

Probabilidade de ocorrência de crimes com Indicator Kriging

Thiago Corrêa de Oliveira Jessé

Trabalho de Conclusão de Curso

MBA em Inteligência Artificial e Big Data

UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Probabilidade de ocorrência de
crimes com Indicator Kriging

Thiago Corrêa de Oliveira Jessé

Thiago Corrêa de Oliveira Jessé

Probabilidade de ocorrência de crimes com Indicator Kriging

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Caetano M. Ranieri

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

Corrêa de Oliveira Jessé, Thiago
C824 Probabilidade de ocorrência de crimes com
p Indicator Kriging / Thiago Corrêa de Oliveira Jessé;
orientador Caetano Ranieri. -- São Carlos, 2024.
55 p.

Trabalho de conclusão de curso (MBA em
Inteligência Artificial e Big Data) -- Instituto de
Ciências Matemáticas e de Computação, Universidade de
São Paulo, 2024.

1. ESTATÍSTICA. 2. GEOESTATÍSTICA. 3. CRIME. 4. BIG
DATA. I. Ranieri, Caetano, orient. II. Título.

DEDICATÓRIA

*A minha mãe pelo suporte e apoio
incansável.*

AGRADECIMENTOS

Ao Prof. Dr. Caetano M. Ranieri que me orientou durante o desenvolvimento de todas as etapas deste trabalho.

Ao ICMC que me proporcionou o conhecimento e a infraestrutura necessária para a realização satisfatória deste estudo.

A minha família, que não hesitou em me apoiar nesse momento de transição de carreira.

RESUMO

Corrêa de Oliveira, T. J. **Probabilidade de ocorrência de crimes com Indicator Kriging**. 2024. 62 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

O estudo apresenta uma metodologia para a estimativa da probabilidade de ocorrência de crimes utilizando a técnica Indicator Kriging, visando lidar com o problema da subnotificação em registros de crimes na cidade de São Paulo. A subnotificação, que ocorre quando vítimas não registram ocorrências, limita a compreensão da criminalidade e impede a aplicação de políticas públicas eficazes. Para solucionar essa lacuna, a técnica de Indicator Kriging foi aplicada em um conjunto de dados de boletins de ocorrência, permitindo calcular a probabilidade de ocorrência de crimes em cada local, mesmo em áreas com muita ou pouca informação disponível. A metodologia incluiu o pré-processamento dos dados geográficos, o agrupamento em grid para viabilizar o processamento computacional e a transformação dos dados contínuos em indicadores binários, com o objetivo de estimar a probabilidade da ocorrência de crimes em cada localização específica.

O modelo foi construído com base em hot spots de criminalidade previamente identificados por meio de um modelo simples de Poisson, que permitiu distinguir áreas com alta concentração de crimes. A técnica Indicator Kriging foi então utilizada para estimar a probabilidade de crimes em cada local, levando em consideração a correlação espacial entre as diferentes áreas. Os resultados demonstraram que o método foi capaz de gerar mapas de probabilidade detalhados e precisos, reproduzindo padrões conhecidos e proporcionando uma análise robusta da criminalidade em diferentes regiões. O modelo desenvolvido se mostrou adequado para trabalhar com grandes volumes de dados, proporcionando resultados que poderão ser utilizados como subsídio para políticas públicas. A combinação de Big Data e técnicas geoestatísticas provou ser uma solução eficiente para a problemática da subnotificação, contribuindo significativamente para a compreensão da criminalidade urbana.

Palavras-chave: Subnotificação; Indicator Kriging; Big Data; Geoestatística; Criminalidade.

ABSTRACT

Corrêa de Oliveira, T. J. **Probability of crime occurrence with Indicator Kriging.** 2024. 62 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2024.

This study presents a methodology for estimating the probability of crime occurrences using the Indicator Kriging technique, aimed at addressing the problem of underreporting in crime records in São Paulo. Underreporting, which occurs when victims do not file reports, limits the understanding of criminal activity and hampers the implementation of effective public policies. To address this gap, the Indicator Kriging technique was applied to a dataset of police reports, enabling the calculation of crime occurrence probabilities for each location, regardless of whether there is a lot or little information available. The methodology included preprocessing geographic data, grid grouping to enable computational processing, and transforming continuous data into binary indicators to estimate the probability of crime occurrences at each specific location.

The model was built based on crime hot spots previously identified using a simple Poisson model, which allowed for distinguishing areas with high concentrations of crimes. Indicator Kriging was then used to estimate the probability of crimes in each location, taking into account the spatial correlation between different areas. The results showed that the method was capable of generating detailed and accurate probability maps, reproducing known patterns, and providing a robust analysis of criminal activity across different regions. The developed model proved suitable for working with large volumes of data, providing results that could serve as a basis for public policies. The combination of Big Data and geostatistical techniques proved to be an efficient solution to the issue of underreporting, significantly contributing to the understanding of urban crime.

Keywords: Underreporting; Indicator Kriging; Big Data; Geostatistics; Crime.

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação das etapas da metodologia proposta.....	32
Figura 2 – Gráfico de linhas com as ocorrências de crimes ao longo dos anos	33
Figura 3 – Gráfico de caixa com a distribuição de crimes por mês.....	34
Figura 4 – Gráfico de barras com a frequência de crimes por hora do dia.....	35
Figura 5 – Localizações informadas em B.O. ao redor do parque Ibirapuera em 2024.....	36
Figura 6 – Conjunto de dados filtrados pela malha territorial do município de São Paulo.....	38
Figura 7 – Grid com as ocorrências de crime agrupadas por célula em São Paulo.....	39
Figura 8 – Distribuição dos <i>hot spots</i> na cidade de São Paulo.....	40
Figura 9 – Exemplos de <i>hot spots</i> encontrados na Zona Sul de São Paulo.....	41
Figura 10 – Mapa de probabilidade gerada pela Indicator Kriging.....	42
Figura 11 – Exemplo de probabilidades de crime na Zona Norte de São Paulo.....	43
Figura 12 – Representação empírica do variograma.....	44
Figura 13 – Histograma de resíduos da Krigagem.....	48
Figura 14 – Comparação da distribuição de crimes reais e previstos.....	59
Figura 15 – Distribuição de crimes reais e previstos com transformação logarítmica.....	50
Figura 16 – Curva de Calibração das probabilidades previstas e observadas.....	51

LISTA DE TABELAS

Tabela 1 – Festas de Carnaval do ano 2017 ao ano 2024.....42

Tabela 2 – Parâmetros otimizados do modelo.....53

Tabela 3 – Comparação de métricas entre modelos.....55

LISTA DE SÍMBOLOS

$\gamma(h)$	Variograma
$I(x_i)$	Indicador no ponto x_i
x_i	Local

SUMÁRIO

1 INTRODUÇÃO	23
1.1 Contextualização do problema	23
1.2 Justificativa e importância.....	23
1.3 Objetivos.....	24
1.4 Objetivos específicos:	24
2 Fundamentação teórica.....	25
2.1 Fundamentos relacionados a criminologia.....	25
2.2 Fundamentos relacionados a geoestatística.....	26
2.3 Fundamentos relacionados ao Big Data	28
3 Trabalhos Relacionados	30
4 Metodologia.....	32
4.1 Materiais necessários.....	32
4.2 Análise Exploratória dos Dados	32
4.3 Pré-processamento dos dados.....	36
4.3.1 Limpeza e tratamento dos dados	36
4.3.2 Filtragem dos dados.....	37
4.3.3 Agrupamento dos dados	38
4.4 Obtenção dos <i>hot spots</i>	39
4.5 Aplicação da Krigagem.....	41
5. Análise dos Resultados	42
5.1 Métricas do próprio modelo	42
5.2 Análise quantitativa do modelo	46
5.2.1 Análise dos padrões de crime entre janeiro e junho de 2024	48
6. Conclusão	52
7. Referências Bibliográficas	53

1 INTRODUÇÃO

1.1 Contextualização do problema

Com o avanço da computação e o surgimento de técnicas de Big Data, há a possibilidade de abordar a criminalidade com estratégias fundamentadas em inteligência analítica que partem da imensa quantidade de dados existentes [1].

De acordo com dados da Secretaria de Segurança pública de São Paulo, somente no ano de 2022, foram feitas 202 mil ocorrências de roubo e furtos de celular na capital do estado de São Paulo, número esse que representa um aumento de 12% em relação ao ano anterior, praticamente 1 roubo ou furto de celular a cada 3 minutos [2].

Além disso, as estatísticas mais recentes indicam que 52% das vítimas de roubo afirmaram não ter registrado boletim de ocorrência [3] evidenciando o problema da subnotificação e trazendo dificuldade para a compreensão da criminalidade através dos dados, que muitas vezes são faltantes.

1.2 Justificativa e importância

Políticas públicas de segurança se tornaram prioridade conforme as áreas urbanas tiveram um gigantesco crescimento na sociedade contemporânea, uma vez que a segurança civil é essencial para o bem-estar da população.

A tentativa de entender o padrão de ocorrência de crimes se tornou extremamente importante nesse cenário, entretanto, abordagens como por exemplo a previsão de crimes baseada em séries temporais possuem a grande limitação de não poder ser utilizada em locais que não possuem dados históricos [4], necessitando então de alguma abordagem de interpolação de dados que consiga estimar valores desconhecidos em um espaço geográfico.

Sendo assim, esse estudo vem para avaliar a aplicabilidade do método de Indicator Kriging como método estatístico para interpolação de dados [5] baseado na probabilidade de ocorrência em cada local analisado, possibilitando não só o desenvolvimento de modelos preditivos, mas também entender melhor os padrões de crime e dessa forma, lidar com o problema da subnotificação presente nos boletins de ocorrência.

1.3 Objetivos

O objetivo principal é utilizar a técnica Indicator Kriging para calcular a probabilidade de ocorrências de crime em um grande volume de dados de Boletins de Ocorrências na cidade de São Paulo e observar os padrões de funcionamento da criminalidade perante a ótica deste mesmo modelo.

1.4 Objetivos específicos:

- Estudar formas de melhor apresentar e analisar os dados públicos disponibilizados pela Secretaria da Segurança Pública do Estado de São Paulo relacionados ao roubo/furto de celular
- Encontrar “hot spots magnéticos” (locais com concentração de ocorrências de crimes, que costumam se repetir ao longo do tempo) na cidade de São Paulo.
- Explicar padrões de ocorrência de crimes a partir da inteligência analítica desenvolvida ao longo do estudo.
- Propor uma solução eficiente de Big Data para a grande quantidade de dados atualmente existentes no contexto do estudo.
- Utilizar técnicas de manipulação de grande volume de dados para possibilitar modelagens geoestatísticas eficientes.

2 Fundamentação teórica

Os fundamentos necessários para a existência de estudo se distribuem entre a compreensão de elementos da criminologia, para entender os padrões da criminalidade e de como o crime funciona, e compreensão de elementos da geoestatística, devido a natureza dos dados analisados e por último, métodos de Big Data, utilizados no estudo para possibilitar a modelagem geoestatística.

2.1 Fundamentos relacionados a criminologia

A criminologia em si, é a ciência que estuda as causas, correções e prevenções da criminalidade, nesse trabalho utilizaremos a Teoria dos Padrões de Crime [6] introduzida por Paul e Patricia Brantingham, propõe que o infrator realiza escolhas racionais ao fazer um crime, por exemplo quando um crime é bem-sucedido, há a motivação do criminoso em cometer outros crimes [7]. O conceito principal dessa teoria, é a ideia de que existem *hot spots* (pontos quentes) nos quais um infrator se sente seguro para cometer crimes pois são íntimos às brechas que aquele determinado ambiente oferece, lugares que são crimogênicos independentemente do fluxo de pessoas no local, justificando os estudos que descobrem que a densidade populacional não demonstrou ser um indicador que explica altas taxas de criminalidade [5]. De forma geral, as vertentes da criminologia argumentam que tempo e espaço são importantes em explicar o comportamento criminoso.

