Clusterização Espacial In-Database para Planejamento Estratégico de Unidades Policiais

In-Database Spatial Clustering for Strategic Planning of Police Units

Gustavo Rodrigues Carvalho
Universidade Tecnológica Federal do Paraná (UTFPR)
Avenida Sete de Setembro – 3165 – 80.230-901 – Curitiba – PR – Brasil
gu171994@gmail.com

Nádia Puchalski Kozievitch
Universidade Tecnológica Federal do Paraná (UTFPR)
Avenida Sete de Setembro – 3165 – 80.230-901 – Curitiba – PR – Brasil
nadiap@utfpr.edu.br

ABSTRACT

The rapid population growth and increasing urban density increase the complexity of managing challenges related to the unequal spatial distribution of crime, the temporal dynamics of incidents, resource limitations, and the operational constraints of security forces. This article presents an in-database spatial clustering approach for optimizing the location of police units using large-scale georeferenced data. The proposed method integrates Geographic Information Systems (SIG) with PostgreSQL/PostGIS, exploring the K-means algorithm executed directly within the spatial database to reduce data transfer overhead and preserve topological integrity. A total of 554,832 crime records from the metropolitan region of London were processed. The selection of the number of clusters was conducted through systematic experimentation varying K from 2 to 10, using Inertia (Elbow method), Silhouette Coefficient, and Calinski–Harabasz Index as evaluation metrics. The results indicated K=4 as the best trade-off between intra-cluster cohesion, inter-cluster separation, and territorial interpretability. The approach demonstrated computational feasibility and reproducibility, contributing as an architectural model for data-driven spatial analysis in large-scale urban scenarios and smart city applications.

CCS Concepts

• **Information systems→Database management system engines**.

Keywords

Crime; public safety; SIG; spatial analysis; urban centers.

RESUMO

O rápido crescimento populacional e a crescente densidade urbana aumentam a complexidade da gestão de desafios relacionados à distribuição espacial desigual da criminalidade, à dinamicidade temporal das ocorrências, às limitações de recursos e às restrições operacionais das forças de segurança. Este artigo apresenta uma abordagem de clusterização espacial *in-database* para otimização da localização de unidades policiais a partir de dados georreferenciados em larga escala. A proposta integra Sistemas de Informação Geográfica (SIG) ao *PostgreSQL/PostGIS*, explorando o algoritmo *K-means*, reduzindo overhead de transferência e preservando a integridade topológica das geometrias. Foram processados 554.832 registros criminais da região metropolitana de Londres. A definição do número de clusters foi conduzida por experimentação sistemática com variação de K entre 2 e 10, utilizando as métricas Inércia *(Elbow)*, Coeficiente de *Silhouette* e Índice de *Calinski-Harabasz*. Os resultados indicaram K=4 como solução de equilíbrio entre coesão *intra-cluster*, separação *inter-cluster* e interpretabilidade territorial.

Palavras-chave

Análise espacial; centros urbanos; criminalidade; segurança pública; SIG.

1. INTRODUÇÃO

A crescente complexidade dos problemas de segurança pública em grandes centros urbanos tem motivado autoridades e pesquisadores a buscar soluções baseadas em evidências para otimizar a alocação de recursos policiais. Estudos clássicos de *hotspot policing* demonstraram que a concentração de esforços em pequenas áreas de alta incidência criminal gera reduções significativas nos índices de crime da área total analisada, sem aumentar os custos operacionais de forma proporcional [29]. Pesquisas mostram que 58% de todos os crimes acontecem nos 10% de lugares com os crimes mais graves[1]. Nesse contexto, a aplicação de análise espacial e o uso de algoritmos analíticos vêm se consolidando como instrumentos-chave para subsidiar decisões estratégicas em segurança pública, permitindo a identificação de tendências espaciais, temporais e espaço-temporais da criminalidade, tais como a persistência ou deslocamento de *hotspots*, variações sazonais de determinados tipos de crime e padrões de concentração associados a características urbanas específicas, o apoio ao diagnóstico antecipado de crimes e a alocação eficiente de recursos, contribuindo para a inteligência das cidades [23].

No plano internacional, cidades como Nova York, Los Angeles e Londres têm investido pesadamente em plataformas de inteligência espacial. A *Metropolitan Police Service (MPS)* de Londres, por exemplo, disponibiliza bancos de dados abertos que incluem registros de ocorrências criminais georreferenciadas[2]. Essa abordagem integrada, baseada na combinação de dados abertos, ferramentas de Sistemas de Informação Geográfica, técnicas de análise espacial e métodos de visualização da informação, favorece a aplicação de técnicas de Visualização da Informação sobre esses

[1]https://youthendowmentfund.org.uk/toolkit/hot-spots-policing/

[2]https://www.met.police.uk/police-forces/metropolitan-police/areas/stats-and-data/stats-and-data/

dados, permitindo a criação de painéis e visualizações, como mapas de calor que apresentam a concentração dos atendimentos, auxiliando na tomada de decisão.

No Brasil, observa-se um movimento crescente em direção à adoção de metodologias quantitativas e geoespaciais no planejamento da segurança pública, ainda que sua implementação permaneça desigual, resultando em sobrecarga de algumas delegacias e subutilização de outras. Apesar de avanços pontuais na utilização de Sistemas de Informação Geográfica (SIG) e de ferramentas de *Visual Analytics* pelas forças policiais, a alocação de efetivos e recursos ainda é predominantemente reativa, com reduzido embasamento preditivo ou estratégico.

Por outro lado, a análise de grandes volumes de dados georreferenciados impõe desafios computacionais relacionados à escalabilidade, eficiência de processamento e preservação da integridade topológica das geometrias. Em aplicações urbanas orientadas a dados, a necessidade de processar centenas de milhares de registros espaciais exige arquiteturas capazes de integrar armazenamento, consulta e análise em um único ambiente computacional. Abordagens que dependem da exportação de dados para ferramentas externas de mineração ou aprendizado de máquina podem introduzir *overhead* de transferência e comprometer a consistência espacial. Nesse contexto, a execução de algoritmos analíticos diretamente em Sistemas Gerenciadores de Banco de Dados tem sido defendida como estratégia para garantir escalabilidade, desempenho e proximidade entre dados e processamento [13].

Na área de segurança pública, a análise espacial é amplamente utilizada para identificação de *hotspots* criminais e apoio ao planejamento territorial. Estudos clássicos demonstram que a criminalidade tende a se concentrar em pequenas áreas geográficas, fenômeno conhecido como *hotspot policing* [34]. Técnicas baseadas em Sistemas de Informação Geográfica (SIG) e estimativa de densidade *kernel* têm sido empregadas para mapear e interpretar esses padrões [12][17]. Mais recentemente, métodos de aprendizado de máquina e análise espaço-temporal vêm sendo explorados para previsão e priorização de áreas de risco [3][10].

Entre os algoritmos de agrupamento, o *K-means* permanece amplamente adotado devido à sua simplicidade, escalabilidade e eficiência computacional [22]. Para a avaliação da qualidade dos agrupamentos, métricas como o Coeficiente de *Silhouette* [28] e o Índice de *Calinski-Harabasz* [7] são frequentemente empregadas, permitindo mensurar coesão *intra-cluster* e separação *inter-cluster*. Embora métodos como *DBSCAN* e *Spectral Clustering* possibilitem a identificação de estruturas espaciais não convexas e padrões mais complexos [41][14], o *K-means* apresenta vantagens operacionais quando integrado a bancos de dados espaciais, especialmente em cenários caracterizados por grande volume de dados e necessidade de processamento eficiente.

Apesar desses avanços, observa-se uma lacuna na literatura quanto à adoção sistemática de estratégias *in-database* para clusterização espacial aplicada ao planejamento territorial em segurança pública. Grande parte dos estudos realiza o processamento analítico fora do ambiente do banco de dados, limitando a integração com consultas espaciais nativas e reduzindo a escalabilidade em bases massivas. São escassas as investigações que combinam execução nativa de algoritmos em *PostGIS*, avaliação comparativa de múltiplas métricas de validação e aplicação em bases abertas de larga escala com metodologia reprodutível.

Diante desse cenário, este trabalho propõe uma abordagem de clusterização espacial *in-database* para otimização da localização de unidades policiais, explorando a função *ST_ClusterKMeans* do *PostGIS* como mecanismo de execução nativa do algoritmo *K-means*. A proposta integra SIG e *PostgreSQL/PostGIS* em uma arquitetura unificada, eliminando a necessidade de exportação de dados para processamento externo. As principais contribuições deste estudo são: (i) implementação e avaliação de clusterização espacial executada diretamente em SGBD-E; (ii) análise sistemática da definição do parâmetro K por meio das métricas *Inertia*, *Silhouette* e *Calinski-Harabasz*; (iii) validação experimental com 554.832 registros criminais georreferenciados; e (iv) proposição de um modelo reprodutível para apoio ao planejamento territorial orientado a dados em cenários urbanos.

Este trabalho se diferencia do anterior [8], por aprofundar a investigação e ampliar os resultados obtidos utilizando métricas e metodologia mais atualizada, além de explorar mais o estado da arte.

