ISBN: 978-628-95207-8-1. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2024.1.1.1757

Use of cluster analysis to study crime in the State of Rio de Janeiro

Max William Coelho Moreira de Oliveira, PhD[1], Miguel Fernández Pérez, PhD[2], Aldo Fernández Pérez, MEng[3], Wagner Santos, MSc[4], Antonio Costa Neto, MSc[5]
[1,4,5]Coordenadoria de Assuntos Estratégicos, Secretaria de Estado de Polícia Militar, Rio de Janeiro, Brasil, mwcoliveira@yahoo.com.br, wgsbusiness@gmail.com, antoniocostaneto@gmail.com
[2]Group of Applied Operations Research, Department of Engineering, Pontifical Catholic University of Peru, Lima, Peru, mfernandezp@pucp.edu.pe
[3]Instituto de Computação, Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brasil, aldofernandez@id.uff.br

Abstract– ***This article aims to construct clusters based on historical data of thefts in the State of Rio de Janeiro, aiming to identify possible similarities among the records. Monthly quantities of vehicle thefts, robberies on public transportation, pedestrian robberies, cell phone thefts, cargo thefts, and robberies at commercial establishments were selected. Using these records, the k-means algorithm was employed to build clusters, resulting in two subsets of records. These subsets present distinct characteristics and are valuable for analyzing the interaction between different types of thefts in a disaggregated manner, thus avoiding statistical fallacies. Additionally, we propose a classification model that establishes criteria for assigning scenarios to a specific cluster. This model can assist in developing more effective strategies in public security, and in the use of human and logistical resources.***

Keywords-- Public security; Clusters; Dimensional reduction; Correlation; Classification; Machine Learning.

Uso da análise de cluster para o estudo da criminalidade no Estado do Rio de Janeiro

Max William Coelho Moreira de Oliveira, PhD[1], Miguel Fernández Pérez, PhD[2], Aldo Fernández Pérez, MEng[3], Wagner Santos, MSc[4], Antonio Costa Neto, MSc[5]
[1,4,5]Coordenadoria de Assuntos Estratégicos, Secretaria de Estado de Polícia Militar, Rio de Janeiro, Brasil, mwcoliveira@yahoo.com.br, wgsbusiness@gmail.com, antoniocostaneto@gmail.com
[2]Group of Applied Operations Research, Department of Engineering, Pontifical Catholic University of Peru, Lima, Peru, mfernandezp@pucp.edu.pe
[3]Instituto de Computação, Universidade Federal Fluminense, Niterói, Rio de Janeiro, Brasil, aldofernandez@id.uff.br

Resumo– O presente artigo tem como objetivo a construção de clusters a partir de dados históricos de roubos no Estado do Rio de Janeiro, buscando identificar possíveis semelhanças entre os registros. Foram selecionadas as quantidades mensais de roubos de veículo, roubos em coletivo, roubos a transeunte, roubos de celular, roubos de carga e roubos a estabelecimento comercial. Utilizando esses registros, empregamos o algoritmo k-means para a construção dos clusters, resultando em dois subconjuntos de registros. Esses subconjuntos apresentam características distintas e são úteis, permitindo uma análise desagregada da interação entre os diferentes tipos de roubos, evitando falácias estatísticas. Adicionalmente, propomos um modelo de classificação que estabelece critérios para atribuir cenários a um cluster específico. Esse modelo pode contribuir para a elaboração de estratégias mais eficazes na segurança pública, e no uso de recursos humanos e logísticos.

Palavras chave-- Segurança pública; Clusters; Redução dimensional; Correlação; Classificação; Aprendizado de máquina.

I. INTRODUÇÃO

O trabalho das forças policiais é um desafio constante em prol da sociedade como um todo. No caso do Brasil, as Polícias Militares, encarregadas da preservação da ordem pública, buscam agir antes do crime ocorrer, tendo a prevenção como a estratégia mais eficaz para retrair custos, minimizar erros e antecipar problemas institucionais. No contexto do Rio de Janeiro, a otimização das estratégias de policiamento desenvolvido pela Polícia Militar do Estado do Rio de Janeiro é uma prioridade devido aos desafios relacionados à criminalidade. Abordagens inovadoras e baseadas em evidências são necessárias para enfrentar a crescente complexidade dos padrões criminais e garantir a eficácia das operações policiais. A Referência [1] ressalta que em 2019, segundo dados do Instituto de Segurança Pública do Governo do Estudo do Rio de Janeiro, foram registradas mais de 100 mil ocorrências de roubos, mais de 90 mil casos de furtos e foram contabilizadas mais de 30 vítimas de latrocínio. Além disso, indica que os crimes de maior impacto na sensação de insegurança da população referem-se a roubos de veículos e de rua, desta forma a prioridade da segurança pública se encontra nas ruas do Rio de Janeiro. Na Referência [2] encontraram uma correlação entre as mortes por operação policial e a atividade criminal, indicando que há uma associação entre o aumento da letalidade policial e a redução dos índices de criminalidade no nível local. A Referência [3] fornece informações relevantes para a mitigação de roubos a pedestres a partir da análise de distribuição temporal de crimes, demostrando que é possível modelar ou relacionar índices de criminais com variáveis temporais para encontrar tendências no aumento do crime em áreas urbanas.