O primeiro estudo que comprovou essa teoria dentro do campo estatística foi conduzido em Minneapolis [8], ao utilizar um modelo simples de Poisson (explicado na próxima subseção) foi descoberto que relativamente poucos lugares produziam a maioria das chamadas para a polícia (50% das chamadas sobre qualquer tipo de crime foi realizada em 3% dos locais) enquanto as chamadas relacionadas ao relato de crimes predatórios (tipo de crime que contém alvos) eram ainda mais específicos, com todos os tipos de roubo presente em apenas 2,2% dos locais e o crime de estupro sendo relatado em apenas 1,2% dos locais. O fato de apenas 3,6% das chamadas serem realizadas em locais não repetidos também trouxeram indícios que comprovavam a promessa de que o crime se concentra em locais atratores e não possui um comportamento aleatório.

Esse estudo desencadeou diversos outros estudos que buscaram encontrar esses pontos críticos atratores de crime para poder tratá-los, em 2019 [9] realizou uma meta-análise entre diversos estudos para avaliar se uma polícia orientada a esses *hot spots* realmente diminuiria o crime ou apenas o dispersaria. Suas descobertas indicaram que as intervenções de policiamento em *hot spots* (principalmente tratar esses locais buscando lidar com os fatores que fazem ele ser um atrator de crime e não apenas só aumentar a patrulha policial neles) resultaram na diminuição do crime e na difusão dos benefícios desse controle criminal para as áreas próximas, e não apenas deslocar o crime para outras regiões.

2.2 Fundamentos relacionados a geoestatística

Em 1951, Danie G. Krige procurou estimar a distribuição mais provável de ouro com base em amostras de alguns furos de sondagem, ao desenvolver o método que leva seu próprio nome, a Krigagem [5] tem como princípio heurístico a ideia de que locus (locais onde alguma coisa acontece, por exemplo o adensamento de pragas em plantações e a distribuição de petróleo no solo) mais próximos tendem a ter valores do fenômeno estudado mais parecido do que os locus distantes. Para utilizar a Krigagem para interpolar dados, é necessário a construção de um variograma (função matemática que descreve como a semivariância dos dados muda com a distância espacial entre amostras) para conhecer a homogeneidade ou heterogeneidade da variação do fenômeno abordado e então estimar valores para locus que não possuem medições.

Diversas variações da Krigagem foram desenvolvidas ao longo dos anos, principalmente a Indicator Kriging ou Krigagem por Indicador ou desenvolvida por Andre G. Journel [10] que será utilizada nesse estudo. Ela se baseia em definir dados categóricos a partir de um limiar de corte para estimar a probabilidade da ocorrência de um determinado evento nos locais analisados, sem necessitar conhecer os parâmetros da distribuição dos dados ou a própria distribuição em si, ideal para o nosso caso que aborda dados incompletos, subnotificados.

Seu cálculo funciona da seguinte forma: Primeiramente, transforma-se os dados originais contínuos em indicadores binários, para cada ponto de amostragem x_i define-se um limiar de corte para dividir os dados, no nosso caso o valor de corte será se o local é um ambiente atrator de crimes (*hot spots*) ou não. Com isso é construído a variável indicadora $I(x_i)$:

$$I(x_i) = \begin{cases} 1, & \text{se } Z(x_i) \leq z_c \\ 0, & \text{se } Z(x_i) > z_c \end{cases}$$

Sendo $Z(x_i)$ a variável de interesse no local x_i , que no nosso caso será a quantidade de boletins de ocorrência que se referem a um furto/roubo de celular no respectivo local, o z_c utilizado para definir se o local é um *hot spot* ou não, será explicado mais para frente na fundamentação teórica.

Com os dados transformados em indicadores binários, calcula-se o variograma $\gamma(h)$ que descreve a variabilidade espacial dos dados, ou seja, a correlação dos pontos com base na distância entre eles, onde “ h ” é igual a distância entre o ponto amostrado e os demais pontos. A função é representada da seguinte forma:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [I(x_i) - I(x_i + h)]^2$$

Com $N(h)$ sendo o número de pares de pontos separados pela distância h e $I(x)$ sendo o valor da variável de interesse no ponto x .

Com o variograma, é possível estimar a probabilidade condicional $\hat{I}(x_0)$ nos pontos pertencentes a área estudada. Essa estimativa é obtida através de uma média ponderada dos pontos amostrais próximos, cada um com seu peso λ_i :

$$\hat{I}(x_0) = \sum_{i=1}^N \lambda_i I(x_i)$$

Os pesos λ_i , que somados são iguais a 1, são calculados resolvendo um sistema de equações lineares baseado no variograma, garantindo que a estimativa não seja enviesada (nem superestimada e nem subestimada) se aproximando do valor real esperado apenas com o valor dos pontos próximos:

$$\sum_{i=1}^N \lambda_i \gamma(x_i - x_j) = \gamma(x_j - x_0), \quad \text{para } j = 1, 2, \dots, N$$

Ao final do cálculo, obtemos a probabilidade da ocorrência de um crime em cada ponto ao longo do espaço analisado.

Para definir o valor do limiar de corte, será utilizado a mesma técnica usada pelo artigo que primeiro encontrou hot spots no âmbito estatístico [8] e posteriormente comprovada ser uma excelente técnica para esse cenário [11], a técnica do modelo simples de Poisson, que

utiliza a distribuição de Poisson (distribuição estatística que melhor descreve uma distribuição de eventos independentes em um intervalo fixo de tempo ou espaço) para calcular a probabilidade de observar um determinado evento caso a distribuição fosse homogênea, ou seja, caso os crimes fossem distribuídos igualmente no espaço ao invés de serem concentrados em *hot spots*, de modo que, valores muito destoantes da média seriam acumulações extremamente improváveis e por consequência *hot spots*, exposto na seguinte fórmula:

$$P(X \leq x) = \sum_{i=0}^x \frac{\lambda^i e^{-\lambda}}{i!}$$

Com λ sendo a média de eventos, $P(X \leq x)$ indicaria a probabilidade da existência de um evento X ser menor ou igual a x , no nosso caso buscaremos valores de locais com a probabilidade de 1×10^{-30} , ou seja, uma probabilidade tão baixa que somente um local que possui algum padrão de ocorrência de crimes satisfaça a condição, possibilitando encontrar locais que são atratores de crimes.

2.3 Fundamentos relacionados ao Big Data

Devido a quantidade de dados em larga escala, é essencial recorrer aos fundamentos de Big Data para um gerenciamento inteligente de recursos e dos próprios dados [12], dentre eles, a Limpeza e Tratamento de dados exerce um papel fundamental ao evitar que recursos sejam gastos desnecessariamente, ao remover dados impróprios (como valores nulos ou erroneamente inseridos) a qualidade do conjunto de dados evolui notavelmente. Esse processo também inclui padronizar os valores para que sempre estejam em um mesmo formato consensual, dessa forma prontificando esse mesmo conjunto de dados para análises futuras.

Por sua vez, a Filtragem de dados também exerce um papel essencial para uma boa utilização do potencial presente nos dados, ao nos depararmos com a pluralidade presente nos dados brutos, é comum a ocorrência de cenários desnecessários para a análise, a filtragem atua de modo a remover dados irrelevantes e focar em amostras que atendem determinadas condições, como no nosso caso, uma determinada localização geográfica, proporcionando uma redução no volume de dados que resulta em uma menor necessidade de processamento computacional.

Por último, o Agrupamento de dados, ou seja, agrupar diferentes dados em uma só métrica, como contagens ou médias, permite que se encontre tendências e padrões em grandes conjuntos de dados sem a necessidade de processar cada registro individualmente, além de organizar as informações, essa técnica aliada as outras técnicas mencionadas, não só permitem que um grande volume de dados sejam analisados e trabalhados, como também o fazem de forma inteligente, de modo a reduzir os recursos necessários em cada operação, proporcionando a realização desse estudo.

3 Trabalhos Relacionados

Muitos estudos interessantes foram feitos tentando observar o fenômeno da criminalidade através de um olhar de IA e Big Data, mas apesar de ser uma área de muita importância para a sociedade, a união entre a Krigagem e a criminalidade foi pouco explorada, justificando o fato do estado da arte na área de compreensão e predição de crimes, trabalho desenvolvido por Boppuru Ruddra Prathap [4], ter sido feito sem a utilização da Krigagem, apesar dela ser a técnica geoestatística mais eficiente para predições [13]. O trabalho de Prathap consiste em utilizar como valores geográficos, os dados de feeds de notícias e mídias sociais (por serem em tempo real), porém o autor utilizou a técnica Kernel Density Estimation (KDE) para encontrar hot spots criminais, ou seja, obteve um grande sucesso em identificar áreas de alta concentração de crimes, mas foi ineficaz em prever valores em locais não amostrados.

Sobre os estudos mais recentes e proveitosos abordando o aspecto espacial da Krigagem dentro do âmbito da criminologia, podemos citar o trabalho relacionado aos dados de movimentação de possíveis infratores [14] utilizados como covariável acrescentada a análise de dados históricos. Embora essa técnica tenha se saído melhor do que ao utilizar apenas uma variável, sua eficácia foi comprovada apenas em previsões de curtas janelas de tempo, muito provavelmente devido o artigo utilizar somente a CoKrigagem para capturar as zonas de alta incerteza, característica que é mais bem compreendida pela Krigagem por Indicador.

Também foram realizadas pesquisas com a Krigagem na Nigéria [13], usando variáveis socioeconômicas (como o desemprego e a taxa de pobreza) para compor uma regressão que ajudasse a explicar a variabilidade da taxa de criminalidade. Apesar do estudo não se preocupar com a obtenção de *hot spots* específicos, ele conseguiu demonstrar que a taxa de criminalidade é maior em regiões com alto grau de desemprego juvenil e maiores taxas de pobreza, justificando o aumento da incidência de crimes nesses locais e identificando outros locais com alta incidência de crime que não foram amostrados.

Além da utilização da Krigagem, outros estudos interessantes obtiveram resultados satisfatórios ao analisar fatores que pudessem explicar a ocorrência de crimes, como por exemplo notícias de jornal sobre o tema [15], redes sociais (mais especificamente *tweets*) atrelado ao clima [16], ocorrências delimitadas por crimes cometidos apenas em ruas e rodovias (de modo que se isole esse fator para uma análise específica) [17], e também avanços na visualização da distribuição espacial do crime, como por exemplo a utilização de indicadores

de risco de crime [18]. Estudos esses que, demonstram o grande potencial do Big Data em melhorar a compreensão sobre os fenômenos da criminalidade.

No Brasil, as pesquisas mais notáveis foram realizadas para tentar prever a ocorrência de crimes como homicídio relacionando a criminalidade com indicadores urbanos [19] e também visualizar as ocorrências de crime de forma mais eficaz, como visto na ferramenta Mirante [20] que recebeu elogios de especialistas na segurança pública e possibilitou a aproximação do desenvolvimento científico com a utilização prática das políticas públicas.

Porém não foi encontrado nenhum estudo que explorasse a Krigagem no contexto da criminalidade brasileira, tão peculiar devido a característica da subnotificação [1] e mesmo no cenário internacional, nenhum estudo utilizou a Krigagem por Indicador aplicada ao tema da criminalidade, ou com um conjunto de dados subnotificados.

