UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MATHEUS IZIDORO DE ALMEIDA

PAINEL CRIMINAL: FERRAMENTA PARA AUXÍLIO NO MONITORAMENTO E ANÁLISE DE OCORRÊNCIAS CRIMINAIS

CURITIBA

2022

MATHEUS IZIDORO DE ALMEIDA

PAINEL CRIMINAL: FERRAMENTA PARA AUXÍLIO NO MONITORAMENTO E ANÁLISE DE OCORRÊNCIAS CRIMINAIS

Criminal Dashboard: assisting software in the monitoring and analysis of criminal incidents

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Luiz Celso Gomes Junior

CURITIBA

2022

MATHEUS IZIDORO DE ALMEIDA

PAINEL CRIMINAL: FERRAMENTA PARA AUXÍLIO NO MONITORAMENTO E ANÁLISE DE OCORRÊNCIAS CRIMINAIS

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 15/dezembro/2022

Prof. Dr. Luiz Celso Gomes Junior
Doutor em Ciência da Computação
Universidade Tecnológica Federal do Paraná

Prof. Dr. Ricardo Lüders
Doutor em Engenharia Elétrica
Universidade Tecnológica Federal do Paraná

Me. Manoel Flavio Leal
Mestre em Computação Aplicada
Universidade Tecnológica Federal do Paraná

CURITIBA
2022

À minha família, que me abraçou e reanimou
em cada vez que minhas forças se esgotaram.

AGRADECIMENTOS

A Deus, onde pela fé pude experimentar Seu descanso e paz durante a jornada.

A minha família, especialmente minha mãe Célia e meu pai Donizete, que não caberia em palavras o que significam e fazem por mim, e minha irmã Mayara e meu irmão Jonatas que ouviram sobre meus medos e inseguranças e os compreenderam.

A minha melhor amiga e meu bem, Celini, que me ensinou sobre o amor e empatia. Obrigado por ter cuidado de mim durante todos os dias, Naninha.

Ao meu orientador Luiz, que compreende a docência como poucos, possibilitando muito mais que a entrega dos resultados aqui relatados mas sim especialmente interessado em meu aprendizado, permitindo que eu pudesse me desenvolver nos temas que mais despertam minha curiosidade.

A UTFPR, que mesmo enfrentando os diversos ataques ocorridos nos últimos anos contra o ensino superior público, é um ambiente onde a democracia, a ciência e o desenvolvimento tecnológico resistem. A todo o corpo docente desta universidade, em especial aos professores do DAINF.

A todos os demais amigos que compartilharam espaço e tempo, contribuindo para esta conquista.

A todos vocês, meu muito obrigado.

O mundo está cheio de coisas óbvias que
ninguém jamais observa.
Onde pensa que estive?
(DOYLE, 1901)

RESUMO

A tomada de decisões acertadas no âmbito da segurança pública é primordial para assegurar a qualidade de vida e bem estar da população. Para que de fato sejam realizadas estratégias, ações e decisões acertadas, é necessário que estas sejam respaldadas por informações precisas e de fácil interpretação. Entretanto, extrair informações de dados nem sempre ocorre de maneira simplificada e portanto pode haver uma lacuna entre as atribuições e especialidades dos agentes de segurança e o conhecimento sobre as ferramentas necessárias para uma abordagem de extração e análise de dados. Este trabalho consiste no desenvolvimento de uma ferramenta computacional que possibilita a visualização de dados fornecidos pela Secretaria de Estado de Segurança Pública do Paraná (SESP) bem como pela Guarda Municipal de Curitiba (GMC) que contém registros de ocorrências atendidas por estes órgãos nos últimos anos na cidade de Curitiba. A ferramenta possui a funcionalidade de monitoramento e detecção de anomalias nos padrões espaciais e temporais de crimes, contando com visualizações de séries temporais de ocorrências e mapas destacando as regiões *hotspots* e *outliers*. Também foi implementada uma aba de detecção de *outliers* espaço-temporais, que destaca o mês e o bairro *outlier* num mapa interativo. Além do desenvolvimento desta ferramenta, foi realizada uma análise de explicabilidade dos *outliers* encontrados.

Palavras-chave: segurança pública; ciência de dados geográficos; detecção de *outliers*; explicabilidade de *outliers*.

ABSTRACT

Making the right decisions in the field of public safety is essential to ensure the quality of life and well-being of the population. In order for the right strategies, actions and decisions to be carried out, it is necessary that these are supported by accurate and easy-to-interpret information. However, extracting information from data is often a complex task and therefore there may be a gap between the attributions and specialties of security agents and the knowledge about the tools necessary for an approach of data mining and analyzing. This work consists in the development of a computational tool that allows the visualization of data provided by the State Secretariat for Public Security of Paraná as well as by the Curitiba Municipal Guard that contains records of occurrences attended by these entities in recent years in the city of Curitiba. The tool has the functionality of monitoring and detecting anomalies in the spatial and temporal patterns of crimes, with visualizations of occurrences temporal series and maps highlighting the hotspots and outlier regions. A spatio-temporal outlier detection tab was also implemented, which highlights the month and outlier neighborhood in an interactive map. In addition to the development of this tool, an explainability analysis of the detected outliers was carried out.

Keywords: public security; geographic data science; outlier detection; outlier explainability.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

ACF	*Autocorrelation Function*
ADF	*Augmented Dickey–Fuller*
ARIMA	*AutoRegressive Integrated Moving Average*
GMC	Guarda Municipal de Curitiba
IPEA	Instituto de Pesquisa Econômica Aplicada
IPPUC	Instituto de Pesquisa e Planejamento Urbano de Curitiba
LISA	*Local Indicator of Spatial Association*
LOF	*Local Outlier Factor*
MLE	*Maximum Likelihood Estimator*
NMF	*Non-Negative Matrix Factorization*
OAM	*Outlying Aspect Mining*
OLS	*Ordinary Least Squares*
PACF	*Partial Autocorrelation Function*
PCPR	Polícia Civil do Paraná
PMC	Prefeitura Municipal de Curitiba
RMC	Região Metropolitana de Curitiba
SESP	Secretaria de Estado de Segurança Pública do Paraná
XAI	*Explainable Artificial Intelligence*

SUMÁRIO

1 INTRODUÇÃO

Muitos dos países considerados emergentes ou mesmo países desenvolvidos enfrentam problemas relativos à segurança pública. Como exemplo, países em desenvolvimento da América Latina - particularmente o Brasil - mantêm elevadas taxas de homicídio desde os anos 90 até a atualidade (ROSER; RITCHIE, 2013) e os Estados Unidos - a maior economia do mundo - apresentou aumento nos assassinatos no ano de 2020 (BECKETT, 2021).

Crimes menores e outras ocorrências são frequentemente notificadas pelas forças policiais, guardas municipais e Secretarias de Estado de Segurança Pública. Existem portais como o da Prefeitura Municipal de Curitiba (CURITIBA, 2015), que permitem acesso a dados de ocorrências atendidas pela Guarda Municipal nos últimos anos e o Atlas da Violência (IPEA, 2017), portal desenvolvido pelo Instituto de Pesquisa Econômica Aplicada Instituto de Pesquisa Econômica Aplicada (IPEA) que disponibiliza informações sobre violência no Brasil.

Há avanços no desenvolvimento de ferramentas para análise de dados, entretanto muitas destas ferramentas exigem habilidades prévias de programação, manipulação de bancos de dados, aplicação de modelos estatísticos, entre outros conhecimentos que não fazem parte das atribuições dos agentes e tomadores de decisão da segurança pública. A existência de dados sobre ocorrências policiais e a possibilidade de tornar a análise e visualização destes dados facilitada para os agentes motivou a pesquisa e desenvolvimento deste trabalho.

Ainda, considera-se também a importância da detecção de *outliers* como forma de enriquecimento de análises de ocorrências criminais, já que a detecção de *outliers* é um passo em direção à descoberta de padrões criminais da cidade.

O objetivo principal deste trabalho é (i) auxiliar os agentes de segurança pública em suas tomadas de decisões e (ii) realizar a detecção de *outliers* nos registros criminais da cidade.

Os objetivos específicos deste trabalho são:

- Desenvolver uma ferramenta de visualização e análise de dados criminais onde é possível realizar a filtragem por data e tipo de ocorrência;
- Aplicar modelos de detecção de anomalias espaciais, temporais e espaço-temporais a fim de se encontrar padrões criminais da cidade;
- Analisar a explicabilidade dos modelos de detecção de anomalias utilizados

O restante do texto está organizado da seguinte forma: o Capítulo 2 aborda brevemente os principais conceitos e técnicas necessários para a compreensão das análises realizadas bem como demonstra e compara trabalhos de teor semelhante a este. O Capítulo 3 aborda a metodologia utilizadas para a realização das implementações e investigações executadas bem como apresenta um relato da análise exploratória dos dados utilizados. A preparação dos dados, criação de variáveis, análises efetuadas, aplicações de modelos e seus resultados para as análises temporal, espacial e espaço-temporal estão descritas no Capítulo 4. Neste Capítulo

também estão relatadas as análises de explicabilidade realizadas sobre os *outliers* encontrados. No Capitulo 5 está relatada a fase de criação da ferramenta e os resultados obtidos por conta da sua implementação e utilização. Por fim, as conclusões e discussões a respeito da obra estão expostas no Capítulo 6.

2 FUNDAMENTOS E TRABALHOS RELACIONADOS

Neste capítulo estão contidos os conceitos dos principais tópicos estudados e abordados na realização deste trabalho bem como são apresentados artigos em que estes fundamentos foram matéria de estudo e aplicação. A primeira Seção apresentada (2.1) trata dos conceitos e métodos relacionados à detecção de *outliers*. Na sequência, são apresentadas três seções correspondentes as três análises desenvolvidas neste estudo: Análises Temporal (2.2), Espacial (2.3) e Espaço-Temporal (2.4). Finalmente, as Seções (2.5) e (2.6) são dedicadas aos fundamentos relacionados à Inteligência Artificial Explicável e à técnica OAM de explicabilidade de *outliers*.

2.1 Detecção de *Outliers*

Optou-se aqui por utilizar o termo em inglês *outlier* em detrimento do termo anomalia da língua portuguesa para se referir as observações num conjunto de dados que são notadamente diferentes das demais observações do conjunto. O uso do termo anomalia pode evocar determinada conotação negativa a depender do domínio e do fenômeno a ser estudado e portanto decidiu-se pela utilização do termo em inglês, que traduzido significa a expressão *fora de série*, resultando assim em maior neutralidade. Uma interessante definição de *outlier*, segundo Hawkins (1980, p.1), indica que "*outlier* é uma observação que desvia tanto das outras observações de modo a levantar suspeita de que foi gerada por um mecanismo diferenciado".

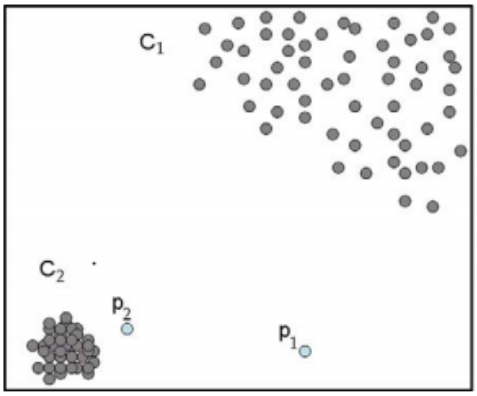
Existem diferentes técnicas para se alcançar o objetivo de detectar um *outlier* e a escolha do método a ser utilizado dependerá de diversos aspectos. Por exemplo leva-se em conta: a natureza dos dados (monovariados, multivariados, contínuos, categóricos, binários, mistos); um *outlier* pode depender do contexto em que está inserido (exemplo: determinados níveis de chuvas podem ser considerados típicos para determinado período enquanto os mesmos níveis podem não ser esperados em demais épocas); em determinadas aplicações busca-se eliminar os *outliers* para melhor modelagem do fenômeno enquanto que em outras o objetivo principal do problema é a detecção dos registros fora de série (exemplo: detecção de fraudes em cartões de crédito); entre outros fatores. (CHANDOLA; BANERJEE; KUMAR, 2009)

Existem abordagens estatísticas e visuais como a utilização de *boxplots* e distribuições, bem como a aplicação de modelos de aprendizado de máquina desenvolvidos para esta finalidade onde as saídas dos modelos podem ser *scores* dos registros avaliando o seu grau de *outlier* (*outlierness*), *labels* indicando se os registros são típicos ou atípicos ou o *cluster* de pertencimento do registro para modelos baseados em distância. Neste trabalho foram aplicadas duas abordagens cujos fundamentos estão descritos a seguir.

2.1.1 *Local Outlier Factor* (LOF)

O modelo *Local Outlier Factor* (LOF) atribui um *score* a cada ponto no conjunto de dados de modo a considerar a densidade de pontos ao redor de um registro individual, já que em fenômenos da vida real os conjuntos de dados podem estar organizados de maneira mais complexa (BREUNIG *et al*., 2000) e *outliers* dependentes do contexto podem se apresentar em detrimento de registros singulares de *outliers* globais. Portanto, o modelo *LOF* permite identificar um *outlier* não de maneira binária, mas associando um grau de desvio do ponto em face a região de densidade em que se encontra. A Figura 1 representa um conjunto de ponto distribuídos em diferentes *clusters* de diferentes densidades. Neste exemplo, pode-se dizer que o ponto p_1 é um *outlier* global enquanto que o ponto p_2 pode se tratar de um *outlier* local tendo em vista o cluster C_2 mas talvez não um *outlier* global já que este se encontra próximo ao *cluster* C_2.

Figura 1 – *Clusters* de pontos com diferentes densidades



Fonte: Mohan (2018).

2.1.2 *Z-Score*

O *Z-Score* - também denominado Escore Padronizado - é uma medida numérica que indica em quantos desvios padrão uma observação se distancia da média. *Z-Scores* positivos indicam que o valor do registro encontra-se acima de média e negativos indicam valor abaixo da média. Este valor é associado à curva de distribuição normal de forma que dado um escore Z, para o intervalo $[+Z, -Z]$, existe uma área correspondente a esta distribuição que indica o percentual da população dentro deste intervalo. Desta forma, é possível dizer que o percentual da população fora do intervalo está mais distante da média do que o percentual da população pertencente ao intervalo. Assim, é possível escolher valores de Z em que a pequena parte da população mais distante da média pode ser classificada como *outlier*. A Equação 1 indica o cálculo do *Z-Score*

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

onde μ e σ são a média e o desvio padrão da amostra respectivamente.

