X Encontro Brasileiro de Administração Pública.
ISSN: 2594-5688
secretaria@sbap.org.br
Sociedade Brasileira de Administração Pública

Ciência de dados e políticas públicas: uma aplicação de algoritmos supervisionados de machine learning para predição de registros de ocorrências de roubos e furtos de celulares na cidade de São Paulo.

Fernando Freire Vasconcelos, Carlos Henrique Ferreira De Souza

[RELATO TÉCNICO] GT 17 – Segurança Pública e Cidadania

Ciência de dados e políticas públicas: uma aplicação de algoritmos supervisionados de machine learning para predição de registros de ocorrências de roubos de celulares na cidade de São Paulo.

Resumo:

Esta pesquisa objetivou implementar dois tipos de algoritmos supervisionados para a predição de registros de ocorrências de roubos de celulares na cidade de São Paulo. Os algoritmos utilizados foram Árvore de Regressão e Regressão Linear Múltipla. Assim, realizou-se 8 experimentos no total, considerando uma base de dados dos Registros de Ocorrências Digitais (RDO) da Polícia Civil e uma base de dados socioeconômicos (IDHM). Os experimentos demonstraram que a dinâmica urbana espacial e social pode estar correlacionada com o número de ocorrências do crime aqui analisados, sendo que a desigualdades econômicas e educacionais são mais relevantes para casos de roubos.
Palavras-chave: roubo, regressão linear múltipla, árvore de decisão, predição de ocorrências.

Orientações gerais de conteúdo do Relato Técnico:

O presente relato técnico busca apresentar uma metodologia capaz de predizer a ocorrência de crime de roubo da cidade de São Paulo, a partir de uma predição utilizando os seguintes algoritmos: Árvore de Decisão e de Regressão Linear Múltipa. A base de dados foi requerido através de solicitação pela Lei de Acesso à Informação à Secretaria de Segurança Publica do Estado de São Paulo. A base de dados ficou com 651.351 observações, contendo dados de 2003 a 2021, distribuídos em 45 variáveis.

Introdução

De acordo com Vargas (2019), em 2018, só a cidade de São Paulo apresentou cerca de 140 mil ocorrências de roubos de celulares. Nesse sentido, pensar o problema social da segurança pública no país é pensar em políticas públicas baseadas em evidências.

Conforme Risso (2016) em uma meta-análise recente, que se baseou principalmente na literatura dos Estados Unidos, identificou o que funciona em termos de redução da violência em comunidades, enfatizando a importância de utilizar dados e evidências na elaboração de políticas públicas.

Nessa linha, entende-se nesta pesquisa que a ciência de dados é importante aliada no processo de construção das políticas públicas, seja no momento de diagnóstico do problema ou até mesmo de avaliação de uma política pública. Apesar de ser um campo pouco utilizado para esta finalidade, ainda mais considerando problemas de segurança pública, como aparece em Campos e Figueiredo (2021).

Verifica-se que existe uma lacuna na literatura científica brasileira que utiliza algoritmos

de machine learning para analisar problemas de políticas públicas, especialmente os de segurança pública. Portanto, buscando ser um complemento dos trabalhos ainda incipientes que versam sobre a temática, essa pesquisa objetiva implementar e comparar o desempenho de dois algoritmos supervisionados de machine learning que consigam predizer a quantidade de ocorrências de roubos de aparelhos celulares nos distritos da cidade de São Paulo e que também sejam capazes de apontar o impacto das variáveis explicativas sobre a variável a ser explicada. Os algoritmos selecionados foram de *Regression Tree* e Regressão Linear Múltipla.

Metodologia

O primeiro procedimento necessário para execução da pesquisa foi a revisão da literatura aplicada que utiliza algoritmos de machine learning para predição de crimes. Essa leitura analítica subsidiou a escolha dos algoritmos, atributos e métricas de avaliação para prever aspectos relativos a ocorrências criminais, sobremaneira roubos. O segundo passo foi realizar a coleta e tratamento dos dados. Nesta etapa, o software estatístico R, em sua versão 4.0.3 foi utilizado, bem como na etapa de descrição e visualização dos dados. Os pacotes utilizados foram tidyverse (Wickham et al., 2019), fastDummies (Kaplan, 2020), naniar (Tierney et al., 2021), rpart (Therneau; Atkinson, 2019), PerformanceAnalytics (Peterson; Carl, 2020) e olsrr (Hebbali, 2020).

