UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

TensorAnalyzer: Identificação de Padrões Urbanos em Grandes Cidades usando Fatoração de Tensores Não-Negativos

Jaqueline Alvarenga Silveira

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Jaqueline Alvarenga Silveira

TensorAnalyzer: Identificação de Padrões Urbanos em Grandes Cidades usando Fatoração de Tensores Não-Negativos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Afonso Paiva Neto

USP – São Carlos
Fevereiro de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A473t

Alvarenga Silveira, Jaqueline
TensorAnalyzer: Identificação de Padrões Urbanos em Grandes Cidades usando Fatoração de Tensores Não-Negativos / Jaqueline Alvarenga Silveira; orientador Afonso Paiva Neto. -- São Carlos, 2023.
81 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. Decomposição de Tucker. 2. Homicídios. 3. Escolas. 4. Clusterização. I. Paiva Neto, Afonso, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Jaqueline Alvarenga Silveira

TensorAnalyzer: Urban Patterns Identification in Big Cities using Non-Negative Tensor Factorization

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Afonso Paiva Neto

USP – São Carlos
February 2023

Dedico essa tese de doutorado a todas as pessoas que me apoiaram durante toda a minha caminhada até aqui. A minha família por ser meu alicerce em todas as fases da minha vida, em especial meus pais que sempre me incentivaram a estudar e seguir em frente!

AGRADECIMENTOS

Primeiramente agradeço a Deus por estar comigo em todos os momentos, sorrindo nos momentos de alegria e segurando a minha mão sempre que eu caio.

À minha família, pelo imenso amor, dedicação e união. Não importa o que aconteça estão sempre prontos para me ajudar, sempre me impulsionando para frente.

À Dieninha que me acompanha sempre, e não importa se passamos 1 ou 2 meses sem nos vermos, quando nos encontramos o amor é exatamente o mesmo.

Aos meus amigos que sempre torceram por mim. Ao Germain, que foi importante em muitos momentos nesta jornada.

Ao prof. Dr. Luis Gustavo Nonato por me orientar, apoiar e direcionar os caminhos possíveis.

Ao meu orientador prof. Dr. Afonso Paiva por me orientar, apoiar, direcionar e compreender em todos os momentos. Agradeço-o imensamente pela oportunidade que me dera de entrar em uma Universidade como a USP.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

“Mires na Lua, porque mesmo que você erre, acabará entre as estrelas.”
(Les Brown)

RESUMO

JAQUELINE, A. S. **TensorAnalyzer: Identificação de Padrões Urbanos em Grandes Cidades usando Fatoração de Tensores Não-Negativos**. 2023. 78 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Grandes cidades como São Paulo (a maior da América Latina), tipicamente apresentam grandes volumes de crimes que vão desde roubos até homicídios. O aumento de atividades criminais em São Paulo tem despertado o interesse de especialistas de criminologia em compreender o relacionamento entre atividades criminais e as características no entorno de escolas, homicídios, etc. No entanto, a extração de relevantes padrões pode se tornar uma tarefa complicada ao empregar algoritmos de clusterização clássicos, isso porque existem muitos parâmetros para configurar e também o usuário ainda tem que se preocupar com os outliers presentes nos dados. Deste modo, este trabalho apresenta uma nova abordagem para detectar padrões relevantes de múltiplas fontes de dados baseada em decomposição de tensor. O desempenho da abordagem proposta é atestada para validar a qualidade dos padrões identificados em comparação com abordagens clássicas. O resultado indica que a abordagem pode efetivamente identificar padrões úteis para caracterizar o conjunto de dados para análise posterior na obtenção de uma boa qualidade de agrupamento. Além disso, um framework genérico nomeado TensorAnalyzer foi desenvolvido, em que a eficácia e a utilidade da metodologia proposta são destacadas por experimentos em conjuntos de dados sintéticos e do mundo real, sendo desenvolvidos dois estudos de caso, em que o primeiro estudo mostra a relação entre o entorno das escolas e os padrões criminais e o segundo estudo busca compreender os padrões no entorno dos homicídios.

Palavras-chave: Decomposição de Tucker, Homicídios, Escolas, Clusterização.

ABSTRACT

JAQUELINE, A. S. **TensorAnalyzer: Urban Patterns Identification in Big Cities using Non-Negative Tensor Factorization**. 2023. 78 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Extracting relevant patterns from multiple data sources can be difficult using classical clustering algorithms since we have to make the suitable configuration of the hyperparameters of the algorithms and deal with outliers. In many contexts, pattern extraction is crucial and should be addressed correctly. In criminology, for example, one of the main interests of the experts of São Paulo is the comprehension of the relationship between crimes and the characteristics around specific targets. This work presents a new approach to detecting the most relevant patterns from multiple data sources based on tensor decomposition. Compared to classical methods, the proposed approach's performance is attested to validate the identified patterns' quality. The result indicates the approach can effectively identify useful patterns to characterize the data set for further analysis in achieving good clustering quality. Furthermore, we developed a generic framework named TensorAnalyzer, where the effectiveness and usefulness of the proposed methodology are tested by a set of experiments and two real-world cases studies showing the relationship between the crime events, urban characteristics, and other variables involved in the analysis.

Keywords: Tucker Decomposition, Homicides, Schools, Clustering.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

APP	Análise de Padrões de Pontos
CEM	Centro de Estudos da Metrópole
CEU	Centros Educacionais Unificados
Emplasa	Empresa Paulista de Planejamento Metropolitano
IPT	Instituto de Pesquisas Tecnológicas do Estado de São Paulo
IPVS	Índice Paulista de Vulnerabilidade Social
NEV-USP	Estudos da Violência na Universidade de São Paulo
NMF	Non-Negative Matrix Factorization
NTD	Non-Negative Tucker Decomposition
OLS	Ordinary Least-Squares
RLO	Regressão Logística Ordinal

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A partir dos anos 60, as relações de trabalho começaram a se modificar tanto no campo quanto nas cidades e as consequências foram o rápido êxodo rural e o crescimento acelerado das cidades. De acordo com Silva (2008) o processo de urbanização sem planejamento gera problemas sociais e ambientais, tais como, ausência de moradias e infraestrutura urbana, crescimento da economia informal, poluição, intensificação do trânsito, periferização dos pobres e ocupação das áreas de proteção. Em geral, regiões metropolitanas desenvolvem mais rapidamente do que o planejado e isto gera um crescimento desordenado que implica em impactos sociais e ambientais.

Em São Paulo, uma das maiores metrópoles do mundo, o crescimento acelerado causou uma conurbação intensa, que culminou em uma grande área urbana conhecida como "Grande São Paulo", que atualmente contém 37 cidades. Essas, por sua vez, formam a principal região industrial do país e têm se desenvolvido sem um planejamento urbano. Deste modo, equipamentos como infraestrutura, habitação e transporte não atendem às suas demandas sociais (JÚNIOR, 2014), o que contribui para o cenário da vulnerabilidade social.

Nesse contexto, uma das consequências da vulnerabilidade social vem sendo o aumento significativo nas taxas de crime (SILVA; GRIGIO; PIMENTA, 2016). De maneira similar às cidades mais populosas do mundo, a vulnerabilidade social tem feito com que São Paulo experimente um alto volume de crimes, variando de pequenos roubos até homicídios. No entanto, crimes tais como roubos e assaltos ocorrem com maior frequência em relação aos demais crimes (ver Tabela 1) e os principais alvos são pessoas descuidadas e que aparentam ricas. Assim, os itens mais almejados pelos criminosos incluem carteiras/bolsas sendo que joias e eletrônicos, como telefones celulares, são de particular interesse (OSAC, 2020).

Diante desse cenário, formuladores de políticas públicas da região metropolitana de São Paulo buscam compreender as relações entre o crime e as outras variáveis no contexto urbano de modo a implementar políticas públicas que contribuam para a redução da criminalidade em geral. Ao longo dos anos, alguns autores têm se empenhado em mostrar que a infraestrutura

Tabela 1 – Esta Tabela resume as ocorrências de crime durante o período de 2000 até 2020.

Fonte: Adaptada de Prefeitura de São Paulo (2018, Prefeitura de São Paulo).

Tipos de Crime	2000	2005	2010	2015	2020
Crimes contra pessoas					
Homicídios/tentativa	15.210	15.226	9.034	7.763	2.766
Lesão corporal	93.631	121.891	114.226	88.805	29.258
Crimes contra propriedade					
Roubos/Tentativa	151.909	155.668	164.951	234.413	187.564
Assalto	324	217	130	406	41
Roubo de Veículo/Tentativa	95.367	64.537	53.006	61.528	12.097
Furto/Tentativa	156.588	240.741	234.160	243.358	113.573
Furto de Veículo/Tentativa	82.537	79.202	61.054	66.625	21.244

urbana pode ser vista como um atrator de crime, principalmente aquelas que permitem um grande fluxo diário de pessoas como metrôs, praças, pontos de ônibus, escolas, etc. Estudos recentes, mostraram que problemas como baixo desempenho acadêmico, evasão escolar, suspensões frequentes e repetição de série, têm sido consistentemente associados a problemas comportamentais de adolescentes, incluindo a delinquência juvenil (ARCHWAMETY; KATSIYANNIS, 2000; KREZMIEN; MULCAHY; LEONE, 2008; PETROCELLI; PETROCELLI, 2005). Além disso, em um estudo (HEISSEL J. A .AND SHARKEY *et al.*, 2018) realizado por pesquisadores da *Northwestern University*, da *New York University* e da *DePaul University* descobriu-se que o crime violento altera os padrões de sono dos jovens na noite imediatamente após o crime e também altera os padrões do hormônio do estresse, cortisol, no dia seguinte. Ambos podem influenciar de maneira negativa o desempenho acadêmico dos alunos. Isso fez com que formuladores de políticas públicas da região metropolitana de São Paulo voltassem seus olhares para o estudo da relação entre atividades criminais e o entorno das escolas, os quais levantaram as seguintes hipóteses:

H1 – Existe uma relação entre eventos criminais e outras variáveis envolvidas nas análises considerando regiões próximas às escolas.

H2 – A criminalidade influencia o desempenho dos estudantes.

H3 – Áreas recreativas próximas às escolas tendem a ser atratores de crime.

Homicídio é considerado outro problema de saúde pública mundial que acomete vários países e tem sido reconhecido como um evento complexo caracterizado por variações no estilo comportamental, nível de violência, motivação e interação pessoal. Assim, compreender a redução de homicídios em alguns lugares pode provar ser instrutivo para a comunidade em geral, por causa das implicações para reduzir crimes violentos. São Paulo, por exemplo, apesar de a taxa de homicídios estar caindo ao longo dos anos, ainda se mantém alta como pode-se observar na Figura 1. Portanto, identificar padrões no entorno de homicídios a fim de compreender a

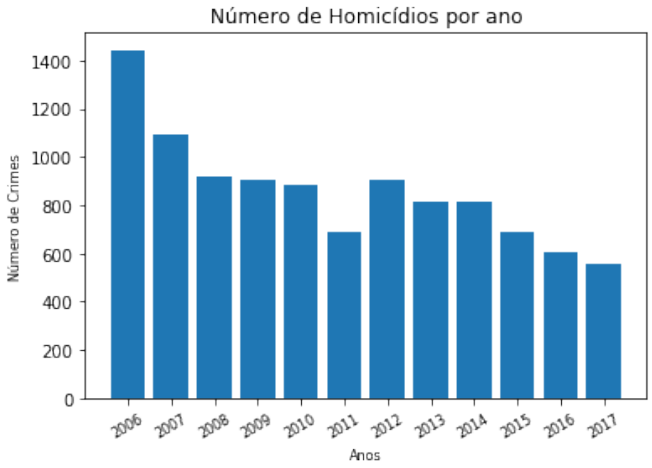


Figura 1 – Quantidade de homicídios por ano na cidade de São Paulo.
Fonte: Elaborada pelo autor.

relação entre eventos criminais e o seu entorno é essencial para o auxiliar políticas públicas na tomada de decisão. Assim, especialistas da área levantaram as seguintes hipóteses:

H4 – Existe uma relação entre homicídios e as outras variáveis envolvidas nas análises em um determinado contexto.

H5 – Em algumas regiões é possível observar tanto uma melhoria na infraestrutura urbana como também uma redução na criminalidade ao longo dos anos.

Ambos os problemas descritos anteriormente estão dentro do escopo da Análise de Padrões de Pontos (APP), a qual busca compreender a relação entre a localização dos eventos e a resposta sobre a distribuição desses locais. Mais especificamente, o interesse consiste em verificar se um padrão de ponto segue algum processo sistemático de agrupamento ou se segue um processo aleatório (YAMADA; ROGERSON, 2010). As informações obtidas a partir desses processos podem permitir adquirir alguns insights iniciais sobre um determinado fenômeno. Não obstante, a APP é considerada uma tarefa desafiadora (GATRELL *et al.*, 1996a; PODUR; MARTELL; CSILLAG, 2003; ALJANABI, 2011; Joshi; Sabitha; Choudhury, 2017), visto que, mesmo que *insights* venham a ser identificados nos dados, os mesmos devem obedecer a um grau de relevância de acordo com o domínio de conhecimento dos especialistas da área que se estuda. Assim, o emprego de técnicas rudimentares para a solução de problemas mais complexos, no contexto da APP, pode levar a uma extração equivocada de padrões e consequentemente os resultados das análises realizadas em torno desses padrões também serão equivocadas.

As técnicas comumente empregadas na solução de problemas no contexto da APP surgiram há mais de 50 anos na ecologia de plantas e foram agrupadas em duas classes: aquelas que examinam a localização dos pontos em relação à área de estudo e aquelas que examinam a localização dos pontos em relação a uns aos outros (BOOTS; GETLS, 2020). Dentro de

cada classe existe uma variedade de técnicas para examinar padrões de pontos, sendo que não existe uma técnica ideal para todas as aplicações e a seleção de um determinado procedimento é influenciada por questões práticas e estatísticas. Devido a frequência de problemas complexos que nascem diariamente e exigem cada vez mais dos pesquisadores soluções mais robustas do que aquelas existentes na literatura, a análise de padrões de pontos é considerada ainda uma área em expansão (PODUR; MARTELL; CSILLAG, 2003).

Por outro lado, embora esteja em um estágio inicial de desenvolvimento, a ciência de dados tem sido cada vez mais reconhecida como uma das principais forças motrizes para diversos campos de estudo, visto que, é um novo campo interdisciplinar que sintetiza e se baseia em estatística, informática, computação, comunicação, gestão e sociologia para estudar dados e seus ambientes, a fim de transformar dados em insights e decisões (CAO, 2017). Nesse sentido, o emprego de técnicas de ciências de dados capazes de manipular diversas fontes de dados de modo a extrair padrões que estejam distribuídos nos dados, parece ser uma alternativa promissora para a solução de problemas mais complexos no contexto da análise de padrões de ponto. Além disso, de acordo com nossos conhecimentos, a ciência de dados ainda não foi aplicada nesse contexto e portanto, esta seria uma das principais contribuições deste trabalho.

Em colaboração com um time de sociologistas, neste trabalho, nós propomos uma nova metodologia baseada na combinação da Non-Negative Tucker Decomposition (NTD) (CICHOCKI *et al.*, 2009) e de algoritmos de *cluster* hierárquicos (RAFSANJANI; VARZANEH; CHUKANLO, 2012; CAMPELLO; MOULAVI; SANDER, 2013) para encontrar padrões no entorno de sites (escolas, homicídios, etc), tendo em conta diferentes fontes de dados, em uma ferramenta visual vinculada que permite a análise dos padrões encontrados. A eficácia e utilidade da metodologia proposta são evidenciadas por um conjunto de experimentos e estudos de caso.

1.1 Contribuições da tese

- Uma metodologia para identificar padrões em diversas fontes de dados georreferenciadas por meio da combinação da fatoração de tensores não-negativos e algoritmo de clusterização hierárquico;

- ferramentas visuais que permitem a exploração visual dos padrões encontrados, sendo que, para cada estudo de caso apresentado neste trabalho foi construída uma ferramenta para exploração dos padrões encontrados pela metodologia proposta;

- resultados usando dados sintéticos e estudo de casos reais que evidenciam o potencial da metodologia proposta, ao revelar padrões escondidos nos dados, em que os mesmos foram validados por especialistas.

1.2 Contribuições do trabalho

A seguir apresenta-se as principais contribuições resultantes deste trabalho de doutorado:

- **Publicação da ferramenta TensorAnalyzer:**
  O desenvolvimento da ferramenta TensorAnalyzer é a principal contribuição desta tese, isso porque, nessa ferramenta nós desenvolvemos uma metodologia baseada na combinação de decomposições e algoritmos de cluster para extrair padrões no entorno de sites. A fim de divulgar os resultados obtidos com essa ferramenta, um artigo foi escrito e será submetido para a revista *ACM Transactions on Intelligent Systems and Technology*.

- **Publicação sobre a aplicação do TensorAnalyzer para dados de homicídios:**
  Após uma apresentação do TensorAnalyzer para o Núcleo de Estudos da Violência, os pesquisadores perceberam que a ferramenta seria muito útil para identificar padrões no entorno de homicídios da cidade de São Paulo. Neste sentido, a fim de divulgar os resultados obtidos um artigo foi escrito e foi submetido para a revista *Journal of Quantitative Criminology*.

1.3 Outras atividades e colaborações

A seguir apresentamos outras atividades e colaborações desenvolvidas ao longo deste trabalho de doutorado:

- **Apresentação do TensorAnalyzer no Congresso Nacional de Matemática aplicada:**
  O TensorAnalyzer foi apresentado no minissimpósio "**Mathematics Against Crime**"que ocorreu durante o CNMAC 2019 realizado no período de 16 a 20 de novembro de 2019 em Uberlândia, MG.

- **Palestra sobre o TensorAnalyzer:**
  Esta palestra aconteceu na Fundação Getúlio Vargas em 11 de Fevereiro de 2020, em que foi apresentada a ferramenta bem como um estudo de caso baseado nas escolas.

- **Colaboração no artigo CrimAnalyzer:**
  Colaboração no estudo sobre identificar hotsposts de crime e as correspondentes evoluções temporais por meio de decomposição de matrizes, que gerou a seguinte publicação:

  Zanabria, G.G., Silveira, J.A., Poco, J., Paiva, A., Nery, M.B., Silva, C.T., Abreu, S.F., Nonato, L. (2019). CrimAnalyzer: Understanding Crime Patterns in São Paulo. IEEE transactions on visualization and computer graphics.

- **Colaboração no artigo MIRANTE:**
  Colaboração no estudo sobre identificar hotspots de crime bem como a correspondente evolução temporal, por meio de grafo, que gerou a seguinte publicação:

Zanabria, G.G., Nieto, E., Silveira, J., Poco, J., Nery, M., Adorno, S., Nonato, L. (2020). Mirante: A visualization tool for analyzing urban crimes. 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 148-155.

Melhor trabalho no SIBGRAPI 2020

1.4 Organização da Tese

- **Capítulo 2 -** Neste Capítulo foi realizado um levantamento bibliográfico dos principais trabalhos encontrados os quais estão relacionados com as áreas de interesse desta tese. Desse modo, este capítulo foi dividido em duas Seções: na Seção 2.1 apresentam-se trabalhos que estão relacionados com a extração de padrões em dados georreferenciados e na Seção 2.2 apresentam-se os trabalhos relacionados com a decomposição de tensores.

- **Capítulo 3 -** Este Capítulo apresenta a metodologia empregada para a extração de padrões, que é a principal contribuição desta tese, tendo em conta diversas bases de dados. Mais especificamente, apresentam-se os conceitos e a interpretação da DTN, a modelagem do tensor, a extração de padrões e as visualizações que foram desenvolvidas.

- **Capítulo 4 -** Neste capítulo, atesta-se a metodologia desenvolvida no Capítulo 3 em que, por meio de dados sintéticos, realiza-se uma comparação com outras abordagens que também permitem identificar padrões nos dados. Contudo, emprega-se a metodologia para encontrar padrões em dois conjuntos de dados: escolas e homicídios, sendo que, os resultados são avaliados por sociólogos especialistas em estudo de crimes em São Paulo.

