UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise Exploratória de Dados de Violência Usando Algoritmos de Agrupamento Particional

Stepheson Alves de Oliveira

Monograph - MBA in Artificial Intelligence and Big Data

ICMC USP
SÃO CARLOS

Stepheson Alves de Oliveira

Análise Exploratória de Dados de Violência Usando Algoritmos de Agrupamento Particional

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Concentration area: Artificial Intelligence and Big Data

Orientador: Profa. Dra. Veronica Oliveira de Carvalho

Versão original

São Carlos

2025

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Alves de Oliveira, Stepheson
S856m Análise Exploratória de Dados de Violência Usando Algoritmos de Agrupamento Particional / Stepheson Alves de Oliveira ; orientador Nome Orientador. – São Carlos, 2025.
75 p. : il. (algumas color.) ; 30 cm.

Monograph (MBA in Artificial Intelligence and Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2025.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. sobrenome orientador, nomes orientador, orient. II. Título.

Stepheson Alves de Oliveira

Exploratory Analysis of Violence Data Using Partitional Clustering Algorithms

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Área de concentração: Inteligência Artificial e Big Data

Advisor: Profa. Dra. Veronica Oliveira de Carvalho

Original version

São Carlos

2025

AGRADECIMENTOS

A Deus, por sentir que me norteia e me equilibra diante dos desafios que escolho superar.

Aos meus pais, que sempre me asseguraram suporte em cada ambição que minha imaginação me convida a ousar perseguir.

À minha orientadora, Profa. Dra. Veronica Oliveira de Carvalho, pela dedicação, pela orientação atenta, pelo respeito e pela confiança com que compartilhou seus conhecimentos.

Aos professores do MBA em Inteligência Artificial e Big Data, pelo empenho em difundir conhecimento com atenção e pela pertinência do que compartilharam.

RESUMO

A. de Oliveira, Stepheson **Análise Exploratória de Dados de Violência Usando Algoritmos de Agrupamento Particional**. 2025. 75p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

A criminalidade urbana constitui um desafio multifacetado para gestores públicos, exigindo ferramentas analíticas capazes de identificar padrões e subsidiar políticas baseadas em evidências. Este trabalho tem como objetivo avaliar a aplicabilidade de algoritmos de agrupamento particional — especificamente k-means, k-DBA e k-shape — para identificar grupos de municípios do Estado de São Paulo com séries históricas mensais de criminalidade semelhantes, utilizando dados da Secretaria de Segurança Pública (SSP-SP) referentes a 18 tipos de crime. A metodologia abrangeu a coleta e organização de séries temporais de 645 municípios entre 2020 e 2025, com foco nos períodos de 1, 2, 3 e 5 anos. Após o pré-processamento, os algoritmos foram executados com k variando entre 2 e 15, testando diferentes números de clusters, e a qualidade dos agrupamentos foi avaliada por meio do coeficiente de silhueta, permitindo identificar o número ótimo de clusters e comparar o desempenho dos algoritmos. Os resultados indicaram que a eficácia dos agrupamentos depende da estrutura temporal e da frequência de cada tipo de crime. O k-means apresentou bom desempenho em crimes de baixo volume e baixa prevalência, em que as diferenças de magnitude foram suficientes para discriminar grupos. Por empregar DTW, o k-DBA mostrou-se mais robusto para capturar semelhanças de forma e dinâmica temporal, sendo particularmente eficaz em crimes de alta prevalência, nos quais o alinhamento elástico revelou padrões defasados que o k-means não capturou. No contexto deste estudo, o k-shape apresentou desempenho inferior, com clusters desbalanceados e comportamento semelhante à detecção de outliers, mostrando-se menos adequado aos dados e ao delineamento experimental aqui adotados. Observou-se predominância de soluções com $k = 2$, refletindo estruturas binárias (um *cluster* principal e outro "residual"). As matrizes de perfis de crime (volume $\times$ prevalência) e de relação algoritmo–perfil criminal, elaboradas neste estudo, consolidaram os achados e funcionam como guias diagnósticos para estimar a "clusterabilidade" de cada crime. Em síntese, o trabalho demonstra a viabilidade de técnicas de agrupamento como ferramentas de apoio à gestão de segurança pública, possibilitando que gestores identifiquem municípios com perfis criminais semelhantes e aloquem recursos de forma estratégica, contribuindo para políticas públicas baseadas em evidências e fortalecendo a capacidade do Estado de prevenir e combater a criminalidade.

Palavras-chave: Agrupamento de séries temporais. Análise criminal. k-means. k-DBA. k-shape. Segurança pública. Coeficiente de silhueta.

ABSTRACT

A. de Oliveira, Stepheson **Exploratory Analysis of Violence Data Using Partitional Clustering Algorithms**. 2025. 75p. Monograph (MBA in Artificial Intelligence and Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2025.

Urban crime represents a multifaceted challenge for public administrators, requiring analytical tools capable of identifying patterns and supporting evidence-based policies. This study aims to evaluate the applicability of partitional clustering algorithms — specifically k-means, k-DBA, and k-shape — in identifying groups of municipalities in the State of São Paulo with similar monthly crime time series, using data from the State Department of Public Safety (SSP-SP) covering 18 types of crime. The methodology encompassed the collection and organization of time series from 645 municipalities between 2020 and 2025, focusing on four temporal windows 1, 2, 3, and 5 years. After preprocessing, the algorithms were executed with k ranging from 2 to 15, testing different numbers of clusters, and the quality of the groupings was evaluated using the silhouette coefficient, allowing the identification of the optimal number of clusters and the comparison of algorithm performance. The results indicated that clustering effectiveness depends on the temporal structure and frequency of each crime type. The k-means algorithm performed well for crimes with low volume and low prevalence, where magnitude differences were sufficient to distinguish groups. By employing DTW, k-DBA proved more robust in capturing shape and temporal dynamics similarities, being particularly effective for high-prevalence crimes in which elastic alignment revealed phase-shifted patterns that k-means could not detect. In the context of this study, the k-shape algorithm exhibited inferior performance, producing unbalanced clusters and showing behavior similar to outlier detection, indicating that it was less suitable for the data and experimental setup used in this work. A predominance of solutions with $k = 2$ was observed, reflecting binary structures (a main cluster and a residual one). The crime profile matrices (volume $\times$ prevalence) and the algorithm–crime profile relation matrix developed in this study consolidated the findings and served as diagnostic guides to estimate each crime's clusterability. In summary, this work demonstrates the feasibility of clustering techniques as decision-support tools for public security management, enabling policymakers to identify municipalities with similar crime profiles and allocate resources strategically, thereby contributing to evidence-based public policies and strengthening the State's capacity to prevent and combat crime.

Keywords: Time series clustering. Crime analysis. k-means. k-DBA. k-shape. Public security. Silhouette coefficient.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AP	Alta Prevalência
AV	Alto Volume
BP	Baixa Prevalência
BV	Baixo Volume
CC	*Cross-Correlation*
DE	Distância Euclidiana
DTW	*Dynamic Time Warping*
ED	*Euclidean Distance*
MP	Média Prevalência
MV	Médio Volume
Prv.	Prevalência
SBD	*Shape-Based Distance*
SC	*Silhouette Coefficient*
SSP-SP	Secretaria da Segurança Pública de São Paulo
S.N.	Séries Nulas
Vol.	Volume de Ocorrências
k	Número de *Clusters*

SUMÁRIO

1 INTRODUÇÃO

"A criminalidade constitui um tema de alta relevância pública, impactando diretamente instituições e cidadãos. Contudo, entender as causas e consequências é algo complexo. Carneiro (2022) apresenta uma revisão da literatura sobre o tema no contexto brasileiro, embora o problema seja global (vide Hunter and Dantzker (2012)). Outros trabalhos estão disponíveis na literatura, como os de Steingraber (2024), Almeida *et al.* (2023), Gomes *et al.* (2023), Oliveira and Silva (2021), entre outros[1].

Não diferente de outras localidades no Brasil, a cidade de São Paulo apresenta um cenário de criminalidade em constante transformação. Diversas são as reportagens que apresentam índices de aumento e/ou diminuição nos mais variados tipos de crime (UOL, 2025; G1, 2025; EXAME, 2025). Os índices refletem a complexidade da violência urbana e reforçam a necessidade de estratégias de segurança pública eficazes, evidenciando a importância do trabalho da polícia para tentar conter e diminuir os índices de violência.

Visando dar transparência aos dados de criminalidade do Estado de São Paulo, a Secretaria de Segurança Pública do Estado de São Paulo (SSP-SP)[2] divulga dados mensais de criminalidade referentes a todos os municípios do estado de São Paulo[3]. Ao todo, são informados dados de 18 crimes distintos, a saber: (1) homicídio doloso, (2) homicídio doloso por acidente de trânsito, (3) homicídio culposo por acidente de trânsito, (4) homicídio culposo outros, (5) tentativa de homicídio, (6) lesão corporal seguida de morte, (7) lesão corporal dolosa, (8) lesão corporal culposa por acidente de trânsito, (9) lesão corporal culposa - outras, (10) latrocínio, (11) roubo de veículo, (12) roubo a banco, (13) roubo de carga, (14) furto - outros, (15) furto de veículo, (16) estupro de vulnerável, (17) estupro e (18) roubo - outros.

Com o intuito de contribuir para o tema apresentado, o **objetivo** deste trabalho é avaliar a aplicabilidade de algoritmos de agrupamento particional visando identificar grupos de municípios com séries históricas mensais de criminalidade semelhantes, avaliando para tanto a qualidade dos agrupamentos gerados a partir dos dados da SSP-SP. A ideia é apoiar os gestores a (i) identificar, assim como visualizar, padrões de similaridade entre os municípios, assim como, com base em um determinado grupo, (ii) facilitar a escolha de municípios para implementar uma determinada política de enfrentamento ao crime. Para tanto, este trabalho se propõe a: (a) coletar e organizar as séries temporais de criminalidade disponibilizadas pela SSP-SP; (b) realizar o pré-processamento das séries temporais para

---

1 A “Revista Brasileira de Segurança Pública” é um veículo especializado em temas relacionados à segurança pública.

2 <https://www.ssp.sp.gov.br/>.

3 <https://www.ssp.sp.gov.br/estatistica>.

posterior análise; (c) aplicar algoritmos de agrupamento particional, especificamente k-means, k-shape e k-DBA; (d) comparar e avaliar os resultados dos agrupamentos com base em métricas de qualidade, especificamente via coeficiente de silhueta; (e) discutir o potencial de uso dos agrupamentos como ferramenta de apoio à tomada de decisão para definição de estratégias de prevenção e combate à criminalidade. Por fim, embora a SSP-SP disponha de algumas visualizações [4] não há visualizações relacionadas à proposta apresentada neste trabalho.

O presente trabalho está estruturado da seguinte maneira: no Capítulo 2 são apresentados os conceitos necessários ao entendimento do mesmo; no Capítulo 3 a metodologia de análise de dados adotada, assim como os resultados obtidos; no Capítulo 4 as considerações finais.

[4] Via Microsoft Power BI.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta os fundamentos necessários ao entendimento deste trabalho, a saber: séries temporais (Seção 2.1), agrupamento de dados (Seção 2.2), além de uma breve descrição de alguns trabalhos relacionados (Seção 2.4).

2.1 Séries Temporais

O processo de observação de eventos para compreender seus comportamentos torna-se eficaz por meio da análise detalhada de seus períodos evolutivos. É uma prática usada há muitos anos e uma solução necessária até os tempos atuais. Estudos que dependem de coletas e observações sequenciais de dados, permitindo a previsão de seus comportamentos, são comuns em diversas áreas, as quais se estendem por campos tão diversos quanto a economia, a meteorologia e as ciências exatas e experimentais, evidenciando a importância das séries temporais como ferramenta de investigação.

Segundo GONÇALVES and Lopes (2008), uma série temporal Z consiste em uma sequência ordenada de observações de uma determinada variável, coletadas ao longo do tempo. Deste modo, $Z = (z_1, z_2, ..., z_m)$, em que Z representa uma série temporal de tamanho m e $z_t \in \mathbb{R}$, uma observação z no instante de tempo t (PARMEZAN; SOUZA; BATISTA, 2019; PARMEZAN; BATISTA, 2016). A ordem em que essas observações são registradas é fundamental, pois reflete a evolução da variável ao longo do período analisado. As observações podem ser obtidas por meio de diferentes métodos, como registros, pesquisas ou instrumentos de medição, e os intervalos de tempo entre as observações podem ser contínuos ou discretos.

Uma série temporal utiliza ativamente recursos como representações gráficas que facilitam observar e analisar a evolução de uma ou mais variáveis definidas por intervalos pressupostos. A Figura 1 apresenta um exemplo de série temporal referente às ocorrências de latrocínios que ocorreram em São Paulo entre os anos de 2001 e 2023.

Segundo Paparrizos, Yang and Li (2024) e Tan *et al.* (2018), séries temporais podem ser categorizadas em dois tipos principais, que orientam tanto o objetivo da análise quanto o tratamento das observações no contexto estudado. Essa categorização depende principalmente da dimensão de cada observação ao longo do tempo, ou seja, do número de variáveis consideradas simultaneamente. Assim, distinguem-se entre: (i) **séries univariadas:** a medição é baseada em uma única variável que varia ao longo do tempo para cada ponto de observação; (ii) **séries multivariadas:** a medição baseia-se em função de múltiplas variáveis ao longo do tempo para cada ponto de observação. Neste trabalho, embora a criminalidade seja influenciada por múltiplas variáveis — como escolaridade e

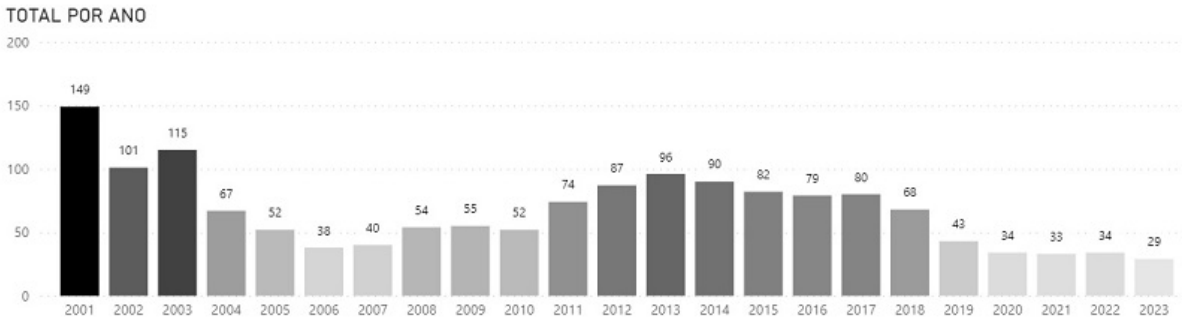

Figura 1 – Ocorrências de latrocínios entre os anos de 2001 e 2023 em São Paulo.

Fonte: SSP-SP (2024).

taxa de pobreza —, optou-se por uma abordagem univariada, analisando exclusivamente a evolução temporal de cada tipo de crime, sem incorporar variáveis externas.

Outra característica reforçada por Paparrizos, Yang and Li (2024) é que as séries temporais podem ser classificadas quanto à regularidade das suas medições. Uma série temporal regular é aquela em que as observações são coletadas em intervalos de tempo fixos, ou seja, o espaçamento entre os pontos de dados é constante. Por exemplo, medições feitas a cada segundo, minuto ou dia, representando uma sequência $\{z_1, z_2, \ldots, z_m\}$ onde o intervalo $\Delta t = t_{i+1} - t_i$ é constante para todo i. Por outro lado, em séries temporais irregulares, as observações ocorrem em instantes de tempo que não seguem intervalos fixos, ou seja, os espaçamentos variam ao longo da série, em que $\Delta t = t_{i+1} - t_i$ não é constante. Essa irregularidade traz desafios adicionais para a análise e o agrupamento dessas séries. Neste trabalho, todas as séries trabalhadas são regulares.

2.2 Agrupamento de Dados

Quando se precisa tomar decisões com base em informações complexas, fragmentá-las pode facilitar a compreensão. A associação para cada detalhe fragmentado por meio das semelhanças guiadas por um contexto é uma das maneiras que podem simplificar a tomada de decisões. Por exemplo, organizar fotos digitais em categorias como “viagens”, “família” e “trabalho” para melhor gerir a galeria de fotos de uma pessoa. Everitt, Landau and Leese (2001, p. 1) mencionam que “Uma das habilidades mais básicas das criaturas vivas envolve o agrupamento de objetos semelhantes para produzir uma classificação. A ideia de classificar coisas semelhantes em categorias é claramente primitiva, já que o homem primitivo, por exemplo, deve ter sido capaz de perceber que muitos objetos individuais compartilhavam certas propriedades, como serem comestíveis, venenosos, ferozes e assim por diante.”.

A própria linguagem contemporânea reflete intrinsecamente este mecanismo cognitivo. Cada palavra que designa uma categoria, seja “árvore”, “veículo” ou “ferramenta”,

funciona como um agrupador conceitual, reunindo sob um único termo diversos elementos que compartilham propriedades essenciais. Assim, quando alguém diz “cadeira”, está implicitamente ativando um agrupamento mental que abarca inúmeros objetos diferentes, mas que compartilham a função básica de servir como assento. Assim, quando se considera o aprendizado que se obtém todos os dias, percebe-se que as pessoas seguem um processo de compreensão padrão quase que universal. As pessoas conseguem identificar certos atributos de algo desconhecido de maneira quase que imediata ao se comparar com propriedades que já estão mapeadas em seu repertório mental. Segundo Xu and Wunsch (2008, p. 1), “Para aprender um novo objeto ou entender um novo fenômeno, as pessoas sempre tentam identificar características descritivas e comparar essas características com as de objetos ou fenômenos conhecidos, com base em sua similaridade ou dissimilaridade, generalizada como proximidade, de acordo com certas normas ou regras.”.

É neste contexto que a técnica de agrupamento (*clustering*) de dados surge: dada uma coleção de objetos, o objetivo do agrupamento é realizar a separação de tais objetos em grupos de tal modo que objetos similares sejam colocados em um mesmo grupo (“categoria”). A análise de um dado agrupamento não é um processo puramente objetivo, pois depende da interpretação e julgamento de quem realiza a análise, como pode ser observado na Figura 2.

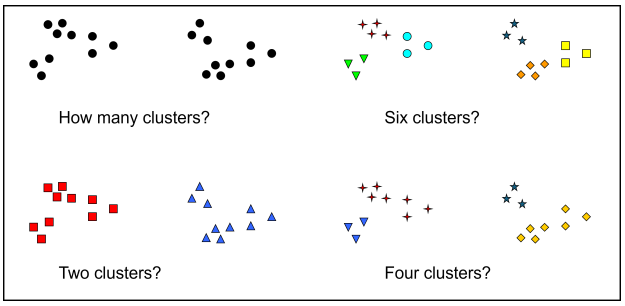

Figura 2 – Ilustração da subjetividade na análise de agrupamento.

Fonte: Tan *et al.* (2018, p. 311).

O agrupamento pode ser aplicado aos mais variados tipos de dados. Neste trabalho, os algoritmos são aplicados a séries temporais, cada uma representando a evolução de um tipo de crime em um município específico. Como discutido anteriormente, optou-se por uma abordagem univariada, o que facilita a interpretabilidade dos padrões identificados. Desse modo, os clusters agrupam municípios com dinâmicas temporais semelhantes para cada tipo de crime analisado separadamente.

