conjuntos de itens

Figura 33 – Divisão territorial por Risp no estado de Minas Gerais.

Fonte: Diagnóstico da Violência Doméstica e Familiar contra a Mulher em Minas Gerais (2013-2015).

frequentes e gerar regras de associação é o algoritmo *Apriori*. Isso auxiliou os pesquisadores a resolverem problemas sobre como encontrar padrões ocultos em análises criminais. Vale ressaltar que este experimento utilizou o algoritmo *Apriori* para a detecção das regras de associação.

De maneira formal, Agrawal, Imieliński e Swami (1993) definiram as regras de associação da seguinte forma: Sejam $I = i_1, i_2, ..., i_m$ um conjunto de m itens distintos e D uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens (*itemset*), tal que $T \subseteq I$. Uma regra de associação é uma expressão na forma $A \Rightarrow B$, onde $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ e $A \cap B = \emptyset$. A é denominado antecedente e B denominado consequente da regra. Tanto o antecedente quanto o consequente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens. Como exemplo, A e B podem ser produtos ou eventos, então, poderíamos ter a regra, "quem compra o produto A também compra o produto B", ou a regra, "quando ocorre um roubo (A) também ocorre um assassinato (B)".

Existem várias medidas de interesse que avaliam regras de acordo com as restrições iniciais do usuário. Diversos pesquisadores propõem medidas que objetivam extrair um padrão específico dos dados (CAMPOS, 2018). Em nossa pesquisa, utilizamos as seguintes medidas de interesse:

- **Suporte:** Suporte: O suporte de um conjunto de itens Z, *Sup(Z)*, representa o percentual de transações da base de dados que contêm os itens do conjunto Z. O suporte de uma regra de associação $A \Rightarrow B$, *Sup* $(A \Rightarrow B)$, é dado por *Sup* $(A \cup B)$, e pode ser visto na Equação

4.1.

$$Sup(A \Rightarrow B) = \frac{Transações\ que\ contêm\ A\ e\ B}{Total\ \ de\ Transações} \tag{4.1}$$

- **Confiança:** A confiança da regra $A \Rightarrow B$, $Conf\,(A \Rightarrow B)$, representa, dentre as transações que contêm A, a porcentagem de transações que também contêm B, é dada por $Conf\,(A \Rightarrow B) = Sup\,(A \cup B) \div Sup\,(A)$, e pode ser vista na Equação 4.2.

$$Conf(A \Rightarrow B) = \frac{Número\ total\ de\ transações\ que\ contêm\ A\ e\ B}{Total\ de\ transações\ que\ contêm\ A} \tag{4.2}$$

- **Lift:** A medida de interesse *Lift*, também conhecida como *Interest*, é utilizada para indicar, dado uma associação $A \Rightarrow B$, o quanto mais frequente torna-se B quando ocorre em conjunto com A. O *Lift* de uma regra de associação $A \Rightarrow B$ é dado pela Equação 4.3. Quando $Lift(A \Rightarrow B) = 1$, significa que A e B são independentes, ou seja, não existe associação entre eles. Se $Lift(A \Rightarrow B) > 1$, então A e B são positivamente dependentes. Se $Lift(A \Rightarrow B) < 1$, então, a regra é enganosa e A e B são negativamente dependentes.

$$Lift(A \Rightarrow B) = \frac{Conf(A \Rightarrow B)}{Sup(B)} \tag{4.3}$$

- **Count:** Representa a frequência da ocorrência de um determinado conjunto de itens.

- **Coeficiente de correlação de *Pearson (r)*:** É uma medida de associação bivariada (força) do grau de relacionamento entre duas variáveis. O coeficiente de correlação *Pearson (r)* varia de -1 a 1. O sinal indica a direção da correlação (negativa ou positiva), enquanto que o valor indica a magnitude. Uma correlação positiva (r=1), por exemplo, indica que quando X aumenta, Y também aumenta, ou seja, valores altos de X estão associados a valores altos de Y (PARANHOS et al., 2014).

- **Qui-Quadrado (χ^2):** É um teste não paramétrico utilizado para avaliar a correlação entre vários itens. Nas regras de associação, é utilizado para testar a independência entre os itens das regras (CAMPOS, 2018). O valor crítico da distribuição Qui-Quadrado, com 1 grau de liberdade (tabela de contingência 2x2), em α=0,05, é 3,84; quanto maior o valor do Qui-Quadrado, mais provável é a correlação das variáveis. O nível de significância é indicado pelo valor do *p-value*, um *p-value* abaixo de 0,05 indica que há dependência entre os itens (WU et al., 2016).

4.4.4 Definição do *Ranking* de Perigosidade

Um grande problema encontrado na análise criminal é quantificar quanto um determinado local é mais perigoso que outro. Isso torna-se ainda mais problemático quando envolvemos diversos tipos de crimes dentro da mesma análise. Diariamente, diferentes regiões são alvos de

diversos tipos de crimes e em várias proporções. Então, como é possível determinar que uma determinada região "*A*" é mais perigosa que uma região "*B*"?.

Neste contexto, (FARIAS et al., 2018) criaram um *ranking* de perigosidade para os bairros de uma cidade brasileira chamada Mossoró, localizada no estado do Rio Grande do Norte. Com o uso dos dados criminais combinado com uma fórmula pré-definida, eles definiram um grau de perigosidade para cada bairro da cidade. Para a definição dessa fórmula, os autores ponderaram cada tipo de crime, a fim de permitir uma comparação entre os diferentes bairros. Esta ponderação foi definida usando o bom senso e a opinião de especialistas. Por exemplo, é senso comum que um homicídio é um crime mais grave que um roubo. A Tabela 19 mostra os pesos que foram definidos para cada crime.

Tabela 19 – Pesos dos crimes.

Tipo de crime	**Peso**
Furto	1,0
Roubo	2,0
Homicídio	3,0
Tráfico de drogas	4,0

Fonte: (FARIAS et al., 2018).

Baseada nos pesos acima, o Índice de Criminalidade (*IC*) para cada bairro foi calculado por meio de uma média ponderada, ou seja, o somatório do produto entre o peso e o número de incidentes de cada tipo de crime, divido pela soma dos pesos. A Equação 4.4 mostra a fórmula utilizada, onde n representa o número de tipos de crime, P_i o valor do peso e N_i o total de ocorrências daquele tipo de crime.

$$IC = \frac{\sum_{i=1}^{n} P_i \times N_i}{\sum_{i=1}^{n} P_i} \tag{4.4}$$

Após o computo do *IC*, o Índice de Criminalidade Normalizado (*ICN*) foi obtido, utilizando o processo de normalização no intervalo [0,1] (LIMA; VIGNATTI; SILVA, 2020). Então, os bairros foram classificados do menos para o mais perigoso, onde *ICN=0* significa o menos perigoso e *ICN=1* representa o mais perigoso. Porém, a fim de permitir uma classificação mais interpretável e clara, os autores transformaram o *ICN*, que são representados por valores contínuos, em níveis de perigosidade representados por valores categóricos. A definição desses níveis foi realizada da seguinte forma:

- Não Perigoso: *0,00 <= ICN < 0,15*;
- Pouco Perigoso: *0,15 <= ICN < 0,30*;
- Perigoso: *0,30 <= ICN < 0,50*;

- Muito Perigoso: *0,50 <= ICN <= 1,00*.

Em nosso trabalho, foi utilizada a mesma estratégia adotada por (FARIAS et al., 2018) para encontrar os níveis de perigosidade. Entretanto, a fim de obter resultados mais precisos, alguns ajustes pontuais no processo foram realizados. Primeiramente, percebe-se que a Equação 4.4 utiliza o número de incidentes criminais. De certa forma, isso favorece regiões (bairros) menos populosas, pois locais com maior concentração populacional podem ter mais crimes. Logo, locais mais populosos poderiam alcançar os maiores *IC* por este fator e não necessariamente pelos níveis de segurança e criminalidade. Para mitigar essa ameaça, foi considerada a estimativa populacional anual no cálculo do *IC*. Ao invés de utilizar diretamente o número de ocorrências criminais, **foi utilizada a taxa de crimes por 100.000 habitantes. Este procedimento atenuará as discrepâncias existentes, trazendo todas as regiões analisadas para o mesmo patamar, ao diminuir o potencial de confundimento com a população (fator de confusão)**. A Equação 4.5 representa o cálculo para encontrar a taxa de crimes por 100.000 habitantes (*TC_HAB*). A estimativa populacional anual de cada estado foi obtida no portal do Instituto Brasileiro de Geografia e Estatística (IBGE), localizado no site https://ibge.gov.br.

$$TC\_HAB = \frac{Número\ \ de\ \ Crimes}{Estimativa\ Populacional} \times 100.000 \tag{4.5}$$

A Equação 4.6 mostra a Equação 4.4 adaptada, utilizando a taxa de crimes por 100.000 habitantes em seu computo.

$$IC_{adaptado} = \frac{\sum_{i=1}^{n} P_i \times TC\_HAB_i}{\sum_{i=1}^{n} P_i} \tag{4.6}$$

Um outro ponto ajustado foram as faixas dos níveis de criminalidade. Diferente do trabalho de (FARIAS et al., 2018), foram adotadas cinco faixas de perigosidade (Baixíssimo, Baixo, Intermediário, Alto ou Altíssimo), com o intuito de também mensurar um nível intermediário. Neste contexto, é importante ressaltar que o *ICN* sempre indicará o nível de perigosidade relativo ao estado de Minas Gerais e ao ano analisado, ou seja, é um *ranking* de perigosidade de Minas Gerais, pois, ainda que os municípios obtivessem níveis de criminalidade desejáveis e compatíveis com os lugares mais seguros do mundo, sempre haverá um município ocupando a posição Altíssimo, representando a pior posição em relação ao resto do estado. Desta forma, os níveis adotados por este artigo foram:

- Baixíssimo: *0,00 <= ICN <= 0,15*;
- Baixo: *0,15 < ICN <= 0,40*;
- Intermediário: *0,40 < ICN <= 0,60*;
- Alto: *0,60 < ICN <= 0,85*;

- Altíssimo: *0,85 < ICN <= 1,00*.

Com intuito de otimizar o *ranking* aqui desenvolvido, vale ressaltar que quatro especialistas em segurança pública foram consultados, para ponderar os novos tipos de crime e calibrar os pesos existentes descritos na Tabela 19. Os especialistas trouxeram as visões da Polícia, do Ministério Público, do Judiciário e dos advogados, com as perspectivas dos crimes de um Coordenador de um Grupo de Combate ao Crime Organizado, de um Promotor, de um Juiz e de um Advogado Criminalista. Os desempates do debate consideraram sempre os intervalos das penas, sem majorantes. Desta forma, para que não houvesse arbitrariedade nos pesos, estes intervalos foram considerados.

O debate dos pesos dos crimes é amplo e profundo, o qual não tem e talvez nunca terá uma conclusão simples, uma vez que as penas nem sempre refletirão a costumaz gravidade de um crime e as opiniões divergem veementemente quando os crimes lidam com vítimas fatais. Neste sentido, apenas a realização de um *Survey* com juristas e especialistas em segurança pública de todo o Brasil poderá estimar o que a sociedade espera e como considera os níveis de gravidade dos crimes, coletando as medianas de pesos atribuídos pelos entrevistados e ampliando a discussão sobre o que pode ser considerada uma localidade mais perigosa e sobre as penas dos crimes. Isto também poderá embasar o poder público e os legisladores, em busca de um código penal que reflita melhor o clamor da sociedade brasileira. A árdua e complexa pesquisa de campo será um trabalho futuro do grupo de pesquisa ao qual os autores deste artigo pertencem.

Em linhas gerais, **foram consideradas as médias dos intervalos das penas dos tipos de crimes, sem majorantes, bem como a uniformidade de distâncias entre os pesos (considerando uma aproximação proporcional às distâncias entre as penas)**. Pensando de maneira simplista, poderia ser considerado, para todos os crimes, apenas a primeira pena, sem majorantes, todavia, infelizmente, os dados abertos fornecidos pelo governo não tipificam detalhadamente os crimes. Em outras palavras, não são contemplados os casos em que o legislador criou um subtipo do crime, por entender que há uma maior reprovabilidade da conduta, ou seja, essa maior reprovabilidade, diferentemente da majorante, faz com que a pena base deixe de ser a do caput para iniciar-se já em um valor maior, análogo a uma nova tipificação. Para estes casos, foi usada a média do intervalo da pena do caput e a média dos intervalos das penas dos subtipos principais, com maior reprovabilidade. No caso de Homicídio, por exemplo, não é possível distinguir entre os Homicídios Culposos e Dolosos, ou seja, não é razoável considerar a menor pena, referente ao Homicídio Culposo.

Exemplificando, considerando as penas de Furto e Roubo, temos o intervalo de pena, para Roubo, sem majorantes, de 4 a 10 anos, bem como temos o intervalo de pena, para Furto, também sem majorantes, de 1 a 4 anos, e, para Furto Qualificado (maior reprovabilidade), de 2 a 8 anos (http://www.planalto.gov.br/ccivil_03/decreto-lei/del2848compilado.htm). Considerando a média 7 do intervalo para Roubo e a média da média dos intervalos de Furto, de 3,75, temos

uma razão de 1,87, aproximadamente 2, mesma razão/proporção utilizada entre os pesos de Furto Consumado e Roubo Consumado. Após essas validações, os novos tipos de crime e os pesos adotados estão descritos na Tabela 20.

Tabela 20 – Pesos dos crimes utilizados neste trabalho.

Tipo de crime	**Peso**
Extorsão Tentado	1,0
Furto Consumado	1,0
Roubo Tentado	1,0
Sequestro e Cárcere Privado Tentado	1,0
Estupro Tentado	1,5
Lesão Corporal Consumado	1,5
Sequestro e Cárcere Privado Consumado	1,5
Estupro de Vulnerável Tentado	2,0
Extorsão Consumado	2,0
Homicídio Tentado	2,0
Roubo Consumado	2,0
Estupro Consumado	3,0
Estupro de Vulnerável Consumado	3,5
Extorsão Mediante Sequestro Consumado	3,5
Homicídio Consumado	4,0

Fonte: Elaborada pelo autor.

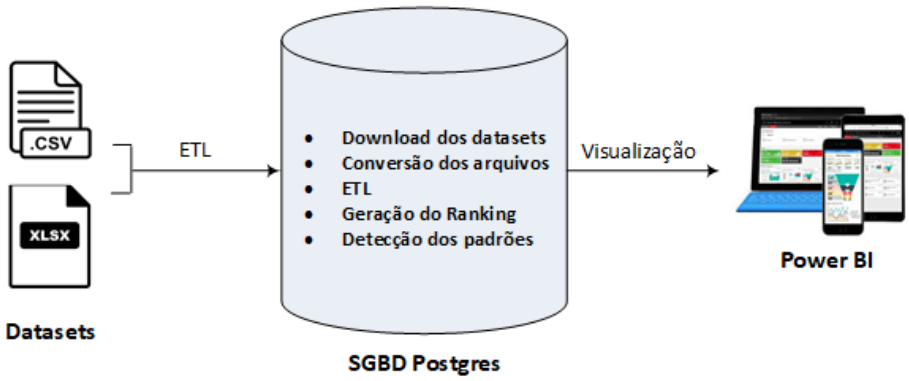
4.4.5 Visão Geral da Arquitetura

O *PostgreSQL* é um Sistema de Gerenciamento de Banco de Dados (SGBD) que permite que as funções definidas pelo usuário sejam escritas em outras linguagens além de SQL e C. Estas linguagens são chamadas genericamente de Linguagens Procedurais (LPs). No caso de uma função escrita em uma Linguagem Procedural (LP), o servidor de banco de dados não possui nenhum conhecimento interno sobre como interpretar o texto do código fonte da função. Em vez disso, a tarefa é passada para um tratador especial que conhece os detalhes da linguagem. O próprio tratador pode fazer todo o trabalho (análise gramatical e sintática, execução, etc) ou pode servir como um "elo de ligação" entre o *PostgreSQL* e a implementação existente de uma linguagem de programação (POSTGRESQL, 2019).

Para facilitar a obtenção, o tratamento, a manipulação e a análise dos dados realizados por esta pesquisa, uma arquitetura centralizada (unificada) foi desenvolvida, tendo como ponto principal o *PostgreSQL*. Dessa forma, procurou-se automatizar ao máximo os processos existentes, que vão desde a obtenção dos *datasets* até a detecção dos padrões criminais. Importante ressaltar que a arquitetura desenvolvida utilizou a ferramenta *Power BI* da *Microsoft* para a apresentação gráfica dos resultados. A Figura 34 exibe uma visão geral da arquitetura desenvolvida.