2. FUNDAMENTAÇÃO TEÓRICA

O escopo desta proposta concentra-se em três eixos de pesquisa principais: Cidades Inteligentes, Banco de Dados e Algoritmos.

2.1 Banco de Dados

A origem do conceito de *hotspot policing* remonta aos estudos de [34], que demonstraram empiricamente que uma pequena parcela de locais concentra a maior parte dos crimes predatórios. Por meio da teoria das atividades rotineiras, que entende o crime como a convergência de um infrator, um alvo e a ausência de um guardião, os autores mostraram que policiamento focado em áreas com alta incidência criminal consegue reduzir a criminalidade sem aumento proporcional dos recursos empregados. Esse marco teórico motivou toda a linha de pesquisa sobre análise espacial de criminalidade e serviu de base para o desenvolvimento de sistemas de mapeamento de crimes.

Dentro da área de banco de dados, [12], sintetizam as principais técnicas de SIG aplicadas à segurança pública, detalhando métodos de geração de mapas de densidade (*kernel density estimation*), interpolação e análise de vizinhança, ressaltando a importância de testar diferentes parâmetros de suavização (*bandwidth*) e de validar espacialmente os resultados para evitar vieses operacionais. Esse trabalho também enfatizou a necessidade de cadências temporais, ou seja, a frequência definida com que os mapas são atualizados, abrindo caminho para sistemas dinâmicos de mapeamento em tempo real. Nesse contexto, a adoção de infraestruturas robustas de tratamento de dados torna-se essencial, [35] detalha como a integração entre sistemas de informação e bancos de dados espaciais permite o processamento automatizado de grandes volumes de dados, facilitando a extração de padrões e a sumarização de informações geoespaciais.

Em [21] pode-se observar a importância de SIG e análise espacial na identificação e interpretação de *hotspots* de crimes. Em consonância [32] detalha como essas ferramentas ajudam a mapear a distribuição dos crimes, enquanto [17] sugerem que a visualização de dados como mapas de densidade e pontos é crucial para a formulação de políticas e estratégias de segurança pública, essas visualizações ajudam a identificar padrões e alocar recursos de forma mais eficaz, conforme detalhado por [1].

Complementando a aplicação prática desses sistemas, [26] discutiu as lacunas de treinamento de agentes de segurança no uso de ferramentas de *crime mapping*. O artigo argumenta que o simples

acesso a um SIG não garante eficácia se operadores não compreendem questões como projeção de coordenadas, escolha de resolução espacial e interpretação de mapas de calor. O artigo ainda propôs módulos de capacitação que unificassem conhecimentos de criminologia ambiental com habilidades técnicas de SIG, promovendo a adoção crítica e informada das tecnologias. A execução de algoritmos analíticos diretamente no Sistema Gerenciador de Banco de Dados (SGBD) fundamenta-se no princípio de *data locality*, segundo o qual a computação deve ocorrer próxima aos dados, minimizando custos de movimentação e overhead de entrada e saída *(I/O)*. Trabalhos como [13] demonstram que a integração entre armazenamento e processamento analítico pode resultar em ganhos significativos de desempenho e escalabilidade.

Em ambientes espaciais, a exportação de grandes volumes de dados para ferramentas externas pode introduzir custos adicionais de serialização, perda de metadados geográficos e inconsistências topológicas.

A abordagem de extensões como o *PostGIS* reduz a necessidade de pipelines externos de processamento e favorece arquiteturas mais integradas, especialmente em cenários caracterizados por bases massivas e atualizações frequentes [13].

2.2 Algoritmos

A clusterização é uma técnica de aprendizado não supervisionado amplamente empregada para identificar estruturas latentes em conjuntos de dados multivariados [22]. Em contextos espaciais urbanos, a aplicação de algoritmos de agrupamento permite detectar padrões territoriais recorrentes e subsidiar processos de planejamento orientados a dados.

Entre os algoritmos particionais, o *K-means* permanece como uma das abordagens mais adotadas devido à sua simplicidade computacional e escalabilidade [22]. O algoritmo busca minimizar a soma das distâncias quadráticas entre cada ponto e o centróide de seu respectivo cluster, formalmente definida como:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - u_j||^2$$

Onde k representa o número de clusters, C_j o conjunto de pontos pertencentes ao *cluster* j, e u_j o centróide correspondente.

A complexidade temporal do *K-means* pode ser aproximada por O(n .k .i), em que n corresponde ao número de observações e i ao número de iterações até convergência. Em bases espaciais com centenas de milhares de registros, essa característica torna-se particularmente relevante sob a perspectiva de desempenho e escalabilidade.

Apesar de sua eficiência, o algoritmo apresenta limitações conhecidas, como sensibilidade à escolha inicial dos centróides, necessidade de definição prévia de k e tendência à formação de *clusters* convexos, o que pode restringir sua capacidade de modelar estruturas espaciais irregulares [41]. Ainda assim, sua eficiência computacional o torna adequado para cenários de grande volume de dados quando integrado a ambientes otimizados de processamento.

A determinação adequada do número de *clusters* constitui um dos principais desafios da clusterização particional. Entre as métricas internas de validação mais consolidadas destacam-se o Coeficiente de *Silhouette* [28] e o Índice de *Calinski-Harabasz* [7].

O Coeficiente de *Silhouette* para uma observação i é definido como:

$$s(i) = \frac{b(i) - a(i)}{max\,(a(i), b(i))}$$

Em que a(i) representa a distância média *intra-cluster* e b(i) a menor distância média ao cluster vizinho mais próximo. Valores próximos de 1 indicam melhor separação entre agrupamentos [28].

O coeficiente de *Silhouette* compara, para cada ponto, a distância média ao próprio *cluster* com a distância média ao *cluster* mais próximo, gerando um índice que varia de -1 (grupo mal definido) a +1 (grupo bem separado). A *Silhouette* fornece uma medida intuitiva de coesão interna e separação externa, sendo particularmente útil para comparar diferentes valores de K.

O Índice de *Calinski-Harabasz* é definido como:

$$CH = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)}$$

Onde Tr(B_k) corresponde à dispersão *inter-cluster* e Tr(W_k) à dispersão *intra-cluster*. Valores elevados indicam melhor separabilidade e compacidade [7].

Essas métricas permitem avaliar simultaneamente coesão interna e separação externa, sendo particularmente relevantes em aplicações espaciais nas quais a interpretabilidade territorial é requisito operacional.

O método *Elbow* é frequentemente empregado para fundamentar a escolha técnica do parâmetro K [25], identificando o ponto de inflexão onde o aumento do número de *clusters* não reduz significativamente a variância total ou a soma dos quadrados dos agrupamentos.

Estudos recentes destacam a relevância de diferentes técnicas de *clustering* para a análise de dados complexos e de alta dimensionalidade. Como detalhado no tutorial fundamental de [41], o *Spectral Clustering* se baseia na construção de grafos de similaridade para transformar o problema de agrupamento, permitindo a identificação de estruturas não lineares e agrupamentos não convexos. De forma complementar, [14] compararam algoritmos como *DBSCAN* e *HDBSCAN*, evidenciando que o *DBSCAN* é eficiente na identificação de clusters com formas arbitrárias e na detecção de ruídos, enquanto o *HDBSCAN* aprimora essa abordagem ao ajustar automaticamente parâmetros e oferecer maior flexibilidade na detecção hierárquica de agrupamentos. Juntas, essas análises, baseadas em métodos de *clustering espectral*, em técnicas de agrupamento por densidade e em estudos comparativos de algoritmos, reforçam a necessidade de selecionar métodos adequados ao tipo, à densidade e à estrutura dos dados, especialmente em cenários de análise espacial urbana e criminal, onde os padrões podem ser irregulares, multifacetados e de distribuição heterogênea.

Do ponto de vista computacional, além da análise exploratória, outros métodos, como modelos de regressão e séries temporais já foram utilizados para a análise de crime [19], por exemplo, para predizer a quantidade de crimes em determinadas regiões e períodos do ano, utilizando o modelo *ARIMA*.

2.3 Cidades Inteligentes

O conceito de cidades inteligentes está associado à capacidade de integrar infraestruturas físicas e digitais por meio de Tecnologias da Informação e Comunicação (TIC), viabilizando monitoramento, análise e suporte à decisão baseados em dados [46][5]. Diferentemente de abordagens meramente conceituais, a

perspectiva contemporânea enfatiza a construção de ecossistemas computacionais capazes de processar grandes volumes de dados urbanos provenientes de múltiplas fontes, incluindo sensores *IoT*, sistemas administrativos e bases georreferenciadas.

Nesse contexto, plataformas de dados urbanos assumem papel central, consolidando informações espaciais e temporais em ambientes integrados, frequentemente estruturados sobre Sistemas de Informação Geográfica (SIG) e bancos de dados espaciais [4]. A convergência entre armazenamento geoespacial e análise computacional permite a construção de modelos analíticos voltados ao planejamento territorial e à otimização de serviços públicos.

A segurança pública constitui uma das dimensões críticas das cidades inteligentes, demandando mecanismos analíticos capazes de identificar padrões espaciais e apoiar a alocação eficiente de recursos [34]. Técnicas de análise espacial, como estimativa de densidade e agrupamento territorial, têm sido empregadas para identificação de áreas de concentração delitiva e suporte ao policiamento orientado a dados [12].

