Research, Society and Development, v. 9, n. 11, e41791110078, 2020
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v9i11.10078

Um Experimento Inicial sobre Associações entre os Crimes Ocorridos no Brasil

An Initial Experiment on Associations between Crimes in Brazil

Un primer experimento sobre asociaciones entre crímenes en Brasil

Recebido: 09/11/2020 | Revisado: 13/11/2020 | Aceito: 16/11/2020 | Publicado: 19/11/2020

Wesckley Faria Gomes
ORCID: https://orcid.org/0000-0003-1278-5506
Universidade Federal de Sergipe, Brasil
E-mail: wesckley.gomes@gmail.com
Methanias Colaço Junior
ORCID: https://orcid.org/0000-0002-4811-1477
Universidade Federal de Sergipe, Brasil
E-mail: mjrse@hotmail.com
Kleber Henrique de Jesus Prado
ORCID: https://orcid.org/0000-0002-3555-9475
Universidade Federal de Sergipe, Brasil
E-mail: kleprado@hotmail.com

Resumo

Contexto: A criminalidade tem sido um problema ao redor do mundo, causando danos às sociedades. Educação, pobreza, emprego e clima são alguns fatores que afetam a taxa de criminalidade, levando as autoridades a gastar, anualmente, milhões com ações de combate à violência e planos estratégicos de prevenção e redução da criminalidade. Objetivo: Aplicar conceitos de *Data Science* para análise de dados governamentais relacionados a crimes no Brasil. Método: Uso de mineração de dados, especificamente regras de associação (RA), em um experimento controlado, para detecção de padrões entre os tipos de crimes, como também entre os tipos de crime e meses do ano. Resultados: No contexto das associações entre crimes, os estados com regras mais interessantes foram: Bahia, com 15 associações, São Paulo, com 12, Goiás, 11, e Paraná, com 9. Destaque para a associação "Latrocínio ⟹ Roubo de Carga", encontrada para o Estado da Bahia, a qual atingiu uma confiança de 99% (0.99). Já no âmbito das associações entre crimes e meses do ano, avultaram-se Paraíba, com 2 associações, Maranhão, Rondônia e São Paulo, com 1 associação cada. Destaque para regra "Dezembro ⟹

Roubo de Veículo", encontrada para o Estado de São Paulo, que alcançou uma confiança de 84% (0,84). Conclusão: Os resultados expostos nesta pesquisa auxiliam analistas criminais no processo de tomada de decisão para prevenção e redução da criminalidade no país, uma vez que podem permitir a inibição de crimes que são antecedentes de outras ocorrências dentro do mesmo estado ou de crimes que ocorrem num mesmo período.

Palavras-chave: Análise criminal; Criminalidade; Data science; Mineração de dados; Python e R juntos.

Abstract

Context: Crime has been a problem around the world, causing damage to societies. Education, poverty, employment and climate are some of the factors that affect the crime rate, leading authorities to spend millions annually on actions to combat violence and strategic plans to prevent and reduce crime. Objective: Applying Data Science concepts to analyze government data related to crimes in Brazil. Method: Use of data mining techniques of association rules (AR), in a controlled experiment, to detect patterns between the types of crimes, as well as the relationship between the types of crime and the months of the year. Results: In the context of associations between crimes, the states with the most interesting rules were: Bahia, with 15 associations, São Paulo, with 12 associations, Goiás, with 11, and Paraná, with 9. Highlight for the association "Robbery Resulting in Death ⟹ Cargo Roberry", found for the State of Bahia, which reached 99% confidence (0.99). In the scope of associations between crimes and months of the year, Paraíba had 2 associations, Maranhão, Rondônia and São Paulo, with 1 association each. Highlight for rule "December ⟹ Vehicle Steal", found for the State of São Paulo, which reached a confidence of 84% (0.84). Conclusion: The results exposed in this research assist criminal analysts in the decision-making process to prevent and reduce crime in the country, since they can allow priority in inhibiting crimes that are antecedents of other occurrences within the same state or crimes that occur in the same period.

Keywords: Criminal analysis; Crime; Data science; Data mining; Python and R together.

Resumen

Contexto: Delincuencia ha sido un problema en todo el mundo y ha causado daños a las sociedades. Educación, la pobreza, el empleo y el clima son algunos de los factores que inciden en la tasa de criminalidad, lo que lleva a las autoridades a gastar anualmente millones en acciones para combatir la violencia y planes estratégicos para prevenir y reducir la

criminalidad. Objetivo: Aplicar conceptos de Data Science para analizar datos gubernamentales relacionados con delitos en Brasil. Método: Utilización de minería de datos, específicamente reglas de asociación (AR), en un experimento controlado, para detectar patrones entre los tipos de delitos, así como entre los tipos de delitos y los meses del año. Resultados: En el contexto de las asociaciones entre delitos, los estados con las reglas más interesantes fueron: Bahía, con 15 asociaciones, São Paulo, con 12, Goiás, 11, y Paraná, con 9. Destacar para la asociación "Robo seguido de muerte ⟹ Robo de Carga", encontrada para el Estado de Bahia, que alcanzó el 99% de confianza (0,99). En el ámbito de las asociaciones entre delitos y meses del año, Paraíba creció, con 2 asociaciones, Maranhão, Rondônia y São Paulo, con 1 asociación cada una. Destacado por regla "Deciembre ⟹ Robo de Vehículos", encontrado para el estado de São Paulo, que alcanzó una confianza del 84% (0,84). Conclusión: Los resultados expuestos en esta investigación ayudan analistas penales en el proceso de toma de decisiones para prevenir y reducir la delincuencia en el país, ya que pueden permitir la inhibición de delitos que son antecedentes de otros sucesos dentro del mismo estado o delitos que ocurren en el mismo período.

Palabras clave: Análisis criminal; Crimen; Ciencia de los datos; Procesamiento de datos; Python y R juntos.

1. Introdução

Todos os dias, vemos nos noticiários relatos sobre crimes cometidos ao redor do mundo. Fatores como educação, pobreza, emprego e clima contribuem para o aumento da taxa de criminalidade. Infelizmente, essa é uma realidade vivida por diferentes sociedades e os transtornos causados por esse fenômeno despertam o interesse de pesquisadores no mundo inteiro (Melo, Teixeira, & Campos, 2012), investigando as ações que, via de regra, são condenadas pela maioria da sociedade e exigem punição legal do governo (Sevri, Karacan, & Akcayol, 2017).

A percepção geral de segurança em uma cidade está entre os principais fatores que afetam diretamente a qualidade de vida dos cidadãos. Os incidentes, por mais relativamente menores que sejam, como um furto, por exemplo, podem causar transtornos significativos, afetando a vida e o bem-estar de uma população (Belesiotis, Papadakis, & Skoutas, 2018). Essa violência causa diversos problemas à sociedade e um dos papéis dos governantes e agências de segurança pública é tentar cada vez mais diminuir a criminalidade, atuando na

prevenção e redução desses índices. Para isso, anualmente, são disponibilizados recursos de combate à criminalidade, que precisam ser utilizados de forma inteligente e eficaz.

Nos últimos anos, a tecnologia vem sendo uma grande aliada no combate à criminalidade. Atualmente, diversas organizações públicas e departamentos de polícia vêm aumentando sua capacidade de coletar e armazenar dados detalhados de eventos criminais (Catlett, Cesario, Talia, & Vinci, 2018). Isso é de vital importância na melhoria dos resultados das investigações e detecções criminais, facilitando o registro, a análise de recuperação e o compartilhamento de informações (Gupta, Chandra, & Gupta, 2014). O próprio Governo Federal brasileiro disponibiliza diversas informações para toda a população, através do Portal Brasileiro de Dados Abertos do Ministério da Justiça e Segurança Pública, disponível em (MSJP, 2020), as quais contêm dados criminais dos últimos anos, para todas as Unidades Federativas (UFs) do Brasil.

Com esse grande volume de informações disponível, surge o desafio da análise e transformação desses dados em conhecimento útil para a tomada de decisão do governo. Assim, torna-se imprescindível o uso de técnicas de mineração de dados para análise de padrões de crimes e apoio à corporação policial, otimizando a alocação de recursos e melhorando a produtividade dos policiais na prevenção e redução da criminalidade (Pereira & Brandão, 2014). Todavia, não obstante as informações sobre estatísticas oficiais de segurança pública disponibilizadas pelos governos federais e estaduais serem uma das formas de levar transparência para população, nenhuma parece promover informações claras suficientes para o entendimento da população em geral, com milhares de fluxos de dados publicados sem métricas que possibilitem ao cidadão deduzir conclusões a respeito do que foi publicado.