2.2 Análise Temporal

2.2.1 Séries Temporais

Uma série temporal é um conjunto de observações de um determinado fenômeno realizadas sequencialmente ao longo do tempo. Geralmente, estas observações estão em intervalos de tempo regulares, isto é, igualmente espaçados (a cada dia, semana, mês, trimestre, ano, etc).

Uma das principais características que diferem as séries temporais de demais conjuntos de dados é ordenamento entre as observações. Por exemplo, em registros de temperaturas percebe-se que as temperaturas variam sazonalmente de acordo com as estações do ano, portanto caso um registro corresponda ao período de inverno, há grandes chances de que as temperaturas registradas no período sejam menores do que as registradas nas demais épocas. Ainda, se em determinada ocasião há um registro de calor, existe grande possibilidade de que o próximo dia registre altas temperaturas novamente. Estes dois fenômenos são chamados de sazonalidade e tendência, que, juntamente à componente chamada de resíduo - erro aleatório geralmente distribuído de forma Normal com média 0 - compõem uma série temporal. (ANISH, 2020)

Existem diferentes abordagens para a modelagem de fenômenos descritos através de Séries Temporais, frequentemente com o objetivo de realizar previsões destes fenômenos. Contudo, neste trabalho objetiva-se a descoberta de registros de *outliers* com o auxílio destes modelos. Para tal, foram utilizadas duas técnicas: ARIMA e Regressão com OLS, desenvolvidas a seguir.

2.2.2 ARIMA

AutoRegressive Integrated Moving Average (ARIMA) é um modelo estatístico para Séries Temporais que introduz o conceito de *lag* em sua modelagem. Determinado valor num instante t, sendo este y_t, pode ser encontrado através da relação entre seus valores passados $y_{t-1}, y_{t-2}, ..., y_{t-n}$. A inclusão destas váriaveis passadas trata-se da relação de atraso - *lag* - com os valores seguintes. Mais especificamente, o modelo ARIMA emprega a combinação linear dos valores y_{t-n} passados correspondente ao modelo AR (*AutoRegressive*) e os erros ε_{t-n} passados correspondente ao modelo MA (*Moving Average*). Além disso, o modelo ARIMA permite a realização de diferenciação dos dados de entrada a fim de torná-los estacionários. O modelo final é estabelecido através dos parâmetros p, d e q, sendo p o número de atrasos da componente AR, d o número de diferenciações realizadas para que a série se torne estacionária e q o número de atrasos da componente MA. Os coeficientes do modelo são encontrados através do *Maximum Likelihood Estimator* (MLE).

A equação 2 a seguir, descreve o modelo final ARIMA:

$$y_t = c + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

onde os valores de ϕ são os coeficientes das componentes AR e os valores de θ são os coeficientes das componentes MA.

2.2.3 Regressão

Usualmente a regressão é utilizada em aplicações em que se deseja realizar a predição ou a explicação de um fenômeno fazendo-se valer dos dados conhecidos deste fenômeno. Em suma, a regressão busca encontrar uma equação linear da forma

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \quad (3)$$

tal que $x_1, ..., x_n$ são variáveis do conjunto de dados e $\beta_0, \beta_1, ..., \beta_n$ são os coeficientes da equação. Esta equação deve modelar o fenômeno de forma que os coeficientes encontrados permitam que as variáveis independentes x expliquem o maior percentual possível de variância na variável dependente y. O percentual da variância explicada por um modelo é indicado pelo valor R-Quadrado (R^2), normalmente calculado pelos métodos computacionais disponíveis atualmente.

Neste trabalho o ajuste de regressão utilizado foi o *Ordinary Least Squares* (OLS). Neste método os coeficientes são encontrados de maneira determinística já que o modelo deve encontrar a equação tal que a soma dos quadrados dos erros (desvio entre o valor real e o valor estimado) seja mínima. Através da derivada da soma dos quadrados dos erros encontra-se fórmulas fechadas para o cálculo dos valores dos coeficientes.

2.2.4 Regressão com *features* temporais

Assim como no modelo ARIMA, ao realizar uma regressão é possível incluir variáveis de *lag* através de *feature engineering*. Um modelo com m variáveis de atraso para a variável x_t é da seguinte forma:

$$y = \beta_0 + \beta_{10} x_t + \beta_{11} x_{t-1} + ... + \beta_{1m} x_{t-m} + ... + \beta_n x_n \quad (4)$$

Assim, $\beta_{11}, \beta_{1m}, ..., \beta_{1m}$ são os coeficientes associados às variáveis de atraso $x_{t-1}, x_{t-2}, ..., x_{t-m}$. Nota-se que as demais variáveis explicativas continuam presentes no modelo.

2.3 Análise Espacial

Um método amplamente utilizado tratando-se de análise espacial de dados é a verificação de autocorrelação espacial através do Índice de Moran (*Moran's I*). Este índice mede a autocorrelação espacial baseando-se na localização de uma variável bem como o valor desta variável, por exemplo, ocorrências atendidas pelos agentes de segurança em determinado bairro de uma cidade. Esta medida - que se encontra no intervalo $[-1, +1]$ - avalia se o padrão espacial é clusterizado (aglomerado), disperso ou aleatório, tal que quanto mais próximo de -1 os dados são dispersos espacialmente, mais próximos de +1 os dados estão clusterizados e mais próximos de 0 os dados estão distribuídos de maneira aleatória. Geralmente, o índice de Moran é acompanhado do gráfico de Moran (como exemplo o gráfico apresentado na Figura 15).

O I de Moran é uma estatística inferencial e portanto os resultados da análise devem sempre ser interpretados no contexto da hipótese nula. No caso do Índice de Moran Global, a hipótese nula é de que o processo espacial em que se formam os padrões estudados é completamente aleatório. É também por isso que este índice é sempre avaliado juntamente ao *p-value* simulado, que realiza múltiplas distribuições aleatórias e verifica-se o grau em que os padrões observados podem ter sido gerados completamente por acaso. Uma vez que o *p-value* é estatisticamente significante, rejeita-se a hipótese nula.

2.3.1 LISA

Os indicadores *Local Indicator of Spatial Association* (LISA) - como o nome sugere - é um indicador que permite a decomposição de indicadores globais (como o *Moran's I*) na contribuição de cada observação (ANSELIN, 2010). Em outras palavras, o indicador LISA permite verificar os padrões espaciais para os subconjuntos de localidades presentes em um *dataset* com variáveis geográficas fornecendo informações de *hotspots* e *coldspots*, bem como o que aqui considerou-se como *outlier*: regiões de baixa incidência em que se apresentou alta de determinado valor ou vice-versa. Um LISA pode ser um dos quatro seguintes: HH ou *High-High* - região de alta incidência de determinada *feature*; LL ou *Low-Low* - região de baixa incidência; HL ou *High-Low* - localidade de alta incidência pertencente a uma região de baixa incidência; e LH ou *Low-High* - localidade de baixa incidência pertencente a uma região de alta incidência.

2.4 Análise Espaço-Temporal

Aspectos geográficos são *features* - isto é, devem ser considerados no ajuste e treinamento de modelos - pois auxiliam diretamente na realização de predições de resultados de processos espaciais. Por exemplo, a localização de um imóvel muito possivelmente irá afetar

seu preço de venda ou aluguel; no caso de doenças infecciosas os lugares em que um indivíduo se desloca ao longo do dia apresentam impacto em sua saúde.

2.4.1 Regressão Espacial

Verificada a existência de autocorrelação espacial como apresentada na Seção 2.3. É possível trazer para os modelos de regressão informações e novas variáveis que levam em conta os aspectos do processo espacial que ocorre. É possível por exemplo ajustar um modelo com variáveis que levem em conta que a incidência de um valor em determinado região pode aumentar ou diminuir baseado em sua redondeza, estas variáveis são chamadas de *spatial lag variables*. Outra possibilidade é a utilização de variáveis categóricas - a maioria das bibliotecas atualmente dão suporte à funcionalidade - em que o modelo realiza o ajuste e as predições baseado nesta variável. Basicamente, um ajuste linear como o da Equação 3 é feito para cada um dos diferentes valores únicos existentes para a variável categórica. Para cada um destes valores únicos existe também um coeficiente que indicará a contribuição desta variável para a explicação da variância da variável dependente.

2.4.2 Regressão Espaço-Temporal

É possível incluir em modelos de regressão tanto os aspectos temporais como os espaciais. Por exemplo, o modelo da Equação 4 pode incluir variáveis categóricas relativas à localidade de um registro em uma base de dados e gerar predições sobre um fenômeno valendo-se de ambos aspectos temporal e espacial.

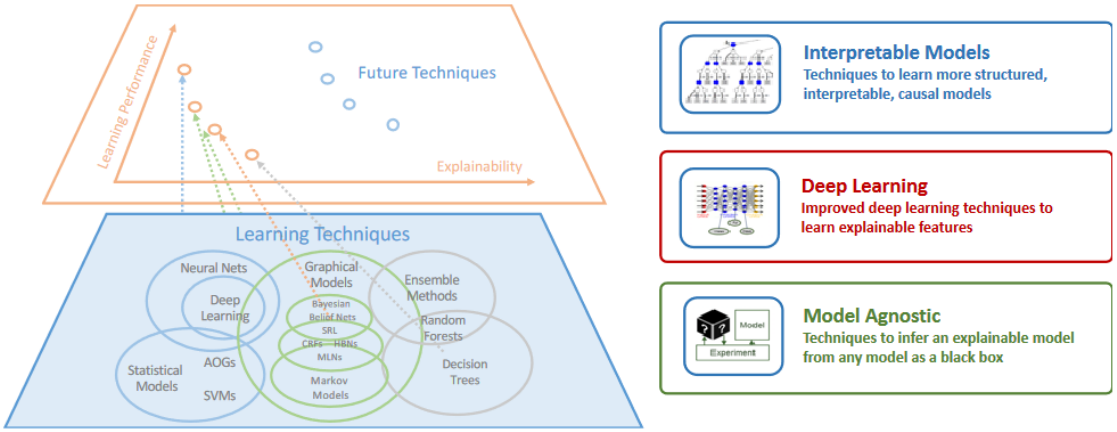
2.5 Inteligência Artificial Explicável

Inteligência Artificial Explicável, em inglês *Explainable Artificial Intelligence* (XAI), é uma Inteligência Artificial (IA) tal que seus resultados podem ser compreendidos por humanos. Muitos dos modelos de IA da atualidade tem funcionamento de caixa preta já que os processos e parametrizações que ocorrem dentro dos algoritmos são fechados ou suficientemente complexos para que uma pessoa não consiga ou apresente grande dificuldade para explicar estes processos. (GUNNING *et al.*, 2019) sugerem que em diversas aplicações críticas a explicabilidade da IA é essencial para os utilizadores entenderem, confiarem e efetivamente manejarem estas ferramentas. Também defendem que um sistema XAI deve ser capaz de explicar suas capacidades e compreensões, bem como informar o que foi, está sendo e será feito, bem como divulgar as informações em que está atuando.

A explicabilidade dos algoritmos utilizados nas aplicações de IA da atualidade têm se demonstrado menor à medida em que o desempenho destes algoritmos melhora. Por exemplo,

as Redes Neurais se demonstram muito eficazes em problemas de classificação, entretanto é muito complexa a tarefa de entender e explicar o porquê os coeficientes e valores de *threshold* dos neurônios determinados pelo treinamento da rede resultam nas classificações fornecidas pelo algoritmo. Este tema tem gerado discussões éticas a respeito do uso deste tipo de técnica e sua confiabilidade. A Figura 2 demonstra o cenário atual da IA onde diferentes modelos apresentam diferentes níveis de explicabilidade em relação ao desempenho destes, sendo que os modelos com maior desempenho, como o caso das Redes Neurais mencionado anteriormente, apresentam baixa explicabilidade devido as características intrínsecas destas. Neste sentido, a XAI é um campo que objetiva o aumento do desempenho bem como da explicabilidade e consequentemente da confiança na IA. Neste trabalho, ainda que os algoritmos e modelos utilizados supostamente apresentem menor dificuldade no fator explicabilidade por se tratarem de modelos estatísticos e de regressão, a XAI continua sendo um campo de relevância, contribuindo para o que se espera dela para o futuro: o desenvolvimento e surgimento de modelos de maior desempenho e maior explicabilidade.

Figura 2 – Diagrama desempenho x explicabilidade dos modelos de IA da atualidade



Fonte: Gunning *et al.* (2019).

2.6 *Outlying Aspect Mining* (OAM)

Relativamente à explicabilidade das técnicas de IA existentes no presente, este trabalho busca também a explicação das razões pelas quais as técnicas utilizadas na detecção de *outliers* espaço-temporais classificaram os registros como sendo *outliers* (dado que grande parte das técnicas atuais possibilitam a descoberta destes registros mas não fornecem explicação sobre eles - é o caso da LOF). Uma das mais recentes áreas que tem colaborado com o desenvolvimento do tema é a área denominada *Outlying Aspect Mining* (OAM).

Samariya, Ma e Aryal (2020) definem a OAM como a tarefa de identificar o subconjunto de variáveis onde um ponto de dados é inconsistente com os demais pontos do conjunto. O ponto de dados a ser analisado é denominado *query* e as variáveis do subconjunto encontrado são denominadas *outlying aspects* da *query*. Existem três diferentes técnicas para a exploração

dos *outlying aspects* de um conjunto de dados. A técnica utilizada neste trabalho, implementada por (SILVA; GOMES-JR, 2022), é baseada em *score-and-search*. Nesta abordagem, uma função de pontuação é passada para o algoritmo OAM que a utilizará para atribuir um *score* a cada subespaço visitado, ou seja, cada subconjunto gerado pelo algoritmo de busca utilizado pelo processo.

Na implementação aqui utilizada, o algoritmo de *score* empregado trata-se do *iPath* (*isolation Path*), trabalho realizado por Vinh *et al.* (2016). O algoritmo baseia-se na realização de cortes no espaço isolando o objeto analisado dos demais, de forma que pontos isolados através da realização de menos cortes são considerados *outliers* em detrimento dos pontos em que necessitou-se a realização de mais cortes no espaço (já que aqueles cercados por uma densidade maior de objetos demandam mais divisões para seu isolamento). O algoritmo *iPath* é uma técnica que estabelece uma métrica dimensionalmente imparcial, já que o OAM compara os *scores* da *query* em diferentes subespaços.

