Denilson da Silva Sousa
Orientador: Glauber Dias Gonçalves

Relações entre Crimes e o Espaço Urbano: Um Estudo de Caso Baseado em Pontos de Interesses Extraídos da Web

Picos - PI
08 de novembro de 2021

Denilson da Silva Sousa
Orientador: Glauber Dias Gonçalves

Relações entre Crimes e o Espaço Urbano: Um Estudo de Caso Baseado em Pontos de Interesses Extraídos da Web

Trabalho de conclusão de curso apresentado na Univerdade Federal do Piauí como parte dos requisitos necessários para a obtenção do grau de bacharel em Sistemas de Informação.

Universidade Federal do Piauí
Campus Senador Helvidio Nunes de Barros
Bacharelado em Sistemas de Informação

Picos - PI
08 de novembro de 2021

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Campus Senador Helvídio Nunes de Barros
Biblioteca Setorial José Albano de Macêdo
Serviço de Processamento Técnico

S719r Sousa, Denilson da Silva
Relações entre crimes e o espaço urbano: um estudo de caso baseado em pontos de interesses extraídos da web / Denilson da Silva Sousa– 2021.

Texto digitado
Indexado no catálogo *online* da biblioteca José Albano de Macêdo - CSHNB
Aberto a pesquisadores, com as restrições da biblioteca

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal do Piauí, Bacharelado em Sistemas de Informação, Picos-PI, 2021.

"Orientador: Glauber Dias Gonçalves."

1. Computação urbana. 2. Crimes-Índices. 3. Aprendizagem de Máquina 4. POI. I. Gonçalves, Glauber Dias. II. Título.

CDD 004.678

Maria José Rodrigues de Castro CRB 3: CE-001510/O

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PIAUÍ
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS
Curso de Sistemas de Informação

RELAÇÕES ENTRE CRIMES E O ESPAÇO URBANO: UM ESTUDO DE CASO BASEADO EM PONTOS DE INTERESSES EXTRAÍDOS DA WEB

DENILSON DA SILVA SOUSA

Monografia apresentada como exigência parcial para obtenção do grau de Bacharel em Sistemas de Informação.

Data de Aprovação

Picos – PI, 17 de Novembro de 2021

Glauber Dias Gonçalves

Prof. Glauber Dias Gonçalves

Deborah Maria Vieira Magalhães

Profa. Deborah Maria Vieira Magalhães

Flávio Henrique Duarte de Araújo

Prof. Flávio Henrique Duarte Araújo

Agradecimentos

Esta fase da minha vida é muito especial e não posso deixar de agradecer a Deus por toda força, ânimo e coragem que me ofereceu para ter alcançado minha meta. Por ter permitido que eu tivesse saúde determinação para não desanimar durante a realização deste trabalho.

Aos professores reconheço um esforço gigante com muita paciência e sabedoria. Foram eles que me deram recursos e ferramentas para evoluir um pouco mais todos os dias. Agradeço especialmente o professor Glauber por ter sido meu orientador e ter desempenhado tal função com dedicação e amizade.

É claro que não posso esquecer da minha família. Agradeço em especial a minha mãe Antônia Valdete, heroína que me deu apoio, incentivo nas horas difíceis, de desânimo e cansaço. Obrigado, Daniel, irmão querido, por ser tão companheiro.

Meus agradecimentos aos amigos de turma Vinicius, Kamargo, Nathan, Pedro e Benedito, companheiros de trabalhos e irmãos na amizade que fizeram parte da minha formação e que vão continuar presentes em minha vida com certeza. Só tenho a agradecer aos meus amigos, Felipe, Laninha, Alana e Jéssica pelas risadas e bons momentos que vocês compartilharam comigo ao longo desses anos nessa etapa tão desafiadora da vida acadêmica. Minha eterna gratidão.

A todos que direta ou indiretamente fizeram parte de minha formação, o meu muito obrigado.

A tecnologia tornou possível a existência de grandes populações. Grandes populações agora tornam a tecnologia indispensável.

Joseph Krutch

Resumo

A quantidade de dados disponíveis na Internet bate recordes a cada ano, incluindo dados de texto, imagem e vídeo (multimídia) com georreferenciamento. A disponibilidade de informações georreferenciadas sobre locais e eventos são mais comuns em serviços web de transporte e mapeamento das vias urbanas como o Open Street Map, WikiMapia e Google Maps. Essas informações são alimentadas pelos próprios usuários dos serviços e podem refletir características e problemas de regiões urbanas. A área de pesquisa que estuda tais problemas é conhecida como computação urbana. Um grande problema da sociedade brasileira são os altos índices de criminalidade que estão dentre os principais problemas que afetam negativamente a qualidade de vida nos centros urbanos. No Brasil, em particular, estima-se uma taxa média de 20 mortes por mês para cada 100 mil habitantes em decorrência de situações de violência. As altas taxas de criminalidade nas cidades brasileiras poderiam ser melhor analisadas e compreendidas a partir de fontes de dados alternativas que exploram características do espaço urbano. Neste trabalho de conclusão de curso, investigamos a relação entre índices de criminalidade e essas características refletidas em pontos de interesse (POIs) que as pessoas registraram em um serviço Web na cidade de São Paulo. Mostramos o potencial desse tipo de dado para predizer índices de crimes por regiões da cidade. Nesse sentido, construímos modelos de regressão com desempenhos satisfatórios para essa predição, no qual obtivemos para os crimes de Furtos e Roubos erros inferiores a 46% dos índices reais. Exploramos também esses modelos para descobrir as categorias de POIs mais importantes para explicar os crimes mais frequentes por regiões das cidades. Adicionalmente, analisamos o ganho de desempenho com o aumento de POIs registrados na cidade de São Paulo em 2012 a 2020 e descobrimos que os erros diminuem em 4% ao passar dos anos.

Palavras-chaves: Computação Urbana, POI, Índice de crimes, Aprendizagem de Máquina.

Abstract

The amount of data available on the Internet breaks records each year, including text, image and video (multimedia) data with georeferences. The availability of georeferenced information about places and events is more common in web transport and urban street mapping services such as Open Street Map, WikiMapia and Google Maps. This information is fed by the service users themselves and may reflect characteristics and problems of urban regions. The area of research that studies such problems is known as urban computing. A major problem in Brazilian society is the high crime rates, which are among the main problems that negatively affect the quality of life in urban centers. In Brazil, in particular, an average rate of 20 deaths per month for every 100,000 inhabitants is estimated as a result of situations of violence. The high crime rates in Brazilian cities could be better analyzed and understood from alternative data sources that explore characteristics of urban space. In this graduation paper, we investigate the relationship between crime rates and these characteristics reflected in points of interest (POIs) that people have registered on a web service in the city of São Paulo. We show the potential of this type of data to predict crime rates by city regions. In this sense, we built regression models with satisfactory performances for this prediction, in which we obtained for the crimes of Theft and Robbery errors lower than 46% of the real indexes. We also explored these models to discover the most important POI categories to explain the most frequent crimes by city regions. Additionally, we analyzed the performance gain with increasing registered POIs in the city of São Paulo from 2012 to 2020 and found that errors decrease by 4% over the years.

Key-words: Urban Computing, POI, Crime Rate, Machine Learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ANOVA	Análise de Variância
API	Application Programming Interface
DBSCAN	Density-Based Spatial Clustering of Applications (Agrupamento espacial de aplicativos baseado em densidade)
DP	Distrito Policial
FA	Floresta Aleatória
GPS	Global Positioning System
IBGE	Instituto Brasileiro de Geografia e Estatística
IDH	Índice de Desenvolvimento Humano
LSBN	Location-Based Social Networks (Redes sociais baseadas em localização)
MVS	Máquina de vetores de suporte
POI	Point of Interest (Ponto de Interesse)
ONU	Organização das Nações Unidas
RF	Random Forest
RL	Regressão Linear
SP	São Paulo
SIG	Sistema de Informação Geográfica
SVC	Support Vector Classifier
SVM	Support Vector Machine
SVR	Support Vector Regression

Sumário

1 Introdução

A computação urbana é uma área de pesquisa interdisciplinar que visa entender e tratar os problemas das cidades para melhorar a qualidade de vida das pessoas que nelas vivem (SILVA et al., 2019). Dentre os diferentes desafios dos centros urbanos, estão o enfrentamento aos altos índices de criminalidade. Para se ter uma ideia da gravidade desse problema no Brasil, no ano de 2019, foram registradas 41.726 mortes por crimes violentos (NEV-USP, 2021), uma taxa média de 20 mortes por mês para cada 100 mil habitantes brasileiros. Essa taxa pode alcançar valores superiores a 40 mortes por mês em estados das regiões norte e nordeste.

As altas taxas de mortes violentas e demais crimes nas cidades brasileiras poderiam ser melhor analisadas e compreendidas a partir de fontes de dados alternativas que exploram características do espaço urbano. Estudos em ciências sociais e geografia indicam que índices de crimes por região têm relações com as características do espaço urbano como meios de transporte, opções de educação, trabalho, saúde e entretenimento (NERY; SOUZA; ADORNO, 2019; ADORNO; NERY, 2019). Tais características, já são consideradas, atualmente, em estatísticas oficiais como o censo. Elas são consideradas através de características urbanísticas do entorno dos domicílios com dados agregados por municípios[1]. Contudo, para subsidiar novos estudos que avancem nesse tema é necessário o desenvolvimento de métodos para a coleta e processamento de características das regiões urbanas em maior amostragem e granularidade por regiões urbanas como bairros e ruas.

Pontos de Interesse (POI) extraídos de serviços Web de mapeamento urbano como *Open Street Maps (OSM)* e *FourSquare* fornecem informações sobre um local da cidade com uma categoria, coordenadas geográficas, popularidade e comentários alimentados por pessoas. A categoria do POI tipicamente identifica um tipo de atividade que ocorre nesse local como restaurantes, lojas, teatros, escolas, etc. Esse tipo de dado vem sendo utilizado em uma variedade de estudos como fonte de informação sobre características espaciais das cidades extraídas da Web (WEISBURD; GROFF; YANG, 2012; YUAN; ZHENG; XIE, 2012; WANG et al., 2021). Logo, POIs também são potencialmente úteis para estudos sobre crimes, visto que características de um local, em especial tipos de atividades desenvolvidas, podem indicar ocorrências de alguns tipos de crimes.

Nesse sentido, a comunidade científica de computação, vem explorando POIs gerados por pessoas e disponíveis publicamente via serviços Web para predição de crimes. Em (WANG et al., 2019) e (BELESIOTIS; PAPADAKIS; SKOUTAS, 2018) foi mostrado que pontos de interesses registrados por pessoas nos serviços *OSM* e *FourSquare* combinados com dados demográficos oficiais possibilitam a predição das taxas criminais em

[1] https://www.ibge.gov.br/geociencias/organizacao-do-territorio/tipologias-do-territorio/24702-caracteristicas-urbanisticas-do-entorno-dos-domicilios.html

menor granularidade espacial, i.e., regiões da cidade. Em (HUANG et al., 2018), dados de crimes oficiais da cidade de Nova York foram incrementados com POIs para prever se uma categoria de crime acontecerá numa região num tempo futuro. No trabalho (CASTRO; RODRIGUES; BRANDAO, 2020), também foi observado que dados não oficiais, extraídos do serviço Web brasileiro "Onde Fui Roubado", adicionados às fontes oficiais melhora significativamente a predição de índices de criminalidade.

Todos esses trabalhos são baseados em fontes de dados oficiais e usam POIs e outros conteúdos gerados por habitantes das cidades em serviços Web apenas como um incremento de informação. Contudo, há ainda uma questão sobre o potencial de POIs a ser esclarecida. Especificamente, até que ponto o uso desse tipo de dado *unicamente* pode refletir os índices de criminalidade das cidades? A investigação dessa questão é importante porque em uma eventual falta ou atraso na coleta de dados oficiais, POIs poderiam oferecer indicações ou estimativas aproximadas de índices de violência por região da cidade para orientar os gestores responsáveis pela segurança pública. Além disso, os POIs podem auxiliar no estudo quando eles estão discrepantes em relação aos índices de violência para uma determinada região. Por exemplo, caso os dados oficiais indiquem uma alta taxa de crime para uma região específica, mas os POIs indicam uma baixa taxa, pode-se ter algum motivo causando este comportamento. Uma outra possibilidade é quando os dados oficiais indicam uma baixa taxa de criminalidade, mas os POIs indicam uma alta taxa, pode ser que as pessoas não estejam reportando incidências de crimes no qual elas foram vítimas ou que ocorreram naquela determinada região. No entanto, é necessário conhecer o nível de acurácia e especificidade desses dados para explicar índices de crimes.

Neste trabalho de conclusão de curso de graduação investigamos essa questão com a utilização de POIs extraídos do serviço Web *OSM* e a avaliação do potencial desse tipo de dado para predizer índices de crimes por regiões da cidade. Nosso foco foi a cidade de São Paulo, que é a maior da América Latina e concentra portanto altos índices de criminalidade.

O estado de São Paulo é um dos poucos que disponibiliza publicamente índices de crimes por categoria e região (distrito policial) de todas as suas cidades mensalmente (SSP-SP, 2021). Adicionalmente, a cidade de São Paulo concentra um vasto número de POIs devido a sua importância econômica a nível internacional. Desse modo, São Paulo é a cidade com mais condições propícias para investigarmos o quanto os POIs conseguem explicar crimes.

Conduzimos essa investigação baseada em oitenta e oito distritos policiais, unidades de espaço, onde relacionamos as ocorrências de crimes com os POIs existentes em cada unidade. Para quantificar essa relação utilizamos modelos de regressão e análises de erros desses em inferências sobre a taxa anual de ocorrência das categorias de crimes mais frequentes na cidade.

Nossos resultados indicam que POIs unicamente podem explicar razoavelmente o nú-

mero de ocorrências de algumas categorias de crimes. Em particular, os crimes mais frequentes em regiões centrais da cidade como furtos, onde a quantidade de POIs é maior. Por exemplo, o melhor modelo de regressão avaliado (floresta aleatória) pode predizer furtos com erro médio de 39% relativo ao índice oficial e curiosamente a categoria de POI ponto de táxi é a mais importante (66%) para tal inferência. Outras categorias de crimes mais comuns em regiões periféricas como homicídios, ainda podem ser explicados via POIs apesar da menor quantidade de dados para a construção de modelos, o que leva a erros médios de até 93% em relação ao índice oficial. Nesse contexto, um total de sete categorias de crimes dentre os mais frequentes em São Paulo podem ser explicados com POIs, considerando coeficientes de determinação (R2) positivos e média de erros relativos (MRE) inferiores ao índice oficial (i.e., menor que 100%). Observamos também ganhos de desempenho dos modelos com a redução de erros absolutos em pelo menos 4% anualmente à medida em que houve crescimento na quantidade de POIs entre os anos de 2012 a 2020.