Entretanto, a escalabilidade das aplicações depende diretamente da infraestrutura computacional subjacente. Em ambientes urbanos caracterizados por bases massivas e atualização contínua, arquiteturas que integram armazenamento e processamento analítico reduzem *overhead* de movimentação de dados e favorecem eficiência operacional [13]. Dessa forma, a adoção de técnicas de clusterização espacial executadas diretamente em bancos de dados geográficos alinha-se aos princípios de cidades inteligentes orientadas por dados, integrando modelagem analítica, persistência geoespacial e suporte à decisão em um único ambiente computacional.

A análise espacial da criminalidade tem sido amplamente empregada para identificação de áreas de concentração delitiva, frequentemente denominadas *hotspots* [34]. Técnicas baseadas em Sistemas de Informação Geográfica (SIG) e estimativa de densidade *kernel* são tradicionalmente utilizadas para mapeamento desses padrões [12].

Mais recentemente, abordagens baseadas em aprendizado de máquina têm sido incorporadas ao planejamento territorial e à priorização de patrulhamento [3]. Contudo, grande parte dessas aplicações realiza o processamento analítico fora do ambiente transacional, o que pode comprometer escalabilidade e integração com consultas espaciais complexas.

Nesse contexto, a adoção de clusterização espacial executada diretamente em SGBD-E representa uma alternativa arquitetural alinhada aos princípios de computação aplicada, integrando modelagem analítica, armazenamento geoespacial e eficiência operacional em um único ambiente computacional.

3. REVISÃO SISTEMÁTICA DA LITERATURA

Esta seção apresenta uma Revisão Sistemática da Literatura (RSL) com o objetivo de identificar e analisar padrões temporais, ocorrência, utilização de SIG e desafios na previsão de crimes.

A condução da RSL seguiu diretrizes metodológicas para revisões sistemáticas em computação [18]. Foram definidas as seguintes questões de pesquisa:

•**QP1:** Onde ocorrem os crimes com maior frequência?

•**QP2:** Quais são os padrões temporais na ocorrência de crimes?

•**QP3:** Como o SIG e dados geoespaciais podem ser usados para análise e previsão de crimes?

•**QP4:** Quais são os desafios associados à visualização e previsão de crimes?

Para esta RSL, utilizamos a fonte de dados IEEE Xplore[3], em idioma inglês, com a seguinte *string* de busca:

((“patrolling” OR “policing” OR “law enforcement” OR “crime” OR “delict”) AND (“geoprocessing” OR “geolocation” OR “spatial analysis” OR “geographic data processing” OR “GIS” OR “geospatial analysis” OR “spatial modeling” OR “location tracking” OR “GPS tracking” OR “spatial location” OR “geographic positioning” OR “location detection”)).

Os critérios foram definidos com base no objetivo da revisão, representados na Tabela 1.

Tabela 1. Critérios de Inclusão e Exclusão.

Inclusão	
CI1	Estudos que respondem às perguntas de pesquisa
CI2	Estudos publicados a partir de 2014
Exclusão	
CE1	Estudos duplicados
CE2	Indisponibilidade do texto completo
CE3	Livros, editoriais, tutoriais, painéis, sessões de pôsteres, prefácios, opiniões, resumos, cartas, apresentações de slides, relatórios técnicos, trabalho com menos de 5 páginas, ou qualquer trabalho classificado como literatura cinza

A busca na plataforma *IEEE* resultou em 1685 artigos, os quais foram ordenados por relevância. Os dados obtidos foram organizados em uma planilha, na qual foram aplicados filtros baseados em critérios previamente definidos. Para os artigos selecionados, seus resumos foram traduzidos e submetidos à leitura inicial. Após essa triagem, 23 artigos foram considerados adequados para leitura integral, de acordo com os critérios estabelecidos.

3.1 Análises e Resultados

Vinte e três artigos foram selecionados para a análise completa, satisfazendo os critérios Tabela 2.

Tabela 2. Artigos Selecionados da RSL.

[3] https://ieeexplore.ieee.org/Xplore/home.jsp

N°	Autor(es), Ano	Título do Artigo
1	Crime Analysis of spatial-temporal distribution based on KNN Algorithm	J. Wang; S. Zhang; Y. Lan; C. Wu; Y. Xia; L. Chen
2	IST: Role of GIS in crime mapping and analysis	J. Sheikh; I. Shafique; M. Sharif; S. A. Zahra; T. Farid
3	Crime Hotspots Mapping and FIR Data Interface	S. G. Lilhare; Y. Kumavat; G. Banait; A. Kurkelli
4	A predictive policing application to support patrol planning in smart cities	A. Araujo; N. Cacho; A. C. Thome; A. Medeiros; J. Borges
5	Fuzzy Based Geo-Spatial Crime Category Prediction for Crime Mapping and Safe Route Travel	S. Sharma; A. Uniyal; P. Srinivasan; S. Chaudhari
6	Comparative Analysis Of Crime Hotspot Detection And Prediction Using Convolutional Neural Network Over Support Vector Machine with Engineered Spatial Features Towards Increase in Classifier Accuracy	T. Sravani; M. R. Suguna
7	Impact of Spatial Correlation on Crime Prediction in Communities with Different Crime Densities	Q. Dong; R. Ye; Y. Fan
8	A Comprehensive Review on Crime Patterns and Trends Analysis using Machine Learning	A. Ratra; A. Agarwal; S. Vats; V. Sharma; V. Kukreja; S. P. Yadav
9	Micro-Level Incident Analysis using Spatial Association Rule Mining	J. S. Yoo; S. J. Park; A. Raman
10	Prediction of Crime in Neighbourhoods of New York City using Spatial Data Analysis	A. A. Almuhanna; M. M. Alrehili; S. H. Alsubhi; L. Syed
11	Crime Detection Using Data Mining Techniques	A. Sayal; A. Gupta; C. Vasundhara; Y. B M; V. Gupta; H. Maheshwari
12	Smart Policing: Using Geospatial Crime Data to Plan Patrol Routes	A. D A; A. S; A. K; A. G. Shenoy; P. K. Auradkar
13	Space-Time Variation of Property Crime in Beijing with ESDA Method	S. Li; W. Tang
14	Crime sequencing: Fighting crime with mathematics and technology	J. Steier; A. Zigarelli; E. Giannini; M. Minimair
15	Time series analysis for crime forecasting	G. Borowik; Z. M. Wawrzyniak; P. Cichosz
16	Crime Prediction Using Auto Regression Techniques for Time Series Data	R. Yadav; S. Kumari Sheoran
17	Visual Analysis of Predictive Policing to Improve Crime Investigation	M. Sathiyanarayanan; A. K. Junejo; O. Fadahunsi
18	Bandwidth Selection of Kernel Density Estimation for GIS-based Crime Occurrence Map Visualization	Y. Kim; G. Kim; Y. Lee; K. Jang
19	Predicting High-Risk Areas for Crime Hotspot Using Hybrid KNN Machine Learning Framework	V. K; R. K. S; V. R. R; N. Mekala; S. P. Sasirekha; R. Reshma
20	The Role of Machine Learning in Crime Analysis and Prediction	M. Geetha Vadav; R. N; E. S. Reddy; M. S. Vishal; G. Vishal
21	An Approach on Cyber Crime Prediction Using Prophet Time Series	A. Nag; R. Ranjan; C. N. S. Vinoth Kumar
22	Geographical crime rate prediction	S. Jain; Riyank
23	Geographical Crime Rate Prediction System	S. Tarlekar; R. Bhosle; E. D'souza; S. Sheikh

Considerando onde ocorrem os crimes com maior frequência **(Pergunta 1)**, o artigo de [29] identifica *hotspots* criminosos por meio de análises espaciais, revelando áreas com alta concentração de crimes. Esse método é corroborado pelos artigos de [43] e [37], que destacam como a análise de densidade e agrupamento podem evidenciar essas áreas de alta incidência. Além disso, o artigo de [39] reforça a ideia de que a análise espacial é crucial para detectar e entender a frequência dos crimes em diferentes regiões. Já para identificar e mapear áreas de alta incidência de crimes, os artigos de [43], [37] e [39] discutem técnicas para mapear *hotspots*, como a estimativa de densidade de *Kernel (KDE)* e análise de agrupamento. O artigo de [43] foca na visualização dos dados criminais, enquanto o artigo de [37] detalha como o *KDE* pode destacar áreas críticas. O artigo de [39] amplia essa perspectiva ao combinar diversas técnicas analíticas para uma visualização completa das áreas com alta incidência de crimes. O artigo de [37] observa que criminosos tendem a seguir um padrão que se estabelece entre 1 a 5 milhas de suas residências, com o último local de crime frequentemente situado próximo de sua residência. Esse padrão é apoiado pelos artigos de [36], [27] e [20], que discutem como a análise de movimento dos criminosos e padrões espaciais ajuda a prever futuras ocorrências baseadas na proximidade dos locais de crime. E quanto às áreas que são mais propensas a crimes, o artigo de [37] identifica áreas próximas a lojas de bebidas, transportes públicos e lojas de conveniência como mais propensas a crimes. O artigo de [29] complementa essa visão ao mostrar que a concentração de crimes também está associada à presença desses estabelecimentos em *hotspots* identificados. O artigo de [39] reforça essa relação ao mapear a incidência de crimes em áreas com alta atividade comercial. O artigo de [37] sugere que criminosos provavelmente utilizam transporte público para facilitar a fuga. Essa ideia é corroborada pelo artigo de [33], que analisa como o acesso a transportes públicos pode influenciar a mobilidade dos criminosos e suas escolhas de locais para cometer crimes. O artigo de [37] recomenda aumentar as patrulhas em áreas de alta criminalidade e utilizar câmeras para monitorar suspeitos. Essa recomendação é apoiada pelo artigo [39], que sugere o uso de dados para orientar a alocação de recursos policiais em pontos críticos e melhorar a eficácia das operações de prevenção.