- **Capítulo 5 -** Conclusão: finalmente concluímos discutindo os principais resultados e contribuições deste trabalho, levantando os bons resultados e vantagens da metodologia desenvolvida bem como as limitações, e apontando alguns possíveis trabalhos futuros.

CAPÍTULO

2

TRABALHOS RELACIONADOS

Como a presente abordagem extrai padrões urbanos de dados georreferenciados usando decomposição tensorial, para melhor contextualiza-la, organiza-se os métodos existentes para extração de padrões a partir de dados georreferenciados e aplicações de decomposição tensorial para processamento de dados.

2.1 Extração de padrões a partir de dados georreferenciados

Dados georreferenciados, também conhecidos como dados espaciais, geográficos ou geoespaciais, são as informações essenciais necessárias para identificar a localização geográfica de fenômenos na superfície da Terra (GATRELL *et al.*, 1996b). Em geral, dados georreferenciados são indexados por coordenadas geográficas (pontos referenciados por latitude e longitude) que permitem a aplicação de técnicas de extração de padrões espaciais para extrair informações valiosas dos dados. A seguir, apresenta-se um panorama dos principais campos investigados na literatura que buscam encontrar padrões considerando diversas fontes de dados georreferenciados.

A análise de padrão de pontos tem sido um interesse crescente na identificação e análise de locais de incêndios florestais. Por exemplo, Podur, Martell e Csillag (2003) estudaram os padrões de raios que causaram incêndios em Ontário de 1976 a 1998. Os autores usaram a função K para avaliar o agrupamento e a suavização da densidade do kernel para fornecer representações gráficas do agrupamento. Wotton e Martell (2005) desenvolveram um modelo logístico para estimar a probabilidade de ocorrência de incêndios causados por raios, considerando variáveis meteorológicas e índices relacionados ao risco de incêndio. A análise de padrões pontuais também tem sido amplamente aplicada no campo da epidemiologia para entender a relação entre eventos de saúde e aspectos relacionados a características individuais (por exemplo, genética, comportamento e demografia) e fatores contextuais (por exemplo, as condições socioeconômicas

de os arredores e o ambiente físico), em que algoritmos de agrupamento são aplicados para detectar grupos espaciais de doenças (WERNECK, 2008; PULLAN *et al.*, 2012). Neste contexto, Bousema *et al.* (2010) usaram estatísticas de agrupamento para mostrar que agrupamentos de indivíduos HIV-positivos podem ser usados para identificar pontos críticos com alta incidência de malária na Tanzânia. Cook *et al.* (2011) empregaram varredura espacial para explorar o agrupamento de infecções por malária em diferentes faixas etárias na Ilha de Bioko, na Guiné Equatorial.

Recentemente, tem-se assistido a uma classe emergente de plataformas de comunicação, como Facebook, Twitter, FourSquare e Flickr, que permitem aos usuários publicar conteúdo georreferenciado, desde mensagens curtas de status até vídeos. Alguns estudos foram desenvolvidos usando essas informações de conteúdo para extrair padrões dos dados. Por exemplo, Huang e Zhang (2006) desenvolveram uma ferramenta que usa técnicas de agrupamento espacial para extrair padrões de dados de dispositivos móveis georreferenciados. Yuan e Raubal (2011) focaram na extração de atividades e mobilidade humana com base em dados de dispositivos móveis. Zhao *et al.* (2015) realizoram uma análise estatística que visa identificar padrões espaciais gerais na ocorrência de tweets relacionados a inundações que podem estar associados à proximidade e gravidade de eventos de inundação. Campelo, Baptista e Jerônimo (2016) apresentoram uma abordagem baseada no algoritmo de agrupamento DBSCAN (ESTER *et al.*, 1996) para extrair padrões de mobilidade de mensagens do Twitter e, em seguida, analisar sua correlação com dados demográficos, econômicos e sociais. Picornell *et al.* (2019) apresentoram uma nova metodologia baseada em algoritmos de agrupamento para estimar a dinâmica populacional a partir de dados de telefones celulares.

No contexto dos crimes, os dados estão cada vez mais disponíveis para a população, o que tem chamado a atenção de pesquisadores para tentar entender os padrões criminais de um determinado local. Registros criminais são geralmente dados georreferenciados; portanto, várias abordagens foram propostas para extrair padrões espaciais desses dados. Por exemplo, o sistema COPLINK (CHEN *et al.*, 2003) é reconhecido como uma das implementações bem-sucedidas da técnica de agrupamento para mineração de dados criminais, na qual o conceito de espaço é aplicado para identificar relacionamentos entre suspeitos e vítimas (HOU *et al.*, 1994) . Brown (1998) desenvolveu uma estrutura chamada ReCAP (Regional Crime Analysis Program) que combina múltiplas fontes de dados para extrair padrões. Chen *et al.* (2004) desenvolveram uma ferramenta geral que permite descobrir associações, identificar padrões e fazer previsões. Bruin *et al.* (2006) descreveram uma ferramenta capaz de extrair características importantes do banco de dados georreferenciado e criar perfis digitais para todos os agressores. Seu método compara todos os indivíduos nesses perfis usando uma nova medida de distância e os agrupa. Malathi e Baboo (2011) desenvolveram uma ferramenta para detectar padrões de crime com base em técnicas de mineração de dados, como agrupamento e classificação. Aljanabi (2011) propôs uma ferramenta para análise de dados criminais que se concentra na identificação de padrões e tendências que usam árvores de decisão para classificar os dados juntamente com algoritmos de

agrupamento. Joshi, Sabitha e Choudhury (2017) empregaram o algoritmo *k-means* (JIN; HAN, 2010) para extrair padrões criminais de grandes conjuntos de dados georreferenciados.

Os trabalhos descritos nesta Seção visam identificar padrões espaciais em dados georreferenciados e a maioria deles utiliza algoritmos de agrupamento padrão para este fim. No entanto, esses algoritmos requerem parâmetros de ajuste do usuário, uma tarefa complicada quando o usuário não tem conhecimento prévio sobre os dados, especialmente em dados ruidosos do mundo real.

2.2 Decomposição de tensor para processamento de dados

Uma das modelagens mais comuns de processamento de dados é por meio de tensores de 2ª ordem (ou seja, matrizes). Peng *et al.* (2012) transformaram viagens de táxi em uma matriz de fuso horário e aplicou um modelo *Non-Negative Matrix Factorization* (NMF) (LEE; SEUNG, 1999) para identificar padrões espaciais e temporais básicos da mobilidade humana. Recentemente, Garcıa *et al.* (2021) projetaram funcionalidades analíticas visuais que permitem aos usuários selecionar e analisar regiões de interesse em termos de pontos de acesso espaço-temporais e padrões de crime com base no NMF. No entanto, em muitos casos, encontramos limitações na modelagem das relações entre os dados com seu uso. Muitos dados são gerados por processos complexos, difíceis de representar sem estruturas adequadas. Portanto, há a necessidade de novos modelos para melhor compreensão desses dados.

O interesse em aplicar tensores na análise de dados tem crescido tanto na academia quanto na indústria. Tensores são arranjos multimodais capazes de representar os vários aspectos dos dados como modos independentes. Existem vários métodos de decomposição tensorial, como *Canonical Polyadic Decompositions* (CPD) (HARSHMAN *et al.*, 1970), *Tensor Train Decomposition* (TTD) (OSELEDETS, 2011) e *Tucker Decompositions* (TD) (KOLDA; BADER, 2009). Cada uma dessas fatorações tem características diferentes e propósitos diferentes quanto à aplicação. Por exemplo, o CPD tem sido usado no processamento de sinais e análise de dados, principalmente por causa de sua facilidade de interpretação de dados (VASILE, 2008). TTD tem sido usado para resolver problemas como equações diferenciais parciais estocásticas e equações elípticas de alta dimensão (OSELEDETS, 2011). Em contraste, TD tem sido usado com mais frequência em aplicações envolvendo compressão de dados (CICHOCKI *et al.*, 2015). Recentemente, Liang et al. (LIANG *et al.*, 2022) propôs uma estrutura de previsão que explora correlações espaço-temporais em dados criminais usando uma técnica de aprendizado de tensor que encontra a solução ótima resolvendo um problema de otimização compacto modelado com CPD.

Por outro lado, em algumas aplicações é essencial manter a propriedade *não negativa* de um tensor (ou seja, se todas as suas entradas forem não negativas) para preservar a interpretabi-

lidade dos dados. Para isso, é necessário introduzir as restrições de não negatividade, e assim tem-se a *Non-Negative Tucker Decomposition* (NTD) (CICHOCKI *et al.*, 2009). A NTD foi aplicado à extração de recursos em muitos contextos. Por exemplo, Phan e Cichocki (2010) propuseram uma redução de modelo e extração de características para problemas de grande escala. Jukić, Kopriva e Cichocki (2013) propuseram um método de extração de características adequado para imagens médicas coloridas.

Zhang *et al.* (2013) propuseram uma estrutura baseada em NTD para extrair padrões de mobilidade urbana de um enorme conjunto de dados de trajetória de táxi em Pequim. Analogamente, Sun e Axhausen (2016) aplicaram um modelo NTD para decompor dados de mobilidade de alta dimensão em um padrão significativo. Wang *et al.* (2022) forneceram uma estrutura para revelar os padrões espaço-temporais da mobilidade urbana, explorando dados maciços e de alta dimensão de telefones celulares.

Apesar de todos os trabalhos descritos acima usarem a decomposição tensorial para extrair os padrões, eles não lidam diretamente com dados urbanos de múltiplas fontes de dados, pois os padrões urbanos requerem análise e compreensão de uma combinação das características significativas de cada modo.

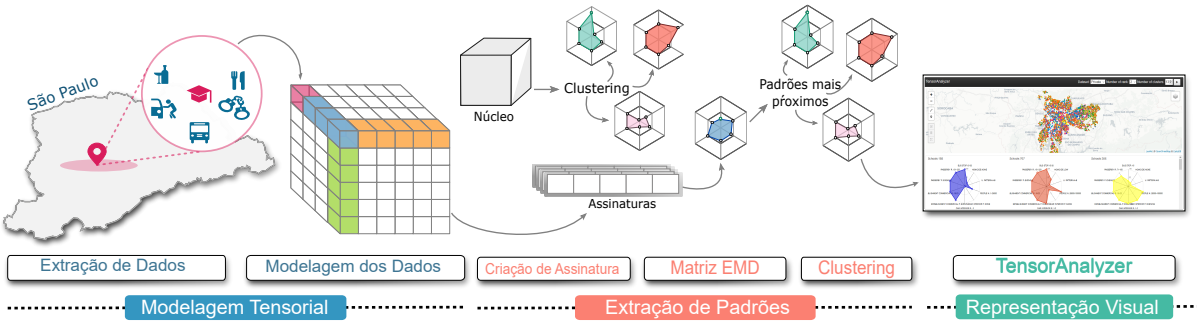


Figura 2 – A visão geral de todo o processo analítico: partiu-se da combinação de diferentes fontes de dados em um tensor; após a modelagem, aplicou-se fatoração tensorial não negativa para extrair padrões relevantes; por fim, os padrões foram analisados e representados usando alguns recursos visuais.

CAPÍTULO 3

METODOLOGIA

Muitas aplicações de ciências de dados geram grandes quantidades de dados com múltiplas features e alta dimensionalidade para os quais tensores fornecem uma modelagem direta (KOLDA; BADER, 2009). Em nossas aplicações, pode-se sumarizar um grupo específico de registros de uma consulta usando uma função de agregação `count`. For exemplo, uma possível agregação para um tipo de crime poderia ser selecionar todos os seus registros em uma região específica e sumarizá-los por contagem. Assim, o tensor resultante destas agregações é não-negativo. Por outro lado, extrair padrões escondidos dos dados do tensor tem um significado concreto apenas quando o tensor é não-negativo. Na prática, fatorações não-negativas são necessarias quando os componentes subjacentes têm um significado físico ou comportamental. Fatorações de tensor não-negativa fornecem uma ferramenta poderosa para descobrir padrões relevantes em conjuntos de dados multidimensionais, proporcionando representações de dados compactas (redução dos dados), e preservar a interpretabildade dos dados (não-negatividade) A metodologia proposta para detecção urbana no entorno de um lugar usando fatoração não-negativa de tensor compreende três principais passos (ver Fig. 2): modelagem do tensor, extração de padrões, e representação visual. Antes de descrever a metodologia desenvolvida, introduz-se

alguns conceitos básicos sobre fatoração não-negativa de tensor.

3.1 Decomposição de Tucker Não-Negativa

Entre as decomposições de tensores mais conhecidas, a NTD pode ser considerada uma generalização de ordem superior de Non-Negative Matrix Factorization (NMF), um método de sucesso para detectar características fundamentais dos dados (LEE; SEUNG, 1999).

Matematicamente, um tensor $\mathscr{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_n}$ de ordem n é um array multidimensional, onde um elemento de $\mathscr{X}$ é indexado por uma n-tupla de índices $(i_1, i_2, \ldots, i_n)$ e denotado por $x_{i_1 i_2 \ldots i_n}$. Cada dimensão i_k indexada corresponde a um modo do tensor e varia de 1 até I_k. O valor $I_k \in \mathbb{N}$ é o tamanho do k-modo. Por exemplo, um tensor de primeira-ordem é um vetor, um tensor de segunda-ordem é uma matriz, e tensores de ordem três ou maior são chamados tensores de alta-ordem.

Matrizes e vetores podem ser multiplicados por tensores. Para o propósito deste trabalho, usa-se o produto de matriz de modo k, ou seja, uma multiplicação ao longo do modo k de um tensor com uma matriz. Considere um tensor $\mathscr{X} \in \mathbb{R}^{I_1 \times \cdots \times I_n}$ e uma matriz $\mathbf{A} \in \mathbb{R}^{I_k \times J}$, o produto de matriz de modo k, denotado por $\times_k$, é definido por:

$$(\mathscr{X} \times_k \mathbf{A})_{i_1 \ldots i_{k-1} j i_{k+1} \ldots i_n} = \sum_{i_k=1}^{I_k} x_{i_1 \ldots i_n} a_{j i_k} .$$

Além disso, de forma análoga à norma de Frobenius da matriz, uma norma de tensor é dada por:

$$\|\mathscr{X}\|_F = \sqrt{\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \cdots \sum_{i_n=1}^{I_n} x_{i_1 i_2 \ldots i_n}^2} .$$

Um tensor $\mathscr{X}$ é não-negativo e denotado por $\mathscr{X} \geq 0$, se todas as suas entradas são não-negativas. Assim, dado um tensor não-negativo $\mathscr{X}$, a NTD retorna um núcleo não-negativo $\mathscr{G} \in \mathbb{R}^{J_1 \times \cdots \times J_n}$ e matrizes não-negativas $\mathbf{F}^{(1)} \in \mathbb{R}^{I_1 \times J_1}, \ldots, \mathbf{F}^{(n)} \in \mathbb{R}^{I_n \times J_n}$, tal que

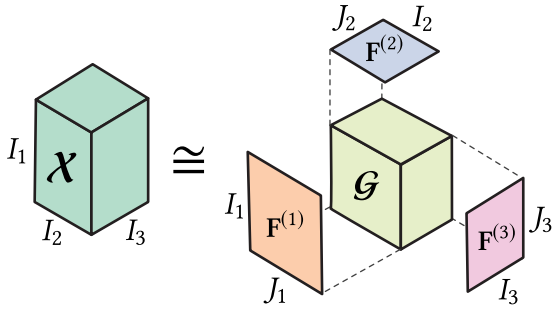
$$\mathscr{X} \cong \widehat{\mathscr{X}} \quad \text{com} \quad \widehat{\mathscr{X}} = \mathscr{G} \times_1 \mathbf{F}^{(1)} \times_2 \cdots \times_n \mathbf{F}^{(n)} . \tag{3.1}$$

o rank J_k dos modos do núcleo são usualmente escolhidos menores do que o rank I_k dos modos do tensor de entrada original. Figura 3 mostra uma NTD de um tensor de ordem três.

Na verdade, a NTD é obtida resolvendo o problema de otimização:

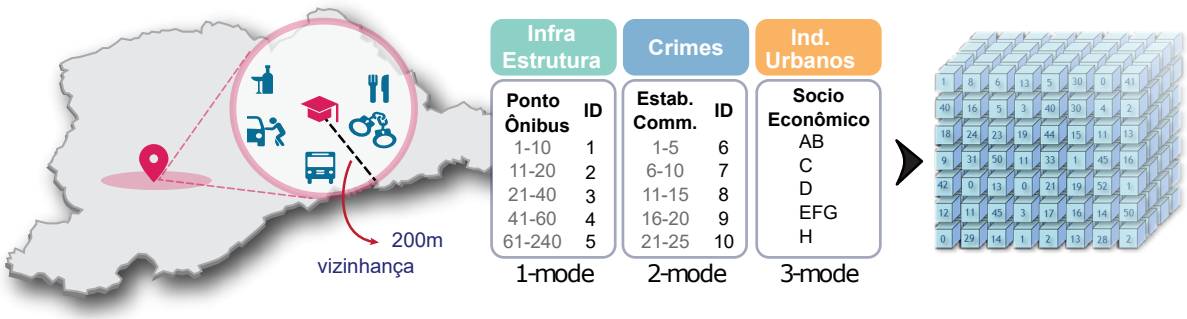
$$\min_{\widehat{\mathscr{X}}} \frac{1}{2} \|\mathscr{X} - \widehat{\mathscr{X}}\|_F^2 , \quad \text{sujeito a} \quad \mathscr{G} \geq 0, \mathbf{F}^{(1)} \geq 0, \ldots, \mathbf{F}^{(n)} \geq 0 .$$

O núcleo $\mathscr{G}$ é uma versão comprimida de $\mathscr{X}$ devido a esparsidade introduzida pelas restrições da não-negatividade. Algoritmos eficientes para resolver o problema de otimização podem ser encontrados em Friedlander e Hatz (2008), Xu e Yin (2013).



Fonte: Silveira *et al.* (2022).

Figura 3 – Ilustração de uma NTD para um tensor de ordem 3 $\mathscr{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$. O principal objetivo é encontrar as matrizes de fator ótimo $\mathbf{F}^{(k)} \in \mathbb{R}^{I_k \times J_k}$ com $k = 1,2,3$ e um núcleo $\mathscr{G} \in \mathbb{R}^{J_1 \times J_2 \times J_3}$, tipicamente $J_k \ll I_k$.



Fonte: Adaptada de Silveira *et al.* (2022).

Figura 4 – Modelagem de tensor de dados com 3 modos. Cria-se uma ROI circular centrada em cada local de destino com um raio de 200 metros (*esquerda*). Conta-se o número de crimes, classes sociais e instalações de infraestrutura em cada ROI (*meio*). As entradas do tensor não negativo resultante $\mathscr{X}$ são o número de lugares com as mesmas características (*right*).

3.2 Modelagem de tensor de dados

Modelar o tensor de dados de entrada dos sites é o primeiro passo da presente metodologia. Inicia-se agrupando os locais em regiões circulares de interesse (ROIs) não sobrepostas centralizadas em cada local de destino. A característica de cada ROI é agregada contando o número de crimes e instalações de infraestrutura ou determinando os indicadores urbanos mais comuns. O raio ROI é um parâmetro de usuário; nas análises desse trabalho emprega-se um raio de 200 metros.

As agregações resultam em um tensor de dados de 10ª ordem, onde seus modos são organizados da seguinte forma:

- seis modas a partir dos dados criminais: número de registros criminais (transeunte, estabelecimento comercial e veículos) e seu período de maior frequência;
- dois modos de dados urbanos: um indicador urbano mais prevalente (socioeconomia e homicídios);

- dois modos de dados de infraestrutura: número de instalações de infraestrutura (paradas de ônibus e fluxo de pessoas).