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

II. BANCO DE DADOS

Os dados de Segurança Pública do Estado do Rio de Janeiro são divulgados pelo Instituto de Segurança Pública [4]. Nesta pesquisa estudam-se seis tipos de roubos presentes no relatório "Estatísticas de segurança: série histórica mensal por área de delegacia desde 01/2003". No qual, basicamente, contabilizam diversos tipos de crimes, desde crimes violentos, crimes contra o patrimônio, sequestros e até o quantitativo de policiais mortos em serviço. Estes dados podem ser agrupados por região, município, ano, mês, entre outros, o que facilita sua extração e segmentação. Os roubos estudados no presente artigo são: roubo de veículo, roubo em coletivo, roubo a transeunte, roubo de celular, roubo de carga e roubo em comércio (variáveis do problema). A série histórica inclui dados mensais no período de 2012 até 2022. Selecionou-se estes registros por ter apresentado correlações fortes em uma análise preliminar. Cabe destacar que esta análise foi verificada em conjunto com profissionais da área de Segurança Pública no que tange na lógica das variáveis explicativas do problema em estudo [5].

III. APLICAÇÃO DO ALGORITMO K-MEANS

A análise de cluster é uma disciplina da estatística multidimensional cujo propósito é identificar grupos (clusters) de elementos que compartilham semelhanças entre si e, ao mesmo tempo, apresentam diferenças em relação aos elementos presentes nos demais clusters.

A Referência [6] adota o termo *k*-means para descrever o processo de atribuição de um elemento a um dos *k* clusters predefinidos, com base no critério do centroide mais próximo.

A chave para esse procedimento é que o centroide é ajustado após cada nova atribuição, levando em consideração os membros do cluster. O algoritmo pode ser resumido da seguinte forma:

Passo 1: Divida os elementos em k clusters iniciais.
Passo 2: Percorrer os elementos da lista, atribuindo um elemento ao cluster cujo centroide é o mais próximo. A distância euclidiana é geralmente usada.
Passo 3: Recalcule o centroide para o cluster que ganha o novo elemento e para o cluster que perde o elemento.
Passo 4: Repita o Passo 2 até que não sejam feitas mais reatribuições.

A continuação, são analisados os registros mensais dos roubos entre 2012 e 2022 no Estado de Rio de Janeiro e é aplicado o algoritmo k-means [7, 8].

A. Escalonamento das variáveis

Dada a diferença de valores entre as variáveis, optou-se por escalonar estas variáveis antes da aplicação do algoritmo k-means. Se uma variável tiver uma magnitude muito maior que as demais variáveis, ela determinará em grande parte o valor distância (semelhança) obtida e alterando negativamente a agrupação final. Assim, cada registro j de uma variável específica y é escalonada como segue:

$$\frac{y_j - \bar{y}}{s_y}$$

Onde $\bar{y}$ e s_y são a média e o desvio padrão dos dados em y, respectivamente.

B. Estimação do número ótimo de clusters

O número ótimo de clusters determinou-se analisando o Within-Cluster Sum of Squares ($WCSS$) e o Calinski-Harabasz Index (CHI) para um número de diferente de clusters.

O $WCSS$ é calculado com base na fórmula abaixo:

$$WCSS = \sum_{i}^{k} \sum_{x \in C_i} \|x - m_i\|^2$$

Onde, k é o número de clusters; x faz referência aos elementos de um cluster; C_i é o i-ésimo cluster; m_i é o centroide do cluster i; e $\|x - m_i\|^2$ é a distância euclidiana ao quadrado entre dois vetores. O $WCSS$ mede a compacidade ou coesão dos clusters, quanto menor é melhor, e normalmente diminui com o aumento do número de clusters. Para achar a faixa potencial do número de clusters a ser considerados, deve-se identificar aquele valor a partir do qual o $WCSS$ não é mais substancial.

O CHI é calculado com base na fórmula abaixo:

$$CHI = \frac{BCSS/(k-1)}{WCSS/(n-k)}$$

Onde, n é o número total de registros; e $BCSS$ (Between-Cluster Sum of Squares) é a soma ponderada das distâncias euclidianas ao quadrado entre cada centroide do cluster e o centroide geral dos dados:

$$BCSS = \sum_{i}^{k} n_i \|m - m_i\|^2$$

Aqui, n_i é o número de registros do C_i; e m é o centroide geral dos dados. O $BCSS$ mede quão bem os clusters estão separados uns dos outros, quanto maior é melhor.

Presume-se que o número ótimo de clusters é igual ao valor máximo do CHI . Um valor elevado deste índice está relacionado com a maximização da razão do $BCSS$ e $WCSS$, o que significa que os clusters determinados apresentam diferenças consideráveis, e os elementos que pertencem a um cluster específico apresentam uma forte similitude. A evolução do $WCSS$ e CHI são mostrados nas figuras 1 e 2.