1.1 Objetivos Gerais e Específicos

Este trabalho tem como objetivo investigar o potencial para predição de índices de violência em regiões urbanas apenas com dados da Web gerados por usuários:

- Análise do potencial de características do espaço urbano extraídas *unicamente* de POIs na Internet para explicar (i.e., predizer) índices de criminalidade.
- Quantificação de categorias de POIs mais relevantes para predição de crimes, assim como o impacto do aumento de POIs por região ao longo do tempo no desempenho da predição.

Os objetivos acima propostos visam não somente a realização de um trabalho acadêmico, mas também iniciar uma pesquisa que pode resultar em benefícios para a sociedade. Podemos citar como possíveis benefícios advindos desse trabalho: (i) descoberta de características do espaço urbano correlacionadas com tipos de crimes; (ii) descoberta de fontes de dados alternativas para monitorar taxas de crimes; (iii) desenvolvimento de modelos de predição de crimes com essas fontes; e (iv) métodos para auxiliar a elaboração de políticas públicas de redução ou controle das taxas de violência em regiões urbanas. Adicionalmente, esse trabalho pode estimular mais órgãos governamentais a divulgar abertamente na Web dados locais sobre violência, assim como a Secretaria de Segurança Pública do estado de São Paulo faz atualmente. A disponibilidade fácil de tais informações abrem um leque de oportunidades para análises e predição de crimes via fontes de dados alternativas extraídas da Web, em especial, dados gerados por usuários.

1.2 Organização do Trabalho

As próximas seções deste trabalho estão organizadas com os seguintes capítulos.

No Capítulo 2, apresentamos o referencial teórico, isto é, as técnicas computacionais utilizadas para a realização deste trabalho. A seguir, no Capítulo 3, descrevemos os trabalhos da literatura técnica especializada relacionados com os temas computação urbana e predição de crimes via conteúdo gerado por usuários na Web que motivaram o desenvolvimento deste trabalho. Prosseguimos então, explicando a metodologia aplicada neste trabalho, desde a etapa de aquisição de dados ao desenvolvimento de modelos preditivos no Capítulo 4.

Os resultados desse trabalho são apresentados no Capítulo 5, e eles foram divididos em três tópicos: (1) Desempenho de diferentes métodos, (2) Importância de POIs e (3) Impacto do Aumento de POIs. Finalmente, no Capítulo 6, apresentamos nossas considerações finais e possíveis trabalhos futuros.

2 Referencial Teórico

Esta Seção descreve conceitos fundamentais para o entendimento deste projeto. Primeiramente, definimos e abordamos os principais aspectos da Computação Urbana na Seção 2.1. Em seguida discutimos um pouco sobre a ferramenta Open Street Maps na Seção 2.2. A seguir abordamos de forma sucinta o Aprendizado de Máquina, tal como os algoritmos de Regressão na Seção 2.3.

2.1 Computação Urbana

O termo "computação urbana" foi utilizado por Eric Paulos pela primeira vez em 2004 na conferência UbiComp (PAULOS; ANDERSON; TOWNSEND, 2004) e em seu artigo "The Familiar Stranger" (PAULOS; GOODMAN, 2004).

A computação urbana é um processo de aquisição, integração e análise de um extenso volume de dados. Esses dados podem estar em diversos formatos e são gerados por inúmeras fontes em espaços urbanos. As fontes podem ser sensores, dispositivos, veículos, softwares e até humanos. Os dados extraídos dessas fontes são utilizados com o objetivo de melhorar o estilo de vida das pessoas que moram na região urbana. Essa melhoria se dá pelo tratamento de problemas urbanos como a poluição do ar, aumento do consumo de energia, falta de água, congestionamento no trânsito e outros (ZHENG et al., 2014).

De uma forma mais ampla, a computação urbana procura ajudar a entender os problemas e causas desses que envolvem os fenômenos urbanos, bem como prever o futuro das cidades. Desse modo, pode-se dizer que computação urbana é uma área interdisciplinar decorrente das semelhanças entre a ciência da computação com áreas tradicionais como economia, sociologia e transporte, no cenário dos espaços urbanos. No âmbito da ciência da computação, a computação urbana tem relações com as áreas de redes de computadores, redes veiculares, redes de sensores, sistemas distribuídos, sistemas colaborativos, interação humano-computador, inteligência artificial e redes sociais (SILVA et al., 2019).

A pesquisa em computação urbana tem ganhado foco nessa última década. Esse foco é devido ao crescimento de fontes públicas de dados na web com informações temporais e espaciais, assim como a evolução de técnicas computacionais para aquisição desses dados (SILVA et al., 2019). Esses dados têm contribuído para estudos, cujo objetivo é entender a evolução de atividades virtuais das pessoas na Internet e sua relação com os problemas das cidades. As atividades podem ser culturais, gastronômicas, festivas e turísticas (VACA et al., 2015). Além dessas, pode-se ainda citar o fluxo origem-destino que são atividades mais comuns dos residentes das cidades (SILVA et al., 2014).

2.1.1 Estrutura da Computação Urbana

Nesta Seção apresentamos a estrutura da computação urbana. Essa estrutura é dividida em 3 componentes principais. Os componentes são: (i) gerenciamento dos dados urbanos; (ii) análise dos dados urbanos; e (iii) desenvolvimento de serviços e aplicações (RODRIGUES et al., 2019). Esses componentes, por sua vez, podem ser subdivididos em inúmeras atividades para alcançar um determinado objetivo. A Figura 1 mostra uma visão geral dessa estrutura.

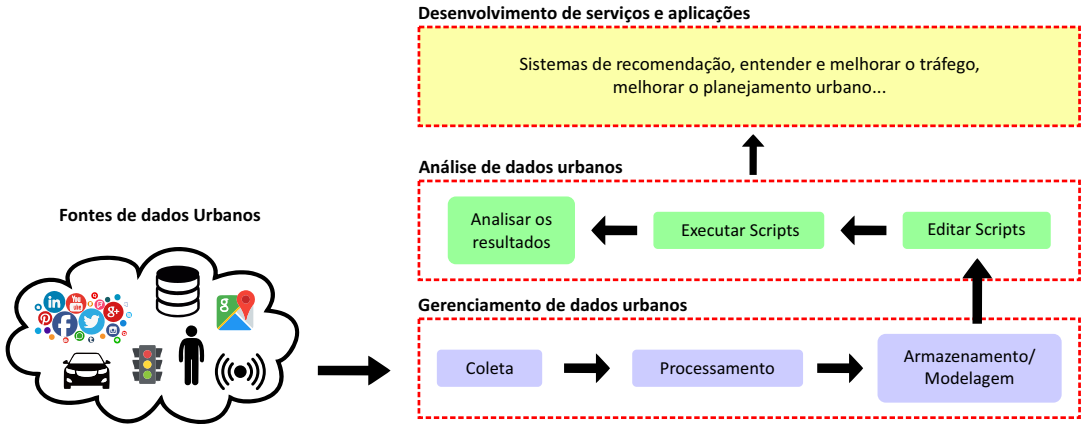


Figura 1 – Visão Geral Da Estrutura Da Computação Urbana. [Inspirada em (SILVA et al., 2019)].

A primeira atividade, (i) gerenciamento dos dados urbanos, é composta, primeiramente, pelo processo de coleta dos dados urbanos. Essa coleta pode ser obtida de diversas fontes que serão discutidas na Seção 2.1.2. A segunda etapa, ainda da primeira atividade, consiste no processamento dos dados. Este processamento tem como objetivo gerar informações que possam ser utilizadas na fase de modelagem. O processamento ocorre de modo individual para cada fonte, já que podemos ter dados em vários formatos e padrões. Eles podem ser modelados em tabelas, estatísticas, grafos e outros.

A segunda atividade ou (ii) análise dos dados urbanos, é constituída do desenvolvimento e execução de scripts, tal como a análise dos resultados. Os dois primeiros passos são semelhantes com as etapas da atividade (i). Aqui eles tem como finalidade processar as informações obtidas de diversas fontes em conjunto, ou seja, combiná-las na procura de um padrão. Já o passo de análise tem como objetivo auxiliar as pessoas a entenderem as características dos dados urbanos para que possam utilizá-los de forma eficiente.

A terceira e última atividade, (iii) desenvolvimento de serviços e aplicações, consiste na execução de métodos para o desenvolvimento de tecnologias inteligentes usando as informações obtidas nas atividades anteriores. Nesta atividade é comum a elaboração de modelos descritivos e preditivos. Os modelos descritivos são utilizados para identificar associações entre variáveis. Os benefícios desse modelo é a caracterização de eventos ocorridos no passado, do mesmo modo que ajuda na tomada de decisões futuras. Já os modelos

preditivos permitem a identificação de padrões omissos além dos efeitos causados por eles.

2.1.2 Fonte de Dados

Nesta Seção apresentamos algumas fontes de dados que podem ser utilizadas para extrair informações relevantes dentro da computação urbana. Essas informações são fundamentais para o desenvolvimento de serviços e aplicações que ajudam no tratamento de problemas nas cidades (SILVA; LOUREIRO, 2015).

- **Dados estatísticos oficiais**: são fontes que fornecem dados provenientes de uma análise estatística sobre uma parte, ou, um todo da população. Esses dados podem ser demográficos, sociais, econômicos e outros. Na maioria dos casos, são dados divulgados como resultado de estudos realizados por organizações do governo, já que sua coleta e compilação tem altos custos (FEIJÓ; VALENTE, 2005). Exemplo de pesquisa em computação com esse tipo de fonte é o trabalho de Quercia et al. (2015), no qual é estudado a relação entre o cheiro relatado por pessoas com os índices de poluição dos centros urbanos. Da mesma forma (BELESIOTIS; PAPADAKIS; SKOUTAS, 2018) utiliza dados do censo inglês juntamente a outras fontes de dados para a predição de crimes em Londres.

- **Redes de sensores tradicionais**: são fontes que fornecem dados adquiridos em dispositivos de determinados sensores para uma aplicação. Pode-se citar como exemplo, sensores de movimentos utilizados em ruas para quantificar o fluxo de pessoas naquela localidade, sensores pluviométricos para medir o volume de chuva em determinada área, sensores para monitoramento da qualidade do ar e entre outros. Em (VLAHOGIANNI et al., 2014) são utilizados sensores de estacionamentos da cidade Santander na Espanha para predizer e verificar a probabilidade de vagas disponíveis em determinado local em um curto período de tempo.

- **Infraestrutura das cidades**: são fontes que fornecem dados a partir de infraestruturas disponíveis nas cidades com outras finalidades. Pode-se citar como exemplo o GPS de veículos (ônibus, táxis, automóveis pessoais). Os dados obtidos desse serviço, em questão, podem auxiliar em estudos sobre o trânsito de uma cidade em pontos desejados. Um outro exemplo é a rede de telefonia celular que pode fornecer dados para auxiliar em estratégias de mobilidade urbana. Por exemplo, nos trabalhos de Oliveira et al. (2017), Naboulsi, Stanica e Fiore (2014) é utilizado os sinais de smarthphones para prever e caracterizar a mobilidade individual e assim ajudar no planejamento urbano.

- **Redes de sensoriamento participativo**: fornecem dados provenientes de aplicações alimentadas por usuários. São fontes de dados mais comuns em pesquisas da computação urbana por consequência das aplicações, em sua maioria, serem redes

sociais com recursos espaciais e temporais. As redes sociais ajudam nesses estudos por facilitar a identificação de relacionamentos e interações entre os usuários. Pode-se citar como exemplo o Facebook, Twitter, Instagram, Google Maps, Open Street Map e outros. Em (REDI et al., 2018) dados provenientes dessas redes foram utilizados para mapear áreas seguras em bairros de Londres.

2.1.3 Sensoriamento participativo: Crowdsourcing e Crowdsensing

A computação urbana, como descrito anteriormente, tem como uma das principais funções o processo de coleta, integração e análise de um grande volume de dados. Esses dados costumam ser heterogêneos, isto é, possuem diversos formatos e podem ser adquiridos de diversas fontes dos espaços urbanos (RODRIGUES et al., 2019). O Sensoriamento participativo permite analisar e armazenar dados em grande escala, quase que em tempo real. Isso possibilita o monitoramento a longo prazo de atividades de pessoas nos grandes centros urbanos. Os dados históricos, armazenados, de atividades humanas ajudam a entender com mais eficiência o comportamento social em diversas regiões do mundo, tal como reagir em tempo hábil a eventos inesperados (SILVA et al., 2014). O sensoriamento participativo compreende também dois paradigmas importantes: *crowdsensing* e *crowdsourcing*.

O *crowdsensing*, ou sensoriamento participativo/móvel, é um paradigma que possui atuação da população gerando dados de forma passiva. Esses dados são coletados em segundo plano por dispositivos com recursos de GPS que possibilitam a extração de informações georreferenciadas (SIMÃO, 2019). Dessa forma, esse paradigma permite o monitoramento de eventos em grande escala como congestionamentos em vias urbanas, poluição sonora, tráfego da população em áreas urbanas e entre outros. A monitoração dos fenômenos ocorre por meio do compartilhamento de dados entre os dispositivos móveis sensoriados.

Os dados compartilhados oferecem grande potencial para aplicações geográficas que ajudam na mobilidade urbana da população. São exemplos de ferramentas: o *Waze* (ver Figura 2, esq.) que compartilha, em tempo real, as condições de tráfego no trânsito além de alertar sobre acidentes, blitz policial, bloqueio de vias, condições climáticas e entre outros (GOOGLE, 2021); *Google Maps* é uma das ferramentas da empresa *Google*, cujo os usuários podem encontrar locais, visualizar as melhores rotas, tempo de viagem, estimar a distância entre dois pontos geográficos e até mesmo verificar se sua região foi afetada pela COVID-19.(ver Figura 2, dir.).

O *crowdsourcing* diferente do *crowdsensing* é um paradigma que possui o envolvimento de uma grande massa de usuários gerando dados de forma ativa. Cada usuário, de modo independente, e sem alguma especialização fica responsável por pequenas tarefas de coleta ou análise de dados. Essas pequenas, ou micro tarefas, são posteriormente analisadas pelos organizadores da tarefa ou verificadas em coletivo (ASMOLOV, 2014; HOWE et al., 2006).