4 Metodologia

4.1 Materiais necessários

Para a aplicação do modelo Indicator Kriging, é necessário que se utilize dados de aspecto geográfico, ou seja, que possuam coordenadas (latitude e longitude), sejam elas já prontas para uso no próprio conjunto de dados ou indiretamente calculadas através de referências geográficas. Para o desenvolvimento do estudo, será utilizado um conjunto de dados disponibilizados pela Secretaria de Segurança Pública do Estado de São Paulo, que cataloga todos os boletins de ocorrência relacionados aos furtos de celulares desde janeiro de 2017 até junho de 2024 [20], o modelo será treinado nos dados de 2017 a 2023 e será testado nos dados de 2024.

Esse conjunto de dados se torna muito útil para a modelagem proposta pois possui informações já explícitas referentes as localizações geográficas (coordenadas do local da ocorrência do crime) que serão utilizadas na modelagem, além de informações relacionadas ao boletim de ocorrência, como por exemplo a data e hora do crime, delegacia que registrou o boletim e o objeto que foi furtado.

Ao ter os dados necessários, é possível iniciar o processo da construção do modelo de interpolação dos dados, demonstrado no diagrama a seguir (Figura 1):



Figura 1 - Representação das etapas da metodologia proposta

4.2 Análise Exploratória dos Dados

O objetivo dessa seção é compreender melhor os dados que serão trabalhados, analisando e os visualizando. Começando pela Figura 2, que demonstra a evolução das ocorrências ao longo dos anos.

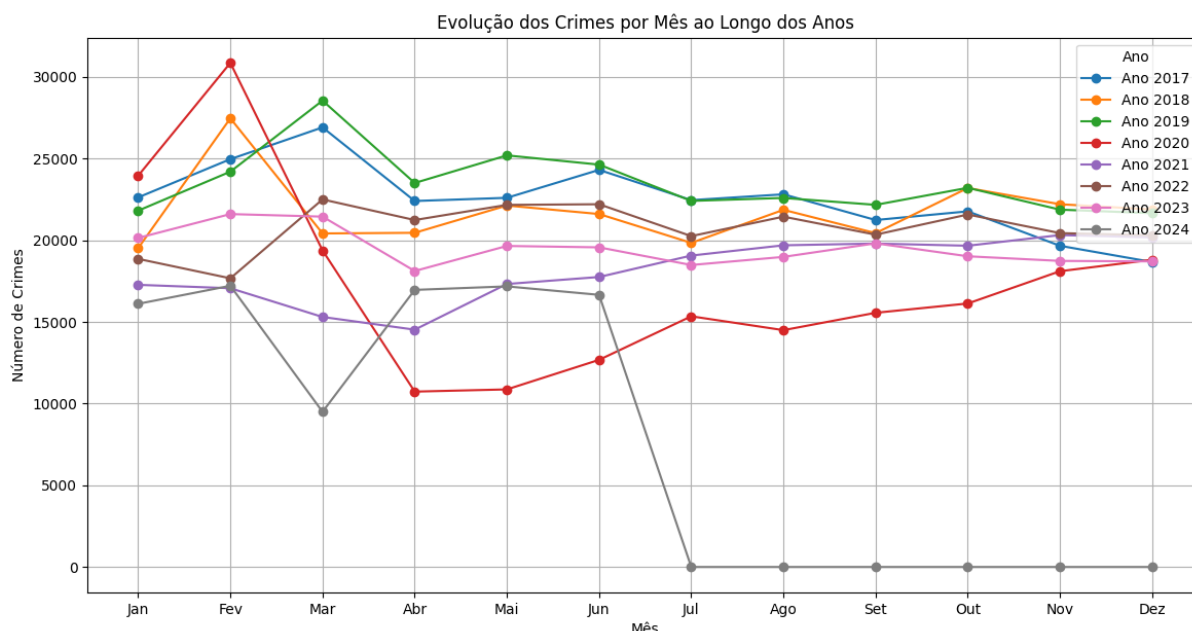


Figura 2 – Gráfico de linhas com as ocorrências de crimes ao longo dos anos

Primeiramente, nota-se que os anos com maiores taxas de criminalidade foram os anos antes da pandemia de 2020 (anos 2017, 2018 e 2019) e que mesmo após ela, a quantidade de ocorrências registradas não chegou ao mesmo valor que o período anterior, demonstrando uma mudança no comportamento marcada pela pandemia do covid-19.

Observando o ano de 2020, é notável a queda abrupta (diminuição de 65%) entre fevereiro e abril, justamente quando se iniciou a contaminação pelo covid-19 no Brasil [22] e de maneira interessante, a quantidade de ocorrências não retornou aos padrões dos anos anteriores até mesmo nos anos posteriores, demonstrando que além de ser um ano extremamente atípico, também moldou o comportamento dos anos seguintes.

Mergulhando mais especificamente na granularidade temporal mensal, temos o seguinte gráfico apresenta na Figura 3:

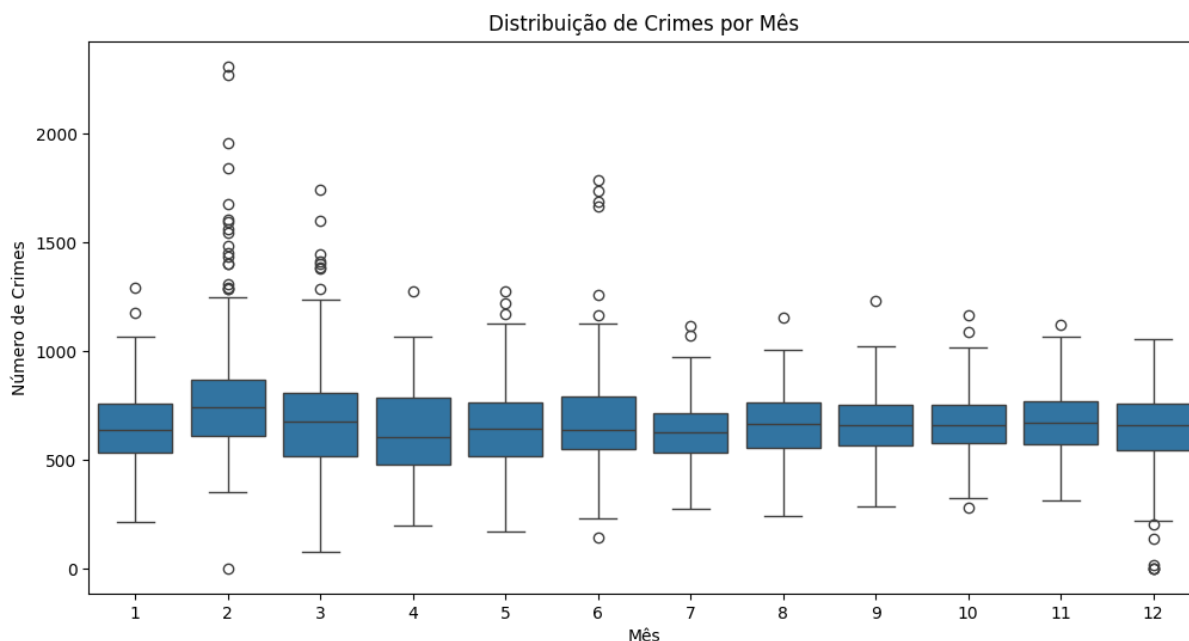


Figura 3 – Gráfico de caixa com a distribuição de crimes por mês

Ao repararmos na forma como os dados se distribuem, entende-se que em todos os meses, sem exceção, há a existência de dados que são *outliers*, ou seja, dados que estão localizados abaixo do primeiro quartil ou acima do terceiro quartil, induzindo que não existe uma relação bem estabelecida entre a variação de mês e a quantidade de ocorrência de crimes.

Porém, é interessante notar que dezembro é o mês com menos ocorrências de crime, mesmo sendo a época na qual as saídas temporárias possuem maior duração, do Natal ao ano novo [23]. Em contrapartida, o mês de fevereiro, seguido do mês de março, são os meses que possuem maior incidência de ocorrência no período analisado, coincidindo com as datas de festa de carnaval.

Tabela 1 – Festas de Carnaval do ano 2017 ao ano 2024

Ano	Data de início
2017	28 de fevereiro
2018	13 de fevereiro
2019	5 de março
2020	25 de fevereiro
2021	16 de fevereiro
2022	1 de março
2023	21 de fevereiro
2024	13 de fevereiro

Analisando a granularidade temporal mais profundamente ainda, podemos visualizar a frequência de eventos registradas em cada hora do dia, na Figura 4:

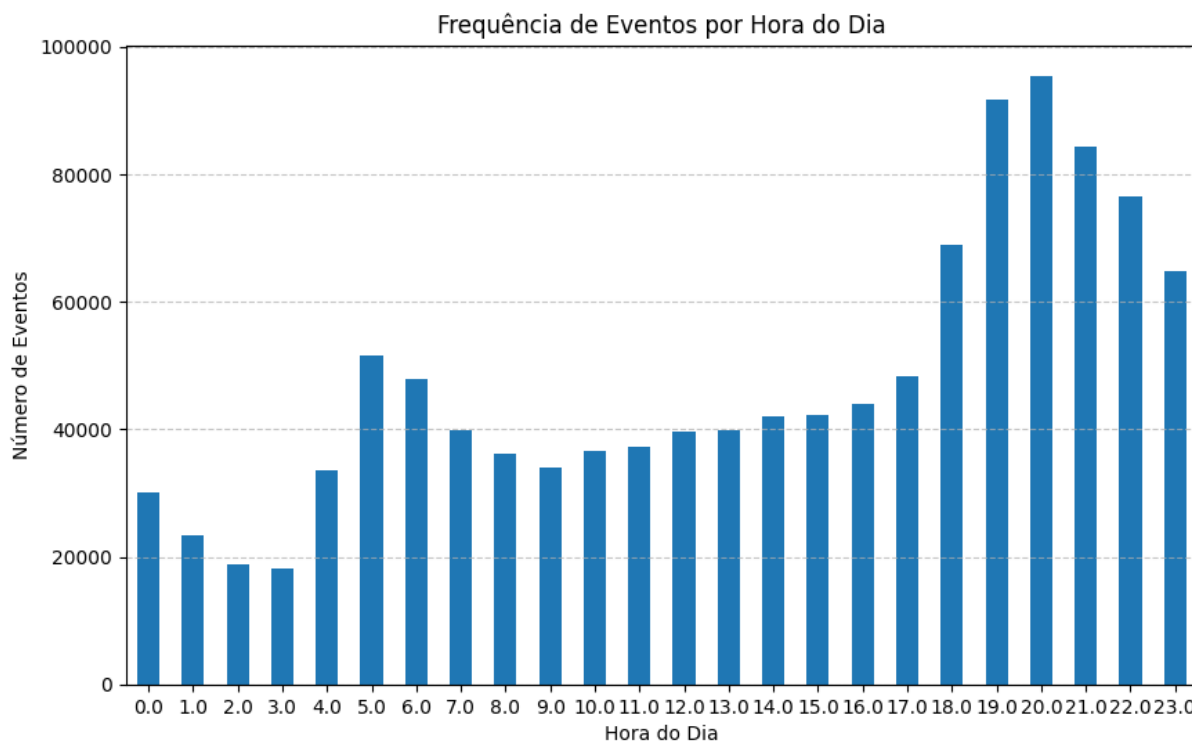


Figura 4 – Gráfico de barras com a frequência de crimes por hora do dia

Ao visualizar como as ocorrências se distribuem ao longo do dia, é possível notar que há um pico entre as 18 e 22 horas, horário no qual as pessoas costumam sair de seus empregos e irem para a casa, ou frequentarem ambientes sociais, se tornando um horário propício para a existência de crimes também pelo fato da escuridão da noite aumentar a vulnerabilidade de cidadãos comuns.