2.7 Trabalhos Relacionados

Os principais conceitos importantes para a realização e compreensão deste trabalho foram já outras vezes abordados por diferentes pesquisadores. Análises espaço-temporais de crimes foram conduzidas por (MASULLO *et al.*, 2017), que realizaram análises de crimes violentos na cidade de São Luís do Maranhão utilizando métodos aqui evidenciados como o Índice Global de Moran e LISA. A partir destas técnicas, eles puderam caracterizar e desenvolver relatórios sobre: mortes violentas em decorrência de intervenção da polícia e latrocínio; violência contra jovens e mulheres; violência e contexto social.

Outros trabalhos também tiveram como foco de desenvolvimento a implementação de ferramentas. Por exemplo a CrimeVis, uma aplicação nos moldes da solução aqui proposta utilizando dados do Instituto de Segurança Pública do Rio de Janeiro de 138 distritos de polícia obtidos ao longo de 12 anos. Na elaboração da CrimeVis (SILVA *et al.*, 2017) utilizaram analises estatísticas e de clusterização para a visualização dos dados semelhantes com os aqui aplicados. Com isto os pesquisadores puderam obter informações para responder perguntas como por exemplo: como as taxas criminais evoluíram ao longo do tempo?; quais são as subdivisões de áreas do estado de acordo com critérios criminais e socioeconômicos?; quais foram os efeitos da implantação das Unidades de Polícia Pacificadora? Outra ferramenta desenvolvida para análise espaço-temporal de crimes, denominada CrimAnalyzer, trabalho de (GARCIA *et al.*, 2019), conta com 7 diferentes tipos de visualizações: mapa; *hotspots*; acumulado; temporal; temporal global; *ranking*; e radial. Eles utilizaram dados de sete anos concedidos pela Polícia do Estado de São Paulo. Para a identificação de *hotspots* eles utilizaram uma abordagem baseada na técnica *Non-Negative Matrix Factorization* (NMF). Também conduziram uma pesquisa com especialistas da área que produziram um relato com suas considerações sobre a aplicação.

Pesquisadores através dos dados da SESP aqui também utilizados produziram análises do impacto da pandemia de COVID-19 nos padrões espaço-temporais de crimes na cidade de Curitiba que demonstraram resultados como a diminuição nas ocorrências de furto e roubo registrados e o aumento de casos de violência doméstica no período da pandemia. (LEAL; GOMES-JR, 2022) também utilizaram métodos para a detecção de *outliers* temporais e espaciais com abordagens estatísticas e com indicadores locais de associação espacial. Leal (2022) ainda desenvolveu o protótipo CityGuardian, que além de possibilitar a análise exploratória dos dados e realizar a decomposição de séries temporais, realiza a emissão de alertas de mudanças nos padrões criminais através de contínuo monitoramento. Neste trabalho, além da visualização e análise exploratória dos dados, objetivos atingidos pelo CityGuardian, o maior empenho realizado foi no sentido de facilitar a visualização e detecção de *outliers* diretamente na ferramenta, isto é, além dos resultados das análises desenvolvidas através de implementações da *scripts*.

3 METODOLOGIA E ANÁLISE EXPLORATÓRIA

O trabalho foi dividido em seis fases:

1. **Análise Exploratória dos dados:** Nesta fase ocorreram os primeiros contatos com os dados e foram geradas percepções iniciais sobre os padrões das ocorrências: bairros com maior número de ocorrências per capita; categorias mais frequentes; aumento/diminuição ao longo do tempo;

2. **Aplicação e análise dos modelos de detecção de *outliers* temporais:** Foram realizados estudos e aplicação de diferentes técnicas e modelos para detecção de *outliers* em séries temporais. Para este fim utilizou-se decomposições de séries para descoberta de tendências e sazonalidade.

3. **Aplicação e análise dos modelos de detecção de *outliers* espaciais:** Foram realizados estudos de diferentes abordagens para a detecção de *outliers* espaciais e aplicação das diferentes técnicas para este fim. Estas análises se inciaram através da compreensão e revisão dos conceitos de autocorrelação espacial de Moran e os indicadores LISA (REY; ARRIBAS-BEL; WOLF, 2020);

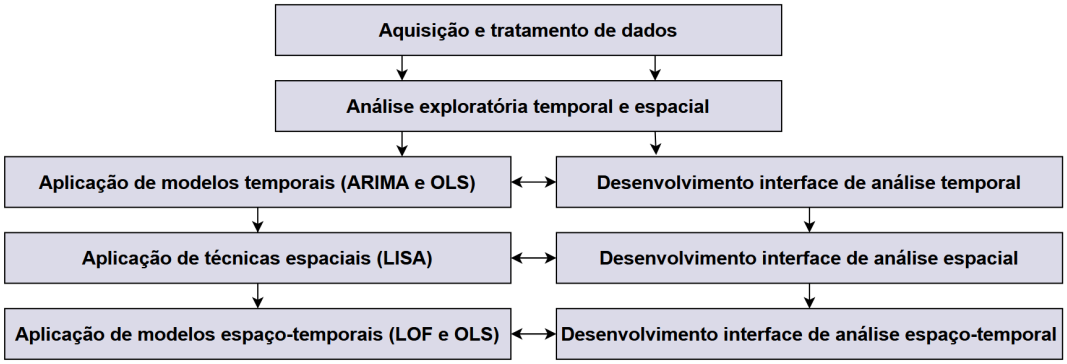
4. **Aplicação e análise dos modelos de detecção de *outliers* espaço-temporais:** Foram testados diferentes conjuntos de variáveis espaço-temporais e ajustes do modelo de regressão OLS com a finalidade de detectar *outliers* espaço-temporais;

5. **Desenvolvimento da interface (*front-end*):** Fase em que ocorreu propriamente a implementação da ferramenta com a criação de *layouts* embasados em boas práticas de interação e visualização;

6. **Análise de explicabilidade dos *outliers* temporais e espaciais:** Foi aplicada a técnica OAM para a mineração dos *outlying aspects* dos *outliers* espaciais e temporais detectados.

A Figura 3 demonstra as etapas de análises e desenvolvimento da ferramenta, de maneira a evidenciar que o processo de análise e implementação da ferramenta para cada tipo de análise ocorreu de forma comutável.

Figura 3 – Etapas de desenvolvimento do trabalho



Fonte: Autoria própria (2022).

3.1 Aquisição dos dados

Duas bases de dados são utilizadas para a realização deste trabalho. A primeira fonte de dados contém ocorrências atendidas pela GMC - chamada SiGesGuarda - e a segunda contém registros de ocorrências atendidas pela SESP. Os dados da GMC são públicos e obtidos diretamente através do portal de dados abertos[1] da Prefeitura Municipal de Curitiba (PMC). Os dados da Secretaria foram obtidos mediante assinatura de termo de confidencialidade entregue ao órgão A confidencialidade dos dados se faz necessária devido à existência de dados considerados sensíveis pela SESP dado que os registros de ocorrências são georreferenciados.

Foram também utilizados dados demográficos e socioeconômicos do Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC). Estes dados contêm informações relativas ao número de habitantes, latitude e longitude médias e renda per capita por bairro. Ainda, foram obtidos dados no formato *shp* que foram posteriormente convertidos em *geojson* contendo as divisões geográficas dos bairros da cidade.

3.2 Análise Exploratória

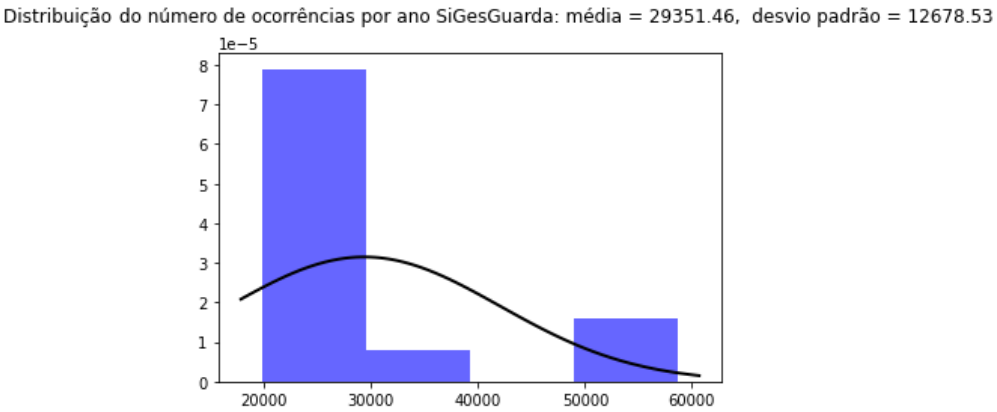
A base SiGesGuarda contém 389.964 registros de ocorrências atendidos pela GMC a partir do ano de 2009. Cada registro é composto por 35 *features* sendo estes, por exemplo: data, hora, bairro, logradouro, flagrante e natureza da ocorrência. Em média foram registradas anualmente 29.351,46 ocorrências com desvio padrão de 12.678,53. A série histórica de ocorrências é mostrada na Figura 4 e a distribuição de ocorrências na Figura 5.

[1] Disponível em: https://www.curitiba.pr.gov.br/dadosabertos

Figura 4 – Série temporal de registros de ocorrências no período de 2009 a 2021

Fonte: Autoria própria (2022).

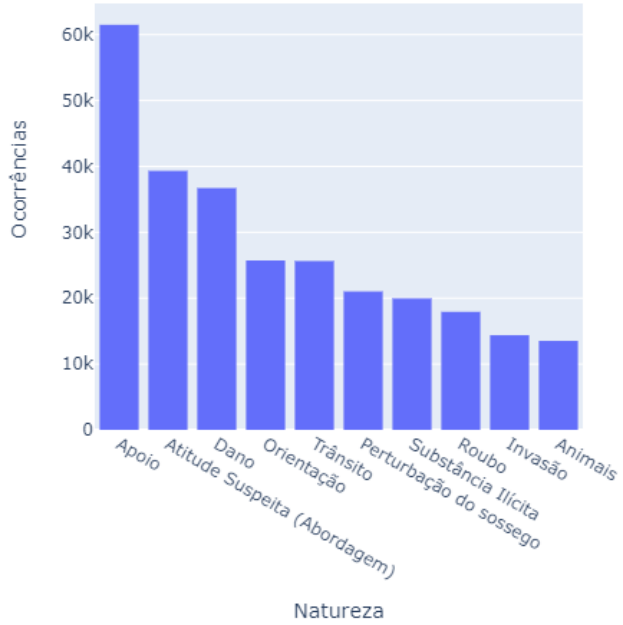
Figura 5 – Histograma de ocorrências por ano



Fonte: Autoria própria (2022).

Foram registradas 185 diferentes naturezas de ocorrência. As naturezas mais atendidas no período de 2009 a 2021 podem ser visualizadas na Figura 6.

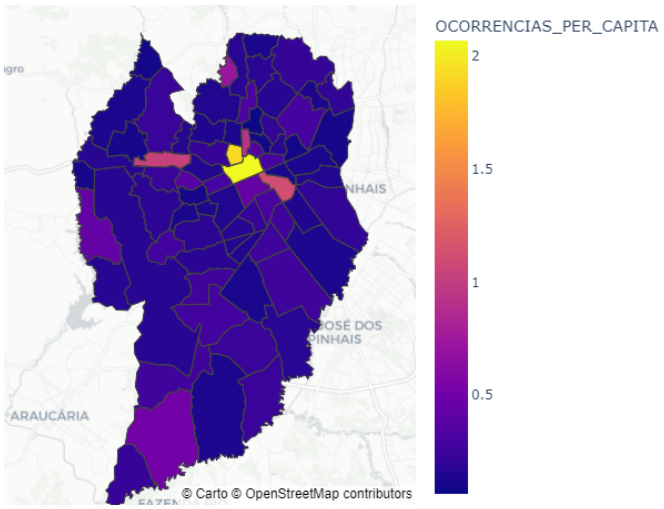
Figura 6 – 10 naturezas de ocorrências mais atendidas - SiGesGuarda (2009 - 2021)



Fonte: Autoria própria (2022).

Através do *geojson* contendo as divisas e dados demográficos, o número de ocorrências foi normalizado por população total dos bairros. Inicialmente pensava-se que as regiões periféricas bem como a região central apresentaria maiores taxas de ocorrências per capita, entretanto esta hipótese se confirmou parcialmente já que as taxas mais altas se concentraram nos bairros da região central da cidade, entretanto não nas regiões periféricas como é mostrado na Figura 7.

Figura 7 – Ocorrências per capita por bairro - SiGesGuarda (2009 - 2021)



Fonte: Autoria própria (2022).

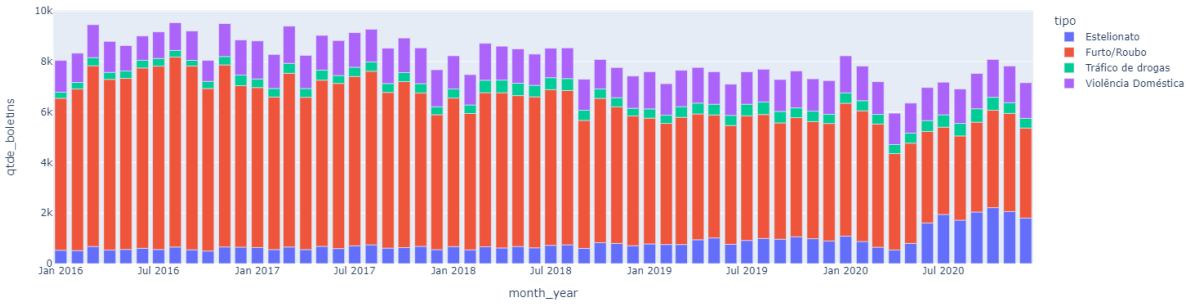
Os dados de ocorrências atendidas pela SESP diferem dos dados da base *SiGesGuarda* (GMC) em estrutura bem como em alguns aspectos de seu conteúdo. A primeira mudança trata-se de que os dados da SESP obtidos não são concentrados num único arquivo *csv* com todos

os registros. São diferentes arquivos de ocorrências enquandradas nas quatro seguintes naturezas: Estelionato, Furto/Roubo, Tráfico de Drogas e Violência Doméstica. Não foram fornecidos dados das demais naturezas. Os dados correspondem ao intervalo de Janeiro de 2016 a Dezembro de 2020.

Outra diferença importante é que os dados de furto e roubo da SESP para o ano de 2020 são georreferenciados e não contém a hora do fato, apenas a data. Para a realização das análises temporais dos dados da GMC foi utilizada uma API para fornecimento dos dados de latitude e longitude através do logradouro da ocorrência. Como a numeração do logradouro do fato não está contido na base SiGesGuarda, a localização retornada pela API é aproximada e não exata como ocorre com os dados da SESP.