Complementarmente, outros materiais utilizados nesta pesquisa foram a base de dados oficiais fornecidas pelo Governo do Estado de São Paulo, através de sua Secretaria de Segurança Pública (SSP), mediante solicitação pela Lei de Acesso à Informação (LAI)[1] e também fornecidas pela Prefeitura Municipal de São Paulo. A principal base de dados utilizada é a mesma já analisada em Vargas (2019): o Registro Digital de Ocorrências (RDO). Isso porque o RDO contém as variáveis targets que desejamos aplicar os algoritmos supervisionados de machine learning, quais sejam, a quantidade de ocorrências de roubos de celulares; além de abarcar variáveis de localização do crime, data, hora, tipo de crime e dados dos indivíduos relacionados com a ocorrência.

Incorporando as críticas realizadas em Duarte e Lobato (2021) quanto a utilização destes algoritmos para estimular o mesmo paradigma de policiamento repressivo e que pode estigmatizar determinadas localidades a partir da criação de manchas criminais, buscou-se variáveis explicativas que também enfatizam características socioeconômicas das localidades

[1] Cujo número do protocolo é o 42937227111.

das ocorrências para que seja possível criar insights que estimulem políticas públicas com foco no longo prazo. Nesse sentido, empregou-se a base de dados abertos que contém variáveis relativas ao Índice de Desenvolvimento Humano Municipal. Estes dados foram agregados por distritos para que fosse possível incorporar a crítica mencionada.

Portanto, os algoritmos foram rodados duas vezes cada. Uma vez considerando uma agregação da base de dados por Delegacias de Circunscrição das ocorrências; e outra vez considerando a agregação por distritos da capital. Como testamos os modelos em duas bases diferentes, então, totalizamos 8 modelos.

Vale dizer que esta última forma de análise ficou prejudicada pela elevada quantidade de valores não disponíveis (NAs), já que na base de dados enviada pela Secretaria de Segurança Pública do Estado de São Paulo não havia nome dos distritos, apenas bairros. E estes dois atributos não são compatíveis. Os distritos foram obtidos mediante *join* com uma outra base de dados da plataforma Geosampa, considerando os atributos de latitude e longitude que tinham no RDO. No entanto, a base de dados foi reduzida quase pela metade para este procedimento funcionar.

No primeiro, realizamos a exclusão das observações duplicadas que se referiam ao mesmo boletim, como recomendado no documento de Metodologia disponibilizado no próprio portal da Transparência da Secretaria de Segurança Pública do Estado de São Paulo[2]. Em seguida, selecionamos apenas os dados para o município de São Paulo. Com estes procedimentos, a base de dados ficou com 651.351 observações, contendo dados de 2003 a 2021, e distribuídos em 45 variáveis. No estágio de integração foi realizado *join* com a base de dados com nomes dos distritos da capital, coletada da plataforma Geosampa. Essa tarefa foi feita utilizando o software QGIS (2022), por meio das variáveis de latitude e longitude.

Já no estágio de transformação, criamos uma variável com nome de dia da semana da ocorrência através da variável de data da ocorrência, utilizando o pacote lubridate, dentro do tidyverse (Wickham et al., 2019). Ainda, precisamos adequar o nome do município de São Paulo representado por categorias diferentes, tais como S.Paulo, Sao Paulo e SÃO PAULO. Também foi necessário adequar nome das categorias da variável Rubrica que diz respeito ao artigo criminal que caracteriza a ocorrência correspondente. Na ocasião, ou 155 (furto), ou 157 (roubo). Importante destacar que neste estágio ainda realizamos recategorizações na variável de local da ocorrência, buscando tornar a visualização mais agradável no momento de plotagem dos algoritmos e de escrita do script.

[2] Disponível em: http://www.ssp.sp.gov.br/transparenciassp/Consulta.aspx. Acesso em: 11/08/2022.

