Dissertação de Mestrado

Uma Nova Metodologia Baseada em Grafos de Ruas Para Investigar a Relação entre Crimes e Pontos de Interesse Usando Dados Georreferenciados: Um Estudo de Caso em Maceió, AL

Débora Barbosa Leite Santos

Orientador:

Thales Miranda de Almeida Vieira

Maceió, Agosto de 2024

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

S586n Silva, Débora Barbosa Leite.
Uma nova metodologia baseada em grafos de ruas para investigar a relação entre crimes e pontos de interesse usando dados georreferenciados : um estudo de caso em Maceió, AL / Débora Barbosa Leite Silva. – 2024.
56 f. : il. color.

Orientador: Thales Miranda de Almeida Vieira.
Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2024.

Bibliografia: f. 54-56.

1. Dados urbanos. 2. Predição de crimes. 3. Grafos de rua (Metodologia computacional). 4. Crimes – Pontos de interesse. I. Título.

CDU: 004

Dedico este trabalho aos meus pais Daniel Barbosa da Silva e Helena Barbosa Leite Silva, e aos meus irmãos Lucas Barbosa Leite Silva e Daniel Barbosa Leite Silva.

Agradecimentos

Gostaria de expressar minha profunda gratidão a todas as pessoas que contribuíram para a realização deste trabalho acadêmico. Cada passo, cada desafio superado, foi possível graças ao apoio e inspiração que recebi ao longo dessa jornada.

Que este trabalho seja uma pequena contribuição ao campo da ciência, inspirando futuras gerações de pesquisadores, assim como eu fui inspirado por aqueles que me cercam.

Muito obrigado a todos.

Resumo

Assim como as cidades, os crimes também evoluíram com o tempo: estão exponencialmente mais intensos, mais violentos e mais modernos, levando à exaustão dos modelos de segurança. Consequentemente, a sociedade, e principalmente os gestores, necessitam de ferramentas sofisticadas para ajudá-los na tomada de decisão. A crescente digitalização de dados da última década tem possibilitado a coleta de dados urbanos em larga escala e com muita agilidade. Isso abre oportunidades de se desenvolverem novas técnicas e ferramentas para análise de dados urbanos massivos, inclusive no escopo da criminalidade urbana. Porém, a análise de dados urbanos georreferenciados em larga escala requer o uso de discretizações espaciais adequadas e o emprego de algoritmos de Aprendizado de Máquina robustos capazes de identificar padrões urbanos complexos. Neste trabalho, apresentamos uma metodologia computacional baseada em grafos de rua para realizar a análise de dados urbanos massivos georreferenciados com o objetivo de investigar a relação entre a ocorrência de crimes e a proximidade de pontos de interesse da cidade (POIs). Em particular, também propomos realizar análises segmentadas de acordo com padrões socioeconômicos das diferentes regiões da cidade, através do uso de algoritmos de agrupamento. Por meio de um estudo de caso realizado na cidade de Maceió, concluímos que que existe, globalmente, correlação entre pontos de interesse e eventos criminais. Além disso, esta correlação é significativamente alterada quando analisados grupos de esquinas segmentadas de acordo com os diferentes padrões socioeconômicos.

Palavras-chave: Predição de Crimes, Dados Urbanos, Probabilidade Condicional, Grafo de Ruas.

Abstract

Just like cities, crimes have also evolved over time: they are exponentially more intense, more violent, and more modern, leading to the exhaustion of security models. Consequently, society, and especially managers, need sophisticated tools to assist them in decision-making. The increasing digitization of data in the last decade has enabled the large-scale and agile collection of urban data. This opens up opportunities to develop new techniques and tools for analyzing massive urban data, including within the scope of urban crime. However, the analysis of large-scale georeferenced urban data requires the use of appropriate spatial discretizations and the employment of robust Machine Learning algorithms capable of identifying complex urban patterns. In this work, we present a computational methodology based on street graphs to perform the analysis of massive georeferenced urban data with the aim of investigating the relationship between the occurrence of crimes and the proximity of points of interest (POIs) in the city. In particular, we also propose to carry out segmented analyses according to the socioeconomic patterns of different city regions using clustering algorithms. Through a case study conducted in the city of Maceió, we conclude that there is, globally, a correlation between points of interest and criminal events. Moreover, this correlation is significantly altered when analyzing groups of corners segmented according to different socioeconomic patterns.

Keywords: Public Security; Data Science; Spatiotemporal data analysis; Urban data; Big Data.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

IA	Inteligência Artificial
PMAL	Polícia Militar do Estado de Alagoas
PCA	Principal Component Analysis

Sumário

Capítulo 1

Introdução

1.1 Contextualização

De acordo com Silva (2008), o processo de urbanização sem planejamento gera problemas sociais, ambientais e vulnerabilidade social, causando um aumento significativo nas taxas de crime. Com o aumento exponencial de crimes e as limitações de recursos, tais como: viaturas, policiais, armamento e tempo, medidas precisam ser tomadas para a rápida resolução, ou melhor gestão, dos problemas oriundos desse processo de urbanização.

Além disso, assim como as cidades, os crimes estão em processo contínuo, dinâmico e complexo, estão mais intensos, violentos e mais modernos, levando à exaustão dos modelos de segurança. Consequentemente, são necessários métodos eficientes que reduzam a ocorrência de crimes e diminuam as perdas econômicas, alocando proativamente recursos policiais. Os gestores e as agências policiais necessitam de ferramentas que agilizem a tomada de decisão e a realização de ações efetivas para garantir a segurança da cidade (Hu et al., 2023b).

Na última década, com a rápida evolução da Internet, houve um crescimento na realização de processos de transformação digital, resultando em uma crescente digitalização de dados que possibilitou a coleta de dados urbanos em larga escala e com muita agilidade. Portando os novos trabalhos tiveram acesso aos mais diversos dados, tais como: censo, uso da terra, pontos de interesse (POIs), dados socioeconômicos, reclamações de serviço público e dados meteorológicos (Hu et al., 2023b). Por outro lado, análises complexas destes dados se tornaram possíveis com o surgimento de técnicas sofisticadas de Aprendizagem de Máquina. Com o advento do Aprendizado Profundo, por exemplo, é possível transformar dados brutos de alta dimensão, coletados das mais diversas fontes, em uma representação latente adequada, ou seja, um vetor de características que posteriormente poderá ser utilizado para resolver problemas complexos de Ciência de Dados (Escovedo e Koshiyama, 2020).

Isso dá a oportunidade de se desenvolverem novas técnicas e ferramentas para análise de

dados urbanos massivos, inclusive no escopo da criminalidade urbana. Em particular, técnicas de análise visual (*visual analytics*) podem ser muito úteis nesse contexto (Salinas et al., 2022).

Atualmente existem poucos estudos que tratam do problema específico de análise da influência de pontos de interesse na ocorrência de eventos criminais usando técnicas modernas de Ciência de Dados. Assim como afirma a pesquisa (Hu et al., 2023a), que ainda é difícil recolher informação suficiente e completa sobre os fatores que influenciam as atividades criminais e a falta de informação afeta o aprendizado dos parâmetros ótimos do modelo. Portanto, a forma como os investigadores podem fazer uso desta informação diversificada resultante da dinâmica das intenções criminosas é uma tarefa muito procurada e essencial para os dias atuais.

Neste trabalho, pretendemos realizar análise de dados urbanos massivos georreferenciados através de métodos de agrupamento e métodos estatísticos, de modo que possamos identificar a relação entre crimes e pontos de interesse (POIs). Para tal objetivo, coletamos os dados do IBGE e da API do Google Maps, através de algoritmos de localização de cobertura máxima. Tais dados foram integrados a um grafo de ruas que une os dados urbanos (socioeconômicos, pontos de interesse etc) à base de crimes do estado de Alagoas, fornecida pela PM-AL. Aplicamos técnicas não supervisionadas nestes dados para agrupar (clusterizar) esquinas com padrões similares socioeconômicos. Por fim, analisamos os dados através do Teste Qui-quadrado, do V de Cramér e Probabilidade Condicional.

1.2 Objetivos

1.2.1 Objetivo Geral

Pretendemos apresentar e experimentar uma metodologia para encontrar padrões urbanos a nível de esquinas, e que possam estar relacionados aos eventos de crimes, usando como recurso as diversas fontes de dados socioeconômicos e pontos de interesse. Estes serão integrados ao grafo de ruas através da biblioteca CityHub. Procuraremos responder às seguintes perguntas de pesquisa:

1. As esquinas da cidade onde há atividade criminal podem ser separadas em grupos com características socioeconômicas distintas?
2. É possível analisar a correlação entre ocorrência criminal e a proximidade de tipos específicos de pontos de interesse integrando os dados nas esquinas do grafo de rua?
3. Existe correlação entre a ocorrência de crimes e a proximidade de tipos específicos de ponto de interesse na cidade de Maceió?
4. Caso exista, quais os tipos de ponto de interesse mais influentes nesse contexto?

5. A influência dos pontos de interesse muda de acordo com características socioeconômicas das esquinas (grupos)?

1.2.2 Objetivos Específicos

- Realizar a engenharia de características para selecionar o melhor subconjunto de características.
- Coletar dados de diversas fontes, para compor os recursos para as análises que serão feitas;
- Integrar todos os dados a um grafo das ruas da cidade de Maceió.
- Utilizar técnicas agrupar as esquinas.
- Experimentar técnicas estatísticas para ranquear os tipos de pontos de interesse de acordo com sua influência na ocorrência criminal.
- Avaliar o resultado das análises aplicando a um estudo de caso da cidade de Maceió, AL.

Este trabalho está estruturado da seguinte forma: No capítulo 2 será explanado os trabalhos já existentes na literatura, relacionados à análise dos dados criminais e ao problema de predição de crimes. Após isso, no capítulo 3 será apresentada a metodologia do trabalho, os passos para obtenção dos dados, representação desses dados computacionalmente, além das técnicas de agrupamento de hotspots e ranqueamento das características. Para respondermos às perguntas de pesquisa, no capítulo 4 apresentaremos um estudo de caso, voltado à cidade de Maceió. Por fim, apresentaremos no capítulo 5 as conclusões e trabalhos futuros.

Capítulo 2

Trabalhos relacionados

2.1 Análise de dados criminais

No âmbito acadêmico é de extrema importância a objetividade na escolha dos parâmetros que serão utilizados na pesquisa científica. Em vista disso, a área da engenharia de características busca investigar quais características são de fato relevantes para as análises que serão realizadas. Por isso, nesta seção apresentaremos alguns artigos que utilizaram dados externos junto com crimes, tais artigos foram usados como base para a escolha das características deste artigo.

O artigo Tensor Analyzer (Silveira et al., 2022) faz uso de múltiplas fontes de dados, através da junção dos dados em um Tensor. Através da combinação dos dados externos e dos números de homicídios, o artigo busca responder algumas hipóteses e analisar os padrões encontrados. Os dados usados são diversificados, dentre eles: infraestrutura, habitações, arborização, cultura, educação, saúde, segurança.

Além disso, o trabalho (Alves et al., 2013) faz uso dos dados de cidades brasileiras do ano de 2000 disponibilizados no DATASUS [1]. Os autores notaram relações entre o número de homicídios, o tamanho da população e as dez métricas urbanas. A hipótese é que as leis naturais de escala de tais atributos são melhores modulados por equação estocástica. Usando a equação estocástica e análise de regressão é possível mensurar os valores das variáveis com mais precisão. As dez métricas urbanas usadas foram: trabalho infantil, população idosa, população feminina, PIB, PIB per capita, número de analfabetos, renda familiar média, população masculina, número de instalações sanitárias e número de desempregados.

O próximo artigo (Wang et al., 2013) utiliza a mineração de dados espaciais para mapear pontos críticos e investigar a relação entre variáveis socioeconômicas e criminais. Tanto o crime alvo como as variáveis relacionadas no conjunto de dados espaciais são plotados em mapas de grade com as mesmas dimensões e logo depois agrupados em padrões. Os dados utilizados

[1] <http://www.datasus.gov.br/>

foram de uma cidade do nordeste dos Estados Unidos, os dados foram: crimes, prisões, casas hipotecadas, densidade habitacional, população, distância para faculdades.

Por último temos o Survey (Hu et al., 2023a) que realiza um levantamento abrangente da última década (2013–2023) dos métodos de análise de eventos criminais (ACE) na perspectiva da fusão de informações. Primeiro investigam as categorias de dados criminais e apresentam brevemente as tarefas existentes, bem como as métricas de avaliação. Com base na leitura dos diversos estudos, o trabalho resumiu as principais características que eram levadas em consideração na análise de crimes: Demografia (Wang et al., 2019), Meteorológicos (Liang et al., 2022), Feriado (Han et al., 2020), Mobilidade humana (Zhao et al., 2022), Locais de interesse (Yang et al., 2018), Redes rodoviárias (Belesiotis, Papadakis e Skoutas, 2018), Reclamação de serviço público (Huang et al., 2018), Policial (Mukhopadhyay et al., 2016).

2.2 Identificação das relações entre os pontos de interesse e crimes

Os dados contêm riquezas de informações, que são usadas atualmente por aqueles que buscam analisar as atuais atividades e ampliar suas visões para o futuro. Os dados são ainda mais ricos quando os integramos às necessidades públicas, eles nos proporcionam entender todo o contexto por trás da dinâmica espacial e temporal do crime e oferecem aos pesquisadores a possibilidade de estudar a relação entre crime e as mais diversas fontes (Hu et al., 2023b).