Foi realizado um agrupamento dos dados das quatro naturezas mencionadas em somente um arquivo e realizou-se a análise com este. A Figura 8 representa a quantidade de ocorrências atendidas a cada mês separadas por natureza do crime. Evidencia-se uma diminuição no número de casos de furto e roubo ao longo dos quatro anos em questão e uma diminuição no número total de ocorrências de todas as espécies a partir dos primeiros meses de 2020, data coincidente com o início da pandemia de COVID-19, o que poderia explicar esta diminuição por conta das medidas de isolamento adotadas no período. Também verificou-se aumento no número de ocorrências de estelionato no ano de 2020.

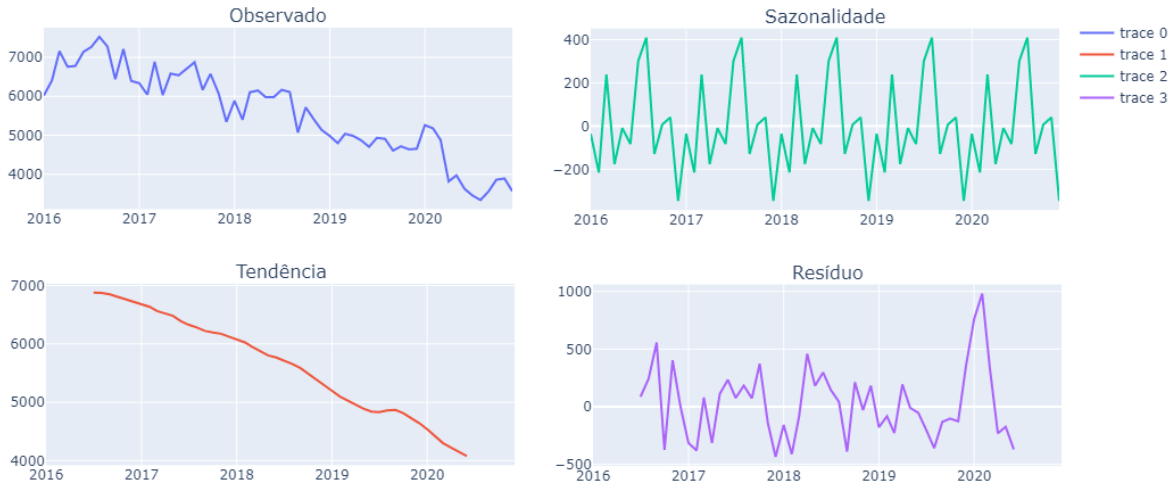
Figura 8 – Total de ocorrências mês a mês por natureza - SESP-PR (2016 - 2020)



Fonte: Autoria própria (2022).

Foi então realizada uma decomposição sazonal para verificação da tendência de diminuição no número de ocorrências de furto e roubo no período.

Figura 9 – Tendência e sazonalidade de ocorrências de furto e roubo - SESP-PR (2016 - 2020)



Fonte: Autoria própria (2022).

3.3 Aplicação de modelos

Foram utilizadas as quatro primeiras semanas após a análise exploratória para a aplicação e treinamento dos modelos e o desenvolvimento das funcionalidades relativas à Análise Temporal dos dados. Os modelos ARIMA e OLS abordados nas Seções 2.2.2 e 2.2.3 foram utilizados para a detecção de *outliers* temporais. Foram agregados os registros por dia para o período contido na base SiGesGuarda - isto é, cada ponto da série temporal corresponde ao número de ocorrências atendidas para um dia. Então os modelos são ajustados sobre a série temporal agregada por dia e havendo significância estatística destes modelos, calcula-se o erro gerado pela diferença entre o número observado de ocorrências atendidas e o número de ocorrências esperado pelo modelo. Os erros são pontuados através do *z-score* e classifica-se como *outlier* os 1% dos dados mais espaçados da média - módulo do *z-score* > 2,575.

Também duas fontes adicionais de dados foram necessárias para a análise temporal, sendo elas informações demográficas e de geometria dos bairros da cidade. Através dos dados demográficos foi realizada normalização das ocorrências registradas nos bairros dividindo-se o número de ocorrências pelo tamanho da população de moradores dos bairros. Entende-se que embora não necessariamente esta normalização consiga representar uma medida completamente segura de comparação entre o número de ocorrência entre os bairros - pois podem haver dinâmicas de concentração de ocorrências na cidade não produzidas por moradores de determinada localidade - ainda assim é necessária a utilização de algum método que permita a comparação do número de ocorrências entre regiões.

4 ANÁLISE DE DADOS E MODELAGEM

4.1 Análise Temporal

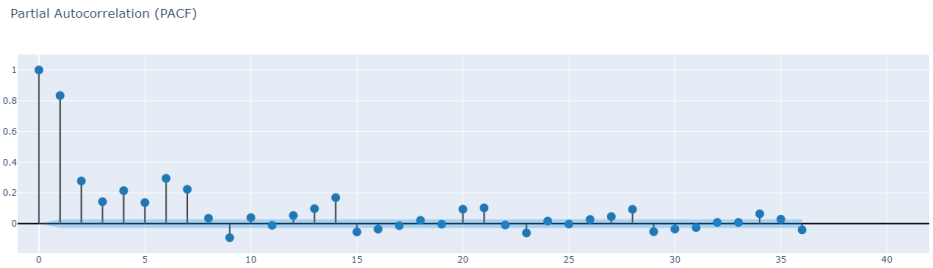
Aqui pretende-se utilizar técnicas de regressão para encontrar modelos que melhor expliquem as variações no número de ocorrências criminais ao longo do tempo. Ao final do ajuste dos modelos, o erro entre o número de ocorrências predito e o registrado é distribuído de forma normal e os registros com *z-score* mais distantes da média são tratados como *outliers*.

4.1.1 ARIMA

Inicialmente foram avaliados diferentes parametrizações do modelo ARIMA com diferentes valores de *p*, *d* e *q*. A série de registros de ocorrências da GMC através do teste *Augmented Dickey–Fuller* (ADF) obteve *score* de -2,941 com *p-value* de 0,041. Embora este *p-value* esteja dentro da significância de 5%, foi elaborada uma análise de estacionaridade com o teste ADF para a mesma série diferenciada uma vez, isto é, $d = 1$. Para este caso, o *score* ADF foi de -16,296 com *p-value* de 3,302e-29, o que permite assumir com mais significância que a série diferenciada é estacionária e portanto aplicável ao modelo ARIMA.

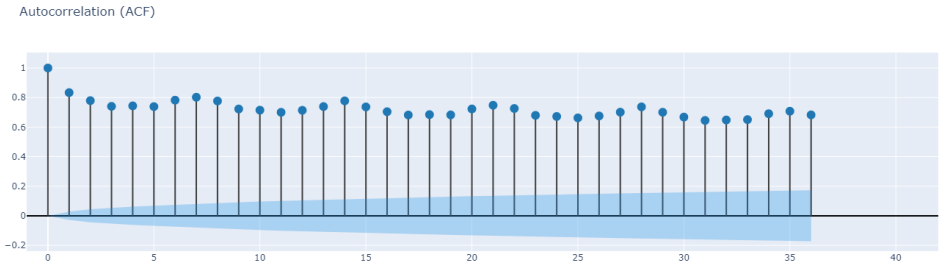
Para o auxílio na escolha dos parâmetros *p* e *q* foram plotados gráficos das *Partial Autocorrelation Function* (PACF) e *Autocorrelation Function* (ACF) respectivamente. Considerou-se para a escolha dos parâmetros somente atrasos com autocorrelação maior que 0,8 o que indicou $p = 2$ e $q = 2$. O gráfico PACF para os dados da GMC é apresentado na Figura 10 e o ACF na Figura 11.

Figura 10 – Gráfico de Autocorrelação Parcial da série temporal da GMC



Fonte: Autoria própria (2022).

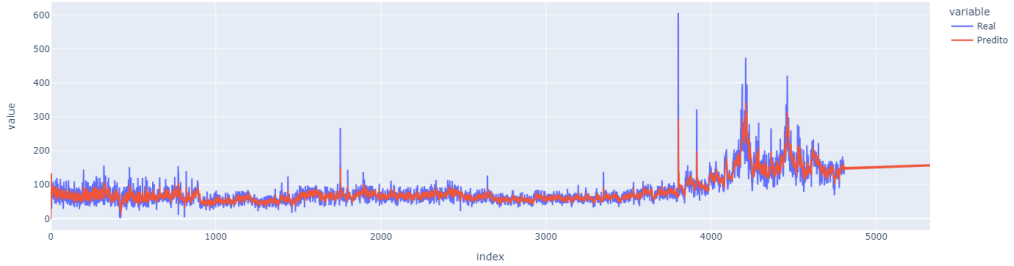
Figura 11 – Gráfico de Autocorrelação da série temporal da GMC



Fonte: Autoria própria (2022).

Além disso, foram gerados modelos ARIMA com diferentes parametrizações para comparação com o modelo (p=2, d=1, q=2). O modelo (2,2,7) apresentou melhor valor de *Log Likelihood*. Este modelo ARIMA foi comparado a aplicações do modelo OLS mais tarde apresentados. A Figura 12 apresenta um gráfico de linhas contendo os valores da série temporal com os valores previstos pelo modelo.

Figura 12 – Comparativo entre valores previstos pelo modelo ARIMA e a série temporal da GMC



Fonte: Autoria própria (2022).

4.1.2 OLS

Foram então criados e testados diferentes ajustes e conjuntos de variáveis sobre o OLS a fim de se obter o modelo com o maior valor de R^2, indicando quais variáveis explanatórias melhor descrevem o fenômeno de ocorrências atendidas pela GMC.

A primeira hipótese considerada para a criação de variáveis foi a de que uma determinada data ser um dia de final de semana, este pode ser um fator para o aumento do número de ocorrências. Portanto, através do campo de *data* da base SiGesGuarda, foi criada uma nova coluna do dataset chamada de *final_de_semana*, uma variável binária que recebe o valor 1 se a data corresponde a um dia de final de semana e 0 caso contrário. O modelo gerado através da inclusão desta variável apresentou R^2 de 0,818 e o coeficiente da variável *final_de_semana* foi de 23,5776 com *p-value* de 0.

Na mesma linha de pensamento utilizada na criação da variável *final_de_semana*, foi gerada uma variável *feriado* - através da busca de uma tabela de feriados na cidade de Curitiba

desde o ano de 2016 - com a hipótese também de que em dias de feriado há aumento no número de ocorrências atendidas. Neste modelo foram então inclusas as duas variáveis - *final_de_semana* e *feriado* - obtendo-se R^2 igual a 0,818 e os coeficientes das variáveis respectivamente sendo 23,6053 e 6,2411. A significância estatística da segunda foi de 4,9%.

De igual modo, foi criada uma variável *pandemia* que recebeu o valor 1 para todo o intervalo de dias a partir do dia 11 de março de 2020, data do primeiro caso da COVID-19 na cidade. O modelo que inclui as variáveis *final_de_semana*, *feriado* e *pandemia* obteve R^2 de 0,847; com os coeficientes das variáveis sendo respectivamente 27,4507; 9,6469 e 51,9131 com *p-value* de 0 para todos os coeficientes. Os valores dos coeficientes indicam que de fato estes fatores levaram a um aumento no número de ocorrências atendidas, em especial o coeficiente de 51,9131 relativo à variável *pandemia*.

Semelhantemente ao modelo ARIMA, que inclui erros e valores do passado em sua modelagem, é possível incluir váriaveis de *lag* em modelos de regressão. Como mencionado, os dados de registros são agregados em séries temporais de ocorrências por dia, portanto as variáveis de *lag* geradas correspondem aos registros de ocorrências referentes à semana anterior a data em que se pretende estimar, isto é, os sete dias anteriores. A Tabela 1 exemplifica uma série de ocorrências de uma semana juntamente com os valores das variáveis de *lag*.

Tabela 1 – Exemplo de variáveis de *lag* dos 7 últimos dias

Ocorrências	Lag1	Lag2	Lag3	Lag4	Lag5	Lag6	Lag7
81	30.0	0.0	0.0	0.0	0.0	0.0	0.0
96	81.0	30.0	0.0	0.0	0.0	0.0	0.0
93	96.0	81.0	30.0	0.0	0.0	0.0	0.0
59	93.0	96.0	81.0	30.0	0.0	0.0	0.0
30	59.0	93.0	96.0	81.0	30.0	0.0	0.0
70	30.0	59.0	93.0	96.0	81.0	30.0	0.0
61	70.0	30.0	59.0	93.0	96.0	81.0	30.0

Fonte: Autoria própria (2022).

O modelo com as variáveis de *lag* foi ajustado e continha também as variáveis anteriormente utilizadas. Este modelo apresentou aumento no valor de R^2 passando para 0,951. Neste modelo entretanto três variáveis não obtiveram significância estatística sendo elas os *lags* de 3 e 5 dias bem como a variável *feriado*, que embora apresentou coeficiente positivo, teve *p-value* de 0,462. Então, foram removidas as variáveis não significantes e o modelo final apresentou os resultados contidos na Tabela 2 a seguir. Nesta Tabela, verifica-se o valor do R^2 obtido pelo modelo onde encontra-se a escrita *R-squared (uncentered)* e os valores dos coeficientes encontrados para as variáveis de entrada, notadamente a contribuição das variáveis final de semana e pandemia na explicação do número de casos que receberam respectivamente os valores 8,7038 e 4,9368.