Baseando-se nesse contexto, o presente artigo propôs aplicar Data Science e realizar um experimento controlado, seguindo as orientações contidas em (Nogueira de Oliveira & Colaço Júnior, 2018) e tendo como base os dados abertos coletados do Ministério da Justiça e Segurança Pública (MJSP), com a finalidade de verificar se existem associações entre os tipos de crimes, bem como entre tipos de crimes e meses do ano no Brasil, visando auxiliar as autoridades na tomada de decisão para prevenção e redução da criminalidade no país.

Os resultados mostraram que São Paulo, com a maior população do país, registrou o maior número de ocorrências criminais, entre os anos de 2015 e 2020, com 1.010.726 incidentes, seguido pelo Rio de Janeiro e Paraná, que apesar de não seguirem essa ordem em termos populacionais (IBGE, 2020), contam com 425.553 e 363.780 incidentes, respectivamente. Alguns estados se avultaram no que diz respeito às associações entre os tipos de crimes, entre os quais, Bahia, com 15 associações, São Paulo, com 12 associações,

Goiás, com 11, e Paraná, com 9 regras de associação encontradas. Destaque para a associação Latrocínio ⟹ Roubo de Carga", encontrada para o Estado da Bahia, a qual atingiu uma confiança de 99% (0.99) e pode ser lida da seguinte forma: para 99% dos casos de latrocínio, dentro do mesmo mês, ocorre um roubo de carga. Já no que diz respeito às associações entre crimes e meses do ano, avultaram-se Paraíba, com 2 regras de associação encontradas, Maranhão, Rondônia e São Paulo, com 1 associação cada. Neste caso, o destaque foi para regra "Dezembro ⟹ Roubo de Veículo", encontrada para o Estado de São Paulo, que alcançou uma confiança de 84% (0,84).

A organização do restante deste artigo sucede da seguinte forma: Na seção 2, são demonstrados os conceitos que foram aplicados no presente artigo. A seção 3 apresenta a metodologia utilizada neste experimento. A seção 4 fornece um levantamento de trabalhos relacionados ao tema. A seção 5 mostra as definições e o planejamento do experimento. A seção 6 detalha a operação do experimento. Na seção 7, são apresentados os resultados experimentais obtidos e, finalmente, na seção 8, discute-se a conclusão e os trabalhos futuros.

2. Base Conceitual

Nesta seção, serão apresentados os conceitos e ferramentas utilizadas para realização do experimento.

2.1 Arquitetura

A principal ferramenta utilizada no experimento foi o Sistema de Gerenciamento de Banco de Dados (SGBD) *PostgreSQL*. A partir deste SGBD, foi possível centralizar a arquitetura para executar os passos do processo, que vão desde a obtenção dos *datasets* até a detecção dos padrões criminais. O *PostgreSQL* é um banco de dados de código fonte aberto, desenvolvido pela universidade da Califórnia, em Berkeley (PostgreSQL, 2020).

Além de ser um banco de dados relacional, o *PostgreSQL*, por meio de suas Linguagens Procedurais (LPs), permite que funções definidas pelo usuário sejam escritas em outras linguagens tais como Java, Python, R, Perl, entre outras. Nesse caso, o servidor de banco de dados não possui nenhum conhecimento interno de como interpretar o código fonte, e, em vez disso, passa a tarefa para um módulo especial que conhece os detalhes da linguagem. As linguagens procedurais utilizadas neste experimento foram:

- ***PL/pgSQL***: Linguagem nativa do *PostgreSQL*, foi utilizada no processo de Extração, Transformação e Carga (ETL) dos dados e desmembramento dos dados para geração das transações.
- ***PL/Python***: Linguagem procedural que permite a execução de funções escritas em *Python* e foi utilizada para realizar o download, descompactação e conversão dos arquivos.
- ***PL/R***: Linguagem procedural que permite a execução de funções escritas em *R* e foi utilizada para executar o algoritmo *Apriori* e gerar as regras de associação do experimento.

2.2 Regras de Associação

Para encontrar associações entre os crimes, utilizou-se uma técnica de mineração de dados chamada de Regras de Associação (RA), que combina itens que ocorrem com uma determinada frequência, e é definida da seguinte forma: Seja I = {*i1, i2, ...im*} um conjunto de itens (produtos, crimes, etc) e D uma base de dados formada por um conjunto de transações (uma venda, ocorrência de crimes em um período de tempo, etc), onde cada transação T é composta por um conjunto de itens (*itemset*), tal que $T \subseteq I$. Uma regra de associação é definida como uma implicação do tipo $A \Rightarrow B$, onde $A \subset I$, $B \subset I$ e $A \cap B = \emptyset$ (Varde, et al., 2004). A é denominado antecedente e B denominado consequente da regra. Esta técnica é útil para descobrir relações entre conjuntos de variáveis, que podem representar produtos em uma loja, sintomas de doenças, palavras-chave, crimes, características demográficas, entre outras (Gillmeister & Cazella, 2007). Por exemplo, a regra "chocolate (A) $\Rightarrow$ adoçante (B)", na qual se pode ler: quem compra chocolate (antecedente) também compra adoçante (consequente). Ou ainda, "roubo $\Rightarrow$ homicídio", na qual se pode ler: os casos de roubo implicam casos de homicídio em um curto espaço de tempo, um dia, uma semana ou um mês.

Uma das abordagens mais utilizadas para encontrar tais regras é o algoritmo *Apriori*. Inicialmente proposto por R.Agrawal e R.Srikant, em 1994, baseia-se no conhecimento prévio das propriedades frequentes dos itens (Sumithra & Paul, 2010). Este algoritmo pode produzir uma grande quantidade de regras desinteressantes. Para que isto seja evitado, segundo (Lallich, Teytaud, & Prudhomme, 2007), é preciso escolher as medidas de interesse mais

adequadas ao problema em questão e, em seguida, validar as regras interessantes em relação às medidas selecionadas. O presente artigo utilizou 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. O suporte de uma regra de associação A ⟹ B, Sup (A ⟹ B), é dado por Sup (A U B) .

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

- **Confiança:** A confiança da regra A ⟹ B, Conf (A ⟹ B) representa, dentre as transações que contêm A, a porcentagem de transações que também contêm B, e é dada por Conf (A ⟹ B) = Sup (A U B) ÷ Sup (A).

$$Conf\,(A \Rightarrow B) = \frac{Número\ total\ de\ transações\ que\ contém\ A\ e\ B}{Total\ de\ transações\ que\ contém\ A}$$

- **Lift:** O Lift avalia a dependência entre o conjunto de itens do antecedente e o conjunto de itens do consequente de uma RA. O valor do Lift de uma regra A ⟹ B indica o quanto mais frequente torna-se B quando ocorre em conjunto com A. Quando *Lift*(A ⟹ B) = 1, significa que A e B são independentes, ou seja, não existe associação entre eles. Se *Lift*(A ⟹ B) > 1, então A e B são positivamente dependentes. Se Se *Lift*(A ⟹ B) < 1, então A e B são negativamente dependentes.

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

- **Count:** É a frequência de ocorrências 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, e varia em intervalo de -1 a 1. O sinal indica a direção da correlação (negativa ou positiva), enquanto o valor indica a magnitude. Se a correlação for positiva (r = 1), indica que se A aumentar, B também aumenta, isto é, valores altos de A estão associados a valores altos de B. Já se a correlação for negativa (r = -1), indica que se A aumentar, B diminui, ou

seja, valores altos de A estão associados a valores baixos de B. Caso a correlação for 0 (r = 0), significa que as variáveis são ortogonais entre si (ausência de correlação) (Paranhos, et al., 2014).

- **Qui-Quadrado (X^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, ou nível de significância (α) igual a 95%, 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).
- **Leverage:** representa o número de transações adicionais cobertas pelos lados direito e esquerdo de uma regra, além do esperado, caso os dois fossem independentes um do outro. O valor do *leverage* maior que 0 indica que os dois lados da regra ocorreriam juntos, em um número de transações maior que o esperado, caso os itens encontrados nas regras fossem completamente independentes. Caso o valor seja menor que 0, os dois lados da regra ocorrem juntos, menos que o esperado, e caso seja igual a 0, os dois lados da regra ocorrem juntos, exatamente o esperado, indicando que os dois lados provavelmente são independentes. Deste modo, quanto maior o *leverage* mais interessante será a regra (Procaci da Silva, 2004).

$$Lev(A \Rightarrow B) = Sup(A \Rightarrow B) - Sup(A) \times Sup(B)$$

3. Metodologia

Nesta seção, será apresentada a metodologia utilizada no experimento.

