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

Denilson da Silva Sousa, Marcos Paulo Fontes Feitosa, Glauber Dias Gonçalves

[1] Universidade Federal do Piauí - CSHNB
{denilsondsousa, marcosmpff, ggoncalves}@ufpi.edu.br

Abstract. *High crime rates are one of the 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 the urban space. In this article, we investigate the relationship between crime rates and these characteristics reflected in points of interest (POIs) that people have registered on a web service for 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 performance for this prediction and explored these models to discover the most important categories of POIs to explain the most frequent crimes by city regions. Additionally, we analyzed the performance gain with the increase of POIs registered in the city over the years.*

Resumo. *Altos índices de criminalidade 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 artigo, investigamos a relação entre índices de criminalidade e essas características refletidas em pontos de interesse (POIs) que as pessoas registraram em um serviço Web para a 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 e exploramos 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 ao longo dos anos.*

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 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 et al. 2019, Adorno and Nery 2019]. Tais características já são consideradas atualmente em estatísticas oficiais como o censo através de características urbanísticas do entorno dos domicílios com dados agregados por municípios[1]. Contudo, para subsidiar novos estudos que avançam 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 alimentado 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 et al. 2012, Yuan et al. 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 gerado por pessoas e disponíveis publicamente via serviços Web para predição de crimes. Em [Wang et al. 2019] e [Belesiotis et al. 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 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 et al. 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 traços de 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. No entanto, é necessário conhecer o nível de acurácia e especificidade desses dados para explicar índices de crimes.

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

Nesse artigo investigamos essa questão com a utilização 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 cidade com mais condições propícias para investigarmos o quanto 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 número 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 taxi 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 desempenhos 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.

Em suma esse artigo traz as seguintes contribuições:

- 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.

As próximas seções desse artigo estão com a seguinte organização. Na Seção 2, discutimos trabalhos relacionados. A descrição detalhada sobre a coleta e o processamento de dados para esse trabalho estão na Seção 3. Nossas análises e resultados são discutidos na Seção 4 ao passo que nossas considerações finais são apresentadas na Seção 5.

2. 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 (e.g censo demográfico, dados de mobilidade urbana, estatísticas sociais e outros). Estudos clássicos nessa área como [Masi et al. 2007] investigam o quanto questões raciais

e o grau de violência influenciam nos resultados de gravidez, enquanto [Tonry 1997, Noronha et al. 1999] analisam a distribuição da violência no espaço quanto a etnia e cor racial. Em [Becker and Kassouf 2017] é analisado se o gasto público do governo em educação impacta na redução da taxa de homicídios. Mais recentemente, em [Adorno and Nery 2019, Nery et al. 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. Em [Weisburd et al. 2012] um estudo baseado no histórico de 16 anos de crimes nas cidades de Seattle e Washington nos EUA é uma referência seminal sobre os esforços de criminologias em análises de crimes por regiões. 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 et al. 2012] os autores mostraram que o uso de informações categóricas de POIs são úteis para traçar o perfil de atividades que caracterizam bairros. 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 de 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. O trabalho de [Mueller et al. 2017] analisa 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 apresentada uma técnica para identificar POIs e, com base neles, reconhecer pontos turísticos em uma região. Diferentemente desses trabalhos o nosso foco é utilizar POIs já definidos e categorizados por serviços Web.

Mais relacionado ao nosso trabalho estão os estudos que mesclam fontes de dados oficiais, i.e., dados adquiridos por requerimento a órgãos de governo 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, Iranmanesh and Alpar Atun 2020]. Já outros trabalhos como [Wang et al. 2017, Belesiotis et al. 2018, 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. Já [Castro et al. 2020] observou o quanto dados do serviço web ”Onde fui Roubado”adicionados a dados oficiais melhorava a predição de índices de criminalidade. Em contra-ponto a esses estudos, nesse trabalho investigamos o quanto POIs unicamente podem predizer, i.e., explicar taxas criminais por regiões das cidades.

3. Bases de Dados e Metodologia

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.