Um exemplo de trabalhos que exploram esses dados é o CrimAnalyzer (Zanabria et al., 2021b), uma ferramenta analítica assistida por visualização que permite aos usuários analisar o comportamento de crimes em regiões específicas de uma cidade, verificando, por tanto, o contexto urbano no qual o crime estava inserido. Outro exemplo é o CityHub, que integra dados urbanos a nível de esquinas para possibilitar a realização de análises visuais e outras tarefas no contexto de Ciência de Dados. Apesar da existência desses artigos, até poucos anos atrás, poucos trabalhos focaram na visualização de dados relacionados a hotspots de crimes de rua.

Além disso, podemos levar em consideração o artigo (Hu et al., 2023b), que fez uma revisão sistemática de métodos de previsão de crime. Os autores analisaram diversos artigos e concluíram que os métodos que usavam combinados dados criminais e dados externos, tinham uma melhora significativa na precisão de suas previsões. O que reforça ainda mais o objetivo deste trabalho de comparar dados socioeconômicos com os dados criminais de Alagoas.

2.3 Predição criminal

Existem alguns trabalhos recentes que tratam da análise visual e predição de crimes usando algoritmos modernos de Ciência de Dados. Um dos mais relevantes é o CrimAnalyzer (Zanabria et al., 2021b), que é uma ferramenta de análise visual que permite analisar o comportamento dos crimes em uma região específica da cidade. Ela tem o objetivo de identificar hotspots, padrões de crimes e como esses evoluiram ao longo do tempo. Na identificação de hotspots, os autores tentam identificar não só aqueles em que os crimes são intensos mas também os que têm alta frequência, ou seja, alta probabilidade. Para isso, fazem uso da Fatoração de Matrizes Não Negativas (Cichocki et al., 2009), obtendo assim os hotspots e a intensidade deles ao longo do tempo. Uma das limitações dessa abordagem foi o espaço de discretização utilizado: as unidades censitárias da cidade de São Paulo. Uma vez que, modificada a unidade de agregação dos dados, os resultados das análises podem mudar significativamente, pois não são espacialmente precisos.

O trabalho Mirante(Zanabria et al., 2020) relata a importância de analisar o micro espaço, pois considera que a maioria dos crimes são concentrados em pequenas regiões cuja escala é próxima a uma única esquina, ou trecho de rua entre duas esquinas (segmento de rua). Por isso, o espaço de discretização usado no Mirante é a nível de rua, onde as esquinas correspondem aos nós de um grafo e os segmentos de rua às arestas. Os dados de crimes são agregados em cada nó do gráfo e representados como um mapa de calor quando uma região de interesse é selecionada, como mostra na figura 2.1.

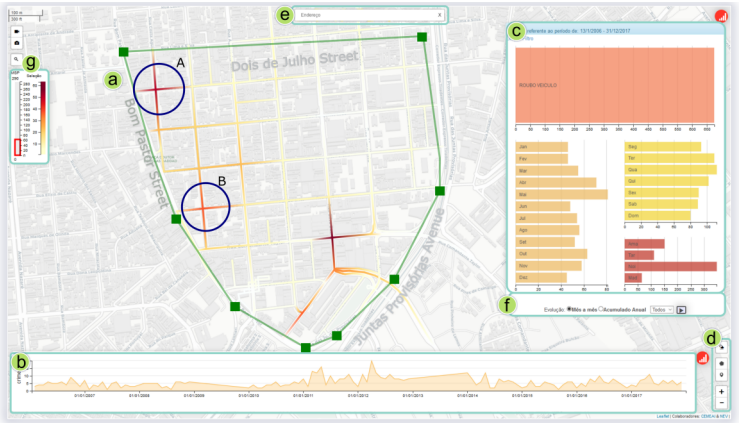

Figura 2.1: Gráfico com um mapa de calor da cidade. Fonte: Mirante (Zanabria et al., 2020)

Para associar cada registro de crime a um nó do grafo encontramos o problema mostrado na figura 2.2. Ao usar a distância Euclidiana, o registro “p”, seria incorporado à esquina de outro segmento, não condizente com a realidade. Porém levando em consideração a estrutura do grafo, se primeiro associarmos ‘p’ à aresta mais próxima e_n, o registro deverá pertencer ao vértice v_n. Então, ao associar à rua mais próxima, e depois procurar a esquina mais próxima, o registro fica na rua(segmento) mais apropriado. No Mirante, os hotspots são identificados em um mapa

Figura 2.2: Grafo da cidade. Fonte: Crime prediction (Vieira et al., 2022)

de calor, de acordo com a intensidade de crimes na esquina. É possível ainda visualizar como os crimes mudaram de esquina ao longo do tempo. O que o artigo traz como limitação é não considerar em suas análises outras bases de dados urbanos, como por exemplo as localizações de bares, escolas, pontos de ônibus, etc. Além disso, percebe-se a diferença com o artigo anterior (CrimAnalyzer), já que apenas as esquinas com maior intensidade são analisadas, sem levar em consideração as esquinas com maior probabilidade, mas baixo índice de crimes.

Continuando na mesma linha de pesquisa, o artigo CRIPAV (Zanabria et al., 2021a) também usa o espaço amostral em grafos de rua, com cada crime atrelado a um nó do grafo de ruas (intersecção entre as ruas). A discretização do tipo grade não foi usada devido ao problema da escolha de sua resolução. A Figura 2.3 mostra como a análise pode ser prejudicada dependendo da resolução da grade, tornando difícil perceber qual local realmente precisa de atenção das autoridades. Além disso, não é possível fazer uma análise temporal detalhada, uma vez que crimes que ocorrem em distâncias menores que a resolução da grade, ainda seriam contados na mesma célula, sem a possibilidade da percepção da mudança no comportamento. Por esses motivos, os trabalhos atuais focam na análise dos dados a nível de grafo de ruas.

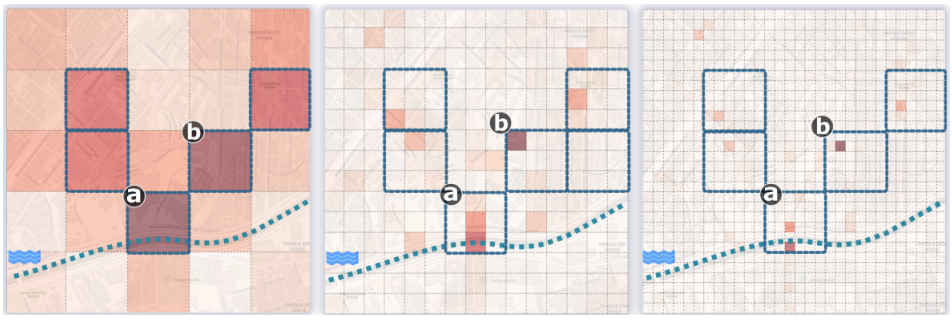

Figura 2.3: Como os crimes mudam a área em que pertencem, de acordo com a mudança na resolução das grades. Fonte: CRIPAV (Zanabria et al., 2021a)

Uma das contribuições do artigo é a forma como o modelo detecta hotspots, pois assim como o foco da análise do CRIPAV é mostrar pontos com alta criminalidade, também é mostrar aqueles que não são intensos (ou volumosos) mas apresentam alta probabilidade. Esta análise é

feita através da matriz estocástica. A probabilidade de crimes em um ponto âncora (nó do grafo) é dada pelo vetor estacionário desta matriz, que é construída com a série temporal de crimes de cada ponto âncora. Além disso, o artigo cria grupos de hotspots com similar dinâmica temporal, usando o mecanismo de aprendizado profundo Hotspot2Vec, desenvolvido pelos próprios autores. Como limitação, assim como os descritos anteriormente, esse trabalho não considera outros tipos de dados urbanos (pontos de ônibus, escolas, bares, etc.).

A CityHub (Salinas et al., 2022) é uma biblioteca que também utiliza como discretização espacial um grafo de ruas da cidade, e permite a integração e análise de diversas fontes de dados urbanos. Esta biblioteca está sendo utilizada na presente pesquisa para manipulação de diversas bases de dados que serão usadas, e para exportar vetores de características associados a cada nó do grafo de ruas. A ferramenta suporta quatro tipos de camadas de dados: dados baseados em pontos (PB), dados de domínio regional (RD), dados agregados de polígonos (PA) e dados esparsos (SD).

Por fim, o TensorAnalyzer (Silveira et al., 2022) apresenta uma nova abordagem para detectar padrões relevantes de múltiplas fontes de dados baseada na combinação da Decomposição de Tucker e de algoritmos de clusterização hierárquica. Neste trabalho, os autores tentaram validar duas hipóteses. A primeira propõe que a infraestrutura urbana pode ser vista como um atrator de crime e a segunda tenta provar que existe uma relação entre o desempenho escolar e a taxa de criminalidade. O artigo conclui que a relação entre homicídios e escolas particulares depende das características do padrão que se está analisando, ou seja, a validação das hipóteses se deu em contextos específicos. Interessante destacar que o próprio usuário escolhe o número de grupos e o posto (rank) da decomposição. Sugere-se que trabalhos futuros poderiam propor um método adequado para selecionar o rank de uma característica automaticamente dependendo do score de Fisher.

Em nossa proposta, utilizaremos a ferramenta CityHub para representar dados urbanos a nível de esquina, incluindo-se dados de ocorrências criminais; as características das vizinhanças das esquinas serão obtidas usando-se distância no grafo.

Capítulo 3

Metodologia

Neste capítulo apresentaremos a metodologia proposta neste trabalho, descrevendo os principais procedimentos que serão seguidos. As etapas do processo de pesquisa incluem:

1. **Compilação da base de dados:** Inicialmente adquirimos os dados das bases criminais. Após isso, coletamos os dados socioeconômicos do censo e também os pontos de interesse da API do google maps.(seção 3.1);

2. **Representação computacional de dados:** Uma vez obtidos os dados, estes precisam ser preprocessados para uma representação computacional concisa, eficiente e precisa. Então nessa seção descrevemos como os dados serão integrados a nível de esquina utilizando a biblioteca CityHub (seção 3.2)

3. **Agrupamento:** A próxima etapa de nossa metodologia é utilizar algoritmos de agrupamento para agrupar as esquinas da cidade. Para ser realizado o agrupamento as características serão aquelas que são relacionadas aos dados socioeconômicos. Utilizamos dois tipos de agrupamento, o primeiro agrupou todas as esquinas da cidade. Já o segundo, levou em consideração apenas as esquinas hotspots (com um limiar para a quantidade de crimes), ou seja agrupamos os hotspots em grupos que apresentam características similares. O agrupamento é opcional e em algumas análises não foi realizado (seção 3.3);

4. **Ranqueamento dos pontos de interesse:** O agrupamento criado na etapa anterior nos possibilitou análise e ranqueamento de pontos de interesse baseado na sua relação com crimes, através de diversas métricas.(seção 3.4).

Cada etapa desempenha um papel crucial no desenvolvimento da pesquisa, com o objetivo de responder às perguntas de pesquisa da seção 1.2.1 . A figura **??** apresenta a visão geral da metodologia. Adiante detalharemos cada passo da metodologia. No capítulo seguinte mostraremos a aplicação dessa metodologia em um Estudo de caso realizado na cidade de Maceió/AL.(Capítulo 4)

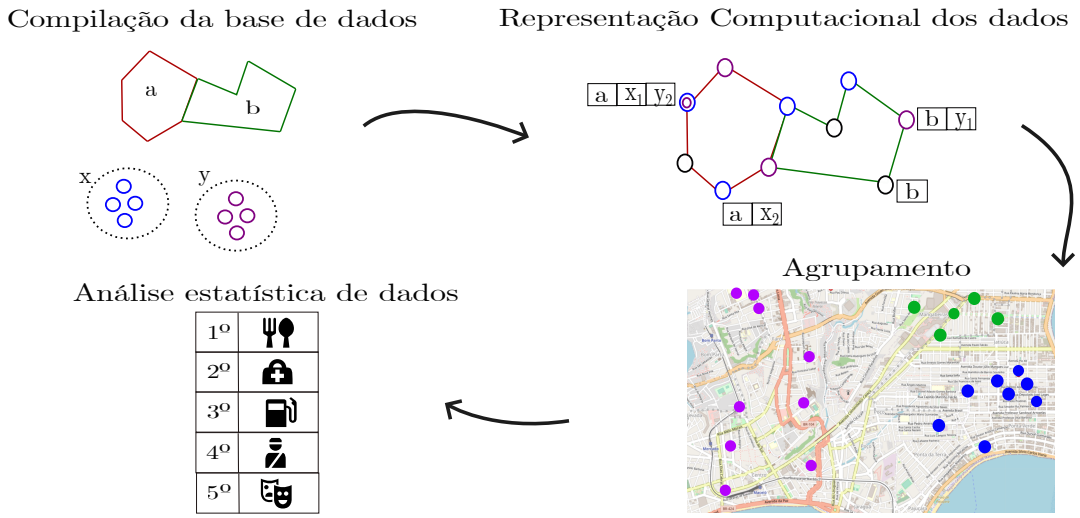

Figura 3.1: Visão geral da metodologia. Fonte: Autor

3.1 Compilação da base de dados

Como base para o desenvolvimento dessa pesquisa usaremos os dados obtidos em três principais fontes: Polícia Militar do Estado de Alagoas (PM-AL), através de um termo de cooperação firmado entre PM-AL e UFAL; Censo do IBGE; e a API do Google Maps. Nas seções abaixo estão descritos os processos para obtenção dos dados de cada uma das fontes.

3.1.1 Obtenção dos dados de Crimes

Através de um acordo de cooperação firmado entre PM-AL e UFAL, a Polícia Militar forneceu os dados de crimes de Alagoas do período de Agosto de 2021 até Novembro de 2021, provenientes de seu banco de registro de ocorrências. Os dados contam com cerca de 2794 amostras da cidade de Maceió apresentando informações detalhadas para cada incidente, incluindo o tipo de crime, localização (cidade, endereço e coordenadas geográficas), data e hora da ocorrência. Destacamos que, como descrito no capítulo 2, as informações espaciais de latitude e longitude requerem a aplicação de técnicas de discretização espacial para a obtenção de representações computacionais eficientes que possibilitem análises robustas.