Relativo à quantidade de pessoas na rua, a madrugada costuma ser o horário com menos crimes justamente por esse motivo, porém conforme as pessoas acordam e vão para o trabalho, há um leve pico durante as 5 e 7 horas da manhã, que volta a ser estabilizado enquanto as mesmas pessoas estão comumente exercendo atividades trabalhistas, das 9 às 17 horas.

Percebendo como a ocorrência de crimes varia ao longo do tempo, podemos seguir com a análise e visualizar os como os dados se comportam geograficamente na Figura 5:

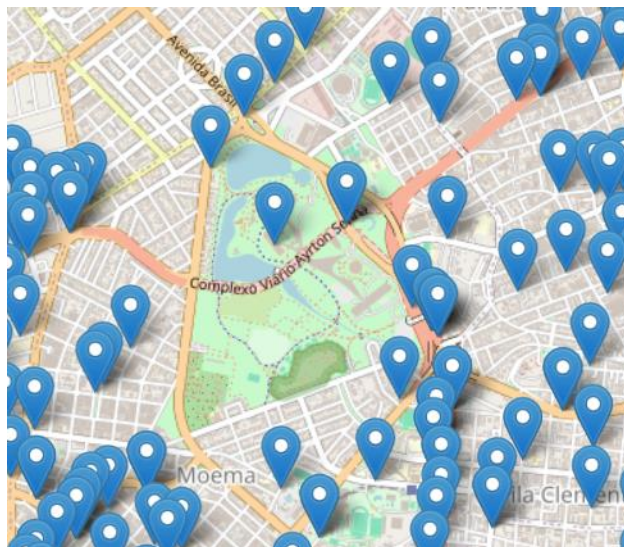


Figura 5 – Exemplo de localizações informadas em B.O. ao redor do parque Ibirapuera

Os dados utilizados, no aspecto geográfico, possuem 3.342.388 registros pertencentes ao município de São Paulo, sendo extremamente úteis por já contarem com a localização exata do local do crime em termos de coordenadas (latitude e longitude) e não precisarem serem calculadas por algum algoritmo ou API, ou formatadas para algum padrão diferente, pois já estão no modelo de graus decimais (exemplo: $\pm DD.DDDDDD$), facilitando o trabalho realizado e justificando a escolha desse conjunto de dados.

Porém, os dados brutos necessitam de uma determinada lapidação para possibilitar a performance do modelo, por exemplo, alguns registros são nulos e existe um grande volume de coordenadas que não direcionam para a cidade de São Paulo, exemplos esses que serão trabalhados na próxima seção.

4.3 Pré-processamento dos dados

Esta seção tem como objetivo esclarecer o pré-processamento de dados que foi realizado para possibilitar a modelagem da Krigagem e lidar com o grande volume de dados espaciais. O pré-processamento de dados seguiu as seguintes etapas:

4.3.1 Limpeza e tratamento dos dados

Teoricamente todo registro deveria ter uma coordenada que se refere ao local do crime, porém essa não é a realidade, alguns dados não possuem coordenadas ou possuem coordenadas fora do espectro geográfico da cidade de São Paulo. Para manter uma boa qualidade dos dados e por consequência uma boa qualidade do modelo, esses registros foram removidos. Nesse processo, 696.833 linhas (20,5% dos dados totais) foram identificadas como inadequadas por não possuírem coordenadas ou terem coordenadas iguais a zero, e por isso foram excluídas do conjunto de dados.

4.3.2 Filtragem dos dados

Devido o conjunto de dados registrar todos os boletins de ocorrência que aconteceram no Estado de São Paulo, há a necessidade de filtrar os dados para as evidências que foram registradas apenas no município da cidade de São Paulo, pois o Estado em si é gigantesco e contém uma pluralidade absurda de diferentes municípios com diferentes características em si, desde regiões rurais até grandes metrópoles como a própria cidade de São Paulo, evidenciando a necessidade de filtrar os dados para uma região geográfica específica, afinal a pluralidade impactaria negativamente na especificidade do modelo.

Para realizar esse procedimento, foi utilizado uma malha territorial disponibilizada pelo IBGE [24] que contém as fronteiras de todos os municípios brasileiros. Portanto, a partir dela realizou-se a escolha de manter apenas os dados que representavam ocorrências de crimes realizados no município da cidade de São Paulo, dessa forma apenas 1.526.082 ocorrências permaneceram no conjunto de dados que alimentará o modelo (45,66% dos dados totais). A implementação dessa etapa será mostrada na Figura 6, a seguir:

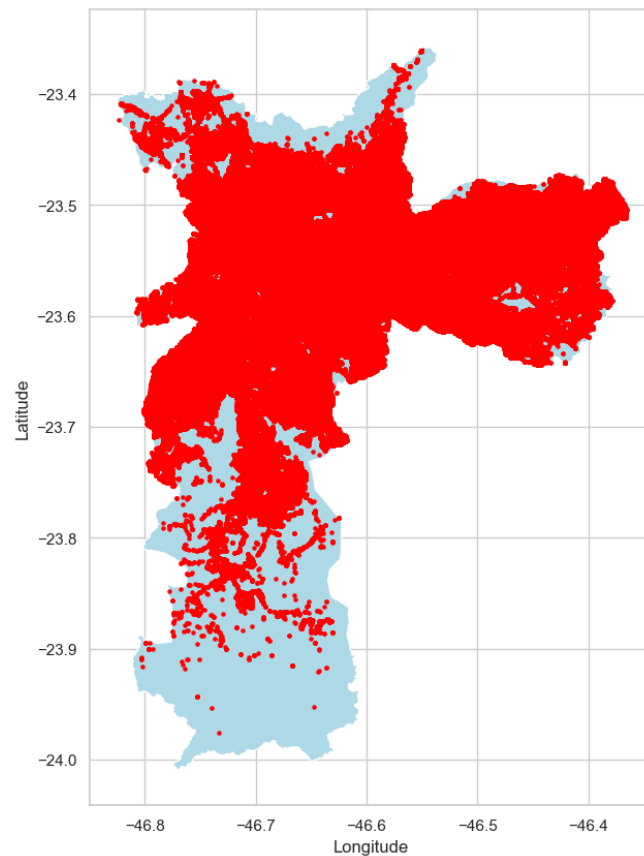


Figura 6 – Conjunto de dados filtrados pela malha territorial do município de São Paulo

4.3.3 Agrupamento dos dados

Apesar das etapas anteriores terem cortado mais da metade do volume dos dados, é extremamente custoso computacionalmente trabalhar com as coordenadas individualizadas, para evitar o altíssimo esforço computacional, as coordenadas disponíveis foram divididas em um grid de 300x300 metros, fazendo com que um grande conjunto de dados fosse resumido em células de menor quantidade, porém que ainda representasse a variabilidade espacial dos dados originais e assim tornando possível calcular operações complexas com um baixo custo computacional, resultado que será visto na Figura 7, a seguir:

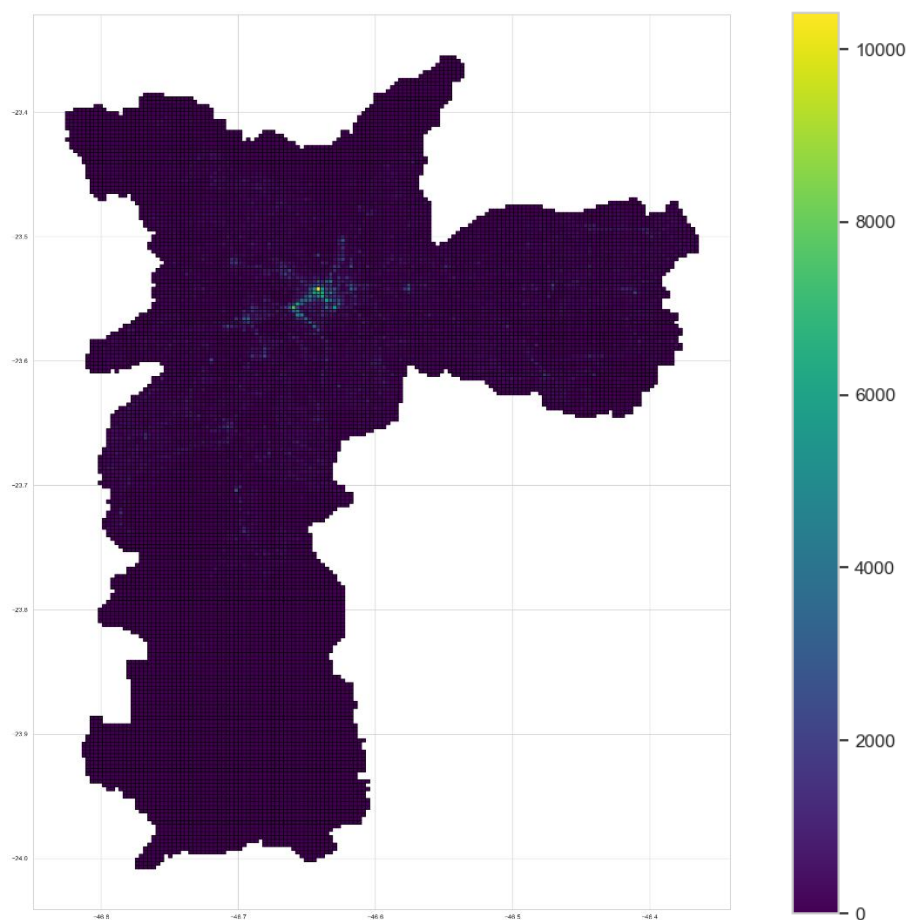


Figura 7 – Grid com as ocorrências de crime agrupadas por célula em São Paulo

Como resultado, obteve-se um grid com 15.598 células, reduzindo absurdamente a quantidade de dados, que inicialmente era de 3.342.388 ocorrências, ou seja, uma redução de mais de 200 vezes no volume do conjunto de dados, otimizando os esforços computacionais e possibilitando a execução de um modelo tão complexo quanto o Indicator Kriging.

Nota-se também que na figura 6, o crime parece estar igualmente distribuído pela cidade, enquanto na figura 7 já se percebe que ele se acumula em pontos específicos

4.4 Obtenção dos *hot spots*

Como demonstrado na seção de 2.2 Fundamentos relacionados a geoestatística, é necessário que os locais analisados sejam pertencentes a uma das possíveis classes, seja ela 0 ou 1, encontrados através do Modelo Simples de Poisson, que permite separar os lugares com quantidades de ocorrências extremamente fora do comum, dos lugares comuns. Vejamos os lugares que possuem a característica de serem atratores de ocorrências criminais, na Figura 8:

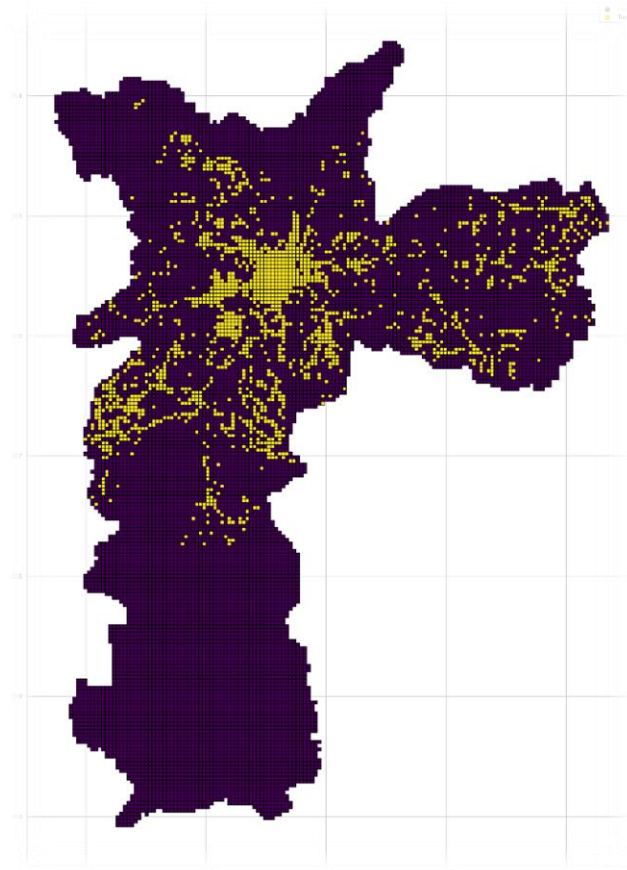


Figura 8 - Distribuição dos *hot spots* na cidade de São Paulo

É possível notar que os pontos críticos se aglomeram em regiões específicas, como na região central e como no Campo Limpo, alguns bairros da Zona Leste como Sapopemba e Itaim Paulista. Observando uma região de exemplo para averiguar os pontos encontrados de maneira mais focalizada:

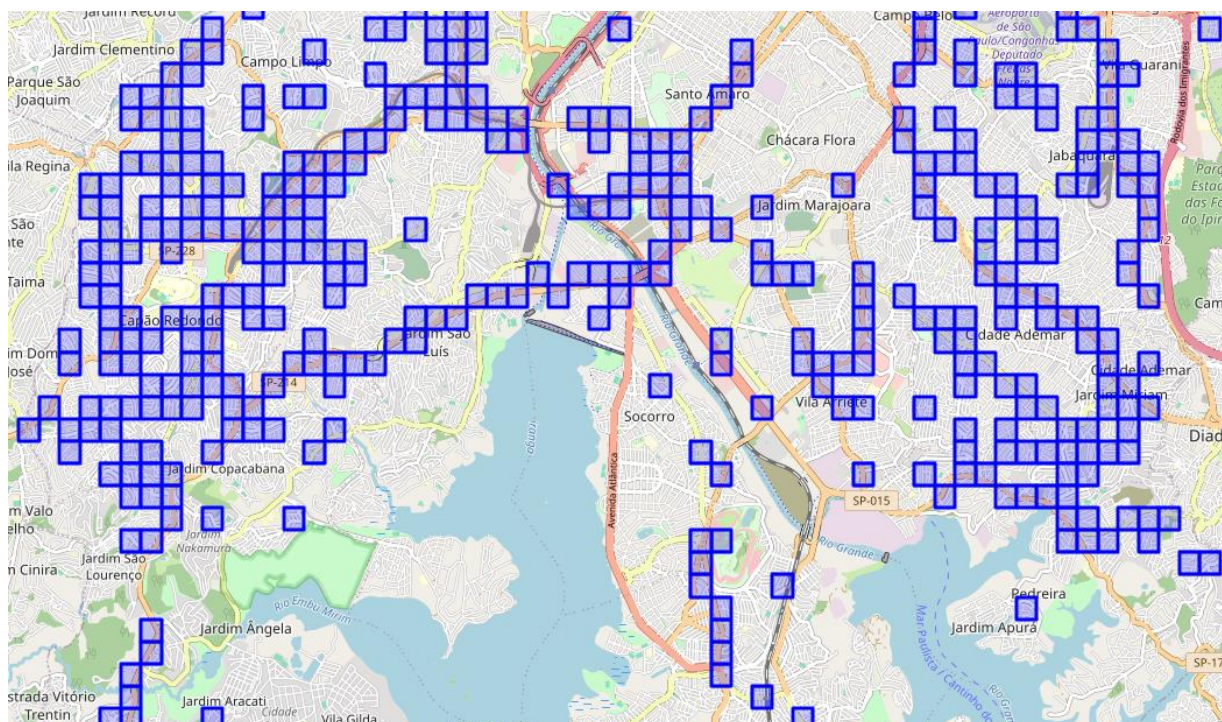


Figura 9 – Exemplo de *hot spots* encontrados na Zona Sul de São Paulo

Como é possível observar na Figura 9, os pontos que aglomeram crimes estão em suma maioria concentrados em Avenidas e saída de metrô, pontos que possibilitam a fácil fuga de um detratador [2], porém alguns lugares se destacam, como a saída do autódromo de Interlagos e o bairro Capão Redondo e Jabaquara, que praticamente são inteiramente preenchidos por *hot spots*.

4.5 Aplicação da Krigagem

Com os dados preparados, será executado o modelo Indicator Kriging utilizando os *hot spots* descobertos pelo Modelo Simples de Poisson, o resultado será um mapa de probabilidades onde cada região terá um valor entre 0 e 1 referente a probabilidade da ocorrência de crime, ou seja, caso um crime ocorra, qual a chance de ele ter acontecido no lugar em questão?

5. Análise dos Resultados

Os resultados do modelo serão analisados visualizando as próprias métricas do modelo, como por exemplo o resultado visualmente obtido, e seu variograma, primeiramente buscando compreender se o modelo conseguiu aprender sobre as características espaciais dos dados e como ele as explica. O modelo também será analisado do ponto de vista quantitativo, observando o quão bem o modelo conseguiu descrever os padrões de crime capturados, atuais, usados para treinar o modelo, e futuros, testado em dados de 2024 (de janeiro até junho).

5.1 Métricas do próprio modelo

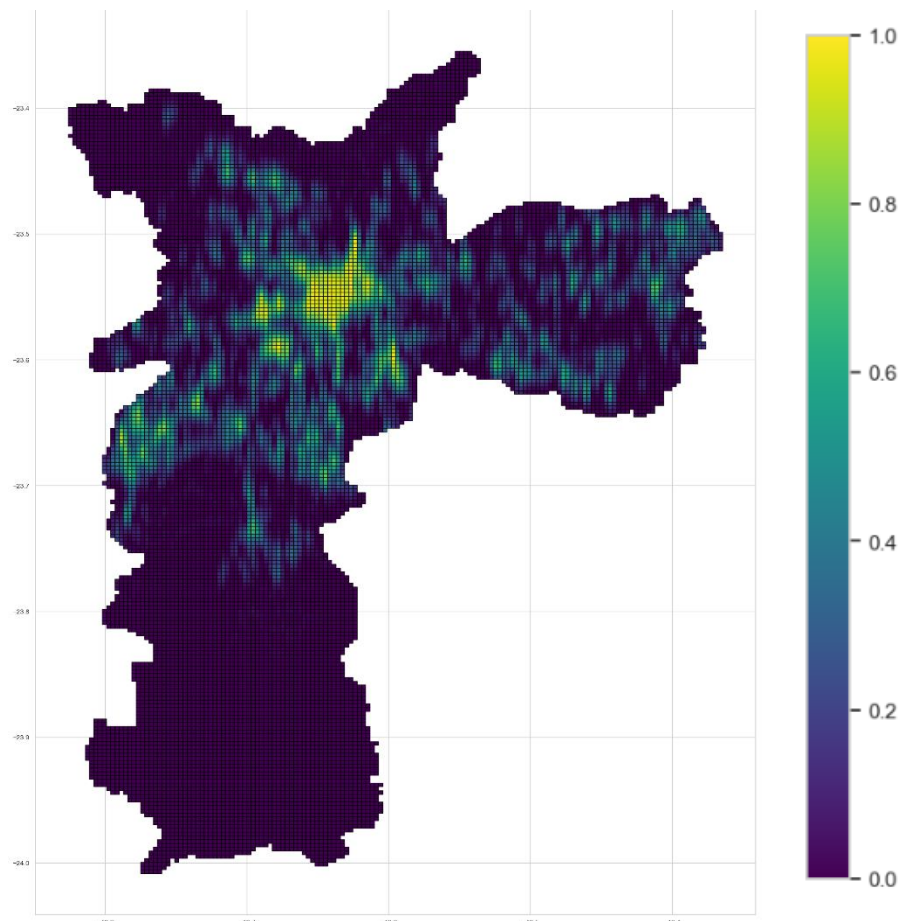


Figura 10 – Mapa de probabilidade gerada pela Indicator Kriging

O mapa gerado que representa a cidade de São Paulo sob a ótica da Krigagem por Indicador contém cada lugar (célula do grid 300x300m) representado por um valor numérico de 0 a 1 e acompanhado visualmente de uma cor na paleta de cores *viridis* (amarelo para valores

mais intensos e roxo para valores menos intensos) de forma que seja possível não só reconhecer padrões de crime qualitativamente, como também com métricas quantitativas.

Ao comparar o mapa da Figura 10 com os *hot spots* da Figura 8, é possível notar que o modelo capturou bem a existência dos *hot spots*, e atrelou eles aos locais mais prováveis de existência de ocorrências, além disso, *hot spots* isolados não possuem uma probabilidade tão alta de ocorrência, indicando zonas seguras que possuem um ou outro lugar perigoso. Por outro lado, *hot spots* cercados por outros *hot spots* indicam zonas de alta criminalidade, caracterizados pelo efeito da “Janela Quebrada” [25] que demonstra que arredores de zonas perigosas costumam ser igualmente perigosos, zonas essas que são bem representadas pelo modelo com valores de risco mais alto (cor amarela, com risco maior que 0,8).

Observemos agora um exemplo mais específico, em uma escala de cor que utiliza uma paleta de cores de tons quentes, na Figura 11:

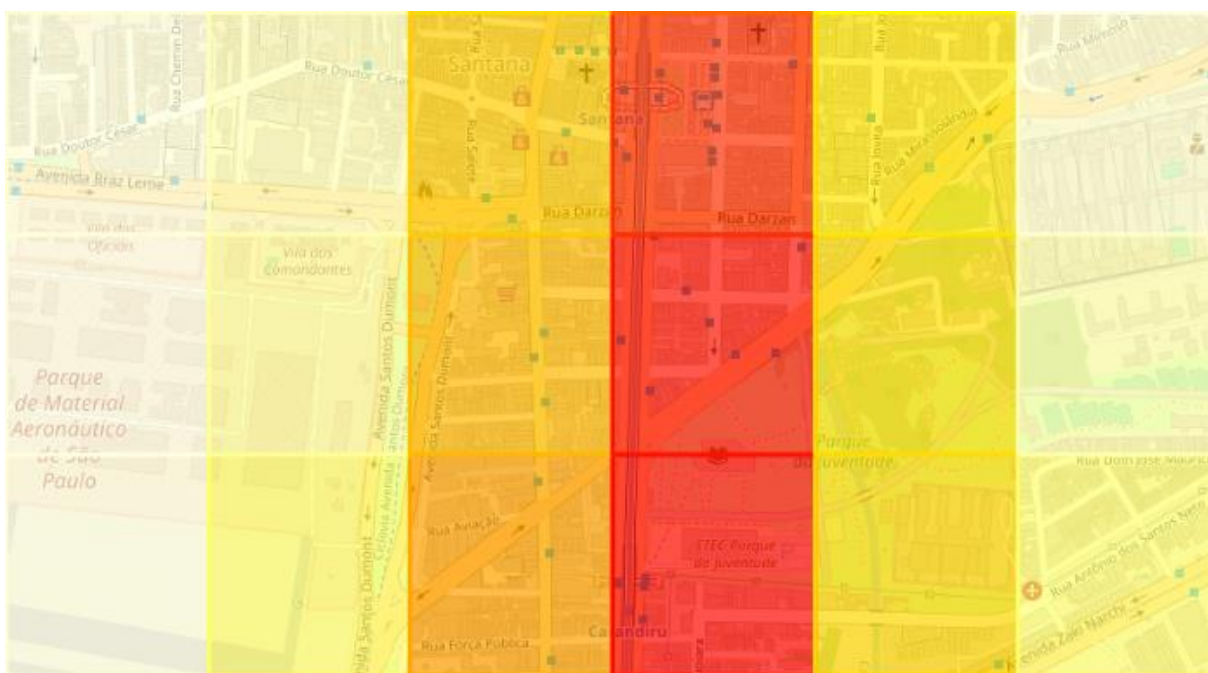


Figura 11 – Exemplo de probabilidades de crime na Zona Norte de São Paulo

Nesse exemplo, é possível identificar duas estações de metrô na Zona Norte paulistana, a Estação Santana e a Estação Carandiru, conectadas pela Avenida Cruzeiro do Sul que concentram crimes e por isso aparecem com a cor vermelha, enquanto seus arredores são alaranjados por conterem riscos de crime, mas não serem tão arriscados quanto aos três *hot spots* identificados. Enquanto as regiões ao canto da imagem, como o Parque de Material Aeronáutico, já aparecem praticamente sem cor, pelo risco de crime presente nele ser nulo.