3.1. 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 organizados em dezessete categorias de crimes contendo a contagem de ocorrências registradas por regiões do estado desde 2001. 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 contem oitenta e oito distritos policiais e coletamos dados desses dos últimos nove anos (2012 - 2020).[2] A Figura 1-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[3], 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.

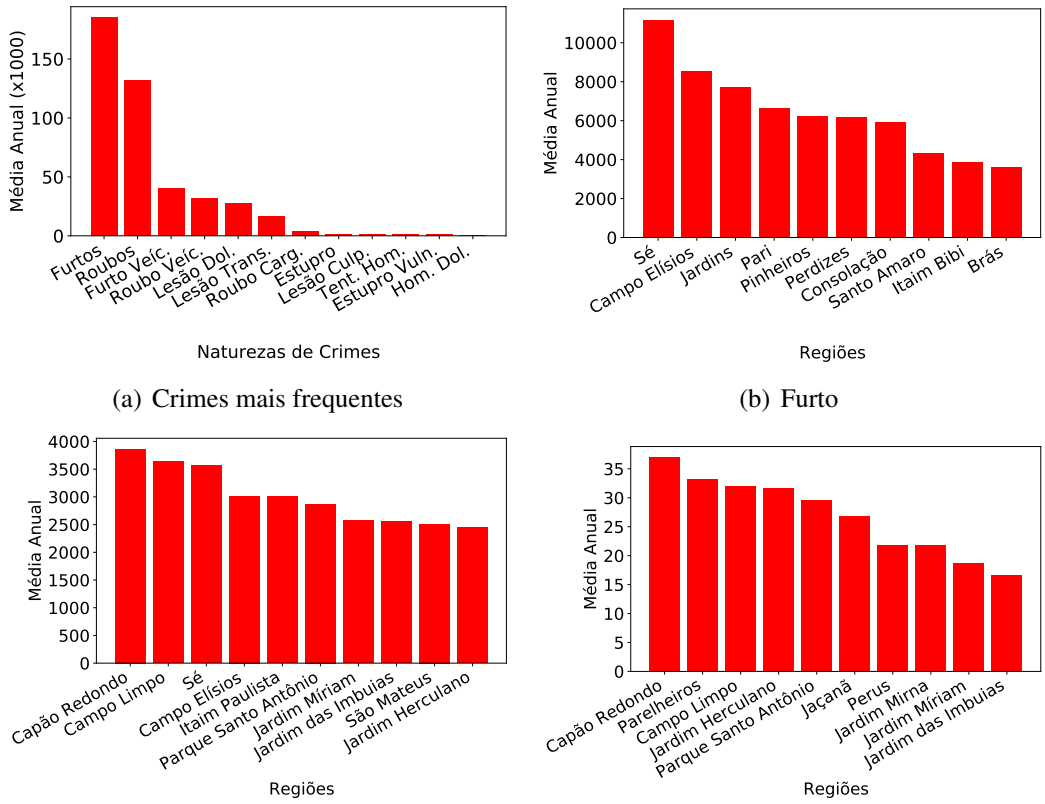
A Figura 1 b-d mostra as regiões mais violentas para os dois crimes mais frequentes (furto e roubo) mais homicídios. Pode-se observar que furto (Figura 1-b) é mais frequente nas regiões centrais da cidade, também regiões de maior poder aquisitivo da população (e.g., Sé, Campos Elísios, Jardins e Pari) com taxas anuais entre 4 mil e 12 mil ocorrências registradas. Contudo, à medida que crimes aumentam o grau de violência, como é o caso de roubos (Figura 1-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 na Figura 1-d, onde as regiões Capão Redondo e Paranheiros 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 and Nery 2019].

3.2. Pontos de Interesse (POIs)

POI é um tipo de dado mantido por serviços Web de mapeamento urbano que provê informações sobre locais da cidade com precisão de coordenadas geográficas [Yuan et al. 2012]. A informação principal de um POI é a sua categoria, que representa atividades econômicas ou culturais do local, inferida

[2] Utilizamos esse período para compatibilizar com os dados de POIs.