Ao se tratar de padrões temporais na ocorrência de crimes **(Pergunta 2)**, os artigos de [43], [11] e [24] identificam padrões de sazonalidade e variações diárias na ocorrência de crimes. O artigo [11] usa dados geográficos para mostrar como esses padrões se manifestam ao longo do tempo, enquanto o artigo de [24] detalha como esses padrões podem variar com base em fatores sazonais e horários específicos. Os artigos de [36], [45] e [24] indicam que crimes tendem a ocorrer em locais de alta atividade durante horários específicos, com picos noturnos e finais de semana. O artigo de [45] oferece uma análise detalhada sobre como esses horários influenciam a localização dos crimes, alinhando-se com os padrões temporais identificados nos artigos de [36] e [24]. Já com relação a previsão de crimes futuros, os artigos de [32], [11], [20], [24] e [15] exploram como combinar dados espaciais e temporais para criar modelos preditivos. O artigo de [20] destaca a eficácia dessa combinação na previsão de áreas e horários de alto risco, enquanto o artigo de [15] analisa a influência desses padrões na criação dos modelos. Considerando a remoção de outliers e a inclusão de feriados nas previsões, o artigo de [24] avalia como há *outliers* e feriados no modelo, considerando esta abordagem crucial para ajustar modelos e garantir que as previsões considerem todas as variáveis relevantes, como os efeitos de eventos sazonais e datas comemorativas.

Os artigos de [32], [21], [33] e [17] mostram a importância de SIG e análise espacial na identificação e análise de *hotspots* de crimes **(Pergunta 3)**. Os artigos de [32] e [33] detalham como essas ferramentas ajudam a mapear a distribuição dos crimes, enquanto o artigo de [17] demonstra como a visualização de dados geoespaciais pode auxiliar nas políticas de segurança pública. Os artigos de [3] e [11] mostram como a análise espacial é essencial para direcionar recursos de forma eficaz e melhorar a resposta a áreas de maior risco. E [17] sugere que a visualização de dados, como mapas de densidade e pontos, é crucial para a formulação de políticas e estratégias de segurança pública. Essas visualizações ajudam a identificar padrões e alocar recursos de forma mais eficaz, conforme detalhado pelos artigos de [32] e [33].

Ao considerar desafios **(Pergunta 4)**, itens como a qualidade dos dados, a diversidade de fontes e questões éticas na análise de dados criminais é mencionada [31] [30] [6]. Nesta direção, modelos preditivos são essenciais para identificar áreas de alto risco e planejar patrulhas mais eficazes [3] [2] [33] [37]. O artigo de [2] detalha como essas previsões podem direcionar a alocação de recursos, enquanto o artigo de [37] recomenda estratégias para aumentar a eficácia do policiamento em áreas identificadas como de alta criminalidade. Algoritmos de *machine learning*, como *Random Forest*, *KNN*, *ARIMA* e modelos de aprendizado profundo podem ser utilizados [36] [2] [30] [40] [15] [39]. O artigo de [30] compara a eficácia desses algoritmos, mostrando que *Random Forest* e métodos de aprendizado profundo frequentemente oferecem maior precisão em comparação com outros métodos. Porém, estes algoritmos têm limitações, como a qualidade dos dados, ajuste de hiperparâmetros e utilização de técnicas avançadas [36] [11] [6], essas abordagens visam aumentar a precisão e a eficácia dos modelos preditivos de crimes. E quanto à alocação estratégica de recursos policiais, [2] [33] [44] [40], a previsão é um método mais utilizado. O artigo de [44] oferece diretrizes práticas para a alocação de recursos, enquanto o artigo de [40] analisa como utilizar dados preditivos para otimizar a resposta policial e a prevenção de crimes.

Os 23 artigos foram classificados entre cinco temas de aplicação, como ilustra a Tabela 3. Os artigos também apresentam estudos empíricos e teóricos Tabela 4, divididos em três abordagens metodológicas Tabela 5: Análise espacial e Visualização de Dados, *Machine Learning* e Algoritmos de Previsão e Análise Temporal e Sazonalidade. As áreas de aplicação encontradas foram Segurança Pública e Policiamento e Tecnologias e Ferramentas de Análise Tabela 6.

As quatro tabelas apresentadas a seguir nos auxiliam a elucidar, como se classificam os artigos com base nas características em comum identificadas, sendo elas: temas e áreas de aplicação representadas pela Tabela 3, tipos de estudos contidos na Tabela 4, abordagens metodológicas listadas na Tabela 5 e áreas de aplicação na Tabela 6.

Tabela 3. Temas e áreas de aplicação.

Temas e Áreas de Aplicação	Artigos	Descrição
Identificação e Mapeamento de Hotspots de Crimes	1, 3, 4, 6, 8, 12, 13, 14, 17, 19, 23	Estes artigos concentram-se na identificação de áreas com alta incidência de crimes e na aplicação de técnicas para mapear hotspots. Eles abordam o uso de técnicas de análise espacial, visualização de dados, e métodos de clustering para identificar e mapear áreas críticas.
Padrões Temporais e Espaciais dos Crimes	1, 2, 4, 6, 7, 9, 10, 13, 14, 15, 21, 22	Esta categoria abrange estudos que investigam como os crimes ocorrem em relação ao tempo e ao espaço. Inclui a análise de sazonalidade, variações diárias, e padrões de crime em diferentes horários e épocas do ano.
Métodos e Algoritmos de Previsão	1, 4, 6, 7, 10, 11, 15, 16, 19, 20, 21, 22	Artigos que discutem diferentes algoritmos e métodos de machine learning aplicados à previsão de crimes. Incluem comparações entre algoritmos e análises sobre como a engenharia de características e o pré-processamento afetam o desempenho dos modelos.
Utilização de Dados Geoespaciais e GIS	2, 3, 4, 6, 7, 9, 12, 18	Foca na aplicação de sistemas de informações geográficas (GIS) e dados geoespaciais na análise e previsão de crimes. Inclui a utilização de GIS para mapear e analisar a distribuição de crimes e a importância das características geográficas na previsão.
Desafios e Limitações na Previsão de Crimes	5, 6, 7, 11, 15, 16, 22	Aborda os desafios e limitações encontradas na visualização e previsão de crimes. Inclui questões relacionadas à qualidade dos dados, diversidade de fontes, e problemas éticos na análise de dados.
Planejamento e Alocação de Recursos Policiais	4, 10, 12, 14, 16, 20	Explora como modelos preditivos podem ser usados para otimizar o planejamento de patrulhamento e a alocação de recursos policiais. Inclui estratégias para melhorar a eficiência e eficácia das operações policiais.

Tabela 4. Tipos de estudos.

Tipos de Estudos	Artigos	Descrição
Estudos Empíricos	1, 2, 3, 6, 7, 8, 12, 13, 14, 15, 19, 21, 22	Artigos que baseiam suas conclusões em dados reais e experimentos. Incluem estudos de caso, análises de dados e experimentos práticos para validar teorias e métodos.
Estudos Teóricos	4, 5, 9, 10, 11, 16, 20	Focam na discussão de teorias, modelos e conceitos relacionados à previsão e análise de crimes. Incluem revisões de literatura, propostas de novos modelos e discussões conceituais sobre métodos e algoritmos.

Tabela 5. Abordagens metodológicas.

Abordagens Metodológicas	Artigos	Descrição
Análise Espacial e Visualização de Dados	1, 3, 4, 6, 12, 13, 14, 17, 18	Utilizam técnicas de análise espacial e visualização para mapear hotspots e identificar padrões de crimes. Incluem métodos como KDE e análise de clustering.
Machine Learning e Algoritmos de Previsão	1, 4, 6, 7, 10, 11, 15, 16, 19, 20, 21, 22	Aplicam algoritmos de machine learning para prever a ocorrência de crimes. Incluem técnicas como Random Forest, KNN, ARIMA, e modelos de aprendizado profundo.
Análise Temporal e Sazonalidade	1, 2, 7, 9, 10, 13, 14, 15, 21	Investigam como os crimes variam ao longo do tempo, considerando padrões sazonais e horários específicos. Incluem a análise de variações diárias e sazonais.

Tabela 6. Áreas de aplicação.

Áreas de Aplicação	Artigos	Descrição
Segurança Pública e Policiamento	4, 10, 12, 14, 16, 20	Focam na aplicação dos resultados da previsão de crimes para melhorar o policiamento e alocação de recursos. Incluem estratégias para otimização de patrulhamento e resposta policial.
Tecnologias e Ferramentas de Análise	2, 3, 4, 6, 7, 12, 18	Discussão sobre o uso de tecnologias como GIS e ferramentas de visualização para análise de crimes. Incluem a aplicação de tecnologias emergentes na análise espacial.