Tabela 2 – Resultados do ajuste final do modelo OLS

Dep. Variable:	OCORRENCIAS_ATENDIDAS	**R-squared (uncentered):**	0.951
Model:	OLS	**Adj. R-squared (uncentered):**	0.951
Method:	Least Squares	**F-statistic:**	1.154e+04
Date:	Mon, 17 Oct 2022	**Prob (F-statistic):**	0.00
Time:	08:07:54	**Log-Likelihood:**	-21342.
No. Observations:	4807	**AIC:**	4.270e+04
Df Residuals:	4799	**BIC:**	4.275e+04
Df Model:	8		
Covariance Type:	nonrobust		

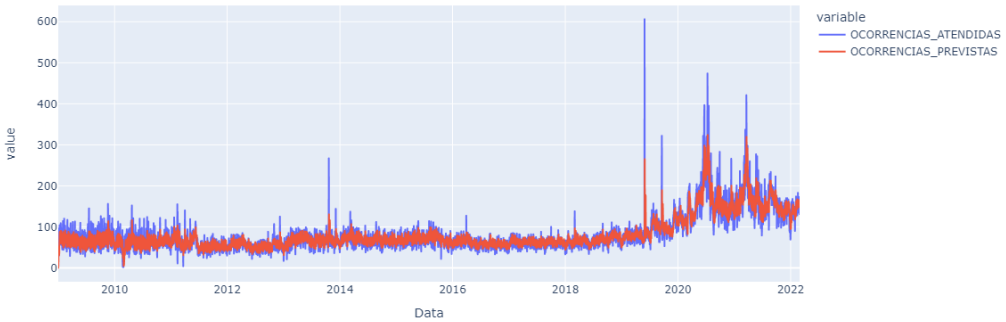
	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
x0_final_de_semana	8.7038	0.688	12.652	0.000	7.355	10.053
x1_time	0.0012	0.000	4.610	0.000	0.001	0.002
x2_lag1	0.3650	0.014	26.298	0.000	0.338	0.392
x3_lag2	0.1010	0.014	7.351	0.000	0.074	0.128
x5_lag4	0.0936	0.013	7.446	0.000	0.069	0.118
x7_lag6	0.1937	0.014	14.021	0.000	0.167	0.221
x8_lag7	0.1661	0.014	11.659	0.000	0.138	0.194
x11_pandemia	4.9368	1.089	4.535	0.000	2.803	7.071

Omnibus:	5044.507	**Durbin-Watson:**	1.949
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	2214031.405
Skew:	4.519	**Prob(JB):**	0.00
Kurtosis:	107.749	**Cond. No.**	1.04e+04

Fonte: Autoria própria (2022).

A Figura 13 mostra a comparação da série temporal de ocorrências atendidas e da série temporal das ocorrências previstas pelo modelo. Ao ampliar o gráfico é possível notar que o modelo percebe a variação de ocorrências para o último dia e então estima esta variação para o presente.

Figura 13 – Gráfico de comparação entre séries de atendidas e previstas pelo modelo final OLS



Fonte: Autoria própria (2022).

Com o modelo de regressão ajustado passou-se então para a tarefa de detecção de *outliers*. Foi criado um *dataframe* com as estimativas geradas pelo modelo e as séries reais de ocorrências e com estes dados foi então gerada uma coluna denominada ERRO resultado da

diferença entre aqueles valores. Calculou-se o *z-score* dos valores contidos na coluna ERRO que foram inseridos em uma coluna denominada Z-SCORE que finalmente, foi utilizada para a criação de uma coluna chamada OUTLIER de modo que os registros com $|z\text{-}score| >= 2{,}17$ receberam o valor 1 enquanto que aos demais foi atribuído o valor 0. A Tabela 3 exibe uma amostra de 5 registros após o tratamento para detecção de *outliers*.

Tabela 3 – Amostra de 5 registros com verificação de *outliers*

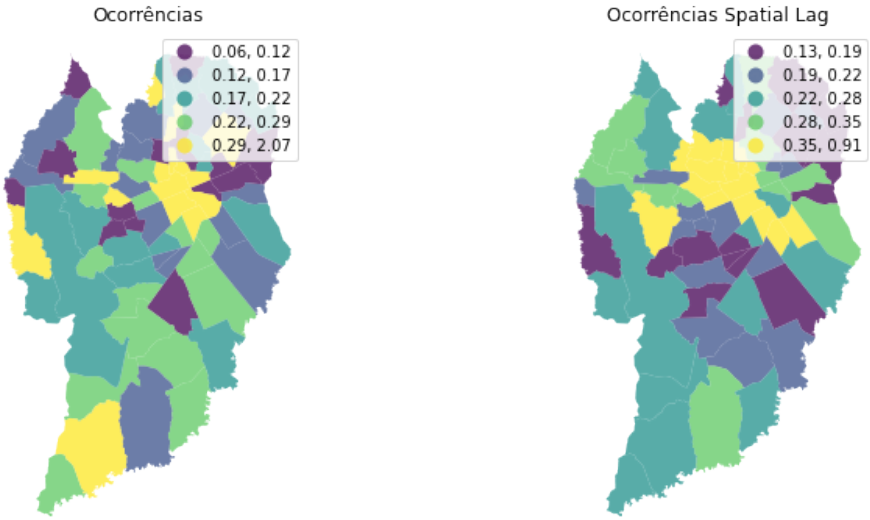
Atendidas	Previstas	Data	ERRO	Z-SCORE	OUTLIER
53.0	69	2009-09-10	16.0	0.827144	0
68.0	57	2009-12-16	-11.0	-0.489326	0
61.0	50	2018-02-21	-11.0	-0.489326	0
217.0	272	2020-06-28	55.0	2.728714	1
168.0	152	2021-04-30	-16.0	-0.733117	0

Fonte: Autoria própria (2022).

4.2 Análise Espacial

Com os dados normalizados, através da matriz de pesos gerada para os 6 vizinhos mais próximos, calculou-se o *spatial lag* para cada bairro através da função de mesmo nome presente na biblioteca *pysal*[1]. A Figura 14 exibe os valores de ocorrências normalizados e *spatial_lag* para os bairros da cidade.

Figura 14 – Mapas de comparação do *spatial lag* de ocorrências normalizadas dos registros da GMC



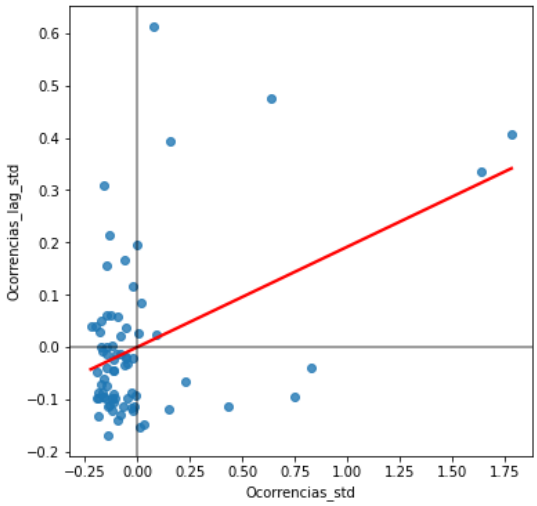
Fonte: Autoria própria (2022).

Os mapas parecem indicar a presença de autocorrelação espacial, entretanto a visualização dos mapas não é suficiente para afirmar esta presença. Portanto, foi utilizado o gráfico de

[1] https://pysal.org/

Moran, presente na Figura 15, para verificar estatísticamente este indicador. A rampa ajustada sobre as variáveis de erro de ocorrências e *spatial_lag* indica a existência de autocorrelação. O valor I de Moran - coeficiente angular da rampa ajustada - é de 0,1916 e o *p-value* simulado, isto é, o *p-value* gerado por simulação de aleatoriedade espacial, é de 0,001, portanto a autocorrelação apresentado tem significância estatística.

Figura 15 – Gráfico de Moran indicando a existência de autocorrelação espacial nos dados da GMC

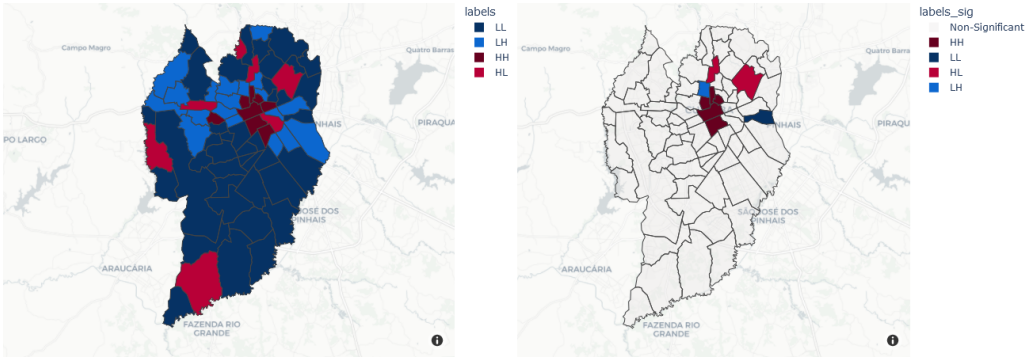


Fonte: Autoria própria (2022).

4.2.1 LISA

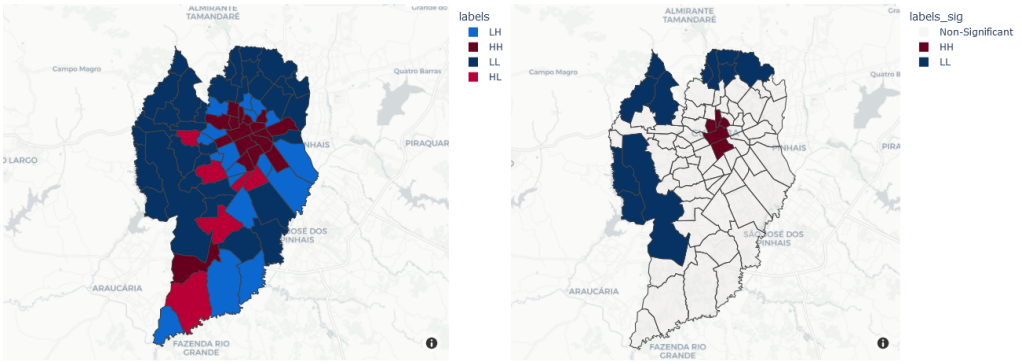
Verificada a existência de autocorrelação espacial, passou-se então para a análise de autcorrelação local entre os bairros. Foi usado portanto o método LISA através da função *Moran_Local* presente no módulo *esda* da biblioteca *pysal*. Como abordado na Seção 2.3.1, são quatro os indicadores em que um bairro pode ser categorizada, sendo: HH um indicador onde o número de ocorrências é alto com sua região de entorno também possuindo índice de ocorrências alto; HL um bairro em que a taxa de ocorrências é alta e a de seu entorno é baixa; LH um bairro com incidência de ocorrências baixa num entorno com alta incidência; e LL um bairro de baixa incidência numa região também de baixa incidência. Os HH são considerados *hotspots* e os LL são considerados *coldspots* enquanto que os demais - HL e LH - são considerados *outliers*. As Figuras 16 e 17 exibem os resultados do método *Moran_Local* nos dados da GMC e SESP respectivamente. Nota-se na Figura à direita a presença da categoria *Non-Significant* que indica os bairros onde não houve *p-value* simulado significativo.

Figura 16 – Técnica LISA aplicada aos dados da GMC para todo o período (2009 - 2021)



Fonte: Autoria própria (2022).

Figura 17 – Técnica LISA aplicada aos dados da SESP para todo o período (2016 - 2020)



Fonte: Autoria própria (2022).

4.3 Análise Espaço-Temporal

A tarefa de incluir as variáveis de espaço e tempo num mesmo modelo possui determinado grau de dificuldade. Em especial, encontrar *outliers* que se manifestam devido as suas características espaço-temporais. Duas modelagens distintas foram aplicadas e estão descritas nas Seções 4.3.1 e 4.3.2 a seguir.

4.3.1 LOF

Nesta abordagem a variável correspondente a data da ocorrência foi separada em três variáveis distintas indicando dia, mês e ano do registro. Também, através da coluna referente ao bairro da ocorrência, foram geradas colunas *dummy* para representar os bairros na forma *hot-encondig* - isto é, para cada bairro foi gerada nova coluna que recebe o valor 1 sendo o registro pertencente ao bairro da coluna, caso contrário, recebe 0. Uma das colunas da forma *hot-enconding* foi retirada para não cair em um problema de correlação entre as variáveis, já que se todas as colunas *dummy* de bairros receberam o valor 0, isto já significa que o registro é de um outro bairro - neste caso correspondente ao bairro da coluna removida.

Após preparados os dados para aplicação do modelo LOF, seu ajuste foi então realizado e cada registro recebeu um valor correspondente ao seu fator de *outlierness*. Foi realizada a distribuição dos *scores* de *outlierness* e baseado nesta, todo registro com *score* menor que -1,1 foi então categorizado como *outlier*.

4.3.2 Regressão Espacial

Para a regressão espacial o mesmo modelo da análise temporal é utilizado (OLS). Entretanto, existem duas fundamentais diferenças entre o modelo de regressão para análise espaço-temporal e o modelo temporal, que são: agregação do número de ocorrências por mês - o modelo temporal agrega o número de ocorrência por dia - e a inclusão de variável categórica referente ao bairro do registro. Assim como na modelagem temporal, variáveis de *lag* são utilizadas para acrescentar ao modelo o histórico de ocorrências em meses passados.

Uma amostra do *dataframe* utilizado para o ajuste do modelo é mostrado na Tabela 4.

Tabela 4 – Amostra do *dataframe* utilizado para o modelo OLS de regressão espacial para os dados da SESP

mes	ano	bairro	qtde_boletins	lag1	lag2	lag3	lag4
4	2017	CENTRO	1059	1224	1190	1318	1190
5	2017	CENTRO	1135	1059	1224	1190	1318
6	2017	CENTRO	1252	1135	1059	1224	1190
7	2017	CENTRO	1347	1252	1135	1059	1224
8	2017	CENTRO	1240	1347	1252	1135	1059

Fonte: Autoria própria (2022).

Os modelos finais de regressão espacial para os dados da GMC e da SESP apresentaram valores de R^2 de 0,894 e 0,972, respectivamente. Em ambos os modelos os coeficientes das variáveis obtiveram significância estatística com *p-value* de 0. Ainda, em ambos os resultados nota-se como os coeficientes das variáveis categóricas relativas a bairros onde observou-se números de ocorrências elevados também obtiveram valores mais altos que de coeficientes de bairros de menor incidência de ocorrências, portanto sendo isto mais uma evidência da predominância de ocorrências nestas regiões. Observa-se também que o coeficiente relativo ao ano da ocorrência foi de -3,9954 nos dados da Secretaria enquanto que para os dados da Guarda Municipal este valor foi de 0,2621 reforçando as informações levantadas na análise exploratória que indicavam aumento de registros atendidos pela GMC - coeficiente positivo - e diminuição de registros atendidos pela SESP - coeficiente negativo. As Tabelas 5 e 6 exibem os resultados obtidos em alguns dos coeficientes para a variável bairro.

Tabela 5 – Parte dos coeficientes de variáveis categóricas de bairro para o modelo de regressão espaço-temporal ajustado sobre os dados da GMC

	coef	std err	*p-value*
bairro[CAXIMBA]	-525.9225	118.909	0.000
bairro[CENTRO]	-471.4077	118.265	0.000
bairro[CENTRO CIVICO]	-522.7323	118.857	0.000
bairro[CIDADE INDUSTRIAL DE CURITIBA]	-505.7536	118.641	0.000
bairro[CRISTO REI]	-525.2374	118.887	0.000
bairro[FANNY]	-525.4652	118.889	0.000
bairro[FAZENDINHA]	-521.0107	118.830	0.000

Fonte: Autoria própria (2022).