A Figura 3 ilustra a diferença entre séries temporais univariadas e multivariadas no contexto do agrupamento. No lado esquerdo, tem-se um exemplo univariado, onde cada série é descrita por um único canal (variável) e os grupos (*clusters*) são formados com base em características simples, como o número e a altura dos picos na série. Já o lado

direito apresenta séries multivariadas, onde múltiplos canais (variáveis) são considerados simultaneamente — neste caso, três canais diferentes que capturam aspectos variados do comportamento temporal. O agrupamento multivariado leva em conta a interação entre esses canais, possibilitando a separação de padrões mais complexos que não seriam evidentes ao analisar cada canal isoladamente.

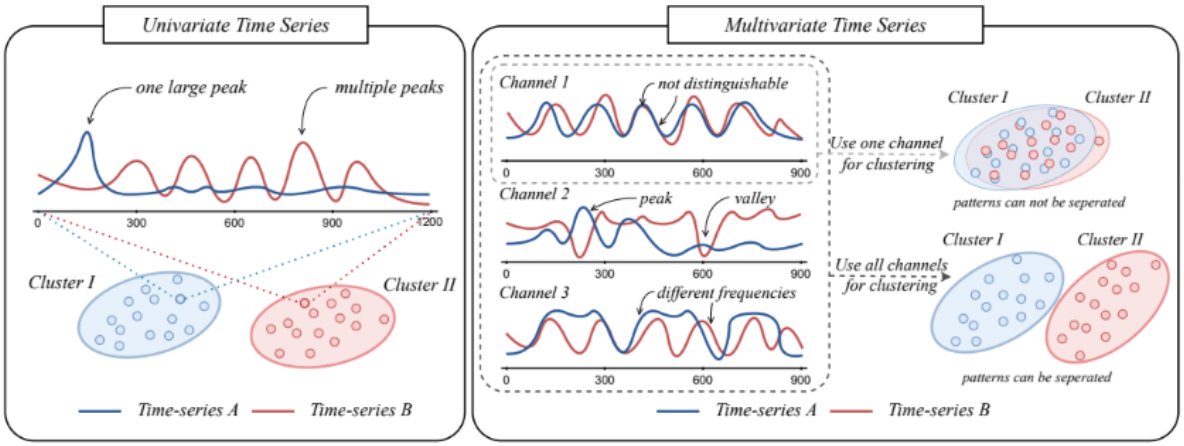

Figura 3 – Comparação de agrupamentos entre séries temporais univariadas e multivariadas.

Fonte: Paparrizos, Yang and Li (2024, p. 4).

2.2.1 Medidas de Distância para Séries Temporais

Para que o agrupamento seja viável, é necessário, inicialmente, definir como a proximidade entre os objetos será computada. Em outras palavras, deve-se especificar a medida de proximidade a ser utilizada. Essa proximidade pode ser instanciada tanto por meio de uma medida de similaridade quanto de dissimilaridade (TAN *et al.*, 2018). As medidas de distância são as mais empregadas quando se trata de dissimilaridade. Nesse caso, quanto mais semelhantes forem dois objetos, menor será a distância entre eles (EVERITT; LANDAU; LEESE, 2001; TAN *et al.*, 2018; ZOU, 2020). Assim, dado um conjunto de dados com n objetos, é possível representá-lo por meio de uma matriz quadrada $n \times n$, na qual cada célula expressa o grau de proximidade entre pares de objetos.

No contexto de séries temporais, as medidas de distância podem ser divididas nas seguintes categorias (PAPARRIZOS; YANG; LI, 2024), conforme ilustrado na Figura 4:

- *lock-step*: o mapeamento entre duas séries é realizado ponto a ponto. Nesta categoria destaca-se a distância Euclidiana;
- *elastic*: o mapeamento é realizado de forma "elástica", permitindo um casamento um-para-muitos ou muitos-para-muitos entre pontos das séries, ou seja, um alinha-

mento flexível ao longo do tempo em diferentes regiões. Nesta categoria destaca-se a DTW (***D**ynamic **T**ime **W**arping*);

- *sliding*: o mapeamento é realizado de modo a criar um alinhamento global em relação à forma das séries. Nesta categoria destaca-se a SBD (***S**hape-**b**ased **D**istance*).

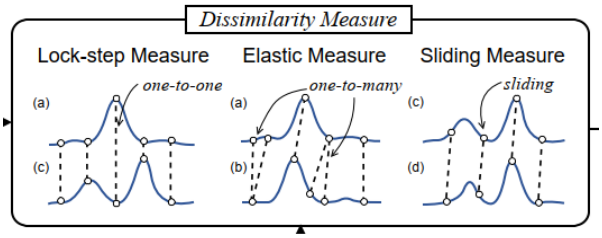

Figura 4 – Visão geral das diferenças entre categorias de medidas de distância no contexto de séries temporais.

Fonte: Adaptado de Paparrizos, Yang and Li (2024).

Antes de discutir os algoritmos de agrupamento em si, é fundamental compreender essas medidas. Ressalta-se que a medida de distância é independente do algoritmo, pois quantifica apenas a diferença entre os objetos, sem definir como eles serão agrupados. A seguir, apresenta-se uma descrição mais detalhada de cada medida. Vale mencionar que as explicações assumem séries previamente normalizadas, de forma que diferenças de *offset* (*translation*) e de amplitude (*scaling*) sejam eliminadas. Ademais, considera-se que todas as séries possuem o mesmo número de pontos.

DE - Distância Euclidiana. A distância Euclidiana entre duas séries $\overrightarrow{x}$ e $\overrightarrow{y}$ é dada pela Equação 2.1, em que m representa o número de pontos presentes nas séries (PAPARRIZOS; GRAVANO, 2016). Para que sejam consideradas idênticas, é necessário que haja um alinhamento perfeito, i.e., o valor de x no instante t deve coincidir com o de y no mesmo instante ao longo de todo o período. Esse alinhamento rígido torna a medida sensível a deslocamentos (*shift*) no tempo e a diferenças de duração. No contexto de séries criminais, duas cidades serão consideradas similares se apresentarem padrões sincronizados mês a mês.

$$DE(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{2.1}$$

A Figura 5 ilustra essa medida, pertencente à categoria *lock-step*. Observa-se que cada valor da série $\overrightarrow{x}$ é comparado diretamente ao valor correspondente da série $\overrightarrow{y}$ no mesmo instante, como indicado pelas conexões verticais tracejadas. Consequentemente, qualquer defasagem temporal ou diferença de duração resulta em aumento significativo da distância, mesmo quando as formas gerais das séries são semelhantes.

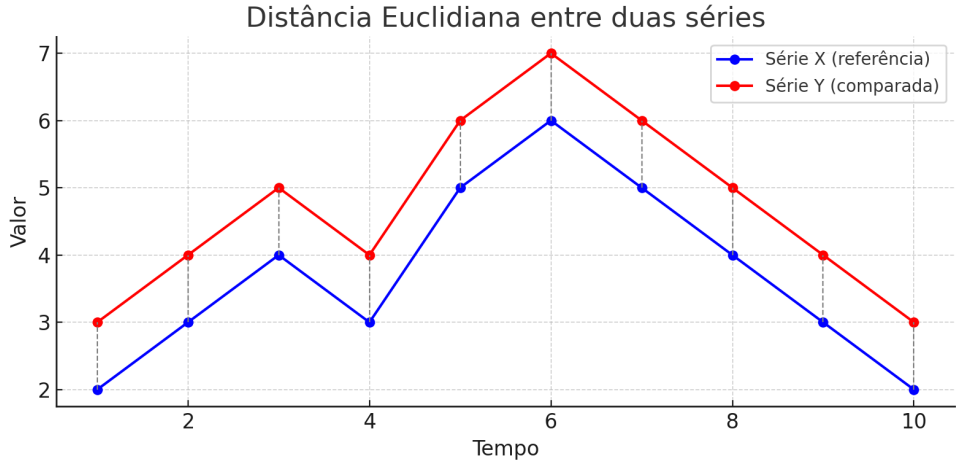

Figura 5 – Exemplo ilustrativo da distância Euclidiana: comparação ponto a ponto entre duas séries temporais.

DTW - Dynamic Time Warping. A DTW é considerada uma extensão da distância Euclidiana que permite um alinhamento local não linear (PAPARRIZOS; GRAVANO, 2016). Para isso, constrói-se uma matriz de custo $M \in \mathbb{R}^{m \times m}$, na qual cada elemento $M(i, j)$ representa a distância Euclidiana entre os pontos x_i e y_j. Um caminho de distorção (*warping path*) $W = \{w_1, w_2, \ldots, w_k\}$, com $k \geq m$, é definido como uma sequência de elementos da matriz M que estabelece um mapeamento entre $\overrightarrow{x}$ e $\overrightarrow{y}$, conforme mostrado na Equação 2.2 (PAPARRIZOS; GRAVANO, 2016). Esse caminho pode ser obtido via programação dinâmica a partir da seguinte recorrência: $\gamma(i, j) = DE(i, j) + min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}$. Em linhas gerais, a DTW "estica" ou "comprime" uma série temporal para que seja alinhada com outra. Assim, torna-se possível comparar séries que apresentam o mesmo padrão, mas com diferentes velocidades ou ritmos. Em outras palavras, diferentemente da distância Euclidiana, que assume um alinhamento ponto a ponto linear no tempo, a DTW permite a criação de um caminho de *warping* W que "mapeia" segmentos das séries de modo que a medida resultante seja robusta a variações na aceleração ou desaceleração dos sinais. Este alinhamento não linear torna a DTW especialmente útil para aplicações onde sequências com padrões similares são deslocadas ou deformadas no tempo, como em reconhecimento de gestos, processamento de fala e análise de sinais médicos.

$$DTW(\overrightarrow{x}, \overrightarrow{y}) = min\sqrt{\sum_{i=1}^{k} w_i} \tag{2.2}$$

A Figura 6 ilustra a aplicação dessa medida, pertencente à categoria *elastic*. Nota-se que a DTW não compara as séries ponto a ponto, como na distância Euclidiana. Em vez disso, ela permite que determinados pontos de uma série sejam alinhados a múltiplos pontos da outra, de modo a minimizar a distância global acumulada. No exemplo, observa-se que

picos e vales de ambas as séries são alinhados mesmo quando ocorrem em instantes de tempo diferentes, o que evidencia a capacidade da DTW de lidar com séries que apresentam o mesmo padrão, mas em ritmos distintos. Essa flexibilidade decorre do *warping path*, representado pelas conexões cinzas, que define a correspondência ótima entre os elementos das duas sequências.

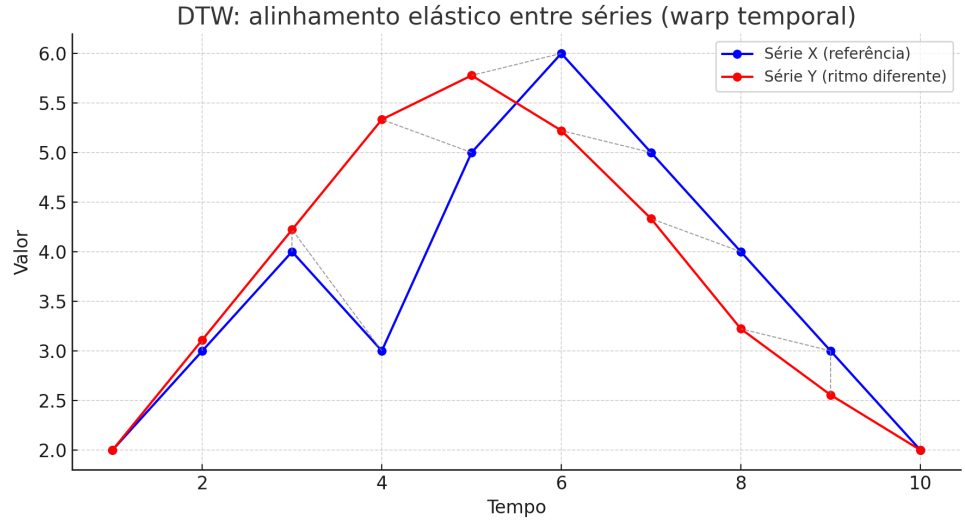

Figura 6 – Exemplo ilustrativo da distância DTW e do caminho de alinhamento (*warping path*).

SBD - Shape-based Distance. A SBD baseia-se na correlação cruzada (*cross-correlation* (CC)), a qual mede a similaridade entre duas sequências $\overrightarrow{x}$ e $\overrightarrow{y}$ mesmo quando não estão alinhadas apropriadamente. Para isso, mantém-se $\overrightarrow{y}$ fixo e "desliza-se" $\overrightarrow{x}$ sobre $\overrightarrow{y}$, computando o produto interno a cada deslocamento s (PAPARRIZOS; GRAVANO, 2016). Considerando todos os deslocamentos possíveis $s \in [-m, m]$, obtém-se $CC_w(\overrightarrow{x}, \overrightarrow{y}) = (c_1, \ldots, c_{2m-1})$, uma sequência de correlações cruzadas, em que $w \in \{1, 2, \ldots, 2m-1\}$ (PAPARRIZOS; GRAVANO, 2016). O objetivo é encontrar o w que maximiza CC_w, isto é, o menor deslocamento necessário para alinhar as sequências. A medida SBD é então definida conforme a Equação 2.3, em que o denominador representa a média geométrica das autocorrelações das séries individuais. Como a CC varia no intervalo $[-1, 1]$, a SBD assume valores em $[0, 2]$, sendo $SBD = 0$ indicativo de casamento perfeito entre as séries. Para detalhes adicionais das equações, vide Paparrizos and Gravano (2016). Em linhas gerais, a SBD visa identificar séries similares tendo como base a forma das mesmas, independentemente de deslocamentos temporais.

$$SBD(\overrightarrow{x}, \overrightarrow{y}) = 1 - \max_{w} \left(\frac{CC_w(\overrightarrow{x}, \overrightarrow{y})}{\sqrt{R_0(\overrightarrow{x}, \overrightarrow{x}) \times R_0(\overrightarrow{y}, \overrightarrow{y})}} \right) \tag{2.3}$$

A Figura 7 ilustra a aplicação dessa medida, pertencente à categoria *sliding*. Nota-se que a SBD não exige que as séries estejam sincronizadas ponto a ponto no tempo. Em

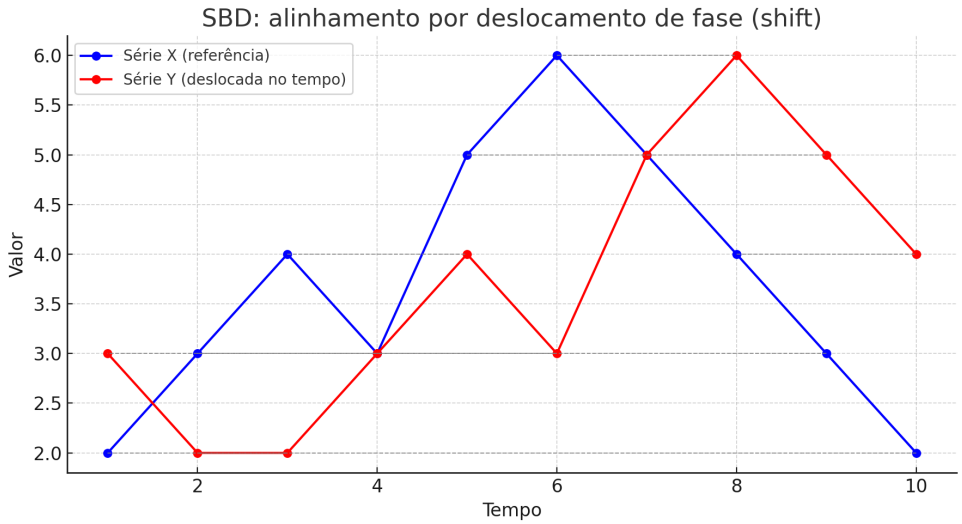

Figura 7 – Exemplo ilustrativo da distância SBD: comparação da forma global entre duas séries temporais por meio de deslocamento de fase (*shift*).

vez disso, permite que uma delas seja deslocada em relação à outra até que se encontre o alinhamento de fase que maximize a correlação cruzada. Assim, mesmo que picos e vales ocorram em instantes diferentes, as séries ainda podem ser consideradas semelhantes caso apresentem a mesma forma global. Essa característica torna a SBD adequada para identificar padrões sazonais ou cíclicos que se repetem em diferentes momentos. No exemplo (Figura 7), a série vermelha foi deslocada em 2 pontos em relação à série azul, e o alinhamento ótimo é obtido quando esse *shift* maximiza a correlação cruzada entre as sequências.

2.3 Algoritmos

Uma vez definida a medida de proximidade (similaridade ou dissimilaridade), é necessária a seleção dos algoritmos de agrupamento a serem utilizados. O objetivo de tais algoritmos é agrupar os objetos do conjunto, neste caso os municípios do estado de SP, em grupos homogêneos, visando minimizar a distância intragrupo e maximizar a distância entre grupos. Paparrizos, Yang and Li (2024) apresentam uma taxonomia de tais algoritmos para séries temporais. Dentre eles encontram-se os particionais, como o k-means, um dos mais conhecidos da área de agrupamento (PAPARRIZOS; YANG; LI, 2024; FORRADELLAS *et al.*, 2020). Assim, este trabalho adotou a seleção de algoritmos particionais, uma vez que os mesmos apresentam menor custo computacional em relação aos hierárquicos, maior facilidade na seleção de hiperparâmetros em relação aos baseados em densidade, além de uma boa interpretabilidade e aderência ao contexto aqui apresentado. Contudo, trabalhos futuros podem incluir outras famílias de algoritmos. Como mencionado, o k-means é um dos algoritmos mais conhecidos da área de agrupamento. Assim, o mesmo é utilizado como base para outros algoritmos. O k-means tem como base a utilização da distância euclidiana.

Contudo, quando tal medida é alterada, assim como alguns outros critérios, têm-se outros algoritmos como o k-DBA, baseado na distância DTW, assim como o k-shape, baseado na distância SBD. Estes três algoritmos, i.e., k-means, k-DBA e k-shape, são amplamente aplicados para agrupamento de séries temporais e, portanto, foram os aqui explorados.

Vale mencionar que cada algoritmo acima mencionado implementa um procedimento iterativo de atribuição e atualização dos grupos, onde as medidas de distância discutidas na seção anterior são usadas para computar a dissimilaridade entre os objetos e os centroides, guiando a formação dos grupos. Assim, nota-se que a escolha da medida influencia diretamente o desempenho e a qualidade dos agrupamentos a serem obtidos.

Diante do exposto, nesta seção uma breve descrição dos algoritmos selecionados será apresentada. A explicação encontra-se fundamentada na estrutura do k-means, o qual será apresentado inicialmente. A partir de seu procedimento, será mostrado como pequenas modificações nas etapas referentes ao cálculo da distância e à atualização dos centroides permitem adaptá-lo para se obter as variantes k-DBA e k-shape. Essa abordagem modular visa evidenciar que as diferentes medidas de distância não alteram o procedimento fundamental do k-means, mas sim sua aplicação particular, permitindo uma compreensão mais clara e organizada das técnicas abordadas, favorecendo sua análise e implementação.

2.3.1 k-means

O objetivo do algoritmo k-means (MACQUEEN, 1967; JAIN, 2010) é particionar os n objetos de um dado conjunto de dados em k grupos, em que cada grupo é sumarizado por um objeto sintético denominado centroide. Assim, pode-se dizer que os centroides representam os valores centrais de cada grupo, facilitando a interpretação e o entendimento dos grupos obtidos. O Algoritmo 1 apresenta o pseudocódigo do k-means. Nota-se que o algoritmo funciona da seguinte maneira:

- k pontos são selecionados de maneira aleatória para representar o "centro" (centroide) de cada um dos k grupos a serem gerados (linha 1);
- em um processo iterativo, calcula-se a **distância euclidiana** de cada ponto a cada um dos centroides e os aloca ao seu grupo mais próximo. Na sequência, os centroides são ajustados tendo como base a **média** dos pontos pertencentes aos seus respectivos grupos. A iteração termina quando os centros não mais se alteram (linhas 2 a 10).