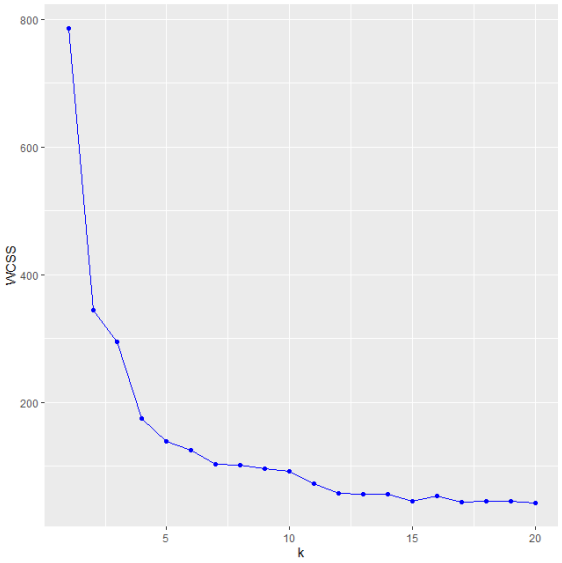

Fig. 1 Valores do $WCSS$ para um número diferente de clusters.

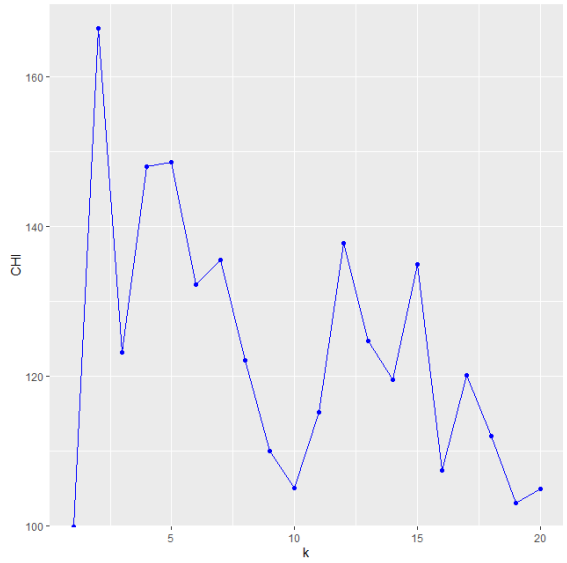

Fig. 2 Valores do CHI para um número diferente de clusters.

Nesta análise, o número ideal de clusters é estabelecido em 2.

C. *Análise de componentes principais*

O seguinte é obter uma visualização dos clusters resultantes. Dado que o número de variáveis do problema em estudo é igual a 6, é necessário realizar uma Análise de Componentes Principais (ACP) y representar as duas primeiras componentes principais. A ACP é uma técnica amplamente empregada em aprendizado de máquina não supervisionado, focada na redução de dimensionalidade em conjuntos de dados, com o objetivo de preservar a máxima quantidade de informação possível. A ACP utiliza uma transformação ortogonal para converter um conjunto de variáveis possivelmente correlacionadas em um conjunto de novas variáveis linearmente não correlacionadas, chamadas componentes principais. O primeiro componente principal tem a maior variância possível (ou seja, representa a maior parte da informação ou variabilidade nos dados), e cada componente subsequente, em ordem, tem a maior variância possível sob a restrição de que seja ortogonal aos componentes anteriores [7, 8]. A Tabela I apresenta de contribuição na variância de cada componente principal gerado.

TABELA I.
DISTRIBUIÇÃO DA VARIÂNCIA - ACP

	Dim1	Dim2	Dim3	Dim4	Dim5	Dim6
Variância	4,3	1,2	0,2	0,2	0,1	0,0
Variância total	6,0					
%Variância	71,4	20,3	4,0	2,6	1,1	0,6
%Variância cumulada	71,4	91,6	95,7	98,2	99,4	100,0

Na prática, um número de dimensões que atinge um %Variância cumulada de 90% ou superior, indica uma boa aproximação para representar o conjunto dos dados. Neste caso, as duas primeiras dimensões capturam 91.6% da variância total, que pode ser considerado adequado. A Figura 3 ilustra o resultado da ACP para representar os dois clusters gerados pelo algoritmo k-means.

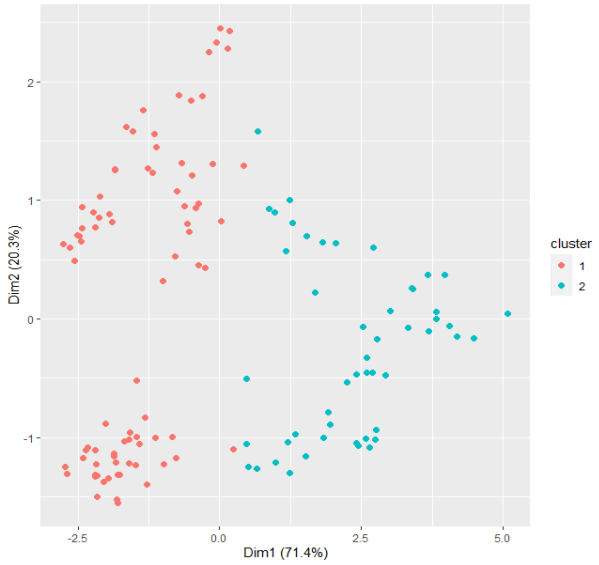

Fig. 3 Gráfico de dispersão das dos primeiras variáveis resultantes da ACP.