Como citado no referencial teórico, o modelo Indicator Kriging se baseia em um variograma, função matemática que descreve como a semivariância dos dados muda com a distância espacial entre amostras, observemos como ela se ajustou aos dados, em uma visualização empírica (Figura 12):

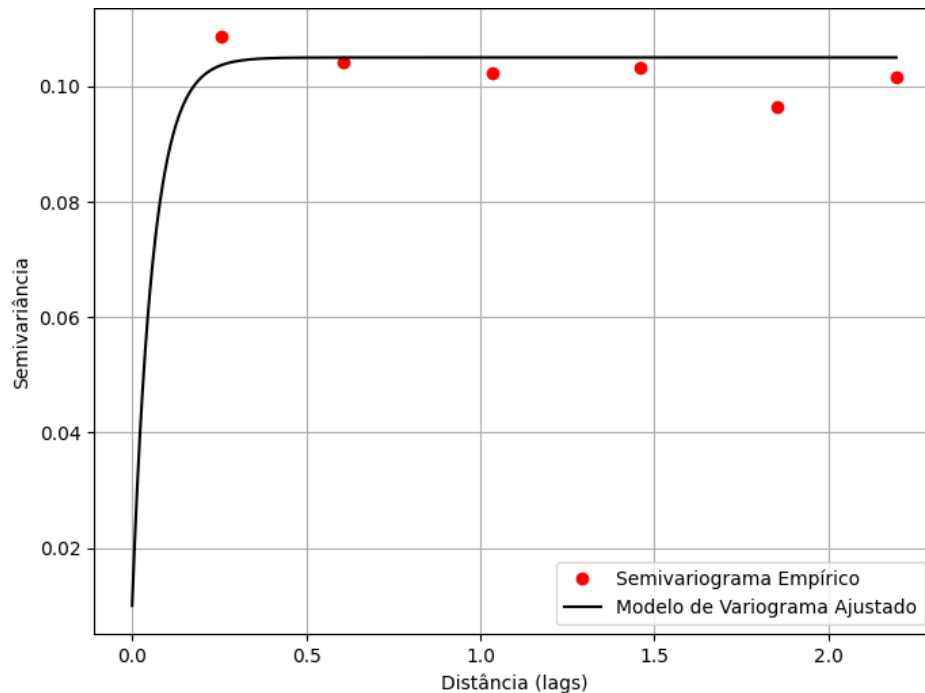


Figura 12 – Representação empírica do variograma

Para compreender como o modelo interpreta a variância espacial dos dados, ou seja, como a dependência espacial muda com a distância entre os dados, foi utilizado uma visualização empírica que constrói pontos artificiais (que no gráfico são os pontos vermelhos) no qual cada um representa a variância em uma determinada distância, e uma curva que se ajusta a esses pontos, sendo ela o próprio variograma em si [26].

A unidade de medida no eixo X se caracteriza como *lags*, que representam a distância entre os dados, como o modelo foi construído com um grid de 300x300m, então cada *lag* é igual a 300 metros. Nota-se que existem mais de um ponto vermelho por *lag*, pois foi identificado que existem diferentes variações entre as direções inferidas, por exemplo na direção norte-sul a variância não necessariamente é a mesma que na direção leste-oeste. Além disso, a partir do segundo *lag*, o variograma deixa de existir, demonstrando que a dependência espacial dos dados, no nosso contexto, se limita a duas células (600 metros).

No eixo Y, é representado a semivariância entre os dados, que é a variância entre eles dividida por 2 (se divide por dois para evitar duplicatas, pois a distância entre os dois pontos é a mesma independente da ordem ser ponto A menos ponto B ou ponto B menos ponto A), no gráfico é possível notar que ela se concentra no valor de 0.1 que significa que valores semelhantes estão fortemente concentrados (o valor de 0 seria a concentração máxima), mas não completamente. Dessa forma, concluímos que o modelo dita que áreas perigosas são, na maioria das vezes, vistas perto de outras áreas perigosas, e essa mesma lei vale para as áreas seguras, interpretação essa que, como visto anteriormente, condiz com a realidade da cidade de São Paulo.

Além do variograma, o modelo também é composto por parâmetros geoestatísticos, que foram encontrados através de um *hyperparameter tuning* que buscou minimizar o erro quadrático médio entre a função do variograma e as variabilidades encontradas nos dados, ou seja, foram testadas diferentes combinações de parâmetros para identificar qual combinação faz o variograma se ajustar melhor nos pontos empíricos. Os valores encontrados foram esses:

Tabela 2 – Parâmetros otimizados do modelo

Parâmetro	Valor
Nugget	0.01
Partial Sill	0.095
Full Sill	0.105
Range	0.06
Anisotropy Scaling	5.5
Anisotropy Angle	90

Segue a explicação de cada parâmetro e seus respectivos impactos no modelo:

Nugget: Variabilidade em distâncias extremamente curtas, que não podem ser explicadas pelo modelo espacial. O *nugget* de 0.01 indica que o ruído em pequenas distâncias é baixíssimo.

Partial Sill: Parte da variância total que é explicada pelo modelo, basicamente *Full Sill* menos o *Nugget*. O *Partial Sill* nesse caso representa cerca de 90.5% da variância total.

Full Sill: Variância total observada nos dados, estabilizada em um valor constante pois a partir dela não há mais aumento significativo na variabilidade entre pontos com o aumento da distância, ou seja, quando a curva se estabiliza.

Range: Alcance da dependência espacial, indicando até qual distância um ponto tem influência sobre outro, considerando que a unidade de medida está em coordenadas decimais, o valor de 0.06 é relativamente alto.

Anisotropy Scaling: A razão entre a direção principal (*anisotropy angle*) e a direção perpendicular a ela, demonstrando que dependendo da direção, a variabilidade muda, de modo que a direção principal é 5,5 vezes mais relevante que a sua perpendicular.

Anisotropy Angle: Levando em consideração que a dependência espacial não é a mesma em todas as direções, o parâmetro *anisotropy_angle* indica qual é a direção que exerce mais influência na variabilidade dos dados, no caso um valor de 90 graus (lembrando que 90 graus no sistema de coordenadas indica a direção leste), muito provavelmente por causa das características da Zona Leste da cidade de São Paulo, que ocupa grande parte da cidade e é relativamente homogênea (tanto em comportamento dos crimes quanto em densidade populacional) quando comparada com as outras zonas.

Tendo em vista o potencial do modelo de descrever as características da criminalidade na área de estudo, a eficácia dos parâmetros otimizados e o modelo em geral terá seus resultados quantitativos analisados na próxima seção.

5.2 Análise quantitativa do modelo

Para aumentar a compreensão sobre a eficiência do modelo em descrever e estimar o padrão de funcionamento do crime sob uma área, se construiu dois modelos análogos para servir de comparação ao modelo com *hyperparameter tuning*: Primeiro, um modelo com parâmetros genéricos, sem otimização, e depois, um modelo com *leave one out cross-validation*, isto é, é retirado uma amostra do modelo (nesse caso um único ponto, “*leave one out*”) e testa-se a eficácia do modelo em prever qual é o valor dessa amostra, que antes estaria no conjunto de dados de treinamento, processo esse que é repetido 100 vezes (“*cross-validation*”) e em estudos anteriores demonstrou ser eficiente em quantificar a capacidade de generalização de um modelo de Krigagem [27].

Para comparar esses três modelos, foram utilizadas métricas academicamente comuns de serem utilizadas para avaliação de modelos espaciais relacionados a criminologia [28], são eles: RMSE (*Root Mean Squared Error*), PAI (*Predictive Accuracy Index*) e MAE (*Mean*

Absolute Error). Além disso, também será calculado a Correlação de Spearman, por causa da natureza não-paramétrica da modelagem do Indicator Kriging. Observemos os resultados:

Tabela 3 – Comparação de métricas entre modelos

<i>Metric</i>	<i>Tuned Model</i>	<i>Generic Model</i>	<i>Cross-validated Model</i>
RMSE	0.214287	0.224017	0.423385
PAI	93.7877%	92.7747%	90.0000%
MAE	0.115442	0.113693	0.244002
Spearman Correlation	0.626688	0.647875	0.456263

Ao analisar os resultados, percebe-se que o modelo tunado obteve melhor performance em quase todas as métricas, superando o modelo genérico e o modelo *cross-validated*, justificando a escolha do *hyperparameter tuning*, porém o modelo genérico se saiu 3,27% melhor na métrica Correlação de Spearman, mostrando que no contexto do estudo, buscar minimizar o erro médio exige uma leve perda na capacidade de detectar tendências gerais.

Observando a capacidade de generalização do modelo Indicator Kriging, a versão com *cross-validation* que também contou com *hyperparameter tuning*, obteve resultados próximos em relação aos outros modelos, com destaque a métrica PAI, que obteve apenas 3.78% pontos de diferença em comparação ao modelo “tunado”. De forma geral, os resultados obtidos foram satisfatórios tanto no que tange a capacidade do modelo de explicar os padrões de crime, quanto de generalizar essas relações para evitar *overfitting*, qualidades que caracterizam bons modelos na área de IA e Big Data [29].

Porém, mesmo no melhor cenário com parâmetros otimizados, o modelo Indicator Kriging apresentou um valor notável de erro, por mais que seja baixo, será melhor analisado na Figura 13, na qual será visualizado resíduos do modelo, com uma transformação logarítmica para uma visualização mais detalhada, a seguir:

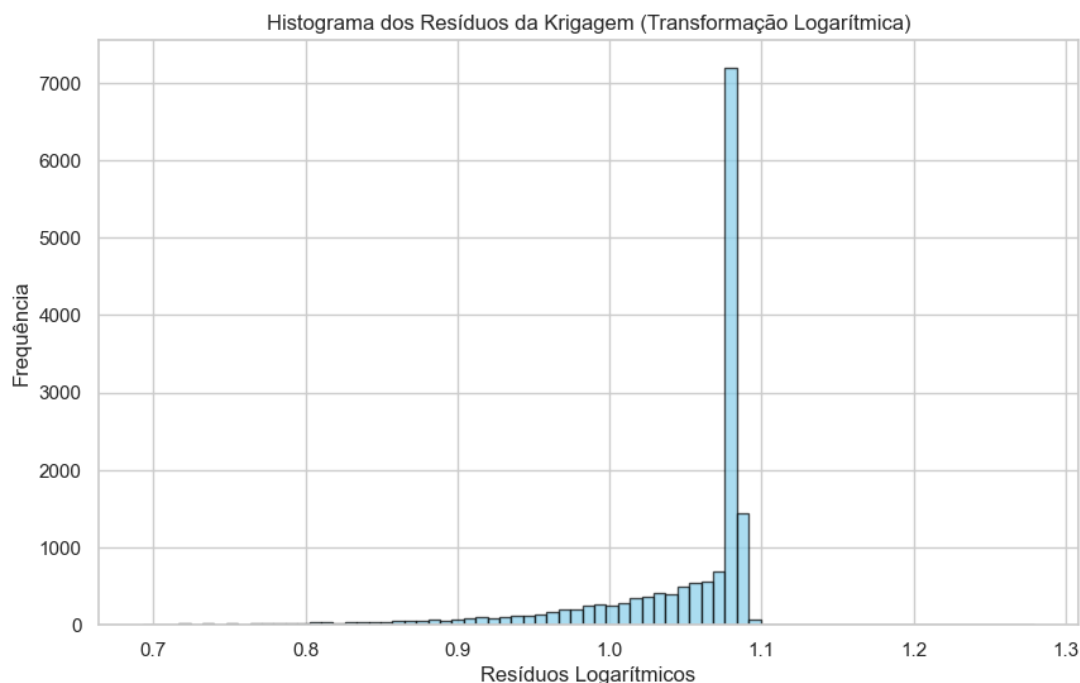


Figura 13 – Histograma de resíduos da Krigagem

Visualizando a distribuição dos resíduos (diferença entre valores inferidos e valores reais), é possível notar que os resíduos se concentram em torno de zero, ou seja, a modelagem do Indicator Kriging realmente fez boas previsões. Entretanto, na distribuição dos resíduos há uma assimetria negativa, indicando que o modelo costuma, mesmo que raramente, subestimar os valores reais, produzindo valores menores do que eles realmente são.

