A Descriptive and Predictive Analysis Tool for Criminal Data: A Case Study from Brazil

Yan Andrade[1(✉)], Matheus Pimenta[3], Gabriel Amarante[3], Antônio Hot Faria[2,3], Marcelo Vilas-Boas[2], João Paulo da Silva[2], Felipe Rocha[2], Jamicel da Silva[2], Wagner Meira Jr.[3], George Teodoro[3], Leonardo Rocha[1], and Renato Ferreira[3]

[1] Universidade Federal de São João del-Rei, São João del Rei, Brazil
yrandrade123@aluno.ufsj.edu.br, lcrocha@ufsj.edu.br
[2] Polícia Militar de Minas Gerais, Belo Horizonte, Brazil
{1265677,1357516,1369016,jamicel.silva}@pm.mg.gov.br
[3] Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
{matheuspimenta,gabriel.amarante,antonio.hot,meira,george,renato}@dcc.ufmg.br

Abstract. Addressing the challenge of crime is crucial for governing bodies, requiring informed strategies. This article examines the underutilization of detailed criminal data, collaborating with the Military Police of Minas Gerais, Brazil. We propose a new methodology, materialized in a tool, that is able to transform raw data into strategic information for public security decision-making. The tool evaluation unfolds in three phases: characterizing the data, a descriptive analysis of a real case study, and a predictive analysis. This work highlights the untapped potential in detailed criminal data, emphasizing the pivotal role of precise analysis in deciphering complex dynamics. Collaborating with law enforcement aims to bridge the gap between data abundance and actionable insights for effective public security strategies.

1 Introduction

Security impacts all dimensions of everyday life, from personal development to the business landscape [3]. The effects of the feeling of insecurity influence the quality of life, restrict mobility, and impact the economy, undermining investor confidence and market stability. The persistent fight against crime poses a significant challenge for governments, requiring well-founded and effective strategies. The inherent complexity of this challenge, connected to a network of social, demographic, and educational factors, highlights the need for a precise understanding for an effective approach [10]. In this context, understanding and characterizing crime and how all the factors inherent to it are related becomes a crucial task for proposals of effective and comprehensive solutions.

Considering this, the capacity of public security organizations to collect and store detailed data on criminal events has experienced significant growth, resulting in the daily generation of large volumes of information. These data, enriched

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2024
O. Gervasi et al. (Eds.): ICCSA 2024, LNCS 14814, pp. 151–169, 2024.
https://doi.org/10.1007/978-3-031-64608-9_10

with spatial and temporal details, offer a comprehensive overview of criminal dynamics. However, within this scenario, with excess information, a problem emerges: the underutilization of these data [12]. The lack of appropriate analysis compromises the extraction of valuable insights and constitutes an obstacle to informed decision-making and the development of effective public policies in the fight against crime.

In the face of this challenging context, in this paper, we present a methodology for characterizing previously stored criminal data that allows it to be transformed into information that helps develop assertive crime-fighting and prevention tactics. Our methodology aims to understand the dynamics of crime occurrence from two different perspectives: temporal and spatial. Based on this understanding, in this work, we also evaluate the application of regression models that can correlate information from the two perspectives considered, helping predict where and when crimes will occur in the future. The entire methodology is materialized in a tool focused on a simple and intuitive interface to be used by public agents responsible for security.

The tool comprises six distinct modules: Scenario Filtering, Trend Analysis, Distribution Analysis, Temporal Heat Map, Layered Event Map, and Regression Models. In the first module, we provide a series of filters that allow us to select the data set for the scenario we want to analyze, for example, theft crimes on weekdays during the afternoon in the downtown neighborhood. Several scenarios can be selected and contrasted. The second module (Trend Analysis) provides visual metaphors that allow us to assess the temporal evolution of a particular type of criminal occurrence. In the scenario previously presented, we can assess the trend of theft occurrences in the city center. The third module (Distribution Analysis) provides intuitive visualizations that allow us to analyze how the distribution of criminal occurrences occurs in different regions and periods (dawn, morning, afternoon, and evening). The tool is flexible enough and allows the inclusion of new distributions for analysis. The fourth module (Temporal Heat Map) provides an intuitive visualization of the distribution of criminal occurrences in a given period by combining a heat map with a calendar. In the fifth module (Layered Event Map), we implemented a georeferenced visualization from a map in multiple layers. In the first layer, the neighborhoods of a given city are delimited according to official city data. Over this first layer, a heat map allows us to identify the neighborhoods with the highest incidence of a specific crime regionally. In the third layer, we provide a map with the distribution of points that represent the incidents, and in the fourth layer, through the integration with a clustering algorithm that groups occurrences according to geographical proximity, identifying and highlighting the critical points according to the concentration of crime incidents. Finally, our last module comprises a series of implemented geospatial regression algorithms that allow us to create predictive models on crime incidence, helping to identify when and where there is the highest probability of incidence of a particular crime. All these modules working together are essential to help the competent authorities decide about optimized public policies to combat and prevent crime.