Tabela 6 – Parte dos coeficientes de variáveis categóricas de bairro para o modelo ajustado sobre os dados da SESP

	coef	std err	*p-value*
bairro[CAXIMBA]	8070.6648	557.309	0.000
bairro[CENTRO]	8683.6210	558.776	0.000
bairro[CENTRO CIVICO]	8101.3390	557.378	0.000
bairro[CIDADE INDUSTRIAL DE CURITIBA]	8391.1255	558.052	0.000
bairro[CRISTO REI]	8112.7419	557.404	0.000
bairro[FANNY]	8093.0007	557.359	0.000
bairro[FAZENDINHA]	8122.9165	557.426	0.000

Fonte: Autoria própria (2022).

Igualmente ao procedimento utilizado na análise temporal, cada registro foi avaliado de acordo com seu *z-score* e os registros com desvio de mais ou de menos 2,575 - equivalente a 1% dos dados - são classificados como sendo *outliers*.

4.4 Discussão

Os resultados das análises espaciais e espaço-temporais demonstram que o centro de Curitiba é o maior *hotspot* da cidade. Isto fica evidenciado por exemplo verificando-se os valores dos coeficientes da regressão espacial para ambas as bases de dados. Estes coeficientes são maiores para a categoria CENTRO (e bairros ao redor do Centro) em comparação com os demais bairros da cidade. A Cidade Industrial de Curitiba em números absolutos é a segunda que mais apresenta ocorrências criminais, entretanto dada a extensão territorial do bairro e sua população o CIC não representou *hotspots* durante os períodos analisados.

Esta foi uma das descobertas realizadas através das análises anteriormente descritas. Outras informações foram reveladas através das análises realizadas, como por exemplo *hotspots* para diversas categorias de crimes e atendimentos, tendências de aumento e diminuição percebidos para diferentes naturezas de ocorrências, *outliers* temporais, espaciais e espaço-temporais foram detectados através dos modelos criados. Entretanto estas descobertas foram

feitas inteiramente com a utilização *scripts* que muito possivelmente seriam demasiadamente difíceis de serem empregados e implementados no dia a dia de um agente, delegado ou mesmo pelos responsáveis pela secretaria e sub secretarias de Segurança Pública.

O próximo capítulo descreve a implementação da ferramenta que possibilitou a descoberta de diferentes informações contidas nestes dados sem a necessidade de entendimento de programação, e sim apenas familiaridade com a utilização dos navegadores de internet modernos. Os modelos utilizados pela ferramenta foram definidos a partir das análises reportadas acima. Neste contexto, os modelos aplicados pelas abas temporal, espacial e espaço-temporal, bem como as variáveis independentes utilizadas em seus ajustes são descritas na Tabela 7 abaixo. A cada interação com o campo de filtragem de dados os modelos são retreinados com os novos dados de entrada e o utilizador pode acompanhar a qualidade dos modelos através de *labels* contendo o valor de R^2 obtido.

Tabela 7 – Modelos e variáveis utilizadas nas abas da ferramenta Painel Criminal (detalhes no Capítulo 5)

		Fonte de dados / Órgão	
Aba	Modelo	SESP	GMC
Temporal	OLS	TEMPO, LAG1, LAG2, PANDEMIA	TEMPO, LAG1, LAG2, LAG3, LAG4, LAG5, LAG6, LAG7, FINAL DE SEMANA, FERIADO, PANDEMIA
Espacial	LISA	-	-
Espaço-Temporal	OLS	MES, ANO, BAIRRO, LAG1, LAG2, LAG3, LAG4	MES, ANO, BAIRRO, LAG1, LAG2, LAG3, LAG4

Fonte: Autoria própria (2022).

4.5 Explicabilidade de *outliers*

Foi preparada uma tabela de dados para a análise de explicabilidade dos *outliers* temporais com os dados da GMC. Nestes dados estão incluídas as mesmas *features* usadas para o treinamento do modelos temporal da guarda municipal indicadas na seção anterior normalizadas.

A biblioteca *python-oam*[2] foi utilizada para a mineração dos *outlying aspects* dos *outliers* detectados pelos modelos OLS treinados, ou seja, após a detecção dos *outliers* através do *z-score* dos erros entre o número de registros reais registrados e os previstos pelo modelo, estes *outliers* foram passados como entrada para o algoritmo OAM. Como mencionado no capítulo de Fundamentos, a técnica OAM aqui utilizada baseia-se em *score-and-search* empregando-se o algoritmo *iPath* para a tarefa de pontuação dos subespaços.

Foram escolhidos os 3 *outliers* de maior número de ocorrências registradas entre todas as naturezas para a análise de explicabilidade, sendo eles referentes aos dias: 30 de Maio

[2] https://rodrigo-fss.github.io/python-oam/_build/html/index.html

de 2019; 20 de Março de 2021; e 11 de Julho de 2020. Cada um dos *outliers* escolhidos são processados pelo algoritmo de busca utilizando duas parametrizações distintas do *iPath*: (i) tamanho dos subespaços de amostra = 2000 e número de caminhos = 100; (ii) tamanho dos subespaços de amostra = 500 e número de caminhos = 250. O número de caminhos indica quantas divisões serão repetidas para cada subespaço de forma a possibilitar o cálculo da média de cortes necessários para isolar uma *query*. Ainda, foi parametrizado o tamanho mínimo e máximo dos subespaços gerados para 2 e 4 respectivamente. O tempo médio de processamento das *queries* para os dois diferentes algoritmos de busca (i) e (ii) foi de 10 minutos e 11 segundos.

4.5.1 Caso 1

Este caso trata-se do dia 30 de Maio de 2019, onde houve fortes chuvas e temporais na RMC. Os 5 melhores subespaços encontrados estão relatados na Tabela 8 a seguir, de modo que os valores menores de *score* estão relacionados ao número de cortes necessários para o isolamento do *outlier* e portanto também indicam o quão *outlier* uma *query* é dado que neste subespaço a *query* foi mais facilmente isolada do que nos demais.

Tabela 8 – Resultados da execução das buscas (i) e (ii) para *outlier* Caso 1

Algoritmo	Outlying Aspects	Score	Tempo de execução
(i)	Final de semana, Tempo, Lag7	12.60	10m 20s
	Tempo, Lag7, Pandemia	12.78	
	Final de semana, Tempo, Lag7, Pandemia	13.18	
	Tempo, Lag7	13.32	
	Final de semana, Pandemia	13.37	
(ii)	Tempo, Lag7, Pandemia	10.50	19m 19s
	Final de semana, Pandemia	10.60	
	Tempo, Lag7	10.70	
	Final de semana, Feriado	10.84	
	Final de semana, Tempo, Lag7, Pandemia	10.88	

Fonte: Autoria própria (2022).

Este *outlier* também se trata do dia com maior número de ocorrências registradas para todo o período do conjunto de dados. Quando a variável do número de ocorrências é adicionada como dimensão para a geração dos subespaços, ela também aparece em como um *outlying aspect*, já que a *query* é facilmente isolada dado que o número de ocorrências nesta data destoa muito dos demais. As *features* contidas nos subespaços de saída das duas buscas realizadas não são suficientes para indicar o motivo da alta de ocorrências neste dia, dado que o motivo da alta foram os temporais ocorridos na cidade. A variável natureza da ocorrência não foi incluída nas variáveis de entrada do modelo, o que muito possivelmente facilitaria o entendimento do registro, entretanto a biblioteca ainda não permite a utilização de variáveis categóricas.

4.5.2 Caso 2

O Caso 2 se trata do dia 20 de Março de 2021. Como mostrado na Tabela 9, a variável Final de semana parece ser a principal característica deste *outlier* já que esta aparece em 9 dos 10 melhores subespaços encontrados pelas duas buscas executadas. Além disso, esta data coincide com o primeiro dia de vigor do Decreto 600/2021 (Brasil de Fato, 2021). Este decreto intensificou o combate à pandemia de coronavírus determinando maiores restrições na abertura de comércios e na circulação de pessoas. A natureza mais registrada na ocasião foi a de orientação. Possivelmente os órgãos atuaram no controle do movimento nos espaços públicos orientando as pessoas sobre a circulação bem como podem ter aplicado multas à organizadores de festas clandestinas na cidade[3]. Como a data ocorreu num sábado, isto pode ter colaborado com o maior número de pessoas utilizando os espaços da cidade a procura de lazer. Ainda, deve-se ressaltar a presença da variável pandemia como dimensão encontrada pelo algoritmo (i) como sendo um *outlying aspect* desta *query*.

Tabela 9 – Resultados da execução das buscas (i) e (ii) para *outlier* Caso 2

Algoritmo	Outlying Aspects	Score	Tempo de execução
(i)	Final de semana, Lag4	4.81	5m 2s
	Final de semana, Lag4, Lag6	5.04	
	Final de semana, Lag4, Feriado	5.38	
	Final de semana, Lag4, Pandemia	5.40	
	Final de semana, Lag6	5.72	
(ii)	Final de semana, Lag4	3.66	9m 36s
	Final de semana, Lag6	3.78	
	Final de semana, Lag4, Lag6	3.81	
	Lag6, Lag7	3.90	
	Final de semana, Lag6, Lag7	4.03	

Fonte: Autoria própria (2022).

4.5.3 Caso 3

Este caso muito se assemelha com o Caso 2. A data de 11 de Julho de 2020 é o segundo fim de semana após o Decreto 875/2020[4] da PMC que suspendeu os serviços não essenciais na cidade. A natureza mais registrada na data foi a de orientação. Os resultados da execução das buscas OAM mostrados na Tabela 10 também indicam a variável Final de semana como o principal aspecto de explicabilidade deste *outlier* e incluem a variável Pandemia.

[3] https://g1.globo.com/pr/parana/noticia/2021/03/13/covid-19-fiscalizacao-encerra-festa-clandestina-e-interdita-cinco-estabelecimentos-em-curitiba.ghtml

[4] https://mid.curitiba.pr.gov.br/2020/00301721.pdf

Tabela 10 – Resultados da execução das buscas (i) e (ii) para *outlier* Caso 3

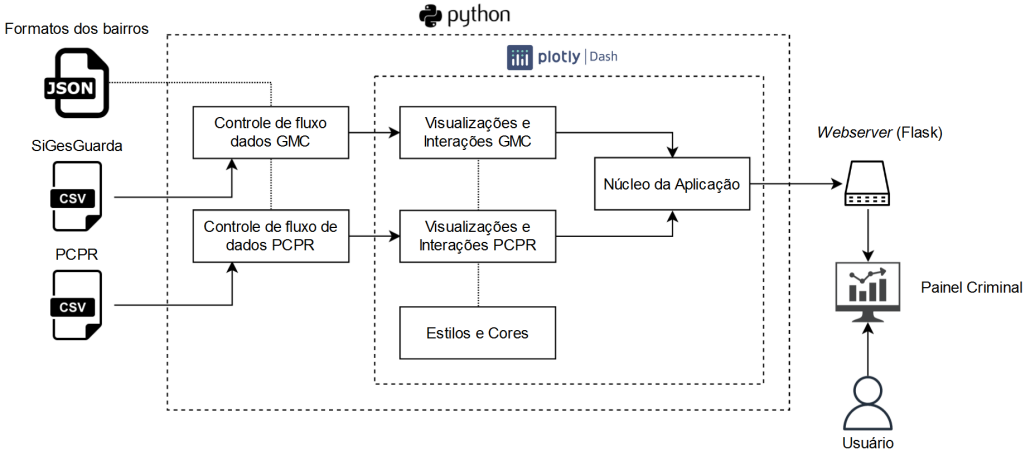
Algoritmo	Outlying Aspects	Score	Tempo de execução
(i)	Final de semana, Lag6	5.68	5m 21s
	Final de semana, Lag4, Lag6	5.98	
	Final de semana, Lag4	6.14	
	Final de semana, Lag1, Lag6	6.21	
	Final de semana, Lag4, Lag6, Feriado	6.23	
(ii)	Final de semana, Lag6	3.89	11m 53s
	Final de semana, Lag4, Lag6	4.22	
	Final de semana, Lag6, Pandemia	4.26	
	Final de semana, Lag1, Lag6	4.31	
	Final de semana, Lag4, Lag6, Lag7	4.35	

Fonte: Autoria própria (2022).

5 IMPLEMENTAÇÃO DA FERRAMENTA

5.1 Arquitetura

Figura 18 – Diagrama da arquitetura do Painel Criminal

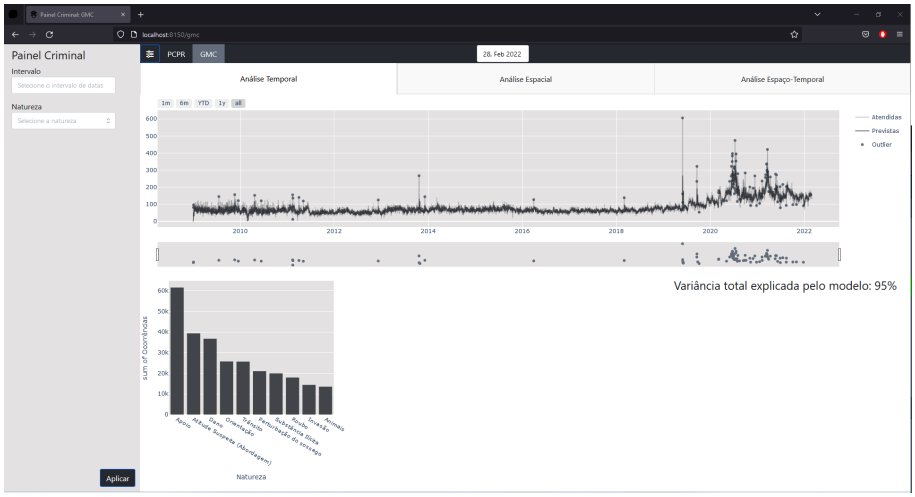


Fonte: Autoria própria (2022).

5.2 Interface

Após realizados os ajustes dos modelos, foi então implementada a ferramenta para visualização dos resultados obtidos. Para esta finalidade foi escolhida a *framework Dash* utilizando-se a linguagem *Python*. A Figura 19 exibe uma visão geral da ferramenta.