Algoritmo 1: Pseudocódigo do k-means.

Entrada: X = Conjunto de pontos a agrupar; k = número de grupos desejados
Saída: C = Grupos formados; μ = Centroides dos grupos

```
// Etapa 1: Inicializar
```
1 Selecionar k pontos aleatoriamente como centros (centroides) iniciais;

```
// Etapa 2: Repetir até convergir
```
2 **enquanto** *os centroides ainda estão se movendo* **faça**

```
    // 2a: Formar grupos
```
3 **para** *cada ponto* $p \in X$ **faça**
4 Calcular a distância euclidiana de p a cada um dos k centroides;
5 Alocar p ao grupo mais próximo, i.e., aquele ao qual o centroide mais próximo representa;

```
    // 2b: Recalcular centroides
```
6 **para** *cada grupo* **faça**
7 **se** *grupo não está vazio* **então**
8 Centroide = média dos pontos do grupo;

```
    // 2c: Verificar critério de parada
```
9 **se** *os centroides não mudaram* **então**
10 Algoritmo convergiu - Pare;

```
// Passo 3: Finalizar
```
11 **retorna** *Grupos e seus respectivos centroides*;

No contexto deste trabalho, como já mencionado, o objetivo é agrupar séries temporais, mais especificamente, municípios com perfis criminais semelhantes. Quando aplicado neste contexto, o k-means utiliza a distância euclidiana conforme descrito na Seção 2.2.1. Em relação aos centroides, os quais são baseados na média, tem-se como base para cálculo o Algoritmo 2. Nota-se, portanto, que o centroide representa o ponto médio de um grupo de séries temporais, sendo calculado pela média aritmética ponto a ponto das séries que compõem o grupo.

Algoritmo 2: Pseudocódigo do cálculo do centroide do k-means.

Entrada: G = Grupo (conjunto) de séries temporais; n = Comprimento (número de pontos) das séries
Saída: C = Centroide do grupo

1 **para** *i de 1* **até** n **faça**
2 C[i] $\leftarrow \frac{1}{|G|} \sum_{s \in G} s[i]$
3 **retorna** C

2.3.2 k-DBA

Uma das variações do algoritmo k-means, comumente utilizada no contexto de séries temporais, é o k-DBA (PETITJEAN; KETTERLIN; GANCARSKI, 2011; AGHABOZORGI; SHIRKHORSHIDI; WAH, 2015). Em relação ao Algoritmo 2, tem-se como principais modificações a alteração da medida de distância (i.e., linha 4) e do cálculo do centroide dos grupos (i.e., linha 8). No caso do k-DBA a medida de distância utilizada é a DTW, enquanto a média baricêntrica (DBA - DTW *Barycenter Averaging*) é a medida utilizada para o cômputo dos centroides. Assim, quando o k-DBA é utilizado neste contexto, a distância DTW é computada como descrito na Seção 2.2.1. Em relação aos centroides, tem-se como base para cálculo o Algoritmo 3. Nota-se que o centroide representa a média generalizada das séries de um dado grupo considerando a medida DTW.

Algoritmo 3: Pseudocódigo do cálculo do centroide do k-DBA - DBA.

Entrada: G = Grupo (conjunto) de séries temporais; max_{it} = número máximo de iterações

Saída: C = Centroide do grupo

```
// Etapa 1: Inicializar
```

1 Criar centroide_atual via média simples de todas as séries

```
// Etapa 2: Processo iterativo de refinamento
```

2 **para** *cada iteração até convergência ou* max_{it} **faça**

```
    // Etapa 2a: Alinhar cada série com o centroide_atual
```

3 **para** *cada série no grupo* **faça**

4 Calcular alinhamento via DTW entre a série e o centroide_atual

5 Extrair caminho ótimo do alinhamento

```
// Associar os pontos da série aos pontos do centroide
```

6 **para** *cada correspondência (ponto_série, ponto_centroide) no caminho* **faça**

7 Adicionar ponto_série à lista do ponto_centroide correspondente

8 centroide_anterior = centroide_atual

```
// Etapa 2b: Atualizar centroide com base nas associações
```

9 **para** *cada posição no centroide* **faça**

10 **se** *existem valores associados a esta posição* **então**

11 Calcular nova_média dos valores associados

12 Atualizar centroide[posição] = nova_média

```
// Etapa 3: Verificar convergência
```

13 **se** *centroide_atual é igual ao centroide_anterior* **então**

14 C = centroide_atual

15 Parar processo - convergência atingida

16 Retornar C

2.3.3 k-shape

Uma outra variação do algoritmo k-means, comumente utilizada no contexto de séries temporais, é o k-shape (PAPARRIZOS; GRAVANO, 2016). Em relação ao Algoritmo 2, tem-se novamente como principais modificações a alteração da medida de distância (i.e., linha 4) e o cálculo do centroide dos grupos (i.e., linha 8). No caso do k-shape a medida de distância utilizada é a SBD, enquanto um algoritmo de extração de forma é utilizado para o cômputo dos centroides. Assim, quando o k-shape é utilizado neste contexto, a distância SBD é computada como descrito na Seção 2.2.1. Em relação aos centroides, tem-se como base para cálculo o Algoritmo 4. Nota-se que os centroides

são recalculados com base nas séries atribuídas a cada grupo, garantindo que os centroides representem fielmente a forma comum do grupo. A forma de cada centroide é extraída de maneira eficiente visando à preservação das características estruturais das séries.

Algoritmo 4: Pseudocódigo do cálculo do centroide do k-shape - extração de forma.

Entrada: G = Grupo (conjunto) de séries temporais; max_{it} = número máximo de iterações; tol = tolerância de convergência

Saída: C = Centroide do grupo

```
// Etapa 1: Inicializar
```

Selecionar como centroide_inicial uma série aleatória do grupo

```
// Etapa 2: Processo iterativo de refinamento
```

para *cada iteração até tol ou* max_{it} **faça**

- Criar uma matriz de alinhamentos Mat_A vazia

```
// Etapa 2a: Alinhar todas as séries com o centroide atual
```

- **para** *cada série no grupo* **faça**
  - Calcular a correlação cruzada (CC) entre a série e o centroide_atual
  - Encontrar o deslocamento ótimo D_O que maximiza a CC
  - Aplicar o deslocamento ótimo D_O à série
  - Adicionar a série alinhada à matriz de alinhamentos Mat_A
- centroide_anterior = centroide_atual

```
// Etapa 2b: Calcular o novo centroide via extração de forma
```

- Computar a matriz de covariância Mat_{Cov} das séries alinhadas
- Encontrar o autovetor principal A_v da matriz de covariância Mat_{Cov}
- centroide_atual = A_v

```
// Etapa 3: Verificar convergência
```

- Calcular a diferença absoluta dif_A entre o centroide_atual e o centroide_anterior
- **se** $dif_A < tol$ **então**
  - C = centroide_atual
  - Parar processo - convergência atingida

Retornar C

2.3.4 Considerações Finais

Como visto nas Subseções 2.3.1, 2.3.2 e 2.3.3, os algoritmos k-means, k-DBA e k-shape baseiam-se na mesma lógica de particionamento, diferenciando-se em relação às medidas de distância utilizadas e, consequentemente, em como o centroide é computado. A Tabela 1 apresenta um comparativo entre os mesmos. Nota-se, portanto, que a semântica dos agrupamentos muda em função das medidas utilizadas. O k-means limita-se a comparar

valores absolutos ponto a ponto, sendo adequado apenas quando o interesse recai em níveis médios de criminalidade. Já o k-DBA permite capturar padrões temporais semelhantes que ocorrem em velocidades ou momentos diferentes, oferecendo uma visão mais flexível da dinâmica criminal. Por fim, o k-shape acaba por comparar padrões globais de forma, considerando deslocamentos lineares no tempo, o que o torna apropriado para identificar tendências sazonais ou estruturais comuns entre municípios ou períodos distintos. Assim, cada um enfatiza um aspecto distinto: **níveis absolutos** (k-means), **dinâmica local flexível** (k-DBA) e **forma global da evolução** (k-Shape).

Tabela 1 – Comparação entre k-means, k-DBA e k-shape: distância, cálculo do centróide e semântica.

Algoritmo	**Distância usada**	**Atualização do centróide**	**Semântica geral**	**Semântica no contexto criminal**
k-means	Euclidiana	Média aritmética ponto a ponto	Agrupa séries com valores absolutos próximos no mesmo instante; o centróide é o perfil médio das séries.	Identifica municípios com níveis médios semelhantes de crimes (ex.: cidades com o mesmo patamar de furtos/mês). Não lida com atrasos ou picos deslocados.
k-DBA	DTW	DBA (média baricêntrica sob DTW)	Agrupa séries que apresentam a mesma sequência de eventos, mesmo em ritmos diferentes (compressão/expansão local do tempo). O centróide é uma média que respeita esses alinhamentos flexíveis.	Permite encontrar municípios onde a evolução do crime segue a mesma dinâmica, mas em períodos diferentes (ex.: municípios em que o aumento de assaltos ocorre mais cedo ou mais tarde no período analisado).
k-shape	SBD	Autovetor principal da matriz de correlação	Agrupa séries com o mesmo padrão global de subida e descida, considerando apenas deslocamentos lineares no tempo. O centróide representa a tendência de variação predominante.	Agrupa municípios com tendências sazonais ou estruturais semelhantes (ex.: municípios com quedas e picos nos mesmos períodos), mesmo que os volumes absolutos de crimes sejam diferentes.

2.3.5 Avaliação

Após a obtenção de um dado agrupamento, é necessário que se avalie sua qualidade a fim de verificar se os grupos gerados são de fato consistentes. Em outras palavras, se o agrupamento de fato minimiza a distância entre os elementos do grupo e maximiza a distância entre os elementos de grupos distintos. Para tanto, diversos índices (medidas) são encontrados na literatura, os quais são divididos em índices externos e internos (PAPARRIZOS; YANG; LI, 2024). Os índices externos avaliam os grupos os comparando

com rótulos de classe, se apoiando, portanto, em recursos externos. Assim, acabam sendo aplicados em contextos onde um *ground truth* encontra-se disponível, o que não é o caso deste trabalho. Já os índices internos realizam a avaliação considerando a estrutura intrínseca do próprio agrupamento.

Paparrizos, Yang and Li (2024) apresentam alguns índices, tanto externos como internos, dentre os muitos disponíveis, como pode ser visto em Hassan *et al.* (2024). Em relação aos internos, Paparrizos, Yang and Li (2024) destaca quatro deles, a saber: Coeficiente de Silhueta (*Silhouette Coefficient* (SC)), Davies-Bouldin (DB), Dunn Index (DI) e Soma dos Quadrados dentro do Cluster (Within-cluster sum of squares (WCSS)). Este trabalho optou por utilizar o coeficiente de silhueta (SC), uma vez que o mesmo fornece uma avaliação simultânea da coesão e da separação dos grupos, apresentando valores no intervalo $[-1, 1]$ que permitem comparação direta entre diferentes execuções e algoritmos. De acordo com Paparrizos, Yang and Li (2024), o SC é um dos índices internos mais clássicos, sendo de fácil interpretação e, portanto, o adotado neste trabalho.

O SC do agrupamento é computado como a média dos valores de SCs (ou coeficientes de silhueta) de cada ponto do conjunto de dados. O SC s de um dado ponto p, $s(p)$, é dado pela Equação 2.4, em que: $b(p)$ representa a menor distância média entre o ponto p em relação aos pontos contidos em qualquer outro grupo diferente do seu; $a(p)$ é a distância média do ponto p em relação aos outros pontos contidos dentro de seu próprio grupo. Segundo Kaufman and Rousseeuw (2009), os valores do índice podem ser interpretados da seguinte maneira:

- $SC \leq 0.25$: não foi encontrada uma estrutura substancial;
- $0.26 \leq SC \leq 0.50$: a estrutura encontrada é fraca e pode ser artificial;
- $0.51 \leq SC \leq 0.70$: uma estrutura razoável foi encontrada;
- $0.71 \leq SC \leq 1$: foi encontrada uma estrutura forte.

$$s(p) = \frac{b(p) - a(p)}{\max(a(p), b(p))} \tag{2.4}$$

2.4 Trabalhos Relacionados

A fim de visualizar a importância da área de agrupamento em contextos criminais, esta seção descreve brevemente alguns trabalhos relacionados. A seleção dos artigos relacionados não seguiu um protocolo sistemático (e.g., mapeamento sistemático). O objetivo foi, primariamente, exemplificar a relevância da temática abordada.

Em Alves *et al.* (2015) os autores analisam a dinâmica espacial dos homicídios nas cidades brasileiras ao longo de mais de trinta anos utilizando ferramentas da física

estatística para investigar padrões de correlação e a formação de aglomerados. O estudo evidencia que o comprimento de correlação característico aumentou significativamente nas últimas décadas, indicando que o número de homicídios em uma cidade exerce influência crescente sobre localidades vizinhas. Para identificar e quantificar os aglomerados espaciais de municípios, os autores aplicam o algoritmo DBSCAN, adequado para detectar padrões de densidade em dados georreferenciados. Os resultados apontam para a ocorrência de uma transição de percolação, em que grupos de cidades conectadas por altas taxas de homicídios se expandem e se reorganizam ao longo do tempo. Esse comportamento reforça a ideia de que as taxas de criminalidade não são independentes, mas fortemente correlacionadas no espaço, o que pode subsidiar políticas públicas mais eficazes de prevenção e controle da violência, orientando a alocação de recursos com base nas dinâmicas regionais.

Em Junior *et al.* (2020) os autores investigam padrões de concentração espacial de roubos de veículos nos municípios da Grande João Pessoa, empregando técnicas de aprendizado de máquina. O estudo analisa dados criminais registrados entre 2017 e 2019 com o objetivo de subsidiar o planejamento de segurança pública por meio da identificação de áreas críticas de ocorrência. A metodologia adotada também inclui o algoritmo DBSCAN, apropriado para detectar agrupamentos de ocorrências em função de coordenadas geográficas e variáveis temporais. Os resultados evidenciam concentrações significativas de roubos em bairros específicos, com maior incidência durante dias úteis e no período noturno. Esses achados reforçam a utilidade da análise espacial como suporte para estratégias de policiamento direcionadas, permitindo uma alocação mais eficiente dos recursos de segurança.

Em Forradellas *et al.* (2020) os autores desenvolvem um modelo de predição de crimes para a cidade de Buenos Aires utilizando uma rede neural MLP, no contexto da metodologia de mineração de dados SEMMA. O estudo analisa dados criminais de 2016 a 2019 com o intuito de apoiar medidas preventivas e a alocação de recursos em um cenário de aumento das taxas de criminalidade decorrente de crises socioeconômicas. Como etapa preliminar, é aplicado o algoritmo de agrupamento k-means, empregado para segmentar e categorizar os registros criminais em grupos com características semelhantes. Essa categorização é fundamental para identificar padrões latentes no comportamento criminal e, consequentemente, aprimorar a capacidade preditiva do modelo de rede neural.

Em Gusmão, Clemente and Nepomuceno (2022) os autores propõem uma metodologia para otimizar a localização de instalações policiais a partir da combinação de técnicas de análise de agrupamento e do modelo de cobertura máxima (*maximal covering location problem*). Inicialmente, aplica-se o algoritmo k-means, utilizado para agrupar ocorrências criminais com base na proximidade geográfica e na similaridade dos registros, de modo que os centroides representem áreas de maior incidência. Essa etapa é fundamental para identificar padrões espaciais do crime e subsidiar a definição de locais estratégicos para

novas instalações policiais. O artigo também destaca a relevância da escolha adequada do número de grupos (k), uma vez que este parâmetro influencia diretamente a qualidade da cobertura obtida. A análise de sensibilidade realizada mostra que a consideração de um número maior de candidatos à instalação tende a ampliar a cobertura das ocorrências, sobretudo quando a distância euclidiana é utilizada como critério de alocação. Assim, a metodologia integrando k-means e modelagem de cobertura máxima oferece um suporte robusto para o planejamento estratégico em segurança pública.

Em Fontalvo-Herrera, Vega-Hernández and Mejía-Zambrano (2023) os autores aplicam técnicas de agrupamento hierárquico para caracterizar e projetar padrões de crimes violentos na Colômbia, utilizando registros da Polícia Nacional entre 2018 e 2022. O método de agrupamento utilizado é hierárquico, com distância euclidiana e ligação Ward, permitindo identificar quatro grupos distintos de delitos violentos, refletindo diferenças significativas entre os departamentos analisados. Essa etapa constitui o núcleo da análise, pois possibilita entender a distribuição espacial e as características comuns dos crimes. Complementarmente, os autores desenvolveram uma rede neural de dupla camada para prever tendências futuras de violência e apoiar a tomada de decisão estratégica. Os resultados indicam que a integração entre agrupamento hierárquico e rede neural não apenas favorece a análise e o monitoramento da dinâmica criminal, mas também amplia o potencial de antecipação de cenários de risco e de formulação de políticas públicas de segurança.

Em Mission (2024) o autor aplica o algoritmo de agrupamento DBSCAN para analisar incidentes criminais relacionados à “Lei Antiviolência contra Mulheres e seus Filhos de 2004”, no período de 2018 a 2023. O estudo faz uso do software ArcGIS Pro para a visualização dos padrões espaciais e recorre à autocorrelação espacial, por meio do Índice de Moran, a fim de avaliar a significância estatística do agrupamento identificado. Os resultados demonstram que o DBSCAN é eficaz para detectar pontos críticos de violência, revelando áreas de maior vulnerabilidade. Esses achados fornecem subsídios importantes para o planejamento de políticas públicas, orientando estratégias de aplicação da lei e a alocação mais direcionada de recursos em ações de prevenção e combate à violência.

3 ANÁLISE EXPLORATÓRIA

Este capítulo apresenta a análise exploratória realizada, detalhando a metodologia empregada na condução deste trabalho. A análise concentra-se na aplicação e comparação de algoritmos de agrupamento particional em séries temporais criminais. O objetivo é, a partir da discussão dos resultados, obter uma compreensão sobre a viabilidade de se aplicar tais algoritmos no referido contexto avaliando para tanto a qualidade dos agrupamentos gerados a partir dos dados da SSP-SP. Na discussão dos resultados, buscou-se compreender os padrões de criminalidade e a dinâmica temporal dos municípios, visando auxiliar futuras decisões de gestão pública. Para tanto, este capítulo encontra-se estruturado da seguinte maneira: na Seção 3.1 apresenta-se o conjunto de dados utilizado, na Seção 3.2 a configuração experimental adotada e na Seção 3.3 os resultados e discussões.

3.1 Conjunto de Dados

O conjunto de dados encontra-se estruturado em registros mensais que associam os tipos de crimes aos municípios, abrangendo o período de janeiro de 2001 a junho de 2025. Cada registro representa o número de ocorrências (quantidade) de uma natureza criminal (crime) em uma localidade (município) para um determinado período (ano e mês), formando um painel de dados sobre a evolução da criminalidade no estado de SP. O conjunto de dados indexa 645 municípios, os quais são comuns entre os crimes, em relação a 18 crimes distintos (listados no Capítulo 1). Todos os dados utilizados neste trabalho foram obtidos no site da SSP-SP[1].