Necessitamos fazer algumas escolhas devido a heterogeneidade dos dados, realizamos duas delimitações. Primeiro quanto à dimensão temporal, estamos usando um recorte temporal do ano de 2021, refletindo nossa preocupação de que os dados sejam condizentes com a temporalidade dos outros dados obtidos. Além disso, focamos nossa análise nos crimes do tipo 'roubos a transeuntes', uma vez que diferentes tipos de crimes podem exibir comportamentos distintos.

3.1.2 Obtenção dos dados do IBGE

Para o enriquecimento das análises dos crimes, buscou-se então os dados socioeconômicos da região estudada, ou seja, a cidade de Maceió. Uma vez que o novo censo do IBGE de 2023 ainda não estava com os microdados disponíveis, utilizamos os dados do censo demográfico do IBGE de 2010, disponíveis no site do IBGE [1].

Esses dados contêm informações de cada setor censitário, que é a menor unidade territorial para a qual o IBGE divulga informações. Para cada setor censitário, os dados continham cerca de 377 variáveis, que abrangiam as seguintes características: proporção de habitantes por sexo, o quantitativo de pessoas com determinados valores de idade, condição do domicílio; pessoas responsáveis pelo domicílio; taxa de alfabetização; e características dos domicílios particulares.

Após as análises, percebeu-se formas de unir colunas e transformar dados, para que cada coluna do dataset agregasse de fato informações relevantes. A base de dados do IBGE e os cálculos deste projeto, fazem uso de códigos para as variáveis, as descrições dessas variaveis se encontram no arquivo "Documentação Agregado dos Setores 2010", disponível no link já mencionado.

O primeiro documento analisado foi *"DomicilioRendaAL.csv"*, para extrair a informação da Renda Média Por Domicílio, usou-se o total do rendimento mensal, dividido pela soma de quantos domicílios existem naquela região. Seguindo a seguinte fórmula:

$$RendaMediaPorDomicilio = \frac{V002}{(V005 + ... + V014)}$$

Fazendo uso ainda do documento *"DomicilioRendaAL.csv"*, calculamos a taxa de responsáveis por domicílio sem renda. No numerador temos a diferença entre o total de pessoas responsáveis e quantos responsáveis existem com rendimento positivo. No denominador o total de responsáveis.

$$ResponsvelSemRenda = \frac{V020 - V021}{V020}$$

Os próximos documentos analisados foram *"Pessoa01AL.csv"* e *"Pessoa13AL.csv"*, com o objetivo de extrairmos a taxa de alfabetizados com idade entre 7 e 15 anos. Na equação a seguir, o numerador da fração é a soma da quantidade de pessoas alfabetizadas (documento Pessoa01AL), que tem a idade entre 7 anos e 15 anos. No denominador temos quantas pessoas com essa idade existem no setor (documento Pessoa13AL). Dividindo o primeiro pelo segundo obtemos a taxa de alfabetização entre 7 e 15 anos de idade.

[1] <https://ftp.ibge.gov.br/Censos/Censo_Demografico_2010/Resultados_do_Universo/Agregados_por_Setores_Censitarios/>

$$Alfabetizados7a15anos = \frac{(V004 + ... + V012)}{(V041 + ... + V049)}$$

Continuando no documento *"Pessoa13AL.csv"*, criamos algumas taxas referente às idades, a primeira foi a taxa de pessoas menores de 18 anos. Na equação a seguir, o numerador da fração é a soma da quantidade de pessoas que são menores de 18 anos. No denominador temos quantas pessoas residem no setor censitário.

$$Menos18anos = \frac{(V022 + V035 + ... + V051)}{V001}$$

A próxima medida é a taxa de pessoas entre 18 e 65 anos de idade. Na equação a seguir, o numerador da fração é a soma da quantidade de pessoas que têm idade entre 18 e 65 anos. No denominador temos quantas pessoas residem no setor censitário.

$$Entre18e65anos = \frac{(V052 + ... + V098)}{V001}$$

Para encerrar temos a taxa de pessoas com idade maior que 65 anos. Na equação a seguir, o numerador da fração é a soma da quantidade de pessoas que têm idade maior que 65 anos. No denominador temos quantas pessoas residem no setor censitário.

$$Maior65anos = \frac{(V099 + ... + V134)}{V001}$$

O dataset resultante dessa base de dados contém 3724 amostras de setores censitários e ao final do pré-processamento apenas 6 variáveis, levando em consideração as que foram mostradas anteriormente e a variável que continha o código do setor.

3.1.3 Problema de localização de cobertura máxima (MCLP)

Para possibilitar a análise da influência da proximidade de diferentes tipos de ponto de interesse nas ocorrências criminais, nos propomos a adquirir dados de diversos tipos de pontos de interesse da região de Maceió utilizando a API Places do Google Maps. Para realizar a extração desses dados usando a API, realizamos diversas consultas usando os seguintes parâmetros: texto de busca, raio, latitude e longitude. O texto de busca é cada ponto de interesse que estamos buscando, raio fizemos alguns teste mas optamos por dois quilômetros. Já latitude e longitude, em cada consulta precisamos fornecer pontos estratégicos que, com um raio de dois quilômetros, cobríssem toda a cidade de Maceió. Vale salientar que há custos para executar as requisições na API para cada local, de modo que se conseguisse uma boa cobertura da cidade com a menor quantidade possível de requisições é importante.

Para conseguirmos tais pontos, nos baseamos no artigo (Church e Revelle, 1974) que aborda o problema de localização de cobertura máxima, exemplificando como podem ser usadas técnicas para fornecer a solução mais desejável para o problema de cobertura de decisão no que diz respeito à localização de instalações públicas. O problema é definido como: dado um conjunto X com N pontos planos, uma quantidade de círculos K e um raio r encontrar o centro dos círculos que maximizem a cobertura de pontos de X, ou seja, que tenha a maior quantidade possível de pontos de X no interior dos círculos, como mostrado na figura 3.2.

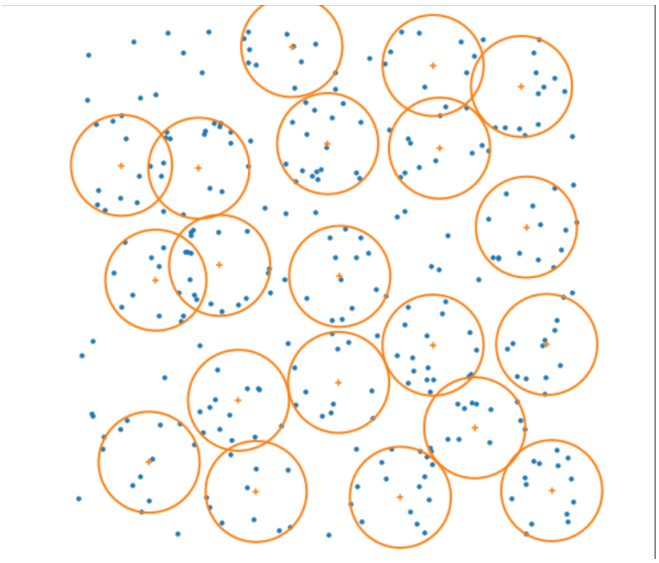

Figura 3.2: Problema de localização de cobertura máxima. Fonte: Can Yang

Para selecionar os pontos âncoras (centros dos círculos) que cobrissem toda a cidade de Maceió nós usamos como base o algoritmo proposto no site(Can Yang, 2019). O método usado neste trabalho gera aleatoriamente um conjunto de candidatos nos dados fornecidos. A seguir mostramos a formulação matemática do método, tal que x_j denota que o $j-simo$ local é selecionado como ponto âncora (1 para sim, 0 para não). O y_i denota se o ponto i é coberto por qualquer um das âncoras escolhidas, dentre um raio r definido, (1 para sim, 0 para não).

$$\sum_{i \in I} y_i$$

$$x_j \in \{0,1\}, j \in J$$

$$y_i \in \{0,1\}, i \in I$$

$$\sum_{j \in J} x_j = K$$

$$\sum_{j \in N_i} x_j \geq y_i, N_i = \{j \in J : d_{ij} \leq r\}$$

Em I temos o conjunto de todos os pontos, estes são as esquinas da cidade, representados

por um par de latitude e longitude gerados a partir da ferramenta CityHub. J é o conjunto de pontos gerados aleatoriamente, que serão os possíveis pontos âncoras que podem ser escolhidos para cobrir a cidade.De forma empírica, concluímos que a quantidade que gera melhores resultados foi 1000 opções de sites, dos quais o algoritmo escolherá a quantidade de K pontos que melhor cobriram a cidade,o valor de K foi determinado como 20 pontos.

Além disso, $d_{(i,j)}$ é a distância de um ponto da cidade i para um local âncora. No artigo original, utilizou-se a distância Euclidiana. Porém, como estamos tratando de dados georreferenciados, adotamos a distância de Haversine[2], que é definida pela equação 3.1 como a distância angular entre dois pontos na superfície de uma esfera. Por fim, K é o número de quantos pontos âncoras deverão ser selecionados e r é o raio do círculo.

$$D(x,y) = 2\arcsin[\sqrt{\sin^2((x_{lat} - y_{lat})/2) + \cos(x_{lat})\cos(y_{lat})\sin^2((x_{lon} - y_{lon})/2)}] \quad (3.1)$$

Finalizando o algoritmo conseguimos um conjunto G com 20 coordenadas (latitude e longitude) que cobre toda a cidade de Maceió, como mostrado na figura 3.3. Inevitavelmente, na escolha desses pontos tivemos intersecções. Porém, as possíveis repetições de pontos de interesse serão tratadas na seção seguinte.

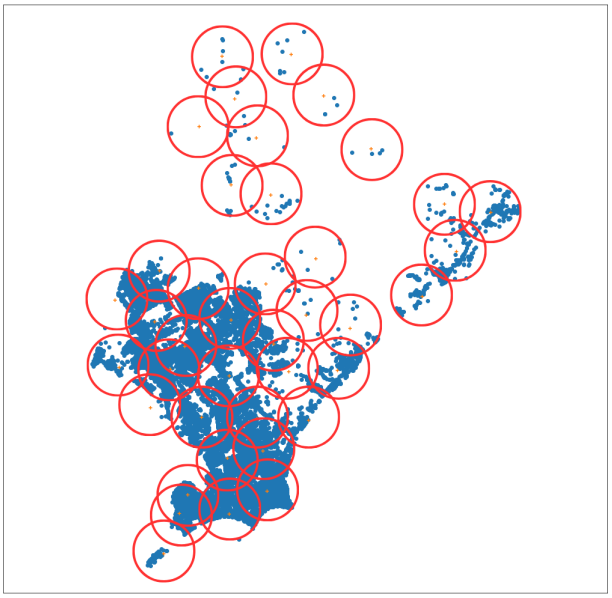

Figura 3.3: Pontos âncoras atingindo toda a cidade de Maceió. Fonte: Autor

[2]<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.haversine_distances.html>

3.1.4 Obtenção dos dados na api do Google Maps

Para obtenção dos diversos pontos de interesses(POI), utilizamos a *Google Places API Web Service* [3]. Dentre os tantos serviços disponíveis de forma gratuita no google, a API Places é um serviço que aceita solicitações para obtenção de dados de um certo local. Através desta API, podemos fazer uma pesquisa de texto que retorna informações sobre um conjunto de locais com base em uma string, por exemplo, "posto de combustível", “escola” e ”faculdade”. O serviço responde com uma lista de locais correspondentes à string de texto e a todos os direcionamentos de localização definidos.

Fornecemos apenas três parâmetros de entrada: raio, localização e consulta. A localização são os pontos âncoras adquiridos na seção anterior (conjunto G). Tais pontos foram calculados com o raio de 2 quilômetros, para assim permitir a cobertura total da cidade. Portanto, em cada requisição da API passamos um ponto âncora por vez, informando o raio de busca de 2 quilômetros.

O terceiro parâmetro passado foi o texto de consulta, ou seja, quais estabelecimentos estávamos procurando. Os POIs usados nessa etapa foram aqueles encontrados nos trabalhos já existentes na literatura, como foi apresentado na seção 2.1. A tabela 3.1 tem o resumo de quais POIs foram escolhidos, além da quantidade de cada ponto que foi conseguida através da API.

Além desses parâmetros, por padrão, para cada solicitação feita, a api retorna até 20 resultados por página. No entanto, cada pesquisa pode retornar até 60 resultados, divididos em três páginas. Se a pesquisa retornar mais de 20 resultados, a resposta incluirá um valor adicional: *next page token*. Quando adicionamos esse novo parâmetro, conseguimos realizar uma nova solicitação para adquirir o próximo conjunto de resultados. Em virtude desse limite de 60 resultados, não garantimos que recuperamos todos os pontos de interesse em um dado círculo. Porém isso não trás prejuízo para nossas análise, já que o que analisamos não é a quantidade de pontos de interesse na vizinhança de uma esquina, mas sim a existência ou não de determinado ponto de interesse na vizinhança, ou seja nossas variáveis são binárias.

Após a realização das pesquisas, observamos a existência de dados repetidos, devido a sobreposição dos raios de cada ponto escolhido. Porém isso foi rapidamente resolvido, excluindo aqueles pontos que eram semelhantes. O resultado final dessa seção é possível observar na tabela 3.1 mostrada a seguir.