Figura 19 – Visão geral da ferramenta



Fonte: Autoria própria (2022).

5.2.1 Cabeçalho

Como os dados utilizados neste trabalho não são buscados periodicamente dos respectivos portais das GMC e SESP e sim refletem intervalos de tempo de dados históricos, adotou-se uma abordagem para que a aplicação considerasse o aspecto de produção constante de dados (a GMC possui frequência de atualização mensal dos dados). A abordagem utilizada portanto consiste em incluir um campo para a seleção de data fazendo com que a aplicação filtre datas posteriores à escolhida utilizando apenas as datas anteriores para a aplicação dos modelos. Também através do cabeçalho é possível selecionar a aba respectiva a qual fonte de dados o utilizador pretende realizar suas análises bem como acessar o menu lateral de filtragem de dados.

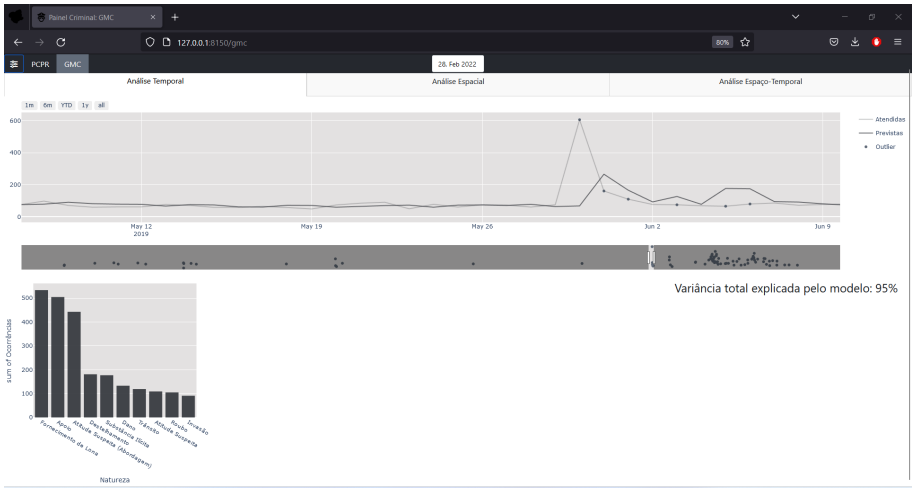
O sistema, além de possibilitar análises das duas diferentes bases de dados utilizadas neste trabalho, permite que o utilizador escolha um entre os três enfoques de análises desenvolvidos: Análise Temporal; Análise Espacial e Análise Espaço-Temporal. Em cada uma delas os componentes apresentados foram escolhidos para que as informações fundamentais a serem apresentadas pela aba sejam a presença e indicação dos *outliers*.

5.2.2 Aba Temporal

Para a aba de Análise Temporal foram desenvolvidas quatro principais componentes dos quais dois possuem capacidade de interação. São os componentes: gráfico de linha contendo as séries temporais de ocorrências atendidas e previstas pelo modelo; *slider* para escolha do intervalo de tempo para a visualização dos gráficos; gráfico de barras contendo as dez naturezas mais atendidas e o número de registros para o intervalo de tempo selecionado; e texto com a indicação do valor de R^2 do modelo ajustado. O gráfico de linha também permite selecionar o período de tempo a qual se deseja analisar através de botões na parte superior do componente bem como ao clicar e arrastar fazendo a ampliação e seleção do intervalo.

Os *outliers* detectados pelo modelo são indicados no gráfico de linhas através da inclusão de um ponto sobre os registros anômalos na série de ocorrências atendidas. A Figura 20 mostra um exemplo de *outlier* detectado pelo modelo e a informação complementar dos registros mais incidentes para o período. Observa-se que dentre as naturezas mais atendidas no intervalo estão "Fornecimento de lona"; "Apoio"e "Destelhamento". A data do *outlier* corresponde a uma grande chuva de granizo ocorrido na Região Metropolitana de Curitiba (RMC), notadamente um evento anômalo.

Figura 20 – Exemplo de *outlier* temporal

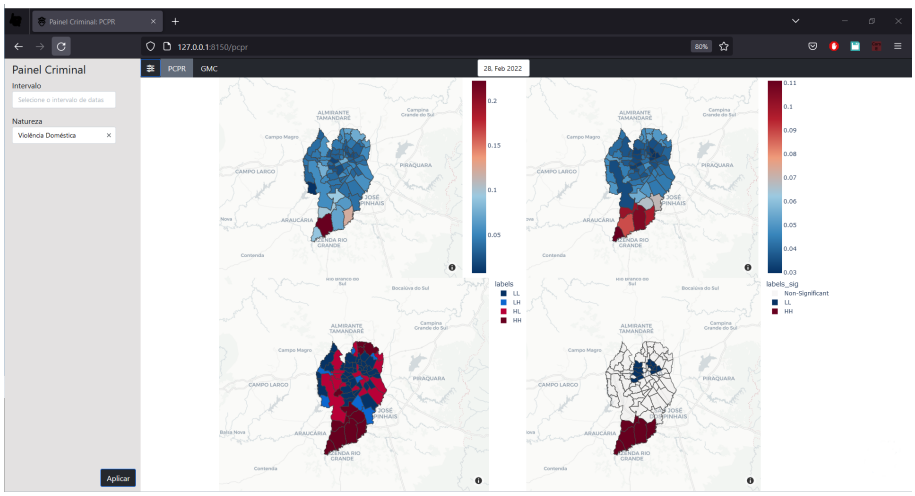


Fonte: Autoria própria (2022).

5.2.3 Aba Espacial

Na aba de Análise Temporal foram implementadas quatro mapas onde são exibidos as etapas de processamento da técnica LISA. O primeiro deles - localizado na parte superior esquerda da tela - exibe o mapa da cidade com a divisão de bairros coloridos de acordo o número de ocorrências registrados no período normalizados pela população do bairro. O segundo mapa, na parte superior direita, contém a divisão de bairros coloridas de acordo com o *spatial-lag* associado. Nos mapas presentes na região inferior da tela os bairros são apresentados de acordo com a categoria em que foram classificados após a aplicação do modelo *Moran*. O mapa à esquerda contém a categoria - como vistas na Seção 4.2.1 - de todos os bairros enquanto que o mapa à direita insere a categoria *Non-Significant* para os bairros que não passaram pelo teste de hipótese do *p-value* simulado. A Figura 21 mostra um exemplo de *hotspots* e *coldspots* processados e exibidos através da Aba Espacial.

Figura 21 – Exemplo de *hotspots* e *coldspots* encontrados pela Aba Espacial

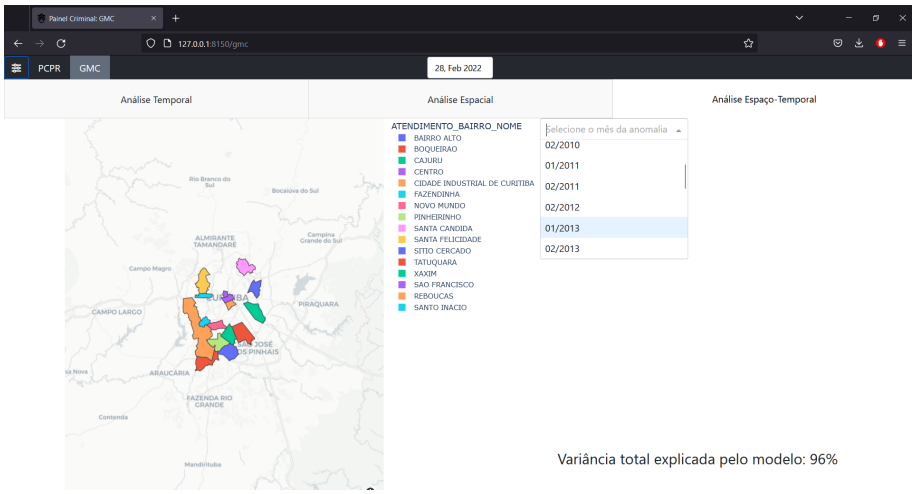


Fonte: Autoria própria (2022).

5.2.4 Aba Espaço-Temporal

Nesta aba é apresentado um mapa com destaque aos bairros considerados *outliers*. Ao lado direito do mapa, há um campo para visualização e seleção das datas em que se encontraram anomalias e abaixo do seletor - nos moldes do que se apresenta na aba temporal - há um texto com a mensagem "Variância total explicada pelo modelo"e o valor de R^2 obtido. Ao selecionar uma data, equivalente ao mês da anomalia já que neste modelo os dados foram agregados por mês como mencionado na Seção 4.3.2, o mapa destaca os bairros dos registros classificados como *outlier*. Ao passar o *mouse* sobre os bairros destacados uma *pop-up* é exibida contendo o número de ocorrências previstas pelo modelo e as ocorrências atendidas. A Figura 22 apresenta a aba Espaço-Temporal.

Figura 22 – Aba Espaço-Temporal da ferramenta

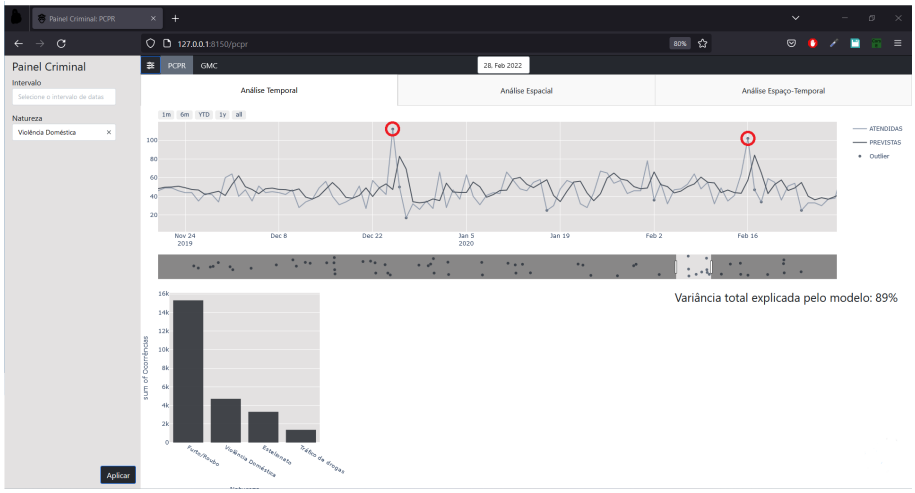


Fonte: Autoria própria (2022).

5.3 Casos de uso

Em um cenário hipotético, um utilizador poderia estar interessado em analisar as ocorrências de Violência Doméstica nos últimos anos. Através da aba temporal este usuário vai identificar uma tendência de aumento nas ocorrências deste gênero. A ferramenta destacaria *outliers* para esta categoria criminal, por exemplo um *outlier* para o Natal do ano de 2019 que registrou 112 ocorrências enquanto o modelo com R^2 de 89% esperava 47, mais da metade do valor registrado. Outro *outlier* que a ferramenta detectaria e apresentaria nesta categoria seria o dia 16 de Fevereiro de 2020, domingo anterior ao Carnaval daquele ano e que também foi data da realização de uma final de competição de futebol no país envolvendo um dos times da capital. Naquela ocasião a SESP registrou 103 ocorrências enquanto o modelo presente na ferramenta indica um número de ocorrências esperado de 57, como exibido na Figura 23.

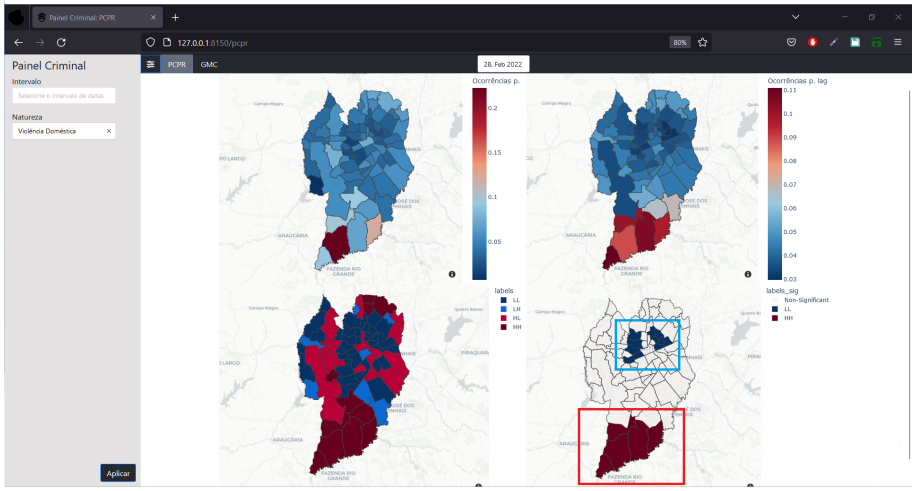
Figura 23 – Destaque da ferramenta dado para os *outliers* ocorridos nas semanas do Natal de 2019 e Carnaval de 2020



Fonte: Autoria própria (2022).

Mais *outliers* semelhantes àqueles são detectados em outros feriados e datas festivas. Na aba espacial, o usuário verificaria que os bairros da região sul da cidade, nomeadamente: Campo de Santana, Caximba, Ganchinho, Tatuquara e Umbará; após o processamento através dos indicadores locais de associação espacial, são *hotspots* para a Violência Doméstica, enquanto que os bairros mais centrais como: Ahu, Alto da Rua XV, Batel, Bigorrilho, Cabral, Hugo Lange, Jardim Social, Juvevê, Mercês, Seminário e Vista Alegre, são *coldspots* para este tipo de ocorrência. Estes *outliers* também seriam encontrados através da aba espaço-temporal. A Figura 24 a seguir demonstra como estes *outliers* são exibidos pelo Painel Criminal.

Figura 24 – *Outliers* espaciais para as ocorrências de Violência Doméstica encontrados pelo Painel Criminal

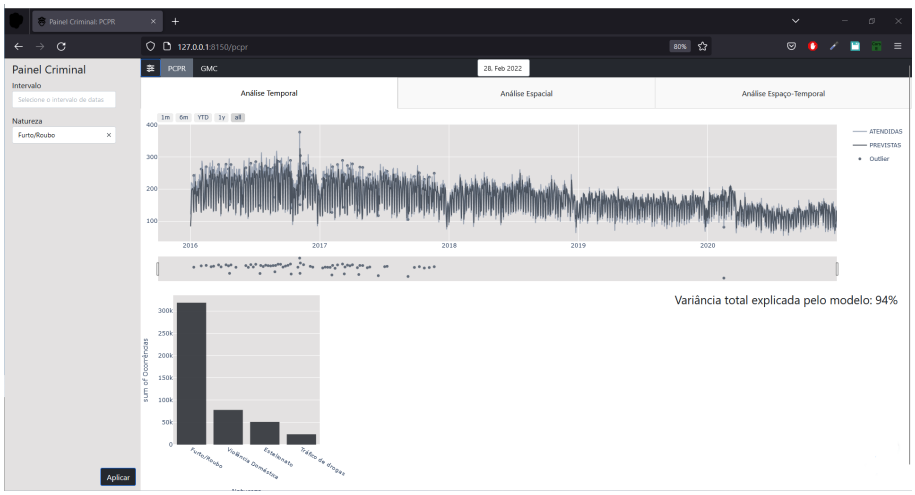


Fonte: Autoria própria (2022).