Em nosso trabalho, três LPs foram utilizadas, são elas: *PL/PgSQL*, *PL/Python* e *PL/R*. A *PL/PgSQL* é uma LP nativa do *PostgreSQL*, a qual foi utilizada, principalmente, nos procedimentos de Extração, Transformação e Carga (ETL) dos dados e geração do *ranking* de perigosidade.

Figura 34 – Visão geral da arquitetura.



Fonte: Elaborada pelo autor.

Já a *PL/Python* permite a implementação de funções utilizando a linguagem *Python*, as quais poderão ser executadas dentro do *PostgreSQL*, sendo útil para realizar os *downloads* dos *datasets*, conversão, remoção e descompactação dos arquivos. E, por fim, a *PL/R*, a qual foi utilizada para detectar padrões de associação com o algoritmo *Apriori*, fazendo uso de códigos escritos na linguagem R.

4.5 Definição e Planejamento do Experimento

Nesta e nas duas próximas seções, este trabalho é apresentado como um processo experimental. O mesmo segue as diretrizes de Wohlin (WOHLIN et al., 2012) e processos experimentais com publicações recentes (SANTOS; COLAÇO JÚNIOR; SOUZA, 2018) (SANTOS et al., 2020). Esta seção focará na definição do objetivo e planejamento do experimento.

4.5.1 Definição dos Objetivos

O objetivo principal deste experimento é analisar as possíveis associações entre municípios, Risps, tipos de crime, alvos de furto e alvos de roubo, utilizando os dados disponibilizados pelo governo do Estado de Minas Gerais, por meio da Secretaria de Estado de Justiça e Segurança Pública (Sejusp). Para atingi-lo, conduziu-se um experimento em ambiente controlado, no qual foram verificadas regras de associação (vide seção 4.4.3), utilizando o algoritmo *Apriori*, para determinação das combinações de itens que ocorrem com determinada frequência, bem como para a geração das medidas de interesse que, estatisticamente, puderam servir de base para medir a força de tais regras. Adicionalmente, *rankings* de perigosidade dos municípios e das Risps também foram desenvolvidos.

Baseado no modelo GQM (*Goal Question Metric*) apresentado em (BASILI; WEISS, 1984)(BASILI et al., 2014), segue a formalização do objetivo desse trabalho: **Analisar** as ocorrências criminais no estado de Minas Gerais, **com o propósito** de avaliá-las, **com respeito**

à detecção de padrões criminais, **do ponto de vista de** cientistas de dados, analistas criminais e cidadãos, **no contexto** de dados abertos da Sejusp/MG.

4.5.2 Planejamento

Formulação das Hipóteses - Não foram encontrados estudos experimentais que analisaram as possíveis associações entre municípios, Risps, tipos de crime, alvos de furto e alvos de roubo. Além disso, também não foram encontradas pesquisas científicas que descrevessem os níveis de perigosidade das Risps ou dos municípios mineiros. Baseadas nestas premissas, oito questões de pesquisa foram formalizadas para esse trabalho. Para as duas primeiras questões, **Q1** e **Q2**, serão realizadas análises estatísticas descritivas, enquanto que, para as questões **Q3**, **Q4**, **Q5**, **Q6**, **Q7** e **Q8** serão aplicados testes de significância das hipóteses. As questões de pesquisa são:

- **Q1**: Quais os municípios vêm se destacando como os mais perigosos?
- **Q2**: Quais as Risps mais perigosas em 2019?
- **Q3**: Há associações entre tipos de crime e municípios?
- **Q4**: Há associações entre tipos de crime e Risps?
- **Q5**: Há associações entre alvos de roubo e municípios?
- **Q6**: Há associações entre alvos de roubo e Risps?
- **Q7**: Há associações entre alvos de furto e municípios?
- **Q8**: Há associações entre alvos de furto e Risps?

Para responder às questões **Q3** a **Q8**, as seguintes hipóteses resumidas serão testadas individualmente, para combinação de antecedente e consequente:

- H_0: Os tipos de crime, alvos de roubo e alvos de furto são independentes dos municípios e das Risps.
- H_1: Os tipos de crime, alvos de roubo e alvos de furto são dependentes dos municípios e das Risps.

Seleção do Contexto - Para a realização do experimento, foram utilizados os conjuntos dos dados criminais de 853 municípios do estado de Minas Gerais.

Seleção dos Participantes e Objetos - Foram selecionados todos dados disponibilizados até o início dessa pesquisa. No total, 6 *datasets* foram obtidos, os quais foram distribuídos em três

categorias de dados: “Ocorrências criminais”, “Alvos de roubo” e “Alvos de furto”. Cada uma dessas categorias é constituída por 2 conjuntos de dados. A categoria “Ocorrências criminais” é composta pelos dados de crimes violentos e crimes de outras naturezas. Já “Alvos de roubo” é formada pelas informações de roubos de veículos e roubos a outros alvos (estabelecimentos comerciais, residências, transporte coletivo, cargas e transeuntes). Por fim, o grupo “Alvos de furto” é formado pelos dados de furtos de veículos e furtos a outros alvos (estabelecimentos comerciais, residências, transporte coletivo, cargas e transeuntes). Tais dados englobam as informações do período de Janeiro de 2012 a Dezembro de 2019, as quais foram obtidas no portal de dados abertos da Secretaria de Estado de Justiça e Segurança Pública de Minas Gerais (Sejusp/MG), localizado em (SEJUSP, 2020).

Variáveis Dependentes - As variáveis dependentes abordadas no experimento, para validação das hipóteses, foram as frequências dos conjuntos de itens analisados e as regras geradas, com seus Suportes e Confianças, das quais podem ser derivadas outras medidas de interesse objetivas para auxiliar na identificação das forças destas regras de associação: *Lift*, *r* (Coeficiente de correlação de *Pearson*) e Qui-Quadrado (χ^2), com seu nível de significância (*p-value*).

Variáveis Independentes - Para este trabalho, foram consideradas, como variáveis independentes: o conjunto de registros compilados e disponibilizados em arquivos, contendo os dados dos incidentes criminais ocorridos nos municípios mineiros, o algoritmo *Apriori* utilizado, bem como Suporte e Confiança mínimos, intervalos aceitáveis de *Lift* e *p-value* máximo.

Instrumentação - O processo de instrumentação teve início com a configuração do ambiente para o experimento, planejamento de coleta de dados, construção de ETL (programa de Extração, Transformação e Carga) e o desenvolvimento dos algoritmos necessários. Os materiais/recursos utilizados foram:

- Arquivo com os incidentes criminais, disponibilizado pela Sejusp/MG;
- Arquivo com as estimativas populacionais dos munícipios, disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE);
- Arquivos com o *script* de criação do projeto de banco de dados;
- Banco de Dados *PostgreSQL*, versão 11.6-3, para armazenar os dados e realizar o processo de ETL, com o uso da linguagem nativa *PL/PgSQL*;
- *Python*, versão 3.7.6;
- Software livre *R*, versão 3.6.0;
- Ferramenta *Power BI*.

4.6 Operação do Experimento

4.6.1 Execução

Os *datasets* disponibilizados pela Sejusp/MG estão sumarizados, ou seja, cada registro informa a quantidade de incidentes criminais ocorridos em um determinado mês, município e tipo de crime. Dessa forma, foi necessário realizar o desmembramento dessas informações, para que cada registro representasse apenas um crime, ou seja, passando a configurar uma transação, a qual é exigida neste formato, no R, e pôde ser avaliada pelo algoritmo *Apriori*. O nome transação representa a ligação existente em cada ocorrência de itens, no nosso caso, a ligação ocorre de forma geográfica e temporal, quando crimes ocorrem em uma determinada cidade, em determinado período. Por exemplo, para a avaliação das associações significativas entre Municípios e Alvo de Roubo, cada linha da nova tabela anual criada possui um par atômico de itens, tal como em (Belo Horizonte, Transeunte), representando cada crime ocorrido em um determinado ano.

De posse dessas transações, o *Apriori* verifica os pares que ocorrem de forma frequente no banco de dados. Essas associações frequentes possuirão um número de ocorrências, em relação ao número total de transações do banco de dados, maior ou igual ao suporte mínimo. Em seguida, a associação será considerada válida, se o seu número de ocorrências (Belo Horizonte, Transeunte), em relação ao número de vezes que o antecedente (Belo Horizonte) ocorrer, for superior à confiança mínima. Por fim, se o consequente é algo comum a todos e gerar regras enganosas, estas serão excluídas pela medida de interesse (*Lift*), pois possuirão um *Lift* menor que zero. Todas essas etapas, bem como a apresentação dos resultados em painéis gráficos na *web* foram automatizadas.

4.6.1.1 Coleta dos dados

Após a execução, com o intuito de facilitar a coleta e análise dos dados, as informações foram publicadas em uma ferramenta gráfica (*Power BI*). Utilizando os diversos recursos deste *framework*, gráficos e tabelas foram construídos, a fim de elucidar as questões de pesquisa propostas.

Com relação às RAs, para cada associação, foi gerado um conjunto de medidas de interesse. Por se tratar de mineração em uma base de dados real, o número de regras gerados foi relativamente alto, como corroborado pelos ensaios apresentados em (ZHENG; KOHAVI; MASON, 2001). Além disso, grande parte destes resultados minerados costuma ser composta por regras óbvias, redundantes ou, até mesmo, contraditórias. Para filtrar as regras interessantes para o estudo, uma confiança mínima de 70% foi adotada, para todas as categorias de dados analisadas (“Ocorrências criminais”, “Alvos de roubo” e “Alvos de furto”). Este valor foi adotado a partir de uma analogia com os intervalos considerados forte (0,70 - 0,89) e muito forte (0,90 - 1,00), para uma correlação (GUIMARÃES et al., 2016) (CHAVES; SHIMIZU et al., 2018).

Assim, analogamente, a confiança mínima de 70% permitirá a geração de regras com confianças fortes e muito fortes.

Em se tratando de suporte mínimo, devido à diferença do número de transações, foram adotados valores individuais para cada categoria. Como visto na Equação 4.1, o valor do suporte de uma RA é dado pela relação entre a quantidade de transações, nas quais aparecem os itens *A* e *B*, e o total de transações. Para o cálculo, adotamos o valor aproximado (para cima) de uma amostra com população infinita, considerando uma margem de erro de 3,5% e 95% de confiabilidade, o que totaliza 784 ocorrências. Desta forma, a quantidade mínima de transações, nas quais aparecem os itens *A* e *B* (crime e município, por exemplo), deve ser de 800 (arredondamento para cima), para o ano em que houve menos transações, ou seja, para a categoria "Ocorrências criminais", cuja a menor quantidade de transações ocorreu em 2019, 405.692 transações, o suporte mínimo foi obtido por meio da divisão, 800/405.692, resultando, aproximadamente, no valor de 0.0019 (0,19%). Nos outros anos, com mais transações, a margem de erro será ainda menor, uma vez que o percentual aplicado foi o mesmo.

De forma semelhante, para a categoria de "Alvos de roubo", cuja a menor quantidade de transações foi de 49.502, em 2019, o suporte mínimo foi obtido por meio da divisão, 800/49.502, resultando, aproximadamente, no valor de 0.0161 (1,61%). E, finalmente, para o conjunto de dados "Alvos de furto", no qual a menor quantidade de transações ocorreu, também, em 2019, o suporte mínimo foi obtido por meio da divisão, 800/187.188, resultando, aproximadamente, no valor de 0.0042 (0,42%).

Para o *Lift*, o valor deverá estar nas seguintes faixas: $Lift > 1$ (significando que existe uma dependência positiva) ou $0 < Lift < 0,1$ (indicando que existe uma dependência negativa considerável). Regras que não estiverem dentro destes limites serão consideradas enganosas e descartadas, pois indicam que, em verdade, o antecedente diminui a probabilidade do consequente ocorrer, ainda que numa pequena proporção.

4.6.1.2 Validação dos dados

Para assegurar a análise, interpretação e validação dos resultados das regras de associação, foram aplicados os testes de significância das hipóteses, utilizando o teste do Qui-Quadrado (χ^2). Além disso, também foi calculado o coeficiente de correlação de *Pearson* (r).

4.7 Análise dos Resultados

4.7.1 Resultados Brutos

Como exposto anteriormente, foi feita uma avaliação que contempla os incidentes criminais de todos os municípios do estado de Minas Gerais. Devido ao espaço reduzido e ao alto número de municípios que Minas Gerias contém (853 municípios), serão apresentadas apenas

as evoluções das taxas de criminalidade (de 2012 a 2019) dos municípios que obtiveram as 10 maiores taxas, no ano de 2019. Como foi descrito, o valor desta taxa foi calculado, utilizando a estimativa populacional de cada localidade, a fim de ser proporcional a 100.000 habitantes, utilizando a Equação 4.5.

Tabela 21 – 10 municípios com as maiores taxas de criminalidade por 100.000 habitantes em 2019 (ordenados da maior para a menor).

Município	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**Variação (2018-2019)**
Belo Horizonte	4092,19	4267,93	4353,28	4664,63	4940,54	4676,13	4103,93	3677,09	-10,40 %
Água Comprida	1538,46	1884,06	2806,00	2228,68	3105,29	3449,95	3192,02	3301,65	3,43 %
Confins	5479,68	6755,05	6537,68	6143,87	6157,37	4963,68	4581,64	3239,23	-29,30 %
Buritis	3715,73	4074,40	4315,45	3108,70	3090,85	3272,71	2846,37	3176,20	11,59 %
Uberaba	3468,67	3794,08	4235,08	4360,72	4429,43	4097,82	3293,67	3055,88	-7,22 %
Campos Altos	2004,72	2345,63	2380,95	2258,66	3453,46	3015,53	2897,89	2923,48	0,88 %
Juatuba	2959,27	3121,01	3596,63	4049,91	4876,42	3806,91	3190,61	2905,81	-8,93 %
Delfinópolis	2198,28	2156,14	1952,25	2393,95	2568,04	3145,00	3043,11	2881,64	-5,31 %
Corinto	4538,39	4039,37	3393,71	3667,32	4109,48	3580,22	3475,23	2878,09	-17,18 %
Rio Novo	1865,63	2285,59	2934,01	2099,45	2900,63	2653,02	2762,55	2815,96	1,93 %

Fonte: Elaborada pelo autor.

A Tabela 21 apresenta a evolução dos 10 municípios com as maiores taxas de criminalidade. A tabela está ordenada, de forma **decrescente**, pelos valores das taxas de 2019. Como pode ser visto, apesar de Belo Horizonte atingir a maior taxa em 2019, nos demais anos (2012 a 2018), o município de Confins liderou o *ranking*. Esta mudança, ocorrida no último ano, pode ser justificada pela queda constante do número de crimes ocorridos em Confins, ao longo dos anos, e, principalmente, pela redução considerável da taxa de criminalidade (29,30%) alcançada em 2019. Além disso, excetuando Água Cumprida, Buritis, Campos Altos e Rio Novo, é notória a redução da taxa de criminalidade em 2019, relativa ao ano anterior, seguindo uma tendência geral de queda em 2018, em menores proporções que em 2019.

Neste contexto, os destaques vão para Confins, Corinto e Belo Horizonte. Nestes municípios, houve uma manutenção da tendência de queda, nos últimos três anos, sendo mais acentuada em 2019. Apesar da queda geral, uma análise detalhada destes resultados é interessante, considerando as ações que foram efetivadas, bem como se estas ações realmente trouxeram efeitos positivos mais significativos para estes munícipios. O raciocínio inverso vale para os municípios de Água Comprida, Buritis, Campos Altos e Rio Novo. No Rio Novo, os valores sobem desde 2018, nas outras cidades, o aumento ocorreu em 2019.

A Tabela 22 apresenta, mais detalhadamente, a taxa de criminalidade, por tipo de crime, dos 10 municípios com as maiores taxas criminais em 2019. Vale a pena destacar que analisando apenas os homicídios e considerando somente as cidades listadas na Tabela 21, os municípios de Corinto e Juatuba assumiriam, respectivamente, a primeira e segunda colocações, merecendo, dessa forma, uma atenção especial por parte das políticas públicas do governo estadual. Adicionalmente, também vale ressaltar o alto número de estupros consumados, i.

Tabela 22 – Taxa dos tipos de crime, dos 10 municípios com as maiores taxas de criminalidade, no ano de 2019, proporcional a 100.000 habitantes.