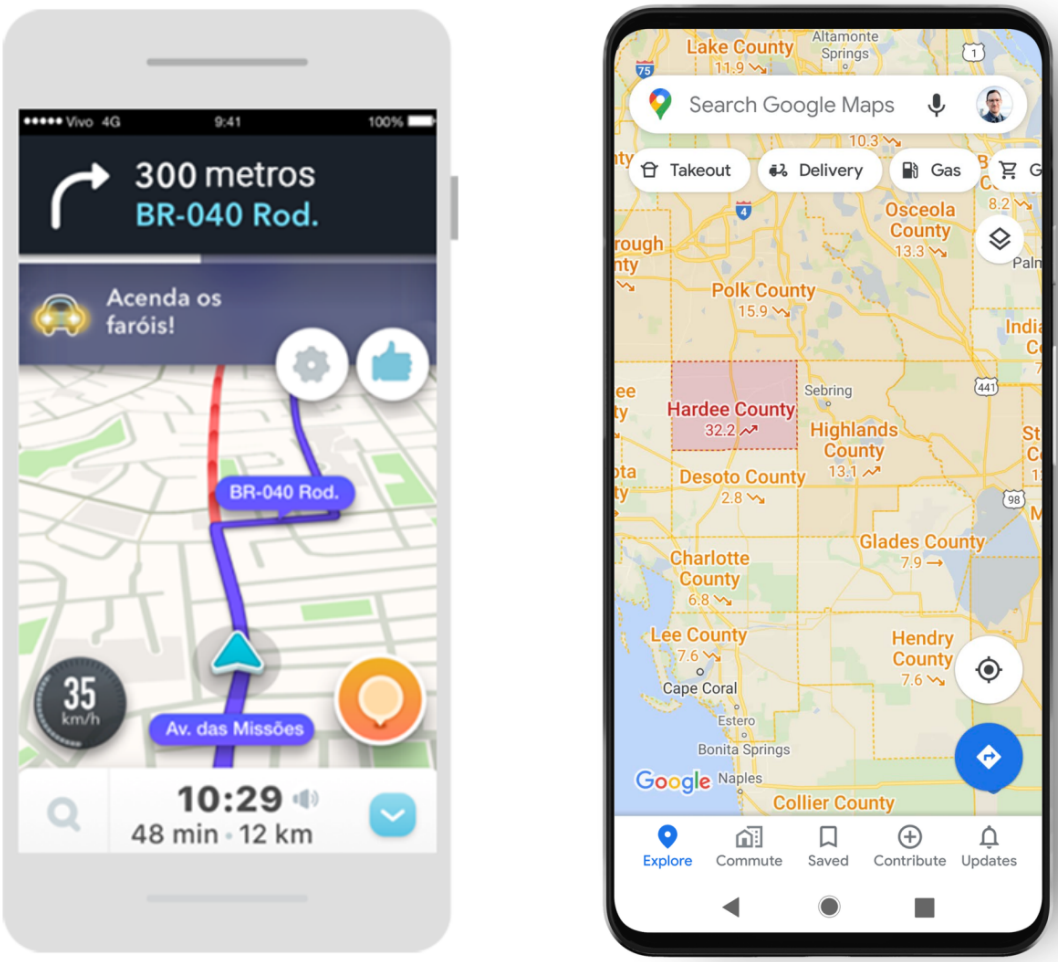


Figura 2 – Telas de Aplicações: Waze e Google Maps. Fonte: Waze (Divulgação Na PlayStore); Google Maps (G1: Google Maps Mostra Quais Estados Possuem Mais Novos Casos de Covid-19).

Um dos maiores exemplos de aplicações colaborativas é o *Wikipedia* - uma espécie de *wiki* cujo textos e outros dados são produzidos por usuários em grande escala - ou a ferramenta de mapeamento geográfico *Open Street Map*.

2.2 Open Street Map

O *Open Street Map* (OSM) é um projeto de mapeamento geográfico colaborativo que fornece um arcabouço munido de dados, mapas e suas respectivas APIs para inúmeros websites, aplicativos móveis e outros dispositivos de hardware. Todo esse arcabouço é desenvolvido por uma comunidade voluntária que mantém seus dados sobre ruas, estradas, trilhos, pontos de interesses, linhas ferroviárias e outros pelo mundo atualizados (OPENSTREETMAP, 2021). Os recursos do OSM são abertos para uso gratuito, o que incentiva a sua utilização em várias aplicações como: navegação GPS, estudos estatísticos, mapas

ilustrativos, SIGs e outros.

Na Figura 3 pode-se observar a ferramenta gráfica do OSM[1]. Os dados sobre algum local no mapa podem ser visualizados ao clicar nos ícones de estabelecimentos, ou, Pontos de Interesses. Esses dados podem ser extraídos e manipulados de uma forma mais fácil utilizando APIs públicas desenvolvidas por terceiros. São exemplos de APIs *Osmium Tool, Osmosis API, Overpass Turbo, OSMPythonTools*, dentre outras. As duas primeiras são as mais populares e foram utilizadas nesse trabalho. *Osmosis API* e *Osmium Tool* são, ambas, ferramentas para manipulação de dados brutos do OSM, tipicamente nomeados com a extensão *.osm*. A *Osmosis API* é uma aplicação *JAVA* de linha de comando, ao passo que a *Osmium Tool* está disponível em várias linguagens como *Node, Python, JAVA* e outros. As funcionalidades dessas ferramentas envolvem:

- Extrair dados de uma caixa delimitadora ou um polígono;
- Carregar dados do planeta para um banco de dados;
- Aplicar mudanças nos dados do OSM;
- Comparar dados (históricos) de dois arquivos OSM;
- Procurar estabelecimentos por categorias;
- Filtrar arquivos por Tag;
- Converter arquivos de um formato para outro;

2.3 Aprendizado de Máquina

O Aprendizado de Máquina é um subcampo de estudo relacionado à área de Inteligência Artificial (IA). O seu objetivo é elaborar técnicas e métodos computacionais que tenham a capacidade de tomar decisões (MICHALSKI; CARBONELL; MITCHELL, 2013). Ou como define (ALPAYDIN, 2020):

> "... são programas de computador utilizados para otimizar um critério de desempenho, usando dados de exemplo ou experiência do passado."

Dessa forma, essas técnicas são aplicadas em problemas que não conseguem ser resolvidos por métodos tradicionais de programação como algoritmos funcionais, imperativos e orientados a objetos (PISTORI, 2003). Podemos citar como exemplo a classificação de uma mensagem em três classes: sentimento positivo, sentimento negativo ou neutro. Para

[1] Disponível em: https://www.openstreetmap.org/#map=19/-23.55172/-46.63469

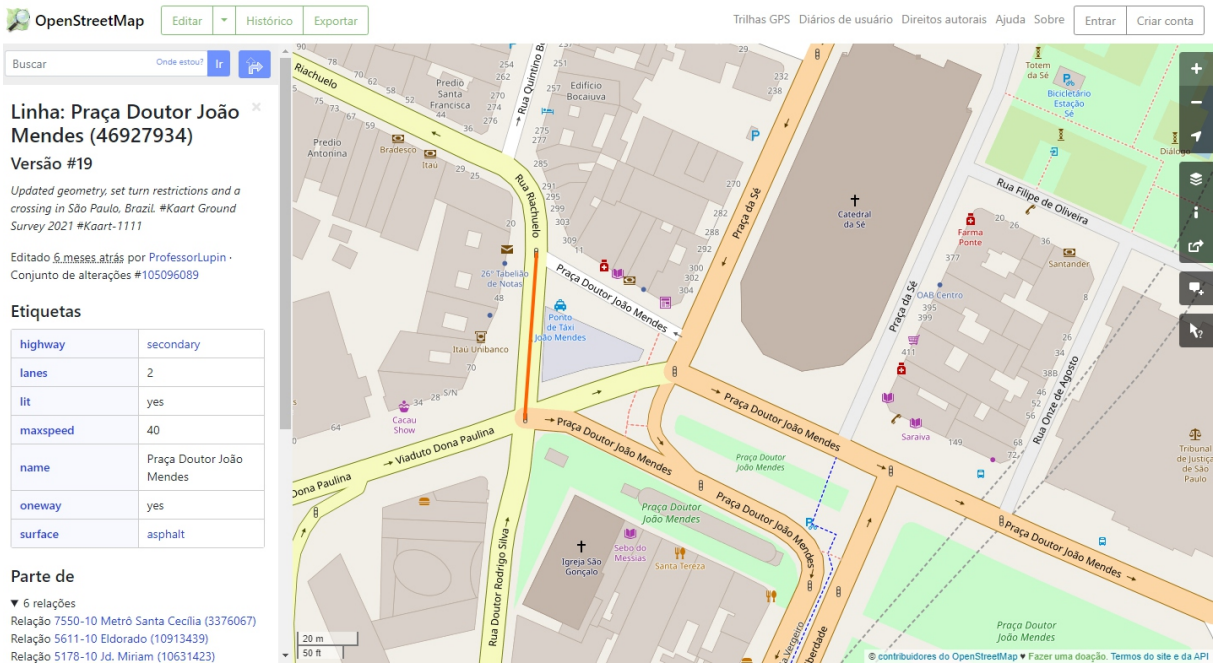


Figura 3 – Tela do Open Street Map no Browser Mostrando Em Destaque a Rua Próxima a Praça Doutor João Mendes Em Sé, SP.

este problema a entrada é conhecida: um texto contendo uma mensagem; e a saída também: o sentimento da mensagem como positivo, negativo ou neutro. A dificuldade está no processo de transformação doa dados de entrada no de saída. Portanto, os algoritmos de aprendizagem de máquina são a melhor solução para resolver esse tipo de problema, já que eles podem aprender padrões a partir de dados reais obtidos.

2.3.1 Caracterização

Algoritmos de aprendizagem de máquina podem ser organizados em três tipos: Supervisionado, Não Supervisionado, e Aprendizado por reforço de acordo com Haykin (2010).

- **Aprendizado Supervisionado** - No aprendizado supervisionado é necessário que a aplicação tenha conhecimento do problema, ou seja, os dados utilizados inicialmente de entrada precisam estar rotulados (pares de entrada-saída) que serão utilizadas pelo computador para tentar chegar no resultado esperado.

- **Aprendizado Não-Supervisionado** - Diferente do aprendizado supervisionado, o aprendizado não-supervisionado não tem conhecimento do problema. Não existe uma saída desejada, logo, é função dos métodos criarem suas próprias saídas (classes) com base nos dados de entrada .

- **Aprendizado por Reforço** - No Aprendizado por reforço, o sistema faz um mapeamento entrada-saída por meio de uma interação contínua com o problema afim de obter o melhor resultado desejado.

2.3.2 Algoritmos de Regressão

Os problemas encontrados no ambiente de aprendizagem de máquina podem ser classificados como problemas de classificação e regressão. O problema é de classificação quando os rótulos ou classes assumem valores discretos de 1 a k. Quando os rótulos são valores contínuos, o problema passa a ser um problema de regressão. A diferença entre problemas de classificação e regressão está entre seus objetivos finais. Os algoritmos de classificação analisam características e atribuem dados a um domínio conhecido, ao mesmo tempo que os algoritmos de regressão procuram realizar previsões por meio de dados adquiridos anteriormente (LORENA; CARVALHO, 2007). Para este trabalho, são utilizados algoritmos de aprendizagem de máquina focados em problemas de regressão.

Os algoritmos de regressão linear são suficientemente eficientes para resolver problemas com baixo grau de dificuldade. É uma das abordagens mais simples do aprendizado de máquina supervisionado (GROSS, 2012). Quando o problema sugere a presença de uma relação entre duas variáveis é possível representá-la por meio de uma reta, ou seja, o modelo assume a função da reta:

$$y = ax + b \tag{2.1}$$

Onde, o coeficiente y, ou termo independente, é o que se deseja prever, x é a variável dependente, a é a inclinação da reta em um plano cartesiano e b é a constante que delimita o valor para y quando x é 0. Essa é a Regressão Linear simples. Entretanto, na maioria dos problemas, a variável dependente precisa de mais de uma variável independente para ser representada. Essa é a Regressão Linear Múltipla que pode ser representada pela função:

$$y_i = \alpha + \beta X_i + e_i, \tag{2.2}$$

onde y_i é a variável dependente, X é uma matriz de variáveis independentes, α é a constante do modelo, β é um vetor de coeficientes angulares para cada variável do modelo e e é o erro, ou variação de y_i não explicada pelo modelo. Nesse caso X_i representa um vetor de variáveis correspondentes à variável independente y_i.

O algoritmo *Support Vector Machine* (SVM), ou Máquina de vetores de suporte, tem a capacidade de reconhecer padrões tênues em bases de dados complexas. A solução ficou conhecida popularmente após Drucker et al. (1997) obter um desempenho considerável na aplicação de predições de regressão e séries de tempo. Uma SVM tem competência para resolver tantos problemas de classificação (SVC) como também de regressão (SVR).

O algoritmo Floresta Aleatória (FA), ou *Random Forest* como é popularmente conhecido, assim como o SVM pode ser utilizado para resolver problemas de classificação e regressão (BREIMAN, 2001). A FA além de criar várias árvores, onde cada uma das árvores usa uma amostra diferente dos dados, ela também altera a forma como as árvores de regressão ou classificação são montadas.

O algoritmo FA divide o conjunto de dados originais em n árvores de decisão, cada uma com x amostras de dados aleatórios. Ao final dos resultados, seja de regressão ou classificação, as árvores escolhem o resultado majoritário como resultado final.

3 Trabalhos Relacionados

Há alguns anos, pesquisadores das áreas de ciências sociais e urbanismo têm investigado a relação entre crimes, características das populações e espaço geográfico a partir de dados oficiais de governos como censo demográfico, dados de mobilidade urbana e estatísticas sociais e econômicas. Masi et al. (2007) realizam um estudo clássico na área e investigam o quanto questões raciais e o grau de violência influenciam nos resultados de gravidez, enquanto Tonry (1997) e (NORONHA et al., 1999) analisam a distribuição da violência no espaço relacionado a aspectos de etnia e cor racial. Em (BECKER; KASSOUF, 2017), os autores analisam se o gasto público do governo em educação impacta na redução da taxa de homicídios. Mais recentemente, em (ADORNO; NERY, 2019; NERY; SOUZA; ADORNO, 2019), investiga-se a violência na cidade de São Paulo ao longo dos anos para analisar a distribuição dos crimes na cidade, confrontando teorias que definem de forma estática bairros violentos e não-violentos, regiões centrais e periferia.