O gráfico de dispersão revela que os registros foram adequadamente separados, sem a presença significativa de dados atípicos.

IV. RESULTADOS E DISCUSSÃO DA CLUSTERIZAÇÃO

Agora, os clusters obtidos podem ser organizados em relação aos meses dos registros coletados. A Tabela II detalha esta organização e as figuras 4-9 mostram o histórico dos roubos enfatizando os clusters. Pode-se observar uma clara divisão dos dados (na maioria dos casos) na quantidade de roubos nos anos 2012, 2013, 2014, 2015, 2020, 2021 e 2022 pertencentes ao Cluster 1, com a quantidade dos roubos nos anos 2016, 2017, 2018 e 2019 pertencentes ao Cluster 2. Com a finalidade de ter uma medida que reflita as diferenças entre os cluster, a Tabela III calcula os valores médios das variáveis do estudo conforme o cluster atribuído. Nesta última tabela, fica claro a diferença entre os dados. A proporção entre a quantidade de roubos do Cluster 2 e o Cluster 1 tem um valor mínimo de 1,28 correspondente aos roubos de comercio e um valor máximo de 2,25 correspondente aos roubos de transeuntes.

TABELA II.
ATRIBUIÇÃO RESULTANTE DOS REGISTROS AOS CLUSTERS

Clusters	N° Registros	Porcentagem	Meses
Cluster 1	82	62,1	Jan-Dez 2012; Jan-Dez 2013; Jan-Dez 2014; Fev-Nov 2015; Fev 2017; Set 2019; Mar-Dez 2020; Jan-Dez 2021; Jan-Dez 2022.
Cluster 2	50	37,9	Jan 2015; Dez 2015; Jan-Dez 2016; Jan 2017; Mar-Dez 2017; Jan-Dez 2018; Jan-Ago 2019; Out-Dez 2019; Jan-Fev 2020;

TABELA III.
VALORES MÉDIOS DAS VARIÁVEIS POR CLUSTER

	Cluster 1	Cluster 2	Proporção
Roubo de veículo	2237	3864	1,73
Roubo em coletivo	605,5	1249	2,06
Roubo de transeunte	4474	7259	1,62
Roubo de celular	895	2018	2,25
Roubo de carga	394	776.8	1,97
Roubo de comercio	418,7	535,3	1,28

Uma vez determinados os clusters, pode-se analisar a correlação entre as variáveis e identificar a dependências entre elas. As figuras 10 e 11 apresentam o gráfico de dispersão para o Cluster 1 e Cluster 2, respectivamente, destacando a correlação entre as variáveis.

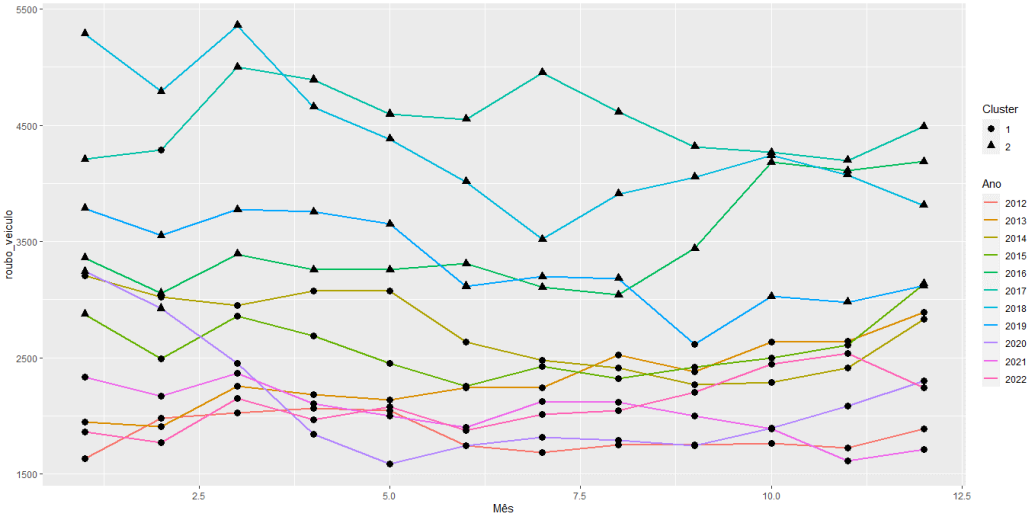

Fig. 4 Quantidade mensal de roubos de veículos no Estado do Rio de Janeiro nos períodos 2012-2022.