Foram identificadas algumas categorias e subcategorias recorrentes nos artigos conforme representa a Tabela 7.

Tabela 7. Categorias.

Categoria	Subcategoria	Número de Artigos
Métodos e Algoritmos	Machine Learning	8
	Análise Estatística	5
Análise Espacial	GIS e Dados Geoespaciais	7
	Visualização de Dados	4
Desafios e Limitações	Qualidade dos Dados	6
	Questões Éticas	3

3.2 Discussão

Os artigos revisados oferecem uma visão abrangente sobre a identificação de *hotspots* de crimes, destacando a importância das técnicas de análise espacial e agrupamento para mapear áreas com alta incidência de criminalidade. O artigo de [37] destaca que a análise de padrões cronológicos e a proximidade com a residência dos criminosos pode revelar informações valiosas para a prevenção de crimes. Além disso, a descoberta de que áreas comerciais tendem a ter mais crimes do que áreas residenciais sugere a necessidade de intervenções específicas. É importante notar que a literatura ainda carece de estudos que integrem completamente dados espaciais e temporais para prever e entender melhor a dinâmica dos *hotspots* criminosos.

Os padrões temporais e espaciais de crimes, abordados em artigos como [43], [11], e [24], revelam variações significativas ao longo do tempo e em diferentes horários. No entanto, ainda há uma lacuna significativa na integração desses padrões para construir modelos preditivos mais robustos. Artigos como [32] e [20] indicam que a combinação de dados espaciais e temporais pode melhorar a previsão de futuros crimes, mas a aplicação prática dessa combinação ainda é um desafio. A literatura sugere que a inclusão de variáveis temporais e a consideração de eventos sazonais são áreas emergentes que precisam de mais desenvolvimento para aprimorar a eficácia das previsões.

A eficácia dos métodos e algoritmos de previsão de crimes é amplamente discutida em artigos como [36], [2], e [15]. A comparação entre diferentes algoritmos, como *Random Forest* e *KNN*, revela que métodos mais avançados, como aprendizado profundo, frequentemente superam outros em termos de precisão conforme [30]. No entanto, a literatura ainda precisa de uma comparação mais abrangente e sistemática dos diferentes algoritmos. A seleção do valor de 'K' no *KNN*, conforme sugerido pelos artigos de [43] e [3], e a engenharia de características conforme [15] são cruciais para o desempenho dos modelos preditivos, mas muitas vezes são negligenciadas.

O uso de SIG e dados geoespaciais é fundamental para a análise e previsão de crimes, conforme discutido nos artigos de [32], [21], [33], e [17]. No entanto, a literatura revela desafios na integração de dados geoespaciais com outras variáveis e na visualização eficaz

desses dados para decisões estratégicas. Artigos como [21] e [17] destacam a importância da visualização avançada de dados, como mapas dinâmicos, para a formulação de políticas de segurança pública.

Os modelos preditivos desempenham um papel crucial no planejamento e na alocação de recursos policiais, como mostrado nos artigos de [3], [2], [33], e [37]. No entanto, a literatura ainda não fornece uma abordagem detalhada sobre como alocar recursos com base nas previsões de forma eficiente. A integração de diferentes fontes de dados e a otimização da alocação de recursos são áreas emergentes que necessitam de mais pesquisa para aprimorar as estratégias de policiamento conforme [2], [44] e [40].

Os desafios associados à visualização e previsão de crimes, conforme discutido em artigos como [31], [30], e [6], incluem questões relacionadas à qualidade dos dados, diversidade de fontes e considerações éticas. Embora haja avanços significativos, a literatura ainda carece de estudos que abordem como superar limitações nos modelos existentes. Artigos de [36], [11], e [6] sugerem que melhorias na qualidade dos dados e ajustes nos hiperparâmetros são essenciais para superar essas limitações. Além disso, questões éticas relacionadas à privacidade e ao uso responsável dos dados, como destacado pelo artigo de [30], são uma preocupação crescente que precisa ser abordada com mais rigor.

4. CASO DE USO – DADOS DA REGIÃO METROPOLITANA DE LONDRES

Esta seção descreve a abordagem proposta para clusterização espacial *in-database*, com um caso de uso prático, utilizando dados abertos da cidade de Londres. O objetivo é identificar os três locais mais adequados para a instalação de novas unidades policiais.

4.1 Modelagem do Problema

Seja D={x_1,x_2,…,x_n} o conjunto de ocorrências criminais georreferenciadas, onde cada ponto x_i ∈ R^2 representa coordenadas espaciais lat,lon.

O objetivo é particionar D em k *clusters* C={C_1,C_2,…,C_k} de modo a minimizar a soma das distâncias quadráticas *intra-cluster*:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - u_j||^2$$

Onde u_j representa o centróide do *cluster* C_J.

A otimização é realizada por meio do algoritmo *K-means* [22], cujo custo computacional é da ordem de O(nki), onde i representa o número de iterações até convergência.

4.2 Metodologia

A implementação foi realizada diretamente no PostgreSQL com extensão *PostGIS*, explorando capacidades de processamento espacial nativo. Diferentemente de abordagens externas, nas quais os dados são exportados para ferramentas analíticas, a execução *in-database* reduz *overhead* de transferência e preserva índices espaciais [13].

A arquitetura proposta segue o paradigma de *bringing computation to the data*, amplamente discutido na literatura de bancos analíticos modernos [38].

As etapas utilizadas são apresentadas na Figura 1. O processo envolve a coleta e o tratamento de dados abertos, a aplicação de filtros espaciais, técnicas de clusterização e o cálculo de centróides para sugerir pontos estratégicos de cobertura.

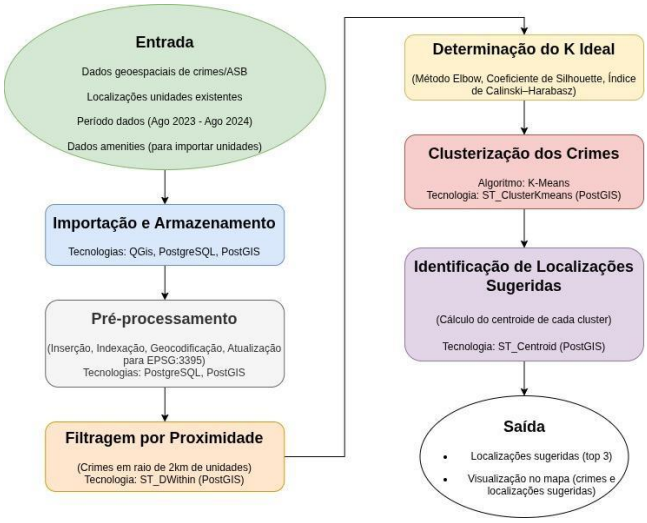

Figura 1. Fluxograma da metodologia.

4.3 Base de Dados e Infraestrutura

A base de dados utilizada para a pesquisa refere-se a dados de crimes individuais e incidentes de comportamentos antissociais *(ASB Anti-social Behaviour)* da região metropolitana de Londres, atendida pela *Metropolitan Police Service (MPS)*[44*], incluindo informações de localização em nível de rua e os resultados subsequentes de ações policiais e judiciais associadas ao crime. No período de Agosto-2023 a Agosto-2024, totalizando 554.832 registros, a Figura 2 apresenta cada coluna contida na tabela referente aos crimes. Os dados fornecidos são anonimizados, motivo pelo qual as ocorrências são agrupadas em meses, a localização é aproximada e o tipo de crime é enquadrado em categorias. O segundo tipo de dado (unidades policiais) foi importado das amenities[5] para o banco de dados utilizando o *QGis 3.22.4 64 bits*[6], selecionando a área tratada no estudo.

Na etapa de importação e armazenamento, foram utilizados o *PostgreSQL 17.2*[7], o *PostGIS 3.5*[8], e o *QGIS*.

[4]https://data.police.uk/data/fetch/0a319eae-59b1-47d6-adcf-b485883123b3

[5]https://overpass-api.de/

[6]https://www.qgis.org/project/visual-changelogs/visualchangelog322

[7]https://www.postgresql.org/docs/17/index.html

[8]https://postgis.net/2024/09/PostGIS-3.5.0

id	Identificador incremental único de cada registro, gerado via planilha
crime_id	Identificador do crime fornecido pela base de dados (nem sempre há valor)
month	Mês e ano em que a ocorrência foi registrada
longitude	Longitude anonimizada de onde o crime ocorreu
latitude	Latitude anonimizada de onde o crime ocorreu
location	Nome da localizacao aproximada em que o crime ocorreu
lsoa_code	Código de referência à LSOA na qual o ponto anonimizado se enquadra
lsoa_name	Nome de referência à LSOA na qual o ponto anonimizado se enquadra
crime_type	Um dos tipos de crimes listados nas Perguntas Frequentes do Police.UK.
last_outcome_category	Uma referência ao resultado associado ao crime ocorrido mais recentemente
geom	Geocodificação gerada a partir da latitude e longitude

Figura 2. Colunas da tabela de crimes.