We instantiated our methodology using data from Brazil, one of the 20 countries with the highest crime rates in the world, according to the Global Organized Crime Index[1]. More specifically, we used the criminal records stored by the Military Police of the State of Minas Gerais (PMMG), Brazil, which is subordinate to the Public Security Secretariat of Minas Gerais, one of the most violent states in Brazil. Since 2005, the PMMG has had a system for recording criminal occurrences and currently has over 40 million cases related to the 853 municipalities in Minas Gerais. These occurrences range from violent crimes, such as theft, robbery, and homicide, to less impactful occurrences from a public security standpoint, such as arguments and verbal aggression. The occurrences are defined employing 224 attributes that provide a wide range of information, from aspects such as temporal data, including the date and time of the occurrence, spatial aspects such as latitude, longitude, municipality, and address, to bureaucratic details associated with each registered incident.

Among the various results we obtained from the instantiation of our tool in this database, we highlight some that demonstrate the potential of our tool. In a more specific analysis, we compared the crime map before, during, and after Carnival, the world's most famous Brazilian popular festival. We focused our analysis on Belo Horizonte, the capital of the State of Minas Gerais. We observed a significant change in criminal behavior during Carnival, with a substantial increase in crimes, concentrated in places where specific events occur (carnival blocks), with a large concentration of tourists. Identifying and characterizing this atypical behavior is essential for developing specific combat and prevention strategies, helping to allocate public security agents more efficiently. In addition to this descriptive analysis, we detail a predictive data analysis in our results. Also, considering the city of Belo Horizonte during January, we trained and evaluated two crime prediction models. Both models demonstrated satisfactory performance, presenting predicted values close to the actual data. These prediction models allow public security to develop preventive actions more assertively. Integrating all modules of the proposed tool offers an approach with complete descriptive and predictive analyses, contributing more effectively to improving public safety.

2 Related Work

Crime data analysis is the subject of several works investigating different data mining techniques for identifying patterns and relationships between criminal events. In [11], the authors propose using association rule algorithms to find correlations between descriptive attributes and crime types. Using Apriori [8], recognized for its effectiveness in several domains, the authors identified rules that correlated attributes such as day of the week, weather conditions, holiday, time of day, and crime type with the location of the occurrence. Notably, most crimes tended to occur on days without rain and in residential areas. Another relevant observation of the authors was the absence of frequent pairs for sexual

[1] https://ocindex.net/rankings?f=rankings&view=List.

crimes, suggesting the possibility of occurring randomly at different times and places. The rules found were contrasted with a heatmap of occurrences showing great correspondence. Based on this last observation, we opted for our tool to use heatmaps.

Another set of techniques widely used in criminal characterizations is clustering [19] and outlier detection [17]. In [16], the authors present a study that analyzes spatio-temporal patterns of crimes in micro-locations (streets) in six major US cities using clustering algorithms. The authors identify the micro-locations with the highest crime incidence at different times and perform a comparative temporal evaluation between them. In their conclusions, the authors identify a pattern, and many regions with the highest crime rates are persistent over time. Our tool also uses clustering algorithms. More specifically, we perform a georeferenced grouping of criminal occurrences, identifying and highlighting the points of highest incidence, which we call critical points, on the map. This analysis is also integrated with a calendar, allowing us to expand the scale of the spatial analysis and provide a more comprehensive view of crime patterns over time.

In another research line, we identified works that analyze the importance of applying pre-processing techniques before conducting analyses on criminal data. In [14], the authors present a taxonomy composed of two types of errors: systematic (underreporting and undercounting of crimes) and random (inconsistencies in the reporting and recording processes of crimes between victims, areas, and police forces). In the evaluations, the authors show that measurement errors can lead to biased estimates in all prediction models. Our work takes a similar stance to the article [14]. Both recognize the importance of processing the data before analyzing it. In our case, we implemented specific processing methods, such as logarithmic transformation, as proposed in this study, for discrepancies in criminal data records.