[3] <https://developers.google.com/maps/documentation/places/web-service/overview?hl=pt-br>

Tabela 3.1: Pontos de Interesse

Tipo de Ponto de Interesse	Quantidade
Escola	632
Faculdade	113
Universidade	124
Teatro	39
Cinema	15.
Museu	91
Ginásio	221
Estádio	87
Hospital	99
Emergência	209
Batalhão	22
Delegacia	99
Bombeiro	43
Hotel	265
Bar	501
Restaurante	981
Posto de gasolina	234
Transporte	628
Agência Bancária	64

3.2 Representação computacional dos dados

Como percebe-se na seção anterior, temos duas principais dificuldades com os dados. Primeiro, eles apresentam diferentes granularidades, uma vez que os dados do IBGE abrangem toda Alagoas e os pontos de interesse fornecidos são apenas para a cidade de Maceió. A segunda dificuldade é a agregação espacial dos dados. Os dados do IBGE são agregados a nível de setor censitário, em forma de polígonos, em contrapartida os POIs e crimes são passados a nível de esquinas, por ponto.

Como forma de integrar todos esses diversos dados, usamos a biblioteca CityHub. Inicialmente, a biblioteca adota um grafo de ruas da cidade como domínio espacial. Para isso, o CityHub explora as funcionalidades do OSMnx [4], e NetworkX [5]. Após carregar o grafo da cidade, incorporamos ao grafo todos os pontos de interesse e os dados do IBGE.

[4] <https://osmnx.readthedocs.io/en/stable/>
[5] <https://networkx.org/>

Os dados socioeconômicos do IBGE são agregados em regiões espaciais, ou seja, o formato como eles são fornecidos é em setores censitários. O CityHub pode importar dados agregados de polígonos através de um arquivo poligonal que descreve regiões espaciais da cidade, fornecidas como um arquivo KML, SHP ou CSV. Para atribuir dados agregados de polígonos a um nó gráfico de rua, o vértice da malha poligonal mais próximo do nó é encontrado usando a árvore de malha poligonal. Então as características atribuídas aos polígonos adjacentes são retornados como lista de características do setor que pertence aquele nó. Caso o nó esteja na fronteira entre dois ou mais setores, Cityhub nos pede para decidir qual a forma de agregação desses valores, no nosso caso escolhemos a média.

Incorporamos os dados de crimes e cada ponto de interesse usando a camadas de dados baseadas em pontos do Cityhub. Dado um canto do grafo da cidade e um raio fixo, pode-se procurar pontos que estejam dentro do disco com o raio determinado. Com essa abordagem conseguimos portanto para cada esquina determinar a quantidade de crimes e quais pontos de interesse ela contém, como apresentado na seção 2.3

Para a criação desse vetor precisamos definir qual o tamanho da vizinhança, qual será o raio que usaremos para afirmar que um ponto pertence a determinada esquina. Porém resolvemos deixar essa análise para o estudo de caso, onde mostraremos se esse raio tem alguma interferência nos resultados.Então para cada esquina e criamos o vetor de características $v_e \subset \mathbb{R}^{26}$, contendo os dados de crimes, do IBGE e POI, o vetor v_e está apresentado na tabela 3.2. Além disso a partir dele criamos mais dois vetores, um com as características relacionadas aos dados socioeconômicos chamado $v_s \subset \mathbb{R}^6$ mostrado na tabela 3.3 e outro vetor apenas com as características binárias $v_b \subset \mathbb{R}^{19}$ mostrado na tabela 3.4.

Tabela 3.2: Todas as características

Características	**Tipo**	**Fonte**
Crime	Inteiro	PM/AL
Renda Média por Domicílio	Ponto Flutuante	IBGE
Responsáveis Sem Renda	Ponto Flutuante	IBGE
Alfabetizados de 7 a 15 anos	Ponto Flutuante	IBGE
Menores de 18 anos	Ponto Flutuante	IBGE
Entre 18 a 65 anos	Ponto Flutuante	IBGE
Maior que 65 anos	Ponto Flutuante	IBGE
Escola	Binário	Google Map
Faculdade	Binário	Google Map
Universidade	Binário	Google Map
Teatro	Binário	Google Map
Cinema	Binário	Google Map
Museu	Binário	Google Map
Ginásio	Binário	Google Map
Estádio	Binário	Google Map
Hospital	Binário	Google Map
Emergência	Binário	Google Map
Batalhão	Binário	Google Map
Delegacia	Binário	Google Map
Bombeiro	Binário	Google Map
Hotel	Binário	Google Map
Bar	Binário	Google Map
Restaurante	Binário	Google Map
Posto de gasolina	Binário	Google Map
Transporte	Binário	Google Map
Agência Bancária	Binário	Google Map

Tabela 3.3: Características Socioeconômicas

Características	**Tipo**
Renda Média por Domicílio	Ponto Flutuante
Responsáveis Sem Renda	Ponto Flutuante
Alfabetizados de 7 a 15 anos	Ponto Flutuante
Menores de 18 anos	Ponto Flutuante
Entre 18 a 65 anos	Ponto Flutuante
Maior que 65 anos	Ponto Flutuante

Tabela 3.4: Características binárias

Características	**Tipo**
Escolas	Binário
Faculdade	Binário
Universidade	Binário
Teatro	Binário
Cinema	Binário
Museu	Binário
Hospital	Binário
Emergência	Binário
Batalhão	Binário
Delegacia	Binário
Bombeiro	Binário
Hotel	Binário
Bar	Binário
Restaurante	Binário
Posto de gasolina	Binário
Transporte	Binário
Ginásio	Binário
Estádio	Binário
Agência Bancária	Binário

3.3 Agrupamento

Para responder a pergunta de pesquisa 4: "analisar a relação entre os crimes e os dados socioeconômicos" iremos agrupar as esquinas por similaridade socioeconômica, usando apenas

os dados socioeconômicos, presentes no vetor v_s (Renda familiar, taxa de alfabetização,etc). Tal separação foi necessária, uma vez que quando colocamos todos os dados para realizar o agrupamento, os POI estavam interferindo negativamente na análise dos grupos, com a retirada deles, conseguimos agrupar de forma mais homogênea os dados. Logo, para os POI iremos trabalhar com eles na forma mostrada na seção seguinte.

Utilizamos dois tipos de agrupamento, o primeiro recebia como dados de entrada todas as esquinas da cidade. Já o segundo, levou em consideração apenas as esquinas hotspots , ou seja, apenas as esquinas com a quantidade de crimes superior ao limiar determinado. Desse modo, conseguimos agrupar os hotspots em grupos que apresentam características similares. Este limiar pode ser ajustado durante a execução para analisar as esquinas com maiores quantidades de crimes.

Um dos algoritmos utilizados para análise dos dados foi o *K-means* (MacQueen et al., 1967). A principal ideia do algoritmo é: dado um agrupamento inicial, modificar as posições de cada ponto para seu novo centro mais próximo, atualizar os centros de agrupamento calculando a média dos pontos membros e repetir o processo de realocação e atualização até que os critérios de convergência determinados inicialmente através de hiperparâmetros. O conjunto de entradas para o K-means será de acordo com o tipo de agrupamento escolhido, um conjunto de pontos, onde cada ponto representa uma esquina, apenas com os dados socioeconômicos.

Para o melhor desempenho do algoritmo, os dados foram normalizados usando a fórmula mínimo pelo máximo, como é mostrado na fórmula a seguir. A figura 3.4 mostra como ficou o agrupamento para a cidade de Maceió usando *Kmeans* , no mapa os pontos são as esquinas da cidade e as cores são os grupos encontrados pelo algoritmo.

$$X_c = \frac{X - X_{min}}{X_{max} - X_{min}}$$

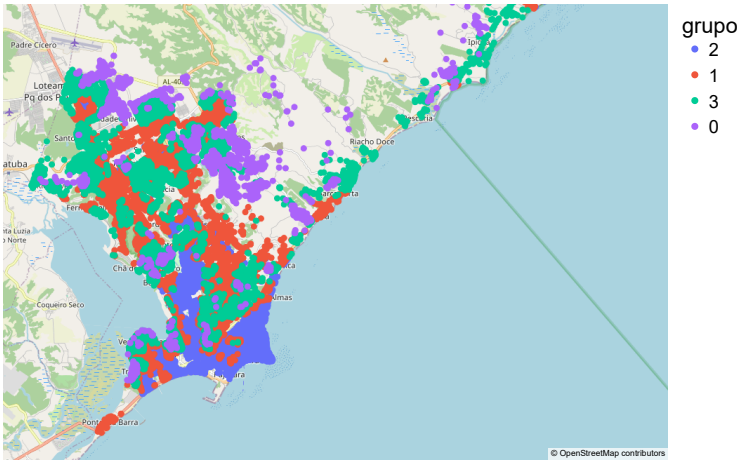

Figura 3.4: Agrupamento da cidade de Maceió usando *Kmeans* . Fonte: Autor

Além disso, utilizamos alguns outros algoritmos, um deles foi a rede neural artificial dos

mapas auto-organizáveis (em inglês: *self-organized maps* - SOM), proposto inicialmente por Teuvo Kohonen Kohonen (2004). Porém como o resultado dos agrupamentos eram similares, optamos por seguir as análises apenas com o agrupamento proviniente do algoritmo *Kmeans*. Os resultados das análises com esse algoritmo serão mostrados na seção seguinte, no estudo de caso.

Como forma de visualização das técnicas usamos o *Boxplot* (Ross, 2014), que é um diagrama de caixa, que representa a variação de dados observados através dos quartis, os limites da caixa correspondem ao 1º quartil e o 3º quartil, a linha interna corresponde ao 2º quartil ou mediana. Os valores atípicos ou outliers (valores discrepantes) podem ser plotados como pontos individuais. No nosso caso, utilizamos os dados já normalizados, montamos um conjunto de vários boxplots para cada grupo. Na figura 3.5 podemos ver visualizar o conjunto de boxplot para os grupos, cada boxplot representa uma variável socioeconômica.

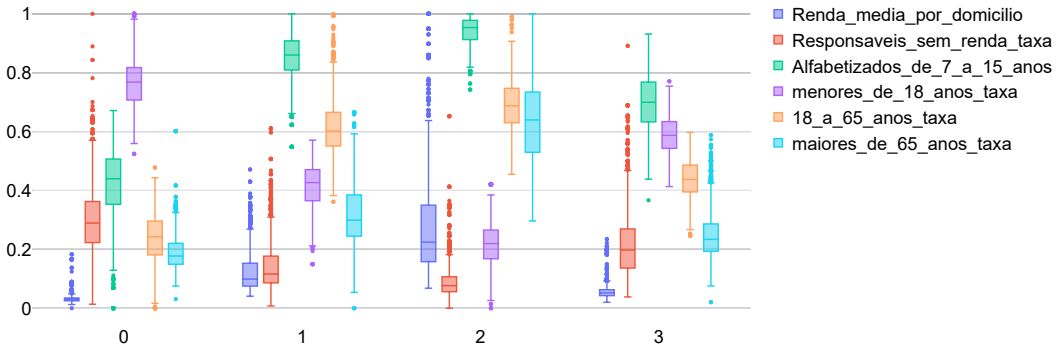

Figura 3.5: Boxplot da cidade de Maceió. Fonte: Autor

3.4 Análise estatística de dados

3.4.1 Definição de variáveis binárias

Para a utilização dos métodos estatísticos foi necessário criar duas classes de variáveis binárias, que nos auxiliarão nas análises futuras. A primeira classe foi crime, para cada esquina e a variável 1 terá o valor registrado como 1, se houver crime, caso não haja crime será registrado como 0. Em uma etapa posterior de análise, fizemos alguns testes usando um limiar para essa binarização dos crimes. Por exemplo, caso determinemos que o limiar é 3, então apenas as esquinas que tem a quantidade de crimes superior ou igual a de 3 serão contadas como 1 (presença de crime) e as demais como 0 (ausência de crimes). Então esta variável agora é definida como ter ocorrido uma quantidade maior ou igual do que o limiar L de crimes em uma dada esquina. Os resultados dessas análises serão mostrados na seção 4.

A segunda classe de variáveis foi POI, ponto de interesse. Para cada esquina e terá um conjunto de variáveis pertencentes ao vetor de características $v_b \subset \mathbb{R}^{19}$. Para estas variáveis

registramos com valor 1 caso a esquina esteja próxima (está a menos de um raio r) de um ponto de interesse, realizando esse processo para cada tipo de ponto de interesse, ou seja, pertencente ao vetor de características $v_b \subset \mathbb{R}^{19}$. Abaixo temos a formulação das variáveis, definidas para cada esquina do grafo de ruas da cidade.

Variável 1: ter ocorrido uma quantidade maior ou igual a L crimes numa dada esquina

Variável 2: a esquina é próxima de um ponto de interesse de um determinado tipo, pertencente ao $v_b \subset \mathbb{R}^{19}$

3.4.2 Teste Qui-quadrado

O teste Qui-quadrado (Pearson, 1900) é um teste de hipótese estatística normalmente usado quando os tamanhos das amostras são grandes. Resumidamente, esse teste é usado quando queremos analisar a relação entre duas variáveis binárias (duas variáveis categóricas), assim dizendo, ele é comumente usado para verificar a independência estatística entre essas variáveis.

O teste Qui-quadrado de Pearson é utilizado para determinar se existe diferença estatisticamente significativa entre as frequências esperadas e as frequências observadas em uma ou mais categorias de uma tabela de contingência. Pensando nisso, utilizamos ele para entender se existe uma correlação forte entre proximidade de um certo POI e a ocorrência de crimes. Para essa análise consideramos as variáveis criadas na seção anterior, variável de classe 1 que contém os crimes e o Conjunto de Variáveis de classe 2, pertencentes ao v_b

Os passos para a utilização do teste qui- quadrado para verificar a independência estatística são: Formulação das Hipóteses, Construção da Tabela de Contingência, Cálculo da Estatística do Qui-Quadrado, Determinação do Grau de Liberdade, Interpretação do Valor-p e Conclusão.