A Tabela 2 ilustra a estrutura relacional dos dados dos municípios para o período de janeiro de 2001 a junho de 2025 (data de extração dos dados), utilizando como exemplo o crime de “Homicídio Doloso” para demonstrar a associação entre as ocorrências, os municípios e os respectivos períodos. As linhas representam os municípios e as colunas correspondem aos períodos mensais (mês/ano), contendo o número de ocorrências.

Para a condução da análise, cada tipo de crime é avaliado de maneira independente. O foco do agrupamento recai sobre os municípios, que são os objetos a serem agrupados. O histórico de ocorrências de um crime para cada município é tratado como uma série temporal, uma vez que consiste em uma sequência de observações coletadas em intervalos de tempo regulares. Esta abordagem é fundamental, pois permite que os algoritmos de agrupamento analisem a dinâmica do comportamento criminal ao longo do tempo, possibilitando a identificação de similaridades no comportamento histórico entre os municípios, algo que não seria capturado por meio de estatísticas agregadas.

[1] <https://www.ssp.sp.gov.br/estatistica>.

Tabela 2 – Casos de “Homicídio Doloso” no período de Janeiro de 2001 a Junho de 2025.

HOMÍCIDIO DOLOSO	Jan-01	Feb-01	Mar-01	Apr-01	...	Jan-25	Jun-25
AGUAÍ	1	0	2	3	...	2	5
ÁGUAS DA PRATA	0	0	0	5	...	1	0
ÁGUAS DE SÃO PEDRO	0	0	0	1	...	1	0
AMERICANA	2	1	1	1	...	105	104
ANALÂNDIA	0	0	0	0	...	0	0
ARARAS	0	2	0	0	...	28	37
...	...	...	...	...	...	...	...
ZACARIAS	1	0	3	1	...	9	7

3.2 Configuração Experimental

Esta seção descreve a configuração experimental adotada neste trabalho. O objetivo é fornecer a transparência necessária para garantir a reprodutibilidade dos experimentos, assim como a validade dos resultados aqui discutidos. Para tanto, a seção encontra-se estruturada da seguinte maneira: na Seção 3.2.1 discute-se sobre o ambiente computacional e bibliotecas utilizadas; na Seção 3.2.2 sobre a preparação dos dados e o pré-processamento realizado; na Seção 3.2.3 sobre a parametrização dos algoritmos utilizados; na Seção 3.2.4 sobre aspectos de avaliação.

3.2.1 Ambiente Computacional e Bibliotecas

Os experimentos foram desenvolvidos na linguagem de programação Python (versão 3.13), utilizando um conjunto de bibliotecas consolidadas no campo da ciência de dados. A biblioteca central foi a *tslearn*[2], que fornece implementações dos algoritmos de agrupamento de séries temporais aqui utilizados, i.e., *TimeSeriesKMeans* (k-means e k-DBA) e k-shape, bem como o índice de coeficiente de silhueta (*silhouette score*).

Em relação à manipulação de dados — carregamento, filtragem, etc. — empregou-se a biblioteca *pandas*[3]. Para os cômputos numéricos, utilizou-se a biblioteca *numpy*[4]. Por fim, em relação aos gráficos e visualizações, para análise dos resultados, foi utilizado as bibliotecas *matplotlib*[5] e *seaborn*[6]. Vale mencionar que os experimentos foram conduzidos em um computador pessoal, refletindo um ambiente de estudo e prototipagem.

[2] <https://tslearn.readthedocs.io>.
[3] <https://pandas.pydata.org>.
[4] <https://numpy.org>.
[5] <https://matplotlib.org>.
[6] <https://seaborn.pydata.org>.

3.2.2 Preparação e Pré-processamento dos Dados

A condução experimental deste trabalho iniciou-se com a separação dos dados por tipo de criminalidade, adotando uma abordagem univariada. Esta divisão foi utilizada uma vez que o padrão de ocorrências de cada delito define um comportamento temporal distinto. Portanto, o foco de cada experimento consistiu em comparar as séries temporais dos diferentes municípios para identificar padrões de similaridade exclusivamente dentro do contexto de um único crime por vez.

Para a realização dos experimentos, foram selecionadas todas as séries temporais correspondentes a cada tipo de crime, abrangendo os 645 municípios do estado de São Paulo. Assim, o escopo da análise contemplou todos os municípios distribuídos entre as 18 naturezas criminais.

Embora os dados completos estivessem disponíveis de janeiro de 2001 a junho de 2025, a análise concentrou-se nos seguintes intervalos temporais: de 2020 a 2024 (5 anos), de 2022 a 2024 (3 anos), de 2023 a 2024 (2 anos) e, por último, o ano completo de 2024 (1 ano). O ano de 2025 foi excluído da análise devido à indisponibilidade de dados completos para o período. Essa escolha visou não apenas capturar as dinâmicas criminais mais recentes, mas também garantir a eficiência computacional dos experimentos, especialmente considerando a complexidade da medida DTW empregada pelo k-DBA. Ademais, a extensão temporal da série influencia diretamente a capacidade dos algoritmos de agrupamento de capturar características relevantes e diferenciadoras dos padrões criminais ao longo do tempo. Embora séries mais longas possam propiciar uma visão mais abrangente das dinâmicas, elas também aumentam a complexidade computacional e podem diluir efeitos pontuais recentes. Por outro lado, séries muito curtas podem não conter informações suficientes para identificar padrões robustos, prejudicando a qualidade dos agrupamentos (JAVED; LEE; RIZZO, 2020, p. 7). Portanto, explorar múltiplos intervalos possibilita equilibrar a riqueza da informação temporal com a eficiência do agrupamento. Essa estratégia está alinhada às melhores práticas em análise temporal, em que a escolha do período de análise deve equilibrar granularidade e relevância temporal dos dados para a aplicação em foco (PAPARRIZOS; GRAVANO, 2016, p. 69).

3.2.2.1 Filtragem de Séries Constantes (Nulas)

Visando assegurar que os índices de avaliação (especificamente o coeficiente de silhueta) reflitam de fato a qualidade estrutural dos agrupamentos, optou-se por excluir antes da etapa de agrupamento as séries totalmente nulas (sem ocorrências). Essa decisão foi motivada por dois argumentos complementares, a saber:

- sequências idênticas (neste caso, vetores com todas as posições zeradas) podem conduzir a um valor médio de silhueta artificialmente alto quando agrupadas,

levando a uma interpretação enganosa da separação e coesão dos *clusters* — logo, sua inclusão pode inflacionar o valor do índice e mascarar problemas reais de separabilidade (ROUSSEEUW, 1987);

- do ponto de vista prático e semântico, séries totalmente zeradas representam municípios sem ocorrência em um determinado tipo de crime no período analisado; agrupar esses municípios entre si não exige metodologia de mineração de séries temporais (a informação é essencialmente "ausência de evento") e tende a reduzir a utilidade interpretativa dos *clusters* para o objetivo do estudo.

Do ponto de vista algorítmico, a filtragem das séries temporais nulas é fundamental para assegurar a consistência numérica e a validade dos agrupamentos. Em algoritmos como o k-shape, cada série é submetida à normalização z-score, processo que requer variância diferente de zero para o cálculo da similaridade de forma; assim, séries compostas inteiramente por zeros tornam-se inviáveis para a aplicação direta do método (PAPARRIZOS; GRAVANO, 2016). Além disso, conforme discutido por Kaufman and Rousseeuw (2009), objetos que não apresentam variação entre as observações devem ser removidos antes da padronização, uma vez que não contribuem para a diferenciação entre grupos e podem gerar distorções nas medidas de distância. A Tabela 3 apresenta a quantidade de séries não nulas, ou seja, séries restantes após a remoção das séries zeradas, por tipo de crime, que foram consideradas durante o agrupamento em cada janela temporal.

Tabela 3 – Quantidade de séries temporais por tipo de crime e período.

Tipo de Crime	**5 anos**	**3 anos**	**2 anos**	**1 ano**
Estupro	590	554	515	420
Homicídio Doloso (2)	554	504	483	381
Lesão Corporal seguida de Morte	159	125	98	74
Estupro Vulnerável	637	631	618	580
Homicídio Doloso por Acidente Transito	39	25	12	8
Roubo Outros	608	582	549	499
Furto Outros	645	645	645	645
Latrocínio	210	155	115	69
Roubo Banco	27	13	6	2
Furto de Veículo	635	619	601	558
Lesão Corp. Culposa - Outras	560	506	454	356
Roubo de Carga	357	295	252	197
Homicídio Culposo Outros	223	177	149	83
Lesão Corp. Culposa Acidente	645	645	640	630
Roubo de Veículo	509	482	432	368
Homicídio Culp. Acidente Trânsito	618	589	564	495
Lesão Corporal Dolosa	645	645	645	645
Tentativa de Homicídio	604	576	538	456

Nota: todos os períodos têm final em 2024.

Para operacionalizar a filtragem (exclusão) acima mencionada, foi aplicado um procedimento de filtragem individual para cada combinação de crime e janela temporal. Consequentemente, o número de municípios (N) submetidos ao agrupamento variou em cada um dos experimentos, refletindo os diferentes padrões de incidência criminal no estado. A seleção de distintas janelas temporais (períodos de 1, 2, 3 e 5 anos) foi um fator determinante para esta variação. Por exemplo, uma série temporal com ocorrências nos primeiros meses de um biênio, mas totalmente nula no segundo ano, seria incluída na análise de 24 meses, mas seria descartada na análise focada apenas nos últimos 12 meses. Essa abordagem metodológica assegurou que os algoritmos de agrupamento operassem exclusivamente sobre conjuntos de dados onde houvesse variabilidade de ocorrências, focando a análise em padrões de comportamento criminal genuínos e contribuindo para a obtenção de resultados mais interpretáveis.

A fim de demonstrar o efeito (impacto) da filtragem de séries temporais nulas, a Tabela 4 compara os resultados do coeficiente de silhueta (SC) para três crimes com características distintas, evidenciando o efeito da remoção de séries nulas. A tabela contrasta os scores máximos obtidos ao se agrupar todos os municípios (sem remoção de séries ("COM S.N.")) em relação ao agrupamento com os municípios filtrados (com remoção de séries ("SEM S.N.")). A principal constatação é que a exclusão das séries nulas diminui os scores para os algoritmos k-means e k-DBA, i.e., nota-se que a presença de séries zeradas gera agrupamentos artificiais, pois municípios sem registros de ocorrência acabam sendo considerados semelhantes, inflando o coeficiente de silhueta de modo não informativo. Ao remover essas séries nulas, evita-se a formação de *clusters* triviais. Em relação ao k-shape, nota-se um comportamento inverso para o crime "Roubo a Banco". O k-shape baseia-se na SBD, a qual, por ser invariante à escala e à translação, necessita que cada série seja padronizada via z-score. No caso de crimes de extrema esparsidade, como "Roubo a Banco", onde a vasta maioria dos 645 municípios não registra ocorrências, a inclusão dessas séries gera instabilidade numérica no processo de padronização, o que inviabiliza o cálculo do SC. A remoção das séries nulas (cenário "SEM S.N.") contorna essa instabilidade. Embora o SC resultante sobre o pequeno número de séries ativas ainda possa ser baixo, ele é calculável e, portanto, é registrado como superior ao resultado do cenário "COM S.N.", onde o cálculo falha ou colapsa devido à incompatibilidade da padronização com os zeros estruturais (PAPARRIZOS; GRAVANO, 2016, p. 71 - 72).

Tabela 4 – Comparação do SC com e sem séries nulas por algoritmo e tipos de crimes nos períodos de 5 e 2 anos.

Per. 5 Anos	**Estupro**		**Homicídio C. Outros**		**Roubo a Banco**	
2020 - 2024	**COM S.N.**	**SEM S.N.**	**COM S.N.**	**SEM S.N.**	**COM S.N.**	**SEM S.N.**
Qtd. Séries	645	590	645	224	645	27
Algoritmo	**SC**	**SC**	**SC**	**SC**	**SC**	**SC**
k-means	0.4462	0.4226	0.9406	0.8619	0.9884	0.7919
k-DBA	0.4406	0.4358	0.9385	0.8971	0.9984	0.9630
k-shape	0.1597	0.1078	-	-	-	*0.4292

Per. 2 Anos	**Estupro**		**Homicídio C. Outros**		**Roubo a Banco**	
2023 - 2024	**COM S.N.**	**SEM S.N.**	**COM S.N.**	**SEM S.N.**	**COM S.N.**	**SEM S.N.**
Qtd. Séries	645	515	645	149	645	6
Algoritmo	**SC**	**SC**	**SC**	**SC**	**SC**	**SC**
k-means	0.7012	0.6625	0.9427	0.8124	0.9926	0.3909
k-DBA	0.7350	0.7208	0.9810	0.9227	0.9984	0.8333
k-shape	-	-	-	-	-	*0.0052

Nota: S.N.: séries nulas; *padrão invertido.

3.2.2.2 Normalização das Séries Temporais

A etapa de escalonamento (*feature scaling*) transforma as variáveis do domínio de modo que suas escalas se tornem comparáveis, constituindo um requisito fundamental para assegurar que todas as séries temporais sejam avaliadas de forma proporcional. Esse procedimento é especialmente crítico em algoritmos baseados em medidas de distância, pois diferenças de ordem de grandeza entre variáveis podem distorcer os resultados e comprometer as interpretações.

Para este trabalho, empregou-se inicialmente o método *TimeSeriesScalerMeanVariance* da biblioteca *tslearn*, o qual aplica padronização z-score, ajustando cada série temporal para média zero e desvio padrão um. Essa padronização foi aplicada de maneira uniforme a todos os dados antes da execução dos modelos, garantindo que as medidas de distância euclidiana, DTW e SBD operassem sob as mesmas condições de escala, proporcionando uma base metodologicamente justa para a comparação dos resultados. A Figura 8 apresenta um exemplo de um conjunto de séries antes e depois da padronização via z-score considerando 6 municípios no intervalo temporal de 10 meses. O z-score é definido pela expressão $z = \frac{x-\mu}{\sigma}$, em que x representa o número de ocorrências em um dado instante de tempo t, μ a média de todas as ocorrências na série temporal ao longo do período considerado e σ corresponde ao desvio padrão.

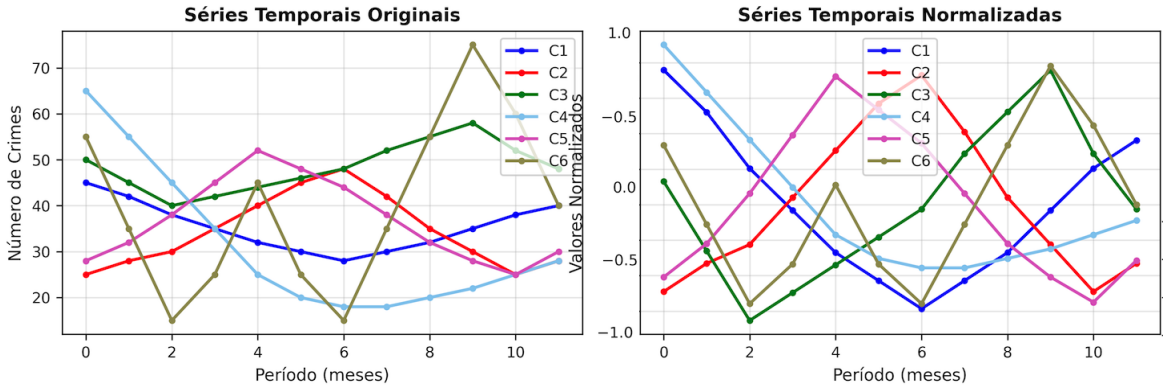

Figura 8 – Exemplo comparativo entre um conjunto de séries não normalizadas e normalizadas via z-score.

Fonte: Elaborado pelo autor.

Contudo, durante os experimentos iniciais, observou-se que a padronização z-score apresentou limitações que impactaram negativamente o desempenho dos algoritmos de agrupamento, manifestando-se através de baixos scores de silhueta (SC). Este comportamento pode ser atribuído à sensibilidade da padronização z-score à presença de valores extremos (*outliers*) nas séries temporais de criminalidade, uma vez que esta técnica utiliza a média e o desvio padrão para a transformação dos dados (ROUSSEEUW, 1987).

No estudo de Hassan *et al.* (2024), que apresenta uma revisão abrangente dos principais índices tradicionais e recentemente propostos para avaliação de agrupamentos, é enfatizado que os índices internos de validação, tais como o índice de silhueta, são amplamente utilizados para medir a coesão dentro do grupo e a separação entre os grupos sem a necessidade de informações externas. Entretanto, é reconhecido que esses índices podem apresentar dificuldades na presença de ruídos, *outliers* e formas arbitrárias dos *clusters*, o que impacta negativamente na avaliação da qualidade dos agrupamentos (HASSAN *et al.*, 2024, p. 4-10).

Diante do exposto, a escolha de métodos que sejam menos sensíveis a esses aspectos do conjunto de dados é recomendada para garantir a obtenção de resultados mais robustos e representativos. Tal recomendação está alinhada com a substituição do método de padronização via z-score pelo método de escalonamento robusto *RobustScaler* da biblioteca *scikit-learn*[7], que utiliza mediana e intervalo interquartil a fim de reduzir o impacto de valores extremos e promover uma transformação mais adequada a séries temporais com características atípicas (HASSAN *et al.*, 2024, p. 5-9).

O *RobustScaler*, definido pela expressão $x' = \frac{x - \text{mediana}(X)}{Q_3(X) - Q_1(X)}$, é um método de escalonamento que centraliza os dados pela mediana e os divide pelo intervalo interquartil (IQR), o qual é definido como a diferença entre o terceiro quartil (Q_3) e o primeiro quartil

[7] <https://scikit-learn.org>.

(Q_1) do conjunto de dados. O método garante que os dados sejam centrados e ajustados em uma escala menos sensível a valores extremos, o que é desejável para algoritmos de agrupamento que podem ser fortemente influenciados por *outliers*. Assim, o *RobustScaler* se apresenta como uma escolha mais adequada para séries temporais e dados com distribuições assimétricas, promovendo maior robustez na formação e avaliação dos *clusters*.

Embora Hassan *et al.* (2024) reconheça que métodos de pré-processamento robusto, como o *RobustScaler*, têm se mostrado uma alternativa efetiva ao z-score para diversos algoritmos de agrupamento, especialmente em dados com *outliers*, sua aplicação no k-shape não é viável devido às características matemáticas específicas deste algoritmo. O k-shape utiliza a correlação cruzada normalizada como medida de similaridade, a qual foi projetada para capturar padrões de forma em séries temporais enquanto garante invariância à escala e deslocamento (PAPARRIZOS; GRAVANO, 2016). Assim, para que essa medida funcione corretamente, é necessário que cada série temporal tenha média zero e desvio padrão um, propriedades que são matematicamente garantidas apenas pela padronização z-score. O *RobustScaler* não assegura tais condições, conforme discutido por Hassan *et al.* (2024) em sua revisão. Assim, neste trabalho utilizou-se o *RobustScaler* para normalizar as séries quando as mesmas foram agrupadas via k-means e k-DBA e z-score para aplicação do k-shape (já automático).