	Municípios									
Tipo de Crime	**Belo Horizonte**	**Água Comprida**	**Confins**	**Buritis**	**Uberaba**	**Campos** Altos	**Juatuba**	**Delfinópolis**	**Corinto**	**Rio Novo**
Homicídio Consumado	1,43	0,00	0,00	0,00	0,90	0,00	3,71	0,00	4,21	0,00
Extorsão Mediante Sequestro Consumado	8,68	0,00	0,00	4,03	4,19	0,00	11,13	0,00	4,21	0,00
Estupro de Vulnerável Consumado	14,89	0,00	14,86	28,18	19,47	25,87	37,11	14,06	12,64	11,17
Estupro Consumado	37,98	0,00	14,86	32,20	28,76	51,74	44,53	14,06	16,86	22,35
Roubo Consumado	207,76	100,05	118,87	213,36	379,29	245,78	322,87	70,28	181,20	100,57
Homicídio Tentado	0,88	0,00	0,00	0,00	0,90	0,00	0,00	0,00	0,00	0,00
Extorsão Consumado	10,63	0,00	0,00	4,03	5,69	6,47	14,84	0,00	8,43	0,00
Estupro de Vulnerável Tentado	14,65	0,00	14,86	24,15	11,68	12,94	25,98	0,00	12,64	0,00
Sequestro e Cárcere Privado Consumado	0,80	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Lesão Corporal Consumado	676,37	1200,6	297,18	382,43	420,93	763,21	653,16	702,84	741,65	502,85
Estupro Tentado	14,09	0,00	14,86	4,03	6,89	6,47	22,27	0,00	12,64	0,00
Sequestro e Cárcere Privado Tentado	0,08	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
Roubo Tentado	0,88	0,00	0,00	0,00	0,60	0,00	0,00	0,00	0,00	0,00
Furto Consumado	2684,76	2001,00	2763,74	2483,80	2173,57	1811,01	1766,5	2080,40	1879,40	2179,01
Extorsão Tentado	3,22	0,00	0,00	0,00	3,00	0,00	3,71	0,00	4,21	0,00

Fonte: Elaborada pelo autor.

e., “Estupros Consumados” somados aos “Estupros de Vulneráveis Consumados”, ocorridos nas cidades de Juatuba, Campos Altos e Buritis, em 2019. Destacar os estupros é importante,

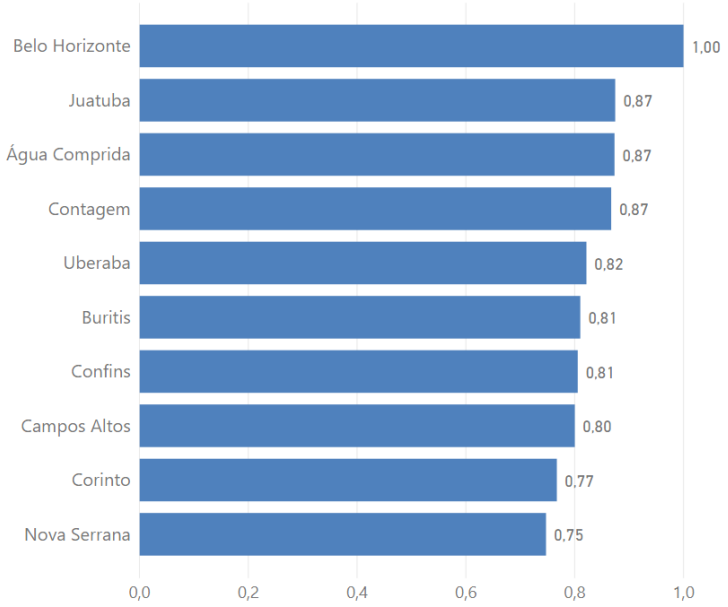
uma vez que a discussão e o questionamento sobre as análises, que consideram a perspectiva única dos assassinatos, indagam a ausência, por exemplo, do registro da barbárie e da natureza hedionda deste tipo de crime. Os homicídios e estupros explicam a posição de Juatuba no *ranking* de perigosidade (vide Figura 35), levando-se em consideração ainda que um estupro pode resultar em morte e que, infelizmente, os dados disponíveis não detalham todos os subtipos de estupro. De qualquer forma, ainda que não tenha sido letal, o estupro pode mudar uma vida ou, metaforicamente, ceifá-la para sempre.

4.7.2 Análise e Interpretação do *Ranking* de Perigosidade

Como relatado na seção 4.4.4, um grande problema encontrado na análise criminal é quantificar quanto um determinado local é mais perigoso que outro. Isso torna-se ainda mais problemático quando envolvemos diversos tipos de crimes dentro da mesma análise. Para responder quais os municípios e Risps vêm se destacando como os mais perigosos (questões **Q1** e **Q2**), um Índice de Criminalidade (*IC*) foi criado, baseado em uma média ponderada, calculada a partir da taxa da criminalidade (proporcional à população) e dos pesos adotados para os diversos tipos de crimes. Posteriormente, para permitir uma classificação mais interpretável e clara, o *IC* foi normalizado e reenquadrado em classes categóricas.

4.7.2.1 *Ranking* de Perigosidade dos Municípios

Figura 35 – 10 municípios com os maiores Índices de Criminalidade Normalizado (*ICN*), em 2019.



Fonte: Elaborada pelo autor.

A Figura 35 apresenta os 10 maiores Índices Normalizados, no ano mais recente (2019). No nível altíssimo, enquadraram-se Belo Horizonte, Juatuba, Água Comprida e Contagem.

Ao confrontarmos esses dados com a Tabela 21, a qual apresenta os 10 municípios com as maiores taxas de criminalidade por 100.000 habitantes em 2019, notam-se algumas alterações, evidenciando que nem sempre os locais que alcançaram altas taxas de criminalidade, apresentaram, também, segundo os pesos adotados, altos níveis de perigosidade. Por esta nova análise, o município de Confins, terceiro local com a maior taxa criminal em 2019, foi classificado como sétimo mais perigoso e seu nível obtido foi "Alto". Por outro lado, o município de Contagem, que não aparece entre os 10 municípios com as maiores taxas em 2019, foi classificado como quarto mais perigoso, atingindo o nível "Altíssimo".

Tabela 23 – Detalhamento anual dos municípios mais perigosos (Top 5).

Ano	**Total de municípios por Nível**	**Município - *ICN* (Nível)**
2012	Altíssimo : 6 Alto : 35 Intermediário : 127 Baixo : 472 Baixíssimo : 213	Confins - 1,00 (Altíssimo) Carmo do Paranaíba - 0,95 (Altíssimo) Corinto - 0,90 (Altíssimo) Belo Horizonte - 0,89 (Altíssimo) Pirapora - 0,88 (Altíssimo)
2013	Altíssimo : 1 Alto : 12 Intermediário : 64 Baixo : 489 Baixíssimo : 287	Confins - 1,00 (Altíssimo) Belo Horizonte - 0,74 (Alto) Carmo do Paranaíba - 0,72 (Alto) Contagem - 0,67 (Alto) Corinto - 0,64 (Alto)
2014	Altíssimo : 1 Alto : 10 Intermediário : 95 Baixo : 477 Baixíssimo : 270	Confins - 1,00 (Altíssimo) Belo Horizonte - 0,82 (Alto) Contagem - 0,77 (Alto) Uberaba - 0,75 (Alto) Juatuba - 0,71 (Alto)
2015	Altíssimo : 3 Alto : 25 Intermediário : 101 Baixo : 487 Baixíssimo : 237	Confins - 1,00 (Altíssimo) Belo Horizonte - 0,94 (Altíssimo) Contagem - 0,89 (Altíssimo) Perdigão - 0,85 (Alto) Juatuba - 0,85 (Alto)
2016	Altíssimo : 5 Alto : 25 Intermediário : 107 Baixo : 478 Baixíssimo : 238	Confins - 1,00 (Altíssimo) Juatuba - 0,97 (Altíssimo) Belo Horizonte - 0,96 (Altíssimo) Contagem - 0,92 (Altíssimo) Perdigão - 0,86 (Altíssimo)
2017	Altíssimo : 6 Alto : 29 Intermediário : 150 Baixo : 494 Baixíssimo : 174	Belo Horizonte - 1,00 (Altíssimo) Conceição do Pará - 0,96 (Altíssimo) Nova Serrana - 0,94 (Altíssimo) Contagem - 0,94 (Altíssimo) Confins - 0,94 (Altíssimo)
2018	Altíssimo : 3 Alto : 34 Intermediário : 191 Baixo : 507 Baixíssimo : 118	Confins - 1,00 (Altíssimo) Belo Horizonte - 0,99 (Altíssimo) Contagem - 0,96 (Altíssimo) São Joaquim de Bicas - 0,84 (Alto) Juatuba - 0,83 (Alto)
2019	Altíssimo : 4 Alto : 49 Intermediário : 192 Baixo : 518 Baixíssimo : 90	Belo Horizonte - 1,00 (Altíssimo) Juatuba - 0,87 (Altíssimo) Agua Comprida - 0,87 (Altíssimo) Contagem - 0,87 (Altíssimo) Uberaba - 0,82 (Alto)

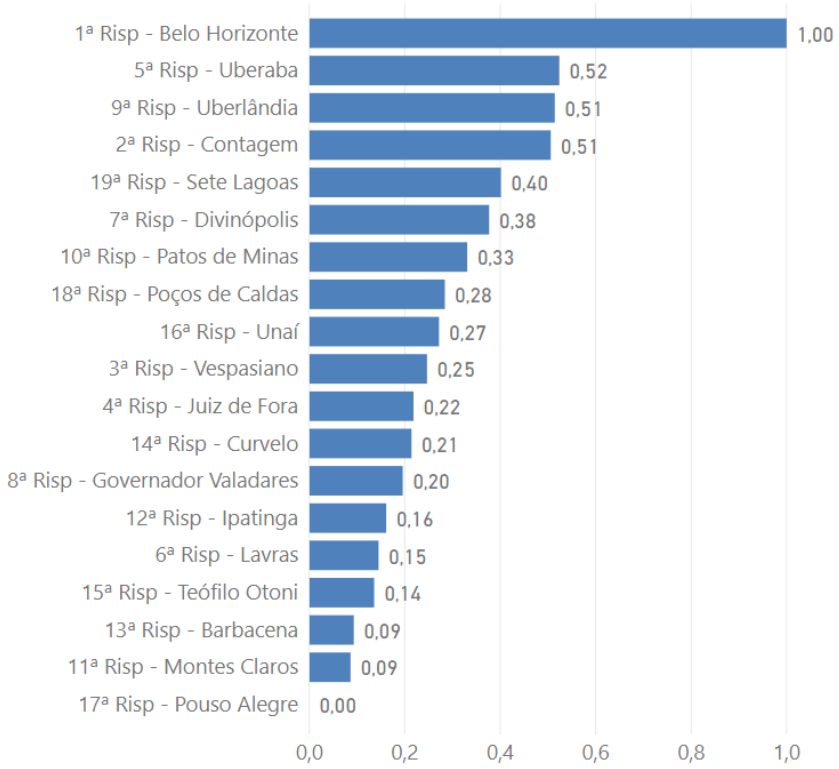
Fonte: Elaborada pelo autor.

O *ICN* indica um *ranking*, uma ordem de perigosidade, relativa aos municípios e ao ano analisado (vide seção 4.4.4), podendo indicar, por exemplo, que as políticas públicas do governo estadual devem considerar uma atenção maior para Belo Horizonte, Juatuba, Água Comprida e Contagem, i.e., cidades com grau "Altíssimo", cumulativas às desenvolvidas para 2019. A classificação de um município, por exemplo, em um ano como alto e em outro ano como altíssimo (vide Tabela 23) não implica um aumento absoluto no nível de perigosidade. Esta troca de posição pode indicar, apenas, uma menor distância para o primeiro colocado daquele ano, aumentando seu nível de perigosidade relativo ao resto do estado. No entanto, essa concorrência salutar só beneficia o cidadão e deve estimular prefeitos na busca pela liderança das cidades mais seguras.

A Tabela 23 detalha, para cada ano, a quantidade de municípios por nível de perigosidade e os cinco mais perigosos, com seus respectivos índices de criminalidade normalizados. Em resposta à questão **Q1**, é importante destacar que Belo Horizonte esteve presente, entre os cinco mais perigosos, em todos os períodos avaliados, sendo classificado no grau "Altíssimo", em seis dos oitos anos. Além disso, os municípios de Confins e Contagem foram classificados entre os cinco mais perigosos, em sete dos oito anos. Uma atenção especial também para Juatuba, presente entre os mais perigosos, em cinco dos últimos 6 anos.

4.7.2.2 *Ranking* de Perigosidade das Risps

Figura 36 – Índice de criminalidade normalizado das Risps em 2019.



Fonte: Elaborada pelo autor.

Em resposta à **Q2**, o gráfico da Figura 36 detalha os índices normalizados, obtidos por cada Risp, no ano de 2019. Nota-se que a região "1ª Risp - Belo Horizonte", única Risp a ser classificada no nível "Altíssimo" (*ICN=1,00*), superou consideravelmente as demais regiões, já que estas foram classificadas no grau "Intermediário" ou em níveis inferiores.

4.7.3 Análise e Interpretação das Regras de Associação

4.7.3.1 Associações entre Tipos de Crime e Municípios/Risps

No total, 136 associações foram encontradas entre "Municípios ⇒ Tipos de Crimes", as quais estão distribuídas, por ano, da seguinte forma: 2012 (19), 2013 (19), 2014 (13), 2015 (12), 2016 (15), 2017 (15), 2018 (20) e 2019 (23). Vale ressaltar que todas as associações encontradas atenderam aos requisitos mínimos exigidos pelo experimento (vide seção 4.6.1.1). Adicionalmente, ficou constatado que as associações "Unaí ⇒ Furto Consumado", "Ouro Preto ⇒ Furto Consumado" e "Araguari ⇒ Furto Consumado" foram encontradas nos oito anos analisados.

Tabela 24 – Regras encontradas para associações entre municípios e tipos de crime, em 2019.

Regra	***supp***	***conf***	***lift***	***count***	***r***	χ^2	***p-value* do χ^2**
Unaí ⇒ Furto Consumado	0,0030	0,78	1,16	1.217	0,014	80,03	0,00
Alfenas ⇒ Furto Consumado	0,0039	0,77	1,15	1.590	0,015	93,39	0,00
Poços De Caldas ⇒ Furto Consumado	0,0075	0,76	1,12	3.037	0,018	126,84	0,00
Ouro Preto ⇒ Furto Consumado	0,0026	0,75	1,11	1.061	0,009	34,97	0,00
Uberlândia ⇒ Furto Consumado	0,0347	0,74	1,10	14.081	0,031	394,86	0,00
Cataguases ⇒ Furto Consumado	0,0027	0,74	1,10	1.081	0,008	28,41	0,00
Itaúna ⇒ Furto Consumado	0,0038	0,73	1,09	1.551	0,010	36,91	0,00
Araguari ⇒ Furto Consumado	0,0044	0,73	1,09	1.797	0,010	38,98	0,00
Governador Valadares ⇒ Furto Consumado	0,0120	0,73	1,09	4.850	0,016	103,27	0,00
Belo Horizonte ⇒ Furto Consumado	0,1662	0,73	1,09	67.443	0,067	1824,80	0,00
Patos De Minas ⇒ Furto Consumado	0,0066	0,73	1,09	2.691	0,012	56,24	0,00
Ipatinga ⇒ Furto Consumado	0,0082	0,73	1,09	3.338	0,013	68,98	0,00
Nova Lima ⇒ Furto Consumado	0,0031	0,72	1,08	1.260	0,007	21,11	0,00
São Sebastião Do Paraíso ⇒ Furto Consumado	0,0026	0,72	1,07	1.065	0,006	14,58	0,00
Ituiutaba ⇒ Furto Consumado	0,0036	0,72	1,07	1.470	0,007	18,31	0,00
Araxá ⇒ Furto Consumado	0,0042	0,72	1,06	1.718	0,007	20,36	0,00
Uberaba ⇒ Furto Consumado	0,0179	0,71	1,06	7.255	0,013	72,69	0,00
Varginha ⇒ Furto Consumado	0,0041	0,71	1,06	1.650	0,006	16,15	0,00
Pirapora ⇒ Furto Consumado	0,0021	0,71	1,06	860	0,004	7,96	0,00
Itajubá ⇒ Furto Consumado	0,0024	0,71	1,05	970	0,004	7,16	0,01
Pará De Minas ⇒ Furto Consumado	0,0030	0,70	1,05	1.215	0,004	7,75	0,01
Passos ⇒ Furto Consumado	0,0048	0,70	1,05	1.955	0,005	12,08	0,00
Pouso Alegre ⇒ Furto Consumado	0,0045	0,70	1,04	1.809	0,005	10,29	0,00

Fonte: Elaborada pelo autor.