Para reduzir os tamanhos dos modos relacionados aos tipos de crime e infraestrutura, usamos histogramas equalizados, ou seja, o tamanho do modo é o número de bins do histograma. Os intervalos bin e os valores categóricos são convertidos em valores numéricos (IDs) por meio de uma função de codificação de rótulo $\varphi_k(i)$, que recupera o ID associado com a i-ésima feature do modo k. Por exemplo, a Fig. 4 mostra a modelagem de um tensor de dados de ordem 3 de um lugar, onde $\varphi_3(2)$ é o ID da classe socioeconômica C. Finalmente, as entradas do tensor de dados $\mathscr{X}$ são o número de lugares com as mesmas características de dados, como ilustrado pela Fig. 4.

3.3 Extração de padrões

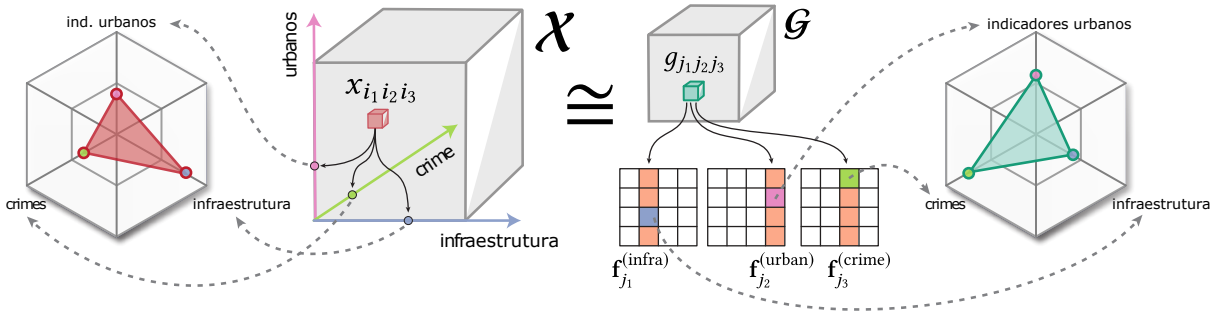
Após criar um tensor de dados $\mathscr{X}$, extrai-se os padrões de dados de sua NTD dada pela Equação (3.1). O tensor core $\mathscr{G} \in \mathbb{R}^{J_1 \times \cdots \times J_n}$ revela padrões significativos, onde a magnitude das sua entradas está diretamente relacionada com a importância das features codificadas pelas matrizes fatores $\mathbf{F}^{(k)}$ e sua interação. Primeiramente, extrai-se os padrões relevantes dos dados usando o tensor core $\mathscr{G}$. Em seguida, atribuímos cada padrão de $\mathscr{X}$ ao seu padrão vizinho mais próximo de $\mathscr{G}$. Seja $g_{j_1 \ldots j_n}$ uma entrada diferente de zero de $\mathscr{G}$, cria-se uma assinatura de padrão como um vetor de recursos usando o atributo mais significativo das colunas $\mathbf{f}^{(k)}_{j_k}$, como segue:

$$\mathbf{s}_{\mathscr{G}} = (\varphi_1(i_1), \ldots, \varphi_k(i_k), \ldots, \varphi_n(i_n))^\top \quad \text{with} \quad i_k = \arg\max_i f^{(k)}_{i j_k}. \tag{3.2}$$

Contudo, o tensor core fornece muitos padrões similares. Assim, é necessário agrupar e reduzir esses padrões em um pequeno conjuto de dados. As assinaturas dos padrões são particionadas em M clusters altamente relacionados usando o algoritmo *Agglomerative Hierarchical Clustering* (AHC) (RAFSANJANI; VARZANEH; CHUKANLO, 2012) onde a matriz de distâncias é calculada usando *Earth Mover's Distance* (EMD) (ZEN; RICCI; SEBE, 2014) como uma medida de similaridade que tem em conta a semântica entre as assinaturas. Finalmente, para cada cluster C, a assinatura $\mathbf{s}_{\mathscr{G}} \in C$ que representa o cluster inteiro é aquele que minimiza $\sum_{\mathbf{s} \in C} d_{\text{EMD}}(\mathbf{s}, \mathbf{s}_{\mathscr{G}})$, where d_{EMD}, onde d_{EMD} é a EMD. Em simples palavras, a assinatura do padrão $\mathbf{s}_{\mathscr{G}}$ é o centróide de C no sentido EMD.

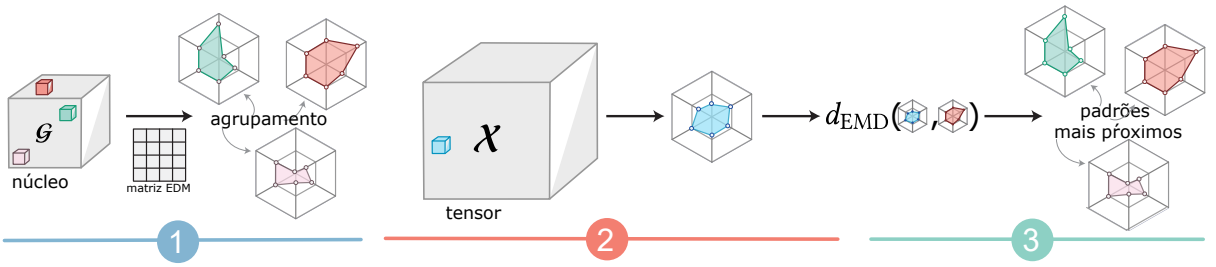
Na presente aplicação, o número de clusters M é um parâmetro de usuário. Além disso, os ranks modo de $\mathscr{G}$ são englobados em um único parâmetro $J \in \mathbb{N}$ (ou seja, $J_1 = \cdots = J_n = J$) também definido pelo usuário.

Próximo, precisa-se criar uma assinatura dos padrões de entrada do tensor $\mathscr{X}$ e atribuí-los a um padrão $\mathbf{s}_{\mathscr{G}}$. Para cada entrada não-zero $x_{i_1 \ldots i_n}$ of $\mathscr{X}$, sua assinatura é alcançada diretamente ao aplicar φ_k em cada modo k, ou seja, $\mathbf{s}_{\mathscr{X}} = (\varphi_1(i_1), \ldots, \varphi_n(i_n))^\top$.



Fonte: Adaptada de Silveira *et al.* (2022).

Figura 5 – Representação visual dos padrões do tensor. Os padrões do tensor de entrada $\mathscr{X}$ são representados diretamente por um gráfico de radar (*no extremo esquerdo*) onde cada ponto de ancoragem é definido pela assinatura $\mathbf{s}_{\mathscr{X}}$. Enquanto as entradas do núcleo do tensor $\mathscr{G}$ fornecem a melhor combinação de fatores de cada matriz de recursos. Esses fatores são combinados para formar uma assinatura $\mathbf{s}_{\mathscr{G}}$ também representada por um gráfico de radar (*na extrema direita*).



Fonte: Adaptada de Silveira *et al.* (2022).

Figura 6 – Processo de extração de padrões: (1) os padrões mais relevantes retornados por $\mathscr{G}$ são agrupados, (2) padrões são extraídos de $\mathscr{X}$, e (3) cada padrão de $\mathscr{X}$ é atribuído ao seu padrão mais próximo de $\mathscr{G}$ em termos de EMD.

Finalmente, atribui-se $\mathbf{s}_{\mathscr{X}}$ ao padrão mais similar do tensor core $\mathscr{G}$. Para este objetivo, encontra-se o vizinho mais próximo $\mathbf{s}_{\mathscr{G}}$ em termos de EMD.

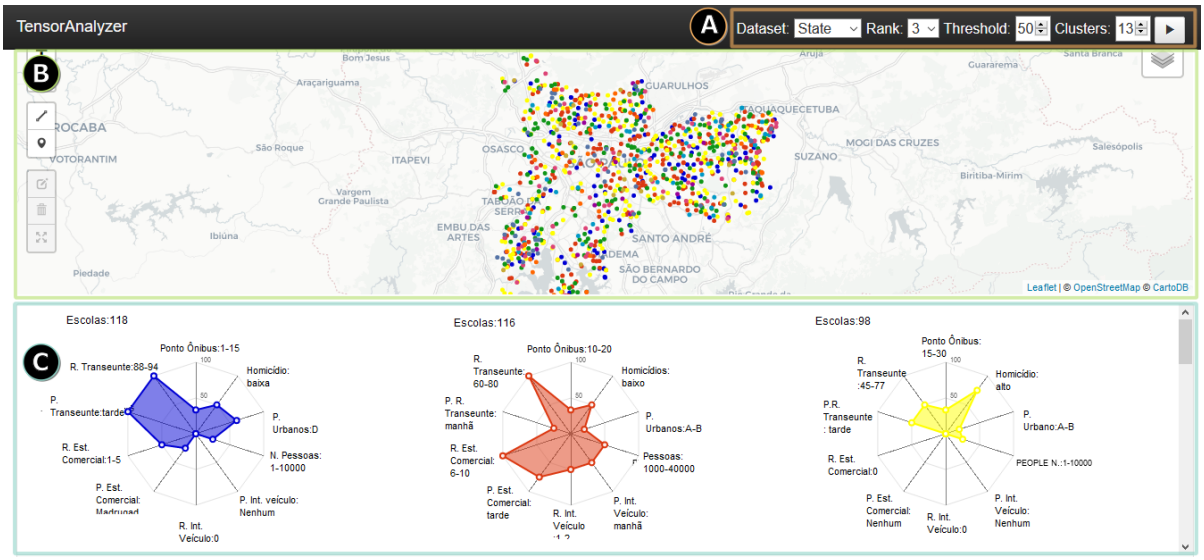
Os padrões são representados visualmente por um simples *radar chart* onde cada

The patterns are represented visually by a simple *radar chart* onde cada ponto de ancoragem representa um modo do tensor, conforme ilustrado por Fig. 5. Os valores em cada modo definem a forma do polígono, produzindo uma identidade visual para a assinatura do padrão. Todo o processo de extração de padrão é resumido por Fig. 6.

3.4 Design Visual

Nesta Seção, descreve-se os componentes visuais do TensorAnalyzer. Na Fig.7 observa-se um sistema baseado na web, o qual compreende um *Menu de Controle* e duas janelas de visualização. Os especialistas do domínio assistiram o presente trabalho no design de cada recurso visual. *Map View* mostra a geolocalização das localizações alvo e seu padrão associado,

enquanto *Patterns View* mostra os padrões (gráfico de radar) extraídos pela presente metodologia.



Fonte: Adaptada de Silveira *et al.* (2022).

Figura 7 – Sistema TensorAnalyzer: a visualização de padrões permite a compreensão do relacionamento entre crimes e outras variáveis envolvidas nas análises. A ferramenta compreende um *Control Menu* (A), *Map View* (B) e *Patterns View* (C).

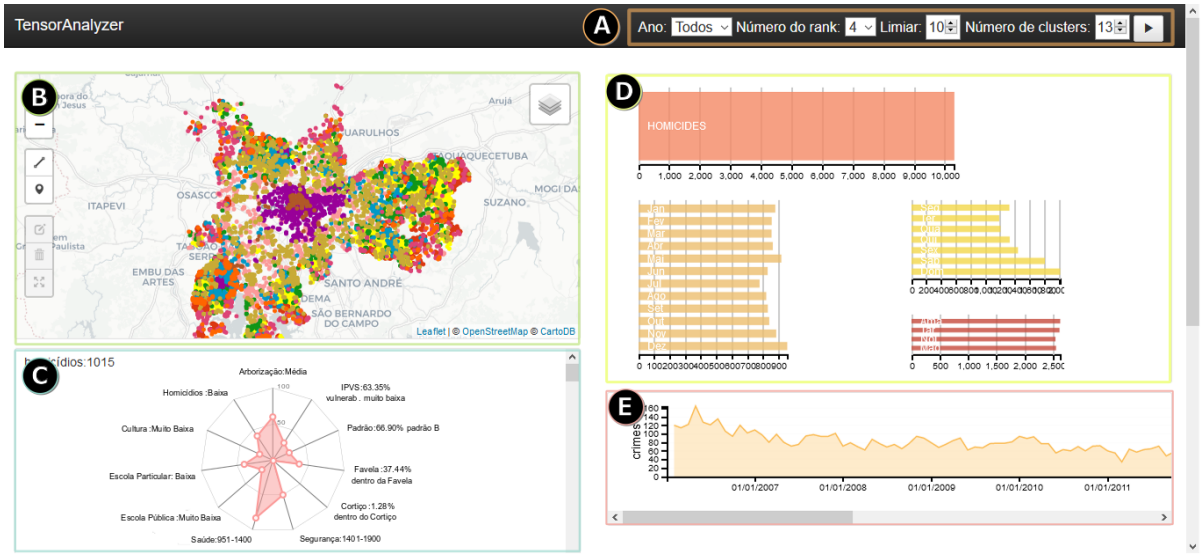
Técnicas de visualização têm sido empregadas com sucesso na análise de crimes para explorar os dados (GARCIA *et al.*, 2021; LIANG *et al.*, 2022). Contudo, novos designs sobre as visualizações existentes são requeridas. Por exemplo, a *Patterns View* é uma nova alternativa de visualização nesse contexto, o que acaba por ajudar a elucidar a relação entre o crime e outras características (indicadores urbanos e instalações de infraestrutura) envolvidas nas análises. Embora a comunidade de visualização de informações conheça bem essa metáfora visual, até onde se sabe, ela nunca foi usada para análise de dados urbanos. Nos parágrafos seguintes, descreve-se cada componente visual.

Menu de controle. Esta barra de ferramentas fornece as seguintes seleções do usuário: o conjunto de dados de entrada, a classificação do modo J para o tensor central da etapa NTD e o número de clusters de padrão M.

Map View. Cada ponto representa um local de destino com sua geolocalização e sua cor corresponde ao padrão associado. Essa visualização permite que os usuários identifiquem e explorem quais sites estão em um padrão específico.

Patterns View. Esta visualização permite ao usuário visualizar todos os padrões detalhados e a quantidade de escolas com as mesmas características. Além disso, é possível selecionar um padrão específico e visualizar apenas os sites que pertencem a ele.

A fim de atender os objetivos especificados pelos especialistas na Subseção 4.4.1 novas visualizações foram adicionadas no TensorAnalyzer. Na Figura 8 pode-se observar o sistema web modificado para apresentar os resultados deste estudo de caso, em que adicionou-se duas



Fonte: Elaborada pelo autor.

Figura 8 – Sistema TensorAnalyzer atualizado: a visualização de padrões combinada com outros gráficos permitem uma exploração temporal dos padrões para uma melhor compreensão da relação entre crimes e a melhoria na infraestrutura da cidade. Além das funcionalidades já mencionadas, foi incrimentado na ferramenta uma Visualização de Histograma Sazonal (D) e uma Visualização Temporal (E).

novas componentes visuais, Visualização de Histograma Sazonal (D) e Visualização Temporal (E), para auxiliar na análise exploratória dos dados de crime a partir de um padrão selecionado. Nos parágrafos seguintes, descreve-se as novas componentes visuais.

Visualização de Histograma Temporal. É composto por quatro gráficos de barras que detalham a distribuição do crime no período do dia, dia da semana e meses do ano. Na parte superior, a barra em vermelho mostra o volume de crimes no padrão selecionado. Na parte inferior, no histograma esquerdo mostra os dados agregados por mês, enquanto à direita, observa-se gráficos de barras em que pode-se visualizar os dados por dia da semana e por período do dia. Cada gráfico de barras serve como um filtro, ou seja, se um usuário clicar em uma das barras, todas as visualizações serão atualizadas para mostrar apenas as ocorrências de crime de acordo com o filtro selecionado.

Visualização Temporal. Esta visualização usa um gráfico de linha interativo para exibir como os eventos de crime evoluem no padrão selecionado. Além disso, essa visualização pode filtrar o foco da análise em um intervalo específico de tempo, que pode ser selecionado arrastando uma janela de intervalo ao longo do eixo horizontal. Depois que um novo intervalo de tempo é selecionado, todas as visualizações vinculadas são atualizadas.

Com a implementação das novas componentes foi possível atender os objetivos dos especialistas descritos na Subseção 4.4.1. Na Seção seguinte descreve-se as análises realizadas utilizando a fim de validar as hipóteses descritas no Capítulo 1 e direcionar os especialistas quanto as decisões a serem tomadas.

CAPÍTULO

4

RESULTADOS

Para validar a eficácia da presente metodologia, apresenta-se um conjunto de comparações usando conjuntos de dados sintéticos e estudos de caso de conjuntos de dados reais da cidade de São Paulo. Em primeiro lugar, mostramos os padrões identificados pelo TensorAnalyzer a partir de dados ruidosos (Seção 4.1) e uma comparação com abordagens de agrupamento de vetores de recursos (Seção 4.2). Em seguida, realizamos estudos de caso para verificar as realizações dos especialistas apresentados na Seções 4.3 e 4.4.

A interface visual foi implementada em JavaScript usando `D3` (Bostock; Ogievetsky; Heer, 2011). Para calcular a NTD, empregou-se a biblioteca Python `Tensorly` (KOSSAIFI *et al.*, 2019). Para o AHC, empregou-se sua implementação fornecida pelo `Scikit-learn` (PEDREGOSA *et al.*, 2011). Todos os experimentos e pré-processamento de dados foram conduzidos em um Intel i7-4980HQ de 4 núcleos e 2,8 GHz com 16 GB de RAM.

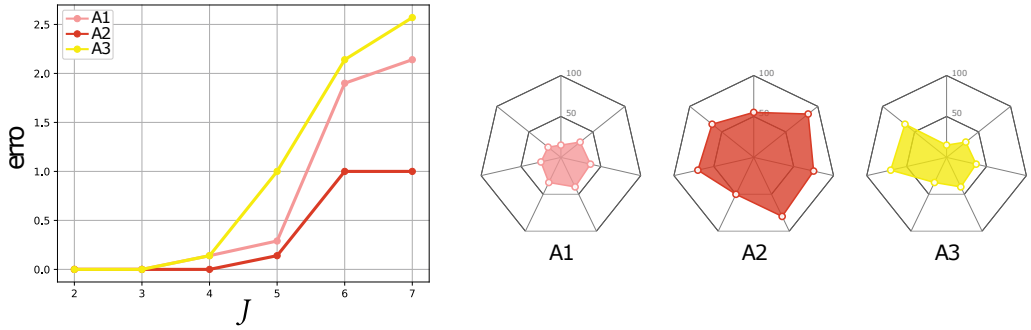
4.1 Identificando padrões com NTD

Por meio da NTD e sua robustez foi possível lidar muito bem com o ruído para identificar padrões. Para simplificar a discussão, apresenta-se a abordagem proposta usando exemplos sintéticos. Em particular, extrai-se amostras aleatórias de uma distribuição normal n dimensional $\mathcal{N}(\mu,\Sigma)$, onde o vetor médio μ é tomado como uma assinatura de padrão de destino (verdade fundamental) e a matriz de covariância Σ é a matriz de identidade.

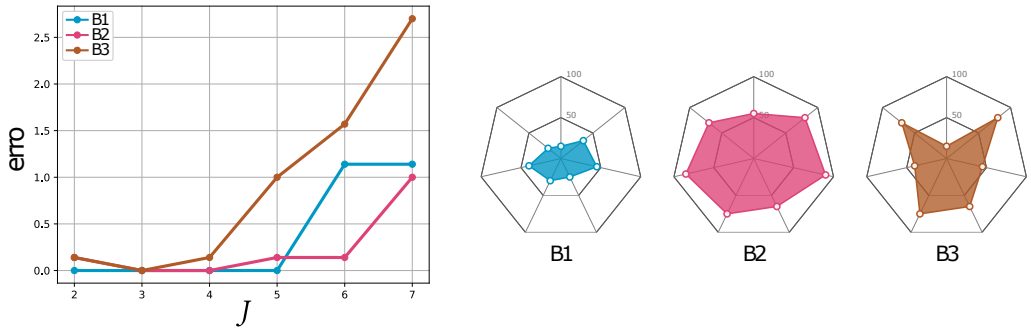
Durante os experimentos, considerou-se três clusters ($M = 3$) em $\mathbb{R}^7$ com diferentes números de amostras e intensidades de ruído. Essa construção simula três regiões onde os dados urbanos foram coletados e deseja-se extrair o padrão de destino de cada cluster usando a abordagem NTD. A Fig. 9 mostra a análise de erro em relação ao rank J em três cenários diferentes, em que o erro é definido por d_{EMD} entre a assinatura alvo e o padrão fornecido por TensorAnalyzer. Como pode ser visto, os valores de classificação baixa podem recuperar padrões

Figura 9 – A análise de erros do TensorAnalyzer em relação ao rank J, o número de amostras e a presença de ruído nos dados.