Há também uma linha de especialistas na área de criminologia cujo foco é analisar crimes por regiões de uma cidade, ou unidades geográficas pequenas e específicas. Weisburd, Groff e Yang (2012) é uma referência seminal sobre os esforços de criminologias em análises de crimes por regiões ao fazer uma análise do histórico de 16 anos de crimes nas cidades de Seattle e Washington nos EUA. O trabalho conclui que o crime está fortemente concentrado em "pontos críticos de crimes", sugerindo que ele pode ser combatido em grande escala se focado nesses pontos críticos.

POIs coletados via serviços de localização baseados em redes sociais (LSBN) é um tipo de dado que vem contribuindo para consolidar análises de crimes em pequena granularidade por regiões da cidade (SILVA et al., 2019). Por exemplo, em (YUAN; ZHENG; XIE, 2012) os autores mostraram que o uso de informações categóricas de POIs são úteis para traçar o perfil de atividades que caracterizam bairros. Os autores tiveram como resultado final a representação de bairros pela distribuição das atividades que os caracterizam enquanto essas atividades foram representadas pela distribuição da mobilidade urbana, ou seja, a frequência de pessoas nesses POIs. Mais recentemente, em (WANG et al., 2021) foi proposto um arcabouço de metódos para identificar as características funcionais de uma determinada região em uma cidade com base em POIs dessas áreas, utilizando o serviço OSM para a extração dos POIs.

Ainda sobre POIs é importante mencionar os trabalhos que propõem técnicas para definir o que é um ponto de interesse com base em informações que as pessoas submetem aos serviços Web e redes sociais, também conhecido como sensoriamento participativo. Mueller et al. (2017) analisam as preferências de gênero por locais usando sensoriamento participativo. Os autores investigaram se *checkins* de usuários em LSBNs podem ser usados para avaliar preferências de gênero por locais em diferentes regiões urbanas no mundo

físico.

Em (SILVA et al., 2017) é apresentado um técnica cuja finalidade é identificar Pontos de Interesses e, com base neles, reconhecer pontos turísticos em uma região. Esse processo associa cada dado (foto) a um par de coordenadas (longitude, latitude) representado por um ponto espacial. Com todos os dados associados, foi calculada a distância entre pares de pontos. Foi utilizada a fórmula Haversine (SINNOTT, 1984) para calcular a distância. Os pares próximos foram agrupados utilizando o critério de ligação "complete-linkage" (SORENSEN, 1948). Para obter os POIs, os autores utilizaram um modelo nulo com o objetivo de excluir grupos isolados. Esses grupos podem surgir devido a eventos singulares, logo, não caracterizam a realidade de uma cidade. Os grupos restantes foram identificados por meio de análise do número de compartilhamentos, utilizando técnicas de estatísticas. Por fim, o trabalho separou pontos turísticos dos pontos de interesse e constatou que os turistas possuem destinos em comum na cidade.

Mais relacionado ao nosso trabalho estão os estudos que mesclam fontes de dados oficiais, i.e., dados coletados ou requisitados por órgãos de governos, com POIs para predição de crimes em regiões urbanas. Alguns trabalhos mesclam fontes de dados oficiais a dados de redes sociais para analisar o quanto *posts* do *Twitter* estão correlacionados com a violência pública (TUCKER et al., 2021). Os autores descobriram que o fluxo de pessoas e turistas em certos pontos de um bairro estão associados ao grau de violência apenas durante os dias da semana, enquanto que os efeitos desses conflitos afetam os índices nos finais de semana. Em iranmanesh2020reading conclui-se que é criticamente importante investigar dados georreferenciados do Twitter para encontrar informações potenciais que estejam ocultas nos espaços urbanos.

Outros trabalhos como (WANG et al., 2017; HUANG et al., 2018) investigaram o quanto a adição de POIs, coletados em serviços de mapeamento (e.g Google Map, Open Street Map) incrementam dados oficiais (e.g, dados demográficos de senso) e fluxos urbanos como táxis e ônibus melhoraram a predição das taxas criminais. Em (BELESIOTIS; PAPADAKIS; SKOUTAS, 2018) é apresentado uma metodologia baseada em dados para mapear pontos críticos de crimes. Essa metodologia permite fornecer modelos de previsão que operam no melhor nível de granularidade, estimando o número de incidentes para áreas de diversos tamanhos. Este último estudo é o principal da nossa literatura. De início investigaremos se o modelo proposto pelos autores pode ser generalizado para outros locais utilizando outras fontes de dados. Já (CASTRO; RODRIGUES; BRANDAO, 2020) observou o quanto dados do serviço web "Onde fui Roubado" adicionados a dados oficiais melhorava a predição de índices de criminalidade.

Em contraponto aos estudos acima mencionados que mesclam dados oficiais a dados gerados por usuários em redes sociais (TUCKER et al., 2021; IRANMANESH; ATUN, 2020) e POIs (CASTRO; RODRIGUES; BRANDAO, 2020; BELESIOTIS; PAPADAKIS; SKOUTAS, 2018; HUANG et al., 2018; WANG et al., 2017), nesse trabalho investigamos

o quanto POIs unicamente podem predizer, ou seja, explicar taxas criminais por regiões das cidades. A Tabela 1 mostra os trabalhos mais relacionados com o nosso. Diferente desses trabalhos que mesclaram dados gerados por usuários com dados oficiais (e.g. censo demográfico, geográficos, sociais), nosso trabalho analisa o potencial na predição de crimes utilizando apenas dados gerados por usuários.

Tabela 1 – Trabalhos Relacionados

Trabalho Relacionado	**Granularidade de Análises**	**Principais técnicas**	**Objetivo**
Belesiotis, Papadakis e Skoutas (2018)	Cidade	Regressão, Caracterização de Dados	Prever crimes em áreas geográficas individuais
Castro, Rodrigues e Brandao (2020)	Bairros	Algoritmos de Classificação	Identificar padrões de comportamento criminoso e prever crimes.
Huang et al. (2018)	Cidade	Redes Neural	Desenvolver uma nova estrutura de previsão de crime - DeepCrime
Iranmanesh e Atun (2020)	Cidade	Regressão	Encontrar dados ocultos nos espaços urbanos usando dados da rede social Twitter.
Tucker et al. (2021)	–	Análise de variância, Algoritmo DBSCAN	Testar a confiabilidade dos dados do Twitter com geo-tag para estimar métricas da população.
Wang et al. (2017)	Bairro	Algoritmos de Regressão, GWNBR e Análises estatísticas	Compreender quais fatores causam maior taxa de criminalidade.
Este trabalho	Distritos Policiais	Regressão, Caracterização de Dados, Análise estatísticas	Prever índices criminais com apenas dados gerados por usuários na Web.

4 Metodologia

Esta Seção descreve a metodologia deste trabalho para análise de informações sobre violência em regiões urbanas utilizando dados extraídos de serviços web. As atividades apresentadas aqui visam alcançar os objetivos mencionados na Seção 1.1. A Figura 4 apresenta o fluxo de trabalho da metodologia descrita.

Pesquisa bibliográfica: Nesta primeira fase, foi realizada uma pesquisa de trabalhos na área da computação urbana. Primeiramente, buscamos entender como poderíamos contribuir com novos estudos e conhecimento na área a partir dos trabalhos existentes. Definindo nossa contribuição que se tornaram os objetivos do projeto, focamos em conhecer as técnicas e informações resultantes de trabalhos de outros autores que possam ser relevantes para o desenvolvimento do nosso trabalho. O resultado dessa etapa foi descrito em nosso referencial teórico na Seção 2.

Seleção de objetos de estudo: Poucos órgãos públicos no Brasil divulgam dados oficiais e estatísticas sobre criminalidade publicamente na Web. Dessa forma, nosso projeto teve de focar em regiões de estados no Brasil que fornecem esses dados. Essa etapa consistiu em estudar e definir a melhor opção de granularidade espacial para nosso trabalho. Além disso, foi selecionado todas as fontes de dados utilizados nesse projeto. A seleção teve como critério a disponibilidade aos dados, informações sobre espaço e tempo contida nos dados, e a existência de conteúdo gerado por usuário e dados oficiais sobre crimes no mesmo tempo e espaço. Descrevemos as fontes de dados selecionadas para este trabalho na Seção 4.1.1 e 4.1.2.

Aquisição dos dados: Após a seleção das fontes de dados, foram desenvolvidas técnicas para coletar os dados. Como a proposta do trabalho é realizar uma investigação com dados gerados por usuários extraídos da web,foram desenvolvidas técnicas com o uso de APIs oficiais das fontes de dados selecionadas. Com isso foi realizada a coleta dos dados que descrevemos com mais detalhes na Seção 4.1.2.1.

Pré-processamento dos dados: Os dados obtidos na fase anterior foram transformados em contagens de frequências, médias, distribuições empíricas, para investigação dos traços de pessoas em serviços web.

Desenvolvimento de modelos preditivos estatísticos: A partir dos dados quantificativos, iniciou a etapa da metodologia que consiste em criar modelos de predições estatísticos. Para isso foram utilizados algoritmos de aprendizado de máquina. Descrevemos com mais detalhes o modelo desenvolvido na Seção 4.2

Avaliação do Modelo: A última etapa do nosso trabalho consistiu em avaliar os resultados obtidos dos modelos de predição e analisar esses resultados tal como descrevemos na Seção 5.

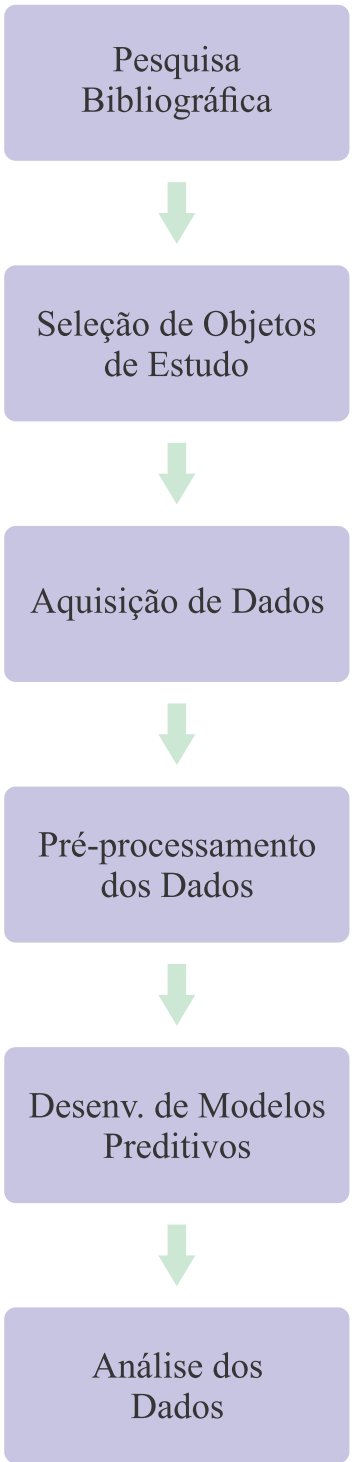


Figura 4 – Fluxograma Da Metodologia Do Projeto

4.1 Coleta de dados

Nesta Seção descrevemos as bases de dados e a metodologia de processamento desses dados para o uso em modelos de predição. Primeiramente, descrevemos os dados sobre índices de criminalidade que foram extraídos de fontes de dados oficiais. A seguir, descrevemos a metodologia para extração de POIs em um serviço Web de mapeamento urbano e sua aplicação na construção da nossa base de dados.

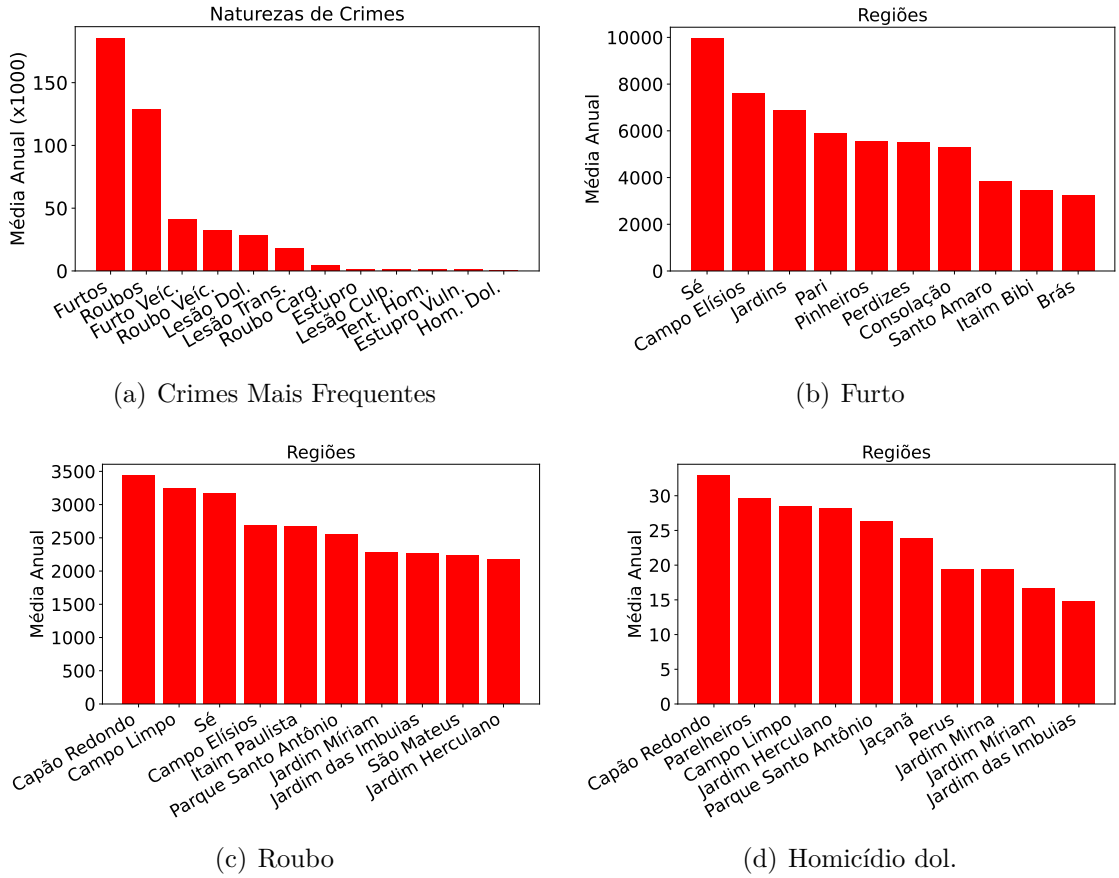
4.1.1 Índices de Crimes Oficiais

Os dados de índices criminais foram extraídos da Secretaria de Segurança Pública do Estado de São Paulo (SSP-SP, 2021). Esses dados são divulgados mensalmente e organizados em vinte e três categorias de crimes contendo a contagem de ocorrências registradas por regiões do estado desde 2001. Todos os crimes estão listados no Apêndice A, tal como, as abreviações dessas utilizadas ao longo deste trabalho em figuras e tabelas. A área geográfica de cada região consiste na delimitação dos *distritos policiais* definidos em lei pelo governo do estado (São Paulo, 2015). Essas regiões são áreas delimitadas por ações estratégicas de segurança pública e não seguem propriamente as definições de bairros conhecidas dessas cidades.