[3] 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 1. 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.

pelo serviço a partir de sensoriamento participativo e redes sociais baseadas em localização [Silva et al. 2019]. Nesse trabalho exploramos POIs do serviço OSM obtidos da API Osmosis [Ramm et al. 2011] que oferece ferramentas de programação para coletar POIs por regiões delimitadas por coordenadas geográficas.

Utilizamos essa API para mapear todos os POIs dos oitenta e oito distritos policiais da cidade de São Paulo, incluindo o crescimento gradativo do volume desses POIs de 2012 até 2020[4]. 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[5].

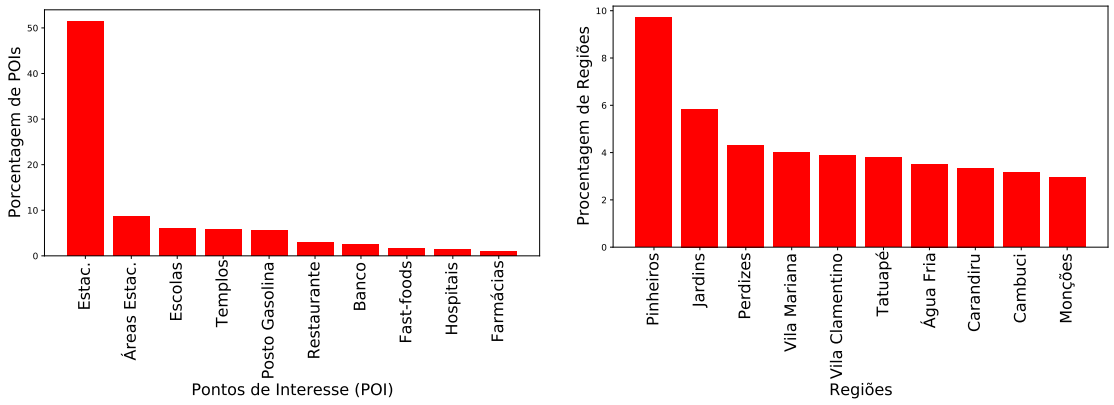
Dado esses procedimentos, coletamos 441.059 POIs cadastrados no OSM entre os anos de 2012 a 2020 na cidade de São Paulo, organizados em 107 categorias. A Figura 2-a mostra as dez categorias mais frequentes em nossa base de dados. Notavelmente estacionamentos é a mais frequente (51,4%), e isso ocorre provavelmente devido ao uso massivo do OSM por motoristas com forte demanda por vagas de estacionamento na ci-

[4]Período total oferecido pelas APIs do OSM

[5]https://www.google.com.br/maps

dade. No entanto, há outras categorias representativas não relacionadas diretamente a veículos como escolas (6,06%), templos (5,82%) e restaurantes (2,52%). Avaliaremos o potencial para predição de crimes de todas as categorias de POIs, pois esperamos que alguma dessas tenham informações preditivas para crimes, ainda que estejam em menor porcentagem.

A Figura 2-b mostra as dez regiões (distritos policiais) da cidade de São Paulo que acumulam maior volume de POIs. 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 categoria estacionamentos é majoritária nessas dez regiões, e desconsiderando essa categoria destacam-se Vila Madalena, Cambuci e Monções com as categorias teatro (0,39%), templos (5,82%) e restaurantes (2,86%) respectivamente, i.e., regiões com volumes representativos de atividades culturais. É importante destacar ainda Capão Redondo como a região com maior predominância de escolas (3,62%) e a região Vila Mariana com predominância de hospitais (11,07%).



(a) (b)

Figura 2. Pontos de Interesses (POIs) coletados do serviço OSM: (a) dez POIs mais frequentes e (b) regiões que acumulam o maior volume de POIs.

4. Resultados Experimentais

Nesta seção apresentamos nossos resultados sobre as base de dados de crimes e POIs apresentadas na seção anterior. Primeiramente, descrevemos as configurações utilizadas nos experimentos, considerando métodos de regressão e métricas de desempenho. A seguir, discutimos os resultados alcançados em termos de desempenho de diferentes métodos, importância de categorias de POIs e o impacto do aumento de POIs ao longo dos anos no desempenho dos métodos de regressão.