3.2.3 Parametrização dos Algoritmos de Agrupamento

A fim de garantir a consistência e a reprodutibilidade dos experimentos, os algoritmos de agrupamento foram instanciados com um conjunto fixo de hiperparâmetros, exceto pelo número de *clusters* (k), que foi variado com o objetivo de identificar o valor mais adequado em cada caso. Os hiperparâmetros que foram fixados foram:

- $random\_state = 42$: este hiperparâmetro funciona como uma semente para os processos de inicialização aleatória dos centroides. Ao fixá-lo, garante-se que os resultados sejam determinísticos e totalmente reprodutíveis. O valor 42 é uma convenção frequentemente adotada na comunidade de ciência de dados para este fim;
- $max\_iter = 200$: número máximo de iterações do algoritmo. Este hiperparâmetro atua como uma condição de parada para garantir que o processo termine, mesmo que a convergência natural não seja atingida dentro do limite especificado (hiperparâmetro *tol*), evitando execuções infinitas;
- $n\_init = 10$: define o número de vezes que o algoritmo é executado com diferentes inicializações de centroides. O resultado final retornado é sempre o da melhor execução em termos de inércia. Esta prática reduz significativamente o risco de convergência para um "ótimo local", aumentando a robustez do agrupamento final. O valor 10 é um padrão que oferece um bom balanço entre qualidade e custo computacional;

- $tol = 1e{-}4$: o hiperparâmetro de tolerância, mantido em seu valor padrão, interrompe o treinamento quando a melhora na inércia entre duas iterações consecutivas se torna insignificante. Manter o valor padrão, que é muito baixo, é ideal para este trabalho, pois o objetivo é que o algoritmo pare preferencialmente por atingir a convergência completa, e não por um limiar de tolerância permissivo.

Já no que se refere ao número de *clusters* (k), sabe-se que sua definição é fundamental para a construção do agrupamento em si. Para este estudo, foi definida uma faixa de valores de k entre 2 e 15, visando um equilíbrio entre a busca por padrões significativos e as limitações computacionais inerentes ao processamento de 645 séries temporais. A própria literatura enfatiza que não existe um método único ou definitivo para a definição do k ideal. A própria avaliação dos agrupamentos obtidos é um processo complexo, pois a noção de um "bom" *cluster* é de certo modo subjetiva e depende diretamente dos objetivos da aplicação em questão (TAN *et al.*, 2018). Assim sendo, o intervalo de valores de k definidos foi aplicado a cada um dos algoritmos.

A definição de um número pequeno de *clusters* (baixos valores de k) visa à obtenção de agrupamentos mais generalizáveis, no sentido de oferecer uma representação mais interpretável e menos sensível a variações locais. Conforme destacam Tan *et al.* (2018, p.308), um dos objetivos primordiais do agrupamento é a "compreensão dos dados", o que é facilitado pela identificação de um número pequeno de grupos que sumarizam a estrutura geral do conjunto de dados. Assim, um k reduzido favorece a interpretabilidade, oferecendo uma visão geral dos padrões de criminalidade entre os municípios. Por outro lado, a definição de um número maior de *clusters* (altos valores de k) permite uma análise com maior granularidade. Esta abordagem é essencial para a descoberta de subgrupos e padrões mais específicos que poderiam ser mascarados em um agrupamento com menor número de grupos. Assim, ao aumentar o valor de k, aumenta-se a sensibilidade do algoritmo para detectar nuances e estruturas mais finas nos dados, revelando padrões de comportamento criminal que podem ser exclusivos de um pequeno conjunto de municípios.

Diante do exposto, a metodologia adotada neste trabalho, ao avaliar uma faixa de valores para k em vez de um único valor, permite uma análise multinível mais rica. Esta abordagem é duplamente vantajosa: primeiramente, possibilita a captura tanto das estruturas de dados mais amplas (com baixos valores de k) quanto dos padrões mais sutis e detalhados (com altos valores de k). Em segundo lugar, ao limitar esta exploração a uma faixa computacionalmente viável (de 2 a 15), a análise se torna prática sem sacrificar a profundidade. Essa abordagem exploratória é fundamental em um estudo de *benchmark* como o proposto, onde o comportamento dos algoritmos (k-means, k-DBA e k-shape) pode variar consideravelmente em diferentes níveis de resolução.

3.2.4 Aspectos de Avaliação

Como já mencionado na Subseção 2.3.5, o coeficiente de silhueta (SC), definido na Equação 2.4, foi o índice interno escolhido para se avaliar os agrupamentos gerados. Assim, para cada crime, os três algoritmos de agrupamento selecionados (k-means, k-DBA e k-shape) foram executados no intervalo de k clusters (de 2 a 15); portanto, para cada crime executou-se 14 vezes cada um dos 3 algoritmos em cada intervalo temporal (14 * 3 = 42 experimentos por crime/intervalo temporal). Ao final de cada experimento, o coeficiente de silhueta foi computado e armazenado, permitindo, ao final, identificar o número ótimo de grupos (k) em cada algoritmo para cada tipo de crime/intervalo temporal. A Tabela 5 apresenta, para cinco tipos de crimes no período de 5 anos, como os resultados finais foram tabulados visando consolidar e comparar os resultados obtidos pelos diferentes algoritmos de agrupamento. Vale mencionar uma limitação específica referente a avaliação do algoritmo k-shape. A função *silhouette score* da biblioteca *tslearn*, utilizada para medir a qualidade dos agrupamentos, não oferece nativamente a medida SBD, inviabilizando, a princípio, a avaliação via SC do algoritmo k-shape. Para contornar essa restrição da biblioteca, a avaliação dos agrupamentos gerados via SC foi conduzida utilizando uma implementação própria do SBD, porém extraída da implementação do k-shape da própria biblioteca (opção "*precomputed*").

Tabela 5 – Padrão de tabulação adotado para avaliação dos resultados obtidos - exemplo considerando apenas cinco crimes no período de 5 anos.

Crime	**Algoritmo**	**Melhor K**	**Melhor SC**	**SC Médio**
FURTO - OUTROS	k-means	2	0.8391	0.0831
FURTO - OUTROS	k-shape	3	0.0375	0.0074
FURTO - OUTROS	k-DBA	2	0.8986	0.0974
FURTO DE VEÍCULO	k-means	2	0.2502	0.1490
FURTO DE VEÍCULO	k-shape	2	0.0746	0.0240
FURTO DE VEÍCULO	k-DBA	2	0.7193	0.1954
ROUBO A BANCO	k-means	2	0.7920	0.2025
ROUBO A BANCO	k-shape	2	0.4292	-0.1940
ROUBO A BANCO	k-DBA	4	0.9630	0.9484
ROUBO DE CARGA	k-means	2	0.6794	0.3714
ROUBO DE CARGA	k-shape	2	0.1953	0.1486
ROUBO DE CARGA	k-DBA	2	0.7374	0.3808
ROUBO DE VEÍCULO	k-means	2	0.7381	0.3188
ROUBO DE VEÍCULO	k-shape	2	0.1299	0.0882
ROUBO DE VEÍCULO	k-DBA	2	0.7451	0.3079

A relevância do uso do coeficiente de silhueta nesta monografia reside em sua capacidade de avaliar objetivamente a estrutura dos agrupamentos obtidos; assim, torna-se fundamental para determinar o número ótimo de *clusters* e comparar diferentes algoritmos

de agrupamento de maneira quantitativa e neutra. Assim, a fim de facilitar a interpretação dos agrupamentos obtidos, assim como a análise comparativa entre os algoritmos, utilizou-se uma categorização qualitativa dos resultados visando classificar a força da estrutura do agrupamento encontrado, conforme intervalos apresentados na Subseção 2.3.5. Considerando que scores mais altos indicam *clusters* mais coesos e bem separados, a categorização divide os resultados em quatro grupos distintos, a saber:

- Grupo A: compreende os resultados (agrupamentos) que obtiveram SC no intervalo $0.71 \leq SC \leq 1$, indicando a descoberta de uma estrutura de *cluster* forte e bem definida;
- Grupo B: compreende os resultados (agrupamentos) que obtiveram SC no intervalo $0.51 \leq SC \leq 0.70$, indicando uma estrutura de agrupamento razoável e claramente identificada;
- Grupo C: compreende os resultados (agrupamentos) que obtiveram SC no intervalo $0.26 \leq SC \leq 0.50$, indicando uma estrutura fraca, mas potencialmente significativa, que merece investigação;
- Grupo D: compreende os resultados (agrupamentos) que obtiveram SC no intervalo $SC \leq 0.25$, indicando a ausência de uma estrutura de agrupamento substancial, onde os *clusters* não são diferenciáveis. Scores negativos, por sua vez, sugerem que os objetos podem ter sido atribuídos a *clusters* incorretos.

Essa categorização é adotada apenas neste estudo, de forma a padronizar a interpretação dos valores de SC, e não corresponde a uma escala universal. Tal categorização permite uma avaliação sumária e consistente do desempenho dos algoritmos em diferentes cenários. A categorização foi utilizada ao longo da seção de resultados para referenciar de maneira concisa a qualidade dos agrupamentos encontrados, permitindo uma rápida comparação entre os algoritmos, os crimes e os diferentes períodos temporais analisados. A Tabela 6 apresenta, para o crime de “Estupro”, como os resultados finais foram tabulados visando facilitar o entendimento, como já mencionado, da qualidade dos agrupamentos obtidos.

Adicionalmente à avaliação via SC, empregou-se também, em algumas análises, visualização. Cada gráfico gerado, como os exemplificados na Figura 9, exibem os resultados de um dado agrupamento, em que cada subgráfico corresponde a um *cluster* do agrupamento para um dado valor de k. O centroide de cada grupo é identificado por meio da linha em vermelho. As demais linhas em preto correspondem às séries individuais de cada município que encontram-se no *cluster*, o eixo y delimita os valores mínimo e máximo de ocorrência das séries no grupo, enquanto o eixo x o intervalo temporal considerado. A inspeção visual

Tabela 6 – Padrão de tabulação adotado, após categorização por grupo, para avaliação dos resultados obtidos - exemplo considerando o crime de “estupro”.

Crime de “ESTUPRO”			
Algoritmo	**Melhor k**	**Melhor SC**	**Grupo**
Qtd. Séries: 690	2020 a 2024 - (5 ANOS)		
k-means	2	0.4226	Grupo C
k-shape	2	0.1078	Grupo D
k-DBA	2	0.4358	Grupo C
Qtd. Séries: 554	2022 a 2024 - (3 ANOS)		
k-means	2	0.5367	Grupo B
k-shape	2	0.1043	Grupo D
k-DBA	2	0.7226	Grupo A
Qtd. Séries: 515	2023 a 2024 - (2 ANOS)		
k-means	2	0.6625	Grupo B
k-shape	2	0.1388	Grupo D
k-DBA	2	0.7208	Grupo A
Qtd. Séries: 420	2024 a 2024 - (1 ANO)		
k-means	2	0.6656	Grupo B
k-shape	2	0.1386	Grupo D
k-DBA	3	0.6544	Grupo B

Qtd. Séries: quantidade de séries não nulas que foram agrupadas.

destes gráficos é relevante, uma vez que permite a interpretação do padrão definido em cada *cluster*.

Por fim, visando possibilitar uma análise transversal e aprofundada dos resultados, este trabalho gerou o que denominamos aqui de “matriz de perfis de crime” ou “matriz de perfis criminais”. Trata-se de uma matriz 3x3 projetada para categorizar cada um dos 18 tipos de crimes em duas dimensões que buscam descrever suas propriedades intrínsecas, a saber: volume, referente à intensidade típica do crime, e prevalência, referente ao espalhamento geográfico do mesmo. O objetivo é tentar caracterizar a natureza de cada fenômeno criminal de forma objetiva, com base nos dados. Posteriormente, tal caracterização é utilizada para investigar como as diferentes propriedades dos crimes influenciam o desempenho dos algoritmos de agrupamento.

A construção da matriz se baseia no cálculo de dois aspectos para cada crime, utilizando-se do período de cinco anos (2020-2024) como base temporal. A prevalência é calculada como o percentual de municípios que registraram ao menos uma ocorrência no período. O volume, por sua vez, é representado pela mediana do total de ocorrências em cinco anos, calculada exclusivamente sobre o subconjunto de municípios ativos (séries não nulas). A escolha da mediana se deu, em detrimento da média, para mitigar o efeito de *outliers*, uma vez que se baseia na posição central dos dados e não na magnitude de

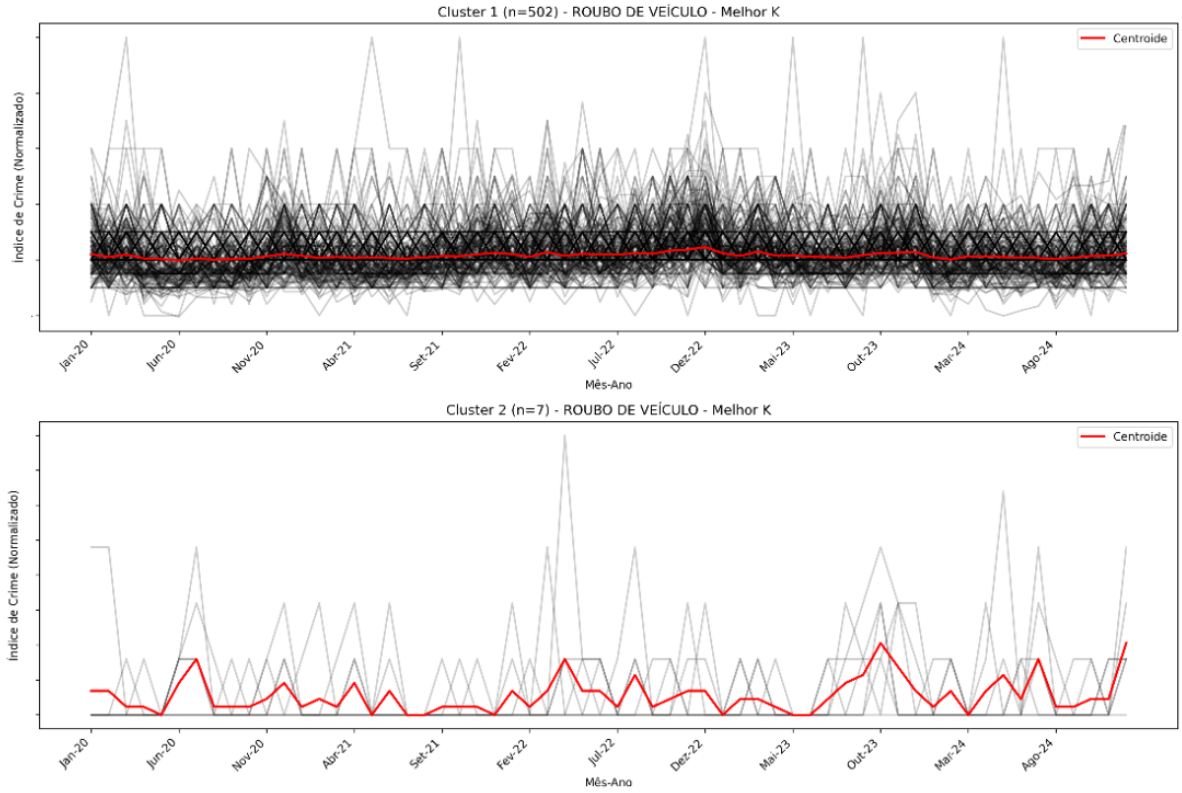

Figura 9 – Gráfico gerado a partir da execução do k-means para k=2 (melhor k) para o crime "Roubo de veículo" no período de 5 anos.

Fonte: Elaborado pelo autor.

todos os valores, não sendo enviesada pelo volume de ocorrências de um único e mesmo município.

Na Tabela 7 encontra-se a matriz de perfis proposta, a qual é composta por nove quadrantes. As colunas referem-se aos valores do aspecto volume e as linhas aos valores do aspecto prevalência. Nota-se que cada um dos aspectos é, portanto, dividido em três níveis: baixo, médio e alto. O cruzamento das linhas e das colunas fornecem os quadrantes que representam um perfil em comum entre o volume e a prevalência dos crimes. Assim, é possível categorizar os crimes por meio de "propriedades similares", como, por exemplo, "Raros e Específicos" (baixo volume, baixa prevalência) ou "Endêmicos" (alto volume, alta prevalência). Nota-se, portanto, que a matriz de perfis de crime atua como um recurso analítico complementar, que pode contribuir para um aprofundamento da análise ao cruzar diferentes resultados. Ela permite relacionar os perfis dos crimes com os grupos obtidos via SC, possibilitando, por exemplo, a identificação de alguma relação entre um determinado perfil e o coeficiente de silhueta. Deste modo, a matriz auxilia na geração de *insights* mais aprofundados sobre o agrupamento de diferentes tipos de fenômenos criminais.

Tabela 7 – Matriz de Perfis de Crime por Volume e Prevalência.

	Baixo Volume (BV) *(Mediana Baixa)*	**Médio Volume (MV)** *(Mediana Média)*	**Alto Volume (AV)** *(Mediana Alta)*
Alta Prevalência (AP) *(Crime espalhado)*	**Perfil: Crimes Crônicos de Baixa Intensidade** Ocorrências frequentes em quase todo o estado, mas com baixo volume em cada local.	**Perfil: Crimes Dispersos e de Atenção** Problema generalizado, afetando muitas cidades com volume considerável.	**Perfil: Crimes Endêmicos** O problema mais desafiador: espalhado por todo o estado e com alto volume.
Média Prevalência (MP) *(Crime regional)*	**Perfil: Crimes Focais de Baixa Intensidade** Ocorre em número significativo de municípios, mas sem grande impacto individual.	**Perfil: Crimes de Incidência Moderada** Afeta parte do estado com volume moderado.	**Perfil: Crimes de Impacto Regional** Afeta regiões específicas do estado, mas de forma muito intensa.
Baixa Prevalência (BP) *(Crime localizado)*	**Perfil: Crimes Raros e Específicos** Ocorre em poucos locais e com pouquíssimas ocorrências. Mais fácil de agrupar.	**Perfil: Crimes de Nicho** Problema localizado em poucas cidades, mas com volume recorrente e moderado.	**Perfil: Crimes Concentrados e de Alto Impacto** O clássico "*hotspot*": problema contido em pouquíssimos locais, mas extremo nesses pontos.

3.3 Resultados e Discussões

Nesta seção são apresentados e discutidos os resultados obtidos a partir dos experimentos realizados. A análise transcende a mera apresentação de índices de desempenho (SC), focando-se em uma investigação transversal que busca relacionar os resultados quantitativos com as propriedades intrínsecas de cada fenômeno criminal. O objetivo é identificar as relações entre as características dos dados e a "clusterabilidade" das séries, a fim de compreender o comportamento de cada algoritmo sobre cada tipo de crime.

3.3.1 Análise do k-shape

Comparando-se os resultados obtidos nos diversos experimentos, notou-se que o k-shape apresentou um desempenho consistentemente inferior aos outros dois algoritmos. Ao longo dos 18 tipos de crimes nos quatro intervalos temporais considerados, os coeficientes de silhueta (SC) para o k-shape raramente ultrapassaram o limiar do Grupo D, indicando a ausência de uma estrutura de agrupamento substancial. Mesmo nos melhores cenários, os SCs dificilmente alcançaram o Grupo C, contrastando fortemente com os resultados obtidos pelos algoritmos k-means e k-DBA. Assim, observou-se inicialmente que as premissas do k-shape poderiam não ser as mais adequadas para a natureza dos dados de criminalidade aqui em estudo.

Em uma investigação mais detalhada na estrutura dos agrupamentos gerados pelo k-shape observou-se um padrão recorrente de formação de *clusters* extremamente desbalanceados, nos quais um único *cluster* agregava a vasta maioria das séries (frequentemente

mais de 95%), enquanto um segundo *cluster* continha um número "residual" de municípios, chegando em alguns casos a apenas uma única série. Este comportamento sugere que o k-shape, em vez de identificar múltiplos padrões de comportamento significativos, atuou primordialmente como um detector de *outliers*. Sua alta sensibilidade à forma o levou a isolar as poucas séries com perfis atípicos, em vez de formar grupos comparáveis e balanceados, resultando em uma estrutura de agrupamento ruim, detectado corretamente pelo coeficiente de silhueta. Na Tabela 8 apresentam-se os resultados obtidos para cada um dos crimes ao longo dos 4 períodos considerados. Nota-se que a classificação de grupos via SC é dominada pelo grupo D.