Caso o usuário estivesse interessado em analisar ocorrências de Furto e Roubo, a ferramenta exibiria uma tendência de diminuição nos registros (Figura 25), assim como relatado

por Leal (2022) que também relacionou a diminuição com dados da pandemia de COVID-19 ocorrida no período.

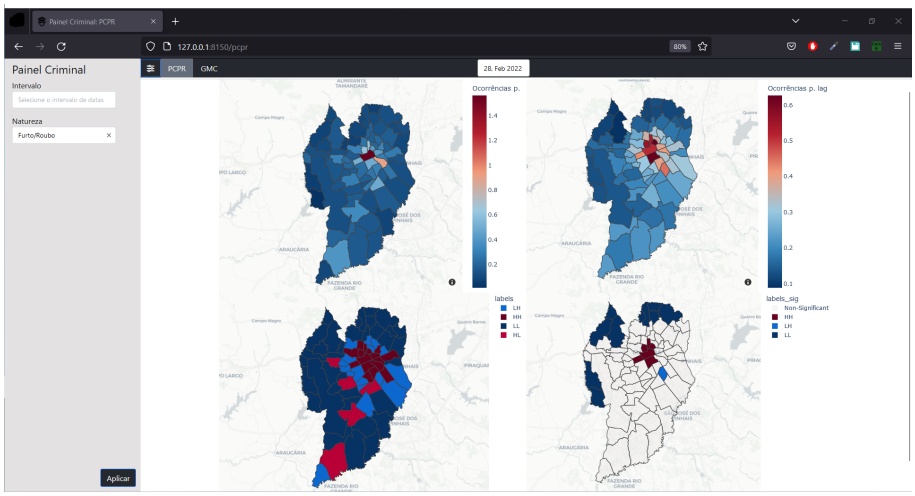
Figura 25 – Aba Temporal do Painel Criminal indicando a tendência de diminuição nos registros de ocorrências de Furto e Roubo



Fonte: Autoria própria (2022).

Neste caso, na aba espacial da GMC com a natureza Roubo selecionada, o bairro Guabirotuba seria destacado como *outlier* LH - aqueles registros em que houve baixo número de ocorrências em regiões de alta correlação para alta incidência, enquanto que o Centro e o São Francisco seriam identificados como *hotspots* ao passo que os bairros da região nordeste como o Bacacheri e Santa Cândida, bem como bairros mais à oeste como Santa Felicidade, Butiatuvinha, Lamenha Pequena, Augusta e São Miguel, seriam mostrados como *coldspots*, como mostrado na Figura 26.

Figura 26 – Aba Temporal do Painel Criminal indicando a tendência de diminuição nos registros de ocorrências de Furto e Roubo



Fonte: Autoria própria (2022).

Analisando os da SESP, o usuário verificaria na aba temporal que os bairros Guabirotuba e Hauer são dois exemplos de *outliers* espaciais para o Tráfico de Drogas pois apresenta-

ram indicador LH, indicador de correlação para baixa incidência em regiões de elevado número de casos, como são os bairros: Guaíra, Jardim Botânico, Parolin e Prado Velho. Já os bairros ao extremo norte da cidade foram classificados como *coldspots*. Para os dados da Guarda Municipal na categoria Substância Ilícita os *hotspots* exibidos pelo Painel Criminal são o Centro, Centro Cívico e São Francisco. Vale notar que o combate ao tráfico de drogas é realizado pela Polícia Civil do Paraná (PCPR) enquanto que muito possivelmente os registros de substância ilícita feitos pela GMC podem tratar somente da posse e não de operações de combate ao tráfico.

O utilizador encontraria *outliers* interessantes ao analisar a série temporal de todos os registros são relativos à Destelhamentos e atendimentos de Fornecimento de Lonas, ocorridos em dias de chuvas intensas registradas na cidade. Concentrado especialmente nos bairros de Atuba, Santa Cândida e Tingui as datas de maior incidência foram os dias: 30 de Maio, 18 e 19 de Setembro do ano de 2019; e 28 de Setembro de 2020, ocasiões que registraram ventos fortes e chuvas de granizo. O usuário poderia encontrar os picos de atendidas que são também indicados como *outlier* através da inclusão de um ponto sobre a série temporal e ao realizar a ampliação do gráfico de linhas, o gráfico de barras indicaria quais os atendimentos mais registrados no período capturado pela ampliação que sinalizaria as duas categorias citadas como as mais frequentes no intervalo de tempo.

6 CONCLUSÃO

Este trabalho empenhou-se em criar uma ferramenta capaz de contribuir na complexa tarefa de analisar dados para a tomada de decisões no âmbito da Segurança Pública. As análises geradas e modelos aqui aplicados contribuem para a verificação das tendências e padrões criminais bem como na detecção de *outliers* nas bases de dados da Guarda Municipal de Curitiba e da Secretaria da Segurança Pública.

Através do Painel Criminal desenvolvido foi possível realizar a descoberta de diversas informações relevantes. Inicialmente, observa-se através da análise temporal dos dados da GMC o aumento nos registros de ocorrências realizados pelo órgão especialmente a partir do ano de 2018. Isto pode indicar uma maior consideração por parte do poder público à respeito da importância dos dados para o planejamento de políticas, aumento este de atenção no valor dos dados que acompanha as tendências dos últimos anos na utilização da *Big Data* como norteadora de estratégias e decisões em empresas e governos.

A ferramenta demonstrou efetividade relativamente a visualização dos dados bem como para o processamento e destaque de *outliers*. Os gráficos apresentados para as visualizações temporais, espaciais e espaço-temporais contribuem efetivamente para o entendimento das quantidades e localizações das ocorrências, indicando assim também padrões espaço-temporais criminais da cidade. Ainda, os modelos utilizados obtiveram de modo geral coeficientes de determinação consideráveis bem como as variáveis explicativas utilizadas alcançaram significância estatística verificadas através do *p-value*. As análises desenvolvidas utilizando a regressão espacial, por exemplo, demonstraram através dos coeficientes das variáveis categóricas dos bairros da cidade que a região central da cidade é onde há a maior incidência criminal *per capita* e a técnica LISA demonstrou como os bairros da região sul da cidade são *outliers* espaciais para as ocorrências de violência doméstica.

Não obstante, há espaço para melhorias e enriquecimento das visualizações e modelos utilizados pelo Painel Criminal. O usuário pode acabar esperando por alguns segundos sem que haja algum indicativo de carregamento e processamento por parte da ferramenta além da label *Updating* no título da aba do navegador em que a aplicação encotra-se aberta. O menu lateral responsável pela filtragem dos dados também poderia incluir um campo para a escolha do valor de *z-score* para a classificação dos *outliers*. As técnicas aplicadas na análise de dados e modelagem, como por exemplo os modelos ARIMA e LOF, na ferramenta não foram incluídos porque os demais modelos apresentaram melhor performance. Entretanto, estes modelos poderiam também ser incluídos de modo a complementar as análises de *outliers* temporais e espaço-temporais respectivamente.

As análises de explicabilidade de *outliers* através do método OAM demonstraram interessante utilidade especialmente na explicação dos registros com elevado número de ocorrências acontecidas em finais de semana e pandemia. Vale ressaltar que os casos analisados foram os de maior desvio no número de registros previstos pelos modelos e a quantidade real de

ocorrências registradas, portanto estas variáveis apresentaram notável grau de explicabilidade destes *outliers*. O sucesso do método - assim como nas demais aplicações de mineração de dados - depende da existência de dados de entrada de qualidade. Isto implica por exemplo no enriquecimento dos dados através da agregação de informações meteorológicas para a explicação dos *outliers* de atendimentos por conta de tempestades e temporais. No entanto, embora exista a possibilidade de agregar mais variáveis de entrada para o algoritmo OAM, ainda não é possível incluir variáveis categóricas no problema, o que agregaria por exemplo na explicabilidade dos *outliers* relacionados com ocorrências de orientação durante os decretos de controle da pandemia. Embora o OAM não permita a inclusão das variáveis categóricas, o Painel Criminal possibilita a verificação das ocorrências mais registradas no período analisado através do gráfico de barras contendo as 10 naturezas de ocorrências mais frequentes no período. Entretanto, é possível agregar mais visualizações ao Painel Criminal que também colaborariam com a explicabilidade destes *outliers*. A ferramenta também poderia fornecer atualizações sobre seu status e o status de execução dos modelos utilizados, favorecendo o aumento de confiança dos usuários relativamente a estes modelos. As mensagens contendo o valor de R^2 dos modelos exibidas pelas páginas da aplicação tentaram de alguma forma colaborar neste sentido.

De maneira geral, as análises aqui realizadas bem como a ferramenta desenvolvida atingiram os objetivos iniciais deste trabalho, de modo que o nível atual de desenvolvimento do Painel Criminal, se disponibilizado para os usuários alvo, já possui grande potencial de colaborar com o fornecimento de informações relevantes para a tomada de decisões por parte da SESP e Guarda Municipal. Casos interessantes de *outliers* foram encontrados e foi possível combinar técnicas relevantes de processamento e análise de dados espaço-temporais.

REFERÊNCIAS

ANISH, A. **Time Series Analysis — medium.com**. 2020. Disponível em https://medium.com/swlh/time-series-analysis-7006ea1c3326.

ANSELIN, L. Local indicators of spatial association-LISA. **Geogr. Anal.**, Wiley, v. 27, n. 2, set. 2010.

BECKETT, L. Us records largest annual increase in murders in six decades. **The Guardian**, 2021. Disponível em: https://www.theguardian.com/us-news/2021/sep/27/us-murder-rate-increase-2020.

BEWICK, V.; CHEEK, L.; BALL, J. Statistics review 7: Correlation and regression. **Crit. Care**, Springer Nature, v. 7, n. 6, p. 451–459, dez. 2003.

Brasil de Fato. **Prefeitura de Curitiba prorroga medidas restritivas por mais uma semana**. 2021. https://www.brasildefato.com.br/2021/03/20/prefeitura-de-curitiba-prorroga-medidas-restritivas-por-mais-uma-semana. Acessado em 27 de Novembro de 2022.

BREUNIG, M. M. *et al.* Lof: identifying density-based local outliers. In: ACM. **ACM sigmod record**. [*S.l.*], 2000. v. 29, n. 2, p. 93–104.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Computing Surveys (CSUR)**, ACM, v. 41, n. 3, p. 15, 2009. Disponível em: http://scholar.google.de/scholar.bib?q=info:jAfBmk-9uAcJ:scholar.google.com/&output=citation&hl=de&ct=citation&cd=0.

CURITIBA, P. M. D. Dados abertos. 2015. Disponível em: https://www.curitiba.pr.gov.br/dadosabertos/.

DOYLE, A. C. **The Hound of the Baskervilles**. United Kingdom: The Strand Magazine, 1901.

GARCIA, G. *et al.* Crimanalyzer: Understanding crime patterns in São Paulo. **IEEE transactions on visualization and computer graphics**, IEEE, v. 27, n. 4, p. 2313–2328, 2019.

GUNNING, D. *et al.* XAI-Explainable artificial intelligence. **Sci. Robot.**, American Association for the Advancement of Science (AAAS), v. 4, n. 37, p. eaay7120, dez. 2019. Disponível em: https://openaccess.city.ac.uk/id/eprint/23405/8/aay7120_article_nearly%20final.pdf.

HAWKINS, D. **Identification of outliers**. London [u.a.]: Chapman and Hall, 1980. (Monographs on applied probability and statistics). ISBN 041221900X. Disponível em: http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+02435757X&sourceid=fbw_bibsonomy.

IPEA, I. Atlas da violência. 2017. Disponível em: https://www.ipea.gov.br/atlasviolencia/.

LEAL, M. F. **Impacto da pandemia da COVID-19 nos padrões de crimes: análises espaço-temporais para avaliar o passado e informar o presente**. 2022. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2022.

LEAL, M. F.; GOMES-JR, L. Cityguardian: Uma ferramenta para monitorar mudanças em padrões de criminalidade nas cidades inteligentes. In: SBC. **Anais Estendidos do XVIII Simpósio Brasileiro de Sistemas de Informação**. [*S.l.*], 2022. p. 366–372.

MASULLO, Y. A. G. *et al.* Diagnóstico espaço-temporal dos crimes violentos letais em São Luís, Maranhão. Universidade Federal da Grande Dourados (UFGD), 2017.

MOHAN, A. **Local Outlier Factor**. 2018. Disponível em: https://arunm8489.medium.com/local-outlier-factor-13784dc1992a.

REY, S. J.; ARRIBAS-BEL, D.; WOLF, L. J. Global spatial autocorrelation. 2020. Disponível em: https://geographicdata.science/book/notebooks/06_spatial_autocorrelation.html.

ROSER, M.; RITCHIE, H. Homicides. **Our World in Data**, 2013. Disponível em: https://ourworldindata.org/homicides.

SAMARIYA, D.; MA, J.; ARYAL, S. **A Comprehensive Survey on Outlying Aspect Mining Methods**. arXiv, 2020. Disponível em: https://arxiv.org/abs/2005.02637.

SILVA, L. J. S. *et al.* Crimevis: An Interactive Visualization System for Analyzing Crime Data in the State of Rio de Janeiro. In: **International Conference on Enterprise Information Systems (CEIS)**. [*S.l.*: *s.n.*], 2017. p. 193–200.

SILVA, R.; GOMES-JR, L. Python OAM: apresentação e uso de uma biblioteca de explicabilidade para processos de detecção de outliers. In: **Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados**. Porto Alegre, RS, Brasil: SBC, 2022. p. 71–76. ISSN 0000-0000. Disponível em: https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/21846.

VINH, N. X. *et al.* Discovering outlying aspects in large datasets. **Data Mining and Knowledge Discovery**, Springer Science and Business Media LLC, v. 30, n. 6, p. 1520–1555, nov. 2016.