Para o nosso problema nós formulamos as seguintes hipóteses:

Hipótese nula (H0): A variável categórica 1 é independente da variável categórica 2 (POI).

Hipótese alternativa (H1): A variável categórica 1 não é independente da variável categórica 2 (POI).

Para ser possível implementar o teste Qui-quadrado é necessário a criação da tabela de contingência cruzando as duas variáveis categóricas. Esta tabela exibe a distribuição de frequência multivariada das variáveis 1 e 2. Ela fornece a inter-relação entre duas variáveis e nos ajuda a encontrar as interações entre elas. A tabela 3.5 mostra a estrutura da nossa tabela de contingência.

Tabela 3.5: Tabela de Contingência

	Crimes = 0	**Crimes = 1**
POI = 0	Count 00	Count 01
POI = 1	Count 10	Count 11

A estrutura básica é a seguinte: Count 00 = quantidade de esquinas que não são próximas de POI e não tem crime; Count 01 = quantidade de esquinas que não são próximas de POI e têm crime; Count 10 = quantidade de esquinas que são próximas de POI e não tem crime; Count 11 = quantidade de esquinas que são próximas de POI e tem crime; Foi gerada uma tabela dessa para cada ponto de interesse investigado.

Tendo calculado a tabela de contingência, passamos ao cálculo do teste Qui-quadrado de Pearson. Aplicamos então o seguinte cálculo:

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Tal Que, O_i são as frequências observadas em cada célula da tabela de contingência e E_i são as frequências esperadas em cada célula sob a hipótese nula de independência. Essas frequências esperadas são calculadas assumindo que as variáveis são independentes.

Próximo passa é determinar o grau de liberdade, este é calculado por: $(r-1)X(c-1)$ onde r é número de linhas na tabela de contingência e c é o número de colunas.

Como já mencionado, a hipótese nula, sugere que as duas variáveis observadas não possuem correlação, ou seja, que as observações ocorreram ao acaso. Para rejeitar essa hipótese usa-se o valor p, quanto menor o valor p ele indica que é improvável que a associação observada seja devida ao acaso. Em geral na literatura, perguntamos se p é menor que 0,05. Caso isso ocorra, não indica causalidade, mas conclui-se que há uma associação significativa entre as duas variáveis binárias. Então concluímos se a variável 2 é ou não correlacionada com a variável 1.

Dado os agrupamentos resultantes da seção 3.3, analisamos também nesta etapa a influência dos grupos nas métricas de ranqueamento. Então buscamos comparar quais pontos de interesse prevaleciam ou mudam dependendo das características da região, ou seja, dos grupos.

3.4.3 V de Cramér

No livro (CRAMÉR, 1999) está a apresentação do V de Cramér, esta medida juntamente com $\phi(Phi)$ mede o grau de associação entre dois campos categóricos. Ambas são medidas de associação entre duas variáveis nominais, o V de Cramér, também chamado de Índice de Cramér,

varia de 0 (correspondente a nenhuma associação entre as variáveis) a 1 (associação completa). Para uma tabela de contingência 2 X 2 a fórmula do V de Cramér é igual a do ϕ. que é:

$$\phi = \sqrt{\frac{X^2}{n}}$$

Sendo X^2 o valor do Qui-quadrado e n o número de esquinas. No nosso trabalho usamos o V de Cramér para determinar quão forte é a relação da ocorrência criminal com um determinado ponto de interesse. Logo, conseguimos montar um ranque dos POI's com maior correlação até os de menor relação com os crimes.

3.4.4 Probabilidade Condicional

A Probabilidade Condicional $P(A|B)$ (Ackerman, Freer e Roy, 2019) é a probabilidade de um evento A acontecer dado que B aconteceu. Isso ocorre porque o evento A é uma das possibilidades de B, ou seja, os dois eventos precisam ter o mesmo espaço amostral e B precisa ser possível. A probabilidade é dada pela fórmula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Na fórmula a seguir temos a simplificação da fórmula da probabilidade condicional, tal que o n é quantidade de esquinas observadas. Portanto, no final, o cálculo feito foi a quantidade de esquinas que apresentam A e B dividido pela quantidade de esquinas que apresentaram B.

$$P(A|B) = \frac{\frac{n(A \cap B)}{n(\Omega)}}{\frac{n(B)}{n(\Omega)}}$$

$$P(A|B) = \frac{n(A \cap B)}{n(B)}$$

Em uma etapa posterior de análise, fizemos alguns testes para calibrar o limiar L que, como já citado em outras seções, diz a partir de qual valor contabilizamos a ocorrência de crimes. No caso deste trabalho, fizemos a análise separadamente para cada POI, ponto de interesse. Calculamos qual a probabilidade de ter ocorrido crime numa esquina, dado que ela está a R metros de um certo tipo de determinado POI. As variáveis consideradas são:

A: ter ocorrido uma quantidade maior ou igual à L de crimes numa dada esquina. Equivalente a variável 1 apresentada na seção

B: a esquina é próxima de um ponto de interesse de um determinado tipo, pertencente ao $v_b \subset \mathbb{R}^{19}$, Equivalente a variável 2 apresentada na seção

Conseguimos uma tabela com uma probabilidade condicional para cada tipo de POI. Então temos que $P(A)$ é a probabilidade de A acontecer, ou seja, dentre as esquinas escolher uma esquina que contém crime. A probabilidade $P(B)$ é a probabilidade de escolhendo uma esquina, esta ser uma esquina próxima de determinado ponto de interesse. $P(A \cap B)$ refere-se a probabilidade de que uma esquina tenha crime e esteja próxima de um ponto de interesse. Por fim obtendo $P(A|B)$ temos a probabilidade de ocorrer crime dado que estejamos próximos de um ponto de interesse. Essa probabilidade é calculada para todos os pontos de interesse presentes no vetor $v_b \subset \mathbb{R}^{19}$. Depois de obtermos essas probabilidades usaremos elas para ranquear os pontos de interesse por ordem de maior probabilidade.

Capítulo 4

Estudo de Caso

Com o objetivo de responder às perguntas de pesquisa apresentadas na Seção 1.2.1, realizamos um estudo de caso na cidade de Maceió, no estado de Alagoas usando a base de dados descrita na Seção 3.1, incluindo dados georreferenciados sobre o registro dos crimes, informações socioeconômicas e pontos de interesse.

A fim de responder à pergunta 1, na Seção 4.1 realizamos um agrupamento das esquinas levando em consideração apenas as características socioeconômicas. Nesse ponto, consideramos duas variações: a primeira agrupando todas as esquinas da cidade, e a segunda agrupando apenas as esquinas com ocorrência criminal.

Com o objetivo de responder às perguntas 2 e 3, na Seção 4.2 descrevemos análises estatísticas realizadas usando os dados dos pontos de interesse. Aplicamos as métricas do Teste Qui-quadrado, V de Cramér e probabilidade condicional. Nesta mesma seção, também respondemos à pergunta de pesquisa 4, através de um ordenamento dos pontos de interesse de acordo com métricas de correlação com a ocorrência criminal.

Para finalizar, na seção 4.3 analisamos se a influência dos pontos de interesse muda de acordo com as características socioeconômicas das esquinas (pergunta 5). Com esse objetivo, aplicamos os testes estatísticos para cada um dos grupos encontrados na seção.

4.1 Características Socioeconômicas

Nesta seção abordamos à pergunta 1 que questiona se é possível separar as esquinas com atividade criminal em grupos com características socioeconômicas distintas. Iniciamos com o algoritmo *K-means*, que requer como entrada as características a serem avaliadas e quantos grupos deveriam ser formados. Passamos como entrada as características relacionadas aos dados socioeconômicos (vetor de características v_s, conforme definido na seção 3.2). Quanto ao número de grupos, a quantidade de grupos escolhida foi quatro, após um estudo empírico

usando visualização de boxplots, procurando a maior quantidade de grupos, mas que não fossem tão similares. Como resultado temos a figura 4.1 , nela apresentamos todas as esquinas da cidade de Maceió. Abaixo apresentamos o gráfico *boxplot* 4.2, que nos ajudou a entender como que o algoritmo agrupou as esquinas.

Notamos que o grupo 2, dentre os outros grupos, é aquele que contém os valores de renda mais altos e em sua maioria idosos, o que condiz com a realidade conhecida da região, pois os pontos marcados como pertencentes ao grupo 2 estão na orla da cidade e ao redor da avenida principal da cidade, ou seja, regiões mais ricas. No grupo 1 observamos a renda média por domicílio com valores medianos, e a taxa de responsáveis sem renda não muito altos, além disso, tem uma presença maior de pessoas maiores de 18 anos.

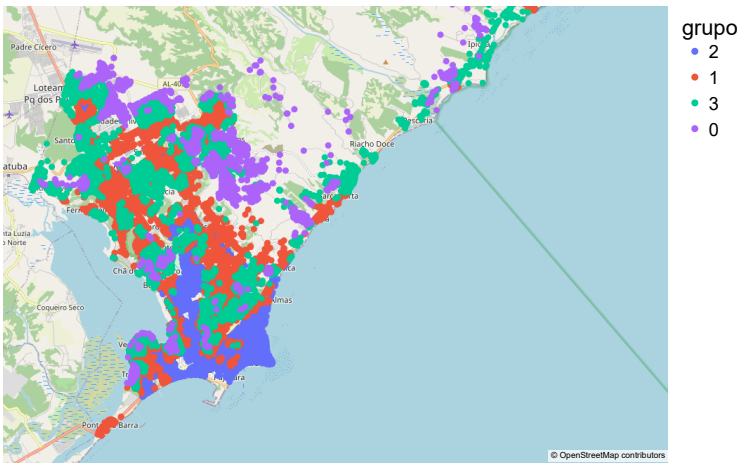

Figura 4.1: Agrupamento todas as esquinas de Maceió usando *Kmeans* . Fonte: Autor

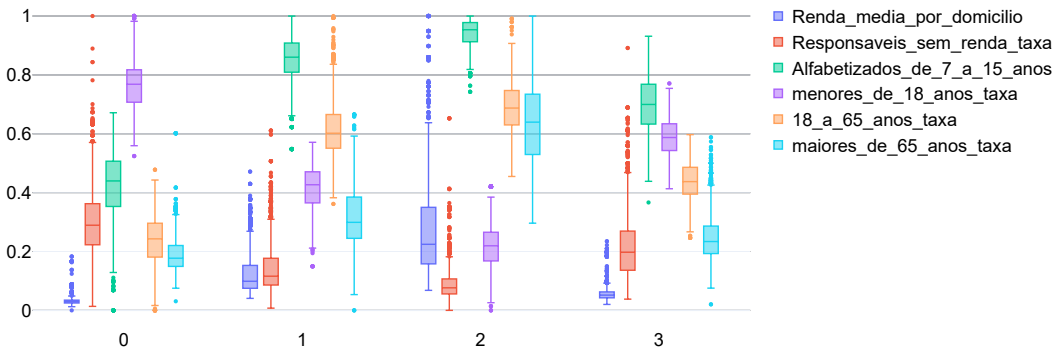

Figura 4.2: Boxplot do agrupamento anterior. Fonte: Autor

Como percebemos nas imagens, notamos que o grupo 3 e 0 apresentam os menores valores de renda e a menor variação (caixa do boxplot pequena), o que indica grande quantidade de rendas médias por domicílio registradas como baixas. Além disso, a taxa de pessoas sem renda e a taxa de menores de idade é alta nas esquinas destes dois grupos. Quando olhamos para o mapa percebemos que as esquinas que estão nesses grupos estão situadas nas periferias da cidade, caracterizando regiões da cidade mais pobres. Os fatores que diferenciam os grupos 3

e 0 são: taxa de menores de 18 anos e alfabetização. O grupo 3 detém um grande número de menores de idade e observamos uma queda na alfabetização com relação aos grupos 1 e 2. Já o grupo 0 apresenta a maior taxa de menores de idade e a alfabetização é a mais baixa entre todos os grupos, mostrando a necessidade de se ampliar a educação nessa região.

A próxima análise realizada foi restrita às esquinas de cidade com ocorrência criminal relevante. Com este objetivo, as esquinas foram filtradas com um limiar para a quantidade de crimes. Em outras palavras, dado um limiar L, consideramos apenas aquelas esquinas que têm a quantidade de crimes maior ou igual do que o limiar. A Figura 4.3 revela a quantidade de esquinas com quantidades de crime entre 0 e 12.