A metodologia adotada para o trabalho envolveu, inicialmente, uma Revisão Sistemática (RS) (Pereira et al., 2018) 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.

Em seguida, para realização do experimento, foram coletados dados históricos sobre crimes, por meio do Portal Brasileiro de Dados Abertos, disponibilizado pelo Ministério da Justiça e Segurança Pública do Governo Federal (MSJP, 2020). O conjunto de dados fornece informações sobre o Estado, tipo de crime, mês, ano e número de ocorrências dos crimes cometidos. Os arquivos contêm informações de todos os estados brasileiros e do Distrito Federal, durante os anos de 2015 a 2020.

Segundo (Marzan, Baculo, Bulos, & Ruiz, 2017), o pré-processamento de dados é um procedimento essencial no processo de mineração de dados, para obtenção dos dados adequados à análise. Neste sentido, os arquivos disponibilizados são grandes e de difícil compreensão por parte do cidadão. Então, para facilitar a leitura e interpretação do conteúdo, utilizando um programa feito em Python, os arquivos, inicialmente em formato .XLSX (formato de planilha eletrônica), foram transformados em arquivos de formato .CSV (arquivo de valores separados por vírgula) e carregados para um banco de dados estruturado, criado no *PostgreSQL*, com a estrutura da Tabela 1. A Tabela 1 reflete uma amostra das ocorrências criminais de janeiro de 2015, no Acre, contendo 6 estupros, 13 homicídios dolosos e 2 tentativas de homicídio.

Tabela 1. Amostra dos dados carregados para o PostgreSQL.

UF	Tipo Crime	Ano	Mês	Ocorrências
Acre	Estupro	2015	janeiro	6
Acre	Furto de veículo	2015	janeiro	0
Acre	Homicídio doloso	2015	janeiro	13
Acre	Lesão corporal seguida de morte	2015	janeiro	0
Acre	Roubo a instituição financeira	2015	janeiro	0
Acre	Roubo de carga	2015	janeiro	0
Acre	Roubo de veículo	2015	janeiro	0
Acre	Roubo seguido de morte (latrocínio)	2015	janeiro	0
Acre	Tentativa de homicídio	2015	janeiro	2

Fonte: Autores.

Com os dados brutos armazenados no banco de dados, foi possível gerar as transações que seriam passadas para o algoritmo *Apriori* identificar as regras de associação. Para tal, foi preciso desmembrar as ocorrências em linhas, conforme Tabela 2, que mostra um exemplo de como as ocorrências criminais de janeiro de 2015 do Acre foram armazenadas após o desmembramento. Percebe-se que a quantidade de ocorrências para os tipos de crimes

estupro, homicídio doloso e tentativa de homicídio continuam as mesmas, apenas armazenadas de forma diferente. Portanto, a amostra que inicialmente continha 21 incidentes criminais, passa a contar com 13 transações que serão passadas ao algoritmo. Em outras palavras, como a granularidade dos dados fornecidos pelo governo é mensal, nós consideramos que os crimes que ocorrem no mesmo mês pertencem à mesma transação, a qual terá seu conjunto usado para determinar os crimes que mais ocorrem juntos, dentro do espaço de um mês. Situação análoga a uma compra feita por um cliente, em que a transação é considerada a partir de uma nota fiscal gerada, sendo os produtos da nota pertencentes à mesma transação, usada para averiguar posteriormente os produtos que mais são comprados juntos, ou seja, que são comprados ao mesmo tempo. Em alguns casos, grandes estruturas de lojas e *Market Places* utilizam transações com intervalo maior, observando produtos que são comprados juntos dentro de uma semana ou dentro de um mês. O fato é que, no nosso trabalho, os produtos são os crimes e os crimes que ocorrem no mesmo mês foram computados como concomitantes.

Tabela 2. Amostra dos dados desmembrados para o algoritmo *Apriori*.

Ano Mês	UF	Crime1	Crime2	Crime3	Crime4	Crime5	Crime6	Crime7	Crime8	Crime9
201501	Acre	Estupro	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	Tentativa de Homicidio
201501	Acre	Estupro	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	Tentativa de Homicidio
201501	Acre	Estupro	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	Estupro	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	Estupro	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	Estupro	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL
201501	Acre	NULL	NULL	Homicidio doloso	NULL	NULL	NULL	NULL	NULL	NULL

Fonte: Autores.

Com as transações prontas, uma nova função foi criada para execução do algoritmo *Apriori*, utilizando a linguagem de programação R. Após a execução do algoritmo, os resultados foram filtrados e armazenados em uma nova tabela. Por fim, com os resultados armazenados, foram extraídas as informações relevantes referentes às regras de associação geradas.

Do ponto de vista da classificação metodológica, este trabalho pode ser classificado como de laboratório e experimental, devido ao planejamento e à execução de um experimento

controlado in vitro, o qual tem o seu método descrito de forma autocontida, no seu planejamento, detalhado na seção 5 deste artigo.

Em resumo, foram realizadas 3 etapas macro para o experimento: (1) identificação e download dos arquivos em formato .XLSX; (2) realização de processo de ETL (Extract, Transform e Load), que consistiu em criar as estruturas de tabelas no banco de dados, transformar e carregar os dados para esta estrutura de armazenamento; (3) programação para execução do algoritmo Apriori, para detecção de padrões e posteriores análise, seleção e validação dos dados.

4. Trabalhos Relacionados

A geração de regras de associação é uma técnica de mineração de dados não supervisionada, utilizada para detecção de conjuntos de itens frequentes, revelando se existe relação entre estes (Sevri, Karacan, & Akcayol, 2017). A técnica tem se mostrado eficaz quando utilizada no campo das análises criminais, e vários países já se beneficiaram dessa abordagem. Nesta seção, apresentaremos alguns trabalhos que utilizaram a mineração de dados e regras de associação como ferramentas para análise criminal.

Em (Marzan, Baculo, Bulos, & Ruiz, 2017), foi conduzida uma análise de dados geoespaciais para detectar atividades criminosas na cidade de Manila, Filipinas. Os autores utilizaram o algoritmo *Apriori* para gerar regras de associação dos crimes com atributos criminais e não encontraram nenhum padrão frequente para o crime abuso sexual, o que pode significar que esse tipo de crime ocorre em ambientes espaciais e temporais aleatórios. Em contrapartida, os distritos de Sampaloc, Tondo, Malate e Ermita foram as consequências comuns para diferentes tipos de crimes e a maioria dos padrões mostraram que os crimes são mais prováveis de acontecer quando não está chovendo, em áreas residenciais. Além disso, segundo os autores, a associação dos crimes com o tempo pode ajudar as agências de aplicação da lei a posicionar mais policiais em um determinado local e em um determinado momento.

(Sevri, Karacan, & Akcayol, 2017) também utilizaram a mineração de dados com o apoio do algoritmo *Apriori*, para encontrar relações entre os atributos de registros criminais, empregando a base de dados NIBRS (National Incident-Based Reporting System), que contém 4.998.574 registros criminais dos EUA, no ano de 2013. Atributos como estado, raça, sexo, idade e data foram usados para detecção das regras, as quais extraíram informações tais como: se o sexo da vítima é feminino, a etnia da vítima não é hispânica e nem latina, e a etnia

do agressor é branca, então, a etnia da vítima também é branca e a cena do crime é o lar; se as etnias da vítima e do agressor forem brancas e a cena do crime for o lar, então, o sexo da vítima é masculino.

O trabalho apresentado em (Yadav, Timbadia, Yadav, Vishwakarma, & Yadav, 2017) utilizou dados abertos retirados do portal online da Índia, listando vários tipos de crimes, tais como assassinatos, sequestros, roubos e estupros, entre outros. Nessa pesquisa, foram aplicados diferentes algoritmos de análise de dados que inferiram as conexões entre os crimes e seus padrões, entre os quais estavam *Apriori*, *K-means*, *Naive Bayes*, por meio das ferramentas Weka e linguagem de programação R. O algoritmo *Apriori* mostrou associação entre pessoas presas e soltas no mesmo ano, indicando que quanto maior o número de pessoas presas, mais pessoas são soltas e, portanto, mais pessoas acabam sendo inocentadas.