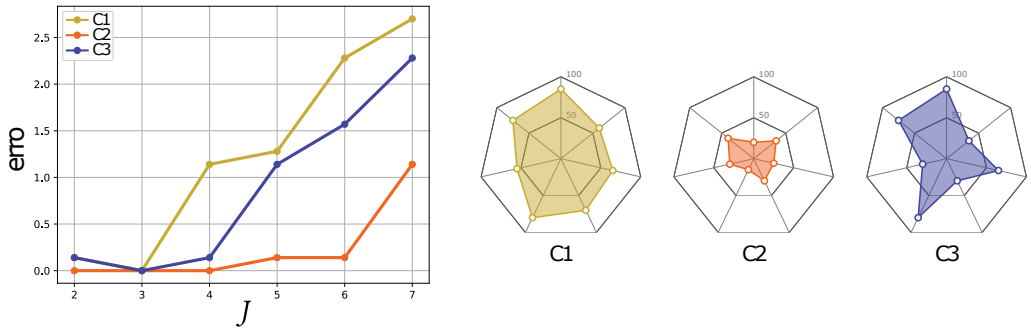
Fonte: Silveira *et al.* (2022).



(a) Padrões aleatórios de uma distribuição normal em torno dos padrões alvo A1, A2 e A3: 500 amostras para cada padrão com 10%, 13% e 15% de ruído para A1, A2 e A3, respectivamente.



(b) Padrões aleatórios de uma distribuição normal em torno dos padrões alvo B1, B2 e B3: 500, 1000, 1500 amostras com 10% de ruído para B1, B2 e B3, respectivamente.



(c) Padrões aleatórios de uma distribuição normal em torno dos padrões alvo C1, C2 e C3: 500 amostras com 16% de ruído para C1, 1000 amostras com 11% de ruído para C2 e 1500 amostras com 13% de ruído para C3.

alvo de dados ruidosos. Além disso, um palpite inicial para a classificação é $J = \lfloor n/2 \rfloor$.

4.2 Comparação com algoritmos de agrupamento

Nesta Seção, compara-se a presente metodologia com técnicas de agrupamento bem conhecidas, como k-means e AHC (sem NTD). Neste experimento, gerou-se um conjunto de dados sintéticos de entrada em $\mathbb{R}^6$ organizados em dez clusters de tamanhos diferentes. Assim, as observações são amostradas aleatoriamente nas regiões $\Omega_k \mathbb{R}^6$ e atribuídas a cada cluster C_k,

conforme detalhado pela Tabela 2.

Fonte: Silveira *et al.* (2022).

Tabela 2 – Conjunto de dados sintéticos com 10 clusters.

cluster	**#elementos**	**região Ω_k**
C_1	500	$[1,10] \times [1,5] \times [1,30000] \times [1,50] \times [1,60] \times [1,7]$
C_2	600	$[11,20] \times [6,10] \times [31000,60000] \times [51,70] \times [61,90] \times [8,11]$
C_3	700	$[21,30] \times [11,15] \times [61000,80000] \times [71,90] \times [91,120] \times [12,16]$
C_4	800	$[31,40] \times [16,21] \times [81000,100000] \times [91,120] \times [121,150] \times [17,21]$
C_5	900	$[41,50] \times [22,30] \times [101000,130000] \times [121,150] \times [151,180] \times [22,27]$
C_6	10	$[51,60] \times [1,5] \times [61000,80000] \times [91,120] \times [121,150] \times [22,27]$
C_7	10	$[21,30] \times [31,40] \times [1,30000] \times [71,90] \times [151,180] \times [17,21]$
C_8	10	$[31,40] \times [11,15] \times [131000,150000] \times [121,150] \times [91,120] \times [12,16]$
C_9	10	$[41,50] \times [1,5] \times [101000,130000] \times [151,180] \times [61,90] \times [17,21]$
C_{10}	15	$[1,10] \times [6,10] \times [61000,80000] \times [91,120] \times [121,150] \times [22,27]$

Considerando o conhecimento do terreno, ou seja, conhecendo os rótulos dos clusters com antecedência, comparou-se as duas abordagens usando medidas tradicionais para avaliar o desempenho do clustering (PEDREGOSA *et al.*, 2011): *Fowlkes-Mallows index* (FMI), *Índice de Rand Ajustado* (ARI), *V-measure*, e *Informação Mútua* (MI). Embora as medidas sejam derivadas da matriz de confusão, elas têm significados diferentes. Por exemplo, FMI e ARI são construídos na contagem de pares de objetos classificados de forma semelhante em ambos os agrupamentos. Por outro lado, V-measure e MI são baseados na análise de entropia condicional. As pontuações de todas as medidas variam de 0,0 a 1,0, e pontuações mais altas levam a melhores resultados.

A Tabela 3 mostra os resultados da comparação entre TensorAnalyzer (com classificação $J = 3$), k-means e AHC aplicados no conjunto de dados sintéticos fornecidos por Table 2. Como pode ser visto, o TensorAnalyzer supera o k-means e o AHC, alcançando pontuações mais altas em todas as medidas.

Fonte: Silveira *et al.* (2022).

Tabela 3 – Comparação do nosso TensorAnalyzer com o agrupamento de vetores de recursos usando k-means e AHC. Os melhores resultados em negrito.

Métricas	**k-means**	**AHC**	**TensorAnalyzer**
FMI	0.6149	0.7987	0.8754
ARI	0.5163	0.7475	0.8527
V-measure	0.7248	0.8485	0.8790
MI	0.6254	0.7578	0.8741

A seguir avalia-se a metodologia proposta por meio de estudos de casos.

4.3 Estudo de caso: criminalidade nas escolas de São Paulo

Nesta Seção apresenta-se um estudo de caso com o objetivo de validar a metodologia proposta e compreender os padrões no entorno de escolas, tendo em conta um conjunto de dados reais da região metropolitana de São Paulo. Mais especificamente, o presente estudo de caso busca verificar alguns objetivos descritos na Subseção 4.3.1 levantados por especialistas, de maneira que a verificação desses objetivos permite validar as hipóteses apresentadas no Capítulo 1.

Uma interface visual foi implementada em *JavaScript* usando `D3` (Bostock; Ogievetsky; Heer, 2011). Para calcular NTD, empregou-se a biblioteca `Tensorly` desenvolvida em *Python*. Para o AHC, empregou-se uma implementação fornecida pela `Scikit-learn` (PEDREGOSA *et al.*, 2011). Todos os pré-processamento de dados e experimentos foram conduzidos em um 4-core 2.8 GHz Intel i7-4980HQ com 16 GB de RAM.

Antes de descrever o estudo de caso em si, apresenta-se os objetivos levantados pelos especialistas e as hipóteses relacionadas. Além disso, uma breve descrição dos dados que foram fornecidos pelos especialistas é feita.

4.3.1 Objetivos e Dados

Primeiramente, foram realizadas várias reuniões e entrevistas com os especialistas para entender os principais objetivos que devem ser atendidos durante as análises dos dados criminais. O resultado destas interações pode ser resumido como segue:

O1 – Compreender as relações entre eventos criminais e outras variáveis envolvidas nas análises considerando regiões próximas às escolas. Os efeitos das características urbanas sobre o crime em cidades da América Latina são pouco estudadas (MONTOLIO, 2018). A compreensão deste relacionamento poderia informar ao planejamento urbano que ajuda a deter o crime. Por exemplo, quantificar as variações do crime e o relacionamento delas com o fluxo da população por meio de diferentes estações de ônibus, poderia ajudar os gestores de políticas públicas compreenderem o impacto que as modificações de infraestrutura urbana teriam no crime, quando estas forem realizadas próximo às escolas. Block e Block (2000) mostraram o efeito de algumas variáveis sobre o aumento da criminalidade no Bronx, e uma das descobertas foi que paradas de metrô podem aumentar quatro vezes os roubos. Com base nos estudos descritos anteriormente, especialistas do NEV gostariam de elucidar se existe uma relação entre crimes e pontos de ônibus próximos às escolas da Região Metropolitana de São Paulo e, consequentemente, validar/rejeitar a hipótese **H1** descrita no Capítulo 1.

O2 – Análise de como a criminalidade influencia o desempenho dos estudantes. O estudo conduzido por pesquisadores das Universidade de Nova York, e da Universidade de DePaul (HEISSEL J. A .AND SHARKEY *et al.*, 2018), encontraram que crimes violentos mudam os padrões de sono dos estudantes que moram próximos a estes eventos criminais, o que aumenta a quantidade do hormônio do estresse cortisol no corpo dos estudantes já no dia seguinte após o incidente violento. Tanto a perturbação do sono quanto o aumento do cortisol mostraram que podem influenciar o desempenho dos estudantes na escolas de forma negativa. Neste contexto, especialistas gostariam de investigar se os crimes violentos tem afetado o desempenho dos estudantes em São Paulo e quais padrões representam este relacionamento. Em caso afirmativo estariam validando a hipótese **H2** descrita no Capítulo 1.

O3 – Compreender padrões criminais de áreas recreativas próximo às escolas.

De acordo com a experiência dos especialistas, parques e áreas de lazer são regiões de incidentes de crimes graves, incluindo agressões, furtos e agressões sexuais. Embora assaltos e roubos sejam crimes esperados nessas regiões durante o dia, altas taxas de crimes também têm sido reportadas durante a noite. Em geral, essas áreas recreativas estão próximas às escolas, e portanto, investigar os padrões de crime de áreas recreativas torna-se uma tarefa imperativa pelos especialistas, diante da proximidade destas regiões com as escolas. Este objetivo verifica a hipótese **H3** descrita no Capítulo 1.

4.3.1.1 Dados

A compreensão dos padrões criminais próximos as escolas de São Paulo tem sido um tópico de interesse de pesquisa de especialistas. De acordo com eles, a extração de padrões deve contar com um mecanismo para extrair padrões escondidos de múltiplas fontes de dados, permitindo o agrupamento das escolas de acordo com esses padrões. Para este propósito, foram realizadas várias reuniões com um renomado grupo de sociologistas do Estudos da Violência na Universidade de São Paulo (NEV-USP) (NEV, 2000). Um dos sociologistas tem uma extensa experiência em sociologia, ênfase em violência e conflitos sociais e urbanos. O outro sociologista é um especialista em geoinformação e sociologia aplicada a análise espacial, planejamento urbano, segurança pública, homicídio e dinâmica criminal. Em parceria com o departamento de polícia do estado de São Paulo, o time de especialistas construiu um conjunto de dados contendo registros criminais referentes a 12 anos em São Paulo.

O conjunto de dados reúne vários tipos de fontes de dados, como mostra a Tabela 4. O departamento de polícia de São Paulo forneceu registros criminais e apenas atos criminais como roubos foram fornecidos, deixando de fora crimes relacionados com drogas e agressão sexual. Cada registro contém as coordenadas geográficas do crime, o tipo do crime, a data e o período (madrugada, manhã, tarde e noite) da ocorrência do crime. O conjunto de dados contém registros de crime de 2006 até 2017 e três tipos de crimes: roubo transeunte, estabelecimento comercial e

Fonte: Silveira *et al.* (2022).

Tabela 4 – Esta Tabela resume o conjunto de dados usado durante as análises, a qual é composta por cinco colunas: os tipos de conjuntos de dados, o período em que o conjunto de dados varia, as categorias que o conjunto de dados foi dividido, as fontes de dados e a quantidade dos dados.

Dados	Anos	Categorias	Fonte	Quantidade
Crime	2006-2017	• Transeunte • Estab. Comercial • Veículo	(NEV, 2000)	1013400 6889 81299
Ind. Urbanos	2010	• Socioeconômicos (A-B,C,D,E-F-G,H) • Homicídios (baixo, alto, outros)	(NEV, 2000)	18953
Infraestrutura	2016 2011-2017	• Pontos de ônibus • Fluxo de Pessoas	(CEM, 2000) (GovernoSP, 2004)	99856 400M
Escolas	2016	• Particular • Estadual • Municipal	(CEM, 2000)	4088 1217 1539

veículo. Contudo, decidiu-se incluir apenas informação de 2011 até 2017 no presente estudo por causa do período de outro conjunto de dados como fluxo de pessoas. Os especialistas também criaram indicadores urbanos para algumas variáveis, tais como expansão urbana, mobilidade, habitação, homicídios e mudanças na população e no ambiente. Desta maneira, para cada indicador, eles atribuíram um rótulo para os sites. Também, os indicadores foram usados para criar um padrão urbano o qual agrupa os setores censitários de São Paulo em oito diferentes padrões urbanos. Os padrões urbanos variam da classe urbana A (onde a população tem o nível alto de saneamento) até a classe urbana H (onde a população tem o pior nível de saneamento básico). Visto que não existem registros de homicídios no conjunto de dados, optou-se por empregar o indicador de homicídios fornecido pelos especialistas do domínio nas análises.

Além do conjunto de dados de crime, nas análises, considera-se também mais fontes de dados, tais como infraestrutura, classes urbanas e fluxo de pessoas. Com relação a infraestrutura, estima-se a base de dados fluxo de pessoas, com base em bilhetes de ônibus, baixados do Portal da Transparência de São Paulo (GovernoSP, 2004). Extraiu-se os dados de pontos de ônibus e escolas do *website* Centro de Estudos da Metrópole (CEM) (CEM, 2000). O conjunto de dados ponto de ônibus tem um total de 99.856 registros. O conjunto de dados das escolas tem 11.931 registros e cada um contém informação a respeito do tipo da escola e os indicadores de desempenho dos estudantes variando de 2012 até 2015, tais como:

- *Exame Nacional do Ensino Médio* (ENEM) é um exame aplicado para todos os estudantes do Brasil para avaliar a qualidade do ensino médio e também é usado para selecionar os estudantes que entram nas universidades públicas;
- *Índice de Desenvolvimento Educacional do Ensino Fundamental nos Anos Iniciais* (IDE-

EFAI) e *Anos Finais* (IDEEFAF) é um indicador nacional para medir a qualidade da educação do ensino fundamental (estudantes do primeiro ao nono ano).

É importante elucidar que todo o conjunto de dados descrito acima é georreferenciado.

4.3.2 Resultados das Análises

Nesta Seção descreve-se os resultados obtidos das análises realizadas a fim de atender os objetivos **O1**, **O2** e **O3** especificados pelos especialistas.

4.3.2.1 O aumento da infraestrutura pode influenciar no aumento do crime (**O1**)

A Grande São Paulo é a maior região metropolitana da América Latina, com altos índices de criminalidade quando comparada a média mudial. É caracterizada por ser não-planejada e com alta desigualdade socioeconômica, sendo que os efeitos das características urbanas sobre o crime não foram estudados. Assim, a compreensão deste relacionamento poderia informar o planejamento urbano que ajuda a prevenir crimes. Na Figura 10 observa-se os padrões das escolas fornecidos pelo TensorAnalyzer.

A fim de investigar se existe relação entre crimes e infraestrutura, foi realizada uma validação para verficar se o aumento das instalações de infraestrutura pode influenciar no aumento dos crimes. Assim, para cada conjunto de dados, aplicou-se uma regressão linear sobre os padrões para explorar o relacionamento entre pontos de ônibus e três tipos de roubo: transeunte, estabelecimento comercial e interior de veículo. Em particular, aplicou-se o modelo de regressão Ordinary Least-Squares (OLS) (MONTGOMERY; PECK; VINING, 2006) neste trabalho, uma vez que esse modelo se adequou aos dados.

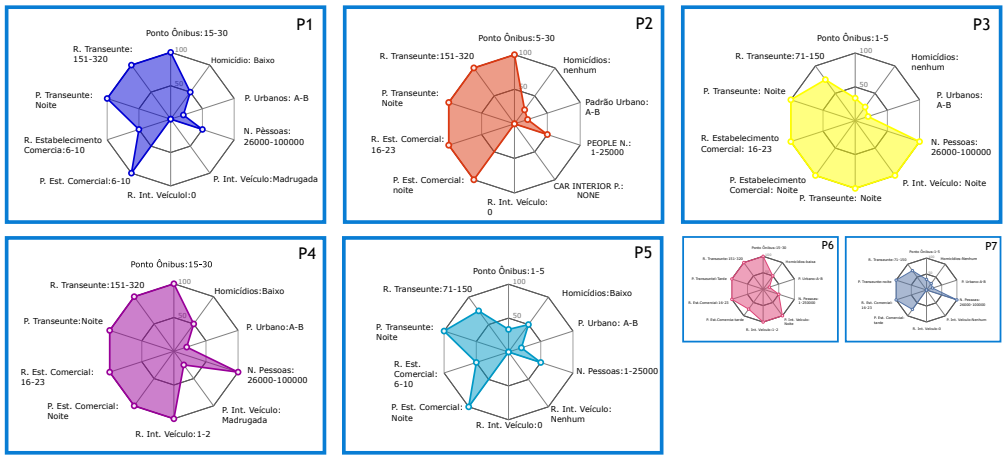
Figura 11 mostra um relacionamento entre os pontos de ônibus e os crimes para cada conjunto de dados em que os eixos horizontais representam os padrões e os eixos verticais representam o número de crimes por pontos de ônibus estimados por OLS. Roubo a transeunte tem um forte relacionamento com pontos de ônibus em todos os conjuntos de dados. Por exemplo, o padrão M2 mostra que o aumento de um ponto de ônibus aumenta o roubo transeunte em 14 vezes na vizinhança de escolas municipais. Por outro lado, pode-se observar que roubos de estabelecimento comercial e veículo não tem um relacionamento forte com pontos de ônibus visto que os valores de regressão estão abaixo de um na maioria dos padrões. Assim os resultados obtidos em **O1** validam a hipótese **H1**.

4.3.2.2 Crimes afetam o desempenho dos estudantes (**O2**)

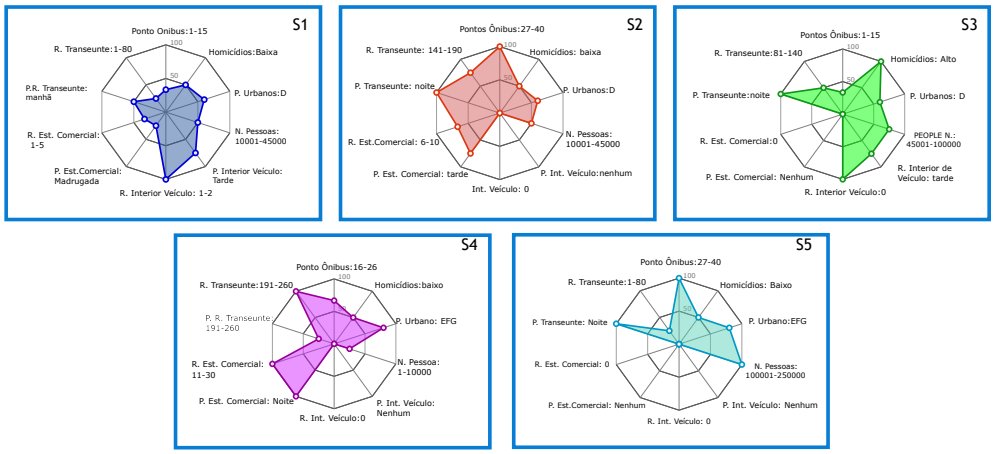
Especialistas do crime gostariam de investigar se crimes violentos (por exemplo, homicídios) têm afetado o desempenho dos estudantes em São Paulo e quais padrões representam este relacionamento.

Fonte: Adaptada de Silveira *et al.* (2022).