O foco de nosso estudo foi nos distritos policiais da cidade de São Paulo, a capital do estado, por se tratar de regiões com os maiores índices criminais. São Paulo contém noventa e três distritos policiais e foi possível coletar dados de 88 regiões dos últimos nove anos (2012 - 2020).[1] A Figura 5-a mostra as doze categorias de crimes mais frequentes em nossa base de dados, considerando a média anual de cada categoria no referido período para reduzir o impacto dos anos com a menor e a maior ocorrência de crimes. Como pode-se observar furtos e roubos são os crimes mais frequentes com uma média de 185.108 e 132.072 ocorrências por ano. Não obstante, homicídio doloso, e.g., assassinatos intencionais, ocupa a décima segunda posição com média de 784 ocorrências por ano[2], o que é considerado um índice alto e alarmante dado a gravidade dessa categoria e o impacto na vida de familiares das vítimas e sociedade.

A Figura 6-a apresenta as naturezas criminais com maior número de ocorrências para cada uma das regiões de São Paulo. Pode-se observar que Furto e Roubo, os dois crimes mais frequentes da nossa base de dados, predominam em todo o território. Furto é a mais frequente ocupando 60% das regiões enquanto Roubo ocupa 35%. As Figuras 5 (b-d) e 6 (b-d) mostram as regiões mais violentas para os dois crimes mais frequentes (furto e roubo) mais homicídios. Pode-se observar que furto (Figuras 5-b e 6-b) é mais frequente nas regiões centrais da cidade (e.g., Sé, Campos Elísios, Jardins e Pari) com taxas anuais entre 4 mil e 12 mil ocorrências registradas. São áreas que se destacam pela alta proporção

[1] Utilizamos esse período para compatibilizar com os dados de POIs.
[2] A média do número anual de homicídios é ligeiramente superior e correlacionado com as ocorrências.



(a) Crimes Mais Frequentes

(b) Furto

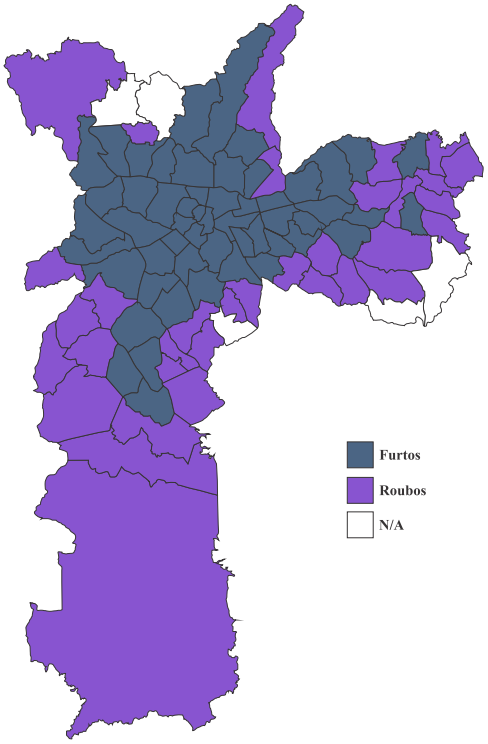
(c) Roubo

(d) Homicídio dol.

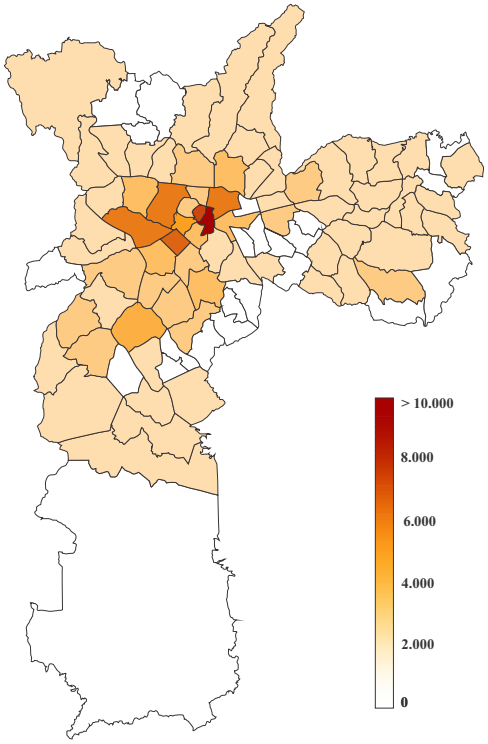
Figura 5 – Médias Anuais Por Categorias De Crimes Na Cidade De São Paulo Entre 2012-2020: (a) Crimes Mais Frequentes, (b-d) Regiões Com Ocorrências Mais Frequentes De Furto, Roubo e Homicídio Doloso Respectivamente.

de chefes de famílias alfabetizados (99,3%), de alta renda (8,6%) com algumas regiões sem registros de homicídios dolosos(NERY; SOUZA; ADORNO, 2019). Contudo, à medida que crimes aumentam o grau de violência, como é o caso de roubos (Figuras 5-c e 6-c), eles diminuem nas regiões centrais em direção às regiões periféricas. Logo, os crimes mais violentos, e.g., homicídio doloso, se concentram nas regiões periféricas como pode ser observado nas Figuras 5-d e 6-d, onde as regiões Capão Redondo e Parelheiros lideram esse índice com média superior a trinta e três homicídios dolosos anuais. Essas regiões têm histórico de altos índices de homicídios devido a contrastes sociais e em especial ao tráfego de drogas (ADORNO; NERY, 2019).

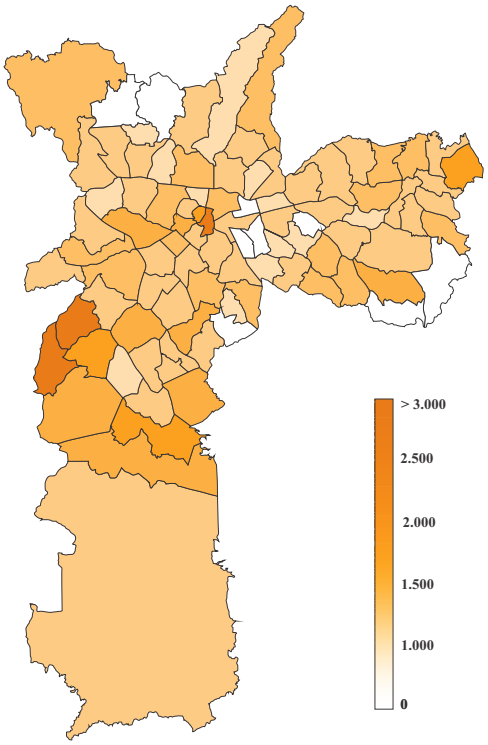
A Figura 7 mostra o mapa de calor da cidade de São Paulo dos piores aos melhores índices de desenvolvimento humano. Os dados são da empresa URBIT que coleta, organiza e facilita o acesso às informações de diversas cidades do Brasil. Como pode-se observar, os bairros centrais, ou próximos do centro, possuem os maiores IDH como Moema com índice de 0.8919, Consolação com 0.8681, Campos Elísios com 0.79 dentre outros. Conforme seguimos do centro da cidade para os bairros periféricos os índices vão diminuindo. Marsilac e Parelheiros, localizados no extremo sul da cidade de São Paulo, apresentam os piores índices com valores de 0.6889 e 0.7219 respectivamente.



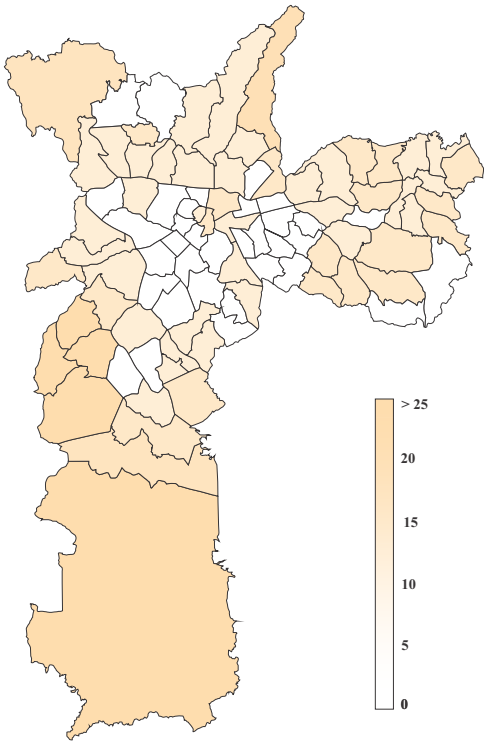
(a) Crimes Predominante



(b) Furto



(c) Roubo



(d) Homicídio dol.

Figura 6 – Distribuição De Categorias de Crimes Na Cidade De São Paulo Entre 2012-2020: (a) Crimes Predominantes Por Região, (b-d) Regiões Com Ocorrências Mais Frequentes De Furto, Roubo e Homicídios Doloso Respectivamente.

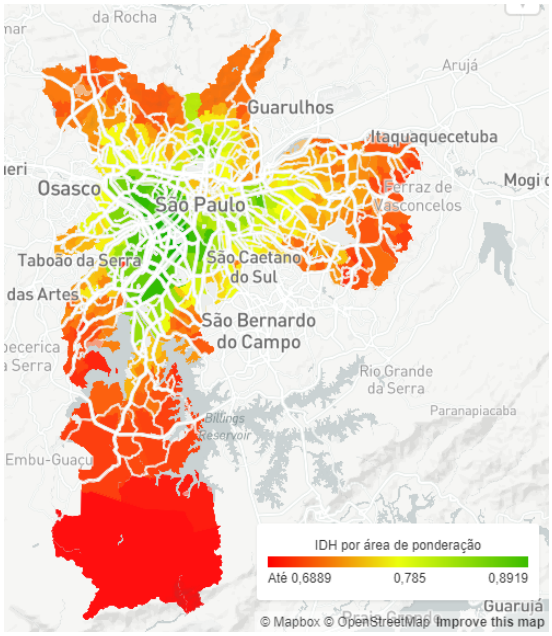


Figura 7 – IDH Dos Bairros De São Paulo. Fonte: https://urbit.com.br/mapa/idh-sp

4.1.2 Pontos de Interesse (POIs)

Aplicações georreferenciadas, i.e, Google Maps e Open Street Map, podem ser utilizadas para diversas funcionalidades como turismo, lazer e serviços úteis. Quando uma pessoa se interessa em obter informação sobre um local específico, diz-se que ela está à procura de um Ponto de Interesse (POI). Geralmente, esses pontos são representados em aplicações de mapeamento geográfico por diversos ícones diferentes associados à sua respectiva categoria. Essas categorias indicam, funcionalmente, o tipo de interesse coletivo por aquele determinado local. Esses locais podem ser classificados com mais de um interesse final, i.e, entretenimento, escolas, restaurantes, serviços públicos e outros (WANG et al., 2021).

4.1.2.1 Extração de POIs

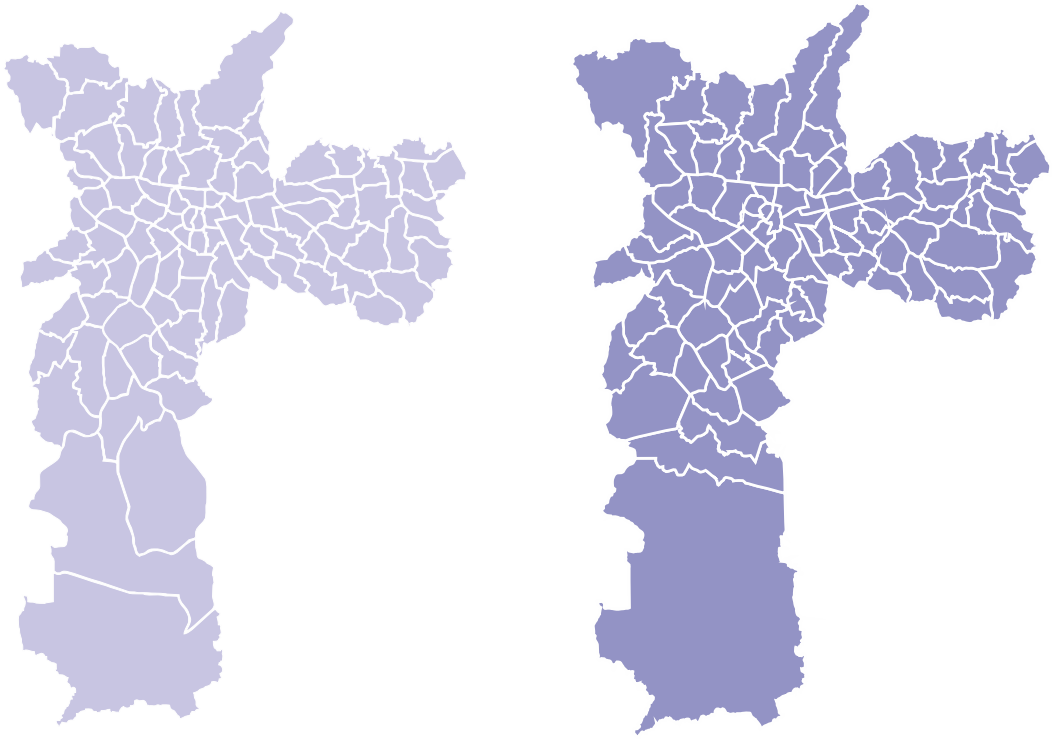
Como mostrado na Seção 2.1.3, existem diversas aplicações de mapeamento geográfico via sensoriamento colaborativo, i.e, Google Maps, Open Street Maps, Foursquare, Wikimapia. Essas aplicações são alimentadas por diversos usuários a todo segundo. Para este trabalho escolhemos a ferramenta OSM, pois trata-se da única aplicação, dentre os citados, cujo mapas, dados e metadados são dados abertos[3]. O Open Street Maps disponibiliza todos os seus dados[4], isto é, do planeta inteiro em blocos anuais. Os dados do ano corrente são atualizados mensalmente.