Dentre os principais problemas com os dados, pode-se citar (apontados pela *Metropolitan Police Service (MPS)*[9]):

•**Precisão da localização:** Políticas inconsistentes de geocodificação nas forças policiais impedem que se tenha plena confiança de que os dados de localização fornecidos sejam totalmente precisos e consistentes. Isso é particularmente relevante em crimes cujo local exato não é conhecido, seja porque o fato ocorreu em um ponto não contemplado no sistema de *gazetteer* da força policial, que se trata do banco de dados oficial que cataloga e localiza endereços e nomes de lugares ou porque a vítima não tem certeza de onde ocorreu. Diferenças na qualidade desses sistemas de referência geográfica também constituem fator determinante. Estimativas da precisão da geocodificação entre diferentes forças variam de 60% a 97%.

•**Correspondência de resultados judiciais:** Não existe um identificador único para crimes que acompanhe o caso desde o serviço policial, passando pelo *Crown Prosecution Service (CPS)*, até os tribunais. Tal ausência torna praticamente impossível o rastreamento automático de um crime ao longo de todo o Sistema de Justiça Criminal. Para contornar essa limitação, utiliza-se um processo de *fuzzy matching*, com taxas de sucesso que variam entre 19% e 97%, dependendo da região em que o crime ocorreu.

•**Contagem dupla de incidentes de comportamento antissocial (ASB) e crimes:** Há indícios de que diferentes forças policiais possam estar duplicando determinados tipos de incidentes de ASB em seus registros.

•**Dados em constante atualização:** As informações enviadas pelas forças policiais a este sistema correspondem a um retrato pontual, capturado no final de determinado mês, para os crimes registrados. É possível que alguns sejam posteriormente reclassificados como outro tipo de infração ou confirmados como denúncia falsa após investigação. De forma semelhante, a localização de um crime pode ser alterada no sistema de tecnologia da informação de origem à medida que novas informações são obtidas. Na maioria dos casos, essas alterações posteriores não chegam ao sistema, a menos que a força policial opte por realizar uma atualização completa da base de dados, procedimento relativamente raro.

•**Ausência de dados sobre desfechos:** Nem a *British Transport Police* nem o *Police Service of Northern Ireland* disponibilizam dados relativos aos desfechos das ocorrências, embora haja tratativas com ambas as instituições para viabilizar tal disponibilização, desafios técnicos indicam que a solução ainda demandará tempo para ser implementada.

Na etapa de pré-processamento, um total de 554.832 registros foram inseridos, indexados, geocodificados e atualizados para a projeção cartográfica *EPSG:3395 (WGS 84/Pseudo Mercator)* na base de dados. Considerando a limitação de gastos e a otimização de recursos em cidades, o objetivo era responder à seguinte pergunta: Quais são as três localizações mais adequadas para a instalação de uma nova unidade policial, considerando a distribuição e a quantidade de crimes na região?

A abordagem utilizada possui as seguintes limitações:

•O raio de abrangência das delegacias foi limitado a 2 km, considerando a rapidez e eficiência da abordagem de policiais.

•No processo de clusterização, foi utilizado K=4 (após vários testes), devido aos agrupamentos moderadamente definidos.

•Os dados, considerando o âmbito temporal, são a nível de mês e ano.

•A escolha de três novas localizações foi adotada como parâmetro experimental, buscando representar um cenário plausível de expansão operacional. Esse número reflete uma meta mínima hipotética de cobertura territorial, suficiente para demonstrar o funcionamento e a aplicabilidade da metodologia proposta.

4.4 Determinação do Número de Clusters

A definição de k foi conduzida por experimentação sistemática variando k ∈[2][10]. Foram utilizadas três métricas de validação: Inércia *(Elbow)*, Coeficiente de *Silhouette* e Índice de *Calinski-Harabasz*.

Em uma primeira etapa, foram selecionadas as ocorrências de crimes em um raio limite de 2 km das delegacias existentes, por meio da função *ST_DWithin*. Dessa forma, foi possível identificar e filtrar, de maneira objetiva, os registros situados dentro da área de influência definida para cada unidade policial, estabelecendo um critério espacial consistente para a análise de cobertura territorial.

Em seguida, foi utilizado o algoritmo *K-means*, por meio da função *ST_ClusterKmeans* do *PostGIS*, com o objetivo de agrupar as ocorrências criminais georreferenciadas em *clusters* espaciais e identificar regiões com maior concentração de eventos, subsidiando a análise de padrões espaciais e a proposição de áreas candidatas para a instalação de novas unidades policiais. Essa função implementa o método de particionamento *K-means* diretamente sobre conjuntos de geometrias, atribuindo cada ocorrência ao cluster cujo centróide se encontra à menor distância, de modo a minimizar a variabilidade interna de cada grupo e maximizar a separação entre os diferentes agrupamentos. Como resultado, obtêm-se *clusters* espacialmente coerentes, representados por seus respectivos centróides, que sintetizam a distribuição espacial das ocorrências e facilitam a interpretação dos padrões de concentração criminal no território analisado.

Os resultados obtidos foram analisados de forma comparativa por meio de gráficos, permitindo avaliar o comportamento de cada métrica em função de K. A análise da curva de inércia evidenciou um ponto de inflexão em K=3, caracterizando o "cotovelo" apresentado na Figura 3. O Coeficiente de *Silhouette* apresentou seu maior valor médio em K=2, conforme mostrado na Figura 4, enquanto o Índice de *Calinski–Harabasz* atingiu seu pico em K=7, como ilustrado na Figura 5. Esses resultados indicam que diferentes métricas sugerem valores distintos para K, reforçando a necessidade de uma decisão baseada em análise conjunta e não em um único critério isolado.

[9]https://data.police.uk/about

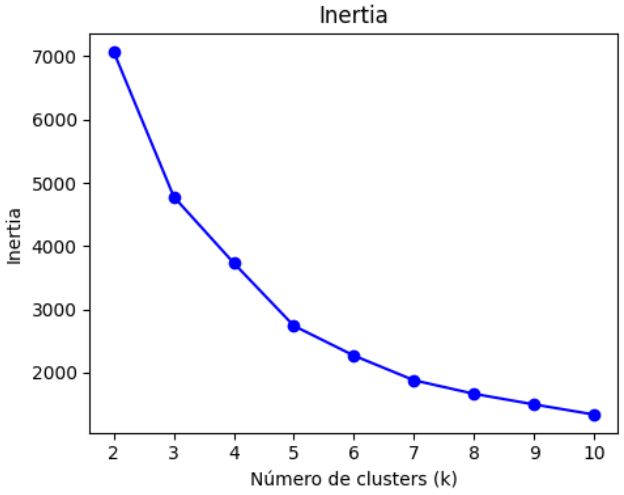

Figura 3. Número de clusters para Inertia.

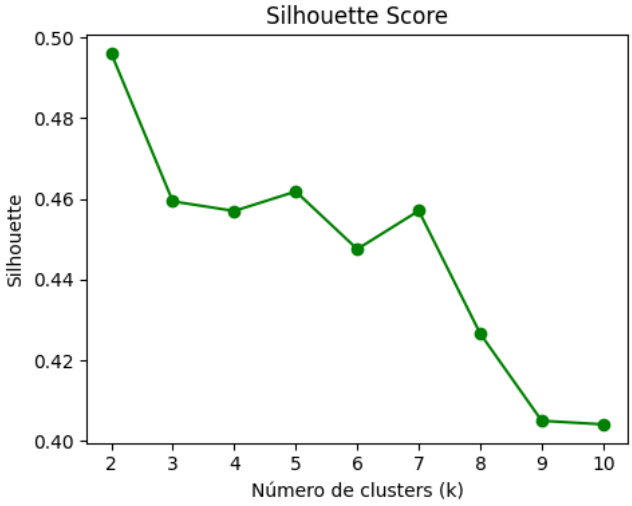

Figura 4. Número de clusters para Silhouette.

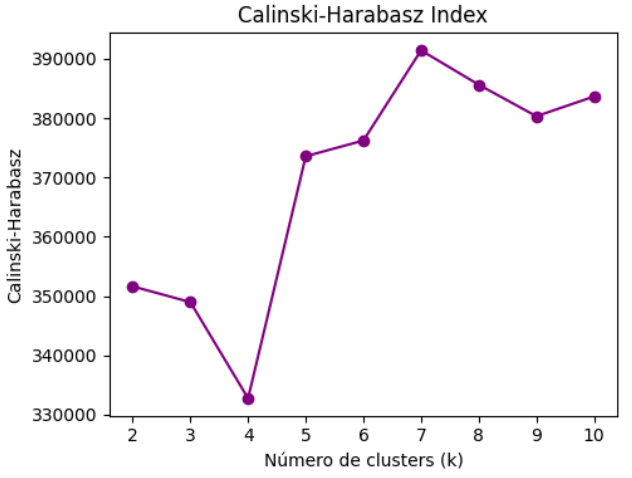

Figura 5. Número de clusters para Calinski Harabasz.