Em (Singh, Kaverappa, & Joshi, 2018), os autores utilizaram várias técnicas de mineração de dados, supervisionadas e não supervisionadas, como regressão linear múltipla, *K-Means* e análise de regras de associação, com o objetivo de mostrar a eficácia da mineração de dados no domínio da prevenção do crime. Utilizando os dados do departamento de polícia de Gujarat, os pesquisadores descobriram que os casos relacionados a roubo são os menos resolvidos e isto pode ser um dos motivos pelos quais roubo é considerado um dos crimes mais ocorridos nesta localidade. Por fim, o trabalho de (Huang, 2013) apresenta um estudo dos pontos críticos de crimes coletados pela polícia de Taipei, capital de Taiwan. Com o uso de informações geográficas e a tecnologia de mineração de dados, essa pesquisa detectou regras de associação entre os pontos críticos dos crimes e a paisagem espacial, bem como a distância entre estes.

Já no âmbito nacional, alguns estudos sobre a criminalidade têm sido realizados. Mais recentemente, o trabalho de (Barros, Baggio, Stege, & Hilgemberg, 2019) teve como objetivo analisar a distribuição espacial da criminalidade em todos os 5.565 municípios brasileiros, investigando a relação dessa variável com o desenvolvimento econômico de cada localidade. Observou-se que, em regra geral, municípios com alto desenvolvimento econômico tendem a ser cercados por municípios com baixos índices de criminalidade (e vice-versa). No entanto, para alguns municípios, como, por exemplo, partes do Rio de Janeiro, Espírito Santo e região metropolitana de Curitiba, o nível de desenvolvimento econômico não é capaz de barrar o avanço da criminalidade.

Em (Pereira & Brandão, 2014), os autores criaram uma nova abordagem de análise de dados criminais. Foi construído um novo modelo multidimensional para armazenamento adequado dos dados de eventos criminais e aplicado o algoritmo *Apriori* para reconhecer

implicações mútuas entre as ocorrências. A base de dados utilizada representava dois anos de dados de eventos criminais fornecidos pelo governo brasileiro e algumas informações relevantes foram extraídas como, por exemplo, inferiu-se, com uma confiança mínima de 0,92, que os crimes violentos que resultam em lesões, cometidos por jovens de 17 a 29 anos, são fortemente motivados pela ganância e por questões financeiras. Outro conjunto de regras mostrou, com uma confiança mínima de 0,82, que os picos de furto ocorrem das 8h às 11h, enquanto os picos de roubo ocorrem por volta de 20h. Segundo os autores, apesar de as regras de associação descobertas por si só não caracterizarem o comportamento criminoso, elas apontam para padrões de crimes não triviais que motivam investigações adicionais.

É importante fazer uma comparação da proposta deste trabalho com os demais trabalhos citados. Vale ressaltar que algumas pesquisas utilizaram bases de dados que, minimamente, possuem dados (informações) semelhantes às disponibilizadas pelo governo brasileiro. Na maioria dos casos, em uma situação diferente do Brasil, o maior detalhamento dos dados fornecidos pelos governos permite o uso de mais opções de algoritmos e melhor uso do algoritmo *Apriori*, o qual foi aqui empregado, na proporção da incompletude dos dados disponíveis. Neste contexto, para a realização do experimento, foram usadas as seguintes informações dos registros criminais: estado, ano, mês, tipo de crime e quantidade de ocorrências. Como comparativo, a pesquisa de (Sevri, Karacan, & Akcayol, 2017) teve outros atributos disponíveis, tais como raça, sexo, idade, etc. Finalmente, vale ressaltar que não foram encontrados trabalhos para descoberta de associações entre os tipos de crimes e entre meses do ano e crime, no Brasil, como é proposto neste artigo. Além disso, este trabalho seguiu um processo experimental, com validação dos dados, o qual facilita o processo de replicação, para aumento da base de conhecimento e para averiguação de que outros cientistas, de forma independente, chegarão aos mesmos resultados.

5. Definição e Planejamento do Experimento

Nesta seção, serão apresentados os objetivos e todo o planejamento da experimentação. O mesmo segue as diretrizes de processos experimentais com publicações recentes (Nogueira de Oliveira & Colaço Júnior, 2018) (Santos, Colaço Júnior, & Souza, 2018).

5.1 Definição do Objetivo

O objetivo deste trabalho é analisar as possíveis associações entre os tipos de crimes ocorridos nos estados brasileiros, bem como entre os tipos de crimes e meses do ano, utilizando dados abertos de ocorrências criminais, disponibilizados pelo Ministério de Justiça e Segurança Pública (MJSP). Para tal, foi realizado um experimento in vitro, em um ambiente controlado elaborado, e aplicado o algoritmo Apriori, para gerar as regras de associação e determinar as combinações de itens que ocorrem com determinada frequência, bem como gerar as medidas de interesse que, estatisticamente, pudesse servir de base para medir a força de tais regras.

Baseando-se no modelo GQM (Goal Question Metric) (Basili, et al., 2014) (Basili & Weiss, A Methodology for Collecting Valid Software Engineering Data, 1984), o objetivo deste trabalho foi formalmente sintetizado como: analisar ocorrências criminais no território brasileiro, com a finalidade de avaliá-las, com relação à detecção de associações entre os tipos de crimes, bem como entre os tipos de crimes e meses do ano, do ponto de vista de analistas criminais, cientistas de dados e cidadãos, no contexto de dados abertos sobre ocorrências criminais do MJSP - Brasil.

5.2 Planejamento

Formulação de Hipóteses: Não foram encontrados estudos experimentais que analisaram as possíveis associações entre os tipos de crimes ocorridos nos estados brasileiros, bem como as associações entre os tipos de crimes e meses do ano. Desta forma, baseadas nessas premissas, duas questões de pesquisa foram formalizadas para esse trabalho, são elas:

- **Q1:** Existem associações interessantes entre os tipos de crimes praticados no Brasil?
- **Q2:** Existem associações interessantes entre os tipos de crimes e os meses do ano?

Para responder à questão **Q1**, é preciso analisar as regras de associação geradas do tipo "Crime $\Rightarrow$ Crime", com o objetivo de refutar a seguinte hipótese nula (H_0) e não refutar seguinte hipótese alternativa (H_1):

- **H_0:** Os crimes são independentes entre si.
- **H_1:** Os crimes são dependentes entre si.

Para responder à questão **Q2**, é preciso analisar as regras de associação geradas do tipo "Mês ⟹ Crime", com o objetivo de refutar a seguinte hipótese nula (H_0) e não refutar seguinte hipótese alternativa (H_1):

- **H_0:** Os crimes são independentes dos meses do ano, ou seja, não estão associados.
- **H_1:** Os crimes são dependentes dos meses do ano, ou seja, estão associados.

Seleção de Contexto: Para realização do experimento, foram coletados dados de ocorrências criminais dos 26 Estados brasileiros e do Distrito Federal.

Seleção dos Participantes e Objetos: Os dados coletados foram obtidos do portal brasileiro de dados abertos do Ministério da Justiça e Segurança Pública, localizado em (MSJP, 2020). O período de informações coletadas foi o total disponível e vai de janeiro de 2015 a março de 2020, totalizando 3.576.663 incidentes criminais.

Variáveis Dependentes: Representa o efeito do tratamento, a saída do processo de experimentação e deriva diretamente da hipótese. As variáveis dependentes utilizadas para validação das hipóteses foram as regras de associação geradas com seus Suporte e Confiança, das quais se pode extrair outras medidas de interesse objetivas para auxiliar na identificação das forças destas regras de associação: Lift, Coeficiente de correlação de Pearson (r), Qui-Quadrado(X^2), com seu nível de significância (*p-value*) e Alavancagem (leverage).

Variáveis Independentes: Representa a causa que afeta o resultado do experimento. As variáveis independentes referem-se à entrada do processo de experimentação (Travassos, Gurov, & Amaral, 2002). Neste experimento, foram consideradas as seguintes variáveis independentes: algoritmo *Apriori*, a base de dados disponibilizada pelo portal do MJSP, as medidas Suporte e Confiança mínimos, intervalos aceitáveis de Lift e *p-value* máximo. Os limiares para os valores das medidas de interesse serão detalhados na seção de coleta de dados.