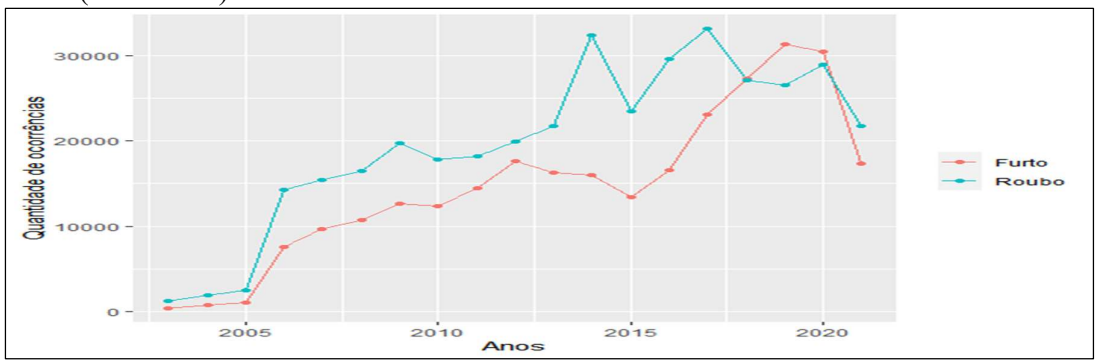
Para transformar os dados em contagem foi necessário transformar em *dummies* as seguintes variáveis: Flagrante, Rubrica, Local da ocorrência, e Dia da ocorrência. Para variável Flagrante havia categorias "Sim" e "Não". Finalmente, agrupamos os dados por Delegacias de Circunscrição e por distritos em uma nova base de dados. Após, criamos taxas de todas as variáveis explicativas dividindo-as pela variável dependente (ou quantidade de ocorrências de roubos).

Resultados

Modelos para roubo de celulares

O gráfico 1 demonstra que o número de ocorrências dos dois tipos de crimes subiu muito na cidade. Saímos de 1656 ocorrências em 2003, para 39.054 em 2021. Uma variação de mais de 2.300%. A partir de 2005 que o número de ocorrências também se acentua consideravelmente. Destaca-se também que de 2013 a 2017 o número de roubos se descolou muito do número de furtos. A partir de 2018 o número de furtos supera o de roubos e, em 2020, há uma queda considerável em ambos os tipos de ocorrências. Aqui pode-se pensar em um possível impacto da pandemia de COVID-19.

Gráfico 1. Quantidade de ocorrências de roubos e furtos de celulares por ano na cidade de São Paulo (2003-2021)



Fonte: elaboração própria a partir de dados do RDO

A tabela 3 apresenta os resultados dos modelos de regressão linear múltipla e de árvore de regressão para predição das ocorrências de roubos de celulares. Todas as bases de dados foram separadas em 70% de observações para treino e 30% para teste. Os modelos de regressão linear múltipla tiveram as variáveis dependentes transformadas pelo lambda de Box-Cox, conforme recomendado em Fávero e Belfiore (2017).

Tabela 3. Métricas de avaliação dos modelos para ocorrências de roubos de celular

Modelo	Base de dados	MAPE	MAE	RMSE	R^2	R^2 ajustado
Regressão linear múltipla com transformação Box-Cox	RDO	**9,40%**	**3,76**	**4,69**	54,01%	50,90%
Regressão linear múltipla com transformação Box-Cox	IDH-M (IBGE)	**25%**	**13,3**	**16,4**	66,70%	61,50%
Árvore de regressão	RDO	36%	1226	1579	-	-
Árvore de regressão	IDH-M (IBGE)	49%	1095.7	1751.8	-	-

Fonte: elaboração própria.

Nota-se que os modelos com Box-Cox, comparativamente às árvores de regressão, em geral, performaram melhor. O MAPE (Mean Absolute Percentage Error) do modelo com Box-Cox para a base de dados RDO foi de 9,4%, enquanto que a árvore para a mesma base de dados obteve um resultado de 36%. As diferenças entre MAE (Mean Absolute Error) e RMSE (Root Mean Square Error) são bastante elevadas também. O modelo de regressão linear múltipla com dados socioeconômicos embora consiga explicar a melhor a variabilidade dos dados (R^2 ajustado de 61,5%) do que o modelo anterior (R^2 ajustado de 50,9%), acaba caindo em poder de predição. Nas árvores, movimento similar acontece.

Não obstante, impende advertir que os modelos de árvores foram implementados com todas as 223 variáveis quantitativas do IDHM, enquanto os modelos com Box-Cox foram implementados apenas com as variáveis apontadas como as mais importantes nas árvores. Este método de seleção de variáveis é similar ao proposto em Sugumaran, Muralidharan e Ramachandran (2006) que utilizaram as variáveis mais importantes de uma árvore de decisão para implementar um modelo de Proximal Support Vector Machine para diagnóstico de falhas em rolamentos utilizados em engenharia mecânica. Ademais, o procedimento stepwise também foi utilizado para selecionar apenas variáveis cujo *p-value* fosse menor do que 0.05.