Dessa forma, foi feito o *download* dos dados dos anos de 2012 a 2020. Esses dados são disponibilizados publicamente com tamanhos que variam de 23 a 97 *gigabytes* entre

[3] Dados disponíveis para uso e publicação por qualquer pessoa.
[4] https://planet.openstreetmap.org/planet/

os anos 2012 a 2020 respectivamente, considerando a compressão no formato *BZip2*. Por serem arquivos grandes com dados do planeta inteiro, primeiramente foi realizado um recorte geográfico espacial a fim de reduzir o processamento nas etapas seguintes. Dessa forma, foi utilizada a ferramenta *Osmium Tool*, descrita na Seção 2.2, para reduzir os dados do planeta apenas aos dados da região de interesse. O processamento dos dados é computacionalmente intensivo e seis máquinas virtuais da Amazon Web Services[5] foram utilizadas para esse processamento em tempo hábil.

Nosso objetivo é extrair POIs de regiões de distritos policiais, mas no OSM não há nenhum dado que associe POIs aos distritos. Por outro lado, essas regiões não necessariamente compreendem um conjunto de bairros. Elas podem ser formadas por bairros inteiros ou partes de outros bairros. Na Figura 8 você pode conferir como a cidade de São Paulo é dividida espacialmente por bairros e pelos distritos policiais. Devido a incompatibilidade espacial, foi necessário realizar o mapeamento dos distritos policiais para mapear os POIs para cada distrito policial.



(a) Bairros (b) Distritos Policiais

Figura 8 – Divisão Da Cidade De São Paulo Por Bairros e Distritos Policiais

O mapeamento foi realizado de forma manual com ajuda do Diário Oficial do Estado de São Paulo. O estado de São Paulo na resolução SSP 52, de oito de maio de 2015 "Altera e compatibiliza os limites territoriais das áreas de atuação da Polícia Civil e Polícia Militar no município de São Paulo e define procedimentos relativos a futuras alterações destes

[5] https://aws.amazon.com/pt/

limites pelos órgãos envolvidos" (São Paulo, 2015). O documento descreve por nomes de ruas, pontos de referências, direções cardeais e outros o limite geográfico de cada uma das regiões dos distritos policiais. Confira logo abaixo, como exemplo, a descrição da região de Bom Retiro.

> Tem início no cruzamento da Rua Florêncio de Abreu com a Rua Mauá, segue por esta (inclusive) até encontrar o Viaduto General Couto de Magalhães; daí, deflete à direita e segue por esta segue por este (inclusive) até encontrar o leito da via férrea; daí, segue por esta (inclusive) até alcançar o Viaduto Engenheiro Orlando Murgel; daí, segue por este (exclusive) até encontrar a Avenida Rudge; daí, segue por esta (inclusive) até encontrar a Praça Torquato Tasso Netto; daí, segue por esta (exclusive) até encontrar a Ponte da Casa Verde; daí, segue por esta (inclusive) até alcançar o leito do Rio Tietê; daí, deflete à direita e segue por este até alcançar a Ponte das Bandeiras; daí, deflete à direita e segue por esta (inclusive) até encontrar a Avenida Santos Dumont; daí, segue por esta (inclusive) até a Avenida Tiradentes, segue por esta (inclusive) até o Viaduto Engenheiro Romerio Zander; daí, segue por este (inclusive) até encontrar o cruzamento da Rua Florêncio de Abreu com a Rua Mauá, ponto inicial do perímetro (São Paulo, 2015).

A partir dessa descrição, utilizamos o serviço *Google Maps*, para coletar o conjunto de coordenadas geográficas (latitude, longitude) que compreendem todo o perímetro do distrito policial. Contudo, algumas ruas e outras localidades tiveram seus nomes alterados ao longo dos anos. Além disso, há pontos de referência descritos no diário oficial que não são encontrados na ferramenta do google, e.g., linhas de energia. A falta de tais informações impossibilitaram realizar o mapeamento das regiões: Parque Bistrol, Jaraguá, Cidade Tiradentes e Parque São Rafael, ambas situadas nas extremidades do município de São Paulo. Dessa forma, nosso estudo foi realizado com oitenta e oito de noventa e três regiões de distritos policiais. Com todas as regiões mapeadas, foi utilizado novamente a ferramenta *Osmium* para recortar do OSM todas as regiões de distritos policiais de São Paulo. Em seguida, utilizamos a API *Osmosis* (ver a Seção 2.2) para extrair as categorias de POIs de cada região. Os dados foram quantificados e utilizados para montar a nossa base de dados que descrevemos na próxima Seção.

Em resumo, utilizamos um conjunto de APIs do OSM e *Google Maps* para mapear oitenta e oito distritos policiais da cidade de São Paulo, incluindo o crescimento gradativo do volume desses POIs de 2012 até 2020, que é o período total oferecido pelas APIs do OSM. As descrições sobre delimitações de cada região foram obtidas no caderno do Estado de São Paulo (São Paulo, 2015). Contudo, tais descrições não contém as coordenadas geográficas necessárias para o mapeamento via API do OSM. Para obtê-las seguimos as

descrições de delimitações, i.e. nome de ruas, pontos de referências, direções, capturando as coordenadas via Google Maps[6].

4.1.2.2 POIs por Distritos Policiais

Dado os procedimentos mostrados na Seção anterior, coletamos 441.059 POIs cadastrados no OSM entre os anos de 2012 a 2020 na cidade de São Paulo, organizados em 107 categorias[7]. Todas as categorias estão listadas no Apêndice B. A Figura 10-a mostra a categoria mais frequente em nossa base de dados para cada uma das regiões, enquanto a Figura 9-a mostra as dez categorias mais frequentes em nossa base de dados. Notavelmente estacionamentos é a mais frequente (51,4%), e isso ocorre provavelmente devido ao uso massivo do OSM por motoristas com forte demanda por vagas de estacionamento na cidade. No entanto, há outras categorias representativas não relacionadas diretamente a veículos como escolas (6,06%), templos (5,82%) e restaurantes (2,52%). Vamos avaliar o potencial para predição de crimes de todas as categorias de POIs, pois esperamos que algumas dessas tenham informações preditivas para crimes, ainda que estejam em menor porcentagem.

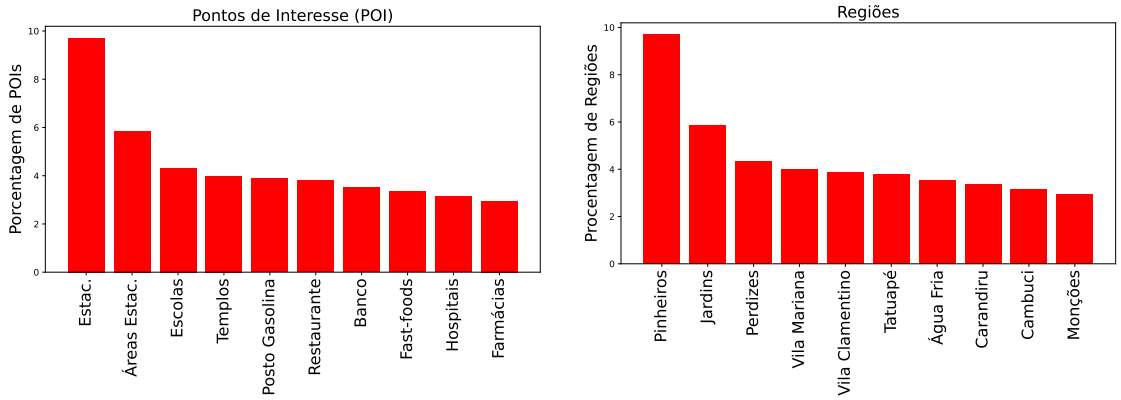
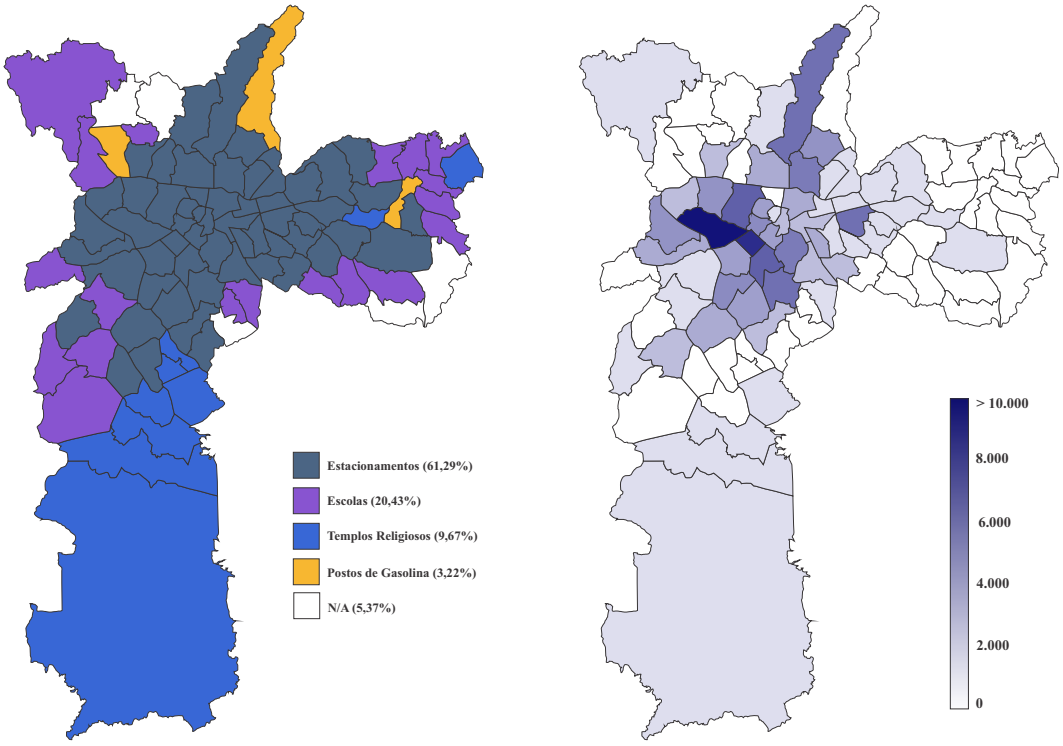


Figura 9 – Pontos de Interesses (POIs) Coletados Do serviço OSM: (a) Dez POIs Mais Frequentes e (b) Regiões Que Acumulam o Maior Volume De POIs.

A Figura 10-b mostra as regiões (distritos policiais) da cidade de São Paulo que acumulam maior volume de POIs, enquanto que a Figura 9-b mostra a porcentagem das 10 regiões com o maior volume de POIs em nossa base de dados. Como esperado, a maior parte dessas regiões fazem parte ou são vizinhas do centro de São Paulo onde há maior atividade econômica e por conseguinte movimentação diária de pessoas. A região de Pinheiros concentra o maior volume de POIs chegando a 11.067 (9,7%), seguido por Jardins com 6.667 (5,8%), Perdizes 4.923 (4,3%) e Vila Mariana 4.543 (3,98%). As áreas com a menor quantidade de POIs estão localizadas na zona leste de São Paulo. A Zona Leste

[6] https://www.google.com.br/maps
[7] https://wiki.openstreetmap.org/wiki/Key:amenity

é vista como uma região periférica, no qual sua população é maioria de baixa renda. As regiões da zona leste apresentaram pouco ou nenhum volume de POIs para as categorias cuja atividade principal é cultura e entretenimento, i.e, teatros, cinemas, escolas de música, restaurantes, sorveterias, fast-foods e outros.



(a) (b)

Figura 10 – Pontos de Interesses (POIs) Coletados Do Serviço OSM (a) Pontos de Interesses Mais Frequentes e (b) Regiões Que Acumulam o Maior Volume De POIs.

4.2 Desenvolvimento do Modelo

O nosso principal objetivo neste trabalho é investigar o potencial de POIs para explicar taxas anuais de ocorrências de crimes por categoria e por região da cidade. Para isso, propomos modelos de regressão em que a taxa anual de ocorrências para uma determinada categoria de crime por região da cidade seja a variável a ser predita (y). Por sua vez, POIs serão utilizados como características para essa predição e levaremos em consideração uma matriz X das cento e sete categorias de POIs (colunas) para as regiões da cidade (linhas) coletadas do serviço OSM.

Contudo, precisamos construir um modelo para predizer a categoria de crime em uma região alvo a, desconsiderando essa região nos valores a serem preditos y e na matriz de POIs X para fins de avaliação do modelo. Nesse sentido, adotamos a metodologia *leave out one* que consiste em predizer a taxa anual de crimes para uma região utilizando dados

das outras regiões. Mais formalmente construímos modelos com o formato:

$$\hat{y}_a = M(\{y\} \setminus y_a, X) \tag{4.1}$$

onde $\hat{y}_a$ é uma categoria de crime de uma região alvo a ter sua taxa anual estimada, a função M representa diferentes métodos de regressão a serem utilizados para o treinamento do modelo. Por sua vez, y_a, que é a taxa anual real de uma categoria de crime na região alvo, é excluída das taxas de crimes y e POIs X, utilizadas no treinamento do modelo. Em outras palavras, retiramos a região alvo dos dados para realizar a sua predição (dados de teste), ao passo que as outras regiões foram utilizadas para treinar o modelo.

Avaliamos o desempenho de diferentes modelos por categoria de crime, ou seja, medimos o erro do modelo para predizer a taxa anual de uma categoria de crime considerando como amostras as regiões da cidade. Esse desempenho foi quantificado pelas métricas média do erro absoluto (MAE), média do erro relativo (MRE) e o índice R2. Especificamente essas métricas foram calculadas da seguinte forma:

$$MAE = \frac{\sum_i |y_i - \hat{y}_i|}{n}; MRE = \frac{\sum_i |y_i - \hat{y}_i|}{\sum_i y_i}; R2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{4.2}$$

onde y_i e $\hat{y}_i$ representam os valores atual e estimados para a taxa anual de um crime na i-ésima região, ao passo que $\bar{y}$ representa a média da taxa de crime considerando todas as n regiões para a construção do modelo, i.e., oitenta e oito distritos policiais da cidade de São Paulo.

O R2 é um coeficiente de determinação utilizado para avaliar modelos de regressão. A métrica pode obter valores entre a faixa 0.0 e 1.0. Valores mais próximos a 1.0 indicam que o modelo pode explicar a variabilidade dos dados, ao passo que valores próximos a 0.0 indicam o inverso. A métrica pode também atingir valores negativos que demonstram um modelo péssimo para explicar os dados. O erro médio absoluto pode obter quaisquer valores positivos, sendo o melhor resultado próximo a 0.0 e menor que os valores reais. Já o erro médio relativo é uma métrica utilizada para calcular a porcentagem de erro do valor real.