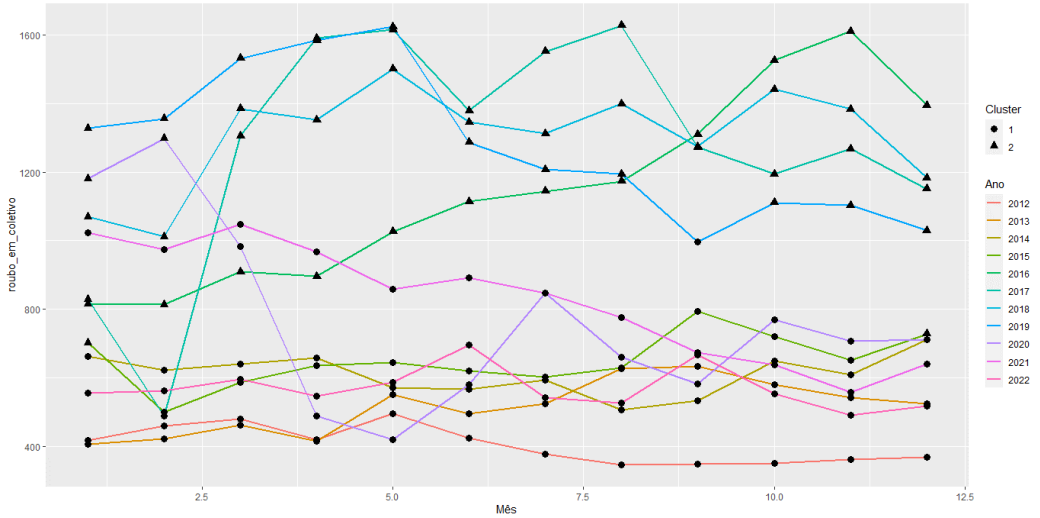

Fig. 5 Quantidade mensal de roubos em coletivo no Estado do Rio de Janeiro nos períodos 2012-2022.

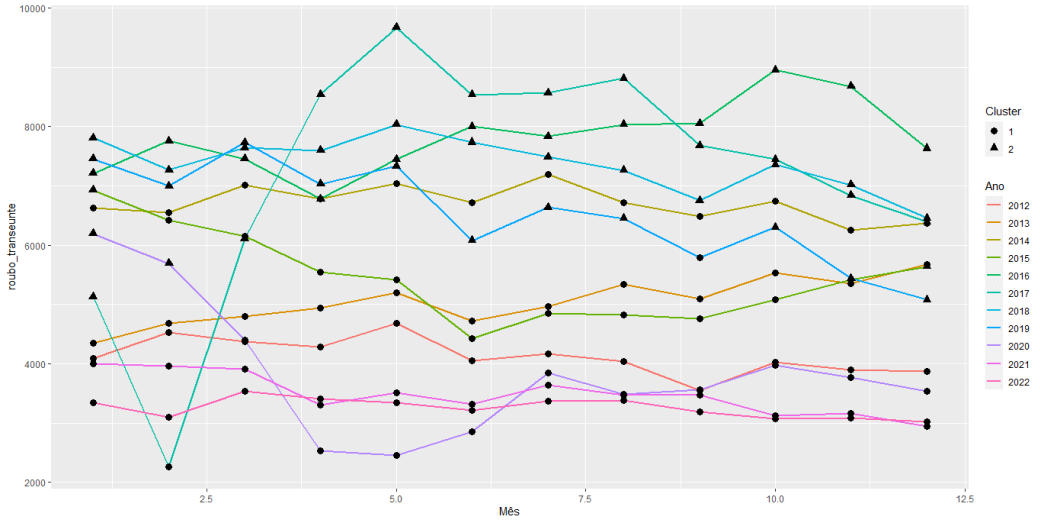

Fig. 6 Quantidade mensal de roubos a transeuntes no Estado do Rio de Janeiro nos períodos 2012-2022.

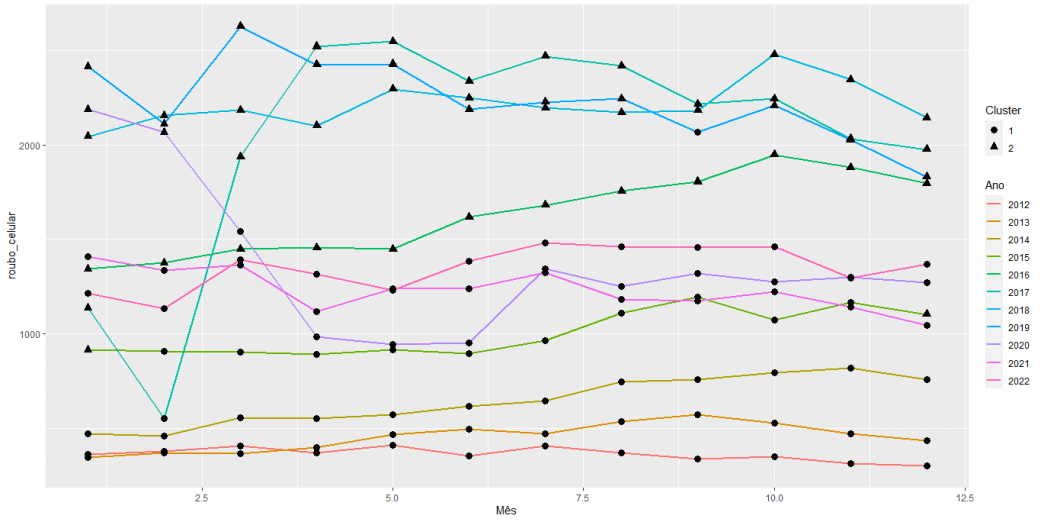

Fig. 7 Quantidade mensal de roubos de celular no Estado do Rio de Janeiro nos períodos 2012-2022.