Esse fenômeno ocorre devido a natureza da criminalidade, na qual locais com baixa probabilidade de crime, podem de fato não ter tido crime algum, enquanto é representado por uma probabilidade baixíssima, porém diferente de zero. O resultado desse fenômeno produz resíduos negativos pois o valor inferido é levemente maior que o valor real, isto é, a cauda à esquerda vista no gráfico.

5.2.1 Análise dos padrões de crime entre janeiro e junho de 2024

Por mais que os padrões de crime mudem conforme se altera a infraestrutura urbana e fosse necessário implementar outras técnicas que capturassem não só as características espaciais do crime, mas também seus padrões temporais [30], para que fosse possível determinar os *hot spots* do futuro, A Indicator Kriging propõe que os crimes em um determinado

lugar, independente do tempo, continuam com a mesma distribuição probabilística, conceito esse que foi estudada no gráfico a seguir (Figura 14):

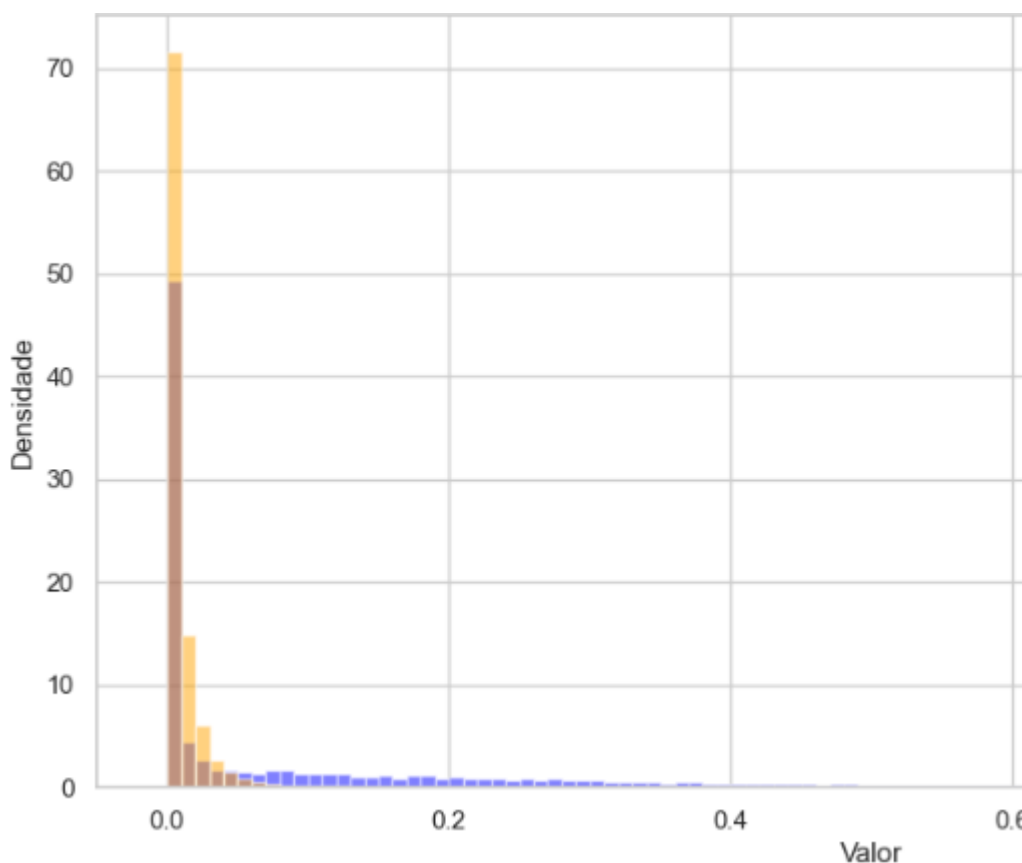


Figura 14 – Comparação da distribuição de crimes reais e previstos

Neste gráfico, foram construídos dois histogramas: Um com a distribuição dos crimes normalizados, que ocorreram no período analisado em 2024, em amarelo, e a distribuição das probabilidades por local na Indicator Kriging de 2017 a 2023, em azul (a sobreposição dos dois histogramas ficou em marrom). Nele, é possível notar que o fenômeno da assimetria negativa persiste, pois a distribuição de probabilidades da Krigagem é suavizada, mas mesmo assim, as duas distribuições são extremamente semelhantes, com muitos lugares com poucos ou nenhum crime, e pouquíssimos lugares concentrando a maior parte das ocorrências de crime.

Entretanto, por causa do intervalo dos dados serem pequenos, a visualização pode ser dificultada, vejamos o mesmo gráfico, porém com uma transformação logarítmica aplicada, na Figura 15:

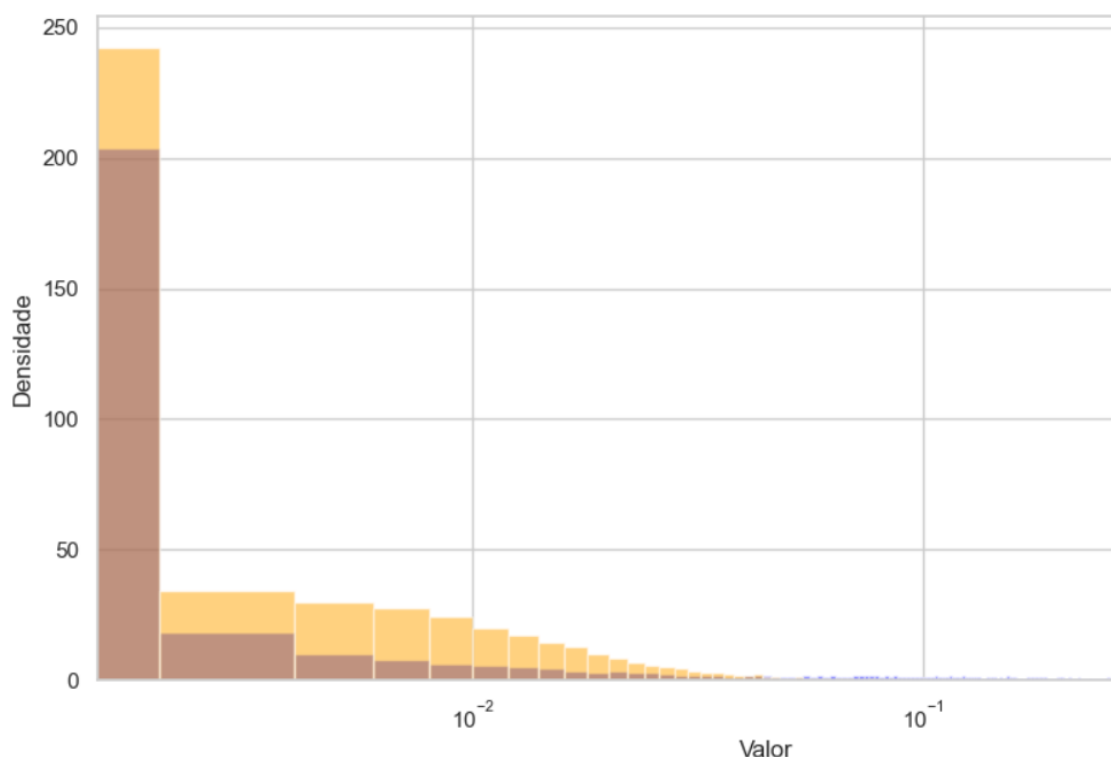


Figura 15 – Distribuição de crimes reais e previstos com transformação logarítmica

Em marrom temos a sobreposição entre as duas distribuições, em amarelo temos a distribuição dos crimes ocorridos, e em azul, a distribuição da probabilidade dos crimes previstos, percebe-se que a probabilidade de crimes prevista está quase que completamente em uma intersecção com o a distribuição real, com exceção de alguns poucos casos que podem ser vistos minuciosamente na direita do gráfico, blocos azuis quase que milimétricos. Porém, a distribuição real em amarelo destoa levemente com valores mais concentrados em torno do zero, sendo uma boa validação da teoria vista anteriormente de que muito dos crimes acontecem em poucos lugares.

A comprovação de que o crime possui relações específicas com o ambiente, não só justifica a utilização de modelos geoestatísticos como a Indicator Kriging, como também demonstra que o ambiente é um fator extremamente importante para a análise criminológica.

Para avaliar mais especificamente o quanto o modelo consegue explicar cada tipo de local, dos mais perigosos aos menos perigosos em 2024, foi construído uma curva de calibração, demonstrada no próximo gráfico (Figura 16):

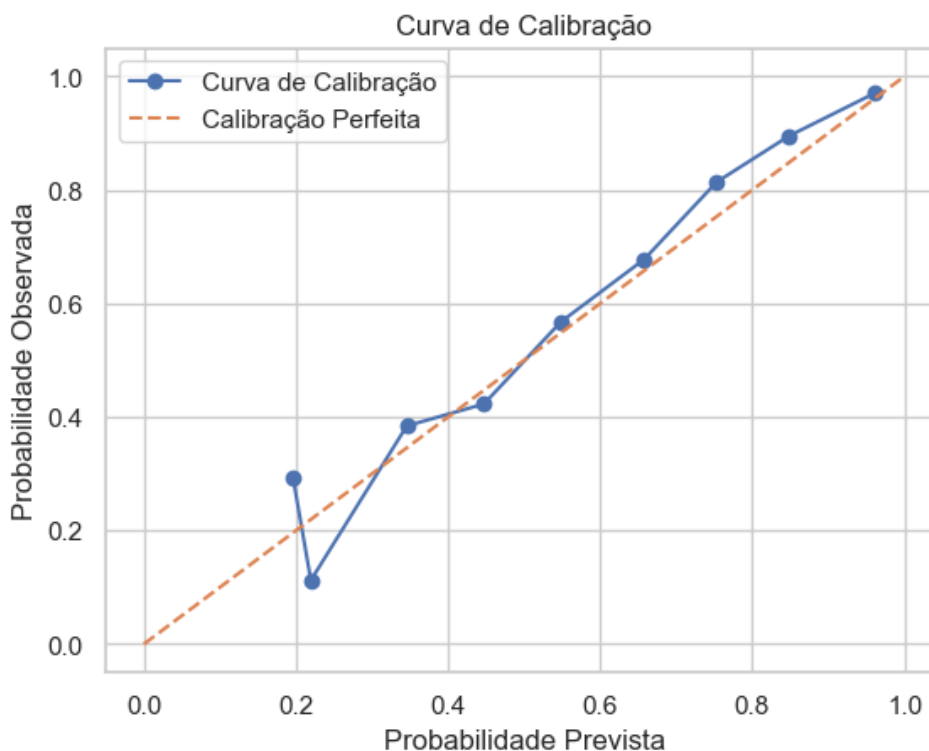


Figura 16 – Curva de Calibração das probabilidades previstas e observadas

Este tipo de gráfico funciona da seguinte forma: O eixo X representa as probabilidades previstas, o quanto o modelo previu a chance de ocorrência em cada ponto do grid, enquanto o eixo Y representa as probabilidades observadas, ou seja, o que realmente aconteceu. Essa relação entre as duas variáveis é vista na linha azul, que mostra a relação entre o valor previsto e o valor observado, e atrás dela há uma linha tracejada em laranja, que representa um modelo perfeitamente calibrado, no qual as previsões feitas se alinham exatamente ao resultado esperado, isto é, quanto mais próximo a linha azul estiver da linha laranja, melhor o modelo é.