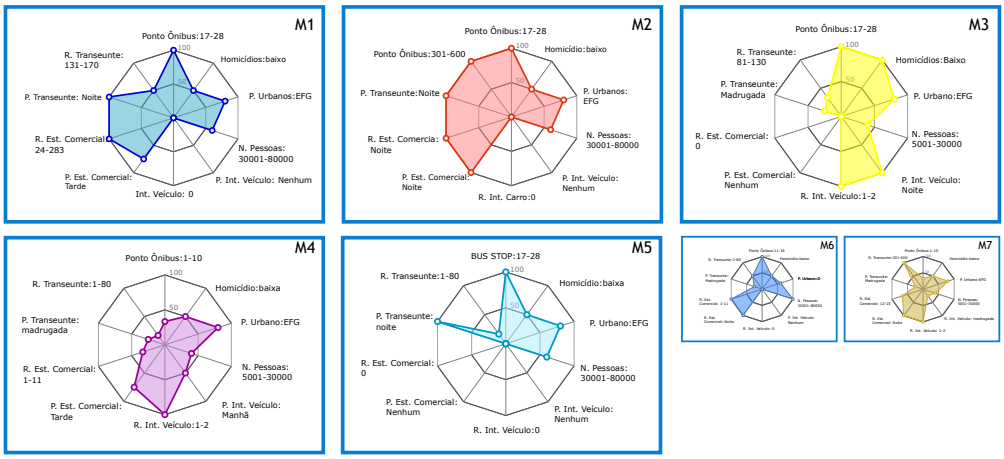
Figura 10 – Padrões fornecidos pelo TensorAnalyzer com rank 4.



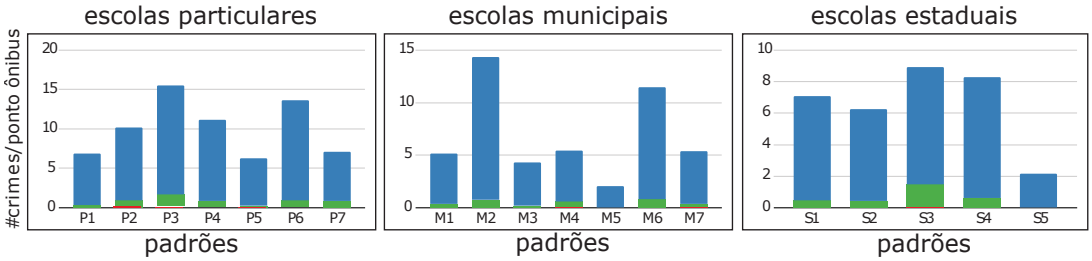
(a) Escolas particulares.



(b) Escolas estaduais.



(c) Escolas municipais.



Fonte: Adaptada de Silveira *et al.* (2022).

Figura 11 – Regressão OLS de pontos de ônibus e roubos: transeunte (■), estabelecimento comercial (■) e veículo (■).

A Tabela 5 mostra como o indicador de desempenho dos estudantes dado por ENEM no ensino médio para escolas particulares e estaduais. O melhor desempenho ocorre em escolas particulares, onde os padrões descritos em Figura 10a apresentam taxas de homicídios muito baixas ou não existentes. Além disso, os crimes destes padrões acontecem à noite, às vezes sem aula. Caso contrário, o padrão S3 das escolas estaduais na Figura 10b mostra altas taxas de homicídios entre outras e os piores desempenhos dos alunos no ENEM.

Fonte: Adaptada de Silveira *et al.* (2022).

Tabela 5 – Desempenho dos estudantes nas escolas do ensino médio dos padrões mostrados nas Figura 10a e Figura 10b. Melhores e piores resultados em negrito e itálico, respectivamente.

Padrão	P1	P2	P3	P4	P5	P6	P7	S1	S2	S3	S4	S5
ENEM	565	575	**601**	569	581	570	584	543	542	*515*	547	536

A Tabela 6 mostra os indicadores de desempenho por IDEEFAI e IDEEFAF para estudantes do ensino fundamental de escolas municipais. Semelhante ao experimento anterior, o padrão M3 na Figura 10c apresenta a maior taxa de homicídios dentre as outras e o pior desempenho dos estudantes em ambos indicadores.

Fonte: Adaptada de Silveira *et al.* (2022).

Tabela 6 – Desempenho dos alunos nas escolas de ensino fundamental a partir dos padrões mostrados na Fig. 10c. Melhor e pior resultado em negrito e itálico, respectivamente.

Padrão	M1	M2	M3	M4	M5	M6	M7
IDEEFAI	**5.65**	5.60	*5.36*	5.48	5.48	5.52	5.48
IDEEFAF	**4.45**	4.40	*4.17*	4.32	4.30	4.32	4.25

Em conclusão, os experimentos mostram um forte relacionamento entre as taxas de homicídios e o desempenho dos estudantes, em que o desempenho é inversamente proporcional às taxas de homicídios. O pior resultado ocorre nas vizinhanças pobres das escolas públicas com um alto número de homicídios. Deste modo, os resultados obtidos em **O2** validam a hipótese **H2**.

4.3.2.3 *Áreas de lazer têm sofrido vários crimes, principalmente durante o período da noite (**O3**)*

A presença de crimes nas ruas de São Paulo é um problema sempre presente, especialmente à noite e tarde da noite. Os visitantes não estão imunes a atos de crime e podem ser suscetíveis a serem alvos de certos crimes, já que tendem a ser mais propensos a exibir riqueza, tornando-os um alvo mais atraente. Parques e áreas recreativas frequentadas por visitantes e cidadãos sofreram crimes graves. Embora assaltos e furtos também sejam comuns durante o dia, os especialistas acreditam que taxas mais altas de crimes foram relatadas durante a noite. Portanto, o objetivo desta análise é compreender os padrões de criminalidade em áreas de lazer e qual é o período em que os crimes ocorrem com mais frequência.

Em escolas particulares, a maioria dos padrões encontra-se na classe urbana A-B, em que nota-se muitas praças, atrações turísticas e áreas recreativas (frequentadas por turistas e cidadãos). A Figura 10a mostra que o roubo a transeunte é intenso em todos os padrões, ocorrendo mais frequentemente no período da noite. Portanto, turistas devem ser mais cuidadosos nas áreas recreativas, principalmente no período da noite. Além disso, os formuladores de políticas públicas devem direcionar esforços para diminuir os roubos a transeunte nessas regiões. Portanto, os resultados obtidos em **O3** validam a hipótese **H3**.

4.3.3 Discussão

Neste capítulo, foi apresentado um estudo de caso cujo objetivo é atestar a eficiência da metodologia proposta no presente trabalho ao possibilitar que os especialistas validassem três hipóteses (**H1**, **H2** e **H3**), descritas no Capítulo 1. Para isso uma ferramenta nomeada TensorAnalyzer foi desenvolvida de modo que atendesse a três principais objetivos descritos na Subseção 4.3.1, os quais foram levantados pelos especialistas. Assim, ao atender a esses objetivos tornou-se possível validar as hipóteses.

Como pode-se observar, por meio da ferramenta desenvolvida foi possível elucidar a todos os objetivos levantados pelos especialistas. Por exemplo, para o objetivo **O1** a ferramenta mostrou que existe uma relação entre pontos de ônibus e a criminalidade no entorno das escolas da região metropolitana de São Paulo validando assim a hipótese **H1**. Já para o **O2** concluiu-se que existe uma relação entre as taxas de homicídios e o desempenho dos estudantes, em que o desempenho é inversamente proporcional às taxas de homicídios. Finalmente, para o **O3** observou-se que em áreas mais frequentadas por turistas o roubo a transeunte é mais intenso, principalmente no período da noite. Portanto, a metodologia desenvolvida foi capaz de elucidar os objetivos levantados, bem como validar as hipóteses apresentadas, atentando sua eficiência.

4.4 Estudo de caso: homicídios em São Paulo

O estudo de caso apresentado nesta Seção está direcionado a compreender os padrões globais no entorno dos homicídios, tendo em conta um conjunto de dados reais da cidade de São Paulo. Mais especificamente, o estudo de caso foi realizado para validar a metodologia proposta e elucidar as hipóteses apresentadas no Capítulo 1.

Para alcançar as especificações dos especialistas, a interface visual foi melhorada no sentido que incrementou-se mais gráficos que permitiram, por exemplo, análises temporais dos homicídios no padrões descobertos pela metodologia desenvolvida. Todos os pré-processamentos de dados e experimentos foram conduzidos em um 4-core 2.8 GHz Intel i7-4980HQ com 16 GB de RAM.

Antes de descrever o estudo de caso em si, uma breve descrição sobre o que os objetivos que os especialistas gostariam que fossem atendidos. Além disso, apresenta-se os dados que foram fornecidos pelos especialistas para as análises.

4.4.1 Objetivos e Dados

Um conjunto de reuniões e entrevistas foram realizadas com os especialistas para entender os principais objetivos que deveriam ser atendidos durante a análise dos dados de homicídios. O resultado destas interações pode ser resumido como segue:

O1 – Compreender as relações entre homicídios e outras variáveis envolvidas nas análises.

Compreender as relações entre as variáveis no entorno dos homicídios e os homicídios é essencial para auxiliar gestores de políticas públicas no planejamento urbano a fim de prevenir esse tipo de crime. Por exemplo, será que o aumento de arborização poderia impactar na redução de homicídios? Existem algumas evidências que sugerem que, nos bairros centrais da cidade, a vegetação pode trazer mais atenção para a rua, aumentando o uso pelos residentes dos espaços ao ar livre dos bairros. Alguns estudos conduzidos em bairros do centro das cidades mostraram que espaços ao ar livre arborizados são consistentemente mais bem usados por jovens, adultos e grupos de idade mista do que espaços sem árvores, sendo que, quanto mais árvores em um espaço, maior será o número de usuários simultâneos (KUO; SULLIVAN, 2001; KUO; BACAICOA; SULLIVAN, 1998). Assim, nesses ambientes, níveis mais elevados de vegetação não só preservam a visibilidade, mas também podem aumentar a vigilância. Muitos estudos têm mostrado que os perpetradores evitam áreas com maior vigilância e maior probabilidade de intervenção (BENNETT; WRIGHT, 1984; CROMWELL; OLSON; AVARY, 1991). Portanto, investigar a relação entre homicídios e arborização e outras variáveis, na cidade de São Paulo, torna-se um requisito imperativo para os especialistas. Com base nos estudos descritos anteriormente, especialistas gostariam de elucidar se existe uma relação entre crimes e as variáveis

envolvidas na análise e, consequentemente, validar/rejeitar a hipótese **H4** descrita no Capítulo 1.

O2 – Entender as dinâmicas espaço-temporais de homicídios que têm o mesmo padrão urbano.

A criminalidade em uma determinada região tende a desaparecer/mudar de lugar de acordo com as mudanças de infraestrutura urbana no seu entorno no decorrer do tempo, como pode-se observar em Garcia-Zanabria *et al.* (2020). Neste contexto, compreender as dinâmicas espaço-temporais dos homicídios nos padrões identificados na cidade de São Paulo é crucial, pois compreender as razões pelas quais o crime reduziu em uma determinada região pode ajudar aos gestores de políticas públicas a tomarem decisões mais assertivas em regiões que possuem um padrão urbano parecido, reduzindo também a criminalidade nestas regiões. Assim, o presente objetivo tem como objetivo validar/rejeitar a hipótese **H5** descrita no Capítulo 1.

4.4.1.1 Dados

Neste estudo de caso foram utilizados dados de habitação, recursos naturais, cultura, educação, saúde e segurança dos últimos anos, disponibilizados pelo GeoSampa (GEOSAMPA, 2000) e registros de homicídios os quais foram fornecidos pelo departamento da polícia de São Paulo. A Tabela 7 consiste em um resumo dos dados utilizados neste trabalho em que constitui-se de 4 colunas: dados, que refere-se aos tipos de bases de dados; categorias, que refere-se a divisão dos dados; ano, que corresponde ao período em que os dados foram coletados; e fonte, que é o órgão fornecedor dos dados. A seguir descreve-se de maneira sucinta os dados empregados neste trabalho.

Habitação: os dados estão divididos em três categorias: cortiço, favelas e loteamentos e são referentes aos anos de 2006, 2016 e 2016, respectivamente. Os cortiços são considerados assentamentos precários que se caracterizam como habitações coletivas precárias de aluguel. Neste sentido, a alta densidade domiciliar, a grande circulação de residentes e a convivência de pessoas de culturas diferentes em espaços pequenos tende a gerar conflitos. As favelas se caracterizam por assentamentos precários que surgem de ocupações espontâneas feitas de forma desordenada, sem definição prévia de lotes e sem arruamento. Já os loteamentos se caracterizam por assentamentos em que a ocupação se deu a partir da iniciativa de um agente promotor sem prévia aprovação dos órgãos públicos. A estrutura viária e habitacional, dificulta tanto o policiamento e a identificação dos crimes como facilita o esconderijo e a fuga de ofensores. Há na literatura evidências da diferenciação na prática e registro de crimes como roubos e furtos dentre outros, em favelas e fora delas.

Recursos Naturais: possui apenas uma categoria nomeada arborização a qual é referente ao ano de 2014. Esta categoria consiste de indivíduos arbóreos localizados no sistema viário

Fonte: Elaborada pelo autor.

Tabela 7 – Resumo dos dados usados durante as análises. A Tabela é composta de quatro colunas: tipo de dado, categorias em que os dados foram divididos, o período correspondente aos dados, as categorias em que os dados foram divididos e a fonte dos dados.

Dados	Categorias	Ano	Fonte
Habitação	• Cortiço • Favelas • Loteamentos	2006 2006 2016	(SEHAB, 2021)
R. Naturais	Arborização	2014	(SMPR, 2017)
Cultura	• Teatro • Cinema • Show • Museu • Artes • Esporte • CEU	2000	(SMDU, 2017)
Educação	• Esc. Públicas • Esc. Particulares	2014	(SME, 2017)
Saúde	• Hospitais • Pronto-Socorro • Emergência	2018	(SMS, 2017)
Segurança	• Casa de Mediação • Batalhões • Delegacias • Bombeiros	2015	(SMSU, 2017)
Crime	Homicídios	2006-2017	(NEV, 2000)
IPVS	Grupo 1-6	2021	(SEADE, 2021)
Padrão Urbano	Grupo A-H	2021	(SEADE, 2021)

do município de São Paulo. Compreendem árvores localizadas em calçadas, canteiros centrais, rotatórias e calçadas de praças. Estão excluídas destes dados árvores localizadas nas áreas internas de parques, parques lineares, reservas e praças bem como em toda e qualquer área interna de lote particular e público. A presença de árvores, quando há boa gestão e serviços públicos, propicia melhor qualidade de vida aos residentes e transeuntes locais.

Cultura, Educação e Saúde: estas três bases de dados são referentes ao ano de 2018. Os dados de Cultura estão divididos em nove categorias: salas de teatro, salas de cinema, show e concertos, bibliotecas, museu, centro culturais, galerias de artes, esportes e Centros Educacionais Unificados (CEU). Já os dados de Educação englobam escolas particulares e públicas. Por fim, os dados de Saúde compreendem hospitais, pronto-socorro e emergência. As manifestações sociais, como a violência, podem ser afetadas pelos serviços públicos que são oferecidos para os indivíduos de um dado local.

Segurança: compreende sete categorias: casas de mediação, inspetorias regionais, unidades de comando gerais, batalhões, companhias, delegacias e corpo de bombeiros as quais são referentes ao ano de 2015. A violência interpessoal tende a estar associada a qualidade, quantidade, diversidade e acessibilidade de instituições presentes em uma comunidade/vizinhança, capazes de suprir, ou não, as necessidades sobretudo dos residentes, tais como facilidades médicas, garantia de ir e vir com segurança, entre outras.

Crime: possui apenas a categoria homicídios a qual abrange o período de 2006 até 2017. Os dados de homicídios contêm informações como data e hora da ocorrência, código da delegacia e a geolocalização.

IPVS: O Índice Paulista de Vulnerabilidade Social (IPVS) expressa a vulnerabilidade da população à pobreza. A tipologia dessas áreas se baseia nas informações do Censo Demográfico e leva em conta variáveis como a renda domiciliar per capita, o percentual de mulheres de 10 a 29 anos responsáveis pelos domicílios e a situação de aglomerado subnormal (favela). Com base nessas variáveis, são definidos sete grupos em que são classificados os setores, levando em conta as diferentes condições de exposição da população residente à vulnerabilidade social. O IPVS é constituído por 6 grupos são eles:

Grupo 1 – Nenhuma Vulnerabilidade: engloba os setores censitários em melhor situação socioeconômica (muito alta), com os responsáveis pelo domicílio possuindo os mais elevados níveis de renda e escolaridade. Seus responsáveis tendem a ser mais velhos, com menor presença de crianças pequenas e de moradores nos domicílios, quando comparados com o conjunto do Estado de São Paulo.

Grupo 2 – Vulnerabilidade Muito Baixa: abrange os setores censitários que se classificam em segundo lugar, no Estado, em termos da dimensão socioeconômica (média ou alta). Nessas áreas concentram-se, em média, as famílias mais velhas.

Grupo 3 – Vulnerabilidade Baixa: formado pelos setores censitários que se classificam nos níveis altos ou médios da dimensão socioeconômica e seu perfil demográfico caracteriza-se pela predominância de famílias jovens e adultas.

Grupo 4 – Vulnerabilidade Média: composto pelos setores que apresentam níveis médios na dimensão socioeconômica, estando em quarto lugar na escala em termos de renda e escolaridade do responsável pelo domicílio. Nesses setores concentram-se famílias jovens, isto é, com forte presença de chefes jovens (com menos de 30 anos) e de crianças pequenas.

Grupo 5 – Vulnerabilidade Alta: engloba os setores censitários que possuem as piores condições na dimensão socioeconômica (baixa), estando entre os dois grupos em que os chefes de domicílios apresentam, em média, os níveis mais baixos de renda e escolaridade. Concentra famílias mais velhas, com menor presença de crianças pequenas.

Grupo 6 – Vulnerabilidade Muito Alta: o segundo dos dois piores grupos em termos da dimensão socioeconômica (baixa), com grande concentração de famílias jovens. A combinação entre chefes jovens, com baixos níveis de renda e de escolaridade e presença significativa de crianças pequenas permite inferir ser este o grupo de maior vulnerabilidade à pobreza.

Padrão Urbano: Os padrões urbanos identificam e analisam as características dos setores reunindo-os em agrupamentos mutuamente exclusivos. A tipologia dessas áreas se baseia nas informações do Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT), Empresa Paulista de Planejamento Metropolitano (Emplasa), Companhia do Metropolitano de São Paulo (Metrô), Secretaria da Segurança Pública do Estado de São Paulo e Censos Demográficos de 1980 a 2010. Com base em 19 variáveis (sensíveis a condições e alterações populacionais, ambientais, criminais, habitacionais, de mobilidade e de expansão urbana), são definidos oito grupos em que são classificados os setores, levando em conta as singularidades da cidade. Cada grupo é descrito sucintamente a seguir:

Grupo A – Comercial e de serviços: engloba os setores censitários que se destacam pelos bons índices de condições sanitárias e higiene, por produzir o maior número de viagens e o maior tempo médio de duração do deslocamento da casa ao trabalho, pela alta proporção de chefes de família alfabetizados, de alta renda e mulheres e pela alta presença de setores sem registros de homicídios dolosos.

Grupo B – Residencial de urbanização consolidada: abrange os setores censitários que foram urbanizados majoritariamente entre os anos 1930 e 1949 e 1950 e 1962, tornando-se totalmente urbanizados no período 1975-1985. Esses setores são distintos pela proporção de domicílios improvisados, pela densidade demográfica, que, além de baixa, apresenta pequena, e pela maior proporção de população masculina jovem.

Grupo C – Urbanização radial: formado pelos setores censitários que se classificam pelo reduzido número de aglomerados subnormais e boas condições sanitárias e de higiene, esse grupo destaca-se pela elevada proporção de chefes de família alfabetizados e de alta renda, diferenciando-se dos outros pela grande variação de crescimento populacional e densidade demográfica.

Grupo D – Residencial disperso: seus setores foram urbanizados mais fortemente entre 1950 e 1962. Esses setores exibem uma elevada variação de verticalização. Além disso, são frequentes as baixas taxas de homicídios.