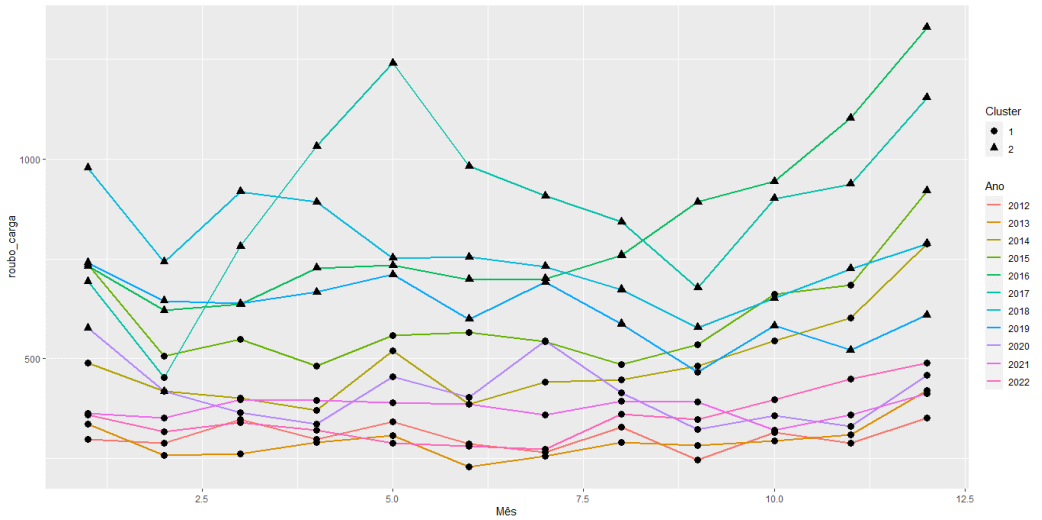

Fig. 8 Quantidade mensal de roubos de carga no Estado do Rio de Janeiro nos períodos 2012-2022.

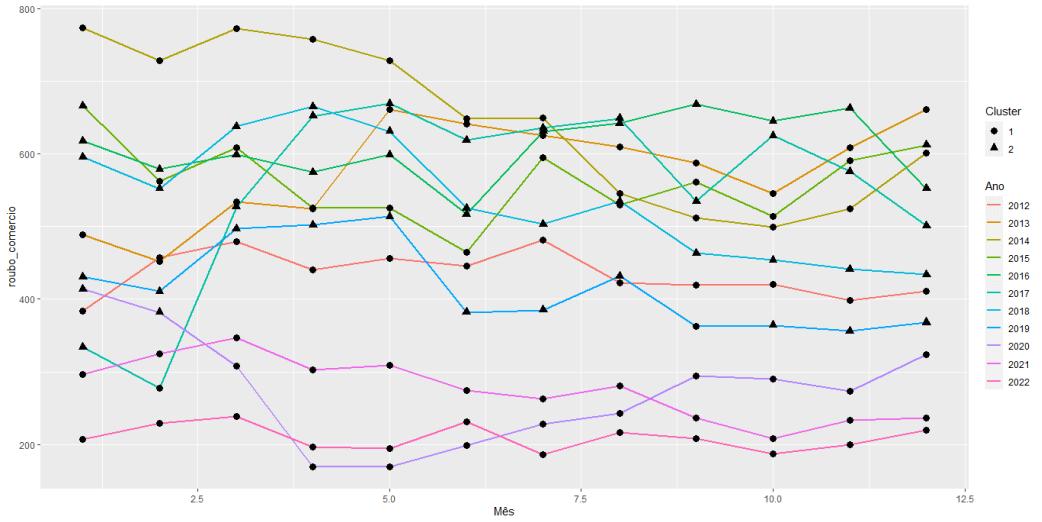

Fig. 9 Quantidade mensal de roubo em comércios no Estado do Rio de Janeiro nos períodos 2012-2022.