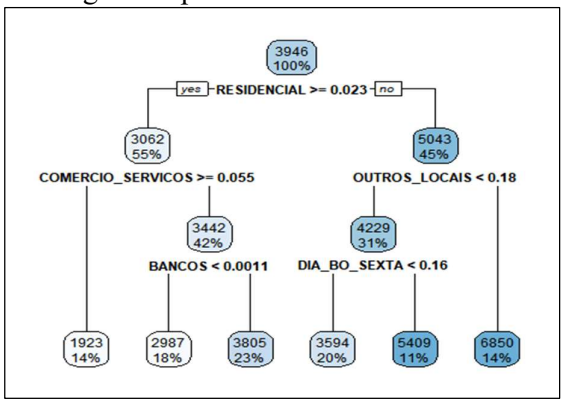
Modelos de Árvore de decisão para roubo de celulares

A figura 1 apresenta a árvore com variáveis do RDO. Este modelo estabelece que, em média, os distritos policiais com a menor quantidade de ocorrências de roubos de celulares são aqueles em que ocorrências de roubos em locais residenciais representam mais do que 2,3%; e em que ocorrências de roubos em estabelecimentos comerciais e de serviços representem mais do que 5,5%. Para estes casos, o modelo prevê, em média, 1923 roubos. Ou seja, isto pode

significar que os locais em que comércios, serviços e residências são os maiores alvos, também são os mesmos locais que possuem menor quantidade de ocorrências de roubos. Embora no modelo não haja uma variável com a quantidade de comércios e serviços por local, um modelo melhorado poderia explorar este insight. É possível que essas ocorrências sejam menores porque nestes locais há maior trânsito de pessoas? O que há para ser explorado neste grupo de localidades? Uma investigação a nível dos logradouros poderia responder esta questão.

Por outro lado, o modelo aponta que, em média, os distritos policiais com maior quantidade de ocorrências de roubos de celulares são aqueles em que 2,3% ou menos dos roubos acontecem em locais residenciais; e em que roubos que ocorrem em "Outros locais" representam mais do que 18%. Para estes casos, o modelo prediz uma média de 6850 ocorrências. A interpretação aqui pode ser inversa à realizada no parágrafo acima.

Figura 1. Árvore de regressão para roubos de celulares com variáveis do RDO



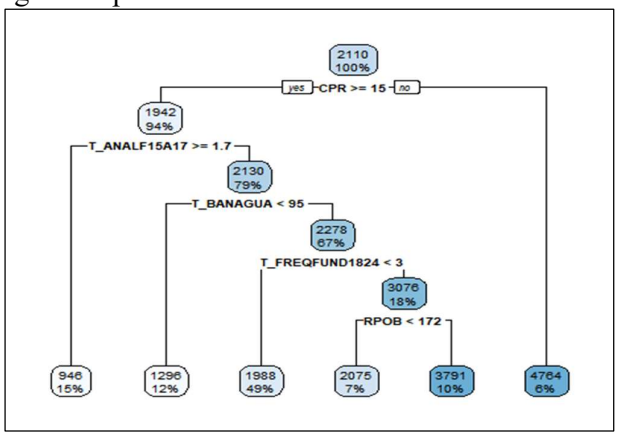
Fonte: elaboração própria.

As variáveis mais importantes para este modelo, considerando aquelas que maximizam a diminuição da soma dos erros quadrados para cada subdivisão da árvore são aquelas que indicam: se a ocorrência foi em locais residenciais, "outros locais", comércios e serviços, no dia de sábado, em uma indústria ou em um terminal/estação de transporte público.

A figura 2 apresenta a árvore de regressão para ocorrências de roubos em distritos (não mais distritos policiais) com dados socioeconômicos do IDHM (IBGE). Segundo este modelo, os distritos que, em média, possuem o menor número de ocorrências de roubos (em média 946) são aqueles cujo CPR (média do percentual de trabalhadores por conta própria) é maior ou igual a 15%; e que a média da taxa de analfabetismo entre pessoas de 15 a 17 anos é maior do que 1,7% da população. Apenas 8 distritos não correspondem à primeira condição, sendo todos eles distritos periféricos da cidade (Jardim Ângela, Capão Redondo, Grajaú, Sapopemba, Anhanguera, Cidade Dutra, Pedreira e Perus). Nestes casos, o modelo prevê, em média, a maior

quantidade de roubos, ou seja, 4764. Para melhor compreensão destas relações seria necessário um estudo mais aprofundado sobre características da espacialização do mercado de trabalho informal e de aspectos educacionais. Principalmente porque todos os 10 distritos com maior média do CPR ficam em regiões centrais ou próximas (Jardim Paulista, Alto de Pinheiros, Consolação, Moema, Pinheiros, Brás, Mooca, Itaim Bibi, Perdizes e Lapa).