Grupo E – Habitação irregular em áreas de urbanização antiga: engloba os setores censitários que se destacam pela proporção de aglomerados subnormais e pelo crescimento deles entre 1991 e 2010. Também, evidenciam-se pelo baixo número de domicílios improvisados e pelo menor índice de chefes de família de alta renda.

Grupo F – Residencial concentrado de urbanização gradativa em áreas de proteção aos mananciais e de risco geológico: marcado pela variação relativamente baixa

de crescimento populacional. Sua urbanização ocorreu paulatinamente entre 1950 e 1985, prosseguindo ao menos até 2010. Ele distingue dos demais pela baixa variação de domicílios particulares permanentes, por setores com padrão alto de homicídios e pela alta variação de população masculina jovem.

Grupo G – Habitação irregular concentrada em áreas de urbanização atual: diferencia-se dos demais essencialmente pela variação de verticalização, extremamente elevada. Com urbanização iniciada no período 1963-1974 ainda preserva parte do seu território como rural e reúne o menor número de setores, 1.198 no total. Destaca-se também pela menor proporção de população masculina jovem e pela evolução da densidade demográfica e dos aglomerados subnormais.

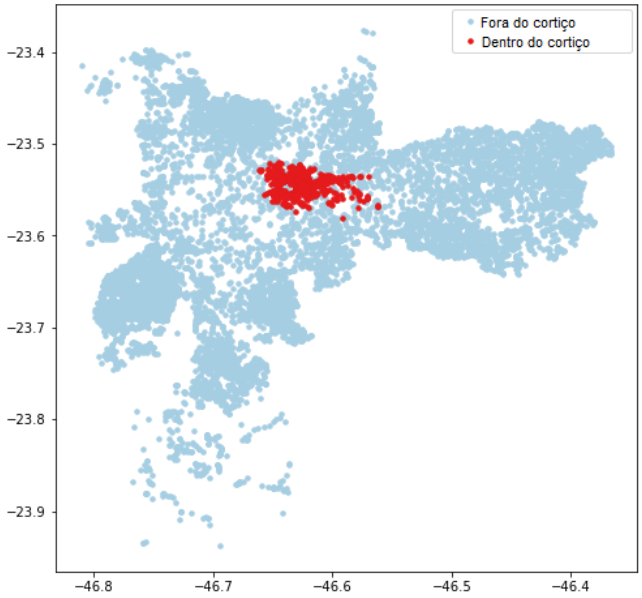
Grupo H – Habitação em áreas periurbanas: caracteriza-se por apresentar o maior conjunto de setores rurais. Exibe o segundo maior no número de setores com áreas de risco geológico e os piores índices de atendimento de água, esgoto e coleta de lixo. Além disso, sua condição está associada a fatores como baixa proporção de chefes de domicílio com rendimento superior a 20 salários mínimos e baixa densidade demográfica. É importante destacar, ainda, que esse é o grupo que mais reúne setores com padrão alto de homicídios.

Neste estudo de caso, os dados de recursos naturais, cultura, educação, saúde, segurança e homicídios estão mensurados em unidades pontuais, já os dados de habitação em unidades de área e por fim padrões urbanos e IPVS em setor censitário.

4.4.2 Índices

Para uma melhor compreensão da relação entre homicídios e as bases de dados descritas anteriormente, alguns índices foram gerados. Os índices foram definidos basicamente de três maneiras: para dados pontuais como (recursos naturais, cultura, educação e homicídios) utilizou-se um estimador de densidade conhecido como Kernel para calcular a intensidade pontual de um evento, situado em uma região de interesse. Já para dados como segurança e saúde, calculou-se a distância entre os pontos da base de homicídios e os pontos das bases de saúde/segurança. Para isso, modelou-se as ruas de São Paulo por meio de um grafo em que foi possível calcular as distâncias mais precisamente do que simplesmente calcular uma distância em linha reta. A seguir descreve-se sucintamente como os índices foram gerados para cada base de dados:

Habitação: Para esta base de dados gerou-se dois mapas: um para Cortiço (Figura 12) e outro para Favelas e Loteamentos (Figura 13) (já que possuem características similares). Os elementos que constituem essa base de dados são representados por polígonos. Neste sentido, para cada polígono traça-se um raio de 200 metros a partir das bordas do polígono. Após isso, verifica-se quais pontos da base de homicídios caem dentro dos raios que foram traçados.



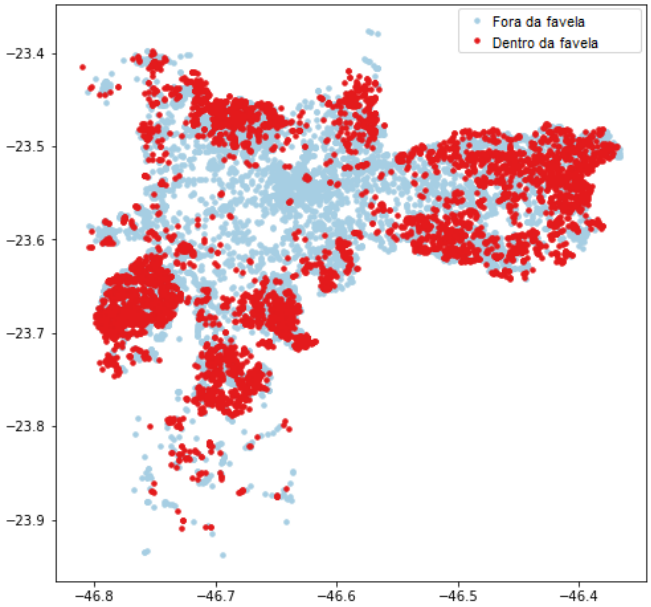
Fonte: Elaborada pelo autor.

Figura 12 – Em vermelho, homicídios que estão na região dos cortiços e em azul homicídios que estão fora.

Aqueles pontos que caem dentro do raio, são coloridos de vermelho, já aqueles que caem fora são coloridos de azul. Portanto, para essa base de dados, os índices gerados são binários.

Recursos Naturais, Cultura, Educação, Homicídios: Mapas temáticos foram gerados para cada uma destas bases de dados utilizando um estimador de densidade conhecido como Kernel. Por meio desse estimador calcula-se a intensidade pontual de um evento situado em uma região de interesse. Neste sentido, para cada base de dados escolheu-se um raio mais apropriado para os dados, visto que um raio muito amplo pode gerar uma superfície excessivamente suavizada e quando excessivamente pequeno pode gerar uma superfície muito fracionada. Neste sentido, mapas temáticos foram gerados para cada uma das bases Recursos Naturais (Figura 14), Cultura (Figura 15), Educação (Figura 16), homicídios (Figura 17), em que os dados foram agrupados em 5 diferentes índices: muito baixa, baixa, média, alta e muito alta.

Saúde e Segurança: Para estes dois equipamentos compreender a que distância os homicídios ocorreram é muito relevante, visto que a violência interpessoal tende a estar associada à qualidade, quantidade, diversidade e acessibilidade de instituições presentes em uma comunidade/vizinhança capazes de suprir ou não as necessidades sobretudo dos residentes tais como facilidades médicas, garantia de ir e vir com segurança, dentre outras. Para isso modelamos a malha viária de São Paulo por meio de um grafo, que permite-nos calcular a distância entre dois pontos de maneira mais próxima da distância real, considerando as ruas e esquinas (ver Figura 18). Neste contexto, as distâncias entre os pontos de homicídios e os pontos de equipamentos de saúde e segurança são calculadas em que a menor distância é atribuída para cada homicídio da base de dados. Assim, foi gerado um mapa temático para a saúde (Figura 19a) e



Fonte: Elaborada pelo autor.

Figura 13 – Em vermelho, homicídios que estão na região das favelas e em azul homicídios que estão fora.

outro para segurança (Figura 19b).

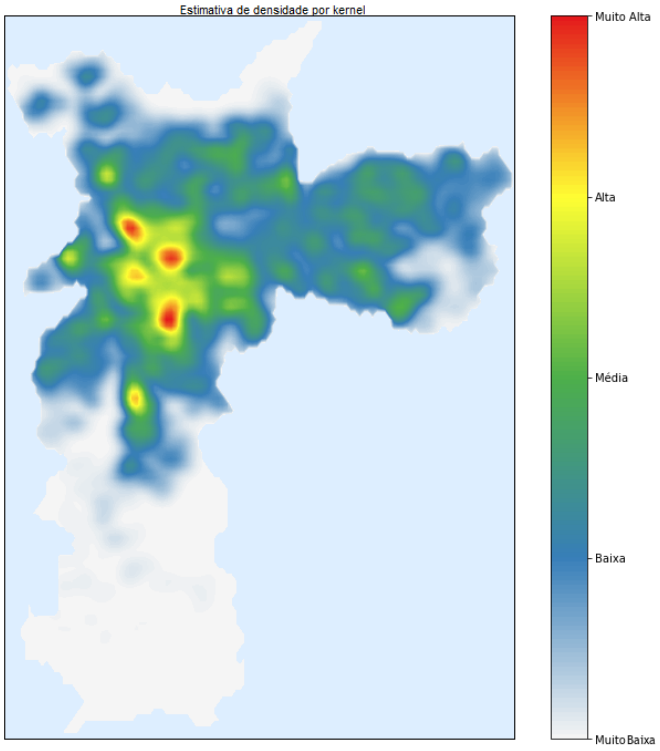
IPVS e Padrão Urbano: Em ambas as bases de dados observa-se uma visão detalhada das condições de vida no município com a identificação e localização espacial dos setores censitários. Neste sentido, para cada setor das bases, verifica-se a localização de cada homicídio o qual é colorido de acordo com classificações de IPVS ou Padrão Urbano. Na Figura 20 pode-se observar a classificação dos homicídios segundo IPVS (Figura 20a) e Padrão Urbano (Figura 20b).

4.4.3 Resultado das Análises

Nesta Seção descreve-se os resultados obtidos das análises realizadas a fim de atender os requisitos **O1** e **O2** especificados pelos especialistas. Para isso, na Figura 21 observa-se os padrões dos homicídios fornecidos pelo TensorAnalyzer, em que as análises a seguir são baseadas.

4.4.3.1 *Compreender as relações entre homicídios e a infraestrutura distribuída na cidade (**O1**)*

Estudos recentes têm mostrado que o aumento/redução das taxas de criminalidade estão associados com a infraestrutura presente nas cidades (LAFORTEZZA *et al.*, 2009; SCOPELLITI *et al.*, 2016). Neste sentido, uma análise das relações entre homicídios e alguns equipamentos de infraestrutura da cidade de São Paulo torna-se imprescindível, pois permite guiar os gestores

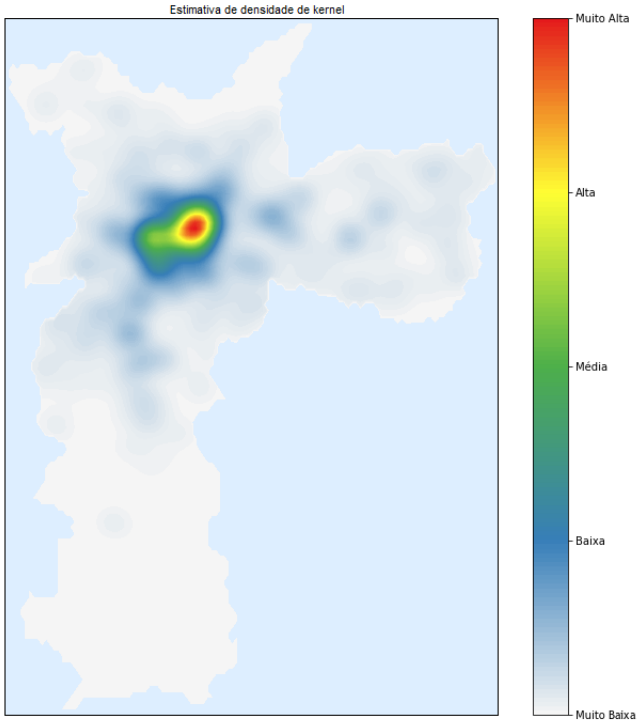


Fonte: Elaborada pelo autor.

Figura 14 – Mapa de densidade gerado para indivíduos arbóreos localizados no sistema viário do município de São Paulo considerando um raio de 800 metros.

de políticas públicas em suas ações como uma tentativa de reduzir a taxa de crimes. A seguir conduz-se uma análise da relação entre as variáveis presentes no estudo de caso e os homicídios em cada um dos padrões encontrados pela abordagem proposta. É importante salientar que os homicídios apresentam características e relações particulares dependendo da região que está sendo estudada. Assim, um grupo de homicídios que está dentro de um padrão urbano A por exemplo, pode ter uma relação com a variável arborização e já outro grupo de homicídios que está no padrão urbano H pode não ter nenhuma relação com arborização e, justamente por isso, essa análise foi feita por padrão.

Primeiramente para encontrar a relação entre os homicídios e as características utilizou-se técnicas baseadas em um modelo de Regressão Logística Ordinal (RLO), isso porque a variável dependente do presente estudo que são os homicídios é uma variável categórica ordinal com mais de duas classes, e, portanto, a RLO é o modelo mais indicado para este caso. Mais especificamente, empregou-se o modelo *LogistAT* presente na biblioteca *mord* do *Python* para encontrar a relação entre homicídios e as variáveis dos padrões. O modelo fornece alguns coeficientes em logarítmo que ao exponenciá-los obtemos a taxa de probabilidade (muito conhecida como *odds ratio* em inglês). A *odds ratio* pode ser interpretada da seguinte maneira: a variável dependente do presente modelo é homicídios e as variáveis independentes são as características restantes presentes nos padrões. Assim, uma *odds ratio* > 1 sugere uma probabilidade crescente de estar

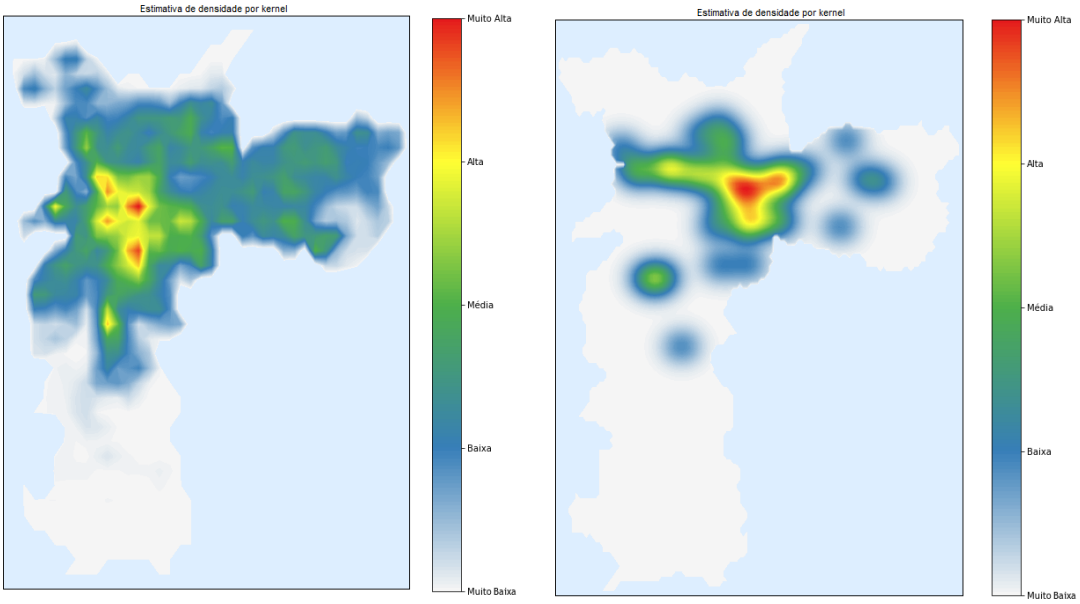


Fonte: Elaborada pelo autor.

Figura 15 – Mapa de densidade gerado para localizações públicas classificadas como cultura considerando um raio de 800 metros.

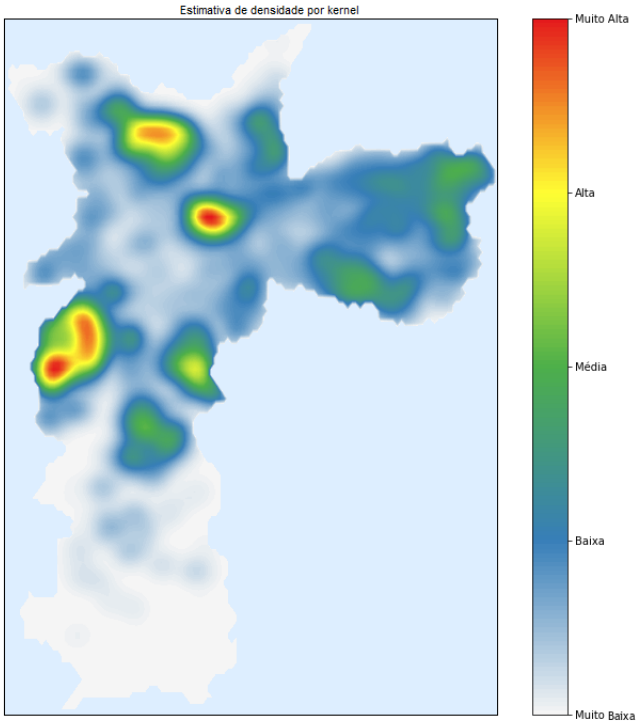
Fonte: Elaborada pelo autor.

Figura 16 – Mapa de densidade gerado para a base de dados de escolas particulares (a) e públicas (b) considerando um raio de 900 e 800 metros respectivamente.



(a) Escolas particulares

(b) Escolas estaduais



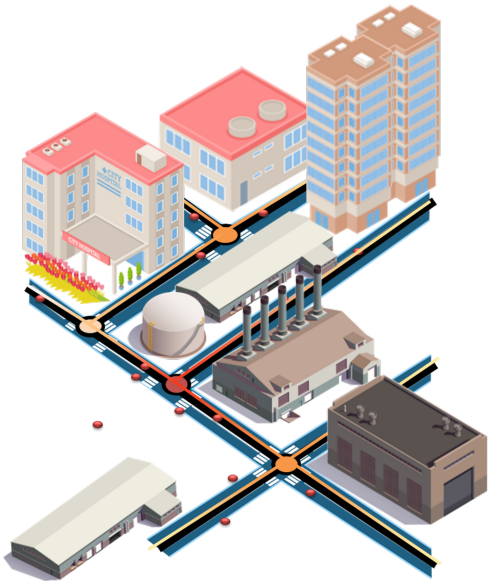
Fonte: Elaborada pelo autor.

Figura 17 – Mapa de densidade gerado para dados de homicídios considerando um raio de 1,5 km.

em um nível superior da variável dependente conforme os valores da variável independente aumentam. Já quando a *odds ratio* < 1 sugere uma probabilidade descrescente de estar em um nível superior da variável dependente conforme os valores da variável independente aumentam. Quando a *odds ratio* $= 1$ sugere que não existem efeitos. O resultado do modelo aplicado aos padrões econtrados pelo TensorAnalyzer pode ser observado na Tabela 8, em que considera-se apenas os resultados que estão dentro de um intervalo de confiança maior ou igual a 95%:

Com base na Tabela 8 parece haver uma relação entre homicídios e outras características do modelo em alguns padrões. A seguir discute-se essas relações com base nos valores das *odds rate* encontrados.