4.1. Configurações

O nosso objetivo nesses experimentos 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 caraterísticas 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:

$$\hat{y}_a = M(\{y\} \setminus y_a, X), \tag{1}$$

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, a taxa y_a será excluída das taxas de crimes y e POIs X a serem utilizados no treinamento do modelo. Em outras palavras, retiramos a região alvo dos dados para realizar a sua predição, ao passo que as outras regiões foram utilizadas para treinar o modelo.

Todas as predições para cada categoria crime foram utilizados para avaliar o desempenho do modelo. A precisão da estimativa foi avaliada através das métricas média do erro absoluto (MAE), média do erro relativo (MRE) e o índice R2. Especificamente essas métricas foram calculadas da seguinte forma:

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

onde y_i e $\hat{y}_i$ representam os valores atual e estimados para a taxa anual de 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.

Nossos experimentos foram realizados utilizando a biblioteca *scikit-learn* da linguagem *python*. Essa biblioteca compreende um conjunto de métodos de regressão [Pedregosa et al. 2011]. Utilizamos para construção do modelo M da Equação 1 os métodos regressão linear (RL), floresta aleatória (FA) e *support vector regression* (SVR). RL busca encontrar correlações entre a média de uma variável dependente com outra ou várias variáveis [Groß 2012]. 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 afim de ajustar seus valores [Drucker et al. 1997].

4.2. Desempenho de Diferentes Métodos

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.

A Tabela 1 mostra o desempenho dos métodos de regressão aplicados aos crimes mais frequentes no 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

Tabela 1. Desempenho dos métodos de regressão aplicados aos crimes mais frequentes do ano 2020: 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

frequentes o que indica que floresta aleatória os explica razoavelmente 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, e 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.

Agora focamos nos resultados da regressão com floresta aleatória, que obteve melhor desempenho, para analisar as categorias de crimes. Nesse sentido, consideramos que os modelos são úteis para explicar crimes quando, além do R2 positivo, apresentam erros relativos médios inferiores a 100%. Para exemplificar esses casos analisamos os erros para furto, roubo e homicídio doloso mostrados na Figura 1. 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, que é um erro razoavelmente baixo, considerando uma média anual superior a duas mil ocorrências anuais em todas as regiões de São Paulo. Por sua vez, roubo tem erro de 513.39 (46%) para uma média anual superior a mil e quatrocentas ocorrências, ao passo que a média de homicídio doloso é cerca de dez ocorrências anuais para um erro do modelo de 4.2 (93%). Levando em consideração as médias anuais de ocorrências mostradas na Figura 1 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.

4.3. 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 2 mostra cada uma das categorias de crimes seguido dos quatro POIs mais importantes para predição com suas respectivas porcentagens de relevância indicadas en-

Tabela 2. 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%)

tre parênteses.

Do total de cento e sete categorias de POIs do serviço OSM, onze aparecem entre as quatro mais importantes para os modelos apresentados. Escolas 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 distribuição desbalanceada de categorias POIs não impacta 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, escolas é a categoria de POI, notavelmente, mais importante para predizer lesão dolosa, mas explicações intuitivas para esse relacionamento podem ser tornar complexas. Possivelmente, há outras características espaciais associadas 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 taxi unicamente ocupa 66% da importância para explicar esse crime. Essas questões sobre características espaciais relacionadas aos POIs serão investigadas em trabalhos futuros.

4.4. 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.

A Tabela 3 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 discutidos 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.

Tabela 3. 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

5. Conclusões e Trabalhos Futuros

Nesse 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 é 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 o 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 desempenhos 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 nesse trabalho e outros tipos de modelos a serem analisados, assim como explicar a importância de diferentes POIs para esses modelos.

Referências

Adorno, S. and Nery, M. B. (2019). Crime e violências em são paulo: retrospectiva teórico-metodológica, avanços, limites e perspectivas futuras. *Cadernos Metrópole*, 21(44):169–194.

Becker, K. L. and Kassouf, A. L. (2017). Uma análise do efeito dos gastos públicos em educação sobre a criminalidade no brasil. *Economia e Sociedade*, 26(1):215–242.