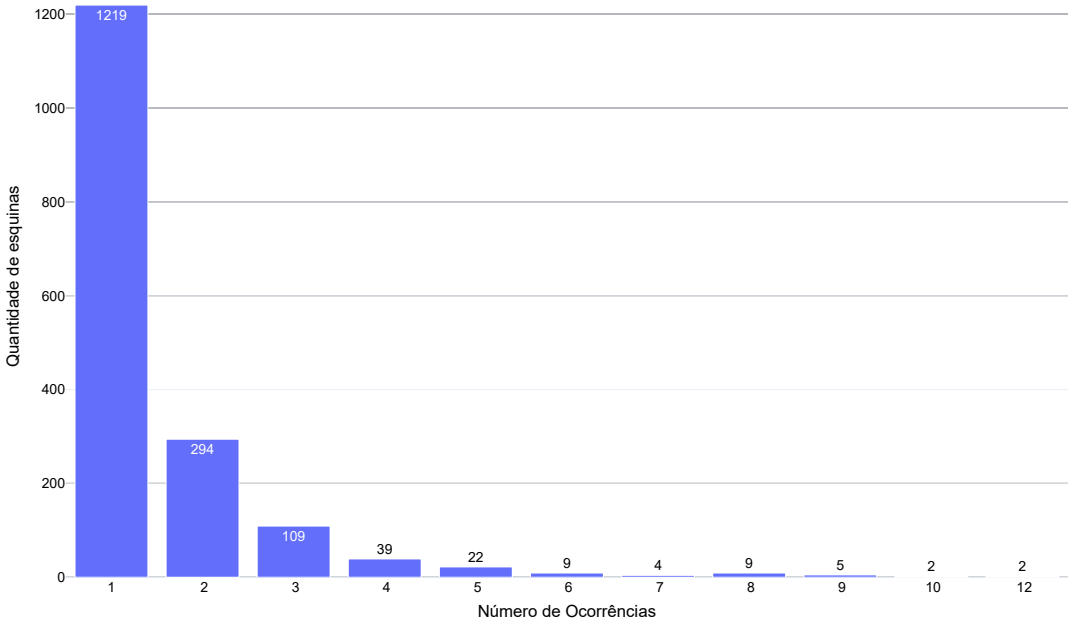

Figura 4.3: Quantidade de esquinas com crimes. Fonte: Autor

Nesse experimento, usamos o limiar igual a 1, então para o agrupamento foram consideradas apenas as esquinas onde havia ocorrido crime. O segundo parâmetro necessário para a execução do *Kmeans* é a definição do número de grupos, no nosso caso usaremos o número quatro, o mesmo número de grupos da etapa anterior. A Figuras 4.4 e a Figura 4.5 apresentam, respectivamente, a localização das esquinas do novo agrupamento e os boxplots destes grupos. Observando os dois mapas percebemos que no novo agrupamento a disposição das esquinas e grupos é semelhante ao agrupamento anterior.

Além disso, os boxplots do agrupamento de todas as esquinas e das esquinas restritas apresentam semelhanças ou então mudanças muito sutis em: renda média, alfabetização e taxa dos maiores de 65 anos. A mudança mais visível é na taxa de responsáveis sem renda que apresentou uma elevação na posição dos boxplots, o que indica aumento dos valores dessa taxa. A taxa de menores de idade teve uma leve descida em alguns grupos. Por fim, na característica

que mostra a taxa de pessoas entre 18 e 65 anos as caixas do boxplot se tornaram menores e posicionadas na parte superior, o que indica que, no agrupamento feito apenas com crimes, as esquinas apresentaram maior concentração nos valores mais altos dessa característica. Dessa forma, podemos responder à pergunta de pesquisa 1, concluindo que existem diferentes perfis socioeconômicos de esquinas com atividade criminal.

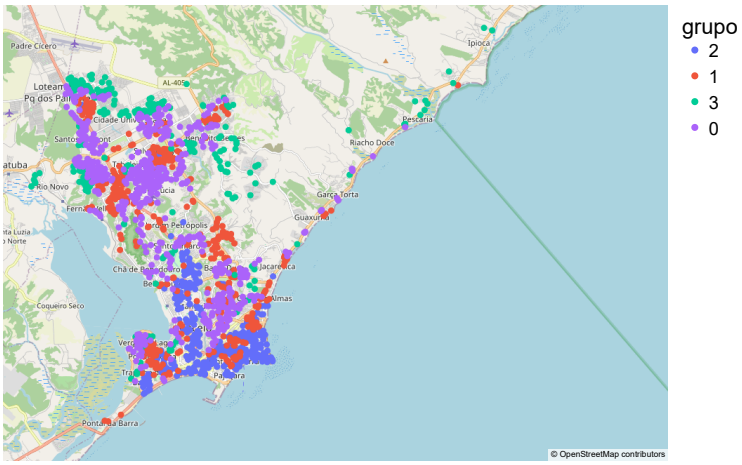

Figura 4.4: Agrupamento com limiar 1 usando *Kmeans* . Fonte: Autor

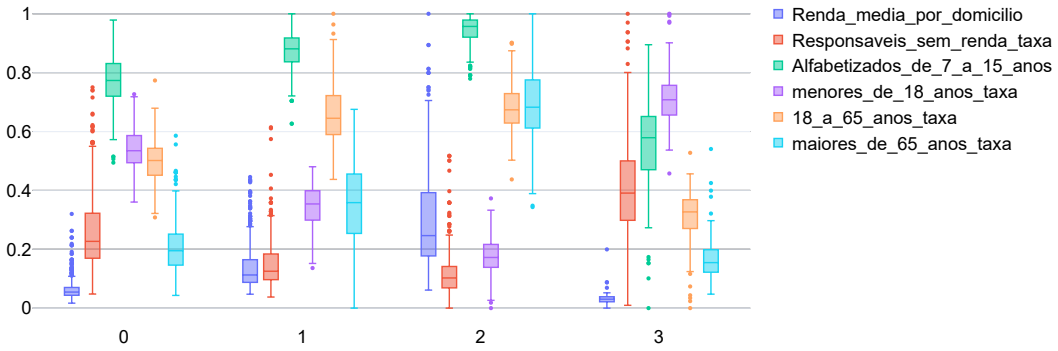

Figura 4.5: Boxplot do agrupamento anterior. Fonte: Autor

4.2 Análises de pontos de interesse (POIs)

Com objetivo de responder às perguntas 2 e 3, nesta seção iremos analisar a relação dos eventos criminais e características relacionadas aos pontos de interesse. As métricas estatísticas utilizadas foram: Qui-quadrado, V de Cramér e a Probabilidade Condicional. Apresentaremos o ranqueamento das características de acordo com cada uma destas métricas nas subseções seguintes para responder também à pergunta de pesquisa 4.

Porém, para melhor entendimento dos resultados, optamos por agrupar alguns pontos de interesse. Acreditamos que a influência desses pontos é similar e por isso podem ser agrupados em

categorias. Na Tabela 4.1 apresentamos um novo vetor de características binárias categorizadas $v_{bc} \subset \mathbb{R}^{12}$.

Tabela 4.1: Características Binárias Categorizadas

Categoria	**Características**
Ensino	Escolas Faculdade Universidade
Entretenimento	Teatro Cinema Museu
Saúde	Hospital Emergência
Segurança	Batalhão Delegacia Bombeiro
Hotel	Hotel
Bar	Bar
Restaurante	Restaurante
Posto de gasolina	Posto de gasolina
Transporte	Transporte
Ginásio	Ginásio
Estádio	Estádio
Agência Bancária	Agência Bancária

4.2.1 Teste Qui-quadrado e Cramér

Iniciamos com o Qui-quadrado, conforme descrito na Subseção 3.4.2. Utilizamos este teste para descobrir se existe diferença estatisticamente significativa entre as frequências esperadas e as frequências observadas entre a proximidade de um certo POI e a ocorrência de crimes. Juntamente com ele aplicamos o V de Cramér para conseguirmos saber a força da correlação entre as variáveis.

Como forma de estabelecer uma base do comportamento da métrica, fizemos um teste com os valores aleatórios. Uma vez que a presença de um POI na proximidade de uma esquina é dado por 0 (não está próximo) ou 1 (está próximo), no teste aleatório consideramos que para cada esquina o valor da variável que descreve a proximidade da esquina a um certo POI seria uma escolha aleatória entre 0 ou 1 com probabilidade de 50% para cada. O resultado encontra-se na

tabela 4.2, como esperado não é possível rejeitar a hipótese nula, que sugere que as observações ocorreram ao acaso, ou seja, a variável crime não apresenta relação com nenhum POI. Na tabela 4.2, a coluna correlação só marcará como 'Sim' quando o valor do Qui-quadrado for menor que 0,05 valor comumente referenciado na literatura.

Tabela 4.2: Aplicação do Qui-quadrado com valores aleatórios

Características	P Valor	Correlação	Cramér
Agência bancária	1.95×10^{-1}	Não	0.0104
Estádio	1.96×10^{-1}	Não	0.0103
Bar	2.11×10^{-1}	Não	0.0100
Entretenimento	2.61×10^{-1}	Não	0.0090
Hotel	2.74×10^{-1}	Não	0.0087
Transporte	4.25×10^{-1}	Não	0.0064
Saúde	5.35×10^{-1}	Não	0.0049
Ensino	5.52×10^{-1}	Não	0.0047
Restaurante	6.39×10^{-1}	Não	0.0037
Ginásio	6.74×10^{-1}	Não	0.0033
Seguranca	8.36×10^{-1}	Não	0.0016
Posto de gasolina	8.64×10^{-1}	Não	0.0013

Em seguida, realizamos o teste usando os valores binários verdadeiros das variáveis que descrevem a proximidade da esquina a cada grupo de POI. Foram calculados então o valor P do Teste Qui-quadrado e o V de Cramér, considerando todas as esquinas da cidade.

Importante lembrar que, na Seção 3.2, quando criamos a representação computacional dos dados, foi necessário definir qual o raio usado para afirmar que um ponto de interesse estava próximo de uma determinada esquina. Para avaliar a influência do raio nas análises, realizamos aqui testes com três valores de raio distintos: 100 metros (Tabela 4.3), 200 metros (Tabela 4.4) e 300 metros (Tabela 4.5), destacamos que essas tabelas são ordenadas pela coluna V de Cramér, em ordem decrescente.

Percebemos que, levando em consideração o P valor, praticamente todas as características tiveram correlação com o crime, exceto a característica estádio que, com um raio de 100 metros, não apresentou correlação. Analisando agora o V de Cramér, ele aponta a força da correlação entre as características, percebemos que a característica ensino se manteve no topo em todos os três resultados e estádio sempre com a menor correlação. Algumas mudanças ocorreram do raio 100 para o raio 200, agência bancária e transporte desceram no ranqueamento, ao mesmo tempo que ginásio e saúde subiram. Entre o raio 200 e 300 as trocas de posições foram suaves e praticamente irrelevantes.

Concluímos que a medida do raio de 200 metros, seria o intermediário entre as mudanças observadas entre 100 e 300 metros. Além disso, tomamos como base os trabalhos CityHub (Salinas et al., 2022) e TensorAnalyzer (Silveira et al., 2022) que usaram 200 metros como

Tabela 4.3: Aplicação do Qui-quadrado com $r = 100m$.

Características	P Valor	Correlação	Cramér
Agência bancária	2.07×10^{-13}	Sim	0.0589
Ensino	5.31×10^{-13}	Sim	0.0579
Transporte	8.01×10^{-11}	Sim	0.0521
Bar	1.06×10^{-10}	Sim	0.0518
Posto de gasolina	1.44×10^{-10}	Sim	0.0514
Saúde	5.21×10^{-9}	Sim	0.0468
Restaurante	6.77×10^{-9}	Sim	0.0465
Hotel	1.03×10^{-6}	Sim	0.0392
Seguranca	1.10×10^{-6}	Sim	0.0391
Ginásio	1.78×10^{-5}	Sim	0.0344
Entretenimento	4.74×10^{-4}	Sim	0.0280
Estádio	2.04×10^{-1}	Não	0.0102

Tabela 4.4: Aplicação do Qui-quadrado com $r = 200m$.

Características	P Valor	Correlação	Cramér
Ensino	7.69×10^{-19}	Sim	0.0711
Ginásio	2.08×10^{-18}	Sim	0.0702
Saúde	8.93×10^{-16}	Sim	0.0645
Posto de gasolina	1.72×10^{-14}	Sim	0.0615
Transporte	5.32×10^{-14}	Sim	0.0604
Agência bancária	8.76×10^{-14}	Sim	0.0598
Restaurante	6.32×10^{-13}	Sim	0.0577
Bar	1.69×10^{-11}	Sim	0.0540
Entretenimento	1.13×10^{-10}	Sim	0.0517
Hotel	2.22×10^{-8}	Sim	0.0449
Seguranca	1.30×10^{-7}	Sim	0.0423
Estádio	4.89×10^{-3}	Sim	0.0225

métrica e Mirante (Zanabria et al., 2020) que usou 300 metros, visto que usar 300 metros é similar a usar 200. As próximas análises serão aplicadas usando os vetores de características formados usando o raio de 200 metros.

Olhando portanto para o ranqueamento exibido na Tabela 4.4, chegamos à conclusão de que existem três tipos de ponto de interesse com correlação mais forte com a ocorrência de crimes: instituições de ensino, ginásios e instituições de saúde.

4.2.2 Probabilidade Condicional

Nesta seção ranquearemos os tipos de ponto de interesse de acordo com probabilidades condicionais, conforme descrito na Seção 3.4.4. Neste caso, um tipo de POI será bem ranqueado

Tabela 4.5: Aplicação do Qui-quadrado com $r = 300m$.

Características	P Valor	Correlação	Cramér
Ensino	2.36×10^{-26}	Sim	0.0852
Ginásio	6.15×10^{-25}	Sim	0.0827
Saúde	4.04×10^{-21}	Sim	0.0757
Posto de gasolina	5.73×10^{-21}	Sim	0.0754
Restaurante	6.14×10^{-18}	Sim	0.0692
Agência bancária	1.50×10^{-16}	Sim	0.0662
Bar	2.14×10^{-15}	Sim	0.0636
Transporte	1.55×10^{-14}	Sim	0.0616
Entretenimento	1.49×10^{-13}	Sim	0.0593
Seguranca	7.67×10^{-10}	Sim	0.0493
Hotel	2.19×10^{-9}	Sim	0.0480
Estádio	4.37×10^{-3}	Sim	0.0228

se a presença de pelo menos uma instância de POI daquele tipo na vizinha de uma dada esquina (a menos de um raio de r metros) implicar em alta probabilidade de ocorrência de crimes naquela esquina. Busca-se, portanto, responder à pergunta de pesquisa 4.

Assim como no teste Qui-quadrado, experimentamos 3 valores de raio distintos para analisar qual o mais adequado ao problema. As tabelas com as probabilidades condicionais para cada ponto de interesse estão apresentadas para valores de raio de 100 metros (Tabela 4.6), 200 metros (Tabela 4.7) e 300 metros (Tabela 4.8).