Homicídios versus Arborização: Alguns trabalhos têm mostrado que a vegetação densa pode ser usada ativamente por criminosos para esconder ou identificar vítimas de crime (TROY; GROVE; O'NEIL-DUNNE, 2012; SANDER; ZHAO, 2015). Por outro lado, outros estudos mostram que árvores em áreas públicas têm sido associadas a menores taxas de criminalidades (DONOVAN; PRESTEMON, 2012; CARRIAZO, 2016). Essa contradição levou os especialistas a questionarem se, em algum dos padrões encontrados pelo TensorAnalyzer, a arborização está associada a uma redução na taxa de homicídios. De acordo com os resultados mostrados na Tabela 8, os valores da *odds ratio* foram significativos e menores do que um para os padrões $P2, P3, P4, P5, P6, P8$ e $P9$. Isso sugere que quando aumenta-se a arborização em uma unidade existe uma probabilidade descrescente de homicídios estar

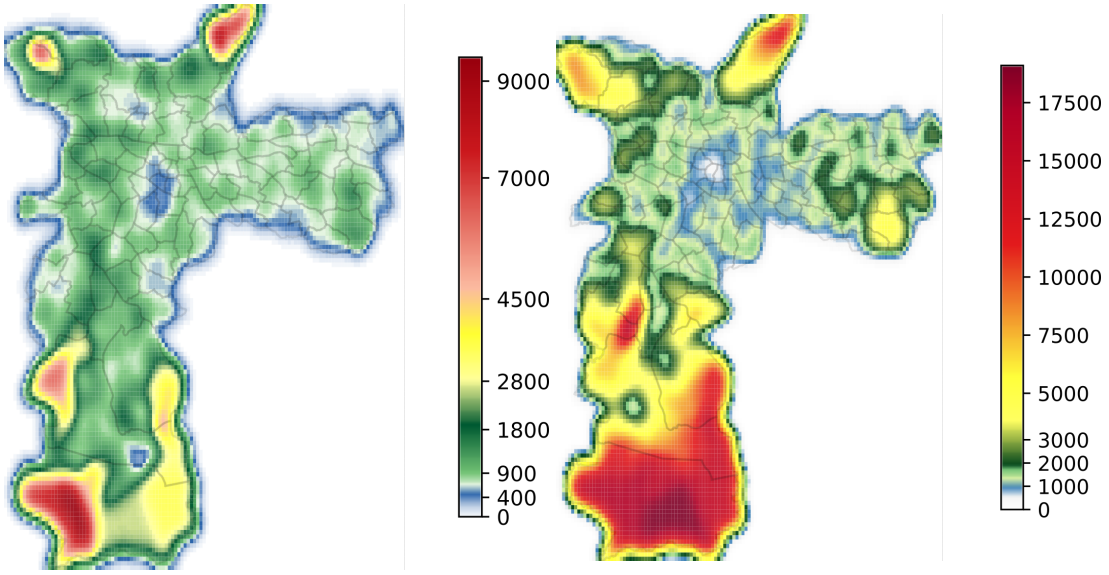


Fonte: Elaborada pelo autor.

Figura 18 – Malha viária representada por um grafo, em que a rede colorida é o grafo e os pontos em vermelho são os homicídios.

Fonte: Elaborada pelo autor.

Figura 19 – Mapa temático gerado para saúde em (a) e segurança em (b), em que as regiões onde os homicídios ocorreram mais distantes dos equipamentos são coloridas com cores quentes. A unidade de medida das distâncias em ambos os mapas é metros.



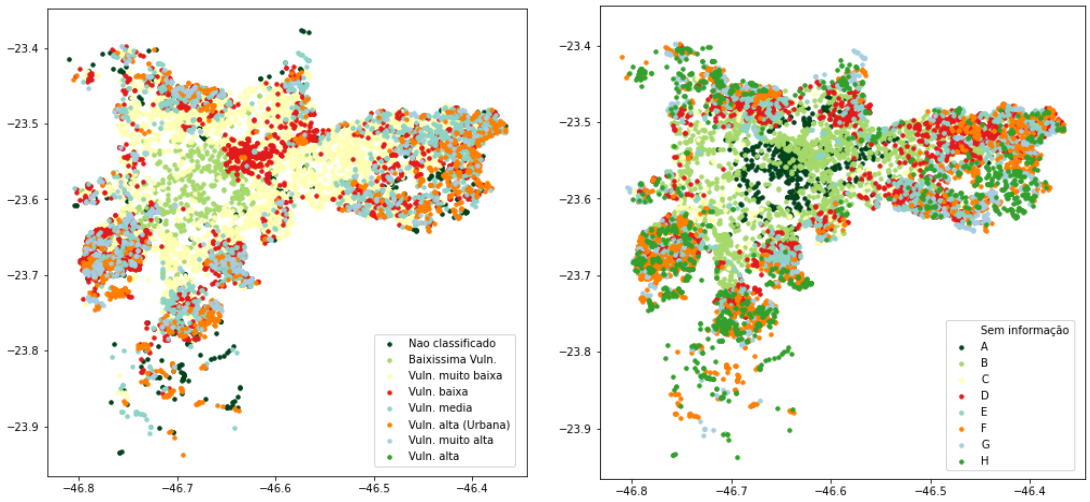
(a) Saúde (b) Segurança

em um nível superior. Mais especificamente, quando há um aumento de arborização em uma determinada região com determinadas características os resultados sugerem haver uma queda na taxa de homicídios naquela região.

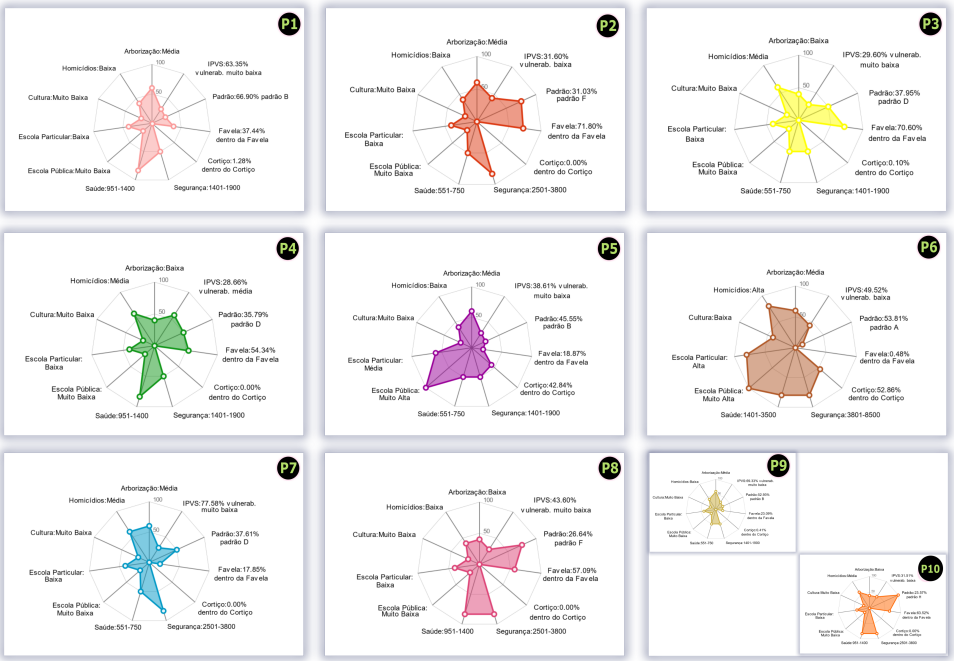
Homicídios versus Cultura: A relação entre homicídios e cultura é bastante específica de al-

Fonte: Elaborada pelo autor.

Figura 20 – Classificação dos homicídios segundo características do IPVS em (a) e Padrão Urbano em (b).



(a) IPVS (b) Padrão Urbano



Fonte: Elaborada pelo autor.

Figura 21 – Padrões fornecidos pelo TensorAnalyzer com rank 4.

guns padrões (*P*2 e *P*8), sendo que a *odds ratio* assume valores maiores do que um nesses padrões, sugerindo que o aumento do número de equipamentos de cultura em uma determinada região com determinadas características aumenta a probabilidade de homicídios estar em um nível mais superior. Mais especificamente, a presença de equipamentos de cultura sugere um aumento na taxa de homicídios em determinadas regiões de São Paulo.

Fonte: Elaborada pelo autor.

Tabela 8 – Resultado da *odds ratio* ao aplicar a regressão logística ordinal sobre as observações presentes em cada padrão um dos padrões, em que considera-se apenas os resultados com um intervalo de confiança maior ou igual a 95%.

Características	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Arborização	-	0,05	0,54	0,34	0,46	0,56	-	0,00	0,47	-
Cultura	-	4,77	-	-	-	-	-	10,47	-	-
Escola Privada.	6,62	0,19	1,97	3,48	-	1,83	-	0,03	0,72	2,04
Escola Pública	0,13	0,61	0,39	0,34	-	0,55	-	0,22	-	0,21
Saúde	-	1,0	-	1,0	1,0	-	1,0	-	-	1,0
Segurança	1,0	1,0	-	-	1,0	1,0	1,0	-	-	1,0
Cortiços	-	-	-	-	-	-	-	-	-	-
Favelas	-	-	1,52	-	1,59	3,10	-	-	1,82	-
Padrões Urbanos	-	1,58	-	-	-	-	-	-	0,89	-
IPVS	-	1,43	1,18	1,48	-	-	-	2,78	-	-

Homicídios versus Esc. Particular: Na relação entre homicídios e escolas particulares a *odds ratio* assume valores maiores que 1 ($P1$, $P3$, $P4$, $P6$ e $P10$) e menores que 1 ($P2$, $P8$ e $P9$). Desta forma, o aumento de escolas particulares podem ter uma probabilidade crescente de homicídios estar em um nível superior, como também uma probabilidade decrescente de homicídios estar em um nível superior. Assim, como nos casos anteriores a relação entre homicídios e escolas particulares depende das características do padrão que se está analizando.

Homicídios versus Esc. Pública: De acordo com a Tabela 8 a *odds ratio* assume valores menores do que um em todos os padrões que estão dentro de um intervalo de confiança maior do que 95%. Isso sugere que o aumento em uma unidade de escolas públicas implica em uma probabilidade decrescente de homicídios estar em um nível superior. Mais especificamente, o aumento de uma unidade de escolas públicas parece ter relação com a redução da taxa de homicídios em determinados contextos em São Paulo.

Homicídios versus Favelas: Nesta relação para todos os padrões que apresentaram um resultado dentro de um intervalo de confiança maior do que 95%, os valores da *odds ratio* são menores do que um, sugerindo que o aumento de favelas implica em uma probabilidade crescente de homicídios estar em um nível superior. Assim as favelas parecem contribuir para o aumento da taxa de homicídios em regiões específicas de São Paulo tendo em conta determinadas características ($P3$,$P5$,$P6$ e $P9$).

Homicídios versus IPVS: Nesta relação, para todos os padrões que estão dentro de um intervalo de confiança superior a 95% a *odds ratio* é maior do que um. Assim, quando o nível de vulnerabilidade aumenta há também uma probabilidade crescente de homicídios estar em um nível superior.

De acordo com os resultados da Tabela 8 equipamentos de saúde e segurança parecem não ter efeito sobre homicídios. Já os resultados em relação a padrões urbanos não são signfificativos na maioria dos padrões.

4.4.3.2 *Entender as dinâmicas espaço-temporais de homicídios que têm o mesmo padrão urbano. (**O2**)*

Como descrito anteriormente, os crimes, em geral, variam no espaço ao longo do tempo e entender as razões que levaram a redução de um tipo de crime em uma determinada região, pode auxiliar os gestores de políticas públicas a tomarem decisões que sejam assertivas no sentido de prevenir a criminalidade. Neste sentido, uma análise temporal foi feita para dois padrões pertecentes ao grupos de padrões retornados pela ferramenta ($P1$ e $P2$). Para ambos padrões a análise consistiu em: selecionar um padrão na área de *Visualização de Padrões* Figura 8 (C), selecionar um intervalo de tempo na área de *Visualização Temporal* Figura 8 (E) e deslizar a janela ao longo do tempo no mesmo instante em que se observa o comportamento do crime na área de *Visualização de homicídios no Mapa* Figura 8 (B). Ao detectar uma redução de crimes em uma determinada região, clica-se em um ponto que pertence aquela região, em seguida uma janela do *Google Street View* é aberta mostrando a região para que o usuário faça uma exploração do local.

Na Figura 22 observa-se em (a) e em (b) uma análise temporal realizada para o padrão $P1$, em que a região de estudo consiste da Avenida das Nações Unidas e Avenida do Rio Bonito. Em (a) o intervalo de tempo de 2006 até 2011 foi selecionado na ferramenta de *Visualização Temporal*, em que percebe-se uma presença considerável de homicídios ao longo de ambas as avenidas e proximidades. As imagens do *Google Street View* coletadas nesse período de tempo revelam uma infraestrutura precária ao longo de ambas as avenidas e também das proximidades, onde existe uma forte presença de lotes, praças e prédios abandonados. Esses lugares são muito atrativos para criminosos cometerem os homicídios, visto que são lugares não muito visados pela polícia. Já em (b) o intervalo de tempo selecionado na ferramenta *Visualização Temporal* foi de 2012 até 2017, nesse período observa-se uma redução dos homicídios nas proximidades de ambas avenidas e uma melhoria significativa na infraestrutura da região como pode-se observar nas imagens coletadas do *Google Street View* que correspondem aos mesmos lugares de (a).

De maneira similar ao padrão $P1$ foi realizada uma análise para o padrão $P2$, e a região de estudo escolhida foi o bairro Jardim São Luis. Na Figura 23 (a) o período selecionado na ferramenta *Visualização Temporal* para a análise varia de 2006 até 2011, em que é possível visualizar uma quantidade de crimes significativa na região de estudo de acordo com o mapa de homicídios. Além disso, as imagens coletadas do *Google Street View* dentro desse mesmo período de tempo mostram uma infraestrutura muito precária. Entretanto, ao selecionarmos outro intervalo de tempo na ferramenta (2012-2017) percebe-se uma redução significativa na criminalidade nesta região e também uma melhoria expressiva da infra-estrututura em diversas

Fonte: Elaborada pelo autor.

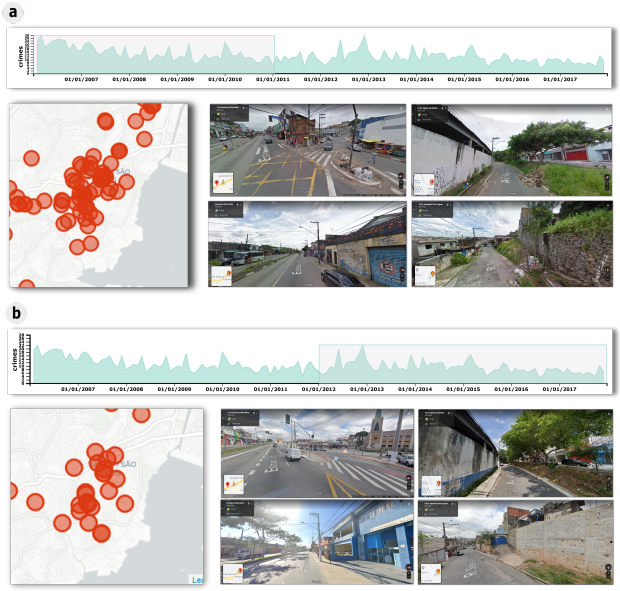
Figura 22 – Análise temporal do padrão $P1$ (ver Figura 21), em que escolheu-se a Avenida das Nações Unidas e a Avenida do Rio Bonito como a região de estudo do padrão. Em (a) e (b) observa-se uma *Visualização Temporal* com um intervalo de tempo selecionado, um mapa das localizações espaciais dos homicídios e algumas imagens do *Google Street View* referentes à região de estudo. É importante salientar que o resultado tanto dos homicídios presentes no mapa quanto das imagens do *Google Street View* estão dentro do intervalo de tempo selecionado na *Visualização Temporal*.

ruas e avenidas da região de estudo.

4.5 Discussão e limitações

Desta forma, os resultados apresentados tanto em Subsubseção 4.4.3.1 quanto em Subsubseção 4.4.3.2 mostraram que existe uma relação entre homicídios e a infraestrutura do entorno em determinados contextos, validando assim a hipótese **H4**. Em Subsubseção 4.4.3.1 foi possível encontrar relações entre homicídios e algumas variáveis dos padrões (arborização, cultura, escolas, favelas e IPVS). Assim, de acordo com os resultados, a presença de arborização parece reduzir a quantidade de homicídios, de um modo geral, visto que essa relação acontece na maioria dos padrões retornados pelo TensorAnalyzer. Outro resultado interessante foi que a presença de escolas públicas também parece contribuir para a redução de homicídios. É importante salientar que as relações identificadas estão dentro de um contexto o qual contém determinadas variáveis.

Em Subsubseção 4.4.3.2, os resultados apresentados na Figura 22 e Figura 23 sugerem uma relação entre a melhoria da infraestrutura e a redução de homicídios em algumas regiões da cidade de São Paulo, validadendo assim a hipótese **H5**. Os resultados mostraram que lugares que estavam mais sujos ou abandonados pelas autoridades públicas em um determinado período de



Fonte: Elaborada pelo autor.

Figura 23 – Análise temporal do padrão *P*2 (ver Figura 23), em que escolheu-se o Jardim São Luis como a região de estudo do padrão. Em (a) e (b) observa-se uma *Visualização Temporal* com um intervalo de tempo selecionado, um mapa das localizações espaciais dos homicídios e algumas imagens do *Google Street View* referentes à região de estudo. É importante salientar que o resultado tanto dos homicídios presentes no mapa quanto das imagens do *Google Street View* estão dentro do intervalo de tempo selecionado na *Visualização Temporal*.

tempo tinham um alto índice de criminalidade. Com o passar do tempo, esses mesmos lugares foram recebendo investimentos em infraestrutura (como limpeza, iluminação pública, construção de praças, etc) e também mais atenção de autoridades públicas. Juntamente com o conjunto de melhorias nas regiões, percebeu-se também uma redução da criminalidade, o que sugere uma relação entre a melhoria da infraestrutura e a redução de homicídios.

CAPÍTULO

5

CONCLUSÕES

A extração de padrões em dados pontuais é imperativa em diversas aplicações, principalmente no campo da criminalidade. No presente trabalho, apresenta-se alguns conceitos da decomposição Não-Negativa de Tucker e mostra-se como a decomposição juntamente com algoritmos de clusterização podem ser combinados para extrair padrões distribuídos de múltiplas fontes de dados.

No Capítulo 3 apresenta-se em detalhes a metodologia proposta neste trabalho. Primeiramente, uma breve descrição da decomposição Não-Negativa de Tucker é feita, bem como apresenta-se a interpretação da decomposição. Após isso, apresenta-se a modelagem do tensor, a extração dos padrões e finalmente uma comparação da metolologia proposta com algoritmos de agrupamento empregando dados sintéticos é realizada.

A eficácia da metodologia proposta também é atestada por meio de experimentos usando dados reais. Assim dois estudos de casos foram realizados: o primeiro estudo de caso explora os padrões detectados e tenta identificar as relações entre pontos de ônibus e a criminalidade no entorno das escolas e também mostra a relação entre a presença de crimes nos padrões identificados com o desempenho escolar. Já o segundo estudo de caso identifica e explora os padrões no entorno dos homicídios da cidade de São Paulo.

De maneira mais específica, os estudos de caso consistiram de elucidar algumas hipóteses levantadas pelos especialistas. A hipótese H1 foi validada ao mostrar que em determinados contextos de São Paulo, existe uma relação entre a criminalidade e a infraestrutura no entorno das escolas. A Hipótese H2, por sua vez, foi validada ao encontrar padrões que mostram que a criminalidade pode influenciar o desempenho dos estudantes nas escolas de forma negativa. Já a hipótese H3 foi validada ao encontrar padrões que mostram altas taxas de criminalidade nas áreas recreativas próximo às escolas de São Paulo, principalmente durante o período da noite. A hipótese H4 foi validada ao mostrar que existia uma relação entre homicídios e as variáveis envolvidas nas análises, em especial homicídios e arborização. Finalmente, a hipótese

5 foi validada ao mostrar que a melhoria da infraestrutura pode contribuir para uma redução nos homicídios. É importante salientar que a validação das hipóteses se deu em contextos específicos, tendo em vista que as dinâmicas criminais dentro cidade de São Paulo, mudam bruscamente mesmo em regiões muito próximas umas das outras.