A definição do número de *clusters* foi, portanto, realizada com base em experimentação orientada por métricas de avaliação, na qual foram efetuados testes com K variando unitariamente, em ordem crescente, de 2 a 10, com tempo total de execução de 3h e 52 min. A escolha final considerou K=4, que apresentou um equilíbrio entre as métricas, com *Silhouette Score* de 0,4570, Índice de *Calinski–Harabasz* de 332760,86 e Inércia de 3,73 × 103, indicando uma maior separação entre os clusters, conforme representado na Figura 6.

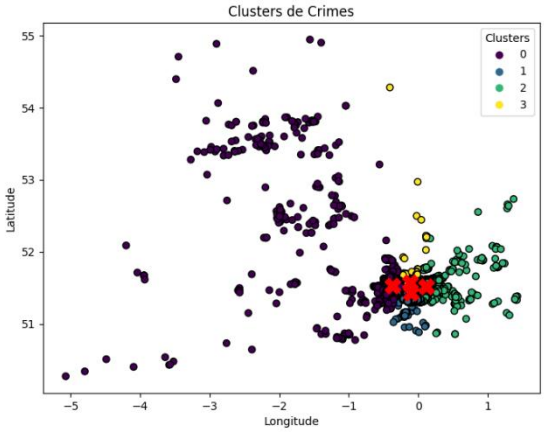

Figura 6. Clusters de crimes e seus centróides.

Do ponto de vista conceitual, a escolha de K=4 mostrou-se mais adequada por representar um compromisso entre qualidade estatística dos agrupamentos e interpretabilidade espacial dos resultados. Embora valores menores de K tenham apresentado melhores escores em métricas específicas, como o Coeficiente de *Silhouette* em K=2, tais configurações resultaram em agrupamentos excessivamente amplos, que tendem a mascarar heterogeneidades espaciais relevantes da distribuição criminal e reduzem a utilidade prática para o planejamento territorial. Por outro lado, valores mais elevados de K, como aqueles sugeridos pelo pico do Índice de *Calinski–Harabasz*, implicaram uma fragmentação excessiva do espaço urbano, produzindo *clusters* muito pequenos e operativamente menos viáveis para a definição de áreas de cobertura de unidades policiais. Nesse sentido, K=4 apresentou uma solução intermediária, capaz de preservar padrões espaciais significativos de concentração de ocorrências, ao mesmo tempo em que gera regiões territorialmente contínuas, coerentes e compatíveis com uma lógica de planejamento e alocação de recursos. Sob a perspectiva operacional, essa configuração favorece a definição de áreas de atuação mais equilibradas, potencialmente associáveis a zonas de patrulhamento ou de implantação de novas unidades, tornando os resultados mais interpretáveis e aplicáveis ao processo decisório em segurança pública.

A segunda etapa teve como objetivo identificar as melhores localizações para as novas unidades policiais, com a função *ST_Centroid*. Na Figura 7 é apresentado o resultado:

•A cor vermelha representa as ocorrências de crimes individuais e incidentes de comportamentos antissociais.

•A distribuição das unidades de delegacias já existentes está representada pelas estrelas rosa.

•As localizações sugeridas para as novas unidades estão representadas pelos símbolos verdes. O objetivo foi otimizar o

atendimento às áreas com maior concentração de ocorrências, e distantes das delegacias já existentes.

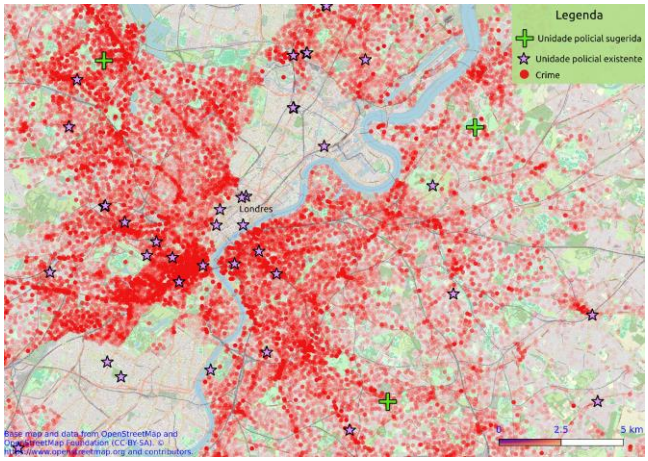

Figura 7. Crimes na região metropolitana de Londres e unidades policiais.

A escolha pelo *K-means* justifica-se por sua eficiência e escalabilidade, características fundamentais para lidar com grandes volumes de dados criminais georreferenciados. A implementação no *PostGIS*, por meio da função *ST_ClusterKmeans*, permite agrupar geometrias com base na proximidade espacial sem exportar dados para ferramentas externas, garantindo processamento interno e integração nativa com consultas *SQL*. Estudos indicam que o *K-means* apresenta desempenho superior em tempo de execução em relação a métodos mais complexos, como *DBSCAN*, especialmente em bases dinâmicas e de grande escala conforme [9]. Além disso, avanços recentes demonstram que o *K-means* continua sendo uma solução eficiente e previsível para vetores espaciais em larga escala, mesmo diante de desafios de memória e custo computacional [16]. Assim, a simplicidade, a rapidez e a integração direta no banco de dados justificam sua adoção em detrimento de abordagens como *DBSCAN* e *Spectral Clustering*, que, embora sofisticadas, exigem maior custo de processamento.

4.5 Critérios de Reprodutibilidade

Todas as consultas *SQL* utilizadas para clusterização e cálculo das métricas foram executadas diretamente no SGBD, permitindo replicabilidade do experimento mediante disponibilização do *script* e da base de dados.

A abordagem proposta mantém integridade topológica das geometrias e evita serialização intermediária dos dados.

5. CONCLUSÃO

A ausência de cobertura policial em determinadas regiões de uma cidade pode impactar a concentração de crimes. Este estudo abordou o desafio de mitigar problemas de segurança pública em grandes centros urbanos através da distribuição estratégica de unidades policiais.

Este trabalho apresentou uma abordagem de clusterização espacial executada diretamente em banco de dados geográfico, explorando a função *ST_ClusterKMeans* do *PostGIS* como mecanismo de processamento analítico *in-database*.

A metodologia empregou a análise espacial utilizando o algoritmo *K-Means*, aplicado a um total de 554.832 registros. Testes com métricas como *Elbow*, *Silhouette* e *Calinski-Harabasz* orientaram a escolha de K=4, o que permitiu revelar concentrações significativas de crimes. Os centróides dos *clusters* foram identificados como localizações sugeridas para novas unidades. Esta identificação proporciona um subsídio quantitativo para a expansão da cobertura policial, visando otimizar o atendimento em áreas com maior concentração de ocorrências e que estão distantes das unidades policiais existentes. Isso representa um avanço em relação a um viés reativo e contribui para um planejamento estratégico baseado em evidências. A 'distribuição estratégica de recursos' mencionada refere-se à capacidade de direcionar, de forma mais assertiva, a instalação física de novas unidades em locais estatisticamente identificados como de maior necessidade.

A flexibilidade desta abordagem permite a generalização do método para além do contexto de segurança pública, uma vez que a criação de regiões de interesse baseadas em divisões matemáticas independe de limites administrativos tradicionais. Essa versatilidade possibilita possíveis aplicações em outros domínios urbanos, como o planejamento de redes de transporte público por meio da análise de fluxos de origem-destino, ou a gestão de recursos econômicos e ambientais, como o monitoramento da pesca artesanal, onde a identificação de centros de demanda e o agrupamento de dados mistos são essenciais para o manejo estratégico.

No entanto, a implementação evidenciou limitações computacionais inerentes ao processamento de grandes massas de dados georreferenciados, exigindo estratégias rigorosas de indexação e otimização de consultas para manter o desempenho em escalas metropolitanas. Além disso, desafios relacionados à compatibilidade entre diferentes bibliotecas de visualização e à dependência de conectividade estável para o acesso a interfaces cartográficas representam gargalos técnicos que devem ser considerados para a universalização do sistema. Para aprimorar ainda mais a eficácia destas medidas, recomenda-se o aperfeiçoamento dos dados, utilizando classificações mais detalhadas dos delitos, considerando padrões temporais e índices de reincidência. A integração da base criminal com informações socioeconômicas e urbanísticas, como densidade populacional, iluminação pública e fluxo de pedestres, é fundamental para uma alocação mais eficiente dos recursos, alinhada às dinâmicas sociais e urbanas.

Como trabalhos futuros, destacam-se os seguintes direcionamentos de pesquisa: 1) desenvolver e comparar diferentes algoritmos de clusterização espacial para avaliar sua adequação em diferentes contextos urbanos e tipos de crime; 2) incorporar dados históricos de crimes e atributos temporais para identificar padrões sazonais e tendências de longo prazo nos *hotspots*; 3) integrar os dados criminais com informações socioeconômicas e urbanísticas para construir modelos mais robustos que ajudem a identificar fatores contribuintes para a criminalidade em áreas específicas; 4) colaborar com especialistas do domínio sendo estes policiais, criminologistas, urbanistas, para validar qualitativamente os *hotspots* identificados e as localizações sugeridas, garantindo a aplicabilidade e relevância prática das recomendações; 5) desenvolver versões otimizadas para dispositivos móveis para facilitar o acesso à informação em locais com infraestrutura de rede limitada; 6) automatizar o fluxo de inserção de dados para permitir o monitoramento dinâmico em tempo real; 7) incorporar modelos de análise complementares para correlacionar os dados espaciais com tendências temporais preditivas. Além disso, pode-se mencionar a análise dos *hotspots* identificados, investigando suas características urbanísticas e socioeconômicas para além da simples concentração de crimes.