Tabela 4.6: Ranqueamento de tipos de ponto de interesse usando probabilidade condicional com $r = 100m$.

Características	Prob.
Agência bancária	0.32
segurança	0.19
Estádio	0.18
Hotel	0.18
Posto de gasolina	0.18
Entretenimento	0.18
Saúde	0.16
Ginásio	0.15
Bar	0.15
Transporte	0.15
Ensino	0.15
Restaurante	0.14

Algumas conclusões podem ser extraídas observando as três tabelas. A primeira é que agência bancária sempre se apresenta no topo da tabela, com maior probabilidade condicional, independente do raio. O segundo e terceiro lugar no ranque se alternam entre: segurança, estádio,

Tabela 4.7: Ranqueamento de tipos de ponto de interesse usando probabilidade condicional com $r = 200m$.

Características	Prob.
Agência bancária	0.23
Estádio	0.18
Entretenimento	0.18
Ginásio	0.16
Hotel	0.16
Posto de gasolina	0.16
Seguranca	0.16
Saúde	0.15
Restaurante	0.13
Bar	0.13
Transporte	0.13
Ensino	0.13

Tabela 4.8: Ranqueamento de tipos de ponto de interesse usando probabilidade condicional com $r = 300m$.

Características	Prob.
Agência bancária	0.20
Entretenimento	0.17
Ginásio	0.15
Estádio	0.15
Hotel	0.15
Posto de gasolina	0.15
Seguranca	0.15
Saúde	0.14
Bar	0.13
Ensino	0.13
Restaurante	0.12
Transporte	0.12

Entretenimento e ginásio. Entre estas tabelas não tivemos mudanças significativas nas posições em que o atributo aparece no ranque. Assim como na seção anterior, concluímos que seria melhor usar o raio $200m$, visto que ele é a ponte entre as duas possibilidades de raios e é usado na literatura. Notamos que essa informação pode ser usada para o planejamento do patrulhamento nas regiões de maior interesse para a diminuição de crimes.

Outro hiperparâmetro relevante neste problema é o limiar para binarização da variável de ocorrência de crimes na esquina. Realizamos experimentos variando este limiar entre 1 e 6 para analisar o que ocorre com as probabilidades condicionais. Neste caso, denotaremos P_i como a probabilidade condicional com limiar de binarização igual a i.

Na Tabela 4.9, exibimos os resultados deste experimento para um raio de 200m, onde as linhas representam os pontos de interesse e as colunas representam a probabilidade de ocorrer um crime para um dado limiar de binarização dos crimes. Por exemplo, na coluna P_4 o limiar é 4, então apenas as esquinas que têm pelo menos 4 crimes serão contadas como 1 (presença de crime), as demais como 0 (ausência de crimes). Importante ressaltar que essa visualização não está ordenada pelas maiores probabilidades, uma vez que as ordenações são distintas para cada coluna. Nessa análise percebemos que as maiores probabilidades sempre são relacionadas à proximidade de agências bancárias.

Tabela 4.9: Probabilidade condicional para todas as esquinas com diferentes limiares. Fonte: Autor

Características	P_1	P_2	P_3	P_4	P_5
Hipótese Nula	0.110	0.031	0.012	0.005	0.003
Agência bancária	0.229	0.083	0.031	0.015	0.010
Ginásio	0.157	0.046	0.018	0.008	0.006
Estádio	0.176	0.051	0.020	0.010	0.005
Hotel	0.157	0.053	0.019	0.007	0.004
Restaurante	0.128	0.038	0.015	0.007	0.004
Bar	0.132	0.039	0.016	0.007	0.004
Posto de gasolina	0.159	0.047	0.017	0.008	0.005
Transporte	0.131	0.038	0.015	0.007	0.004
Ensino	0.132	0.038	0.016	0.006	0.003
Saúde	0.148	0.045	0.020	0.007	0.002
Seguranca	0.156	0.050	0.023	0.010	0.007
Entretenimento	0.179	0.070	0.029	0.013	0.006

Para enriquecer esse estudo, calculamos também uma hipótese nula, ou seja, a probabilidade de ocorrer crime em uma esquina qualquer, independente da proximidade de POIs. Nesse caso o cálculo foi a probabilidade de uma esquina qualquer satisfazer o limiar de binarização de crimes, independente de proximidade de POIs. Mais especificamente, essa probabilidade pode ser calculada como a quantidade de esquinas que satisfaz o limiar dividido pela quantidade total de esquinas. O valor da chamada hipótese nula para a primeira coluna foi de 0.11, como observado na Tabela 4.9, onde o limiar para crimes é 1.

Importante ressaltar que nas colunas da tabela, o fato de ter ocorrido crime muda de acordo com um limiar, portanto ao longo das colunas, a probabilidade nula muda, já que o número de esquinas com crimes muda. Praticamente todos os valores das probabilidades condicionais para cada ponto de interesse estão acima dessa probabilidade. A única exceção foi saúde na colunas 'P_5'. Por tanto todos os POIs possuem algum tipo de relação com os crimes, conforme aumentamos o número de crimes essa relação vai diminuindo.

A figura 4.6 exibe uma visualização da evolução das probabilidades condicionais quando

o limiar de binarização aumenta, onde cada curva representa um tipo de POI. Vale ressaltar que o eixo vertical está em escala logarítmica.

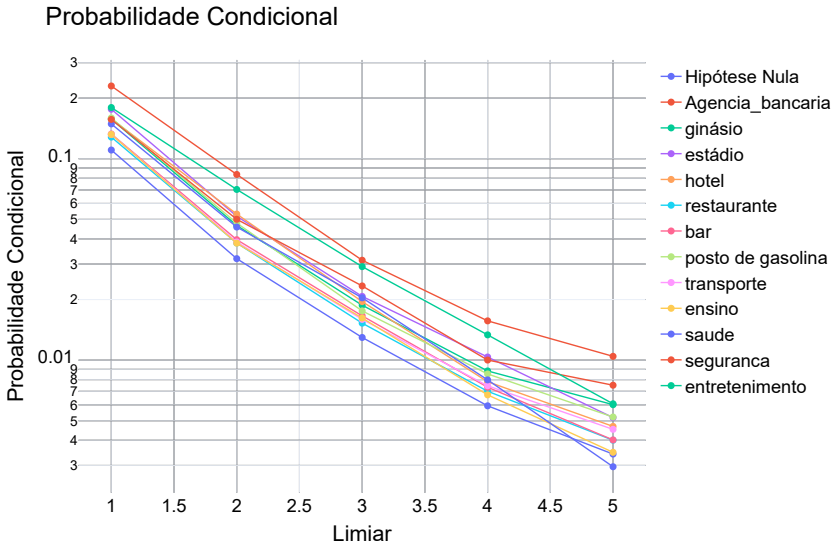

Figura 4.6: Gráfico para a probabilidade condicional para todas as esquinas com diferentes limiares. Fonte: Autor

4.3 POI e Agrupamentos

Essa seção tem como intuito responder à pergunta 5 que questiona se a influência dos pontos de interesse muda de acordo com características socioeconômicas das esquinas, ou seja, se cada grupo possui um comportamento diferente. Foram aplicadas a análise da Probabilidade Condicional nos grupos de toda a cidade.

4.3.1 Probabilidade nos grupos de esquina

Nesta seção aplicamos a probabilidade condicional para cada grupo obtido nos experimentos da Seção 4.1 separadamente. Portanto, todas as probabilidades condicionais e a hipótese nula foram calculadas separadamente para cada grupo, considerando espaços amostrais diferentes. Para esse experimento, usamos o agrupamento que envolvia todas as esquinas de Maceió, não apenas as que tinham ocorrência criminal. Porém, um cuidado deve ser tomado, na abordagem por grupos, uma vez que as métricas podem não funcionar muito bem quando existem valores muito baixos da ocorrência dos pontos calculados.

Nas Tabelas 4.10, 4.11, 4.12 e 4.13 apresentamos os resultados dessa análise. Sabendo que em cada tabela os valores estão ordenados, notamos que na maioria dos casos os POI estavam acima da hipótese nula.

Tabela 4.10: Aplicação da Probabilidade Condicional apenas para as esquinas do grupo 0

Características	Pr. Condicional	Quantidade
Posto de gasolina	0.083	12
Saúde	0.064	234
Bar	0.063	442
Ensino	0.061	623
Transporte	0.051	758
Restaurante	0.049	712
Hotel	0.048	21
Hipotese Nula	0.043	1926
Ginásio	0.011	89
Agência bancária	0.000	0
Estádio	0.000	4
Seguranca	0.000	71
Entretenimento	0.000	6

Tabela 4.11: Aplicação da Probabilidade Condicional apenas para as esquinas do grupo 1.

Características	Pr. Condicional	Quantidade
Agência bancária	0.223	94
Estádio	0.171	82
Ginásio	0.170	1185
Seguranca	0.162	382
Posto de gasolina	0.150	908
Saúde	0.148	1327
Entretenimento	0.145	179
Hotel	0.144	409
Transporte	0.142	2707
Bar	0.140	2116
Ensino	0.137	3135
Restaurante	0.134	3113
Hipotese Nula	0.120	5326

No grupo 0, ginásio, agência bancária, estádio, segurança e entretenimento estão abaixo da hipótese nula. As probabilidades mais altas foram posto de gasolina e saúde. Lembramos que nas características socioeconômicas esse grupo tinha uma das menores rendas e a mais baixa taxa de alfabetização. Mostrando que localidades com renda muito baixa, quando existem postos de gasolina, podem ser locais perigosos.

Já no grupo 1, todos os pontos ficaram acima da hipótese nula, na probabilidade condicional, os três com a maior probabilidade foram agência bancária, estádio e ginásio. Quanto às características, a renda média por domicílio apresenta valores mais medianos, e tem uma presença maior de pessoas maiores de 18 anos, notamos que esse grupo tem mais relação com o esporte apresentado como primeiros colocados estádio e ginásio.

Tabela 4.12: Aplicação da Probabilidade Condicional apenas para as esquinas do grupo 2.

Características	Pr. Condicional	Quantidade
Estádio	0.243	37
Agência bancária	0.233	266
Entretenimento	0.208	472
Hotel	0.197	660
Seguranca	0.186	576
Saúde	0.186	1040
Bar	0.177	1082
Transporte	0.177	1479
Posto de gasolina	0.177	864
Restaurante	0.176	1565
Ginásio	0.171	938
Ensino	0.169	1715
Hipotese Nula	0.163	2843

Tabela 4.13: Aplicação da Probabilidade Condicional apenas para as esquinas do grupo 3.

Características	Pr. Condicional	Quantidade
Agência bancária	0.217	23
Estádio	0.157	70
Posto de gasolina	0.142	324
Entretenimento	0.140	143
Ginásio	0.135	617
Saúde	0.126	783
Ensino	0.119	2561
Bar	0.117	2087
Restaurante	0.114	2604
Transporte	0.114	1896
Seguranca	0.112	170
Hipotese Nula	0.098	5420
Hotel	0.063	190

No grupo 2, a probabilidade condicional apontou como pontos relevantes estádio, agência bancária e entretenimento, neste grupo as esquinas contém os valores de renda mais altos e em sua maioria idosos. Ou seja, em regiões mais ricas da cidade a atenção deve ser focada nesses pontos, o que é preocupante uma vez que nos bairros ricos da cidade de Maceió temos muitos pontos turísticos e agências bancárias.

Na Figura 4.7 observamos a disposição das ocorrências do POI estádio e dos crimes, apenas nas esquinas do grupo 2. Destacamos três regiões com base na ocorrência de estádios, a fim de melhorar a visualização. Notamos que sempre que existe a ocorrência de 'Apenas estádio' (ponto azul), nas proximidades existem pontos com crimes(verdes) e até mesmo de forma simultânea (vermelho), ou seja, crimes e estádio no mesmo ponto. Mostrando assim de

forma mais clara a forte relação entre a ocorrência de crimes na vizinhança de estádios.