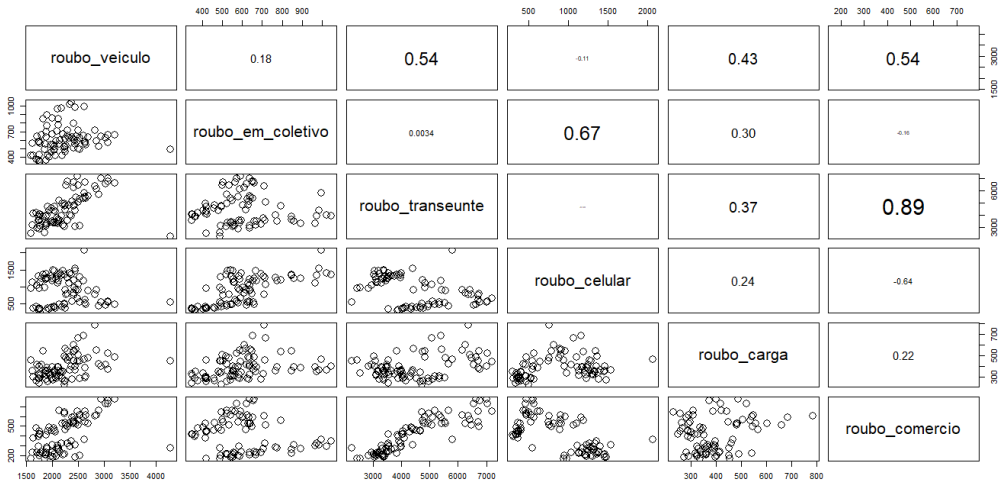

Fig. 10 Gráfico de dispersão e correlação dos roubos no Estado do Rio de Janeiro no Cluster 1.

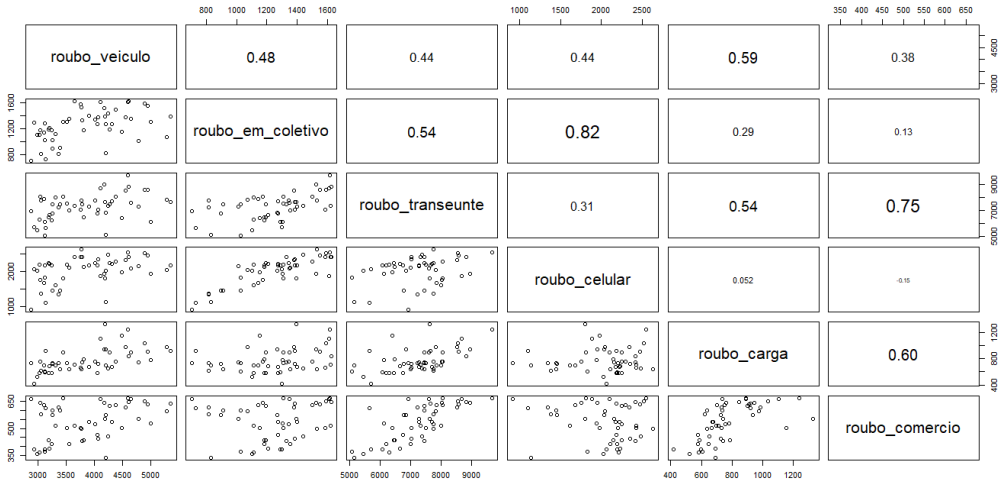

Fig. 11 Gráfico de dispersão e correlação dos roubos no Estado do Rio de Janeiro no Cluster 2.

No Cluster 1, a dependência mais forte se dá entre a quantidade de roubos a transeuntes e a quantidade de roubos a comércios, apresentando uma correlação positiva. No Cluster 2, a dependência mais forte se dá entre a quantidade de roubos de celular e a quantidade de roubos em coletivo, apresentando uma correlação positiva.

V. APLICAÇÃO DO ALGORITMO DE CLASSIFICAÇÃO

Ao analisar a Figura 3, a decisão foi desenvolver um algoritmo de classificação devido à clara separação entre os clusters. O objetivo é propor uma regra para determinar o nível da quantidade de roubos em cenários futuros. O algoritmo selecionado é a Análise Discriminante Linear (ADL) [8]. As variáveis $Dim1$ y $Dim2$ geradas pela ACP serão empregadas, pois estas variáveis representam adequadamente dos dados do problema em estudo. A Figura 12 mostra o resultado de aplicar o ADL. Apenas, 3 das 132 predições feitas pelo modelo do ADL estavam erradas, com um erro de treinamento baixo de 2,27%, o que indica que o modelo é bom. A equação da ADL é dada por:

$$1{,}0776Dim1 - 0{,}3919Dim2 = 0{,}5082$$

Então, a classificação de um novo registro, obtém-se substituindo suas coordenadas (após o escalonamento e aplicação da transformação ortogonal da ACP) na equação $1{,}0776Dim1 - 0{,}3919Dim2 = Z$. Se o valor Z for menor que $0{,}5082$, o novo registro pertence ao Cluster 1; se o valor de Z for maior o igual que $0{,}5082$, o novo registro pertence ao Cluster 2.

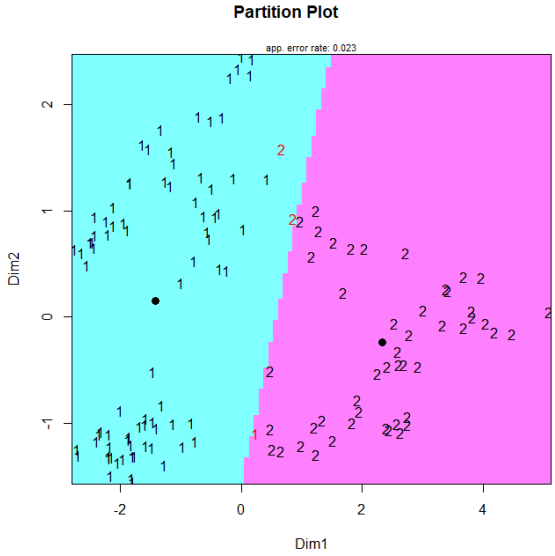

Fig. 12 Gráfico de classificação das dos primeiras variáveis resultantes da ACP.