Tabela 8 – Resultados obtidos via k-shape em cada crime e período.

Período	**Estupro**	**Homicídio Doloso (2)**	**Lesão Cr.[1] Sg. Morte**	**Estupro Vulnerável**	**Homicídio[2] Dol. Acdt**	**Roubo Outros**
5 anos	Grupo D	Grupo D	Grupo C	Grupo D	Grupo C	Grupo D
3 anos	Grupo D	Grupo D	Grupo D	Grupo D	Grupo C	Grupo D
2 anos	Grupo D	Grupo D	Grupo D	Grupo D	Grupo C	Grupo D
1 ano	Grupo D	Grupo D	Grupo D	Grupo D	Grupo C	Grupo D
Período	**Furto Outros**	**Latrocínio**	**Roubo Banco**	**Furto de Veículo**	**Lesão Cr. Cp. Outras[3]**	**Roubo Carga**
5 anos	Grupo D	Grupo C	Grupo C	Grupo D	Grupo D	Grupo D
3 anos	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D
2 anos	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D
1 ano	Grupo D	Grupo D	-	Grupo D	Grupo D	Grupo D
Período	**Homicídio Cp. Outros[4]**	**Lesão Cr. Cp. Acdt[5]**	**Roubo de Veículo**	**Homicídio Cp. Acdt[6]**	**Lesão Cr. Dolosa**	**Tentativa Homicídio**
5 anos	Grupo C	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D
3 anos	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D
2 anos	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D
1 ano	Grupo C	Grupo D	Grupo D	Grupo D	Grupo D	Grupo D

Notas: [a] Todos os períodos finalizam no ano de 2024.
[b] O símbolo "-" indica que não houve agrupamento.
Legenda:
[1]Lesão Corporal Seguida de Morte
[2]Homicídio Doloso Por Acidente de Trânsito
[3]Lesão Corporal Culposa - Outras
[4]Homicídio Culposo Outros
[5]Lesão Corporal Culposa por Acidente de Trânsito
[6]Homicídio Culposo por Acidente de Trânsito

O desempenho ruim do k-shape, observado nos experimentos, pode ser atribuído à premissa do k-shape, que busca agrupar séries que compartilham uma forma comum e bem definida (PAPARRIZOS; GRAVANO, 2016). Utilizando a analogia de que o k-shape é um microscópio, focado na "textura" exata de cada padrão, ele encontra dificuldades ao analisar séries heterogêneas, onde uma mesma tendência geral se manifesta por meio de múltiplas formas distintas — como picos agudos ("V" invertido), subidas graduais ("rampa") ou ondas suaves ("U" invertido). O resultado é a formação de um *cluster* principal altamente heterogêneo, contendo séries com eventos extremos, como picos de criminalidade com desvios padrão altos, acima da sua própria média, o que degrada severamente a coesão do

grupo. Dado que a análise não revelou a existência de padrões de forma claros e distintos, e que os agrupamentos gerados não ofereceram uma sumarização significativa dos dados, optou-se por focar nas análises subsequentes nos resultados obtidos pelo k-means e pelo k-DBA, os quais se mostraram mais robustos.

Em um esforço final para avaliar a robustez do k-shape, foi conduzido um experimento adicional sob condições otimizadas, projetadas para maximizar a chance de encontrar padrões de forma significativos. Para este experimento, a janela temporal foi ampliada para 20 anos (2005-2024) e foi selecionado um crime de alto volume e alta prevalência, cujas séries temporais, após a remoção das nulas, ofereciam um conjunto de dados denso e extenso. A premissa era que um horizonte temporal mais longo, com mais pontos de dados, poderia revelar padrões de forma mais definidos e favorecer o mecanismo do k-shape. Contudo, mesmo sob estas condições, o resultado obtido não demonstrou uma melhora substancial na qualidade do agrupamento, o qual apresentou um SC referente ao Grupo D, detectando k=2 como melhor número de *cluster*. É possível observar na visualização da Figura 10 a alta variação de séries em cada um dos grupos, assim como um centroide não representativo das mesmas.

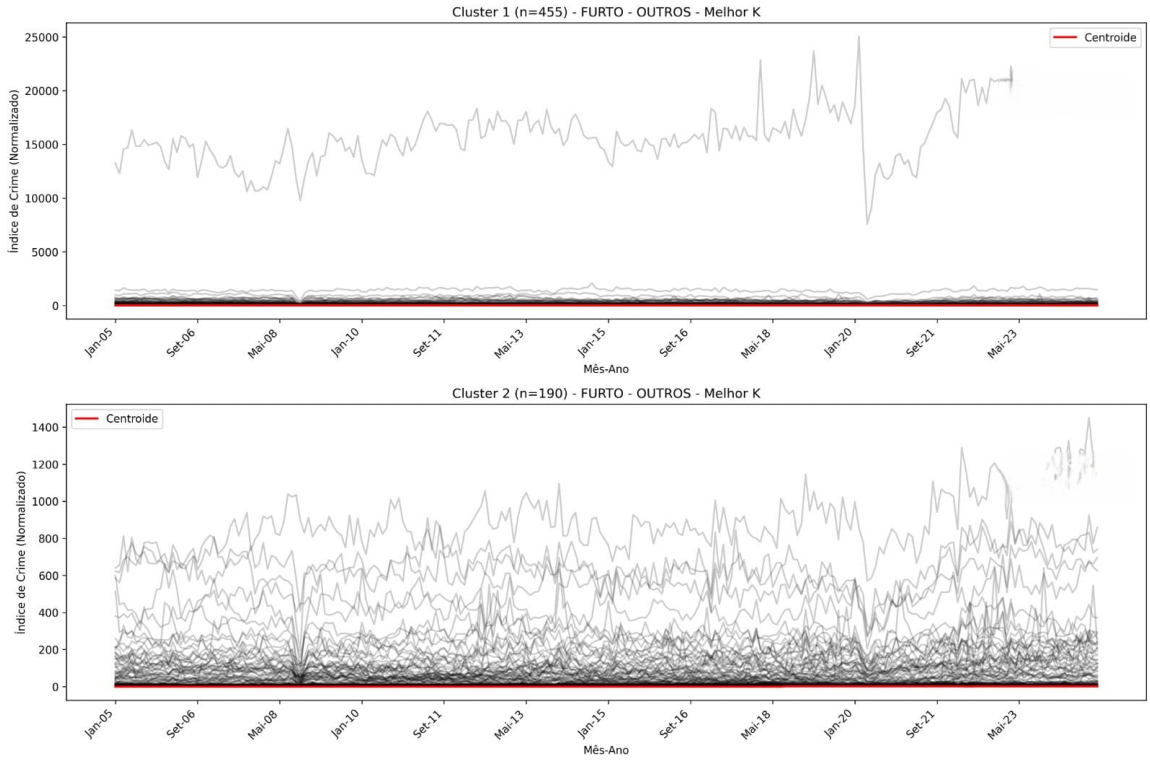

Figura 10 – *Clusters* resultantes para o crime "Furto-Outros" obtidos via k-shape com k=2 no intervalo de 20 anos (2005-2024).

Fonte: Elaborado pelo autor.

3.3.2 Análise do k-means

Os resultados obtidos a partir da aplicação do k-means demonstram que a distância euclidiana (DE) se revela eficaz sob condições específicas, sendo, em outros cenários, inadequada (conforme já relatado nas seções anteriores deste trabalho). Por definição, a DE baseia-se na diferença da magnitude absoluta das ocorrências e na sua sincronia temporal ponto a ponto. A análise de cada um dos crimes, nas diferentes janelas temporais, revelou que, embora o k-means tenha atingido consistentemente as categorias de qualidade Grupo A (estrutura forte) ou Grupo B (estrutura razoável) via SC em múltiplos cenários (Tabela 9), a aplicabilidade do mesmo é estritamente condicionada à estrutura intrínseca do fenômeno criminal, conforme detalhado na Tabela 10. Na maior parte dos resultados notou-se uma predominância de soluções com k =2, sugerindo que, para a maioria dos crimes, a estrutura intrínseca dos dados tende a se dividir de forma binária entre um grupo majoritário, com um perfil de volume, e um grupo "residual".

Os resultados mais favoráveis do k-means foram observados considerando dois padrões distintos de séries. O primeiro padrão refere-se a crimes com uma baixa quantidade de séries. Nestes casos, a identificação de grupos com alto contraste de magnitude foi facilitada, separando de maneira clara os municípios com perfis de ocorrência ativos do vasto conjunto de municípios com volumes muito baixos. Crimes como "Latrocínio", "Roubo a Banco" e "Homicídio Culposo Outros" ilustram este sucesso (vide Tabelas 9 e 10 (padrão 1)). O segundo padrão refere-se a crimes de alta frequência quando o horizonte temporal de análise é encurtado. A melhoria significativa do coeficiente de silhueta em períodos menores, como observado em "Homicídio Doloso (2)" (transição de Grupo C em 5 anos para Grupo A a partir de 3 anos até 1 ano), sugere que a sincronia nos níveis de ocorrência é um fenômeno mais recente e frágil para esses crimes (vide Tabelas 9 e 10 (padrão 2)). Nestes casos, o k-means consegue capturar essa coerência pontual quando a heterogeneidade histórica é removida. Entretanto, o desempenho do k-means demonstrou ser vulnerável em grandes conjuntos de dados, mesmo quando a estrutura parece ser inicialmente forte. O terceiro padrão observado (padrão 3) diz respeito a essa variabilidade em grandes conjuntos e é exemplificado pelo crime "Furto - Outros". Embora este crime apresente resultado forte (Grupo A) em longo prazo (5 anos), sua qualidade de agrupamento se degrada para Grupo D em curto prazo (1 ano) (vide Tabelas 9 e 10 (padrão 3)). Esta flutuação demonstra a vulnerabilidade do k-means à perda anual da sincronia ponto a ponto, apesar de o crime manter um alto volume e prevalência em todos os períodos.

Em contraste com os resultados mais favoráveis, o k-means apresentou resultados ruins (Grupo D) em crimes de alta frequência que exibem forte heterogeneidade temporal ou dessincronização de eventos. O baixo desempenho em separar crimes como "Lesão Corporal Dolosa" (vide Tabelas 9 e 10 (padrão 4)) revela sua principal limitação: o requisito *lock-step* da distância euclidiana em penalizar qualquer variação no ritmo ou no *timing*

dos eventos, mesmo que a forma ou o padrão do crime seja semelhante. Tais resultados, que confirmam a ausência de um padrão na aplicabilidade deste algoritmo, reforçam que o desempenho do mesmo é altamente dependente da adequação da DE às características do conjunto de dados (JAVED; LEE; RIZZO, 2020, p. 4).

Tabela 9 – Resultados obtidos via k-means em cada crime e período.

Período	**Estupro**	**Homicídio Doloso (2)**	**Lesão Cr.[1] Sg. Morte**	**Estupro Vulnerável**	**Homicídio[2] Dol. Acdt**	**Roubo Outros**
5 anos	Grupo C	Grupo C	Grupo B	Grupo D	Grupo A	Grupo D
3 anos	Grupo B	Grupo A	Grupo A	Grupo C	Grupo B	Grupo B
2 anos	Grupo B	Grupo A	Grupo A	Grupo B	Grupo B	Grupo B
1 ano	Grupo B	Grupo A	Grupo B	Grupo C	Grupo B	Grupo B
Período	**Furto Outros**	**Latrocínio**	**Roubo Banco**	**Furto de Veículo**	**Lesão Cr. Cp. Outras[3]**	**Roubo Carga**
5 anos	Grupo A	Grupo A	Grupo A	Grupo C	Grupo A	Grupo D
3 anos	Grupo D	Grupo A	Grupo B	Grupo C	Grupo A	Grupo C
2 anos	Grupo D	Grupo A	Grupo C	Grupo D	Grupo A	Grupo A
1 ano	Grupo D	Grupo B	-	Grupo B	Grupo B	Grupo A
Período	**Homicídio Cp. Outros[4]**	**Lesão Cr. Cp. Acdt[5]**	**Roubo de Veículo**	**Homicídio Cp. Acdt[6]**	**Lesão Cr. Dolosa**	**Tentativa Homicídio**
5 anos	Grupo A	Grupo D	Grupo A	Grupo C	Grupo D	Grupo C
3 anos	Grupo A	Grupo D	Grupo B	Grupo B	Grupo D	Grupo B
2 anos	Grupo A	Grupo D	Grupo B	Grupo C	Grupo D	Grupo B
1 ano	Grupo A	Grupo C	Grupo A	Grupo B	Grupo D	Grupo B

Notas: [a] Todos os períodos finalizam no ano de 2024.
[b] O símbolo "-" indica que não houve agrupamento.
Legenda:
[1]Lesão Corporal Seguida de Morte
[2]Homicídio Doloso Por Acidente de Trânsito
[3]Lesão Corporal Culposa - Outras
[4]Homicídio Culposo Outros
[5]Lesão Corporal Culposa por Acidente de Trânsito
[6]Homicídio Culposo por Acidente de Trânsito

Diante do exposto, pode-se dizer que o k-means, devido à sua simplicidade e ao baixo custo computacional (linear) (JAVED; LEE; RIZZO, 2020, p. 1123), é particularmente útil quando a finalidade da análise é a divisão de municípios com perfis de volume distintos ou quando o interesse está em padrões de sincronia em janelas temporais curtas. No entanto, sua sensibilidade a defasagens temporais exige que sua aplicação seja complementar a outros algoritmos (k-DBA), cuja flexibilidade temporal pode capturar as similaridades de forma que a DE rigidamente ignora.

Tabela 10 – Categorização de padrões de desempenho ocorridos nos resultados do k-means exemplificados por alguns crimes.

Crime	Período	Qtd. Séries	Grupo	Observação sobre o Padrão Identificado
Padrão 1: Alto Desempenho em Baixo Volume de Séries (Forte Contraste de Magnitude)				
Latrocínio	5 Anos	210	A	Demonstra a eficácia do k-means em cenários de poucas ocorrências, onde a separação por volume (ativo vs. inativo) cria *clusters* de estrutura forte.
Roubo a Banco	5 Anos	27	A	Desempenho máximo mesmo com pouquíssimas séries, reforçando que o k-means é altamente robusto para isolar fenômenos raros com claro contraste de magnitude.
Homicídio Culposo Outros	5 Anos	223	A	Atingiu o Grupo A em todos os períodos, indicando que, para este crime, a similaridade por níveis absolutos é altamente estável e sincronizada no longo prazo.
Padrão 2: Sincronia Frágil (Melhora de Qualidade com a Redução do Período)				
Homicídio Doloso (2)	5 Anos	554	C	Resultado fraco em longo prazo, sugerindo que a sincronia histórica de volume é heterogênea.
Homicídio Doloso (2)	3,2,1 Anos	381	A	Forte aumento de qualidade, indicando que a sincronia de magnitude (níveis de crime) se estabeleceu de forma mais nítida apenas nos períodos mais recentes.
Padrão 3: Desempenho Variável em Grandes Conjuntos				
Furto - Outros	5 Anos	645	A	Resultado forte em longo prazo, mostrando que os perfis de volume se mantiveram sincronizados e estáveis na maioria dos municípios.
Furto - Outros	1 Ano	645	D	Perda de qualidade (Grupo D) em curto prazo, demonstrando a vulnerabilidade do k-means quando a sincronia ponto a ponto é perdida anualmente, apesar da alta quantidade de séries.
Padrão 4: Falha Consistente (Alta Heterogeneidade Temporal)				
Lesão Corporal Dolosa	5 Anos	645	D	Falha consistente em todos os períodos. O agrupamento por magnitude sincronizada é ineficaz, pois o crime é dominado por heterogeneidade temporal e não por volumes absolutos alinhados.

Legenda: DE: Distância euclidiana.

3.3.3 Análise do k-DBA

A utilização da medida DTW no k-DBA se apresenta como uma alternativa mais robusta para o agrupamento de séries de criminalidade em comparação com o k-means. Por ser uma medida elástica, a DTW lida com distorções inerentes às séries temporais, como variações de ritmo e deslocamentos temporais (*shifting*), ao buscar o caminho de alinhamento que minimiza a distância total entre as sequências, independentemente de os eventos ocorrerem no mesmo instante. Esta capacidade de alinhamento não linear permite que a DTW se concentre na similaridade da forma e na dinâmica evolutiva dos crimes, o que é crucial em cenários onde a dinâmica criminal entre municípios não é perfeitamente sincronizada.

Tabela 11 – Resultados obtidos via k-DBA em cada crime e período.

Período	Estupro	Homicídio Doloso (2)	Lesão Cr.[1] Sg. Morte	Estupro Vulnerável	Homicídio[2] Dol. Acdt	Roubo Outros
5 anos	Grupo C	Grupo A	Grupo A	Grupo A	Grupo A	Grupo B
3 anos	Grupo A	Grupo A	Grupo A	Grupo B	Grupo A	Grupo B
2 anos	Grupo A	Grupo A	Grupo A	Grupo B	Grupo A	Grupo B
1 ano	Grupo B	Grupo A	Grupo A	Grupo B	Grupo A	Grupo B
Período	**Furto Outros**	**Latrocínio**	**Roubo Banco**	**Furto de Veículo**	**Lesão Cr. Cp. Outras[3]**	**Roubo Carga**
5 anos	Grupo A	Grupo A	Grupo A	Grupo A	Grupo A	Grupo A
3 anos	Grupo A	Grupo A	Grupo A	Grupo B	Grupo A	Grupo C
2 anos	Grupo A	Grupo A	Grupo A	Grupo B	Grupo A	Grupo A
1 ano	Grupo A	Grupo A	-	Grupo B	Grupo A	Grupo B
Período	**Homicídio Cp. Outros[4]**	**Lesão Cr. Cp. Acdt[5]**	**Roubo de Veículo**	**Homicídio Cp. Acdt[6]**	**Lesão Cr. Dolosa**	**Tentativa Homicídio**
5 anos	Grupo A	Grupo B	Grupo A	Grupo A	Grupo D	Grupo A
3 anos	Grupo A	Grupo B	Grupo B	Grupo A	Grupo D	Grupo A
2 anos	Grupo A	Grupo B	Grupo B	Grupo A	Grupo D	Grupo A
1 ano	Grupo A	Grupo B	Grupo B	Grupo B	Grupo B	Grupo B

Notas: [a] Todos os períodos finalizam no ano de 2024.
[b] O símbolo "-" indica que não houve agrupamento.
Legenda:
[1]Lesão Corporal Seguida de Morte
[2]Homicídio Doloso Por Acidente de Trânsito
[3]Lesão Corporal Culposa - Outras
[4]Homicídio Culposo Outros
[5]Lesão Corporal Culposa por Acidente de Trânsito
[6]Homicídio Culposo por Acidente de Trânsito

Observando-se os resultados (Tabela 11) nota-se a eficácia do k-DBA em identificar padrões de alta qualidade (Grupos A e B) em crimes que o k-means classificou como de baixa qualidade (Grupos C ou D) (comparar com a Tabela 9). Um caso notável é o crime "Homicídio Doloso (2)": enquanto o k-means apresenta desempenho Grupo C no horizonte temporal de 5 anos, o k-DBA apresenta desempenho Grupo A para o mesmo período (Tabela 12 (padrão 1)). Este contraste evidencia que os municípios que compartilham a mesma dinâmica criminal apresentam defasagens temporais que o k-means ignora. Resultados semelhantes foram observados em "Lesão Corporal Culposa por Acidente de Trânsito" (de Grupo D no k-means para Grupo B no k-DBA) (Tabela 12 (padrão 1)) e "Roubo de Carga" (de Grupo D no k-means para Grupo A no k-DBA). O crime "Roubo a Banco", por sua vez, um crime de baixo volume, manteve a alta qualidade (Grupo A) do agrupamento obtida pelo k-means, confirmando que a elasticidade da DTW preserva a alta separabilidade de magnitude em dados esparsos (Tabela 12 (padrão 2)). Estas observações sugerem que, para esses crimes, a dinâmica e a forma do padrão de ocorrências são mais coesas entre os municípios do que a sincronização da magnitude absoluta, conforme sintetizado na Tabela 12.