Instrumentação: As ferramentas e materiais necessários para a execução do experimento foram:

- Arquivos com scripts de criação do projeto de banco de dados;
- Arquivo com ocorrências criminais, disponibilizado pelo MJSP;
- Banco de Dados PostgreSQL, versão 11.6-3;
- Python, versão 3.7.6;
- Software livre R, versão 3.6.0.

6. Operação do Experimento

Nesta seção, será descrita como foi a preparação e execução do experimento.

6.1 Preparação

Na etapa de preparação do ambiente, foram instalados o SGBD *PostgreSQL* e os ambientes de execução do *Python* e do *R*. O banco de dados serviu para armazenar todos os dados e executar todas as etapas da experimentação, tendo sido configurado para executar códigos *Python* e *R*, por meio das suas respectivas linguagens procedurais *PL/Python* e *PL/R*. Em seguida, foram criados scripts para criação das tabelas e demais procedimentos necessários para realização do experimento. Com a preparação, os arquivos puderam ser baixados e o ambiente estava pronto para iniciar a execução.

6.2 Execução

De posse do ambiente preparado, conforme seção anterior, o primeiro passo foi carregar os arquivos contendo os incidentes criminais para uma tabela do banco de dados. Para isso, uma função escrita em *PL/Python* realizou o download dos arquivos com as ocorrências criminais, como também a conversão de arquivos Excel (.XLSX) em arquivos .CSV, formato aceito pelo *PostgreSQL*. Na etapa seguinte, um script executou o processo de ETL, no qual os dados foram transformados, para eliminação de possíveis inconsistências, e armazenados em uma tabela no banco de dados. Como o arquivo trouxe a soma dos crimes ocorridos em um determinado ano, mês e estado, foi necessário realizar um desmembramento dos dados, para gerar as transações. Um script *PL/pgSQL* realizou o desmembramento do bando de dados, que incialmente contava com 15.180 registros, cuja soma de ocorrências era de 3.576.663 incidentes e passou a ter 1.765.534 transações, após desmembramento, conforme explicado na seção 3. Em seguida, um procedimento implementado usando a linguagem procedural *PL/R* foi executado para gerar as regras de associação, por meio do algoritmo *Apriori*. Vale ressaltar que todas as etapas podem ser executadas por meio de um único procedimento, o módulo principal, no banco de dados.

6.2.1 Coleta dos dados

Como visto na seção 2.2, o suporte de uma RA é dado pela relação entre a quantidade de transações em que aparecem os conjuntos de itens A e B, e o total de transações. Como os

Estados possuem uma quantidade diferente de ocorrências criminais, foi calculado o percentual de suporte mínimo para cada Estado, em outras palavras, considerando o tamanho da base de cada estado. Os percentuais estipulados para as bases de cada estado representam o mínimo de transações que uma regra, para cada estado, deve ter. Para tal, como não temos como prever o número de crimes que realmente ocorrem em um estado (população desconhecida), adotamos o valor aproximado 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 para uma regra ser considerada, nas quais os itens A e B da regra aparecem juntos (crime e mês ou crime A e crime B, por exemplo), para cada estado, deve ser de 800 (arredondamento para cima).

A Tabela 3 mostra a quantidade de transações e o suporte mínimo para cada Estado, cujo cálculo do percentual se deu por meio da divisão 800/(quantidade de transações). Neste mesmo contexto de parametrização do algoritmo e com o intuito de filtrar as regras de associação mais relevantes para o estudo, uma confiança mínima de 70% (0,70) foi adotada. Este valor foi considerado 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, 2018). Assim, analogamente, a confiança mínima adotada nesta pesquisa permitirá a geração de regras com confianças fortes e muito fortes.

Tabela 3. Número de transações e suporte mínimo por Estado.

Estado	Transações	Sup. Mínimo
Acre	4.139	0.1932
Alagoas	16.696	0.0479
Amazonas	17.054	0.0469
Amapá	3.892	0.2055
Bahia	70.233	0.0113
Ceará	47.960	0.0166
Distrito Federal	30.762	0.0260
Espírito Santo	22.897	0.0349
Goiás	62.386	0.0128
Maranhão	21.409	0.0373
Minas Gerais	128.966	0.0062
Mato Grosso do Sul	19.267	0.0415
Mato Grosso	15.714	0.0509
Pará	32.623	0.0245
Paraíba	11.108	0.0720
Pernambuco	76.417	0.0104
Piauí	18.768	0.0426
Paraná	185.953	0.0043
Rio de Janeiro	227.564	0.0035
Rio Grande do Norte	31.485	0.0254
Rondônia	14.172	0.0564
Roraima	3.936	0.2032
Rio Grande do Sul	89.707	0.0089
Santa Catarina	60.634	0.0131
Sergipe	12.742	0.0627
São Paulo	530.225	0.0015
Tocantins	8.825	0.0906
Total	**1.765.534**	

Fonte: Autores.

Para o Lift, o valor deve 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 fora destes limites serão consideradas enganosas e descartadas, pois indicam que, em verdade, o antecedente diminui a probabilidade de o consequente ocorrer, ainda que numa pequena proporção. Por fim, como foi adotada uma confiabilidade de 95%, ou seja, um $\alpha = 0,05$, o *p-value* do qui-quadrado, para rejeição da hipótese nula de independência e para aceitação da regra, deve ser menor que 0,05.

6.2.2 Validação dos dados

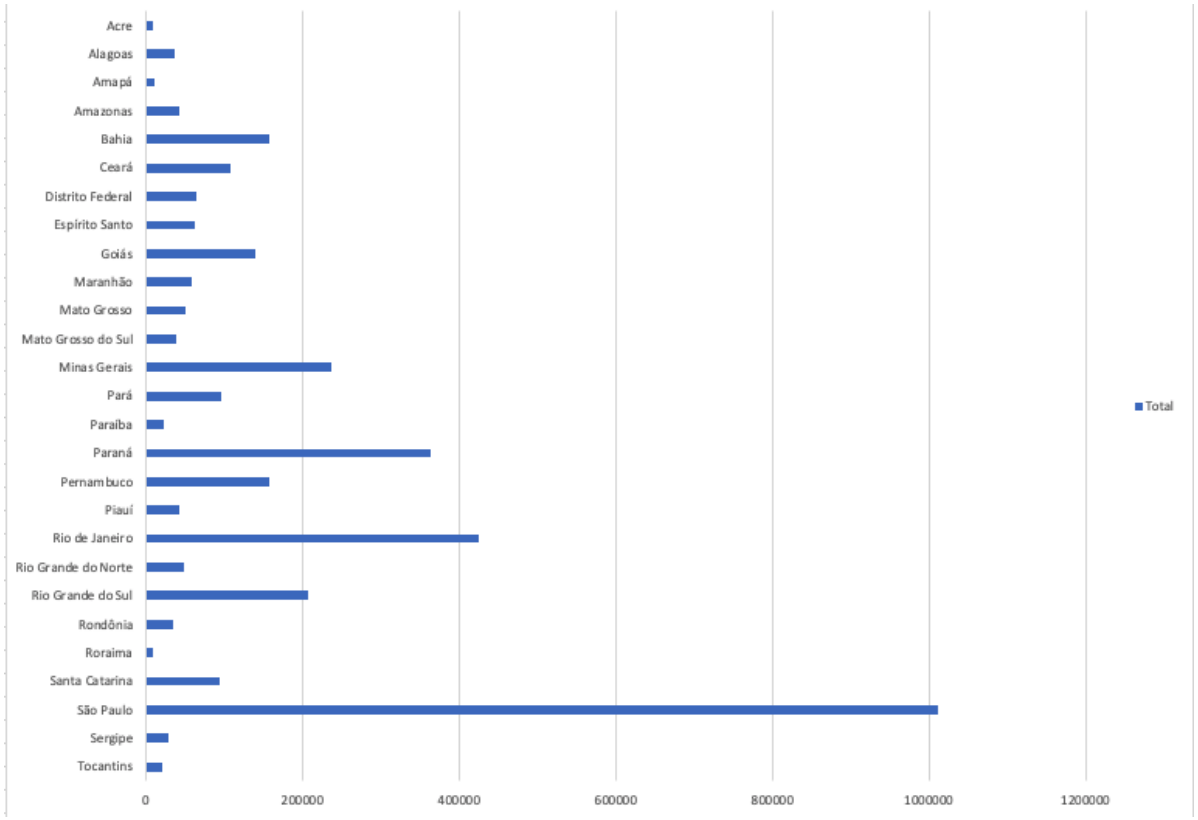
Para validação das regras de associação, foi aplicado o teste de significância da hipótese, por meio do teste do Qui-quadrado (X^2), ou seja, só foram aceitas regras cujos itens (crimes ou meses) estavam além do limite da independência, em outras palavras, com dependência estatisticamente significante. Além disso, também foi calculado o coeficiente de correlação de Pearson (r).