Figura 2. Árvore de regressão para roubos de celulares com variáveis socioeconômicas



Fonte: elaboração própria.

Por fim, entre as variáveis mais importantes estão: CPR, Índice de Theil dos rendimentos do trabalho, taxa de analfabetismo da população entre 15 a 17 anos, percentual da população das pessoas de 18 a 24 anos que frequentam o ensino fundamental, percentual de pessoas com água encanada e renda domiciliar per capta dos mais vulneráveis.

Regressão com BOX-COX

A tabela 5 apresenta os resultados dos modelos de regressão com Box-Cox para ocorrências de roubos, de acordo com suas variáveis e parâmetros. O modelo com variáveis RDO possui R^2 de 54% e MAPE de 9,4% e RMSE de 4,69. Destaca-se que a variável COMERCIO_SERVICOS também apareceu como importante para este modelo, possuindo uma correlação inversa com a quantidade de ocorrências roubos de celulares. Ou seja, em média, quanto maior a ocorrência de roubos em comércios e serviços, menor a ocorrência de roubos totais em um distrito policial. Essa correlação também apareceu na árvore de regressão. A correlação com a variável INDUSTRIA também apresenta o mesmo padrão. Já as variáveis DIA_BO_SABADO e BANCOS apresentam correlação positiva. Como demonstra a tabela, todas as variáveis são estatisticamente significativas a pelo menos 5% de nível de significância.

Tabela 5. Parâmetros das regressões com Box-Cox para ocorrências de roubos

Modelo	Variáveis	Coeficientes	p-value
Roubos com RDO	Intercepto	42.389	0.00
	COMERCIO_SERVICOS	-246.311	0.00
	DIA_BO_SABADO	67.691	0.02
	INDUSTRIA	-2.723.616	0.00
	BANCOS	720.941	0.03
	Estatística F		0.00
Roubos com IDH-M (IBGE)	Intercepto	-11432945	0.00
	PREN60	217445	0.00
	GINI	14531229	0.00
	T_SUPER25M	-632507	0.05
	P_SUPER	-42089	0.02
	AGUA_ESGOTO	-90541	0.00
	RAZDEP	-18009	0.00
	T_ANALF18M	928234	0.00
	T_ANALF15M	-981949	0.00
	T_FREQ25A29	727565	0.01
	Estatística F		0.00

Fonte: elaboração própria.

O modelo com variáveis socioeconômicas possui R^2 de 66,7%, MAPE de 25% e RMSE de 16,4. Neste caso, as correlações inversas ocorreram com as variáveis: T_SUPER25M (média do percentual da população com 25 anos ou mais com ensino superior), P_SUPER (média do percentual de ocupados com ensino superior), AGUA_ESGOTO (média do percentual de pessoas em domicílios com água e esgoto adequados), RAZDEP (média da razão de dependência) e T_ANALF15M (média da taxa de analfabetismo da população de 15 anos ou mais). Portanto, temos 3 de 5 variáveis que, em média, contribuem para um menor número de ocorrências de roubos que são educacionais-econômicas, uma econômica e uma de infraestrutura. Complementarmente, as variáveis que possuem correlação positiva são: PREN60 (média do percentual da renda total apropriada pelo 60% mais pobres), GINI (índice de Gini), T_ANALF18M (média da taxa de analfabetismo da população de 18 anos ou mais) e T_FREQ25A29 (média do percentual de pessoal de 25 a 29 anos matriculadas em escolas). Por conseguinte, temos aqui uma variável estritamente econômica, um indicador social e duas variáveis educacionais.

Recomendações

Em geral, os modelos implementados permitiram compreender as relações entre quantidade de ocorrências de furtos e roubos de celulares e variáveis explicativas definidas nas duas bases de dados utilizados nesta pesquisa. Tal como nos trabalhos de Castro (2020) e Rosa (2019), nossa pesquisa também identificou variáveis de localização do crime como relevantes

para explicar a variabilidade dos dados. No entanto, em Rosa (ibidem), os fatores socioeconômicos apresentaram baixo poder de explicação dos roubos. No nosso caso, todos os modelos apresentaram resultados significativos, principalmente considerando um problema social tão complexo.