Em razão do espaço reduzido, a Tabela 24 apresenta as associações detectadas para o ano de 2019. Percebe-se que todas estão associadas ao crime "Furto Consumado", entre as quais se destacaram as regras dos municípios Unaí, Alfenas e Poços de Caldas, alcançando uma confiança superior a 75%. Apesar de não necessariamente implicar causa, essas regras coadunam com seus níveis de perigosidade mais baixos, uma vez que furto possui o menor peso para perigosidade. Além disso, como já publicado na literatura (ERVILHA; LIMA, 2019), em Minas Gerais, crimes

menos violentos, tais como furtos, tendem a estar mais associados a menores oportunidades de empregos para os jovens, menores gastos com segurança pública e um efetivo menor do policiamento militar em relação à população.

Estas associações foram avaliadas por meio da modelagem econométrica de dados em painel, além disto, vale ressaltar mais uma vez que as regras aqui apresentadas não implicam causa e efeito, ou seja, todos os dados aqui apresentados são apenas contextuais às regras ou aos índices, sem implicar conclusões definitivas, mas apenas um embasamento maior para a tomada de decisão em segurança pública.

Nesta mesma linha, essas regras podem contribuir para a priorização das políticas públicas (SILVARES, 2019) para estes municípios, bem como para o planejamento de experimentos e planos pilotos em locais com padrões mais explícitos. Em algumas situações, a segurança pode imitar os negócios, nos quais as ações imediatas já partem do "O quê" para, depois, entender o "Porquê".

Finalmente, com relação às associações "Risps ⇒ Tipos de Crimes", não foram encontradas RAs para os anos de 2014 e de 2017. Na Tabela 25, as informações encontradas para as associações são detalhadas, juntamente com os valores de suas medidas de interesse.

Tabela 25 – Regras encontradas para associações entre Risps e tipos de crimes.

Ano	**Regra**	***supp***	***conf***	***lift***	***count***	***r***	χ^2	***p-value* do χ^2**
2012	16º Risp - Unaí ⇒ Furto Consumado	0,0155	0,72	1,12	7.242	0,024	272,40	0,00
2013	16º Risp - Unaí ⇒ Furto Consumado	0,0148	0,70	1,11	7.211	0,022	234,52	0,00
2015	9º Risp - Uberlândia ⇒ Furto Consumado	0,0400	0,71	1,18	20.977	0,054	1525,53	0,00
2016	17º Risp - Pouso Alegre ⇒ Furto Consumado	0,0174	0,70	1,17	9.771	0,033	628,76	0,00
2016	9º Risp - Uberlândia ⇒ Furto Consumado	0,0381	0,70	1,17	21.398	0,051	1462,96	0,00
2018	16º Risp - Unaí ⇒ Furto Consumado	0,0123	0,70	1,08	5.658	0,014	95,07	0,00
2018	9º Risp - Uberlândia ⇒ Furto Consumado	0,0455	0,70	1,08	20.956	0,029	378,02	0,00
2019	16º Risp - Unaí ⇒ Furto Consumado	0,0116	0,71	1,05	4.725	0,010	37,27	0,00
2019	18º Risp - Poços de Caldas ⇒ Furto Consumado	0,0362	0,70	1,04	14.699	0,015	88,26	0,00
2019	1º Risp - Belo Horizonte ⇒ Furto Consumado	0,1662	0,73	1,09	67.443	0,067	1824,80	0,00
2019	9º Risp - Uberlândia ⇒ Furto Consumado	0,0474	0,73	1,08	19.210	0,031	384,67	0,00

Fonte: Elaborada pelo autor.

Em resposta às questões **Q3** e **Q4** considerando as regras selecionadas pelas medidas de interesse adotadas, também ficou constatado que os *p-values* ficaram abaixo do nível de significância adotado ($p\text{-value} < 0{,}05$). Dessa forma, para estas questões, a hipótese H_0 pôde ser rejeitada, indicando que há uma dependência entre o antecedente e o consequente.

4.7.3.2 Associações entre Alvos de Roubo e Municípios/Risps

No contexto das associações entre "Municípios ⇒ Alvos de roubo", 3 regras foram encontradas. A Tabela 26 detalha as informações encontradas para essas associações.

Além de serem municípios com alta atividade econômica, vale a pena destacar que a associação "Belo Horizonte ⇒ Transeunte", ocorrida no último ano (2019), apresentou um

Tabela 26 – Regras encontradas para associações entre e municípios e alvos de roubo.

Ano	**Regra**	***supp***	***conf***	***lift***	***count***	***r***	χ^2	***p-value do*** χ^2
2016	Montes Claros ⇒ Transeunte	0,0204	0,70	1,27	2.535	0,052	332,06	0,00
2017	Montes Claros ⇒ Transeunte	0,0189	0,72	1,27	2.022	0,051	281,24	0,00
2019	Belo Horizonte ⇒ Transeunte	0,2180	0,71	1,25	10.792	0,189	1.759,48	0,00

Fonte: Elaborada pelo autor.

suporte no valor aproximando de 21,80%, o que indica 10.792 ocorrências, de uma amostra de 49.502 transações. Além disso, o valor da confiança, ou seja, 0,71, confere-nos dizer que mais de 70% dos roubos do município de "Belo Horizonte" têm como alvo o "Transeunte", pessoas em movimento pela capital mineira. Isto pode implicar necessidade de mudanças nas estratégias de policiamento e monitoramento por meios eletrônicos, como também pode implicar realocação de esforços de outros tipos de segurança, tal como a segurança patrimonial.

Tabela 27 – Regras encontradas para associações entre Risps e alvos de roubo.

Ano	**Regra**	***supp***	***conf***	***lift***	***count***	***r***	χ^2	***p-value do*** χ^2
2019	1ª Risp - Belo Horizonte ⇒ Transeunte	0,2180	0,71	1,25	10.792	0,189	1.759,48	0,00

Fonte: Elaborada pelo autor.

Por fim, com relação às associações "Risps ⇒ Alvos de roubo", apenas uma associação foi encontrada, para o ano de 2019. A Tabela 27 detalha as informações encontradas para esta regra. A associação "1ª Risp - Belo Horizonte ⇒ Transeunte", na Tabela 27, valida a associação "Belo Horizonte ⇒ Transeunte", do ano de 2019, apresentada na Tabela 26, uma vez que a "1ª Risp - Belo Horizonte" é constituída apenas por este município.

Consequentemente, considerando as regras selecionadas pelas medidas de interesse adotadas, também ficou constatado, para as questões **Q5** e **Q6**, que os *p-values* ficaram abaixo do nível de significância adotado ($p\text{-value} < 0{,}05$). Dessa forma, para todas as regras aqui expostas, a hipótese H_0 pôde ser rejeitada, indicando que há uma dependência entre o antecedente e o consequente.

4.7.3.3 Associações entre Alvos de Furto e Municípios/Risps

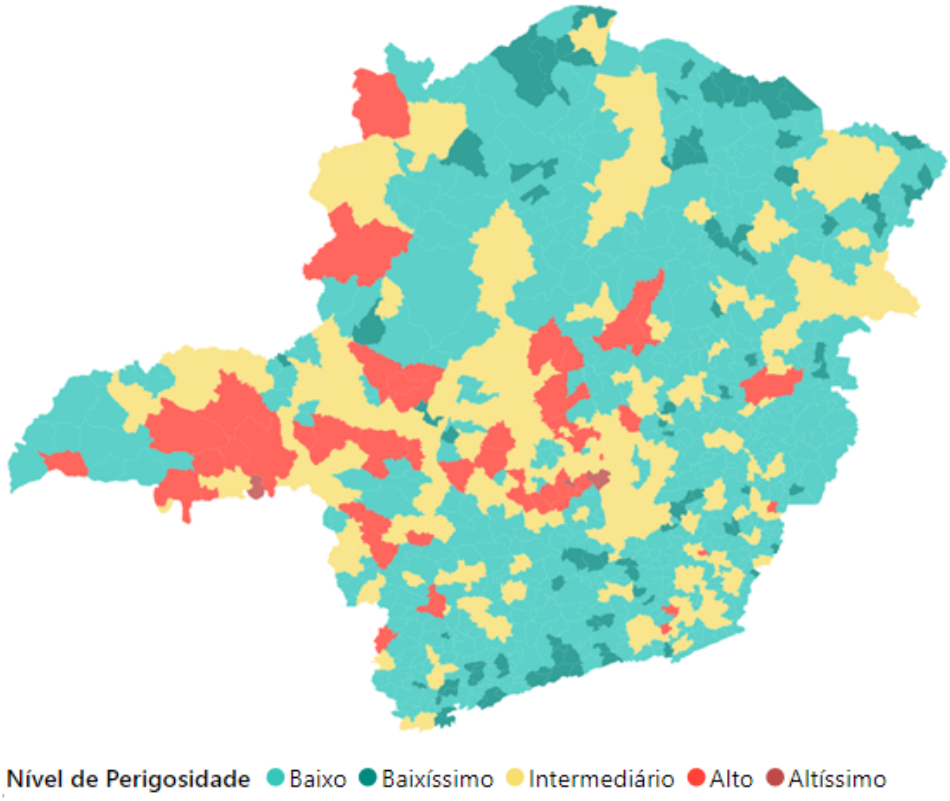
Não foram encontradas associações para "Municípios ⇒ Alvos de furto" e "Risps ⇒ Alvos de furto". Logo, para as questões **Q7** e **Q8**, as hipóteses H_0 não podem ser rejeitadas.

4.7.4 Análise Espacial e Socioeconômica

Na Figura 37 os índices de perigosidade de 2019 podem ser vistos por município e sua localização geográfica. De imediato, é notória a confirmação de boa parte das evidências

publicadas em (BARROS et al., 2019), as quais apontaram uma tendência de regiões com desenvolvimento econômico à baixa criminalidade, tais como as regiões de Juiz de Fora, Ipatinga e Montes Claros, todavia, que há casos em que o nível de desenvolvimento econômico não consegue conter o avanço do crime, tais como em Uberlândia e na grande Belo Horizonte, cujos índices de perigosidade foram alto e altíssimo. Neste contexto, como os crimes contra pessoa têm uma influência maior no índice de perigosidade, em (ERVILHA; LIMA, 2019), usando modelagem de dados em painel, foram encontradas evidências de que, em Minas Gerais, as variáveis população de 15 a 24 anos (relação positiva com criminalidade), mortalidade padronizada (positiva), escolarização líquida (negativa), emprego formal (positiva) e razão de dependência (negativa) foram estatisticamente significativas em pelo menos 5%, bem como gasto per capita com assistência social (negativa), ao nível de 10%.

Figura 37 – Índice de criminalidade por localização em 2019.



Fonte: www.transparenciatraduzida.ufs.br.

O índice de desenvolvimento econômico destes locais deveria refletir em investimentos na educação, políticas de proteção social (SILVARES, 2019) e soluções para aumento da qualidade de vida e ocupação de jovens, os quais participam precocemente e são vítimas da criminalidade. Neste contexto, (CARAZZA; NETO; EMANUEL, 2020) evidenciaram que a adoção de uma política de toque de recolher juvenil reduziu a taxa de furto em 17,9%.

Fugindo um pouco às tendências já publicadas, uma vez que o nível de perigosidade não considera apenas os Homicídios, por exemplo, Chácara, vizinha a Juiz de Fora, alcançou

nível de perigosidade alto, com destaque paras as Lesões Corporais. Também chama a atenção as cidades do Noroeste do estado que alcançaram índices altos, Buritis e Paracatu, próximas ao Distrito Federal e ambas da Risp Unaí, cuja a cidade de mesmo nome alcançou um índice intermediário. Essa tendência é clara no mapa, com cidades mais perigosas vizinhas a cidades com níveis intermediário e baixos, o que indica a necessidade de soluções preventivas contra o avanço da criminalidade e pode indicar um deslocamento da criminalidade para locais vizinhos, com maiores ofertas para prática criminal e/ou menor assistência social. Além disso, o município de Paracatu é o principal município da região, com presença de muitas empresas, mais empregos, escolas e mineração de ouro, o que atrai gente de toda a região e a presença de jovens, infelizmente, correlacionados com os crimes mais violentos. Por outro lado, Buritis, apesar de ser menor, chamou a atenção, conforme destacado anteriormente, pela alta taxa de Estupros, principalmente de vulneráveis. Para coibir os estupros, esta é uma cidade cujas particularidades municipais devem ser exploradas (ERVILHA; LIMA, 2019), além das variáveis que explicam o aumento da criminalidade já publicadas.

Na mesma linha, Água Comprida, localizada em uma região de grande desenvolvimento econômico, também chama atenção pelo seu índice Altíssimo. Em (ERVILHA; LIMA, 2019), também se verificou que os municípios com características intrínsecas não observadas e taxa de crimes contra pessoa positivamente relacionadas encontram-se, via de regra, nas regiões com maiores desigualdades sociais. Neste sentido, por exemplo, na contramão, com melhor distribuição de renda, Água Comprida e Uberaba tiveram uma relação negativa entre os efeitos locais, fixos no tempo, não observáveis, e as taxas de crime. Isto implica que estas cidades podem concentrar-se em gerir melhor as variáveis explicativas já estudadas para a região de Minas Gerais. No caso de Água Comprida, é alarmante a taxa de Lesões Corporais Consumadas, sendo aproximadamente o dobro da taxa de Belo Horizonte (vide Tabela 22), para uma cidade que não teve nenhum Homicídio em 2019 e que já pertence a uma Risp com destaque para as lesões corporais. A administração pública pode analisar a viabilidade de mais políticas de proteção social e melhores investimentos em Educação.

Por fim, vale a pena situar Minas Geras em relação ao resto do país. De acordo com Lima e Bueno (2020), considerando o número de homicídios dolosos e um comparativo entre os primeiros semestres dos anos de 2019 e 2020, os estados do Amapá (0,0%) e Minas Gerais (0,6%) são destaques, com índices abaixo de 1%. No mesmo contexto, alguns estados se destacaram pela redução dos seus números. O estado de Roraima foi o que apresentou a maior redução, -31,1%. Em seguida, vieram Pará, com -24,5%, Goiás, com -16,6%, Rio de Janeiro, com -9,2%, e Distrito Federal, com -8,4%. Por outro lado, o Ceará mais que dobrou o número de homicídios, atingindo a maior majoração entre as Unidades Federativas, ou seja, 106,9%. Este índice pode ser justificado pelo fato de que este viveu uma crise de segurança pública no início de 2020, com a greve da Polícia Militar, a qual durou 13 dias, no mês de fevereiro. Essa crise teve impactos importantes nos indicadores da segurança pública estadual, no primeiro semestre.

4.7.5 Ameaças à Validade

De acordo com Chapetta (2006), ameaças à validade podem limitar a habilidade de interpretar e/ou descrever resultados dos dados obtidos em um experimento. Neste sentido, as seguintes ameaças, encontradas durante a experimentação, devem ser consideradas.

- **Ameaças às validades de construção e interna -** Considerando que os dados foram obtidos, por meio de *download*, tratados e analisados pelos autores, existem ameaças a serem consideradas. Para mitigar possíveis erros, todos os artefatos de software construídos para o tratamento dos dados e os resultados por eles gerados foram homologados e revisados por mais de um pesquisador, considerando amostras de cálculos feitos pelos artefatos, contra amostras de cálculos replicadas em planilhas, manualmente e diretamente no banco de dados. Tais testes foram feitos na fase de construção (validade de construção) dos artefatos e na fase de execução (validade interna).

  Foi utilizada uma taxa bruta para o cálculo dos indicadores, ou seja, para municípios muito pequenos que se apresentem como muito perigosos, as autoridades públicas devem fazer uma avaliação melhor dos valores absolutos destes municípios e definir prioridades. Uma alternativa para isto é usar **taxas bayesianas**, no entanto, neste trabalho, essa situação foi relativamente mitigada pela consideração de diversos tipos de crimes para a criação do índice.

- **Ameaças à validade de conclusão -** Os pesos considerados para os crimes podem não refletir a cultura, o nível de violência e o clamor social de algum lugar específico. Esta ameaça foi mitigada com a consideração das penas dos crimes. Além disso, as distâncias dos pesos são aproximadamente proporcionais às distâncias das penas. Vale ressaltar que os pesos podem variar a fórmula de cálculo, de acordo com o objetivo do estudo.

- **Ameaças à validade de conclusão - (1)** A falta de correção dos dados criminais fornecidos pelo governo de Minas Gerais e das estimativas populacionais fornecidas pelo IBGE, bem como possíveis subnotificações poderão influenciar diretamente o resultado do experimento. **(2)** A agregação, imposta pelo governo de Minas Gerais, de diversos tipos de crime, também pode influenciar os resultados, uma vez que não é possível pesar e ponderar com maior precisão pela reprovabilidade. Por fim, os *rankings* se limitam aos tipos de crimes disponibilizados.