7. Análise dos Resultados

Nesta seção, serão apresentados e discutidos os resultados obtidos, bem como serão descritas as ameaças à validade do experimento.

Para realização do experimento, foram utilizados dados contendo ocorrências criminais de todos os estados do Brasil, incluindo o Distrito Federal. Com um total de 3.576.663 ocorrências criminais, que aconteceram de janeiro de 2015 a março de 2020 e estão distribuídas conforme Figura 1. Percebe-se que o Estado de São Paulo supera, largamente, as demais UFs, com 1.010.726 ocorrências, representando 28,26% do total de crimes. O Rio de Janeiro vem em segundo lugar, com 425.553 (11,90%) ocorrências, seguido do Paraná, com 363.780 (10.17%). Já os Estados do Acre, Roraima e Amapá contêm os menores índices de ocorrências criminais, com 8.226 (0,23%), 9.222 (0.26%) e 10.487 (0.29%), respectivamente.

Figura 1. Número de ocorrências criminais por Estado.



Fonte: Autores.

A Tabela 4 apresenta a distribuição de ocorrências criminais, por ano, em cada um dos estados. Assim como exposto na Figura 1, observa-se que os Estados de São Paulo, Rio de Janeiro e Paraná vêm se mantendo no topo em número de casos ao longo dos anos, e o mesmo vale para os estados do Amapá, Roraima e Acre, com os menores números de casos. Excetuando Piauí e Acre, é notória a queda do número de crimes em 2019, em relação ao ano anterior, seguindo uma tendência geral de queda em 2018, em menores proporções que em 2019. Para 2019, é interessante verificar se houve ações diferentes de combate à criminalidade, bem como averiguar porque os Estados do Piauí e do Acre não seguiram a tendência de queda e estão com os índices em tendência de alta.

Tabela 4. Número de ocorrências criminais anual, para cada Estado brasileiro.

Estado/Ano	2015	2016	2017	2018	2019	2020	Total
São Paulo	215.832	217.018	202.053	183.532	156.030	36.261	**1.010.726**
Rio de Janeiro	67.759	81.590	94.046	90.592	75.000	16.566	**425.553**
Paraná	65.628	79.742	78.026	69.046	61.448	9.890	**363.780**
Minas Gerais	50.871	54.498	49.319	39.617	33.539	8.396	**236.240**
Pernambuco	23.311	30.189	37.195	30.544	28.845	6.873	**156.957**
Rio Grande do Sul	46.957	45.582	43.159	36.803	28.336	7.101	**207.938**
Bahia	31.573	33.213	31.480	29.478	24.359	7.049	**157.152**
Goiás	27.984	33.658	29.627	26.649	17.587	3.744	**139.249**
Pará	16.285	19.108	22.304	19.804	15.324	3.129	**95.954**
Ceará	21.486	21.529	24.268	20.955	14.971	5.888	**109.097**
Santa Catarina	20.982	21.512	18.989	14.957	14.231	3.796	**94.467**
Distrito Federal	13.312	14.791	12.808	10.862	10.449	2.573	**64.795**
Espírito Santo	10.557	10.843	15.368	12.712	10.163	3.176	**62.819**
Maranhão	10.844	12.659	11.959	11.095	9.745	2.219	**58.521**
Piauí	6.364	7.831	7.761	8.975	9.106	2.467	**42.504**
Mato Grosso	11.405	10.830	9.503	8.243	7.992	1.869	**49.842**
Rio Grande do	7.721	11.040	10.767	10.017	6.855	2.181	**48.581**
Mato Grosso do	7.839	7.905	7.740	7.489	6.775	1.611	**39.359**
Amazonas	7.240	8.787	10.292	7.783	6.632	1.589	**42.323**
Rondônia	7.234	7.716	6.415	6.420	6.191	-	**33.976**
Alagoas	6.939	7.802	7.428	6.997	5.947	1.679	**36.792**
Sergipe	4.606	6.176	5.990	5.540	5.102	1.565	**28.979**
Paraíba	5.213	2.666	2.335	7.152	4.634	881	**22.881**
Tocantins	3.291	3.706	3.639	4.720	3.836	1.051	**20.243**
Acre	260	487	1.796	2.550	2.598	535	**8.226**
Amapá	2.125	2.042	1.996	2.025	1.949	350	**10.487**
Roraima	1.457	2.064	1.782	1.895	1.550	474	**9.222**
Total	**695.075**	**1.410.474**	**748.045**	**676.452**	**569.194**	**132.913**	**3.576.663**

Fonte: Autores.

Para realização do experimento, a partir dos incidentes criminais, foram geradas 1.765.534 transações que associam os meses e tipos de crimes ocorridos nas UFs brasileiras, conforme explicado na seção 3. Para cada UF, foram geradas regras de associação utilizando o algoritmo *Apriori*, o qual recebeu como entrada as informações do mês, crimes, suporte mínimo calculado, conforme Tabela 3 e a confiança mínima adotada, que foi de 70% (0,70). A Tabela 5 detalha as informações encontradas pelas trinta regras de associação com melhores medidas de interesse e qualidade, do tipo "Crime $\Rightarrow$ Crime", juntamente com os valores destas medidas: suporte, confiança, lift, count, coeficiente de correlação de Pearson (r), qui-quadrado (X^2) e seu *p-value*. Confirmando o que foi explicado anteriormente, a fim de encontrar evidências que ratifiquem, ou não, as regras de associação encontradas, foi adotado um nível de confiança de 95% ($\alpha = 0.05$) para o experimento.

Ao analisar a Tabela 5, percebe-se que todas as associações atenderam aos requisitos mínimos exigidos pelo experimento. Os valores de suporte superaram os valores de suporte mínimo para cada Estado, demonstrados na Tabela 3, e o mesmo acontece com os valores de confiança, que superam a confiança mínima estipulada, de 70% (0,7). Em relação ao lift, os

valores atingidos para as regras selecionadas foram maiores que 1, o que indica que existe uma relação de dependência positiva entre os tipos de crimes. Além disso, os valores do coeficiente de correlação de Pearson (*r*) foram positivos, o que demonstra que há uma correlação positiva, ou seja, quando a frequência do "Antecedente" aumentar, a frequência do "Consequente" aumentará, na proporção indicada por *r*.

Vale a pena destacar a associação "Latrocínio ⟹ Roubo de Carga" para o estado da Bahia. Com um lift muito superior aos demais, a regra nos diz que, juntamente com 99% dos crimes mensais de latrocínio, ocorre também roubo de carga. O poder público deve investigar as causas e criar ações necessárias para coibir este tipo de crime. Como vimos no parágrafo anterior, existe uma correlação positiva entre os crimes, ou seja, caso o crime de latrocínio aumente, roubo de carga também aumentará. O mesmo raciocínio vale para a associação "Homicídio doloso ⟹ Tentativa de homicídio", para o estado de São Paulo. Com o segundo maior lift, a regra nos diz que juntamente com 99% dos crimes mensais de homicídio doloso, ocorrem também novas tentativas de homicídio. Medidas preventivas devem ser tomadas pelas autoridades, principalmente no sentido de tentar antecipar a prevenção dos crimes consequentes quando um latrocínio ou tentativa de homicídio ocorre, tomando como exemplo os casos aqui destacados.

Tabela 5. Regras com os maiores lifts para associações entre crimes, por Estado.