Nossos experimentos foram realizados utilizando a biblioteca *scikit-learn* da linguagem *python*. Utilizamos para construção do modelo M da Equação 4.1 os métodos regressão linear (RL), floresta aleatória (FA) e vetores de suporte a regressão, ou *support vector regression* (SVR), implementados nessa biblioteca. Na Seção 2.3, descrevemos os fundamentos teóricos desses métodos. Em suma, RL busca encontrar correlações entre a média de uma variável dependente com outra ou várias variáveis (GROSS, 2012). A FA combina um conjunto de preditores de árvores de modo que cada árvore depende dos valores de um vetor aleatório da amostragem com a mesma distribuição (BREIMAN, 2001). Por fim, SVR, busca predizer um valor real após traçar duas retas paralelas, chamadas de limites. O modelo ainda traça uma reta linear entre as duas outras retas a fim de ajustar seus valores (DRUCKER et al., 1997).

5 Resultados

Nesta Seção, comparamos o desempenho de diferentes métodos de regressão. A seguir, discutimos as categorias de crimes com os melhores resultados preditivos, ou seja, os crimes que podem ser melhor explicados a partir de POIs.

5.1 Desempenho de Diferentes Métodos

A Tabela 2 mostra o desempenho dos métodos de regressão aplicados às nove categorias de crimes mais frequentes e a categoria homicídio doloso, ambos do ano de 2020. É importante observar primeiramente o desempenho dos três métodos de regressão utilizados. Floresta aleatória é o método que obteve melhor desempenho pois alcançou maiores índices R2 e menores erros absolutos e relativos. O R2 alcançou valores positivos e não próximos a zero para oito dentre os dez crimes mais frequentes, o que indica que a floresta aleatória os explica razoavelmente em relação a uma simples média das variáveis explicativas (X) e, por conseguinte, obtém erros menores que os demais métodos. Tomando o R2 como referência, SVR foi o segundo método em termos de desempenho e regressão linear foi o pior método. Ambos alcançaram valores positivos apenas para quatro e três categorias de crimes respectivamente. Logo, nota-se que os erros, em geral, diminuem à medida que o R2 cresce, e esses erros são menores para os modelos com floresta aleatória.

Tabela 2 – Desempenho Dos Métodos De Regressão Aplicados Aos Crimes Mais Frequentes Do Ano 2020 e Homicídios Dolosos: Floresta Aleatória (FA), *Support Vector Regression* (SVR) e Regressão Linear (RL).

Crimes	R2			MAE			MRE (%)		
	FA	SVR	RL	FA	SVR	RL	FA	SVR	RL
Furto	**0,56**	0,03	-0,64	**650,63**	762,20	1072,8	0,39	0,40	0,70
Roubo	0,10	-0,33	-1,2	**513,39**	605,49	731,69	0,46	0,56	0,68
Furto veíc.	0,02	-0,25	-0,15	122,37	130,04	127,80	0,54	0,54	0,52
Lesão dol.	0,28	0,14	-0,47	91,20	88,72	118,04	0,42	0,44	0,60
Roubo veíc.	0,14	-1,0	-0,82	69,61	99,42	91,38	0,95	1,36	1,27
Lesão trans.	0,17	-0,24	-0,003	29,55	38,31	33,00	0,34	0,42	0,38
Roubo carg.	0,13	-0,12	0,04	21,15	23,37	24,44	1,42	1,64	1,99
Estupro vuln.	0,53	0,22	0,08	7,17	10,19	10,92	0,73	1,07	1,13
Lesão culp.	-0,14	-0,15	-1,0	8,1	7,31	9,4	1,34	0,84	1,29
Homicídio dol.	0,26	0,04	0,06	4,2	4,97	4,91	0,93	1,22	1,28

Agora focamos nos resultados da regressão com floresta aleatória, que obteve melhor desempenho, para analisar as categorias de crimes. Nesse sentido, consideramos que o modelo é útil para explicar crimes quando, além do R2 positivo, apresentam erros relativos médios inferiores a 100%. Isso indica que para a maioria das estimativas da taxa de um

crime o erro é menor do que a taxa real em valores absolutos. Excluímos do cálculo do erro relativo médio, as regiões com a taxa anual de crime zero, visto que esses valores impossibilita o cálculo.

Para exemplificar os casos em que modelos são úteis analisamos os erros para furto, roubo e homicídio doloso mostrados na Figura 5. Essa última categoria é classificada como um dos crimes mais violentos no país (NEV-USP, 2021), enquanto as outras duas são as categorias mais frequentes, embora sejam crimes menos violentos. Furto tem erro absoluto (relativo) médio de 650.63 (39%) ocorrências por ano. É um erro razoavelmente baixo, se compararmos com a média anual (Figura 5-b) que supera duas mil ocorrências anuais em todas as regiões de São Paulo. Por sua vez, roubo tem erro de 513.39 (46%), com uma média anual (Figura 5-c) superior a mil e quatrocentas ocorrências. Já a média de homicídio doloso é cerca de dez ocorrências anuais para um erro do modelo de 4.2 (93%), ver Figura 5-c. Levando em consideração as médias anuais de ocorrências mostradas na Figura 5 b-c para as regiões mais violentas, os erros dos modelos tornam-se ainda menores. Logo, nota-se que os erros absolutos médios dos modelos estão bem abaixo das taxas de crimes reais. Nossa observação, portanto, é que modelos com erros em níveis razoáveis como os apresentados acima são bem indicados por MRE abaixo de 100% e R2 positivos.

5.2 Importância de POIs

Nesta seção, analisamos a importância de diferentes categorias de POIs para predição de crimes por região, baseados no melhor método de regressão obtido, i.e., floresta aleatória. Nesse sentido, selecionamos sete categorias de crimes em que esse método pôde explicar variação da taxa anual de crimes via POIs, conforme o critério observado na Seção anterior. A Tabela 3 mostra cada uma das categorias de crimes seguido dos quatro POIs mais importantes para predição com suas respectivas porcentagens de relevância indicadas entre parênteses.

Do total de cento e sete categorias de POIs do serviço OSM, onze aparecem entre as quatro mais importantes para os modelos apresentados. Escola é a categoria predominante aparecendo entre a primeira e a segunda ordens de importâncias para os modelos. Por outro lado, estacionamentos que são as categorias de POIs mais frequentes na maioria das regiões de São Paulo ocupam a primeira e segunda ordem de importância para apenas dois crimes (roubo de veículos e estupro). Isso indica que a quantidade desproporcional entre o número de POIs por categoria não impactam decisivamente na importância desses para predizer crimes. No entanto, as explicações sobre a importância de alguns POIs para determinados crimes não são triviais e requer a análise de especialistas experientes em criminologia e urbanismo. Por exemplo, escola é a categoria de POI, notavelmente, mais importante para predizer lesão dolosa, mas explicações intuitivas para esse relacionamento podem ser complexas. Possivelmente, há outras características espaciais associadas

Tabela 3 – Relação De Quatro Categorias De POIs Mais Importantes Para Predição Da Taxa Anual De Crimes (Sete Categorias De Crimes Com os Melhores Modelos).

Crime	Ordem de importância com sua respectiva porcentagem (%)			
	1a.	2a.	3a.	4a.
Furto	Táxi (66%)	Escolas (7%)	Bares (6%)	Estac. (3%)
Roubo veíc.	Escolas (26%)	Estac. (20%)	Farmácias (6%)	Áreas de estac. (4%)
Roubo	Escolas (29%)	Ônibus (27%)	Templos (5%)	Estac. (3%)
Estupro vuln.	Estac. (30%)	Escolas (12%)	Bancos (11%)	Táxi (8%)
Lesão trans.	Escolas (16%)	Gasolina (16%)	Fast-foods (12%)	Farmácias (6%)
Lesão dol.	Escolas (31%)	Ônibus (12%)	Estac. (8%)	Áreas de estac. (5%)
Homicídio dol.	Gasolina (29%)	Escolas (16%)	Estac. (11%)	Ônibus (7%)

à frequência de escolas em algumas regiões que levam a relações indiretas com crimes de lesão dolosa. Por sua vez, furto é uma categoria de crime que pode ser facilmente relacionada a vários POIs que expressam característica no espaço, mas pontos de táxi unicamente ocupam 66% da importância para explicar esse crime. Essas questões sobre características espaciais relacionadas aos POIs serão investigadas em trabalhos futuros.

5.3 Impacto do Aumento de POIs

Finalmente, é importante analisar o impacto no aumento gradativo da quantidade de POIs no espaço ao longo do tempo. Para essa análise observamos o desempenho do melhor modelo de regressão (floresta aleatória) considerando POIs existentes para cada região iniciando do ano de 2012 até 2020. Esse é o período em que o serviço OSM disponibiliza desde então dados sobre POIs.

Tabela 4 – Impacto Do Aumento De POIs Entre Anos 2012-2020 No Erro Dos Modelos: O Cabeçalho Mostra o Percentual De POIs Em Relação a 2020 e As Linhas Mostram A Média Do Erro Absoluto (MAE) Para A Taxa Anual Dos Crimes, Considerando Os Modelos De Regressão Com Os Melhores Desempenhos.

Crime	2012 (6%)	2013 (8%)	2014 (10%)	2015 (12%)	2016 (39%)	2017 (56%)	2018 (67%)	2019 (86%)	2020 (100%)
Furto	1008,81	749,40	1013,02	949,55	896,67	717,21	791,01	913,43	650,63
Roubo	429,78	490,28	680,97	681,22	663,12	551,17	480,00	504,00	513,39
Lesão dol.	154,13	160,00	126,00	106,34	111,96	96,05	81,23	82,62	91,20
Roubo veíc.	228,98	311,20	274,74	222,80	229,12	179,77	141,03	113,17	69,61
Lesão trans.	111,75	97,23	96,39	76,93	66,97	51,30	42,02	41,12	29,55
Estupro vuln.	15,13	16,23	11,78	11,04	3,94	7,34	7,05	8,29	7,17
Homicídio dol.	10,00	8,08	7,67	7,35	5,64	4,21	4,01	3,99	4,20

A Tabela 4 mostra em seu cabeçalho a porcentagem de POIs a partir de 2012 de forma cumulativa até atingir o total de POIs observado em 2020 (100%). Os dados dessa tabela representam em cada linha a média do erro absoluto (MAE) para as sete categorias de crimes em que os modelos de regressão obtiveram melhores desempenhos, conforme discutido na Seção anterior. É notável a tendência de ganho em desempenho, i.e., redução do erro, à medida em que se aumenta o volume de POIs, a despeito de algumas perdas

em anos isolados. Os ganhos mais expressivos ocorrem nos anos de 2016 e 2020 onde as reduções nos erros alcançam entre 28% até 72% em relação ao ano anterior. De modo geral, observamos uma redução na média dos erros absolutos de pelo menos 4% por ano.

6 Conclusão

Neste trabalho investigamos o potencial de POIs para explicar taxas anuais de ocorrências de crimes por categoria e por região da cidade. Os trabalhos da literatura sobre computação urbana e análise de crimes tipicamente utilizam POIs como um incremento à dados oficiais sobre demografia e censo urbano para predizer crimes. O nosso desafio nesse trabalho foi predizer crimes unicamente baseado em POIs extraídos de fontes de dados abertas da Internet como um recurso adicional na ausência ou atraso de dados oficiais. Nesse sentido, conduzimos uma investigação baseada em oitenta e oito distritos policiais (regiões) da cidade de São Paulo onde relacionamos as ocorrências de crimes com os POIs existentes em cada região. Quantificamos essa relação com modelos de regressão e análises da média de erros absolutos, média de erros relativos e o índice R2.

Nossos experimentos evidenciam que o uso de POIs unicamente podem explicar razoavelmente algumas categorias de crimes. Observamos que crimes mais frequentes em regiões centrais da cidade como furto e roubo, onde a quantidade de POIs é maior, obtiveram melhores desempenhos via modelos de regressão com média de erros inferiores a 46% dos dados oficiais e índices R2 que alcançam 0,56. Outras categorias como homicídio doloso, que ocorrem em regiões com menor quantidade de POIs, ainda podem ser explicadas com erros inferiores a 93% dos dados oficiais. O POI escola, curiosamente, predomina como o mais importante para predições, aparecendo entre a primeira e a segunda ordens de importâncias para a maioria dos modelos. Observamos também ganhos de desempenho dos modelos com a redução de erros absolutos em pelo menos 4% anualmente à medida em que ocorreu o aumento na quantidade de POIs entre os anos de 2012 a 2020. Logo, esses modelos podem se tornar mais eficientes futuramente.

Trabalhos futuros incluem a identificação de características das cidades e funcionalidades de regiões via um conjunto de POIs visando melhorar o desempenho dos modelos de regressão analisados neste trabalho e outros tipos de modelos a serem analisados, assim como explicar a importância de diferentes POIs para esses modelos.

7 Publicações

Abaixo são listadas as publicações científicas do autor deste trabalho:

- SOUSA, D. da S.; FEITOSA, M. P. F.; GONÇALVES, G. D. Relações entre crimes e o espaço urbano: Um estudo de caso baseado em pontos de interesses extraídos da web. In: SBC. *Anais do V Workshop de Computação Urbana.* [S.l.], 2021. p. 196–208
- SOUSA, D.; GONÇALVES, G. Um estudo sobre compartilhamentos entre contatos via d2d em serviços de armazenamento pessoal em nuvem. In: SBC. *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí.* [S.l.], 2020. p. 244–251.
- SOUSA, D.; GONÇALVES, G. Análise do uso de comunicação dispositivo a dispositivo para transferência de dados compartilhados em serviços de armazenamento pessoal em nuvem. *Revista de Sistemas e Computação-RSC*, v. 10, n. 3, 2020.
- ROCHA, S. C. ; SOUSA, D. S. ; GONCALVES, G. D. ; LEAL, I. H. . Guardião: Um Sistema de Informação para Apoiar Decisões Estratégicas na Área de Segurança Pública. In: *Simpósio de Sistemas de Informação*, Picos, 2019.

Referências

ADORNO, S.; NERY, M. B. Crime e violências em são paulo: retrospectiva teórico-metodológica, avanços, limites e perspectivas futuras. *Cadernos Metrópole*, SciELO Brasil, v. 21, n. 44, p. 169–194, 2019. Citado 3 vezes nas páginas 13, 26 e 32.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 22.