4.8 Conclusão e Trabalhos Futuros

Este trabalho teve o intuito de detectar padrões, bem como promover maior transparência, visando auxiliar o processo de apoio às decisões estratégicas e operacionais dos governantes e agentes da lei, no combate efetivo da criminalidade. Foram utilizados dados dos 853 municípios,

relacionados a incidentes criminais, disponibilizados, de forma aberta, pelo Governo de Minas Gerais, por meio da Secretaria de Estado de Justiça e Segurança Pública (Sejusp).

Como relação ao nível de perigosidade, ficou constatado que Belo Horizonte esteve presente entre os cinco mais perigosos, em todos os períodos avaliados. Além disso, os municípios de Confins e Contagem também foram classificados entre os cinco mais perigosos, em sete dos oito anos. Como destaque fora desse eixo, Água Comprida e Buritis apresentaram casos muito particulares a serem explorados, considerando localização, desenvolvimento regional, educação e políticas sociais.

No âmbito das regras de associação, considerando as associações "Municípios ⇒ Tipos de Crimes", destaque paras as associações "Unaí ⇒ Furto Consumado", "Ouro Preto ⇒ Furto Consumado" e "Araguari ⇒ Furto Consumado", as quais foram encontradas nos oito anos analisados. Para as associações "Municípios ⇒ Alvos de Roubo, o destaque ficou por conta da associação "Belo Horizonte ⇒ Transeunte", ocorrida no último ano (2019), a qual obteve o suporte de 21,80% e uma confiança de 0,71, indicando a necessidade do poder público investigar as causas e ações necessárias para coibir o roubo a este tipo de alvo.

Além da pesquisa desenvolvida neste trabalho e como trabalhos futuros, outros possíveis desdobramentos poderão ser analisados. Dados criminais levando em consideração regiões menores, como, por exemplo, as regiões metropolitanas ou os bairros, poderão ser investigados, desde que estados ou municípios os disponibilizem. Nesta linha, podem ser encontradas as associações entre crimes, considerando a localidade e os espaços temporais disponíveis, bem como permitindo a prioridade na inibição de crimes que são antecedentes de outras ocorrências dentro de uma mesma cidade. Sobre pesos de crimes e suas penas, sugerimos a realização de um *Survey* com juristas, criminalistas, profissionais e especialistas em segurança pública em todo o Brasil, coletando as medianas de pesos atribuídos pelos entrevistados e ampliando a discussão sobre o que pode ser considerada uma localidade mais perigosa e sobre as penas dos crimes. Isto também poderá embasar o poder público e os legisladores, em busca de um código penal que reflita melhor o clamor da sociedade brasileira.

À guisa de conclusão, destacamos a relevância desta pesquisa na produção de evidências da necessidade da disponibilidade de melhores e mais completas arquiteturas de dados abertos para os governos estaduais e de um modelo para outros estados de extração, produção de conhecimento e publicação automatizadas, baseadas nestes dados. Em paralelo, também é importante a produção de indicadores de perigosidade atualizados, que não se baseiam apenas em homicídios, e regras que sirvam para apoiar a definição de prioridades nas avaliações e aplicações de políticas públicas, auxiliando o planejamento estratégico e o processo de apoio à decisão, buscando conter o orçamento empregado na área de segurança pública. Os dados aqui apresentados poderão ser consultados, nessas e em outras perspectivas, no site do projeto Transparência Traduzida, em (www.transparenciatraduzida.ufs.br).

5
Conclusão

As recentes políticas de disponibilização de dados públicos fornecem um canal direto entre o cidadão e o governo, as quais proporcionam o fortalecimento do processo democrático e melhoram a qualidade de vida da população. Com a publicação de dados, é possível o controle social mais efetivo, pois a sociedade terá mais condições de cobrar, exigir e fiscalizar seus governantes. Entretanto, esta participação pode ser potencializada e estimulada, desde que os dados fornecidos pelos governos sejam estruturados e dispostos de uma forma entendível, ou efetivamente transparente, para o cidadão. Quando isto não ocorre, pesquisas sobre dados abertos podem contribuir com a estruturação, autenticação, interpretação, organização, tradução, descoberta de padrões e disponibilização de informação útil para tomada de decisão e para o controle social exercido pela população.

5.1 Contribuições

A principal contribuição deste estudo foi aplicar *Data Science*, para analisar dados abertos governamentais relacionados a incidentes criminais, ocorridos nos estados brasileiros e nos municípios de Minas Gerais, com objetivo de auxiliar o planejamento estratégico governamental e o processo de tomada de decisão no combate efetivo da criminalidade. Este trabalho focou em encontrar aberrações (*outliers)* e correlações, bem como, desenvolver *rankings* dos estados, municípios e Risps mais perigosos.

Em resumo, entre as principais contribuições deste trabalho, destacam-se:

- Revisão Sistemática Quantitativa que teve como objetivo identificar, caracterizar e metanalisar trabalhos científicos que fizeram uso de algoritmos, técnicas e abordagens inteligentes sobre dados criminais. Ressalta-se que os resultados foram publicados no periódico *Journal of Applied Security Research* (PRADO et al., 2020);

- Criação de uma arquitetura centralizada, com objetivo de automatizar as etapas de *download*, ETL (Extração, Transformação e Carga) e mineração dos dados;
- Dois experimentos controlados sobre os dados criminais ocorridos no Brasil, disponibilizados pelo MJSP e pela Sesjusp/MG, para investigar a existência de correlações e anomalias que auxiliem no combate ostensivo e preventivo da criminalidade. Ressalta-se que desses experimentos resultaram dois artigos, os quais foram, respectivamente, publicados nos periódicos *Journal of Applied Security Research* (PRADO; COLAÇO JÚNIOR, 2020b) e *Research, Society and Development* (PRADO; COLAÇO JÚNIOR, 2020a);
- Duas aplicações públicas, disponíveis no site Transparência Traduzida [1], que realizam análises inteligentes sobre dados criminais dos estados brasileiros e dos municípios de Minas Gerais, facilitando a consulta e a visualização dos resultados.
- Como desdobramento desta pesquisa, outro experimento foi realizado para verificar a existência de associações entre os tipos de crimes, bem como entre tipos de crimes e meses do ano no Brasil. Ressalta-se que os resultados deste experimento foram publicados no periódico *Research, Society and Development* (GOMES; COLAÇO JÚNIOR; PRADO, 2020).

Como consequência destas contribuições, conseguimos responder às questões de pesquisa elaboradas no início do trabalho:

- **Q1**: Existem discrepâncias entre as taxas de criminalidade (taxa por cem mil habitantes) das Regiões Brasileiras? Sim. No total, 7 valores aberrantes (*outliers*) foram detectados nas regiões Norte (2) e Nordeste (5), distribuídos nos anos 2015 (3), 2016 (2) e 2017 (2). Adicionalmente, constatou-se que nos dois últimos anos (2018 e 2019) não foram encontrados valores atípicos. O destaque negativo foi o estado de Pernambuco, em 2017, que excedeu o limite máximo, com a taxa de 392,63 por 100.000 habitantes. Os demais valores atípicos configuram como pontos positivos, i.e., ficaram abaixo do limite mínimo.
- **Q2**: Quais os estados vêm se destacando como os mais perigosos? O estado do Paraná foi o local mais perigoso em todos os períodos avaliados, sendo sempre classificado no grau "Altíssimo". Destaque também para Rio de Janeiro, presente em todos os anos, sempre na segunda colocação. Além disso, os estados de Goiás, Pernambuco e Rondônia foram classificados entre os cinco mais perigosos, em três dos cinco anos avaliados.
- **Q3**: Há associações entre tipos de crime e os estados? Sim. Foram encontradas 10 associações entre estados e tipos de crimes, distribuídas da seguinte forma: 2015 (1), 2016 (2), 2017 (2), 2018 (2) e 2019 (4). Vale destacar que a associação "Rio Grande do Norte

[1] Transparência Traduzida - <http://www.http://www.transparenciatraduzida.ufs.br>

⇒ Roubo de veículo" esteve presente em 4 dos 5 anos, representado mais de 64% das ocorrências criminais do estado, aproximadamente o dobro dos percentuais do país nos últimos anos.

- **Q4**: Quais os municípios vêm se destacando como os mais perigosos? Entre os municípios mineiros, Belo Horizonte esteve presente, entre os cinco mais perigosos, em todos os períodos avaliados, sendo classificado no grau "Altíssimo", em seis dos oitos anos. Além disso, os municípios de Confins e Contagem foram classificados entre os cinco mais perigosos, em sete dos oito anos. Destaque também para o município de Juatuba, o qual esteve presente entre os mais perigosos, em cinco dos últimos seis anos.

- **Q5**: Quais as Risps mais perigosas em 2019? Em 2019, a região "1ª Risp - Belo Horizonte" foi única Risp a ser classificada no nível "Altíssimo", superando consideravelmente as outras Risps, as quais foram classificadas no grau "Intermediário" ou em níveis inferiores.

- **Q6**: Há associações entre tipos de crime e municípios? Sim. No total, 136 associações entre tipos de crime e municípios foram encontradas e distribuídas, por ano, da seguinte forma: 2012 (19), 2013 (19), 2014 (13), 2015 (12), 2016 (15), 2017 (15), 2018 (20) e 2019 (23). Ficou constatado que as associações "Unaí ⇒ Furto Consumado", "Ouro Preto ⇒ Furto Consumado" e "Araguari ⇒ Furto Consumado" foram as mais frequentes, encontradas em todos os anos analisados.

- **Q7**: Há associações entre tipos de crime e Risps? Sim. No total, 11 associações entre tipos de crime e Risps foram encontradas e distribuídas, por ano, da seguinte forma: 2012 (1), 2013 (1), 2015 (1), 2016 (2), 2018 (2) e 2019 (4). Não foram encontradas RAs para os anos de 2014 e de 2017. As regras "9º Risp - Uberlândia ⇒ Furto Consumado" e "16º Risp - Unaí ⇒ Furto Consumado" foram as mais frequentes.

- **Q8**: Há associações entre alvos de roubo e municípios? Sim. No total, 3 associações entre alvos de roubo e municípios foram detectadas, distribuídas nos anos 2016, 2017 e 2019. Não foram encontradas RAs para os anos de 2015 e de 2018. Destaque para a associação "Belo Horizonte ⇒ Transeunte", ocorrida no último ano (2019), com suporte de 21,80% e confiança de 71%, indicando que a grande maioria dos roubos ocorridos no município de "Belo Horizonte" têm como alvo o "Transeunte".

- **Q9**: Há associações entre alvos de roubo e Risps? Sim. Apenas uma RA entre alvos de roubo e Risps foi encontrada. Trata-se da associação "1ª Risp - Belo Horizonte ⇒ Transeunte", ocorrida em 2019.

- **Q10**: Há associações entre alvos de furto e municípios? Não. Nenhuma regra de associação foi encontrada entre alvos de furto e municípios, nos dados de incidentes criminais, ocorridos nos munícipios de Minas Gerais, entre os anos de 2012 e 2019.

- **Q11**: Há associações entre alvos de furto e Risps? Não. Nenhuma regra de associação foi encontrada entre alvos de furto e Risps, nos dados de incidentes criminais, ocorridos nos munícipios de Minas Gerais, entre os anos de 2012 e 2019.

5.2 Limitações

O estudo realizado apresentou limitações importantes quanto ao tratamento dos dados, a disponibilidade das informações e a quantidade de atributos contidos nos *datasets*.

Os dados utilizados nesta pesquisa foram obtidos, por meio de *download*, tratados e analisados pelos autores. Sendo assim, todos os artefatos de software construídos para o tratamento dos dados e os resultados por eles gerados foram revisados por mais de um pesquisador, com o intuito de mitigar possíveis erros. No entanto, esta tarefa de revisão, realizada pelos pesquisadores, introduz um fator subjetivo, visto que depende de uma observação cuidadosa do pesquisador para o perfeito enquadramento das informações. Desta forma, interpretações errôneas podem gerar análises inconsistentes.

Em relação à disponibilidade das informações, após uma busca minuciosa de *datasets* estaduais, constatou-se que a maioria dos estados ainda não disponibilizam, de forma aberta, seus dados relacionados a incidentes criminais. Além disso, grande parte das UFs que publicam seus *datasets* não seguem os princípios de dados abertos, ou seja, disponibilizam suas informações em formatos não abertos e sem padronização, dificultando a leitura por máquina.

Adicionalmente, em relação à área criminal, vale ressaltar que o Brasil, em comparação a outros países, precisa evoluir em termos de política e arquitetura para dados abertos. O maior detalhamento dos dados fornecido pelos governos permite o uso de mais opções de algoritmos, não explorados neste artigo pela incompletude dos dados fornecidos pelos órgãos brasileiros.

5.3 Trabalhos Futuros

Além da pesquisa desenvolvida neste trabalho, outros possíveis desdobramentos são:

- Ampliar a arquitetura centralizada, desenvolvida neste trabalho, para suportar outros tipos de *datasets* que serão depositados no portal Transparência Traduzida [2];
- Realizar um *Survey* com juristas, criminalistas, profissionais e especialistas em segurança pública em todo o Brasil, com intuito de calibrar os pesos dos tipos de crimes, coletando as medianas de pesos atribuídos pelos entrevistados e ampliando a discussão sobre o que pode ser considerada uma localidade mais perigosa e sobre as penas dos crimes;

[2] Transparência Traduzida - <http://www.http://www.transparenciatraduzida.ufs.br>

- Realizar uma análise para encontrar as associações entre crimes, considerando a localidade e os espaços temporais disponíveis, com o objetivo de inibir crimes que são antecedentes de outras ocorrências dentro de uma mesma cidade.
- Realizar uma análise criminal levando em consideração regiões menores, como, por exemplo, as regiões metropolitanas ou os bairros.

5.4 Considerações Finais

Este trabalho apresentou os resultados da análise dos dados abertos governamentais relacionados a incidentes criminais, com objetivo de encontrar aberrações (*outliers)* e correlações, bem como desenvolver *rankings* dos estados, municípios e Risps mais perigosos. Os resultados aqui apresentados poderão ser consultados, nessas e em outras perspectivas, no site do projeto Transparência Traduzida.

Por fim, destacamos a relevância desta pesquisa e dos órgãos aqui envolvidos e citados, os quais têm a responsabilidade social de apresentar resultados que servem como direcionamento para as decisões sobre segurança pública em todo o país. As proeminências destes órgãos e de análises como as descritas neste trabalho devem servir de alerta e direcionamento para os nossos governantes, auxiliando o planejamento estratégico e o processo de apoio à decisão, buscando conter o orçamento empregado na área de segurança pública. Vale ressaltar que o Brasil ainda precisa formular melhores diretrizes para a publicação de dados abertos governamentais, padronizando estruturas e protocolos, impondo limites temporais e punindo não conformidades.

Referências

ABRAMOVICI, M. et al. Competing fusion for bayesian applications. In: CITESEER. *Proceedings of IPMU*. [S.l.], 2008. v. 8, p. 379. Citado na página 24.

AGHABABAEI, S.; MAKREHCHI, M. Temporal topic inference for trend prediction. In: IEEE. *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. [S.l.], 2015. p. 877–884. Citado 2 vezes nas páginas 35 e 143.

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: ACM. *Acm sigmod record*. [S.l.], 1993. v. 22, n. 2, p. 207–216. Citado 2 vezes nas páginas 62 e 99.

AGRAWAL, S.; SEJWAR, V. Crime identification using fp-growth and multi objective particle swarm optimization. In: IEEE. *2017 International Conference on Trends in Electronics and Informatics (ICEI)*. [S.l.], 2017. p. 727–734. Citado 5 vezes nas páginas 35, 56, 94, 142 e 144.

ALBUQUERQUE, D. J. S. et al. Implementing e-government processes distribution with transparency using multi-agent systems. *iSys-Revista Brasileira de Sistemas de Informação*, v. 9, n. 1, p. 118–138, 2016. Citado na página 97.

ALKHAIBARI, A. A.; CHUNG, P.-T. Cluster analysis for reducing city crime rates. In: IEEE. *2017 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*. [S.l.], 2017. p. 1–6. Citado 3 vezes nas páginas 34, 141 e 142.

ANDRIENKO, G. et al. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. In: WILEY ONLINE LIBRARY. *Computer Graphics Forum*. [S.l.], 2010. v. 29, n. 3, p. 913–922. Citado 2 vezes nas páginas 34 e 142.

ANSARI, M. Y.; PRAKASH, A. et al. Application of spatiotemporal fuzzy c-means clustering for crime spot detection. *Defence Science Journal*, v. 68, n. 4, p. 374–380, 2018. Citado 3 vezes nas páginas 34, 141 e 142.

APPOLINÁRIO, F. Dicionário de metodologia científica: um guia para a produção do conhecimento científico. In: *Dicionário de metodologia científica: um guia para a produção do conhecimento científico*. [S.l.: s.n.], 2007. p. 300–300. Citado na página 24.