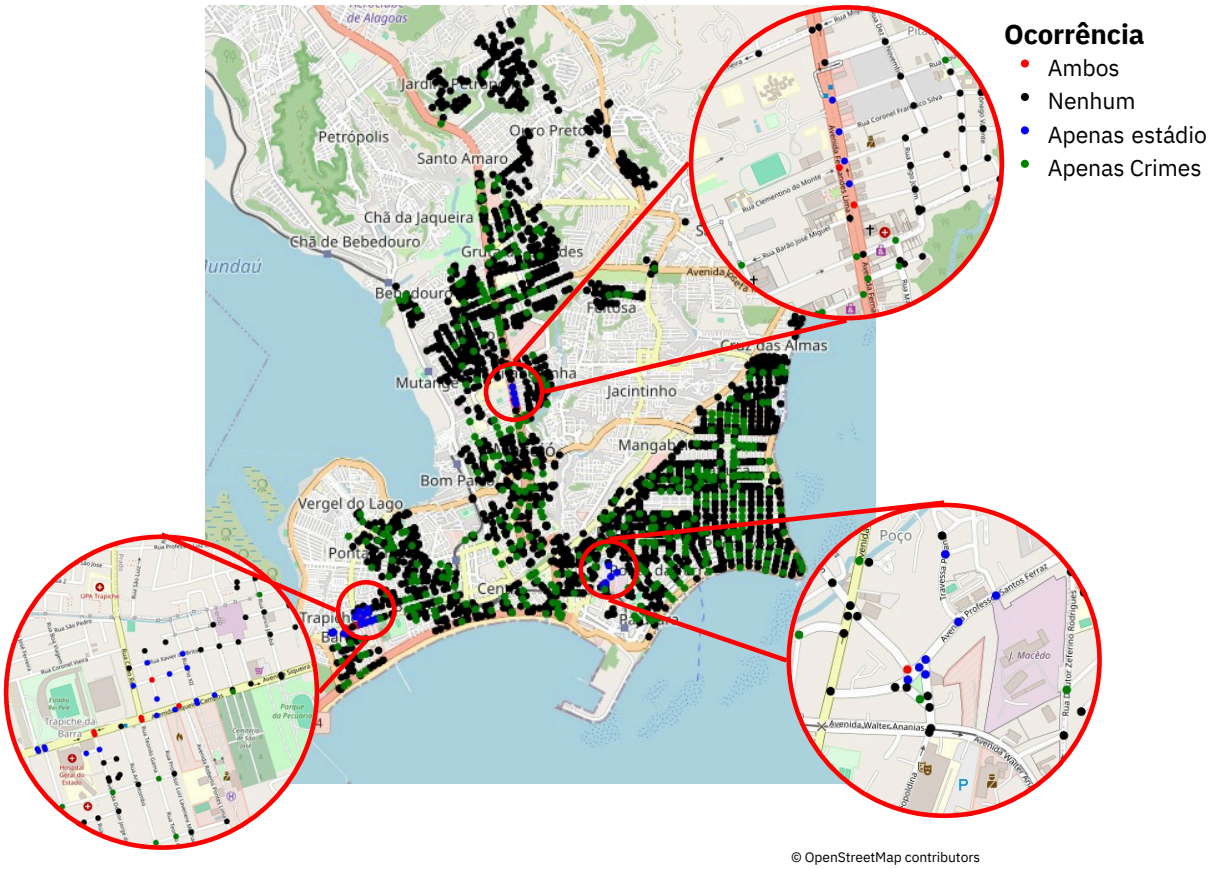

Figura 4.7: Esquinas da cidade pertencentes ao Grupo 2 junto com o POI estádio e crimes. Fonte: Autor

Por fim, o grupo 3 apresentou probabilidade alta para agência bancária, estádio e posto de gasolina, além disso hotel está abaixo da hipótese nula. Este grupo está dentre os mais pobres, e apresenta características socioeconômicas parecidas com grupo 0, diferindo apenas em alfabetização e taxa de menores de idade. A diferença agora é mais notória graças aos pontos destacados na probabilidade que são pontos diferentes para os grupos.

Percebemos que, ao mudar o grupo, ou seja, mudar as características socioeconômicas, ocorre uma mudança também no ranque dos pontos mais relevantes para a ocorrência criminal. Como forma de visualizar as probabilidades para cada grupo, criamos por fim o gráfico na Figura 4.8 , no eixo horizontal tempos os POI's separados por grupos, no eixo vertical encontra-se a probabilidade condicional.

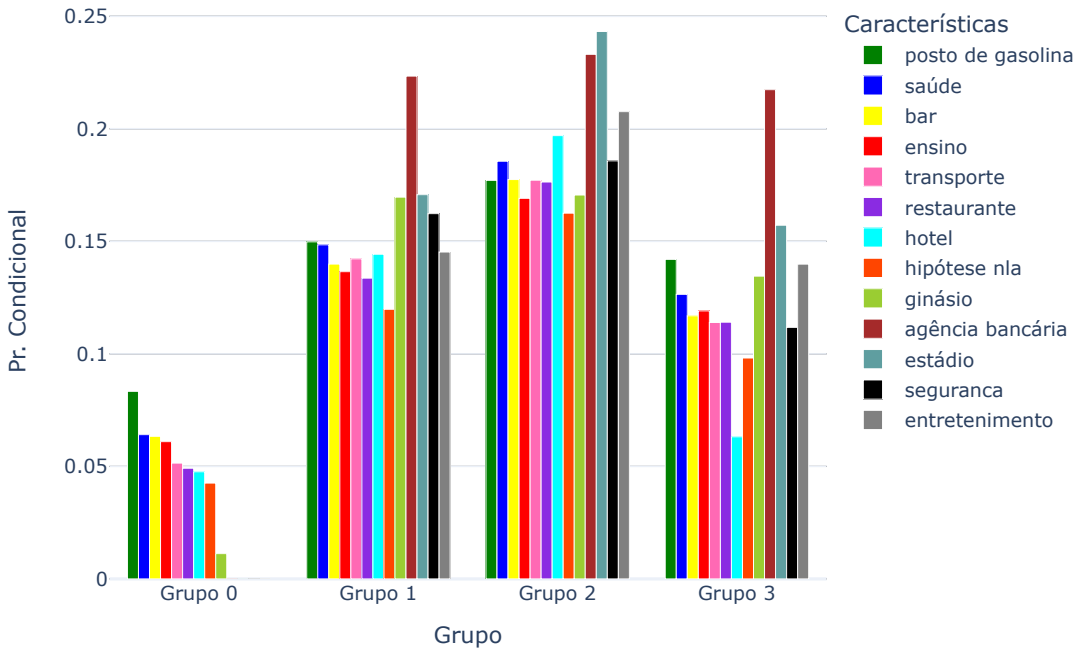

Figura 4.8: Gráfico para a probabilidade condicional para os grupos. Fonte: Autor

Capítulo 5

Conclusão e Trabalhos Futuros

Neste trabalho investigamos a relação entre a ocorrência de crimes e pontos de interesse (POIs) realizando análises de dados urbanos massivos georreferenciados. Para isso, realizamos engenharia de características para selecionar o melhor subconjunto de características para esse problema. Em seguida, coletamos os dados das mais diversas fontes e os integramos ao grafo das ruas da cidade de Maceió. Neste grafo aplicamos técnicas não supervisionadas para agrupar (clusterizar) esquinas com padrões similares socioeconômicos. Por fim, analisamos os dados através do Teste Qui-quadrado, do V de Cramér e Probabilidade Condicional.

Avaliamos que conforme proposto pela pergunta 1 as esquinas da cidade onde há atividade criminal podem de fato ser separadas em grupos com características socioeconômicas distintas. Tendo em vista a pergunta 2 e 3, analisamos a correlação entre ocorrência criminal e a proximidade de tipos específicos de pontos de interesse. A técnica do Qui-quadrado mostrou que a maioria dos POI's apresentaram correlação com os dados dos crimes. Além disso, notamos que, relacionando POI's e crimes, a probabilidade condicional da maioria dos POI's está acima da probabilidade da hipótese nula, ou seja, a probabilidade de ocorrer um crime dado que estamos próximos a um POI é maior do que a probabilidade da ocorrência criminal estando em qualquer esquina da cidade. Buscando responder à pergunta 4 utilizamos também o V de Cramer e a Probabilidade Condicional para ordenar uma lista com os POI's mais influentes.

Por fim, para responder a pergunta 5, calculamos as mesmas métricas para cada grupo de esquinas que apresentavam características socioeconômicas diferentes. Notamos que a influência dos pontos de interesse também alterou-se nas esquinas de um grupo para outro. Como por exemplo, no grupo 0, que é majoritariamente representado por localidades com renda muito baixa, quando existem postos de gasolina, estes podem ser locais perigosos. No grupo 2, as análises mostram que os pontos: estádio, agência bancária e entretenimento, são importantes, uma vez que próximos a eles sempre existe a ocorrência criminal, apresentando uma alta probabilidade. O que é preocupante uma vez que nos bairros ricos da cidade de Maceió temos muitos pontos turísticos e agências bancárias.

Finalmente, concluimos por meio deste estudo de caso realizado na cidade de Maceió, que a metodologia proposta é eficaz para realizar a análise de correlação entre POI's e ocorrências criminais, permitindo adquirir conhecimento relevante para o combate à criminalidade.

Como trabalhos futuros, visando potencializar a capacidade analítica e a extração de novas relações, pretendemos aprimorar e acrescentar algumas funcionalidades específicas:

1. **Captura de Dados Urbanos**: A expansão da variedade de dados urbanos a serem integrados, como informações sobre fluxo de tráfego, iluminação pública, padrões de uso de espaço público e dados meteorológicos, aprimorará a compreensão da dinâmica urbana e suas relações com as ocorrências criminais.

2. **Análise de Imagens**: A implementação de técnicas de inteligência artificial para análise de imagens capturadas, associadas a algoritmos avançados de reconhecimento de padrões. Proporcionará uma interpretação automatizada e mais sofisticada das condições reais daquela esquina.

3. **Aprimoramento da Interface de Usuário**: A melhoria da interface do usuário, com ferramentas intuitivas e amigáveis, garantirá uma experiência mais eficiente e acessível para os usuários da ferramenta.

Essas propostas de desenvolvimento futuro visam não apenas fortalecer a eficácia da ferramenta existente, mas também expandir suas capacidades analíticas, proporcionando uma plataforma mais abrangente e avançada para o entendimento e a gestão de questões relacionadas à segurança urbana.

Referências bibliográficas

ACKERMAN, N. L.; FREER, C. E.; ROY, D. M. On the computability of conditional probability. *J. ACM*, Association for Computing Machinery, New York, NY, USA, v. 66, n. 3, jun 2019. ISSN 0004-5411. Disponível em: <https://doi.org/10.1145/3321699>.

ALVES, L. et al. Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics. *PloS one*, v. 8, p. e69580, 12 2013.

BELESIOTIS, A.; PAPADAKIS, G.; SKOUTAS, D. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Trans. Spatial Algorithms Syst.*, Association for Computing Machinery, New York, NY, USA, v. 3, n. 4, apr 2018. ISSN 2374-0353. Disponível em: <https://doi.org/10.1145/3190345>.

Can Yang. *Maximum coverage location problem (MCLP)*. 2019. Disponível em: <https://github.com/cyang-kth/maximum-coverage-location>. Acesso em: 02 de outubro 2023.

CHURCH, R. L.; REVELLE, C. S. The maximal covering location problem. *Papers of the Regional Science Association*, v. 32, p. 101–118, 1974. Disponível em: <https://api.semanticscholar.org/CorpusID:154435809>.

CICHOCKI, A. et al. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. [S.l.: s.n.], 2009. ISBN 978-0-470-74666-0.

CRAMÉR, H. *Mathematical Methods of Statistics (PMS-9)*. Princeton University Press, 1999. ISBN 9780691005478. Disponível em: <http://www.jstor.org/stable/j.ctt1bpm9r4>.

ESCOVEDO, T.; KOSHIYAMA, A. *Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise*. Casa do Código, 2020. ISBN 9788572540551. Disponível em: <https://books.google.com.br/books?id=cL7TDwAAQBAJ>.

HAN, X. et al. Risk prediction of theft crimes in urban communities: An integrated model of lstm and st-gcn. *IEEE Access*, v. 8, p. 217222–217230, 2020.

HU, K. et al. Information fusion in crime event analysis: A decade survey on data, features and models. *Information Fusion*, v. 100, p. 101904, 2023. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253523002208>.

HU, K. et al. A systematic review of multi-scale spatio-temporal crime prediction methods. *ISPRS International Journal of Geo-Information*, v. 12, p. 209, 05 2023.

HUANG, C. et al. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2018. (CIKM '18), p. 1423–1432. ISBN 9781450360142. Disponível em: <https://doi.org/10.1145/3269206.3271793>.

KOHONEN, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, v. 43, p. 59–69, 2004. Disponível em: <https://api.semanticscholar.org/CorpusID:206775459>.

LIANG, W. et al. Towards hour-level crime prediction: A neural attentive framework with spatial–temporal-categorical fusion. *Neurocomputing*, v. 486, p. 286–297, 2022. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S092523122101729X>.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MUKHOPADHYAY, A. et al. Optimal allocation of police patrol resources using a continuous-time crime model. In: ZHU, Q. et al. (Ed.). *Decision and Game Theory for Security*. Cham: Springer International Publishing, 2016. p. 139–158. ISBN 978-3-319-47413-7.

PEARSON, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Taylor Francis, v. 50, n. 302, p. 157–175, 1900. Disponível em: <https://doi.org/10.1080/14786440009463897>.

ROSS, S. *Introduction to Probability and Statistics for Engineers and Scientists*. Elsevier Science, 2014. ISBN 9780123948427. Disponível em: <https://books.google.com.br/books?id=BaPOv33uZCMC>.

SALINAS, K. et al. Cityhub: A library for urban data integration. In: *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2022. v. 1, p. 43–48.

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

SILVEIRA, J. et al. *TensorAnalyzer: Identification of Urban Patterns in Big Cities using Non-Negative Tensor Factorization*. 2022.

VIEIRA, T. et al. Crime prediction and prevention using police patrolling data: challenges and prospects. In: SBC. *Anais Estendidos do XXXV Conference on Graphics, Patterns and Images*. [S.l.], 2022. p. 183–186.

WANG, D. et al. Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems*, v. 39, p. 93–106, 2013. ISSN 0198-9715. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0198971513000185>.

WANG, H. et al. Non-stationary model for crime rate inference using modern urban data. *IEEE Transactions on Big Data*, v. 5, n. 2, p. 180–194, 2019.

YANG, D. et al. Crimetelescope: crime hotspot prediction based on urban and social media data fusion. *World Wide Web*, v. 21, 09 2018.

ZANABRIA, G. G. et al. Mirante: A visualization tool for analyzing urban crimes. In: . [S.l.: s.n.], 2020. p. 148–155.

ZANABRIA, G. G. et al. Cripav: Street-level crime patterns analysis and visualization. *IEEE Transactions on Visualization and Computer Graphics*, PP, p. 1–1, 09 2021.

ZANABRIA, G. G. et al. Crimanalyzer: understanding crime patterns in são paulo. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

ZHAO, X. et al. Multi-type urban crime prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, v. 36, n. 4, p. 4388–4396, Jun. 2022. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/20360>.