O k-DBA não apenas demonstrou ser mais eficaz em contornar o desalinhamento temporal, mas também indicou uma maior capacidade de desvendar a estrutura complexa

dos dados, mantendo um alto grau de qualidade em quase todos os cenários analisados. Embora o número de clusters ótimo (k) tenha permanecido predominantemente baixo (k=2) para muitos crimes, o k-DBA revelou uma sensibilidade para estruturas mais detalhadas em crimes de menor prevalência. Por exemplo, para "Lesão Corporal Seguida de Morte", o algoritmo encontrou soluções ótimas com k=15 em períodos mais curtos, sugerindo que a elasticidade da medida DTW permitiu a descoberta de sub-padrões e nuances na dinâmica que são mascarados pela rigidez do alinhamento ponto a ponto (Tabela 12 (padrão 2)). Este desempenho robusto do k-DBA, especialmente em dados com variações inerentes, o posiciona como um algoritmo adequado para a análise exploratória de séries temporais no domínio aqui apresentado.

Tabela 12 – Categorização de padrões de desempenho ocorridos nos resultados do k-DBA exemplificados por alguns crimes.

Crime	**Período**	**Qtd. Séries**	**Grupo**	**Observação sobre o Padrão Identificado**
Padrão 1: Alta Robustez em Cenários de Dessincronização				
Homicídio Doloso (2)	5 ANOS	554	**A**	Notou-se a eficácia da flexibilidade do k-DBA, pois a alta qualidade (Grupo A) sugere que a similaridade de forma existia, mas estava dessincronizada no tempo, o que fez com que o k-means a penalizasse.
Lesão Corporal Culposa por Acidente de Trânsito	5 ANOS	645	**B**	A melhoria de Grupo D (k-means) para Grupo B indica que as séries possuem formas dinâmicas similares, mesmo que ocorram em momentos distintos entre os municípios.
Padrão 2: Capacidade de Descoberta de Estruturas Complexas (k Alto)				
Lesão Corporal Seguida de Morte	5 ANOS	159	**A**	O k ótimo encontrado (k=15) sugere que o k-DBA (DTW), com sua elasticidade, consegue identificar múltiplas sub-estruturas ou nuances na dinâmica de crimes raros, que não seriam visíveis com a rigidez do k-means (DE).
Roubo a Banco	5 ANOS	27	**A**	O k-DBA manteve a qualidade alta (Grupo A) em relação ao k-means, indicando que a similaridade por magnitude absoluta em dados de baixo volume é preservada mesmo com o alinhamento elástico.
Padrão 3: Falha Consistente (Ausência de Padrão Estrutural)				
Lesão Corporal Dolosa	5 ANOS	645	**D**	A falha em obter um agrupamento coerente (Grupo D) sugere que, para este crime, não há nem sincronia de magnitude (falha do k-means (DE)) nem similaridade de forma/dinâmica (falha do k-DBA (DTW)) entre os municípios.

Apesar do bom desempenho geral apresentado pelo k-DBA, a análise identificou um limite claro para sua aplicabilidade. O crime "Lesão Corporal Dolosa" persistiu no Grupo D (ausência de estrutura substancial) em quase todos os períodos analisados (Tabelas 11 e 12 (padrão 3)). Este resultado, combinado com o desempenho ruim também no k-means (Tabela 9), sugere que para este tipo de crime as séries dos municípios carecem tanto de sincronia de volume quanto de similaridade de forma ou dinâmica. Nesses casos de alta desordem e heterogeneidade estrutural, nenhum dos algoritmos se mostrou adequado. Em suma, o k-DBA mostrou-se altamente vantajoso para capturar a semelhança de padrões temporais e de forma (dinâmica) onde o k-means falha devido à sensibilidade ao *timing*,

mas suas limitações persistem em dados onde a variação não é apenas temporal, mas fundamentalmente estrutural.

3.3.4 Perfis Criminais x Qualidade dos Agrupamentos

O objetivo desta seção é relacionar as propriedades intrínsecas dos crimes — volume (mediana de ocorrências) e prevalência (percentual de municípios ativos) — com a estabilidade da qualidade dos agrupamentos (Grupo) (vide definição da matriz de perfis de crime na Subseção 3.2.4). Esta estratégia permite investigar como a estrutura do conjunto de séries, seja ela marcada pela sincronia rígida de magnitude ou pela similaridade elástica de forma, condiciona a eficácia do agrupamento. Assim, a identificação de grupos de municípios é, portanto, interpretada como estritamente dependente da natureza do fenômeno criminal subjacente, reforçando que a aplicabilidade dos algoritmos é condicionada aos dados, especialmente em face da heterogeneidade e do desalinhamento temporal inerentes às séries (PAPARRIZOS; YANG; LI, 2024).

A fim de se obter a matriz de perfis de crime (Tabela 14), a Tabela 13 foi criada. A mesma relaciona as propriedades intrínsecas dos crimes — volume e prevalência — com a estabilidade dos agrupamentos via qualidade (Grp), intervalo temporal (IT) e a moda de grupos (Mo). Por exemplo, a primeira linha indica que o crime “Roubo a Banco” tem volume 1 e prevalência de 4%, i.e., apresenta mediana igual a 1 em relação ao total de ocorrências em cinco anos considerando os municípios ativos (séries não nulas), assim como um total de 4% de municípios que registraram ao menos uma ocorrência neste período. Ademais, em relação ao k-means, que o melhor k encontrado, de acordo com o coeficiente de silhueta, se encontra no Grupo A (Grp), o qual foi obtido no intervalo temporal (IT) de 5 anos e com moda (Mo) indefinida considerando todos os possíveis k nos diferentes ITs. Já em relação ao k-DBA, o melhor k encontrado se encontra no Grupo A (Grp), o qual foi obtido no intervalo temporal (IT) de 5 anos e com moda (Mo) de grupos igual a A considerando todos os possíveis k nos diferentes ITs. Deste modo, a tabela condensa as descobertas sobre a eficácia de cada algoritmo para diferentes aspectos dos dados (volume e prevalência), servindo como base para se avaliar os agrupamentos. Em linhas gerais pode-se dizer que o volume representa o quanto um dado crime é recorrente entre os municípios e a prevalência, a proporção com que um dado crime ocorre nos municípios. Por fim, vale mencionar que enquanto o volume é obtido por meio da mediana das somas das ocorrências de cada município, a prevalência computando-se a porcentagem de ocorrência do crime considerando o total de séries não nulas.

Tabela 13 – Sumário da eficácia dos algoritmos em cada um dos crimes considerando seus aspectos de volume e prevalência.

Crime	**Vol.**	**Prv.**	**k-means**			**k-DBA**		
	M_d	(%)	**Grp**	**IT**	Mo	**Grp**	**IT**	Mo
Roubo a Banco	1	4 %	A	5	*	A	5	A
Homicídio Doloso A. Trânsito	1	6 %	A	5	B	A	3	A
Lesão Crp.[1] Seguida de Morte	1	25 %	A	3	B	A	3	A
Latrocínio	1	33 %	A	5	A	A	2	A
Homicídio Culposo Outros	1	35 %	A	5	A	A	1	A
Roubo de Carga	4	55 %	A	2	A	A	2	A
Lesão Crp.[1] Culposa - Outras	4	87 %	A	5	A	A	5	A
Homicídio Doloso (2)	5	86 %	A	1	A	A	2	A
Estupro	6	91 %	B	2	B	A	3	A
Tentativa de Homicídio	7	94 %	B	2	B	A	5	A
Homicídio Culposo A. Trânsito	8	96 %	B	3	C	A	5	A
Roubo de Veículo	10	79 %	A	1	A	A	5	B
Estupro de Vulnerável	22	99 %	B	2	C	A	5	B
Furto de Veículo	26	98 %	B	1	C	A	5	B
Roubo - Outros	29	94 %	B	3	B	B	2	B
Lesão Crp.[1] Culposa A. Trânsito	73	87 %	C	1	D	B	5	B
Lesão Crp.[1] Dolosa	214	100 %	D	1	D	B	1	D
Furto - Outros	434	100 %	A	5	D	A	2	A

Notas:
O símbolo "*" indica a ausência de moda.
Legenda:
Grp: Grupo
Mo: Moda Grupo
IT: Intervalo Temporal
Vol.: Volume das ocorrências representado pela mediana
Prv.: Prevalência do crime em percentual
[1]Crp.: Corporal

Tendo como base a Tabela 13 foi necessário se definir os limiares que separassem o volume em faixas. Para tanto, considerou-se os extremos das maiores variações observadas, i.e., de 1 a 10 para baixo, de 22 a 73 para médio, de 214 a 434 para alto. Deste modo, valores muito próximos não são utilizados para se definir as margens e gerar incertezas. O mesmo raciocínio foi utilizado para se definir os limiares da prevalência, a saber: de 4% a 35% baixa, 55% média, de 79% a 100% alta. Assim, a partir dessa "discretização" criou-se a matriz de perfis de crime, apresentada na Tabela 14. Tal "discretização" permite a categorização do volume e da prevalência em valores categóricos que podem então ser relacionados com os resultados dos agrupamentos para ambos os algoritmos (k-means e k-DBA).

Tabela 14 – Matriz de Perfis de Crime por Volume e Prevalência considerando os resultados dos experimentos realizados.

	BV	**MV**	**AV**
AP	• Estupro • Roubo de Veículo • Homicídio Doloso (2) • Tentativa de Homicídio • Lesão Corporal Culposa - Outras • Homicídio Culposo Por Acidente de Trânsito	• Estupro de Vulnerável • Furto de Veículo • Roubo - Outros • Lesão Corporal Culposa Por Acidente de Trânsito	• Lesão Corporal Dolosa • Furto - Outros
MP	• Roubo de Carga	-	-
BP	• Roubo a Banco • Latrocínio • Homicídio Culposo Outros • Lesão Corporal Seguida de Morte • Homicídio Doloso por Acidente de Trânsito	-	-

Notas: O símbolo "-" indica ausência de relação.
Legenda:
BV: Baixo Volume BP: Baixa Prevalência
MV: Médio Volume MP: Média Prevalência
AV: Alto Volume AP: Alta Prevalência

Diante do exposto, a matriz de perfis de crime (Tabela 14) se apresenta como uma síntese da caracterização empírica dos 18 crimes investigados. Cada crime foi posicionado em um dos nove quadrantes resultantes do cruzamento entre volume e prevalência. Essa categorização revela que a natureza estrutural de cada fenômeno criminal pode ser caracterizada por dois eixos independentes: a intensidade das ocorrências (volume) e a dispersão geográfica (prevalência). Essa matriz funciona como um diagnóstico estrutural que permite identificar, com base nas propriedades intrínsecas de cada delito, o grau de desafio que ele representará para os algoritmos de agrupamento.

Os crimes posicionados no quadrante de Baixo Volume (BV) e Baixa Prevalência (BP), classificados como "Crimes Raros e Específicos", incluem "Roubo a Banco" (Vol. 1, Prv. 4%), "Homicídio Doloso por Acidente de Trânsito" (Vol. 1, Prv. 6%) e "Latrocínio" (Vol. 1, Prv. 33%). Esses crimes caracterizam-se por ocorrências esparsas tanto em magnitude quanto em distribuição geográfica, resultando em séries temporais com alto contraste entre municípios ativos e inativos. Este contraste facilita a separação dos grupos. No extremo oposto tem-se o quadrante de Alto Volume (AV) e Alta Prevalência (AP), representado por "Lesão Corporal Dolosa" (Vol. 214, Prv. 100%) e "Furto - Outros" (Vol. 434, Prv. 100%), abrigando os "Crimes Endêmicos", que afetam praticamente todos os municípios com frequências elevadas. Esses perfis apresentam séries temporais densas, com variações sutis entre os municípios, dificultando a identificação de padrões distintos. Já os quadrantes intermediários revelam perfis diferentes. No quadrante de Baixo Volume (BV) e Alta Prevalência (AP) encontram-se crimes como "Homicídio Doloso (2)" (Vol. 5, Prv. 86%), "Estupro" (Vol. 6, Prv. 91%) e "Tentativa de Homicídio" (Vol. 7, Prv. 94%),

categorizados como “Crimes Crônicos de Baixa Intensidade”. Embora ocorram em quase todo o estado, apresentam baixo volume em cada local, caracterizando um problema generalizado, mas de intensidade moderada. O quadrante de Baixo Volume (BV) e Média Prevalência (MP) é composto pelo crime “Roubo de Carga” (Vol. 4, Prv. 55%), classificado como “Crimes Focais de Baixa Intensidade”. Por fim, o quadrante de Médio Volume (BV) e Alta Prevalência (AP) inclui “Furto de Veículo” (Vol. 26, Prv. 98%) e “Roubo - Outros” (Vol. 29, Prv. 94%), classificado como “Crimes Dispersos e de Atenção”.

Em uma última análise, visando relacionar as Tabelas 13 e 14, a Tabela 15 foi criada. A mesma apresenta uma matriz que relaciona os algoritmos de agrupamento com os perfis criminais, fornecendo um guia entre cada perfil criminal (definido por Volume × Prevalência) ao algoritmo de agrupamento mais adequado (via coeficiente de silhueta (SC)). Deste modo, analisa-se de maneira comparativa os resultados obtidos ao longo dos quatro intervalos temporais (1, 2, 3 e 5 anos), evidenciando a condicionalidade da eficácia dos algoritmos em função da densidade e da sincronia das séries criminais.

Tabela 15 – Sumário da relação entre os algoritmos e cada perfil criminal (volume x prevalência).

	BV	**MV**	**AV**
AP	**k-DBA (necessário)** **Grp** A - * ***Mo*** A - **IT** = 2/3/5 **Motivo:** resgata a forma em séries densas e dessincronizadas.	**k-DBA (obrigatório)** **Grp** A/B - * ***Mo*** A/B - **IT** = 2/3/5 **Motivo:** k-DBA essencial para perfis densos com AV/MP.	**k-DBA (desafiador)** **Grp** A/B - * ***Mo*** A/D - **IT** = 1/2 **Motivo:** k-DBA é o único viável, *Mo*: instável (A ou D).
MP	**k-DBA (superior)** **Grp** A - * ***Mo*** A - **IT** = 2 **Motivo:** k-means e k-DBA são igualmente adequados para perfis MP/BV, ambos atingindo *Mo*=A no IT = 2.	-	-
BP	**k-means (suficiente)** **Grp** A - * ***Mo*** A/B - **IT** = 3/5 **Motivo:** k-means é suficiente (*Mo*=A) devido ao alto contraste de magnitude.	-	-

Notas:
O símbolo “*” indica variação em todos os anos ou Intervalos Temporais (IT).
Legenda:
Mo: Moda
Grp: Grupo — IT: Intervalo Temporal
BV: Baixo Volume — BP: Baixa Prevalência
MV: Médio Volume — MP: Média Prevalência
AV: Alto Volume — AP: Alta Prevalência

Para crimes de Baixa Prevalência e Baixo Volume (BP/BV), como “Latrocínio”, “Roubo a Banco”, “Homicídio Culposo Outros”, “Lesão Corporal Seguida de Morte” e “Homicídio Doloso por Acidente de Trânsito”, o k-means mostrou-se suficiente, alcançando consistentemente Grupo A (estrutura forte) com moda A/B nos intervalos temporais de 3 e 5 anos (IT = 3/5). Nesse perfil o alto contraste de magnitude entre municípios ativos e inativos simplifica a separação dos grupos, tornando desnecessária a flexibilidade de alinhamento temporal oferecida pela DTW. Em contrapartida, para crimes de Alta Prevalência e Baixo Volume (AP/BV), como “Estupro”, “Roubo de Veículo”, “Homicídio Doloso (2)”, “Tentativa de Homicídio”, “Lesão Corporal Culposa - Outras” e “Homicídio Culposo Por Acidente de Trânsito”, o k-DBA tornou-se necessário: sua capacidade de alinhamento elástico resgatou padrões dinâmicos que o k-means não capturou, elevando a qualidade dos agrupamentos para Grupo A com moda A nos intervalos temporais de 2, 3 e 5 anos (IT = 2/3/5). Este achado demonstra que, em séries densas (alta prevalência) e dessincronizadas, o k-means falha ao impor uma correspondência rígida ponto a ponto, enquanto o k-DBA reconhece similaridades mesmo quando os eventos ocorrem em momentos distintos.

Para crimes de Alta Prevalência e Médio Volume (AP/MV), como “Estupro de Vulnerável”, “Furto de Veículo”, “Roubo - Outros” e “Lesão Corporal Culposa Por Acidente de Trânsito”, o k-DBA revelou-se obrigatório, alcançando Grupo A/B com moda A/B nos intervalos temporais de 2, 3 e 5 anos (IT = 2/3/5), enquanto o k-means apresentou desempenho insuficiente. Para crimes de Alta Prevalência e Alto Volume (AP/AV), como “Lesão Corporal Dolosa” e “Furto - Outros”, o k-DBA foi o único algoritmo viável, embora a moda tenha se mostrado instável (oscilando entre A e D nos intervalos temporais de 1 e 2 anos (IT = 1/2)), evidenciando que a alta densidade das séries dificulta a formação de *clusters* coesos mesmo com alinhamento elástico. Para o único crime de Média Prevalência (MP) e Baixo Volume (BV), “Roubo de Carga”, tanto o k-means quanto o k-DBA alcançaram desempenho equivalente, ambos proporcionando Grupo e moda A no intervalo temporal de 2 anos (IT = 2). Este resultado demonstra que, para perfis MP/BV, não há diferença significativa entre as duas abordagens: o contraste de magnitude (que favorece o k-means) e a capacidade de alinhamento temporal (que favorece o k-DBA) convergem para estruturas de agrupamento igualmente robustas. Este comportamento sugere que perfis AP/AV podem beneficiar-se de algoritmos complementares, como os hierárquicos ou baseados em densidade. Em conjunto, a Tabela 15 consolida a condicionalidade da eficácia dos algoritmos: o k-means é suficiente apenas em perfis esparsos (BP/BV), enquanto o k-DBA é necessário para perfis AP/BV, obrigatório para AP/MV, essencial (mas desafiador) para AP/AV, e superior para MP/BV. Este protocolo metodológico permite que gestores públicos identifiquem, sem necessidade de testes exaustivos, qual algoritmo aplicar para cada tipo de crime, consolidando um guia diagnóstico baseado em evidências empíricas.

3.3.5 Considerações Finais

Determinar padrões de comportamento coerentes entre os crimes provou ser um desafio complexo, uma vez que a previsibilidade dos acontecimentos criminais envolve múltiplos fatores, incluindo a magnitude das ocorrências, os momentos em que os eventos acontecem e a região geográfica. Este cenário se torna significativamente mais difícil pela variedade de perfis entre os diferentes tipos de crime, o que impede a adoção de uma solução de agrupamento universal. O coeficiente de silhueta (SC) foi adotado como índice de avaliação, por ser de fácil interpretação e amplamente utilizado para medir a coesão e separação dos grupos.