ARYAL, A. M.; WANG, S. Sparksnn: A density-based clustering algorithm on spark. In: IEEE. *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*. [S.l.], 2018. p. 433–437. Citado 3 vezes nas páginas 34, 56 e 142.

AWAL, M. A. et al. Using linear regression to forecast future trends in crime of bangladesh. In: IEEE. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. [S.l.], 2016. p. 333–338. Citado 2 vezes nas páginas 35 e 142.

BACULO, M. J. C. et al. Geospatial-temporal analysis and classification of criminal data in manila. In: IEEE. *2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA)*. [S.l.], 2017. p. 6–11. Citado 5 vezes nas páginas 34, 35, 141, 143 e 144.

BALOIAN CORONEL ENRIQUE BASSALETTI, M. F. O. t.-c. O. F. P. F. R. M. M. O. S. P. J. A. P. M. V. N. Crime prediction using patterns and context. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017. Citado 2 vezes nas páginas 35 e 141.

BAPPEE, F. K.; JÚNIOR, A. S.; MATWIN, S. Predicting crime using spatial features. In: SPRINGER. *Advances in Artificial Intelligence: 31st Canadian Conference on Artificial Intelligence, Canadian AI 2018, Toronto, ON, Canada, May 8–11, 2018, Proceedings 31*. [S.l.], 2018. p. 367–373. Citado 9 vezes nas páginas 26, 35, 43, 44, 50, 52, 55, 141 e 142.

BARROS, P. H. B. de et al. Economic development and crime in brazil: a multivariate and spatial analysis. *Revista Brasileira de Estudos Regionais e Urbanos*, v. 13, n. 1, p. 1–22, 2019. Citado 2 vezes nas páginas 95 e 118.

BASILI, V. et al. *Aligning Organizations Through Measurement: The GQM+ Strategies Approach*. [S.l.]: Springer, 2014. Citado 2 vezes nas páginas 67 e 105.

BASILI, V.; WEISS, D. M. A methodology for collecting valid software engineering data. *IEEE Transactions on software engineering*, IEEE, n. 6, p. 728–738, 1984. Citado 2 vezes nas páginas 67 e 105.

BELESIOTIS, A.; PAPADAKIS, G.; SKOUTAS, D. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, ACM, v. 3, n. 4, p. 12, 2018. Citado 4 vezes nas páginas 34, 141, 142 e 144.

BENGTSSON, F.; HEIN, A.; OLSSON, C. M. Exploring the potential for using artificial intelligence techniques in police report analysis. 2012. Citado 4 vezes nas páginas 34, 35, 142 e 143.

BERTOT, J. C. et al. Big data, open government and e-government: Issues, policies and recommendations. *Information polity*, IOS Press, v. 19, n. 1, 2, p. 5–16, 2014. Citado na página 97.

BERWANGER, O. et al. Como avaliar criticamente revisões sistemáticas e metanálises. *Rev Bras Ter Intensiva*, SciELO Brasil, v. 19, n. 4, p. 475–80, 2007. Citado 4 vezes nas páginas 27, 29, 41 e 42.

BHARATHI, S.; INDRANI, B.; PRABAKAR, M. A. A supervised learning approach for criminal identification using similarity measures and k-medoids clustering. In: IEEE. *2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT)*. [S.l.], 2017. p. 646–653. Citado 2 vezes nas páginas 34 e 141.

BLACKBURN, S. *The Oxford Dictionary of Philosophy*. 2016. Accessed: 2020-09-23. Disponível em: <http://www.oxfordreference.com/view/10.1093/acref/9780199541430.001.0001/acref-9780199541430-e-1645>. Citado 2 vezes nas páginas 23 e 96.

BOGAHAWATTE, K.; ADIKARI, S. Intelligent criminal identification system. In: IEEE. *2013 8th International Conference on Computer Science & Education*. [S.l.], 2013. p. 633–638. Citado 2 vezes nas páginas 34 e 141.

BONI, M. A.; GERBER, M. S. Area-specific crime prediction models. In: IEEE. *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.], 2016. p. 671–676. Citado 2 vezes nas páginas 35 e 141.

BONI, M. A.; GERBER, M. S. Predicting crime with routine activity patterns inferred from social media. In: IEEE. *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [S.l.], 2016. p. 001233–001238. Citado 2 vezes nas páginas 35 e 143.

BUCZAK, A. L.; GIFFORD, C. M. Fuzzy association rule mining for community crime pattern discovery. In: ACM. *ACM SIGKDD Workshop on Intelligence and Security Informatics*. [S.l.], 2010. p. 2. Citado 4 vezes nas páginas 19, 34, 35 e 141.

CALVO, H. et al. Forecasting, clustering and patrolling criminal activities. *Intelligent Data Analysis*, IOS Press, v. 21, n. 3, p. 697–720, 2017. Citado 3 vezes nas páginas 34, 141 e 142.

CAMPOS, O. S. F. Data analytics transparente para descoberta de padrões e anomalias na realização de convênios e contratos de repasse federais [Transparent data analysis for pattern discovery and anomalies in the convenant of agreements and contracts for federal transfer]. Pós-Graduação em Ciência da Computação, 2018. Citado 5 vezes nas páginas 16, 62, 63, 99 e 100.

CAPES. *Portal de periódicos CAPES/MEC*. 2019. Disponível em: <http://www.periodicos.capes.gov.br/>. Citado 3 vezes nas páginas 23, 29 e 57.

CARAZZA, L.; NETO, R. da M. S.; EMANUEL, L. Juvenile curfew and crime reduction: Evidence from brazil. *Papers in Regional Science*, Wiley Online Library, 2020. Citado 2 vezes nas páginas 95 e 118.

CATLETT, C. et al. A data-driven approach for spatio-temporal crime predictions in smart cities. In: IEEE. *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*. [S.l.], 2018. p. 17–24. Citado 8 vezes nas páginas 15, 25, 34, 35, 53, 54, 92 e 142.

CAVADAS, B.; BRANCO, P.; PEREIRA, S. Crime prediction using regression and resources optimization. In: SPRINGER. *Portuguese Conference on Artificial Intelligence*. [S.l.], 2015. p. 513–524. Citado 4 vezes nas páginas 34, 35, 141 e 144.

CERQUEIRA, D. et al. Atlas da violência 2019 [Atlas of violence 2019]. Instituto de Pesquisa Econômica Aplicada (IPEA), 2018. Citado 3 vezes nas páginas 19, 21 e 84.

CHAN, S.; LEONG, K. An application of cyclic signature (cs) clustering for spatial-temporal pattern analysis to support public safety work. In: IEEE. *2010 IEEE International Conference on Systems, Man and Cybernetics*. [S.l.], 2010. p. 2716–2723. Citado 5 vezes nas páginas 34, 35, 141, 142 e 143.

CHANDRA, B.; GUPTA, M. Novel multivariate time series clustering approach for e-governance of crime data. In: IEEE. *2013 Sixth International Conference on Developments in eSystems Engineering*. [S.l.], 2013. p. 311–316. Citado 2 vezes nas páginas 35 e 144.

CHAPETTA, W. A. *Uma Infra-estrutura para Planejamento, Execução e Empacotamento de Estudos Experimentais em Engenharia de Software [An Infrastructure for Planning, Execution and Packaging of Experimental Studies in Software Engineering]*. Tese (Doutorado) — Dissertação de Mestrado, Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil, 2006. Citado 2 vezes nas páginas 89 e 120.

CHAVES, M. S. R. S.; SHIMIZU, I. S. et al. Síndrome de burnout e qualidade do sono de policiais militares do piauí. *Revista Brasileira de Medicina do Trabalho*, Revista Brasileira de Medicina do Trabalho, v. 16, n. 4, p. 436–441, 2018. Citado na página 108.

CHEN, Z. et al. Using map-based interactive interface for understanding and characterizing crime data in cities. In: SPRINGER. *International Conference in Swarm Intelligence*. [S.l.], 2015. p. 479–490. Citado 3 vezes nas páginas 35, 56 e 144.

CHI, H. et al. A decision support system for detecting serial crimes. *Knowledge-Based Systems*, Elsevier, v. 123, p. 88–101, 2017. Citado 3 vezes nas páginas 35, 143 e 145.

CONFORTI, R.; ROSA, M. L.; HOFSTEDE, A. H. ter. Noise filtering of process execution logs based on outliers detection. 2015. Citado na página 60.

CRANDELL, I.; KORKMAZ, G. Link prediction in the criminal network of albuquerque. In: IEEE. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. [S.l.], 2018. p. 564–567. Citado 2 vezes nas páginas 35 e 141.

DADOS. *Portal Brasileiro de Dados Abertos [Brazilian Open Data Portal]*. 2020. Accessed: 2020-01-15. Disponível em: <http://www.dados.gov.br>. Citado 2 vezes nas páginas 97 e 98.

DAMASCENO, M.; TEIXEIRA, J.; CAMPOS, G. A prediction model for criminal levels using socio-criminal data. *International Journal of Electronic Security and Digital Forensics*, Inderscience Publishers, v. 4, n. 2/3, p. 201–214, 2012. Citado 7 vezes nas páginas 15, 21, 25, 35, 53, 92 e 144.

DAS, P.; DAS, A. K. Behavioural analysis of crime against women using a graph based clustering approach. In: IEEE. *2017 International Conference on Computer Communication and Informatics (ICCCI)*. [S.l.], 2017. p. 1–6. Citado 4 vezes nas páginas 34, 35, 141 e 143.

DAS, P.; DAS, A. K. Application of classification techniques for prediction and analysis of crime in india. In: *Computational Intelligence in Data Mining*. [S.l.]: Springer, 2019. p. 191–201. Citado 5 vezes nas páginas 30, 34, 35, 141 e 142.

DELAMARO, M.; JINO, M.; MALDONADO, J. *Introdução ao teste de software*. [S.l.]: Elsevier Brasil, 2017. Citado na página 28.

DERSIMONIAN, R.; LAIRD, N. Meta-analysis in clinical trials. *Controlled clinical trials*, Elsevier, v. 7, n. 3, p. 177–188, 1986. Citado na página 42.

DUAN, L.; XU, T. A short text similarity algorithm for finding similar police 110 incidents. In: IEEE. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. [S.l.], 2016. p. 260–264. Citado 4 vezes nas páginas 34, 35, 142 e 145.

ERVILHA, G. T.; LIMA, J. E. D. Um método econométrico na identificação dos determinantes da criminalidade municipal: a aplicação em minas gerais, brasil (2000-2014). *Economía, sociedad y territorio*, El Colegio Mexiquense AC, v. 19, n. 59, p. 1059–1086, 2019. Citado 4 vezes nas páginas 95, 115, 118 e 119.

FARIAS, A. M. G. de et al. Definition of strategies for crime prevention and combat using fuzzy clustering and formal concept analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, World Scientific, v. 26, n. 03, p. 429–452, 2018. Citado 8 vezes nas páginas 18, 34, 56, 64, 65, 101, 102 e 143.

FARIAS, M. A. d. F. et al. Identifying technical debt through a code comment mining tool. In: *Proceedings of the XV Brazilian Symposium on Information Systems*. [S.l.: s.n.], 2019. p. 1–8. Citado na página 28.

FENG, M. et al. Big data analytics and mining for crime data analysis, visualization and prediction. In: SPRINGER. *International Conference on Brain Inspired Cognitive Systems*. [S.l.], 2018. p. 605–614. Citado 3 vezes nas páginas 34, 35 e 141.

FILHO, D. B. F. et al. O que é, para que serve e como se faz uma meta-análise? *Teoria & Pesquisa: Revista de Ciência Política*, v. 23, n. 2, 2014. Citado na página 41.

FRAGA, L. d. S. et al. Transparência da gestão pública: Análise em pequenos municípios do rio grande do sul. *Gestão & Planejamento-G&P*, v. 20, 2019. Citado na página 97.

FREITAS, R. K. V. de; DACORSO, A. L. R. Inovação aberta na gestão pública: análise do plano de ação brasileiro para a open government partnership. *Revista de administração pública*, v. 48, n. 4, p. 869–888, 2014. Citado na página 16.

GALBRUN, E.; PELECHRINIS, K.; TERZI, E. Urban navigation beyond shortest route: The case of safe paths. *Information Systems*, Elsevier, v. 57, p. 160–171, 2016. Citado na página 30.

GOMES, W. F.; COLAÇO JÚNIOR, M.; PRADO, K. H. d. J. Um experimento inicial sobre associações entre os crimes ocorridos no brasil. *Research, Society and Development*, v. 9, n. 11, p. e41791110078–e41791110078, 2020. Citado na página 123.

GONSALVES, E. C. Regras de associações e suas medidas de interesse objetivas e subjetivas [Association rules and their objective and subjective measures of interest]. *INFOCOMP*, v. 4, n. 1, p. 26–35, 2004. Citado na página 62.

GUIMARÃES, F. F. et al. Comparison phenotypic and genotypic identification of staphylococcus species isolated from bovine mastitis. *Pesquisa Veterinária Brasileira*, SciELO Brasil, v. 36, n. 12, p. 1160–1164, 2016. Citado na página 108.

GUPTA, M.; CHANDRA, B.; GUPTA, M. A framework of intelligent decision support system for indian police. *Journal of Enterprise Information Management*, Emerald Group Publishing Limited, v. 27, n. 5, p. 512–540, 2014. Citado 8 vezes nas páginas 16, 26, 34, 35, 54, 93, 141 e 142.

HARDY, K.; MAURUSHAT, A. Opening up government data for big data analysis and public benefit. *Computer Law & Security Review*, Elsevier, v. 33, n. 1, p. 30–37, 2017. Citado 2 vezes nas páginas 16 e 97.

HARRISON, T. M. et al. Open government and e-government: Democratic challenges from a public value perspective. *Information Polity*, IOS Press, v. 17, n. 2, p. 83–97, 2012. Citado na página 16.

HIGGINS, J. P. et al. Measuring inconsistency in meta-analyses. *Bmj*, British Medical Journal Publishing Group, v. 327, n. 7414, p. 557–560, 2003. Citado na página 42.

HUANG, S.-M. A study of the application of data mining on the spatial landscape allocation of crime hot spots. In: *Geo-Informatics in Resource Management and Sustainable Ecosystem*. [S.l.]: Springer, 2013. p. 274–286. Citado 2 vezes nas páginas 35 e 141.

HUANG, Y.-Y.; LI, C.-T.; JENG, S.-K. Mining location-based social networks for criminal activity prediction. In: IEEE. *2015 24th Wireless and Optical Communication Conference (WOCC)*. [S.l.], 2015. p. 185–189. Citado 3 vezes nas páginas 35, 141 e 142.

IBGE. *Instituto Brasileiro de Geografia e Estatistica [Brazilian Institute of Geography and Statistics]*. 2020. Accessed: 2020-03-01. Disponível em: <https://www.ibge.gov.br>. Citado na página 69.

IBRAHIM, R.; SHAFIQ, M. O. On the measurement and analysis of safety in a large city. In: IEEE. *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*. [S.l.], 2017. p. 1068–1075. Citado na página 30.

ISAFIADE, O. E.; BAGULA, A. B. Citisafe: Adaptive spatial pattern knowledge using fp-growth algorithm for crime situation recognition. In: IEEE. *2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing*. [S.l.], 2013. p. 551–556. Citado 3 vezes nas páginas 19, 35 e 142.

JAMES, K. L.; RANDALL, N. P.; HADDAWAY, N. R. A methodology for systematic mapping in environmental sciences. *Environmental evidence*, BioMed Central, v. 5, n. 1, p. 7, 2016. Citado na página 29.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. *Information systems management*, Taylor & Francis, v. 29, n. 4, p. 258–268, 2012. Citado 3 vezes nas páginas 16, 54 e 93.

JOHANSSON, E.; GÅHLIN, C.; BORG, A. Crime hotspots: An evaluation of the kde spatial mapping technique. In: IEEE. *2015 European Intelligence and Security Informatics Conference*. [S.l.], 2015. p. 69–74. Citado 2 vezes nas páginas 34 e 142.

JURISTO, N.; MORENO, A. M. *Basics of software engineering experimentation*. [S.l.]: Springer Science & Business Media, 2013. Citado 3 vezes nas páginas 23, 58 e 96.

KADAR, C.; PLETIKOSA, I. Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science*, Springer, v. 7, n. 1, p. 26, 2018. Citado 3 vezes nas páginas 35, 141 e 143.

KEELE, S. et al. *Guidelines for performing systematic literature reviews in software engineering*. [S.l.], 2007. Citado na página 29.