UF	Regra	Suporte	Confiança	Lift	Count	r	X^2	p-value	Leverage
BA	Latrocinio → Roubo de carga	0.012572	0.997740	38.651010	883	0.692370	33668.095357	0.000000	0.012247
SP	Tentativa de homicidio → Homicidio doloso	0.032148	0.846291	26.304286	17.046	0.916796	445662.226393	0.000000	0.030926
SP	Homicidio doloso → Tentativa de homicidio	0.032148	0.999237	26.304286	17.046	0.916796	445662.226393	0.000000	0.030926
PR	Roubo de carga → Tentativa de homicidio	0.011406	0.764875	20.781839	2.121	0.475771	42092.134130	0.000000	0.010857
GO	Roubo de carga → Estupro	0.032699	0.956848	17.883139	2.040	0.755082	35569.359089	0.000000	0.030871
MG	Roubo de carga → Estupro	0.022199	1.000000	17.172569	2.863	0.605950	47353.294177	0.000000	0.020906
SC	Roubo de carga → Homicidio doloso	0.013589	1.000000	14.358039	824	0.428990	11158.667729	0.000000	0.012643
SP	Tentativa de homicidio → Roubo de carga	0.037987	1.000000	11.335891	20.142	0.638858	216406.297297	0.000000	0.034636
SP	Homicidio doloso → Roubo de carga	0.032173	1.000000	11.335891	17.059	0.586167	182181.321169	0.000000	0.029334
PR	Tentativa de homicidio → Homicidio doloso	0.036267	0.985388	11.258047	6.744	0.621045	71721.585132	0.000000	0.033045
PR	Roubo de carga → Homicidio doloso	0.014482	0.971150	11.095375	2.693	0.384698	27519.759552	0.000000	0.013176
RS	Roubo de carga → Estupro	0.018905	1.000000	10.717682	1.696	0.432738	16798.786769	0.000000	0.017141
RJ	Tentativa de homicidio → Homicidio doloso	0.085953	0.988328	9.922260	19.560	0.915890	190893.188781	0.000000	0.077291
RJ	Homicidio doloso → Tentativa de homicidio	0.085953	0.862928	9.922260	19.560	0.915890	190893.188781	0.000000	0.077291
SC	Roubo de carga → Estupro	0.013589	1.000000	9.734146	824	0.346886	7296.088822	0.000000	0.012193
SC	Homicidio doloso → Estupro	0.069053	0.991475	9.651165	4.187	0.800927	38895.760754	0.000000	0.061898
RJ	Tentativa de homicidio → Estupro	0.086854	0.998686	9.512181	19.765	0.899789	184240.509325	0.000000	0.077723
RJ	Estupro → Tentativa de homicidio	0.086854	0.827264	9.512181	19.765	0.899789	184240.509325	0.000000	0.077723
SP	Homicidio doloso → Estupro	0.032173	1.000000	9.322473	17.059	0.525985	146692.636457	0.000000	0.028721
SP	Tentativa de homicidio → Estupro	0.037851	0.996425	9.289149	20.070	0.570971	172857.979587	0.000000	0.033777
SP	Estupro → Roubo de carga	0.085854	0.800372	9.072938	45.522	0.870425	401719.713435	0.000000	0.076391
SP	Roubo de carga → Estupro	0.085854	0.973232	9.072938	45.522	0.870425	401719.713435	0.000000	0.076391
RJ	Homicidio doloso → Estupro	0.094681	0.950544	9.053649	21.546	0.917450	191544.021454	0.000000	0.084223
RJ	Estupro → Homicidio doloso	0.094681	0.901808	9.053649	21.546	0.917450	191544.021454	0.000000	0.084223
DF	Homicidio doloso → Estupro	0.079253	0.949007	8.474126	2.438	0.801216	19747.607329	0.000000	0.069901
DF	Estupro → Homicidio doloso	0.079253	0.707692	8.474126	2.438	0.801216	19747.607329	0.000000	0.069901
SC	Homicidio doloso → Tentativa de homicidio	0.069647	1.000000	7.915665	4.223	0.719524	31391.169782	0.000000	0.060848
SC	Roubo de carga → Tentativa de homicidio	0.013589	1.000000	7.915665	824	0.308669	5777.016743	0.000000	0.011872
RS	Roubo de carga → Homicidio doloso	0.018905	1.000000	7.673167	1.696	0.358600	11535.787427	0.000000	0.016442
RS	Homicidio doloso → Estupro	0.092991	0.713540	7.647498	8.342	0.825488	61129.116266	0.000000	0.080831

Fonte: Autores.

Ao analisarmos as regras, percebemos uma confirmação das evidências publicadas em (Barros, Baggio, Stege, & Hilgemberg, 2019), as quais apontaram uma tendência de regiões com baixo desenvolvimento econômico à alta taxa de homicídios (e vice-versa), como é o caso do estado da Bahia (AtlasBR, 2020), que liderou o ranking das associações. No entanto, como exceções, o estudo também aponta regiões em que o nível de desenvolvimento econômico não é capaz de conter o avanço da criminalidade. Isso fica evidente nos estados do Rio de Janeiro e Paraná, que, apesar de possuírem um alto desenvolvimento econômico, apresentaram a totalidade das regras importantes relacionadas a homicídios.

Na Tabela 6, são apresentadas as regras de associação do tipo "Mês ⟹ Crime", revelando que tipo de crime tende a ocorrer mais em um determinado mês. Observa-se que, da mesma forma que na tabela anterior, os valores também atendem os requisitos mínimos exigidos pelo experimento. Os valores de lift atingidos também foram maiores que 1, indicando que existe uma relação de dependência entre os meses do ano e os tipos de crimes. Vale a pena destacar a associação "Dezembro ⟹ Roubo de veículo", para o estado de São Paulo, que mostra que 84% das ocorrências do mês de dezembro são roubos de veículo. Assim como também podemos destacar a associação "Novembro ⟹ Homicídio doloso", para o estado da Paraíba, que nos diz que 76% das ocorrências do mês de novembro são homicídios dolosos.

Vale ressaltar que o valor do Lift acima de 1 garante que estas regras não foram geradas para crimes que também ocorrem de forma comum em outros meses, ou seja, não são regras enganosas, realmente há um destaque para estes crimes, nesses meses. Uma regra enganosa poderia ser gerada para um crime que sempre ocorre com frequência em todos os meses, como é o caso dos furtos. O poder público deve ficar atento e maximizar as ações de combate ao roubo de veículo, principalmente no mês de dezembro, para o estado de São Paulo, bem como ações de combate ao homicídio doloso, principalmente no mês de novembro, para o estado da Paraíba.

Tabela 6. Regras encontradas para associação entre mês e crime, por Estado.

UF	Regra	Suporte	Confiança	Lift	Count	r	X^2	p-value	Leverage
SP	Dezembro → Roubo de veiculo	0.058520	0.845868	1.320965	31.029	0.116758	7228.328628	0.000000	0.014219
PB	Novembro → Homicidio doloso	0.074990	0.769870	1.315447	833	0.123099	168.325900	0.000000	0.017983
MA	Fevereiro → Furto de veiculo	0.069970	0.867400	1.260276	1.498	0.114558	280.964972	0.000000	0.014450
RO	Março → Roubo de veiculo	0.068303	0.777510	1.236547	968	0.095539	129.358982	0.000000	0.013066
PB	Março → Roubo de veiculo	0.097767	0.838610	1.153737	1.086	0.091106	92.201768	0.000000	0.013027

Fonte: Autores.

Finalmente, com o que foi exposto acima, percebemos que as questões **Q1** e **Q2** podem ser respondidas. Foram encontrados valores de *p-value* abaixo do nível de significância adotado (*p-value* < 0.05) e, desta forma, a hipótese H_0 pode ser rejeitada, indicando que há dependência entre os tipos de crimes e entre os meses do ano e os crimes.

7.1 Ameaças à validade

Para validação de um experimento, é preciso considerar questões que influenciam o resultado final. 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 externa: Apesar dos dados serem disponibilizados pelo Ministério da Justiça e Segurança Pública, não se pode garantir que não exista algum tipo de subnotificação ou informação incorreta nos arquivos, o que pode influenciar diretamente o resultado da pesquisa. Além disso, 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).

8. Conclusão e Trabalhos Futuros

Neste artigo, foi apresentado um experimento cujo objetivo principal consistiu em detectar padrões entre os crimes no Brasil, verificando se existe alguma relação entre os tipos de crimes e entre os crimes praticados e os meses do ano, visando auxiliar o Governo e as agências de polícia na adoção de medidas preventivas, para minimizar as ameaças à sociedade. Foram utilizados dados abertos, relacionados a incidentes criminais, dos 26 Estados brasileiros mais o Distrito Federal, disponibilizados pelo Governo Federal através do portal do MJSP.