VI. CONCLUSÕES

A análise dos dados do Instituto de Segurança Pública do Governo do Estudo do Rio de Janeiro permitiu a construção de dos clusters da série histórica de crimes no período de 2012 até 2022. O 62,1% dos registros pertencem ao Cluster 1 e 37,9% pertencem ao Cluster 2, eles apresentam níveis de roubos distintos. Principalmente, o Cluster 1 abarca os anos de 2012-2015 e 2020-2022, e o Cluster 2 os anos 2016-2019.

A clusterização permitiu determinar uma correlação mais confiável entre as variáveis do que se fossem analisados de forma agregada. Além disso, a correlação não implica causalidade; ou seja, embora duas variáveis estejam correlacionadas, uma não causa necessariamente a outra.

Analisando Figura 10, o conjunto de interações moderadas a fortes do Cluster 1 são:

- Roubos a veículos - roubos em comercio - roubos a transeuntes.
- Roubos a veículos - roubos de carga.
- Roubos em coletivo - roubos de celulares.

Analisando Figura 11, o conjunto de interações moderadas a fortes do Cluster 2 são:

- Roubos a veículos - roubos em coletivo - roubos a transeuntes - roubos de celulares.
- Roubos a veículos - roubos em comercio - roubos de carga.
- Roubos em comercio - roubos de carga - roubos a transeuntes.

As correlações resultantes dos clusters apresentam coerência com a realidade, já que podem ser interpretadas da seguinte maneira, por exemplo, em um caso de roubo em um ônibus, este inclui o roubo de celular dos passageiros e em alguns casos o roubo de pessoas presentes num ponto de ônibus ou prestes a embarcar. Quando é realizado um roubo de veículo, o criminoso rouba também os pertences da vítima, entre eles o celular. Complementarmente, quando é realizado roubo de veículo, este pode ser utilizado pelos criminosos para realizar crimes mais complexos, como são o roubo em comércio e o roubo de carga, onde os veículos podem servir como veículo de fuga ou até para carregamento de material roubado. Por outro lado, quando é realizado roubo no local comercial, tanto clientes, funcionários e transeuntes que passam perto do local, tornam-se vítimas do grupo criminoso.

Depois, o modelo de classificação é uma iniciativa para identificar padrões (atribuição a um cluster específico) entre as quantidades de roubos mensais, e pode servir para a elaboração de estratégias de segurança pública, mediante a estimativa do uso de recursos humanos e logísticos baseadas em experiências passadas.

O artigo destaca a viabilidade da aplicação de técnicas de aprendizado de máquina em dados criminais, proporcionando a compreensão sobre os fatores que influenciam o comportamento dos criminosos, a dinâmica dos roubos, e a definição de prioridades nas ações policiais, revelando conhecimentos valiosos para cada situação. Ressalta-se que essas abordagens e análises são fundamentais, pois fornecem informações valiosas que podem ser utilizadas na formulação de medidas de segurança para a sociedade. Isso não apenas aprimora as operações da Polícia Militar do Estado do Rio de Janeiro, mas também pode beneficiar outras forças policiais militares nacionais e internacionais encarregadas da aplicação de Polícia Ostensiva. Além disso, o artigo contribui significativamente para enriquecer o debate acadêmico sobre o papel das análises de dados na melhoria da segurança pública.

REFERÊNCIAS

[1] C. Grillo, and L. Martins, "Indo até o problema: Roubo e circulação na cidade do Rio de Janeiro," *Dilemas: Revista de Estudos de Conflito e Controle Social*, vol.13, no. 3, pp. 565-590, 2021.

[2] J. Monteiro, E. Fagundes, and J. Guerra, "Letalidade policial e criminalidade violenta," *Revista de Administração Pública*, vol. 54, pp. 1772-1783, 2020.

[3] T. Reis, and A. Kipper, "Análise da distribuição temporal de roubos a pedestres em áreas urbanas: o caso de Porto Alegre," *Revista Brasileira de Segurança Pública*, vol. 17, no. 2, pp. 348-369, 2023.

[4] Dados abertos do Instituto de Segurança Pública do Governo do Estado do Rio de Janeiro, Estatísticas de segurança. http://www.ispdados.rj.gov.br/estatistica.html

[5] M. W. Coelho, A. Pérez, and M. Fernández, "Emprego de Correlação Estatística na Análise Criminal para Otimização do Emprego do Policiamento Ostensivo na PMERJ," *Seminário Internacional: Ciência, Tecnologia e Inovação em Segurança Pública*, 2023.

[6] J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, no. 14, pp. 281-297, 1967.

[7] K. Faceli, A. C. Lorena, J. Gama, and A. Carvalho, *Inteligência Artificial: Uma abordagem de aprendizado de máquina*, LTC, Rio de Janeiro, 2011.

[8] M. J. Zaki, and W. Meira, *Data mining and analysis: Fundamental concepts and algorithms*, Cambridge University Press, 2014.