6. AGRADECIMENTOS

Os autores agradecem especialmente à Universidade Tecnológica Federal do Paraná e aos órgãos governamentais pela disponibilização de dados abertos.

7. REFERÊNCIAS

[1] Abhay, D. A., Akash, S., Ashwin, K., Shenoy, A. G., and Auradkar, P. K. (2023). Smart policing: Using geospatial crime data to plan patrol routes. 2023 4th International Conference for Emerging Technology (INCET), pages 1–7.

[2] Almuhanna, A. A., Alrehili, M. M., Alsubhi, S. H., and Syed, L. (2021). Prediction of crime in neighbourhoods of new york city using spatial data analysis. 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA).

[3] Alves Junior, A. A., Cacho, N., Thome, A. C., Medeiros, A., and Borges, J. (2017). A predictive policing application to support patrol planning in smart cities. In IEEE International Smart Cities Conference (ISC2), pages 1–8.

[4] Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., and Portugali, J. (2012). Smart cities of the future. The European Physical Journal Special Topics, 214(1):481–518.

[5] Bibri, S. E. and Krogstie, J. (2017). Smart sustainable cities of the future: An extensive interdisciplinary literature review. Sustainable Cities and Society, 31:183–212.

[6] Borowik, G., Wawrzyniak, Z. M., and Cichosz, P. (2018). Time series analysis for crime forecasting. 2018 26th International Conference on Systems Engineering (ICSEng).

[7] Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. Communi-cations in Statistics, 3(1):1–27.

[8] Carvalho, G. R., Kozievitch, N. P., and Zavadzki, L. S. (2025). Análise de otimização de pontos para abertura de novas unidades policiais. ERBD 2025 Proceedings – Escola Regional de Banco de Dados, pages 161–164.

[9] Chakraborty, S., Nagwani, N. K., and Dey, L. (2014). Performance comparison of incre- mental k-means and incremental dbscan algorithms. International Journal of Advan-ced Research in Computer and Communication Engineering.

[10] Dong, Q., Ye, R., and Fan, Y. (2022a). Impact of spatial correlation on crime prediction in communities with different crime densities. In IEEE International Conference on Communications (ICCC), pages 1–6.

[11] Dong, Q., Ye, R., and Fan, Y. (2022b). Impact of spatial correlation on crime prediction in communities with different crime densities. 2022 IEEE 8th International Conference on Computer and Communications (ICCC).

[12] Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M., and Wilson, R. E. (2005). Mapping crime: Understanding hot spots. Technical report, National Institute of Justice, U.S. Department of Justice.

[13] Hellerstein, J. M., Re, C., Schoppmann, F., Wang, D. Z., Fratkin, E., Gorajek, A., Ng, K. S., Welton, C., Feng, X., Li, K., and Kumar, A. (2012). The madlib analytics library: Or mad skills, the sql. Proceedings of the VLDB Endowment, 5(12):1700–1711.

[14] Hunt, E. L. and Reffert, S. (2020). Improving the open cluster census i: Comparison of clustering algorithms applied to gaia dr2 data. Astronomy & Astrophysics, 633:A114.

[15] Jain, S. and Riyank (2023). Geographical crime rate prediction. 2023 4th International Conference on Intelligent Engineering and Management (ICIEM).

[16] Ji, Y., Liu, Z., Wang, S., Sun, Y., and Peng, Z. (2025). On simplifying large-scale spatial vectors: Fast, memory-efficient, and cost-predictable k-means. Proceedings of the 41st IEEE International Conference on Data Engineering (ICDE).

[17] Kim, Y., Kim, G., Lee, Y., and Jang, K. (2020). Bandwidth selection of kernel density estimation for gis-based crime occurrence map visualization. In International Con- ference on Information and Communication Technology Convergence (ICTC), pages 1705–1708.

[18] Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University and Durham University Joint Report, UK.

[19] Leal, M. and Gomes-Jr, L. (2022). Impacto da pandemia da covid-19 nos padrões de crimes no município de curitiba. In Anais da XVII Escola Regional de Banco de Dados, pages 101–108, Porto Alegre, RS, Brasil. SBC.

[20] Li, S. and Tang, W. (2019). Space-time variation of property crime in beijing with esda method. 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC).

[21] Lilhare, S. G., Kumavat, Y., Banait, G., and Kurkelli, A. (2024). Crime hotspots mapping and fir data interface. ICICET 2024, pages 1–7.

[22] MacQueen, J. (1967). Some methods for classification and analysis of multivariate obser- vations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pages 281–297.

[23] Mantovani, D. M. N., Santos, L. T., Machado, C. J., and Leal, G. A. (2024). Inteligência em segurança pública municipal: modelos para previsão de crimes e planejamento de zeladoria urbana. ISLA 2024 Proceedings Latin America (ISLA), pages 1–10.

[24] Nag, A., Ranjan, R., and Kumar, C. N. S. V. (2022). An approach on cyber crime pre- diction using prophet time series. 2022 IEEE 7th International conference for Conver-gence in Technology (I2CT).

[25] Parcianello, Y., Kozievitch, N. P., Fonseca, K. V. O., and Rosa, M. (2021). Origin- destination data: a prototype and related scenarios. Revista Brasileira de Computação Aplicada, 13(2):16–27.

[26] Ratcliffe, J. H. (2010). Crime mapping and the training needs of law enforcement. Police Practice and Research, 11(1), pages 3–15.

[27] Ratra, A., Agarwal, A., Vats, S., Sharma, V., Kukreja, V., and Yadav, S. P. (2023). A com- prehensive review on crime patterns and trends analysis using machine learning. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS).

[28] Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20:53–65.

[29] Sathiyanarayanan, M., Junejo, A. K., and Fadahunsi, O. (2019). Visual analysis of pre- dictive policing to improve crime investigation. IC3I 2019, pages 197–203.

[30] Sayal, A., Gupta, A., Vasundhara, C., M, Y. B., Gupta, V., and Maheshwari, H. (2024). Crime detection using data mining techniques. 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT).

[31] Sharma, S., Uniyal, A., Srinivasan, P., and Chaudhari, S. (2022). Fuzzy based geo-spatial crime category prediction for crime mapping and safe route travel. 2022 IEEE Region 10 Symposium (TENSYMP).

[32] Sheikh, J., Shafique, I., Sharif, M., Zahra, S. A., and Farid, T. (2017). Role of gis in crime mapping and analysis. ComTech 2017, pages 126–131.

[33] Shenoy, A. G. and Auradkar, P. K. (2023). Smart policing: Using geospatial crime data to plan patrol routes. 2023 4th International Conference for Emerging Technology (INCET).

[34] Sherman, L. W., Gartin, P. R., and Buerger, M. E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. Criminology, 27(1):27–56.

[35] Silva, R. S., Silva, R. R., Kuribayashi, H. P., Cunha, C. V., Francês, C. R. L., and Sousa, K. N. S. (2019). Clusterização de dados mistos para análise da atividade pesqueira artesanal na bacia araguaia-tocantins. Revista Brasileira de Computação Aplicada, 11(3):155–164.

[36] Sravani, T. and Suguna, M. R. (2022). Comparative analysis of crime hotspot detection and prediction using convolutional neural network over support vector machine with engineered spatial features towards increase in classifier accuracy. 2022 International Conference on Business Analytics for Technology and Security (ICBATS).

[37] Steier, J., Zigarelli, A., Giannini, E., and Minimair, M. (2017). Crime sequencing: Figh- ting crime with mathematics and technology. 2017 IEEE MIT Undergraduate Research Technology Conference (URTC).

[38] Stonebraker, M. (2018). The end of an architectural era: (it's time for a complete rewrite). Communications of the ACM, 61(11):22–24.

[39] Tarlekar, S., Bhosle, R., D'souza, E., and Sheikh, S. (2021). Geographical crime rate prediction system. 2021 IEEE India Council International Subsections Conference (INDISCON).

[40] Vadav, M. G., N, R., Reddy, E. S., Vishal, M. S., and Vishal, G. (2024). The role of machine learning in crime analysis and prediction. 2024 International Conference on Expert Clouds and Applications (ICOECA).

[41] von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing,

[42] 17(4):395–416.

[43] Wang, J., Zhang, S., Lan, Y., Wu, C., Xia, Y., and Chen, L. (2020). Crime analysis of spatial-temporal distribution based on knn algorithm. ICISE-IE 2020, pages 150–154.

[44] Yadav, R. and Sheoran, S. K. (2018). Crime prediction using auto regression techniques for time series data. 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE).

[45] Yoo, J. S., Park, S. J., and Raman, A. (2019). Micro-level incident analysis using spa- tial association rule mining. 2019 IEEE International Conference on Big Knowledge (ICBK).

[46] Zanella, A., Bui, N., Castellani, A., Vangelista, L., and Zorzi, M. (2014). Internet of things for smart cities. IEEE Internet of Things Journal, 1(1):22–32.