A abordagem proposta também contem algumas limitações, como por exemplo, atualmente, no framework, a escolha do *rank* e o número de grupos são parâmetros definidos pelo usuário, porém um conhecimento prévio dos dados e da região de estudos não é necessária. Gerar o tensor é o passo mais demorado de todo o pipeline, no entanto, este passo é considerado com um pré-processamento para o TensorAnalyzer. Todos os padrões extraídos para os estudos de caso (Seções 4.3 e 4.4) demandam 12 segundos, tornando o tempo computacional viável para um agente de segurança ou um especialista em crime realizar uma análise de dados interativa. Um método conveniente para selecionar o rank de uma característica automaticamente depende do score de Fisher (PHAN; CICHOCKI, 2010), no entanto o algoritmo é um pouco custoso de ser implementado, sendo que, a inclusão desta seleção para nosso framework é deixada como trabalho futuro.

REFERÊNCIAS

ALJANABI, K. b. S. A proposed framework for analyzing crime data set using decision tree and simple k-means mining algorithms. **Journal of Kufa for Mathematics and Computer**, v. 1, p. 8–24, 2011. Citado nas páginas 27 e 32.

ARCHWAMETY, T.; KATSIYANNIS, A. Academic remediation, parole violations, and recidivism rates among delinquent youths. **Remedial and Special Education - REM SPEC EDUC**, v. 21, p. 161–170, 05 2000. Citado na página 26.

BENNETT, T.; WRIGHT, R. **Burglars on Burglary: Prevention and the Offender**. Gower, 1984. ISBN 9780566007569. Disponível em: <https://books.google.com.br/books?id=nJDaAAAAMAAJ>. Citado na página 53.

BLOCK, R.; BLOCK, C. R. The bronx and chicago: street robbery in the environs of rapid transit stations. In: **Analyzing Crime Patterns: Frontiers of Practice**. [S.l.]: SAGE, 2000. p. 137–152. Citado na página 46.

BOOTS, B. N.; GETLS, A. Point pattern analysis. Regional Research Institute, West Virginia University, 2020. Citado na página 27.

Bostock, M.; Ogievetsky, V.; Heer, J. D3 Data-Driven Documents. **IEEE Trans. Vis. Comput. Graph.**, v. 17, n. 12, p. 2301–2309, 2011. Citado nas páginas 43 e 46.

BOUSEMA, T.; DRAKELEY, C.; GESASE, S.; HASHIM, R.; MAGESA, S.; MOSHA, F.; OTIENO, S.; CARNEIRO, I.; COX, J.; MSUYA, E. *et al.* Identification of hot spots of malaria transmission for targeted malaria control. **The Journal of infectious diseases**, The University of Chicago Press, v. 201, n. 11, p. 1764–1774, 2010. Citado na página 32.

Brown, D. E. The regional crime analysis program (recap): a framework for mining data to catch criminals. In: **SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)**. [S.l.: s.n.], 1998. v. 3, p. 2848–2853 vol.3. Citado na página 32.

BRUIN, J. D.; COCX, T.; KOSTERS, W.; LAROS, J.; KOK, J. Data mining approaches to criminal career analysis. In: . [S.l.: s.n.], 2006. p. 171–177. Citado na página 32.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J.; TSENG, V. S.; CAO, L.; MOTODA, H.; XU, G. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. Citado na página 28.

CAMPELO, C.; BAPTISTA, C.; JERôNIMO, C. Analyzing mobility patterns from social networks and social, economic and demographic open data. In: . [S.l.: s.n.], 2016. Citado na página 32.

CAO, L. Data science: A comprehensive overview. Association for Computing Machinery, New York, NY, USA, v. 50, n. 3, jun. 2017. ISSN 0360-0300. Disponível em: <https://doi.org/10.1145/3076253>. Citado na página 28.

CARRIAZO, F. Arborización y crimen urbano en bogotá. 12 2016. Citado na página 63.

CEM. **Centro de Estudos da Metrópole**. 2000. <http://centrodametropole.fflch.usp.br/pt-br>. Accessed: 2019-11-26. Citado na página 48.

CHEN, H.; ZENG, D.; ATABAKHSH, H.; WYZGA, W.; SCHROEDER, J. Coplink: Managing law enforcement data and knowledge. **Commun. ACM**, Association for Computing Machinery, New York, NY, USA, v. 46, n. 1, p. 28–34, jan. 2003. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/602421.602441>. Citado na página 32.

CHEN, H.-c.; CHUNG, W.; XU, J.; WANG, G.; QIN, Y.; CHAU, M. Crime data mining: A general framework and some examples. **Computer**, v. 37, p. 50–56, 05 2004. Citado na página 32.

CICHOCKI, A.; MANDIC, D.; LATHAUWER, L. D.; ZHOU, G.; ZHAO, Q.; CAIAFA, C.; PHAN, H. A. Tensor decompositions for signal processing applications: from two-way to multiway component analysis. **IEEE Signal Process. Mag.**, IEEE, v. 32, n. 2, p. 145–163, 2015. Citado na página 33.

CICHOCKI, A.; ZDUNEK, R.; PHAN, A. H.; AMARI, S.-i. **Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation**. [S.l.]: John Wiley & Sons, 2009. Citado nas páginas 28 e 34.

COOK, J.; KLEINSCHMIDT, I.; SCHWABE, C.; NSENG, G.; BOUSEMA, T.; CORRAN, P. H.; RILEY, E. M.; DRAKELEY, C. J. Serological markers suggest heterogeneity of effectiveness of malaria control interventions on bioko island, equatorial guinea. **PloS one**, Public Library of Science, v. 6, n. 9, p. e25137, 2011. Citado na página 32.

CROMWELL, P.; OLSON, J.; AVARY, D. **Breaking and Entering: An Ethnographic Analysis of Burglary**. SAGE Publications, 1991. (Communication and Human Values). ISBN 9780803940260. Disponível em: <https://books.google.com.br/books?id=AckPAQAAMAAJ>. Citado na página 53.

DONOVAN, G.; PRESTEMON, J. The effect of trees on crime in portland, oregon. **Environment and Behavior - ENVIRON BEHAV**, v. 44, p. 3–30, 01 2012. Citado na página 63.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**. [S.l.]: AAAI Press, 1996. (KDD'96), p. 226–231. Citado na página 32.

FRIEDLANDER, M. P.; HATZ, K. Computing non-negative tensor factorizations. **Optim. Method. Softw.**, Informa UK Limited, v. 23, n. 4, p. 631–647, 2008. Citado na página 36.

GARCIA, G.; SILVEIRA, J.; POCO, J.; PAIVA, A.; NERY, M. B.; SILVA, C. T.; ADORNO, S.; NONATO, L. G. Crimanalyzer: understanding crime patterns in são paulo. **IEEE transactions on visualization and computer graphics**, p. 2313–2328, 2021. Citado nas páginas 33 e 40.

GARCIA-ZANABRIA, G.; GOMEZ-NIETO, E. M.; SILVEIRA, J. A.; POCO, J.; NERY, M. B.; ADORNO, S.; NONATO, L. G. Mirante: a visualization tool for analyzing urban crimes. In: **Conference on Graphics, Patterns and Images - SIBGRAPI**. [S.l.]: IEEE, 2020. Citado na página 54.

GATRELL, A. C.; BAILEY, T. C.; DIGGLE, P. J.; ROWLINGSON, B. S. Spatial point pattern analysis and its application in geographical epidemiology. **Transactions of the Institute of British geographers**, JSTOR, p. 256–274, 1996. Citado na página 27.

____. Spatial point pattern analysis and its application in geographical epidemiology. **Transactions of the Institute of British Geographers**, [Royal Geographical Society (with the Institute of British Geographers), Wiley], v. 21, n. 1, p. 256–274, 1996. ISSN 00202754, 14755661. Disponível em: <http://www.jstor.org/stable/622936>. Citado na página 31.

GEOSAMPA. **GeoSampa, Prefeitura de São Paulo**. 2000. <http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx>. Acessado: 01-11-2020. Citado na página 54.

GovernoSP. **Portal Da Transparência Estadual**. 2004. <http://www.transparencia.sp.gov.br/>. Accessed: 2020-03-04. Citado na página 48.

HARSHMAN, R. A. *et al.* **Foundations of the PARAFAC procedure: Models and conditions for an"explanatory"multimodal factor analysis**. [S.l.]: University of California at Los Angeles Los Angeles, 1970. Citado na página 33.

HEISSEL J. A .AND SHARKEY, P. T.; TORRATS-ESPINOSA, G.; GRANT, K.; ADAM, E. K. Violence and vigilance: The acute effects of community violent crime on sleep and cortisol. **Child Dev.**, v. 89, n. 4, p. e323–e331, 2018. Citado nas páginas 26 e 47.

HOU, R. Y.; PATT, Y. N.; WORTHINGTON, B. L.; GANGER, G. R. Disk arrays: high-performance, high-reliability storage subsystems. **Computer**, IEEE Computer Society, Los Alamitos, CA, USA, v. 35, n. 03, p. 30–36, mar 1994. ISSN 1558-0814. Citado na página 32.

HUANG, Y.; ZHANG, P. On the relationships between clustering and spatial co-location pattern mining. In: . [S.l.: s.n.], 2006. p. 513–522. Citado na página 32.

JIN, X.; HAN, J. K-means clustering. In: ____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 563–564. ISBN 978-0-387-30164-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_425>. Citado na página 33.

Joshi, A.; Sabitha, A. S.; Choudhury, T. Crime analysis using k-means clustering. In: **2017 3rd International Conference on Computational Intelligence and Networks (CINE)**. [S.l.: s.n.], 2017. p. 33–39. Citado nas páginas 27 e 33.

JUKIĆ, A.; KOPRIVA, I.; CICHOCKI, A. Noninvasive diagnosis of melanoma with tensor decomposition-based feature extraction from clinical color image. **Biomedical Signal Processing and Control**, p. 755–763, 2013. Citado na página 34.

JÚNIOR, J. C. U. Planejamento da paisagem e planejamento urbano: reflexões sobre a urbanização brasileira. **Revista Mato-Grossense de Geografia**, v. 17, n. 1, 2014. Citado na página 25.

KOLDA, T. G.; BADER, B. W. Tensor decompositions and applications. **SIAM Rev.**, SIAM, v. 51, n. 3, p. 455–500, 2009. Citado nas páginas 33 e 35.

KOSSAIFI, J.; PANAGAKIS, Y.; ANANDKUMAR, A.; PANTIC, M. Tensorly: tensor learning in python. **J .Mach. Learn. Res.**, v. 20, n. 26, p. 1–6, 2019. Citado na página 43.

KREZMIEN, M.; MULCAHY, C.; LEONE, P. Detained and committed youth: Examining differences in achievement, mental health needs, and special education status. **Education and Treatment of Children**, v. 31, p. 445–464, 11 2008. Citado na página 26.

KUO, M.; BACAICOA, M.; SULLIVAN, W. Transforming inner-city landscapestrees, sense of safety, and preference. **Environment and Behavior - ENVIRON BEHAV**, v. 30, p. 28–59, 01 1998. Citado na página 53.

KUO, M.; SULLIVAN, W. Environment and crime in the inner city: Does vegetation reduce crime? **Environment and Behavior - ENVIRON BEHAV**, v. 33, p. 343–367, 05 2001. Citado na página 53.

LAFORTEZZA, R.; CARRUS, G.; SANESI, G.; DAVIES, C. Benefits and well-being perceived by people visiting green spaces in periods of heat stress. **Urban Forestry & Urban Greening**, Elsevier, v. 8, n. 2, p. 97–108, 2009. Citado na página 60.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, v. 401, n. 6755, p. 788–791, 1999. Citado nas páginas 33 e 36.

LIANG, W.; WU, Z.; LI, Z.; GE, Y. Crimetensor: Fine-scale crime prediction via tensor learning with spatiotemporal consistency. **ACM Transactions on Intelligent Systems and Technology (TIST)**, p. 1–24, 2022. Citado nas páginas 33 e 40.

MALATHI, A.; BABOO, S. An enhanced algorithm to predict a future crime using data mining. **International Journal of Computer Applications**, v. 21, 05 2011. Citado na página 32.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 4. ed. [S.l.]: Wiley, 2006. Citado na página 49.

MONTOLIO, D. The effects of local infrastructure investment on crime. **Labour Econ.**, v. 52, p. 210–230, 2018. Citado na página 46.

NEV. **Núcleo de Estudos da Violência da USP**. 2000. <https://nev.prp.usp.br/>. Accessed: 2020-05-11. Citado nas páginas 47, 48 e 55.

OSAC. **Brazil 2020 Crime Safety Report: São Paulo**. [S.l.], 2020. Citado na página 25.

OSELEDETS, I. V. Tensor-train decomposition. **SIAM Journal on Scientific Computingg**, p. 2295–2317, 2011. Citado na página 33.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: machine learning in Python. **J. Mach. Learn. Res.**, v. 12, p. 2825–2830, 2011. Citado nas páginas 43, 45 e 46.

PENG, C.; JIN, X.; WONG, K.-C.; SHI, M.; LIO, P. Collective human mobility pattern from taxi trips in urban area. **PloS one**, p. e34487, 2012. Citado na página 33.

PETROCELLI, M.; PETROCELLI, J. School performance and crime: Theoretical and empirical links. **Southwest Journal of Criminal Justice**, v. 2, n. 2, 2005. Citado na página 26.

PHAN, A. H.; CICHOCKI, A. Tensor decompositions for feature extraction and classification of high dimensional datasets. **Nonlinear Theory and Its Applications, IEICE**, p. 37–68, 2010. Citado nas páginas 34 e 72.

PICORNELL, M.; RUIZ, T.; BORGE, R.; GARCíA-ALBERTOS, P.; PAZ, D.; LUMBRERAS, J. Population dynamics based on mobile phone data to improve air pollution exposure assessments. **Journal of Exposure Science Environmental Epidemiology**, v. 29, 03 2019. Citado na página 32.

PODUR, J.; MARTELL, D. L.; CSILLAG, F. Spatial patterns of lightning-caused forest fires in ontario, 1976–1998. **Ecological modelling**, Elsevier, v. 164, n. 1, p. 1–20, 2003. Citado nas páginas 27, 28 e 31.

Prefeitura de São Paulo. **Sistema Integrado de Informações ao Cidadão**. 2018. Disponível em: <http://www.sic.sp.gov.br/RelatorioEstatistico.aspx>. Acesso em: 16/01/2023. Citado na página 26.

PULLAN, R. L.; STURROCK, H. J.; MAGALHAES, R. J. S.; CLEMENTS, A. C.; BROOKER, S. J. Spatial parasite ecology and epidemiology: a review of methods and applications. **Parasitology**, p. 1870–1887, 2012. Citado na página 32.

RAFSANJANI, M. K.; VARZANEH, Z. A.; CHUKANLO, N. E. A survey of hierarchical clustering algorithms. **Int. J. Math. Comput. Sci.**, v. 5, n. 3, p. 229–240, 2012. Citado nas páginas 28 e 38.

SANDER, H.; ZHAO, C. Urban green and blue: Who values what and where? **Land Use Policy**, v. 42, p. 194–209, 01 2015. Citado na página 63.

SCOPELLITI, M.; CARRUS, G.; ADINOLFI, C.; SUAREZ, G.; COLANGELO, G.; LAFORTEZZA, R.; PANNO, A.; SANESI, G. Staying in touch with nature and well-being in different income groups: The experience of urban parks in bogotá. **Landscape and urban planning**, Elsevier, v. 148, p. 139–148, 2016. Citado na página 60.

SEADE. **IPVS**. 2021. <http://ipvs.seade.gov.br/view/index.php?prodCod=2>. Acessado: 02-11-2020. Citado na página 55.

SEHAB. **Companhia Metropolitana de Habitação de São Paulo**. 2021. <http://www.cohab.sp.gov.br/dadosabertos>. Acessado: 03-11-2020. Citado na página 55.

SILVA, C. S. P.; GRIGIO, A. M.; PIMENTA, M. R. C. Levantamento e espacialização da criminalidade urbana do município de mossoró-rn. **HOLOS**, v. 3, p. 352–362, 2016. Citado na página 25.

SILVA, J. da. **Direito urbanístico brasileiro**. Malheiros Editores, 2008. ISBN 9788574208428. Disponível em: <https://books.google.com.br/books?id=9olHPwAACAAJ>. Citado na página 25.

SILVEIRA, J.; GARCíA, G.; PAIVA, A.; NERY, M.; ADORNO, S.; NONATO, L. G. **TensorAnalyzer: Identification of Urban Patterns in Big Cities using Non-Negative Tensor Factorization**. arXiv, 2022. Disponível em: <https://arxiv.org/abs/2210.02623>. Citado nas páginas 37, 39, 40, 44, 45, 48, 50 e 51.

SMDU. **Secretaria Municipal de Desenvolvimento Urbano**. 2017. <https://www.prefeitura.sp.gov.br/cidade/secretarias/urbanismo/>. Acessado: 03-11-2020. Citado na página 55.

SME. **Secretaria Municipal de Educação**. 2017. <https://educacao.sme.prefeitura.sp.gov.br/>. Acessado: 03-11-2020. Citado na página 55.

SMPR. **Secretaria Municipal de Prefeituras Regionais**. 2017. <http://dados.prefeitura.sp.gov.br/organization/coordenacao-das-subprefeituras>. Acessado: 03-11-2020. Citado na página 55.

SMS. **Secretaria Municipal de Saúde**. 2017. <https://www.prefeitura.sp.gov.br/cidade/secretarias/saude/>. Acessado: 03-11-2020. Citado na página 55.

SMSU. **Secretaria Municipal de Segurança Urbana**. 2017. <https://www.prefeitura.sp.gov.br/cidade/secretarias/seguranca_urbana/>. Acessado: 03-11-2020. Citado na página 55.

SUN, L.; AXHAUSEN, K. Understanding urban mobility patterns with a probabilistic tensor factorization framework. **Transportation Research Part B Methodological**, p. 511–524, 2016. Citado na página 34.

TROY, A.; GROVE, M.; O'NEIL-DUNNE, J. The relationship between tree canopy and crime rates across an urban-rural gradient in the great baltimore region. **Landscape and Urban Planning**, v. 106, p. 262–270, 06 2012. Citado na página 63.

VASILE, F. Uncovering the structure of hypergraphs through tensor decomposition: an application to folksonomy analysis. Digital Repository@ Iowa State University, http://lib. dr. iastate. edu/, 2008. Citado na página 33.

WANG, D.; CAI, Z.; CUI, Y.; CHEN, X. Nonnegative tensor decomposition for urban mobility analysis and applications with mobile phone data. **Transportmetrica A: transport science**, p. 29–53, 2022. Citado na página 34.

WERNECK, G. Georeferenced data in epidemiologic research. **Ciência saúde coletiva**, v. 13, p. 1753–66, 11 2008. Citado na página 32.

WOTTON, B.; MARTELL, D. L. A lightning fire occurrence model for ontario. **Canadian Journal of Forest Research**, NRC Research Press Ottawa, Canada, v. 35, n. 6, p. 1389–1401, 2005. Citado na página 31.

XU, Y.; YIN, W. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. **SIAM J. Imaging Sci.**, v. 6, n. 3, p. 1758–1789, 2013. Citado na página 36.

YAMADA, I.; ROGERSON, P. An empirical comparison of edge effect correction methods applied to k-function analysis. **Geographical Analysis**, v. 35, p. 97 – 109, 11 2010. Citado na página 27.

YUAN, Y.; RAUBAL, M. Extracting clustered urban mobility and activities from georeferenced mobile phone datasets. In: . [S.l.: s.n.], 2011. Citado na página 32.

ZEN, G.; RICCI, E.; SEBE, N. Simultaneous ground metric learning and matrix factorization with earth mover's distance. In: **22nd International Conference on Pattern Recognition**. [S.l.: s.n.], 2014. Citado na página 38.

ZHANG, F.; WILKIE, D.; ZHENG, Y.; XIE, X. Sensing the pulse of urban refueling behavior. In: . [S.l.]: Association for Computing Machinery, 2013. p. 13–22. Citado na página 34.

ZHAO, K.; MUSOLESI, M.; HUI, P.; RAO, W.; TARKOMA, S. **Explaining the Power-law Distribution of Human Mobility Through Transportation Modality Decomposition**. 2015. Citado na página 32.


ICMC
USP
SÃO CARLOS