Belesiotis, A., Papadakis, G., and Skoutas, D. (2018). Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, 3(4):1–31.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Castro, U. R., Rodrigues, M. W., and Brandao, W. C. (2020). Predicting crime by exploiting supervised learning on heterogeneous data. In *ICEIS (1)*, pages 524–531.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. (1997). Support vector regression machines. *Advances in neural information processing systems*, 9:155–161.

Groß, J. (2012). *Linear regression*, volume 175. Springer Science & Business Media.

Huang, C., Zhang, J., Zheng, Y., and Chawla, N. V. (2018). Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1423–1432.

Iranmanesh, A. and Alpar Atun, R. (2020). Reading the urban socio-spatial network through space syntax and geo-tagged twitter data. *Journal of Urban Design*, 25(6):738–757.

Masi, C. M., Hawkley, L. C., Piotrowski, Z. H., and Pickett, K. E. (2007). Neighborhood economic disadvantage, violent crime, group density, and pregnancy outcomes in a diverse, urban population. *Social science & medicine*, 65(12):2440–2457.

Mueller, W., Silva, T. H., Almeida, J. M., and Loureiro, A. A. (2017). Gender matters! analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science*, 6(1):5.

Nery, M. B., Souza, A. A. L. d., and Adorno, S. (2019). Os padrões urbano-demográficos da capital paulista. *Estudos Avançados*, 33(97):5–36.

NEV-USP (2021). Monitor da violência. Disponível em: https://nev.prp.usp.br/projetos/projetos-especiais/monitor-da-violencia/. Acesso em 07 de jun. 2021.

Noronha, C. V., Machado, E. P., Tapparelli, G., Cordeiro, T. R. F., Laranjeira, D. H. P., and Santos, C. A. T. (1999). Violência, etnia e cor: um estudo dos diferenciais na região metropolitana de salvador, bahia, brasil. *Revista Panamericana de Salud Pública*, 5:268–277.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Ramm, F., Topf, J., and Chilton, S. (2011). *OpenStreetMap: using and enhancing the free map of the world*. UIT Cambridge Cambridge.

Silva, T. H., de Melo, P. O. V., Almeida, J. M., and Loureiro, A. A. (2017). Uma fotografia do instagram: Caracterização e aplicação. *Revista Brasileira de Redes de Computadores e Sistemas Distribuídos*.

Silva, T. H., Viana, A. C., Benevenuto, F., Villas, L., Salles, J., Loureiro, A., and Quercia, D. (2019). Urban computing leveraging location-based social network data: a survey. *ACM Computing Surveys (CSUR)*, 52(1):1–39.

SSP-SP (2021). Dados estatísticos do estado de são paulo. Disponível em: http://www.ssp.sp.gov.br/estatistica/pesquisa.aspx. Acesso em 10 de mai. 2021.

São Paulo (2015). Diário oficial do estado de são paulo. Disponível em: https://www.imprensaoficial.com.br. Acesso em 07 de jul. 2021.

Tonry, M. (1997). Ethnicity, crime, and immigration. *Crime and justice*, 21:1–29.

Tucker, R., O'Brien, D. T., Ciomek, A., Castro, E., Wang, Q., and Phillips, N. E. (2021). 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*, 37(2):333–359.

Wang, H., Jenkins, P., Wei, H., Wu, F., and Li, Z. (2019). Learning task-specific city region partition. In *The World Wide Web Conference*, pages 3300–3306.

Wang, H., Yao, H., Kifer, D., Graif, C., and Li, Z. (2017). Non-stationary model for crime rate inference using modern urban data. *IEEE transactions on big data*, 5(2):180–194.

Wang, Z., Ma, D., Sun, D., and Zhang, J. (2021). Identification and analysis of urban functional area in hangzhou based on osm and poi data. *Plos one*, 16(5):e0251988.

Weisburd, D., Groff, E. R., and Yang, S.-M. (2012). *The criminology of place: Street segments and our understanding of the crime problem*. Oxford University Press.

Yuan, J., Zheng, Y., and Xie, X. (2012). 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*, pages 186–194.