A exploração detalhada dos dados e das técnicas de pré-processamento — como a filtragem de séries nulas e a investigação de diferentes métodos de normalização (z-score e *RobustScaler*) — foi fundamental para assegurar consistência e interpretabilidade aos resultados. A construção da matriz de perfis de crime (Tabela 14) e, posteriormente, da matriz de relação algoritmo por perfil criminal (Tabela 15) sumarizou os achados empíricos, permitindo compreender as especificidades metodológicas de cada perfil criminal. Pode-se dizer que tais matrizes constituem artefatos diagnósticos que permitem prever a "clusterabilidade" de um crime e orientar a escolha do algoritmo. Os experimentos realizados abrangeram 18 tipos de crime, 4 janelas temporais (1, 2, 3 e 5 anos), 3 algoritmos (k-means, k-DBA e k-shape) e variação de k entre 2 e 15, totalizando 3.024 execuções de agrupamento. Ao se analisar comparativamente os algoritmos notou-se que:

- o k-means mostrou-se adequado apenas em perfis de Baixo Volume (BV) e Baixa Prevalência (BP), nos quais o contraste de magnitudes favorece o agrupamento direto, alcançando consistentemente moda A. Exemplos incluem "Latrocínio" e "Roubo a Banco", onde a esparsidade das séries facilita a separação por magnitude absoluta;
- o k-DBA demonstrou ser essencial para crimes de Alta Prevalência (AP), pois sua capacidade de alinhamento elástico reconhece padrões semelhantes mesmo quando ocorrem em momentos distintos. Em perfis como "Homicídio Doloso (2)" e "Estupro", o k-DBA alcançou moda A ou B, resgatando padrões dinâmicos que o k-means não capturou (Grupo D);
- o k-shape apresentou desempenho inferior em praticamente todos os cenários analisados, indicando que sua noção de similaridade de forma não capturou adequadamente o comportamento das séries de criminalidade dentro do pipeline experimental adotado, atuando mais como um detector de padrões anômalos do que como um agrupador eficiente.

Observou-se ainda uma predominância de soluções com k=2, indicando que as estruturas internas dos dados tendem a se organizar de maneira binária: um *cluster*

principal, representando o comportamento predominante, e um *cluster* “residual”, composto por municípios com padrões atípicos, cuja separação contribuiu para as melhores pontuações de silhueta obtidas. Esta configuração (k=2) maximiza o coeficiente de silhueta, pois a separação entre o *cluster* “principal” e o “residual” é naturalmente mais nítida.

4 CONCLUSÕES E TRABALHOS FUTUROS

A criminalidade urbana constitui um desafio multifacetado para gestores públicos, exigindo ferramentas analíticas que permitam identificar padrões espaciais e temporais para subsidiar políticas baseadas em evidências. Este trabalho teve como objetivo avaliar a aplicabilidade de algoritmos de agrupamento particional — especificamente k-means, k-DBA e k-shape — para identificar grupos de municípios do Estado de São Paulo com séries históricas de criminalidade semelhantes, utilizando dados mensais da Secretaria de Segurança Pública (SSP-SP) referentes a 18 tipos de crime.

A metodologia adotada abrangeu a coleta e organização de séries temporais de 645 municípios entre 2001 e 2025, com foco nos períodos de 1, 2, 3 e 5 anos (2020-2024). Após o pré-processamento, os algoritmos selecionados foram executados variando-se o k entre 2 e 15, totalizando 3.024 experimentos. A qualidade dos agrupamentos foi avaliada via coeficiente de silhueta, permitindo identificar o número ótimo de *clusters* e comparar o desempenho dos algoritmos entre si.

Os resultados evidenciaram que a eficácia do agrupamento é condicional à natureza estrutural do crime. O k-means mostrou-se suficiente para crimes de Baixo Volume/Baixa Prevalência (BV/BP). Já o k-DBA foi essencial para crimes de Alta Prevalência (AP), onde o alinhamento elástico do k-DBA resgatou padrões dinâmicos que o k-means não capturou. O k-shape apresentou desempenho inferior, não capturando adequadamente a similaridade de forma nas séries criminais. Observou-se predominância de k=2 como solução ótima, refletindo uma estrutura binária (*cluster* “principal” + *cluster* “residual”). A matriz de perfis de crime (Tabela 14) e a matriz de relação algoritmo por perfil criminal (Tabela 15) consolidaram essas observações, funcionando como guias diagnósticos para prever a “clusterabilidade” de cada crime.

Vale mencionar três limitações principais do presente trabalho, a saber: (i) a análise foi conduzida de maneira univariada, considerando exclusivamente a evolução temporal de cada tipo de crime; uma abordagem multivariada, incorporando variáveis externas como escolaridade e taxa de pobreza, poderia revelar perfis municipais mais complexos e suas inter-relações com os padrões criminais; (ii) a análise baseou-se exclusivamente no coeficiente de silhueta; embora amplamente reconhecido, a complementação com outros índices internos poderia enriquecer a avaliação; (iii) a análise restringiu-se à família de algoritmos particionais; algoritmos hierárquicos e baseados em densidade poderiam capturar estruturas não-esféricas não detectadas pelos algoritmos aqui avaliados. Tais aspectos podem ser explorados em trabalhos futuros. O tratamento refinado de *outliers* e *singletons*, especialmente em crimes de baixíssima frequência, também merece investigação futura.

Em relação ao impacto (contribuição) deste trabalho, pode-se mencionar o estudo da viabilidade de algoritmos de agrupamento como ferramentas de apoio à gestão de segurança pública. Os resultados obtidos podem subsidiar políticas baseadas em evidências, permitindo que gestores identifiquem municípios com perfis criminais semelhantes e aloquem recursos de maneira estratégica. Por exemplo, municípios agrupados no mesmo *cluster* para "Roubo de Veículo" compartilham dinâmicas temporais similares, sugerindo que intervenções testadas em um município podem ser eficazes nos demais do grupo. Já a visualização dos centroides permite comunicar padrões criminais de maneira acessível a tomadores de decisão não especialistas, traduzindo resultados analíticos em *insights* acionáveis. Ademais, a matriz de relação algoritmo por perfil criminal (Tabela 15) consolida um protocolo metodológico replicável, o qual pode ser aplicado, por exemplo, a outros estados brasileiros. Ao transformar dados históricos em conhecimento acionável, este estudo contribui para a construção de uma gestão pública mais eficiente, transparente e orientada por dados, fortalecendo a capacidade do Estado em prevenir e combater a criminalidade.

Por fim, vale mencionar que em paralelo a este trabalho foi sendo construído um sistema para permitir o agrupamento das séries por parte dos tomadores de decisão. O sistema permite, por meio de interface interativa, a parametrização dos dados de modo a obter os resultados visuais dos agrupamentos. O sistema facilita tanto a compreensão dos dados quanto a interpretação, por parte de profissionais, do comportamento entre os diferentes tipos de criminalidade. O sistema disponibiliza aos usuários: (i) seleção dinâmica de tipos de crimes, períodos temporais e algoritmos de agrupamento; (ii) visualização de mapas de calor e gráficos de centroides; (iii) exportação de relatórios automatizados; e (iv) atualização automática conforme novos dados mensais sejam disponibilizados pela SSP-SP. Este sistema transforma a análise exploratória aqui apresentada em ferramenta operacional para o planejamento estratégico de segurança pública, ampliando seu alcance e utilidade prática. A Figura 11 apresenta uma das interfaces do sistema desenvolvido. Tal sistema está sendo desenvolvido por alunos de graduação sob a supervisão da supervisora deste trabalho.

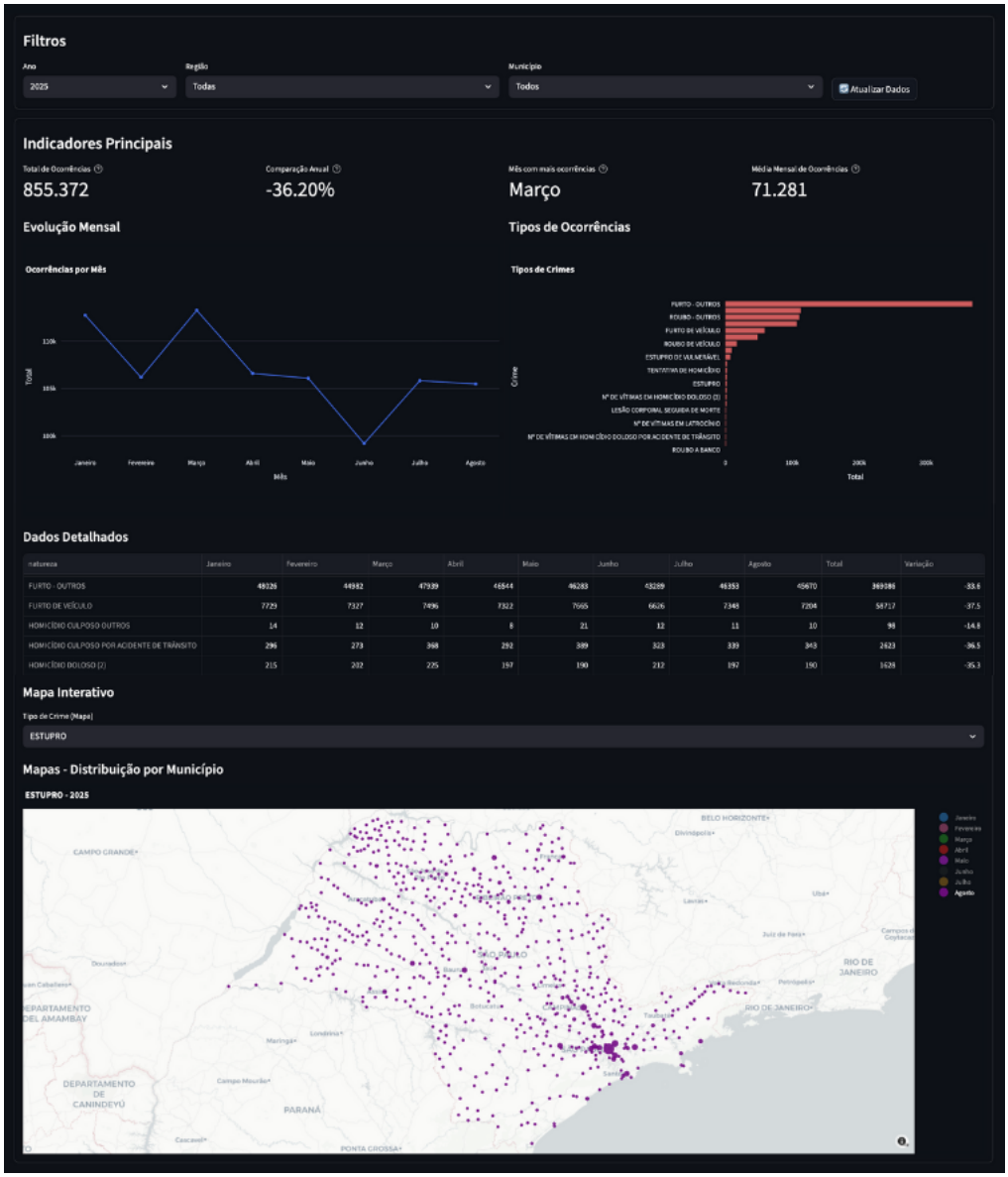

Figura 11 – Uma das interfaces do sistema em desenvolvimento.

Fonte: Disponível em <https://sspdata.streamlit.app/>.

REFERÊNCIAS

AGHABOZORGI, S.; SHIRKHORSHIDI, A. S.; WAH, T. Y. Time-series clustering – A decade review. **Information Systems**, Elsevier, v. 53, p. 16–38, 2015.

ALMEIDA, F. *et al.* Análise temporal de roubos e furtos a residência em Cuiabá, Brasil. **Revista Brasileira de Segurança Pública**, v. 17, n. 1, p. 208–231, 2023. Available at: <https://revista.forumseguranca.org.br/rbsp/article/view/1456>.

ALVES, L. G. A. *et al.* Spatial correlations, clustering and percolation-like transitions in homicide crimes. **EPL (Europhysics Letters)**, IOP Publishing, v. 111, n. 1, p. 18002, 2015. Available at: <http://dx.doi.org/10.1209/0295-5075/111/18002>.

CARNEIRO, L. d. A. Causas e consequências da criminalidade no Brasil: Uma revisão da literatura. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, v. 8, n. 7, p. 20–44, 2022. Available at: <https://periodicorease.pro.br/rease/article/view/6215>.

EVERITT, B.; LANDAU, S.; LEESE, M. **Cluster Analysis**. Wiley, 2001. (A Hodder Arnold Publication). Available at: <https://books.google.pt/books?id=htZzDGlCnQYC>.

EXAME. **São Paulo registra alta de homicídios e de estupros no primeiro trimestre**. 2025. Available at: <https://exame.com/brasil/sao-paulo-registra-alta-de-homicidios-e-de-estupros-no-primeiro-trimestre/>.

FONTALVO-HERRERA, T. J.; VEGA-HERNáNDEZ, M. A.; MEJíA-ZAMBRANO, F. Método de clustering e inteligencia artificial para clasificar y proyectar delitos violentos en colombia. **Revista Científica General José María Córdova**, v. 21, n. 42, p. 551–572, 2023. Available at: <https://revistacientificaesmic.com/index.php/esmic/article/view/1117>.

FORRADELLAS, R. *et al.* Applied machine learning in social sciences: Neural networks and crime prediction. **Social Sciences**, v. 10, 12 2020.

G1. **Número de homicídios e estupros sobe na cidade de SP no 1º trimestre de 2025; roubos e latrocínios caem**. 2025. Available at: <https://g1.globo.com/sp/sao-paulo/noticia/2025/05/01/numero-de-homicidios-e-estupros-sobe-na-cidade-de-sp-no-1o-trimestre-de-2025-roubos-e-latrocinios-c
ghtml>.

GOMES, L. D. *et al.* Crimes na era Covid-19 : evidências para o estado de São Paulo. **Revista Brasileira de Segurança Pública**, v. 17, n. 2, p. 370–393, 2023. Available at: <https://revista.forumseguranca.org.br/rbsp/article/view/1720>.

GONÇALVES, E.; LOPES, N. M. **Séries Temporais. Modelações lineares e não lineares**. 2008. Sociedade Portuguesa de Estatística. 2a edição. Disponível em: <https://spestatistica.pt/storage/app/uploads/public/600/482/db0/600482db02914978566173.pdf>.

GUSMãO, A.; CLEMENTE, T.; NEPOMUCENO, T. Optimizing police facility locations based on cluster analysis and the maximal covering location problem. **Applied System Innovation**, v. 5, p. 1–19, 07 2022.

HASSAN, B. A. *et al.* From A-to-Z review of clustering validation indices. **Neurocomputing**, Elsevier, v. 601, p. 128198, 2024. Available at: <https://arxiv.org/pdf/2407.20246.pdf>.

HUNTER, R.; DANTZKER, M. **Crime and Criminality: Causes and Consequences**. Lynne Rienner Publishers, 2012. Available at: <https://books.google.com.br/books?id=lveFZwEACAAJ>.

JAIN, A. K. Data clustering: 50 years beyond k-means. **Pattern Recognition Letters**, Elsevier, v. 31, n. 8, p. 651–666, 2010.

JAVED, A.; LEE, B.; RIZZO, D. A benchmark study on time series clustering. **Machine Learning with Applications**, v. 1, 09 2020.

JUNIOR, O. R. dos A. *et al.* Padrões de concentração espacial de roubos de automóveis em municípios da grande João Pessoa a partir de técnicas de aprendizado de máquinas. **Teoria e Prática em Administração**, v. 11, n. 2, p. 28–45, 2020. Available at: <https://periodicos.ufpb.br/ojs2/index.php/tpa/article/view/50891>.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: An Introduction to Cluster Analysis**. [*S.l.: s.n.*]: John Wiley & Sons, 2009. (Wiley Series in Probability and Statistics).

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In*: UNIVERSITY OF CALIFORNIA PRESS. **Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability**. [*S.l.: s.n.*], 1967. v. 1, p. 281–297.

MISSION, R. Spatial patterns of violence against women and children using geographic information system and density-based clustering algorithm. **International Journal of Emerging Technologies and Advanced Applications**, v. 1, p. 1–7, 05 2024.

OLIVEIRA, C. A. d.; SILVA, D. M. Os impactos do medo do crime sobre o consumo de atividades de lazer no Brasil. **Revista Brasileira de Segurança Pública**, v. 15, n. 1, p. 156–173, 2021. Available at: <https://revista.forumseguranca.org.br/rbsp/article/view/1179>.

PAPARRIZOS, J.; GRAVANO, L. k-shape: Efficient and accurate clustering of time series. **SIGMOD Rec.**, Association for Computing Machinery, v. 45, n. 1, p. 69–76, 2016. Available at: <https://doi.org/10.1145/2949741.2949758>.

PAPARRIZOS, J.; YANG, F.; LI, H. **Bridging the Gap: A Decade Review of Time-Series Clustering Methods**. 2024. Available at: <https://arxiv.org/abs/2412.20582>.

PARMEZAN, A. R. S.; BATISTA, G. E. A. P. A. **Descrição de Modelos Estatísticos e de Aprendizado de Máquina para Predição de Séries Temporais**. 2016. Relatório Técnico da Universidade de São Paulo, São Carlos. Available at: <https://repositorio.usp.br/item/002772986>.

PARMEZAN, A. R. S.; SOUZA, V. M. A. de; BATISTA, G. E. A. P. A. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. **Inf. Sci.**, v. 484, p. 302–337, 2019.

PETITJEAN, F.; KETTERLIN, A.; GANCARSKI, P. A global averaging method for dynamic time warping, with applications to clustering. **Pattern Recognition**, Elsevier, v. 44, n. 3, p. 678–693, 2011.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, Elsevier, v. 20, n. 1, p. 53–65, 1987. Available at: <https://wis.kuleuven.be/stat/robust/papers/publications-1987/rousseeuw-silhouettes-jcam-sciencedirectopenarchiv.pdf>.

SSP-SP. **Com média de 2,4 casos por mês, Grande SP encerra 2023 com menor número de latrocínios da história**. 2024. Acessado em 12 de março de 2025. Available at: <https://www.ssp.sp.gov.br/noticia/56688>.

STEINGRABER, R. Homicídios no brasil: análise do indivíduo no período 2006-2019. **Revista Brasileira de Segurança Pública**, v. 18, n. 1, p. 72–91, 2024. Available at: <https://revista.forumseguranca.org.br/rbsp/article/view/1744>.

TAN, P.-N. *et al.* **Introduction to Data Mining**. [*S.l.: s.n.*]: Pearson, 2018.

UOL. **Cidade de SP tem alta nos homicídios, na contramão do Estado; veja dados**. 2025. Available at: <https://noticias.uol.com.br/ultimas-noticias/agencia-estado/2025/04/30/cidade-de-sp-tem-alta-nos-homicidios-na-contramao-do-estado-veja-dados.htm>.

XU, R.; WUNSCH, D. **Clustering**. Wiley, 2008. (IEEE Press Series on Computational Intelligence). Available at: <https://books.google.com.br/books?id=kYC3YCyl_tkC>.

ZOU, H. Clustering algorithm and its application in data mining. **Wireless Personal Communications**, v. 110, n. 1, p. 21–30, 2020. Available at: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85070962369&doi=10.1007%2fs11277-019-06709-z&partnerID=40&md5=64ca04109f0f20e71cc8d217f57b6c15>.