Ao interpretar o gráfico, é perceptível que nas menores faixas, próximo de 0.2, a linha azul está abaixo da linha laranja, sugerindo que o modelo subestima o valor real das ocorrências em locais com baixa probabilidade de ocorrência, devido a suavização natural de modelos probabilísticos [31], porém é notável que a partir da faixa de probabilidade de 0.3 para cima, o modelo possui um bom ajuste, resultando em um bom desempenho geral.

6. Conclusão

O modelo Indicator Kriging se mostrou eficiente em explicar as características da variabilidade espacial da criminalidade, levando em consideração que o crime é um fenômeno extremamente complexo e que foi utilizado somente a variável espaço para tentar explicar e prever a quantidade de ocorrências. O modelo demonstra resultados animadores sobre a técnica utilizada, como por exemplo um PAI de 93,78%. Resultado esse que demonstra também a importância do estudo ambiental para compreender a criminalidade, afinal, na área de IA e Big Data são pouquíssimos casos que apenas uma variável consegue determinar tanta influência em um modelo. Além disso, a escolha da Indicator Kriging para resolver esse problema também confirmou a hipótese do modelo servir bem ao problema proposto, principalmente no quesito características dos dados, pois os indicadores da Krigagem (0 ou 1) podem ser traduzidos em ser um *hot spot* ou não ser um *hot spot* e a subnotificação (crimes que não são relatados para as autoridades) exigem o uso de um modelo não-paramétrico, assim como a Indicator Kriging.

Apesar de ser um modelo tecnicamente complexo e os dados serem extremamente volumosos, apesar de incompletos, a realização desse estudo só foi possível graças as técnicas de Big Data, que permitiram tratar e manipular os dados de maneira que operações matemáticas não exigissem um alto grau de esforço e poder computacional. Possibilitando que o modelo proposto por esse estudo seja implementado na infraestrutura pública exigindo uma baixíssima quantidade de recursos e performance de hardware, de modo a ser uma boa alternativa para a problemática de identificação de *hot spots* e previsão de crimes no cenário brasileiro.

No que tange trabalhos futuros sobre esse estudo, há a possibilidade de coletar diferentes tipos de dados que pudessem informar sobre a existência de ocorrências de crimes que foram subnotificados, como dados de redes sociais, contendo postagens de pessoas relatando furtos e suas respectivas localizações, também é possível utilizar o modelo desenvolvido nesse estudo para aprimorar outros modelos de previsão de crime, aumentando a interpretabilidade que eles detêm sob diferentes padrões de crimes, lidando principalmente com a subnotificação.

7. Referências Bibliográficas

- [1] Carvalho, Beatriz. O. (2021). Atlas da Dinâmica Criminal em São Paulo (SP): roubos a transeuntes e de veículos. São Paulo: Núcleo de Estudos da Violência, Universidade de São Paulo. <https://doi.org/10.11606/9786587621791>
- [2] G1. (2023) Cidade de SP tem um registro de celular roubado ou furtado a cada 3 minutos. G1 - Monitor da Violência. <https://g1.globo.com/monitor-da-violencia/noticia/2023/03/26/cidade-de-sp-tem-um-registro-de-celular-roubado-ou-furtado-a-cada-3-minutos.ghhtml>
- [3] FOLHA DE S. PAULO. (2018) Maioria das vítimas de roubo e furto em SP não relata crime à polícia. <https://www1.folha.uol.com.br/cotidiano/2018/09/maioria-das-vitimas-de-roubo-e-furto-em-sp-nao-relata-crime-a-policia.shtml>
- [4] Prathap, B.R. (2022). Artificial Intelligence and Machine Learning for EDGE Computing, Chapter 7 - Geospatial crime analysis and forecasting with machine learning techniques. <https://doi.org/10.1016/B978-0-12-824054-0.00008-3>
- [5] Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the South African Institute of Mining and Metallurgy, volume (52), 119 - 139. https://hdl.handle.net/10520/AJA0038223X_4792
- [6] Brantingham, P.L., Brantingham, P.J (1993). Nodes, Paths and Edges: Considerations on the Complexity of Crime and the Physical Environment. Journal of Environmental Psychology, volume (130), número (2), 3–28. [https://doi.org/10.1016/S0272-4944\(05\)80212-9](https://doi.org/10.1016/S0272-4944(05)80212-9)
- [7] Algahtany, M., Kumar, L., Barclay, E. et al (2018). The spatial distribution of crime and population density in Saudi Arabia. Crime Prev Community Saf, volume (20), 30–46. <https://doi.org/10.1057/s41300-017-0034-3>
- [8] Sherman, Lawrence W.; Gartin, Patrick R.; Buerger, Michael E. (1989). Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology*, volume (27), número (1), 27-55. DOI: 10.1111/j.1745-9125.1989.tb00862.x.
- [9] Braga, Anthony A. et al. (2019) Hot spots policing and crime reduction: an update of an ongoing systematic review and meta-analysis. Journal of Experimental Criminology. Volume (15), 289-311. DOI: 10.1007/s11292-019-09372-3.
- [10] Journel, A.G. (1983) Nonparametric estimation of spatial distributions. Mathematical Geology, volume (15), 445–468. <https://doi.org/10.1007/BF01031292>
- [11] Osgood, D.W. (2000). Poisson-Based Regression Analysis of Aggregate Crime Rates. Journal of Quantitative Criminology, volume (16), 21–43. <https://doi.org/10.1023/A:1007521427059>

[12] Freitas Junior J. C.; Gastaud Maçada A. C.; Oliveira M.; Brinkhues R. A. (2016). Big Data e Gestão do Conhecimento: Definições e Direcionamentos de Pesquisa. *Revista Alcance*, volume (23), número (4), 529-546, DOI: [alcance.v23n4.p529-546](https://doi.org/10.1590/alcance.v23n4.p529-546).

[13] Usman, Umar & Suleiman, Shamsuddeen & Yar'adua, S & Alfa, Abduljalil. (2021). Spatial Analysis on the Crimes Rate Using Regression Kriging Model. (A Case Study of Katsina State), volume (8). 64-69. <https://www.researchgate.net/publication/355470006>

[14] Yu, Hongjie, Lin Liu, Bo Yang, and Minxuan Lan. (2020). Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method. *ISPRS International Journal of Geo-Information*, volume (9), número (12) 732. <https://doi.org/10.3390/ijgi9120732>

[15] Jayaweera, I. et al. (2015). Crime Analytics: Analysis of Crimes Through Newspaper Articles. *Proceedings of the 2015 IEEE International Conference on Information and Automation for Sustainability*. 1-6. DOI: [10.1109/ICIAFS.2015.7194661](https://doi.org/10.1109/ICIAFS.2015.7194661).

[16] Chen, X.; Cho, Y.; Jang, S. Y (2015). Crime prediction using Twitter sentiment and weather. *IEEE Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 63-68. DOI: [10.1109/SIEDS.2015.7117012](https://doi.org/10.1109/SIEDS.2015.7117012).

[17] Khalid, S. et al. (2018). Network constrained spatio-temporal hotspot mapping of crimes in Faisalabad. *Applied Spatial Analysis and Policy*, volume (11), número 4, 599-622. DOI: [10.1007/s12061-017-9230-x](https://doi.org/10.1007/s12061-017-9230-x).

[18] Lisowska-Kierepka A. (2022). How to analyse spatial distribution of crime? Crime risk indicator in an attempt to design an original method of spatial crime analysis. *Cities*, volume (120), 103403, <https://doi.org/10.1016/j.cities.2021.103403>

[19] Alves, L.G.A.; Ribeiro H. V.; Rodrigues F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A*, volume (505), 435-443, DOI: [10.1016/j.physa.2018.03.084](https://doi.org/10.1016/j.physa.2018.03.084).

[20] Garcia-Zanabria, G.; Gomez-Nieto, E.; Silveira, J.; Poco, J.; Nery, M.; Adorno, S.; Nonato, L. G. (2020). Mirante: A visualization tool for analyzing urban crimes. *33rd SIBGRAPI Conference on Graphics, Patterns and Images*. DOI: [10.1109/SIBGRAPI51738.2020.00028](https://doi.org/10.1109/SIBGRAPI51738.2020.00028).

[21] Secretaria de Segurança Pública do Estado de São Paulo, "Celulares Subtraídos de 2017 a 2024," Portal SSP, 2024. [Online]. <https://www.ssp.sp.gov.br/assets/estatistica/transparencia/baseDados/celularesSub/>

[22] BRASIL. Ministério da Saúde. Primeiro caso de COVID-19 no Brasil permanece sendo o de 26 de fevereiro. Ministério da Saúde, 17 jul. 2020. www.gov.br/saude/pt-br/assuntos/noticias/2020/julho/primeiro-caso-de-covid-19-no-brasil-permanece-sendo-o-de-26-de-fevereiro.

[23] PGE - Procuradoria Geral do Estado de São Paulo. Presos – Parte 9. Biblioteca Virtual, 2024. <https://www.pge.sp.gov.br/centrodeestudos/bibliotecavirtual/presos/parte9.htm>.

[24] IBGE - Instituto Brasileiro de Geografia e Estatística. Malhas territoriais. Disponível em: <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html>.

[25] Wilson, J. Q.; Kelling, G. L. (1982). Broken windows: the police and neighborhood safety. *The Atlantic*, Boston, volume (249), número (3), 29-38. https://media4.manhattan-institute.org/pdf/_atlantic_monthly-broken_windows.pdf

[26] Matheron, G. Principles of Geostatistics. (1963). *Economic Geology*, volume (58), 1246-1266. <https://doi.org/10.2113/gsecongeo.58.8.1246>

[27] Lundgren, W. J. C.; Silva, J. A. A.; Ferreira, R. L. C. (2017). A precisão da estimativa do erro da krigagem pela validação cruzada. *Floresta e Ambiente*, volume (24), e00124114. DOI: 10.1590/2179-8087.124114.

[28] Du, Yingjie, and Ning Ding. (2023). A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods. *ISPRS International Journal of Geo-Information*, volume (12), número (6), 209. <https://doi.org/10.3390/ijgi12060209>.

[29] Sullivan, E. (2020). Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, volume (73), 109 - 133. <https://doi.org/10.1093/bjps/axz035>.

[30] García-Zanabria, G., Raimundo, M. M., Poco, J., Nery, M. B., Silva, C. T., Adorno, S., & Nonato, L. G. (2022). CriPAV: Street-Level Crime Patterns Analysis and Visualization. *IEEE Transactions on Visualization and Computer Graphics*, volume (28), número (12), 4000-4012. DOI: 10.1109/TVCG.2021.3111146

[31] Olea, Ricardo A., e Vera Pawlowsky (1996). Compensating for Estimation Smoothing in Kriging. *Mathematical Geology*, volume (28), número (4), 407-415. <https://doi.org/10.1007/BF02083653>.