KELER, A.; MAZIMPAKA, J. D. Safety-aware routing for motorised tourists based on open data and vgi. *Journal of location Based services*, Taylor & Francis, v. 10, n. 1, p. 64–77, 2016. Citado 2 vezes nas páginas 34 e 141.

KEYVANPOUR, M. R.; EBRAHIMI, M. R.; JAVIDEH, M. Designing efficient ann classifiers for matching burglaries from dwelling houses. *Applied Artificial Intelligence*, Taylor & Francis, v. 26, n. 8, p. 787–807, 2012. Citado 3 vezes nas páginas 35, 143 e 144.

KIM, S. et al. Crime analysis through machine learning. In: IEEE. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*. [S.l.], 2018. p. 415–420. Citado 4 vezes nas páginas 35, 55, 141 e 142.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, v. 33, n. 2004, p. 1–26, 2004. Citado 4 vezes nas páginas 23, 27, 28 e 29.

KUMAR, M. et al. Forecasting of annual crime rate in india: A case study. In: IEEE. *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. [S.l.], 2018. p. 2087–2092. Citado 2 vezes nas páginas 35 e 142.

KUO, N.; CHANG, C.-M.; CHEN, K.-T. Exploring spatial and social factors of crime: A case study of taipei city. In: SPRINGER. *Asian Conference on Intelligent Information and Database Systems*. [S.l.], 2017. p. 3–13. Citado 7 vezes nas páginas 26, 35, 43, 50, 52, 141 e 142.

LAURIKKALA, J. et al. Informal identification of outliers in medical data. In: *Fifth international workshop on intelligent data analysis in medicine and pharmacology*. [S.l.: s.n.], 2000. v. 1, p. 20–24. Citado na página 61.

LI, S.-T.; KUO, S.-C.; TSAI, F.-C. An intelligent decision-support model using fsom and rule extraction for crime prevention. *Expert Systems with Applications*, Elsevier, v. 37, n. 10, p. 7108–7119, 2010. Citado 2 vezes nas páginas 34 e 142.

LI, X. et al. Development of crime in england and wales 1898–2001: Data mining using self-organising map. In: IEEE. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*. [S.l.], 2017. p. 1–8. Citado 4 vezes nas páginas 34, 35, 141 e 142.

LI, X. et al. Gdp growth vs. criminal phenomena: data mining of japan 1926–2013. *AI & SOCIETY*, Springer, p. 1–14, 2018. Citado 4 vezes nas páginas 34, 35, 141 e 142.

LI, Z. et al. Spatio-temporal pattern analysis and prediction for urban crime. In: IEEE. *2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)*. [S.l.], 2018. p. 177–182. Citado 4 vezes nas páginas 35, 142, 143 e 144.

LIMA, A.; VIGNATTI, A.; SILVA, M. Reconhecimento de grafos power-law por algoritmos de aprendizagem de máquina utilizando um conjunto reduzido de propriedades estruturais [Recognizing power-law graphs by machine learning algorithms using a reduced set of structural features]. In: SBC. *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.], 2020. p. 611–621. Citado 2 vezes nas páginas 64 e 101.

LIMA, R. S.; BUENO, S. 13 º anuário brasileiro de segurança pública. *São Paulo: Fórum Brasileiro de Segurança Pública*, 2019. Citado 2 vezes nas páginas 21 e 22.

LIMA, R. S.; BUENO, S. Anuário brasileiro de segurança pública 2020. *Fórum Brasileiro de Segurança Pública. São Paulo*, 2020. Citado na página 119.

LINKEDIN. *Prova de conceito (PoC) em projetos*. 2015. Disponível em: <https://www.linkedin.com/pulse/prova-de-conceito-poc-em-projetos-silva-pmp-prince2-practitioner>. Citado na página 28.

MA, L.; CHEN, Y.; HUANG, H. Ak-modes: A weighted clustering algorithm for finding similar case subsets. In: IEEE. *2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*. [S.l.], 2010. p. 218–223. Citado 3 vezes nas páginas 34, 141 e 143.

MACIEL, G. S. et al. Eficiência técnica da polícia militar: um estudo dos comandos de policiamento regionais do distrito federal por meio da análise envoltória de dado. Universidade Federal de Goiás, 2019. Citado na página 98.

MARZAN, C. S. et al. Time series analysis and crime pattern forecasting of city crime data. In: ACM. *Proceedings of the International Conference on Algorithms, Computing and Systems*. [S.l.], 2017. p. 113–118. Citado 14 vezes nas páginas 15, 25, 30, 35, 53, 56, 62, 92, 94, 98, 141, 142, 143 e 144.

MEDINA, E. U.; PAILAQUILÉN, R. M. B. A revisão sistemática e a sua relação com a prática baseada na evidência em saúde. *Revista Latino-Americana de Enfermagem*, SciELO Brasil, v. 18, n. 4, p. 1–8, 2010. Citado 2 vezes nas páginas 41 e 42.

MJSP. *Portal Brasileiro de Dados Abertos [Brazilian Open Data Portal]*. 2020. Accessed: 2020-01-11. Disponível em: <http://www.dados.gov.br/dataset/sistema-nacional-de-estatisticas-de-seguranca-publica>. Citado 2 vezes nas páginas 17 e 68.

MOHAN, P. et al. Cascading spatio-temporal pattern discovery. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 24, n. 11, p. 1977–1992, 2011. Citado 2 vezes nas páginas 34 e 141.

MONTEIRO, R. N. M. *Metodologias de meta-análise aplicadas nas Ciências da Saúde*. Tese (Doutorado) — Universidade da Beira Interior, 2010. Citado na página 41.

NASCIMENTO, D. O. d. Transparência pública: uma análise do município de alagoa nova-pb. Universidade Estadual da Paraíba-UEPB, 2019. Citado na página 97.

NETO, J. V. et al. Boxplot: um recurso gráfico para a análise e interpretação de dados quantitativos [Boxplot: A graphic resource for the analysis and interpretation of quantitative data]. *Revista Odontológica do Brasil Central*, v. 26, n. 76, 2017. Citado na página 60.

NOOR, N. M. M. et al. A review on a classification framework for supporting decision making in crime prevention. *Journal of Artificial Intelligence*, Asian Network for Scientific Information (ANSINET), v. 8, n. 1, p. 17, 2015. Citado na página 26.

OLIVEIRA, M. et al. Spatio-temporal variations in the urban rhythm: the travelling waves of crime. *EPJ Data Science*, SpringerOpen, v. 7, n. 1, p. 29, 2018. Citado 2 vezes nas páginas 35 e 144.

OLIVEIRA, R. N. d.; COLAÇO JÚNIOR, M. Experimental analysis of stemming on jurisprudential documents retrieval. *Information*, Multidisciplinary Digital Publishing Institute, v. 9, n. 2, p. 28, 2018. Citado na página 67.

ORONG, M. Y.; SISON, A. M.; HERNANDEZ, A. A. Mitigating vulnerabilities through forecasting and crime trend analysis. In: IEEE. *2018 5th International Conference on Business and Industrial Research (ICBIR)*. [S.l.], 2018. p. 57–62. Citado 3 vezes nas páginas 35, 141 e 142.

OZGUL, F. et al. Combined detection model for criminal network detection. In: SPRINGER. *Pacific-Asia Workshop on Intelligence and Security Informatics*. [S.l.], 2010. p. 1–14. Citado 3 vezes nas páginas 35, 142 e 143.

OZGUL, F. et al. Mining hate crimes to figure out reasons behind. In: IEEE COMPUTER SOCIETY. *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*. [S.l.], 2012. p. 887–889. Citado 4 vezes nas páginas 34, 141, 143 e 145.

PARANHOS, R. et al. Desvendando os mistérios do coeficiente de correlação de pearson: o retorno [Unraveling the mysteries of pearson's correlation coefficient: the return]. *Leviathan (São Paulo)*, n. 8, p. 66–95, 2014. Citado 2 vezes nas páginas 63 e 100.

PEREIRA, M. G.; GALVÃO, T. F. Heterogeneidade e viés de publicação em revisões sistemáticas. *Epidemiologia e Serviços de Saúde*, SciELO Public Health, v. 23, p. 775–778, 2014. Citado na página 27.

PETERSEN, K. et al. Systematic mapping studies in software engineering. In: *Ease*. [S.l.: s.n.], 2008. v. 8, p. 68–77. Citado na página 29.

PETERSEN, K.; VAKKALANKA, S.; KUZNIARZ, L. Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, Elsevier, v. 64, p. 1–18, 2015. Citado na página 28.

PHILLIPS, P.; LEE, I. Crime analysis through spatial areal aggregated density patterns. *Geoinformatica*, Springer, v. 15, n. 1, p. 49–74, 2011. Citado 8 vezes nas páginas 15, 26, 35, 93, 141, 142, 143 e 144.

PHILLIPS, P.; LEE, I. Mining co-distribution patterns for large crime datasets. *Expert Systems with Applications*, Elsevier, v. 39, n. 14, p. 11556–11563, 2012. Citado 2 vezes nas páginas 35 e 143.

POSTGRESQL. *Installing Procedural Languages*. 2019. Accessed: 2019-10-01. Disponível em: <https://www.postgresql.org/docs/current/xplang-install.html>. Citado 2 vezes nas páginas 59 e 104.

PRADO, K. H. d. J.; COLAÇO JÚNIOR, M. Data science aplicada à análise criminal baseada nos dados abertos governamentais de minas gerais. *Research, Society and Development*, v. 9, n. 11, p. e36391110044–e36391110044, 2020. Citado 3 vezes nas páginas 24, 92 e 123.

PRADO, K. H. d. J.; COLAÇO JÚNIOR, M. Data science applied to crime analysis based on brazilian open government data. *Journal of Applied Security Research*, Taylor & Francis, 2020. Citado 3 vezes nas páginas 24, 53 e 123.

PRADO, K. H. d. J. et al. Applied intelligent data analysis to government data related to criminal incident: A systematic review. *Journal of Applied Security Research*, Taylor & Francis, p. 1–35, 2020. Citado 6 vezes nas páginas 23, 24, 25, 57, 96 e 122.

RAJESWARI P.SURYA TEJA, P. H. T. K. D. Enhancing the performance of crime prediction technique using data mining. *International Journal of Engineering & Technology*, n. 7 (2.32), p. 424–426, 2018. Citado 2 vezes nas páginas 35 e 141.

READ, S. Applications of case study research. *Nurse Researcher*, Royal College of Nursing Publishing Company (RCN), v. 10, n. 4, p. 93–95, 2003. Citado na página 28.

REWARI, S.; SINGH, W. Systematic review of crime data analytics. In: IEEE. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. [S.l.], 2017. p. 3042–3045. Citado na página 27.

RIED, K. Interpreting and understanding meta-analysis graphs: a practical guide. Royal Australian College of General Practitioners, 2006. Citado na página 41.

RODRIGUES, C. L. Metanálise: um guia prático. 2010. Citado na página 42.

R.SUJATHA, D. D. An adaptive method for analyzing and predicting the crime locations by means of amabc and arm. *Journal of Theoretical and Applied Information Technology*, p. 45–56, 2014. Citado 4 vezes nas páginas 34, 35, 141 e 142.

RUIZ, M. D. et al. Meta-association rules for fusing regular association rules from different databases. In: IEEE. *17th International Conference on Information Fusion (FUSION)*. [S.l.], 2014. p. 1–7. Citado 3 vezes nas páginas 35, 56 e 143.

RUMI, S. K.; DENG, K.; SALIM, F. D. Crime event prediction with dynamic features. *EPJ Data Science*, SpringerOpen, v. 7, n. 1, p. 43, 2018. Citado 3 vezes nas páginas 35, 141 e 142.

RUMI, S. K.; DENG, K.; SALIM, F. D. Theft prediction with individual risk factor of visitors. In: ACM. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. [S.l.], 2018. p. 552–555. Citado 2 vezes nas páginas 35 e 141.

SALTOS, G.; COCEA, M. An exploration of crime prediction using data mining on open data. *International Journal of Information Technology & Decision Making*, World Scientific, v. 16, n. 05, p. 1155–1181, 2017. Citado 3 vezes nas páginas 35, 141 e 142.

SAMPAIO, R. F.; MANCINI, M. C. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica [Systematic review studies: a guide for careful synthesis of the scientific evidence]. SciELO Brasil, 2007. Citado na página 29.

SANDIG, J. D. E. et al. Mining online gis for crime rate and models based on frequent pattern analysis. In: *Proceedings of the World Congress on Engineering and Computer Science*. [S.l.: s.n.], 2013. v. 2, p. 23–27. Citado 2 vezes nas páginas 35 e 141.

SANTOS, B. S.; COLAÇO JÚNIOR, M.; SOUZA, J. G. de. A initial experimental evaluation of the neuromessenger: A collaborative tool to improve the empathy of text interactions. In: *Information Technology-New Generations*. [S.l.]: Springer, 2018. p. 411–419. Citado 2 vezes nas páginas 67 e 105.

SANTOS, E. J. F. D.; CUNHA, M. Interpretação crítica dos resultados estatísticos de uma meta-análise: Estratégias metodológicas. *Millenium*, Instituto Politécnico de Viseu, n. 44, p. 85–89, 2013. Citado 2 vezes nas páginas 41 e 42.

SANTOS, R. M. et al. Long term-short memory neural networks and word2vec for self-admitted technical debt detection. In: *ICEIS (2)*. [S.l.: s.n.], 2020. p. 157–165. Citado na página 105.

SCOPUS. *Scopus - Elsevier Database*. 2019. Disponível em: <https://www.scopus.com>. Citado 3 vezes nas páginas 29, 39 e 51.

SEJUSP. *Portal da Secretaria de Estado de Justiça e Segurança Pública de Minas Gerais*. 2020. <http://www.seguranca.mg.gov.br>. Accessed: 2020-01-19. Citado 2 vezes nas páginas 98 e 107.

SharePSI. *Share-PSI: Uses of Open Data Within Government for Innovation and Efficiency: Report*. 2014. <https://www.w3.org/2013/share-psi/workshop/samos/report>. Accessed: 2019-10-16. Citado na página 16.

SILVA, L. J. S. et al. Crimevis: An interactive visualization system for analyzing crime data in the state of rio de janeiro. In: *ICEIS (1)*. [S.l.: s.n.], 2017. p. 193–200. Citado 3 vezes nas páginas 34, 143 e 144.

SILVARES, A. C. Políticas públicas em segurança no brasil: Avanços e novos desafios. *Revista Científica Doctum Direito*, v. 1, n. 3, 2019. Citado 2 vezes nas páginas 116 e 118.

SINGH, C. B. K. N.; JOSHI, J. D. Data mining for prevention of crimes. In: . [S.l.]: Springer Verlag, 2018. Citado 10 vezes nas páginas 18, 19, 30, 34, 35, 55, 57, 62, 94 e 141.

SIVARANJANI, S.; AASHA, M.; SIVAKUMARI, S. Hot spot identification using kernel density estimation for serial crime detection. In: SPRINGER. *International Conference on Soft Computing Systems*. [S.l.], 2018. p. 253–265. Citado 2 vezes nas páginas 35 e 144.

SIVARANJANI, S.; SIVAKUMARI, S.; AASHA, M. Crime prediction and forecasting in tamilnadu using clustering approaches. In: IEEE. *2016 International Conference on Emerging Technological Trends (ICETT)*. [S.l.], 2016. p. 1–6. Citado 4 vezes nas páginas 34, 55, 141 e 142.

SRIDHAR, R.; SATHYRAJ, S.; BALASUBRAMANIAM, S. Analysis and pattern deduction on linguistic, numeric based mean and fuzzy association rule algorithm on any geo-referenced crime point data integrated with google map. In: SPRINGER. *Proceedings of the International Conference on Soft Computing for Problem Solving (SocProS 2011) December 20-22, 2011*. [S.l.], 2012. p. 15–27. Citado 2 vezes nas páginas 34 e 141.

SUN, C.-c. et al. Detecting crime types using classification algorithms. *JDIM*, v. 12, n. 5, p. 321–327, 2014. Citado 3 vezes nas páginas 35, 141 e 142.

SURVE, R.; LU, M.; DAI, J. Xquery data mining tool for campus security. In: IEEE. *2015 IEEE International Conference on Information Reuse and Integration*. [S.l.], 2015. p. 193–196. Citado 2 vezes nas páginas 35 e 141.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to data mining*. [S.l.]: Pearson Education India, 2016. Citado na página 60.

TAYAL, D. K. et al. Crime detection and criminal identification in india using data mining techniques. *AI & society*, Springer, v. 30, n. 1, p. 117–127, 2015. Citado 3 vezes nas páginas 34, 35 e 141.

TEAM, R. C. *R: A Language and Environment for Statistical Computing*. 2014. Disponível em: <https://www.r-project.org/>. Citado na página 43.