ASMOLOV, G. Crowdsourcing as an activity system: Online platforms as mediating artifacts. In: *Sintelnet WG5 Workshop on Crowd Intelligence: Foundations, Methods and Practices*. [S.l.: s.n.], 2014. Citado na página 20.

BECKER, K. L.; KASSOUF, A. L. Uma análise do efeito dos gastos públicos em educação sobre a criminalidade no brasil. *Economia e Sociedade*, SciELO Brasil, v. 26, n. 1, p. 215–242, 2017. Citado na página 26.

BELESIOTIS, A.; PAPADAKIS, G.; SKOUTAS, D. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, ACM New York, NY, USA, v. 3, n. 4, p. 1–31, 2018. Citado 4 vezes nas páginas 13, 19, 27 e 28.

BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001. Citado 2 vezes nas páginas 24 e 39.

CASTRO, U. R.; RODRIGUES, M. W.; BRANDAO, W. C. Predicting crime by exploiting supervised learning on heterogeneous data. In: *ICEIS (1)*. [S.l.: s.n.], 2020. p. 524–531. Citado 3 vezes nas páginas 14, 27 e 28.

DRUCKER, H. et al. Support vector regression machines. *Advances in neural information processing systems*, Morgan Kaufmann Publishers, v. 9, p. 155–161, 1997. Citado 2 vezes nas páginas 24 e 39.

FEIJÓ, C.; VALENTE, E. As estatísticas oficiais e o interesse público. *Bahia Análise & Dados, Salvador*, v. 15, n. 1, p. 43–54, 2005. Citado na página 19.

GOOGLE, S. *Google: Ajuda com Waze - Como o Waze funciona?* 2021. Disponível em: https://support.google.com/waze/answer/6078702?hl=pt-BR. Acesso em 26 de out. 2021. Citado na página 20.

GROSS, J. *Linear regression*. [S.l.]: Springer Science & Business Media, 2012. Citado 2 vezes nas páginas 24 e 39.

HAYKIN, S. *Neural networks and learning machines, 3/E*. [S.l.]: Pearson Education India, 2010. Citado na página 23.

HOWE, J. et al. The rise of crowdsourcing. *Wired magazine*, v. 14, n. 6, p. 1–4, 2006. Citado na página 20.

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*. [S.l.: s.n.], 2018. p. 1423–1432. Citado 3 vezes nas páginas 14, 27 e 28.

IRANMANESH, A.; ATUN, R. A. Reading the urban socio-spatial network through space syntax and geo-tagged twitter data. *Journal of Urban Design*, Taylor & Francis, v. 25, n. 6, p. 738–757, 2020. Citado 2 vezes nas páginas 27 e 28.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007. Citado na página 24.

MASI, C. M. et al. Neighborhood economic disadvantage, violent crime, group density, and pregnancy outcomes in a diverse, urban population. *Social science & medicine*, Elsevier, v. 65, n. 12, p. 2440–2457, 2007. Citado na página 26.

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: An artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 22.

MUELLER, W. et al. Gender matters! analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science*, Springer, v. 6, n. 1, p. 5, 2017. Citado na página 26.

NABOULSI, D.; STANICA, R.; FIORE, M. Classifying call profiles in large-scale mobile traffic datasets. In: IEEE. *IEEE INFOCOM 2014-IEEE conference on computer communications*. [S.l.], 2014. p. 1806–1814. Citado na página 19.

NERY, M. B.; SOUZA, A. A. L. d.; ADORNO, S. Os padrões urbano-demográficos da capital paulista. *Estudos Avançados*, SciELO Brasil, v. 33, n. 97, p. 5–36, 2019. Citado 3 vezes nas páginas 13, 26 e 32.

NEV-USP. *Monitor da violência*. 2021. Disponível em: https://nev.prp.usp.br/projetos/projetos-especiais/monitor-da-violencia/. Acesso em 07 de jun. 2021. Citado 2 vezes nas páginas 13 e 41.

NORONHA, C. V. et al. Violência, etnia e cor: um estudo dos diferenciais na região metropolitana de salvador, bahia, brasil. *Revista Panamericana de Salud Pública*, SciELO Public Health, v. 5, p. 268–277, 1999. Citado na página 26.

OLIVEIRA, E. M. R. et al. Mobile data traffic modeling: Revealing temporal facets. *Computer Networks*, Elsevier, v. 112, p. 176–193, 2017. Citado na página 19.

OPENSTREETMAP. 2021. Disponível em: https://www.openstreetmap.org/about. Acesso em: 07 de nov. 2021. Citado na página 21.

PAULOS, E.; ANDERSON, K.; TOWNSEND, A. Ubicomp in the urban frontier. Carnegie Mellon University, 2004. Citado na página 17.

PAULOS, E.; GOODMAN, E. The familiar stranger: anxiety, comfort, and play in public places. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. [S.l.: s.n.], 2004. p. 223–230. Citado na página 17.

PISTORI, H. Tecnologia adaptativa em engenharia de computação: Estado da arte e aplicações. *Universidade de São Paulo (USP), São Paulo*, 2003. Citado na página 22.

QUERCIA, D. et al. Smelly maps: the digital life of urban smellscapes. *arXiv preprint arXiv:1505.06851*, 2015. Citado na página 19.

REDI, M. et al. The spirit of the city: Using social media to capture neighborhood ambiance. *Proceedings of the ACM on Human-Computer Interaction*, ACM New York, NY, USA, v. 2, n. CSCW, p. 1–18, 2018. Citado na página 20.

RODRIGUES, D. O. et al. Computa\c {c}\˜ ao urbana da teoria\a pr\'atica: Fundamentos, aplica\c {c}\˜ oes e desafios. *arXiv preprint arXiv:1912.05662*, 2019. Citado 2 vezes nas páginas 18 e 20.

SILVA, T. H.; LOUREIRO, A. A. Computaçao urbana: Técnicas para o estudo de sociedades com redes de sensoriamento participativo. *Anais da XXXIV JAI*, v. 8329, p. 68–122, 2015. Citado na página 19.

SILVA, T. H. et al. Definiçao, modelagem e aplicaçoes de camadas de sensoriamento participativo. In: *Brazilian Symposium on Computer Networks and Distributed Systems (SBRC'14), Florianópolis, Brazil*. [S.l.: s.n.], 2014. Citado na página 20.

SILVA, T. H. et al. Uma fotografia do instagram: Caracterização e aplicação. *Revista Brasileira de Redes de Computadores e Sistemas Distribuídos*, 2017. Citado na página 27.

SILVA, T. H. et al. Revealing the city that we cannot see. *ACM Transactions on Internet Technology (TOIT)*, ACM New York, NY, USA, v. 14, n. 4, p. 1–23, 2014. Citado na página 17.

SILVA, T. H. et al. Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 52, n. 1, p. 1–39, 2019. Citado 5 vezes nas páginas 8, 13, 17, 18 e 26.

SIMÃO, R. S. Computação urbana: as camadas de dados urbanos em florianópolis/sc. 2019. Citado na página 20.

SINNOTT, R. W. Virtues of the haversine. *S&T*, v. 68, n. 2, p. 158, 1984. Citado na página 27.

SORENSEN, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, v. 5, p. 1–34, 1948. Citado na página 27.

SSP-SP. *Dados Estatísticos do Estado de São Paulo*. 2021. Disponível em: http://www.ssp.sp.gov.br/estatistica/pesquisa.aspx. Acesso em 10 de mai. 2021. Citado 2 vezes nas páginas 14 e 31.

São Paulo. *Diário Oficial Do Estado De São Paulo*. 2015. Disponível em: https://www.imprensaoficial.com.br. Acesso em 07 de jul. 2021. Citado 2 vezes nas páginas 31 e 36.

TONRY, M. Ethnicity, crime, and immigration. *Crime and justice*, University of Chicago Press, v. 21, p. 1–29, 1997. Citado na página 26.

TUCKER, R. et al. Who 'tweets' where and when, and how does it help understand crime rates at places? measuring the presence of tourists and commuters in ambient populations. *Journal of Quantitative Criminology*, Springer, v. 37, n. 2, p. 333–359, 2021. Citado 2 vezes nas páginas 27 e 28.

VACA, C. K. et al. Taxonomy-based discovery and annotation of functional areas in the city. In: *Ninth International AAAI Conference on Web and Social Media*. [S.l.: s.n.], 2015. Citado na página 17.

VLAHOGIANNI, E. I. et al. *Exploiting new sensor technologies for real-time parking prediction in urban areas*. [S.l.], 2014. Citado na página 19.

WANG, H. et al. Learning task-specific city region partition. In: *The World Wide Web Conference*. [S.l.: s.n.], 2019. p. 3300–3306. Citado na página 13.

WANG, H. et al. Non-stationary model for crime rate inference using modern urban data. *IEEE transactions on big data*, IEEE, v. 5, n. 2, p. 180–194, 2017. Citado 2 vezes nas páginas 27 e 28.

WANG, Z. et al. Identification and analysis of urban functional area in hangzhou based on osm and poi data. *Plos one*, Public Library of Science San Francisco, CA USA, v. 16, n. 5, p. e0251988, 2021. Citado 3 vezes nas páginas 13, 26 e 34.

WEISBURD, D.; GROFF, E. R.; YANG, S.-M. *The criminology of place: Street segments and our understanding of the crime problem*. [S.l.]: Oxford University Press, 2012. Citado 2 vezes nas páginas 13 e 26.

YUAN, J.; ZHENG, Y.; XIE, X. Discovering regions of different functions in a city using human mobility and pois. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2012. p. 186–194. Citado 2 vezes nas páginas 13 e 26.

ZHENG, Y. et al. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, ACM New York, NY, USA, v. 5, n. 3, p. 1–55, 2014. Citado na página 17.

Apêndices

APÊNDICE A – Lista De Crimes

Crime	Abreviações
Homicídio Doloso	Hom. dol.
Nº de Vítimas em Homicídio Doloso	N° Hom. dol.
Homicídio Doloso por Acidente de Trânsito	Hom. trans.
Nº de Vítimas em Homicídio Doloso por Acidente de Trânsito	N° Hom. trans.
Homicídio Culposo por Acidente de Trânsito	Hom. culp. trans.
Homicídio Culposo Outros	Hom. culp.
Tentativa de Homicídio	Ten. hom.
Lesão Corporal Seguida de Morte	Lesão seg. mort.
Lesão Corporal Dolosa	Lesão dol.
Lesão Corporal Culposa por Acidente de Trânsito	Lesão trans.
Lesão Corporal Culposa - Outras	Lesão culp.
Latrocínio	Lat.
Nº de Vítimas em Latrocínio	N° Lat.
Total de Estupro	Tot. Estupros
Estupro	Estupro
Estupro de Vulnerável	Estupro vuln.
Total de Roubo - Outros	Tot. Roubos
Roubo - Outros	Roubos
Roubo de Veículo	Roubo veíc
Roubo a Banco	Roubo banc.
Roubo de Carga	Roubo carg.
Furto - Outros	Furtos
Furto de Veículo	Furto veíc.

APÊNDICE B – Lista De POIs

Alimentação	**Educação**	**Transporte**
		Estacionamento de Bicicletas
		Estação de Conserto de Bicicletas
		Aluguel de Bicicletas
		Aluguel de Barcos
		Compartilhamento de Barcos
Bar	Faculdade	Estação de Ônibus
Churrascaria	Auto Escola	Aluguel de Carros
Biergarten	Pré-escola	Compartilhamento de Carros
Cafeteria	Escola de Idiomas	Lava-jato
Fast Food	Biblioteca	Inspeção de Veículos do Governo
Praça de Alimentação	Brinquedoteca	Posto de Veículos Elétricos
Sorveteria	Escola de Música	Terminal de Barcas
Pub	Escola	Posto de Gasolina
Restaurante	Universidade	Caixa de sal de grãos
		Estacionamento de Motos
		Estacionamentos
		Entrada de Estacionamentos
		Áreas de Estacionamentos
		Táxi
Financeiro	**Saúde**	**Entretenimento, Arte e Cultura**
		Centro de Artes
		Bordel
		Cassino
	Baby Hatch	Cinema
	Clinica	Centro Comunitário
	Dentista	Fonte Decorativo
Caixa Eletrônico	Médico	Jogos de Azar
Banco	Hospital	Boate
Casa de câmbio	Casa de Repouso	Planetarium
	Farmácia	Public Bookcase
	Serviços Sociais	Centro Social
	Veterinário	Clube de Strip
		Estúdio
		Swingerclub
		Teatro

Serviços Públicos	**Facilidades**	**Gerenciamento de Resíduos**
Tribunal		
Corpo de Bombeiros	Banco de Sentar	
Policia	Drinking Water	
Caixa Postal	Give Box	Estação Sanitária
Depósito de Correio	Grave Yard	Instalações de Reciclagem
Correios	Banheiro Público	Cesto de Lixo
Prisão	Telefone Público	Depósito de Lixo
Piscina pública	Banheiro Público (com taxa)	Estação de Transferência de Lixo
Prefeitura	Abrigo	
Consulado		
Outros		
Embarque de Animais		
Abrigo de Animais		
Forno de Cozimento		
Relógio		
Crematório		
Centro de Mergulho		
Posto de Caça		
Cyber-café		
Cozinha Pública		
Kneipp Water Cure		
Mercado		
Mosteiro		
Cabine de Fotos		
Templos Religiosos		
Centro de Mergulho		
Área para refugiados		
Childcare		
Posto de Guarda Florestal		
Máquina de Vendas		
Regador		

TERMO DE AUTORIZAÇÃO PARA PUBLICAÇÃO DIGITAL NA BIBLIOTECA "JOSÉ ALBANO DE MACEDO"

Identificação do Tipo de Documento

( ) Tese

( ) Dissertação

(X) Monografia

( ) Artigo

Eu, **Denilson da Silva Sousa**, autorizo com base na Lei Federal nº 9.610 de 19 de Fevereiro de 1998 e na Lei nº 10.973 de 02 de dezembro de 2004, a biblioteca da Universidade Federal do Piauí a divulgar, gratuitamente, sem ressarcimento de direitos autorais, o texto integral da publicação **Relações entre Crimes e o Espaço Urbano: Um Estudo de Caso Baseado em Pontos de Interesses Extraídos da Web** de minha autoria, em formato PDF, para fins de leitura e/ou impressão, pela internet a título de divulgação da produção científica gerada pela Universidade.

Picos-PI **17** de **Maio** de **2023**.

Denilson da Silva Sousa

Assinatura

Denilson da Silva Sousa

Assinatura