Neste contexto, as seguintes medidas de interesse e qualidade foram utilizadas para avaliar as forças das regras de associação geradas no experimento: Suporte, Confiança, Lift,

Coeficiente de correlação de Pearson (r), Qui-Quadrado (X^2), com seu nível de significância (*p-value*) e a Alavancagem (leverage). Considerando a validação de qualidade das regras geradas, os resultados expostos nesta pesquisa podem auxiliar os analistas criminais no processo de tomada de decisão para prevenção e redução da criminalidade no país, uma vez que inferem a prioridade na inibição de crimes que são antecedentes de outras ocorrências dentro de um mesmo estado, ou de crimes que ocorrem num mesmo período.

Como resultado à questão **Q1**, o destaque ficou por conta da associação "Latrocínio ⟹ Roubo de carga", para o estado da Bahia. O poder público deve investigar as causas de um crime tal como roubo seguido de morte (latrocínio) estar associado ao crime de roubo de carga, criando ações necessárias para coibir o antecedente, bem como prevenir o consequente, quando o latrocínio não tiver sido evitado e já estiver ocorrido no estado. Já como reposta à questão **Q2**, o destaque ficou por conta da associação "Dezembro ⟹ Roubo de veículo", para o estado de São Paulo, o que indica que as autoridades precisam analisar o motivo deste tipo de crime ter um número de ocorrências significativo no mês de dezembro.

Como trabalhos futuros, ainda no âmbito de crimes, outros tipos de associações podem ser encontrados, com outros desdobramentos dos dados. Dados criminais levando em consideração regiões menores, tais como, por exemplo, os municípios de um Estado, ou até mesmo bairros de um Município poderão ser investigados, desde que os dados sejam disponibilizados por alguma entidade federativa, estadual ou municipal. Neste mesmo sentido, assim como visto na seção 4, podem ser encontradas associações entre os crimes, levando em consideração outras informações tais como clima, sexo, idade, horário do crime, entre outros.

Vale ressaltar a importância 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. Todos os dados aqui apresentados, assim como novas atualizações, poderão ser consultados, nessas e em outras perspectivas, no site do projeto Transparência Traduzida, em (www.transparenciatraduzida.ufs.br).

Referências

AtlasBR. (2020). *AtlasBR*. Recuperado de http://www.atlasbrasil.org.br

Barros, P., Baggio, I., Stege, A., & Hilgemberg, C. (2019). Economic development and crime in Brazil: a multivariate and spatial analysis.

Basili, V. R., & Weiss, D. M. (1984). A methodology for collecting valid software engineering data. *IEEE Transactions on software engineering*, (6), 728-738.

Basili, V., Trendowicz, A., Kowalczyk, M., Heidrich, J., Seaman, C., Münch, J., & Rombach, D. (2014). *Aligning Organizations Through Measurement: The GQM+ Strategies Approach*. Springer.

Belesiotis, A., Papadakis, G., & Skoutas, D. (2018). Analyzing and Predicting Spatial Crime Distribution Using Crowdsourced and Open Data. *ACM Trans. Spatial Algorithms Syst.*

Campos, O. S. (2018). Data analytics transparente para descoberta de padrões e anomalias na realização de convênios e contratos de repasse federais.

Catlett, C., Cesario, E., Talia, D., & Vinci, A. (2018). A Data-Driven Approach for Spatio-Temporal Crime Predictions in Smart Cities. *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, 17-24. Taormina.

Chapetta, W. A. (2006). *Uma Infra-estrutura para Planejamento, Execução e Empacotamento de Estudos Experimentais em Engenharia de Software* (Doctoral dissertation, Dissertação de Mestrado, Programa de Engenharia de Sistemas e Computação, COPPE/UFRJ, Universidade Federal do Rio de Janeiro. Rio de Janeiro, RJ, Brasil).

Chaves, M. S. R. S., & Shimizu, I. S. (2018). Síndrome de burnout e qualidade do sono de policiais militares do Piauí. *Revista Brasileira de Medicina do Trabalho*, 16(4), 436-441.

Gillmeister, P., & Cazella, S. (2007). Uma análise comparativa de algoritmos de regras de associação: minerando dados da indústria automotiva. *Escola Regional de Banco de Dados (ERBD)*. Caxias do Sul, RS.

Guimarães, F. F., Joaquim, S. F., Manzi, M. P., Silva, R. C. D., Bruder-Nascimento, A. C. M. D. O., Costa, E. O., & Langoni, H. (2016). Comparison phenotypic and genotypic

identification of Staphylococcus species isolated from bovine mastitis. *Pesquisa Veterinária Brasileira*, 36(12), 1160-1164.

Gupta, M., Chandra, B., & Gupta, M. (2014). A framework of intelligent decision support system for Indian police. *J. Enterp. Inf. Manag., 27*, 512-540.

Huang, S.-M. (2013). A Study of the Application of Data Mining on the Spatial Landscape Allocation of Crime Hot Spots. *Communications in Computer and Information Science, 398*, 274-286.

IBGE. (2020). *Instituto Brasileiro de Geografia e Estatística*. Recuperado de https://www.ibge.gov.br

Lallich, S., Teytaud, O., & Prudhomme, E. (2007). Association Rule Interestingness: Measure and Statistical Validation. *Studies in Computational Intelligence, 43*, 251-275.

Marzan, C. S., Baculo, M. J. C., de Dios Bulos, R., & Ruiz Jr, C. (2017). Time series analysis and crime pattern forecasting of city crime data. *International conference on algorithms, computing and systems*, 113-118.

Melo, M., Teixeira, J., & Campos, G. (2012). A prediction model for criminal levels using socio-criminal data. *International Journal of Electronic Security and Digital Forensics, 4*, 201-214.

MSJP. (2020). *Portal Brasileiro de Dados Abertos*. Recuperado de http://dados.gov.br

Nogueira de Oliveira, R., & Colaço Júnior, M. (2018). Experimental Analysis of Stemming on Jurisprudential Documents Retrieval. *Information, 9*, 28.

Paranhos, R., Figueiredo Filho, D. B., da Rocha, E. C., da Silva Júnior, J. A., Neves, J. A. B., & Santos, M. L. W. D. (2014). Desvendando os mistérios do coeficiente de correlação de Pearson: o retorno. *Leviathan (São Paulo)*, (8), 66-95.

Pereira, A. S., Shitsuka, D. M., Parreira, F. J., & Shitsuka, R. (2018). Metodologia da pesquisa científica.

Pereira, B., & Brandão, W. (2014). ARCA: Mining Crime Patterns Using Association Rules.

PostgreSQL. (2020). *PostgreSQL*. Recuperado de https://www.postgresql.org.

Prado, K. H. de J., Souza, L. S., de Jesus Junior, I. D., & Colaço Júnior, M. (2020). Applied Intelligent Data Analysis to Government Data Related to Criminal Incident: A Systematic Review. *Journal of Applied Security Research*, 1-35.

Procaci da Silva, A. (2004). Regras de associação quantitativas em intervalos não contínuos.
Santos, B. S., Júnior, M. C., & de Souza, J. G. (2018). A initial experimental evaluation of the neuromessenger: a collaborative tool to improve the empathy of text interactions. In *Information Technology-New Generations* (pp. 411-419). Springer, Cham.

Sevri, M., Karacan, H., & Akcayol, M. (2017). Crime Analysis Based on Association Rules Using Apriori Algorithm. *International Journal of Information and Electronics Engineering, 7*, 99-102.

Singh, N., Kaverappa, C. B., & Joshi, J. D. (2018). Data mining for prevention of crimes. In *International Conference on Human Interface and the Management of Information*, 705-717.. Springer, Cham.

Sumithra, R., & Paul, S. (2010). Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery. *2010 Second International conference on Computing, Communication and Networking Technologies*, 1-5.

Travassos, G., Gurov, D., & Amaral, E. (2002). Introdução à engenharia de software experimental.

Varde, A., Takahashi, M., Rundensteiner, E., Ward, M., Maniruzzaman, M., & Jr., R. (2004). Apriori Algorithm and Game-of-Life for Predictive Analysis in Materials Science. *International Journal of Knowledge-Based and Intelligent Engineering Systems, 8*, 213-228.

Wu, J., He, Z., Gu, F., Liu, X., Zhou, J., & Yang, C. (2016). Computing exact permutation p-values for association rules. *Information Sciences*, 346, 146-162.

Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017). Crime pattern detection, analysis & prediction. In *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, 1, 225-230. IEEE.

Porcentagem de contribuição de cada autor no manuscrito

Wesckley Faria Gomes – 33,34%

Methanias Colaço Júnior – 33,33%

Kleber Henrique de Jesus Prado – 33,33%