THOTA, L. S. et al. Cluster based zoning of crime info. In: IEEE. *2017 2nd International Conference on Anti-Cyber Crimes (ICACC)*. [S.l.], 2017. p. 87–92. Citado 3 vezes nas páginas 34, 35 e 141.

TOMAR, N.; MANJHVAR, A. K. An improved optimized clustering technique for crime detection. In: IEEE. *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. [S.l.], 2016. p. 1–5. Citado na página 27.

TOMAR, N.; MANJHVAR, A. K. Role of clustering in crime detection: Application of fuzzy k-means. In: *Advances in Computer and Computational Sciences*. [S.l.]: Springer, 2018. p. 591–599. Citado 2 vezes nas páginas 34 e 141.

TOPPIREDDY, H. K. R.; SAINI, B.; MAHAJAN, G. Crime prediction & monitoring framework based on spatial analysis. *Procedia computer science*, Elsevier, v. 132, p. 696–705, 2018. Citado 8 vezes nas páginas 15, 19, 34, 53, 54, 92, 93 e 141.

TRANSPARÊNCIA Portal. *Portal da Transparência do Governo Federal (Federal Government's Transparency Portal)*. 2020. Disponível em: <http://www.portaltransparencia.gov.br/>. Citado na página 97.

TRAVASSOS, G. H.; BARROS, M. O. Contributions of in virtuo and in silico experiments for the future of empirical studies in software engineering. In: *2nd Workshop on Empirical Software Engineering the Future of Empirical Studies in Software Engineering*. [S.l.: s.n.], 2003. p. 117–130. Citado 3 vezes nas páginas 68, 89 e 96.

TURET, J. G.; COSTA, A. P. C. S. Big data analytics to improve the decision-making process in public safety: A case study in northeast brazil. In: SPRINGER. *International Conference on Decision Support System Technology*. [S.l.], 2018. p. 76–87. Citado 2 vezes nas páginas 35 e 144.

VINEETH, K. S.; PANDEY, A.; PRADHAN, T. A novel approach for intelligent crime pattern discovery and prediction. In: IEEE. *2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*. [S.l.], 2016. p. 531–538. Citado 2 vezes nas páginas 35 e 141.

WANG, M. et al. Hybrid neural network mixed with random forests and perlin noise. In: IEEE. *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. [S.l.], 2016. p. 1937–1941. Citado 2 vezes nas páginas 35 e 143.

WANG, T. et al. Finding patterns with a rotten core: Data mining for crime series with cores. *Big Data*, Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, v. 3, n. 1, p. 3–21, 2015. Citado 2 vezes nas páginas 34 e 143.

WANG, X.; BROWN, D. E.; GERBER, M. S. Spatio-temporal modeling of criminal incidents using geographic, demographic, and twitter-derived information. In: IEEE. *2012 IEEE International Conference on Intelligence and Security Informatics*. [S.l.], 2012. p. 36–41. Citado 2 vezes nas páginas 35 e 145.

WEI, X. et al. Analysis of crime rate distribution based on tpml-wma. In: IEEE. *2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. [S.l.], 2016. p. 157–160. Citado 2 vezes nas páginas 34 e 145.

WOHLIN, C. et al. *Experimentation in software engineering*. [S.l.]: Springer Science & Business Media, 2012. Citado 8 vezes nas páginas 17, 23, 28, 54, 58, 66, 96 e 105.

WU, J. et al. Computing exact permutation p-values for association rules. *Information Sciences*, Elsevier, v. 346, p. 146–162, 2016. Citado 2 vezes nas páginas 63 e 100.

YADAV MEET TIMBADIA, A. Y. R. V. S.; YADAV, N. Crime pattern det ection, analysis & prediction. In: . [S.l.]: Institute of Electrical and Electronics Engineers Inc., 2017. Citado 5 vezes nas páginas 34, 35, 57, 94 e 141.

YANG, C. et al. A rough-fuzzy c-means using information entropy for discretized violent crimes data. In: IEEE. *13th International Conference on Hybrid Intelligent Systems (HIS 2013)*. [S.l.], 2013. p. 23–27. Citado 2 vezes nas páginas 34 e 142.

YU, P.-H.; LAY, J.-G. Exploring non-stationarity of local mechanism of crime events with spatial-temporal weighted regression. In: IEEE. *Proceedings 2011 IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services*. [S.l.], 2011. p. 7–12. Citado 2 vezes nas páginas 35 e 144.

ZHENG, Z.; KOHAVI, R.; MASON, L. Real world performance of association rule algorithms. In: ACM. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2001. p. 401–406. Citado 2 vezes nas páginas 70 e 108.

ZHU, S.; XIE, Y. Crime incidents embedding using restricted boltzmann machines. In: IEEE. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.], 2018. p. 2376–2380. Citado 2 vezes nas páginas 35 e 143.

ZHUANG, Y. et al. Crime hot spot forecasting: A recurrent model with spatial and temporal information. In: IEEE. *2017 IEEE International Conference on Big Knowledge (ICBK)*. [S.l.], 2017. p. 143–150. Citado 7 vezes nas páginas 26, 34, 43, 50, 52, 142 e 145.

Apêndices

APÊNDICE A – Lista de referências por algoritmo

Tabela 28 – Referências por algoritmo.

Algoritmo	Referências
K-Means	(TAYAL et al., 2015), (OZGUL et al., 2012), (BHARATHI; INDRANI; PRABAKAR, 2017), (MA; CHEN; HUANG, 2010), (CHAN; LEONG, 2010), (SRIDHAR; SATHYRAJ; BALASUBRAMANIAM, 2012), (ANSARI; PRAKASH et al., 2018), (DAS; DAS, 2017), (ALKHAIBARI; CHUNG, 2017), (THOTA et al., 2017), (YADAV MEET TIMBADIA; YADAV, 2017), (TOPPIREDDY; SAINI; MAHAJAN, 2018), (SIVARANJANI; SIVAKUMARI; AASHA, 2016), (CAVADAS; BRANCO; PEREIRA, 2015), (SINGH; JOSHI, 2018), (CALVO et al., 2017), (BUCZAK; GIFFORD, 2010), (ORONG; SISON; HERNANDEZ, 2018) e (TOMAR; MANJHVAR, 2018)
K-Nearest Neighbors (KNN)	(KIM et al., 2018), (TAYAL et al., 2015), (SUN et al., 2014), (CHAN; LEONG, 2010), (SALTOS; COCEA, 2017), (BELESIOTIS; PAPADAKIS; SKOUTAS, 2018), (DAS; DAS, 2019), (FENG et al., 2018), (THOTA et al., 2017), (TOPPIREDDY; SAINI; MAHAJAN, 2018), (SIVARANJANI; SIVAKUMARI; AASHA, 2016), (CAVADAS; BRANCO; PEREIRA, 2015), (LI et al., 2017), (LI et al., 2018a) e (KELER; MAZIMPAKA, 2016)
Apriori	(PHILLIPS; LEE, 2011), (MARZAN et al., 2017), (SURVE; LU; DAI, 2015), (HUANG, 2013), (R.SUJATHA, 2014), (DAS; DAS, 2017), (YADAV MEET TIMBADIA; YADAV, 2017), (BALOIAN CORONEL ENRIQUE BASSALETTI, 2017), (SINGH; JOSHI, 2018), (RAJESWARI P.SURYA TEJA, 2018), (BUCZAK; GIFFORD, 2010), (CRANDELL; KORKMAZ, 2018) e (SANDIG et al., 2013)
Naive Bayes	(DAS; DAS, 2019), (DAS; DAS, 2017), (FENG et al., 2018), (MOHAN et al., 2011), (YADAV MEET TIMBADIA; YADAV, 2017), (TOPPIREDDY; SAINI; MAHAJAN, 2018), (CAVADAS; BRANCO; PEREIRA, 2015), (LI et al., 2017), (CALVO et al., 2017), (LI et al., 2018a), (BACULO et al., 2017) e (BOGAHAWATTE; ADIKARI, 2013)
Random Forest (RF)	(VINEETH; PANDEY; PRADHAN, 2016), (RUMI; DENG; SALIM, 2018a), (KUO; CHANG; CHEN, 2017), (KADAR; PLETIKOSA, 2018), (HUANG; LI; JENG, 2015), (RUMI; DENG; SALIM, 2018b), (BAPPEE; JÚNIOR; MATWIN, 2018), (DAS; DAS, 2019), (FENG et al., 2018), (CAVADAS; BRANCO; PEREIRA, 2015), (LI et al., 2017) e (LI et al., 2018a)
Support Vector Machine (SVM)	(VINEETH; PANDEY; PRADHAN, 2016), (RUMI; DENG; SALIM, 2018a), (KUO; CHANG; CHEN, 2017), (HUANG; LI; JENG, 2015), (BAPPEE; JÚNIOR; MATWIN, 2018), (BONI; GERBER, 2016a), (THOTA et al., 2017), (CAVADAS; BRANCO; PEREIRA, 2015), (LI et al., 2017) e (LI et al., 2018a)
Decision Tree	(GUPTA; CHANDRA; GUPTA, 2014), (KUO; CHANG; CHEN, 2017), (HUANG, 2013), (DAS; DAS, 2019), (FENG et al., 2018), (LI et al., 2018a) e (BACULO et al., 2017)

Continuação na próxima página

Tabela 28 – *Referências por algoritmo (continuação).*

Algoritmo	Referências
Self-Organizing Map (SOM)	(BENGTSSON; HEIN; OLSSON, 2012), (DUAN; XU, 2016), (R.SUJATHA, 2014), (LI; KUO; TSAI, 2010), (LI et al., 2017), (LI et al., 2018a) e (ANDRIENKO et al., 2010)
DBCSAN	(CATLETT et al., 2018), (CHAN; LEONG, 2010), (ALKHAIBARI; CHUNG, 2017), (SIVARANJANI; SIVAKUMARI; AASHA, 2016) e (CALVO et al., 2017)
AutoRegressive Integrated Moving Average (ARIMA)	(CATLETT et al., 2018) (LI et al., 2018b) (KUMAR et al., 2018) e (ORONG; SISON; HERNANDEZ, 2018)
Hierarchical Agglomerative Clustering (HAC)	(BELESIOTIS; PAPADAKIS; SKOUTAS, 2018), (ALKHAIBARI; CHUNG, 2017) e (SIVARANJANI; SIVAKUMARI; AASHA, 2016)
Linear Regression	(HUANG; LI; JENG, 2015), (MARZAN et al., 2017) e (AWAL et al., 2016)
Logistic Regression (LR)	(RUMI; DENG; SALIM, 2018a), (KUO; CHANG; CHEN, 2017) e (BAPPEE; JÚNIOR; MATWIN, 2018)
Naive Bayesian	(GUPTA; CHANDRA; GUPTA, 2014), (SUN et al., 2014) e (KUO; CHANG; CHEN, 2017)
Adaptive Boosting (AdaBoost)	(KIM et al., 2018) e (DAS; DAS, 2019)
Artificial Neural Network (ANN)	(GUPTA; CHANDRA; GUPTA, 2014) e (RUMI; DENG; SALIM, 2018a)
C-Fuzzy Rough-means	(YANG et al., 2013) e (ANSARI; PRAKASH et al., 2018)
Ensemble method	(BAPPEE; JÚNIOR; MATWIN, 2018) e (RUMI; DENG; SALIM, 2018a)
Frequent Pattern growth (FP-growth)	(ISAFIADE; BAGULA, 2013) e (AGRAWAL; SEJWAR, 2017)
Kernel Density Estimation (KDE)	(JOHANSSON; GÅHLIN; BORG, 2015) e (LI et al., 2017)
M5 model trees (M5P)	(CHAN; LEONG, 2010) e (SALTOS; COCEA, 2017)
Recurrent Neural Network (RNN)	(OZGUL et al., 2010) e (ZHUANG et al., 2017)
Spark-based shared nearest neighbor clustering algorithm (SparkSNN)	(ARYAL; WANG, 2018)
Adaptive Mutation-based Artificial Bee Colony (AMABC)	(R.SUJATHA, 2014)
Artificial Bee Colony (ABC)	(R.SUJATHA, 2014)
Association Rule Mining (ARM)	(R.SUJATHA, 2014)
Breadth First Search (BFS)	(PHILLIPS; LEE, 2011)
C4.5	(SUN et al., 2014)
Classification And Regression Trees (CART)	(LI et al., 2017)

Continuação na próxima página

Tabela 28 – *Referências por algoritmo (continuação).*

Algoritmo	Referências
Criado pelos autores (não nomeado)	(WANG et al., 2015)
Cyclic Signature (CS)	(CHAN; LEONG, 2010)
Density tracing technique	(PHILLIPS; LEE, 2012)
Depth First Search (DFS)	(PHILLIPS; LEE, 2011)
Ensemble Neural Network Crime Classifier	(KEYVANPOUR; EBRAHIMI; JAVIDEH, 2012)
Expectation maximization (EM)	(CHAN; LEONG, 2010)
Extremely Randomized Tree	(KADAR; PLETIKOSA, 2018)
Feedforward backpropagation network	(BENGTSSON; HEIN; OLSSON, 2012)
Fusing Information with Meta-association rules	(RUIZ et al., 2014)
Fuzzy K-Means	(FARIAS et al., 2018)
Fuzzy C-means	(OZGUL et al., 2010)
Gaussian Processes	(MARZAN et al., 2017)
Gaussian-Bernoulli Restricted Boltzmann Machine (GB RBM)	(ZHU; XIE, 2018)
Global Model(GM)	(LI et al., 2018b)
Gradient-Boosting	(KADAR; PLETIKOSA, 2018)
Guided Local Search (GLS)	(PHILLIPS; LEE, 2011)
Hierarchical similarity algorithm	(CHI et al., 2017)
Hybrid Neural Network	(WANG et al., 2016)
Infomap	(DAS; DAS, 2017)
J48	(BACULO et al., 2017)
K-Medoids	(SILVA et al., 2017)
K-Modes	(MA; CHEN; HUANG, 2010)
Kohonen	(OZGUL et al., 2012)
LibLinear	(BONI; GERBER, 2016b)
Linear SVC	(AGHABABAEI; MAKREHCHI, 2015)

Continuação na próxima página

Tabela 28 – *Referências por algoritmo (continuação).*

Algoritmo	Referências
MDS+k-medoids	(SILVA et al., 2017)
Modified Graph Cut Clustering (MGCC)	(SIVARANJANI; AASHA; SIVAKUMARI, 2018)
Morlet wavelet	(OLIVEIRA et al., 2018)
Multiclass MLP Crime Classifier	(KEYVANPOUR; EBRAHIMI; JAVIDEH, 2012)
Multilayer Perceptron (MLP)	(MARZAN et al., 2017)
Multi-Objective Particle Swarm Optimization (MOPSO)	(AGRAWAL; SEJWAR, 2017)
Multiple-Level Spatial Association Rules	(CHEN et al., 2015)
Multivariate Adaptive Regression Splines (MARS)	(CAVADAS; BRANCO; PEREIRA, 2015)
Multivariate Time Series Clustering (MTS Clustering)	(CHANDRA; GUPTA, 2013)
Não informado	(DAMASCENO; TEIXEIRA; CAMPOS, 2012)
Nearest Neighbor (NN-5)	(PHILLIPS; LEE, 2011)
OneR	(BACULO et al., 2017)
Pooled Model (PM) and Hierarchical Model (HM)	(LI et al., 2018b)
Promethee II	(TURET; COSTA, 2018)
Sequential Minimum Optimization Regression (SMOreg)	(MARZAN et al., 2017)
Smote	(CAVADAS; BRANCO; PEREIRA, 2015)
Social crime data aware Kernel Density Estimation based Serial crime Detection (SAKDESD)	(SIVARANJANI; AASHA; SIVAKUMARI, 2018)
Space-time kernel density (STKD)	(YU; LAY, 2011)
Space-time weighted regression (STWR)	(YU; LAY, 2011)
Spatial Kluster Analysis by Tree Edge Removal (SKATER)	(SILVA et al., 2017)
Spatio Temporal Clustering	(BELESIOTIS; PAPADAKIS; SKOUTAS, 2018)

Continuação na próxima página

Tabela 28 – *Referências por algoritmo (continuação).*

Algoritmo	**Referências**
Spatio-Temporal Generalized Additive Modeling (STGAM)	(WANG; BROWN; GERBER, 2012)
Spatio-Temporal Neural Network (STNN)	(ZHUANG et al., 2017)
Taxonomic similarity algorithm	(CHI et al., 2017)
Two-way	(OZGUL et al., 2012)
Weighted majority algorithm (WMA)	(WEI et al., 2016)
Word Mover's Distance (WMD)	(DUAN; XU, 2016)
Word2Vec	(DUAN; XU, 2016)