Considerando os insights de Risso (2016), fatores socioeconômicos foram perseguidos nesta pesquisa para que ações denominadas pela autora como situacionais - que modificam o contexto onde mais as ocorrências acontecem - e sociais - direcionadas para grupos de indivíduos com elevado potencial de desenvolverem comportamentos agressivos. Segundo a autora os municípios possuem uma grande vantagem competitiva na produção e interpretação de dados sobre violência, pois muitos dos fatores de risco associados a diferentes tipos de violência não estão limitados ao âmbito criminal. Por exemplo, dados desagregados sobre evasão escolar são fortemente relacionados ao envolvimento de adolescentes com a violência

Nessa linha, os modelos de árvores de decisão, apesar do baixo poder preditivo, foram relevantes para encontrar melhores variáveis explicativas de acordo com sua importância relativa. Tanto os modelos de árvore, quanto os modelos de regressão linear múltipla sugerem que existe uma relação entre ocorrências de roubos e distritos mais periféricos da cidade de São Paulo.

Para os modelos de ocorrências de roubos, essas relações se apresentam, principalmente, através de variáveis econômicas que expressam desigualdades sociais como renda domiciliar per capta, índice de Theil, percentual de trabalhadores no mercado informal e índice de Gini; e variáveis educacionais como taxas de analfabetismo, percentual de indivíduos com ensino superior e percentual de adultos matriculados em escolas.

Finalmente, cabe ainda destacar algumas das limitações desta pesquisa. Por ser um tema com poucas referências anteriores é muito provável que não tenhamos incluído variáveis explicativas mais interessantes. Por exemplo, como a qualidade da variável de hora da ocorrência não estava adequada, isto impossibilitou o seu uso, o que pode ter prejudicado o poder preditivo dos modelos, já que para a literatura discutida esta é uma das principais variáveis. Ademais, como apontado em Breiman (2009), Random Forest pode ser uma técnica mais adequada para se obter ganhos em acurácia, mesmo que nosso dataset tenha poucas observações já que é baseado em distritos policiais ou administrativos. A escolha que fizemos por criar taxas para todas as variáveis do RDO reduziu o poder preditivo dos modelos, ao passo que nos deu uma maior compreensão do contexto. Fato que difere muito das pesquisas de Castro (2020) e Vargas (2019) que optaram por dados de contagem.

Referências

CAMPOS, Sandro Luís Brandão; DE FIGUEIREDO, Josiel Maimone. Aplicação de Inteligência Artificial no Ciclo de Políticas Públicas. Cadernos de Prospecção, v. 15, n. 1, p. 196-214, 2022.

CASTRO, Ursula Rosa Monteiro de. Explorando aprendizagem supervisionada em dados heterogêneos para predição de crimes. 2020.

DUARTE, Daniel Edler; LOBATO, Luisa Cruz. A política do policiamento preditivo: pressupostos criminológicos, técnicas algorítmicas e estratégias punitivas. Revista Brasileira de Ciências Criminais, 29, 2021. p. 57-98.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil, 2017.

HEBBALI, Aravind; HEBBALI, Maintainer Aravind. Package 'olsrr'. Version 0.5, v. 3, 2017.

KAPLAN, Jacob. fastDummies: Fast creation of dummy (binary) columns and rows from categorical variables. R package version, v. 1, n. 1, 2020.

PETERSON, Brian G. et al. Package 'performanceanalytics'. R Team Cooperation, v. 3, p. 13-14, 2018.

RISSO, Melina Ingrid. Prevenção da violência: construção de um novo sentido para a participação dos municípios na segurança pública. Revista Brasileira de Segurança Pública, v. 10, n. 2, 2016.

ROSA, Amanda Gadelha Ferreira. Multicritério em segurança pública: uma aplicação no contexto de roubos. 2019. Dissertação de Mestrado. Universidade Federal de Pernambuco.

TIERNEY, N. et al. naniar: Data Structures, Summaries, and Visualizations for Missing Data. R package version 0.6.1.

VARGAS, Wagner Augusto Lopes de. Data science & segurança pública: padrões estatísticos sobre as ocorrências de flagrantes em roubo de celular na cidade de São Paulo. Artigo para obtenção do título de Mestre. Escola de Administração de Empresas de São Paulo, Fundação Getúlio Vargas, São Paulo, SP, Brasil. 2019.