As mentioned in the previous paragraph, in addition to works on characterization in understanding criminal data, we observe a growing line of works that aim to build spatio-temporal predictive models [2], with emphasis on time series. These are notable for the chronological indexing of their observations [20] and, consequently, crucial in analyzing trends and patterns over time. This approach is divided into three main components: Trend, which reflects the long-term structural direction of the data; Seasonality, related to repetitive patterns at fixed intervals, capturing regular fluctuations; and Noise, representing residuals after removing seasonality and trend, expressing non-systematic and random variation in the data. A relevant study [5] employed the ARIMA model [21] to perform temporal regression, exploring different parameters and their impacts. Focusing on the Chicago region, the goal was to identify critical areas of the state. Using hotspot and clustering techniques, the study outlined these regions and applied regression algorithms to identify specific trends in each one. They demonstrated the need to adjust the model's hyperparameters for each region. This integrated approach offered a deeper understanding of the temporal behavior in specific regions. Similar work is presented in [9] using ARIMA with criminal data from India. In [15], the authors present a temporal analysis technique for criminal

data called Heartbeat, which overlays spatial, temporal, and intensity information, separated by crime type, into a predictive model. The authors observed that the Heartbeat was able to indicate specific moments of higher occurrence for different types of crime. Going a little further, [18] analyzes the impact of well-made criminal predictions on the proper allocation of patrolling and, consequently, on crime reduction.

As we will see in the next section, our tool also allows training models capable of performing spatio-temporal predictions. We provide a series of filters that allow the creation of various scenarios for training predictive models, from more specific scenarios (i.e., theft crimes during the night in central neighborhoods) to more general scenarios (i.e., crimes in general throughout the city). This allows the development of crime-fighting tactics customized according to region, time, and type of crime, the development of articulated strategies between different public security agents from different regions, and even the development of medium- and long-term public safety policies. We will detail our proposal below.

3 Criminal Analysis Tool

The objective of the proposed tool is to transform criminal records into valuable information that can assist public security officials in the development of (1) assertive crime-fighting strategies, as well as (2) the construction of crime prevention public policies. Our tool uses both descriptive models, which aim to continuously map occurrences (crime-fighting strategies), and predictive models, which aim to anticipate where and when crimes will occur (crime prevention public policies). Thinking about these objectives, we created a complete tool that, in addition to geographic visualization, provides in-depth data visualization, scenario simulation, identification of trends in the number of occurrences, data distribution analysis, visual representation through temporal heat maps, and methodology to highlight points that require special attention. The tool was implemented using the Python programming language and the Streamlit library due to their ability to create dashboards that comprehensively and intuitively present data. An overview of the tool is presented in Fig. 1, and in this section, we present a detailed description of each component of the tool, highlighting how they complement each other to provide a comprehensive and effective decision-making tool. Our focus is on the ability of the tool to transform complex data into actionable insights and increase a deep understanding of the crime landscape to optimize public security strategies.

3.1 Scenario Filtering

Typically, criminal records are composed of several attributes that detail the occurrences. These attributes can be related to temporal perspectives (date and time of the occurrence), geographic (city, neighborhood, street, etc.), and descriptive (type of crime, total victims, etc.). Performing a global analysis of all data can be computationally infeasible and also uninformative from the point

of view of the conclusions that can be generated. In addition, public security actions and responsibilities are segmented in most countries. That is, specific organizations are responsible for combating and preventing crime according to the locality. For this reason, the first component of our tool is filters (item 1 in Fig. 1) that allow users to analyze different scenarios by segmenting the data representation according to specific criteria. The filters provided by our tool are:

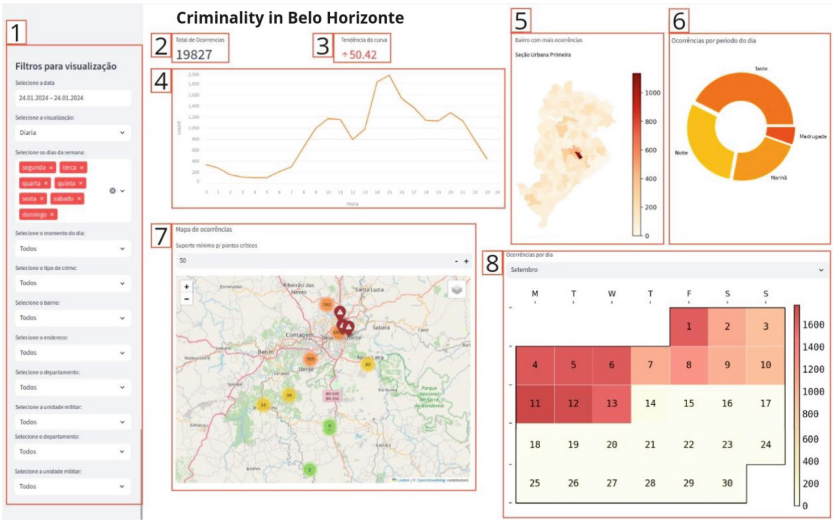


Fig. 1. Main page of the tool

- **Time interval**: Provides a custom time visualization by selecting a specific date range to perform the analysis;
- **Granularity**: Offers the ability to visualize data at different temporal granularities, such as daily, monthly, or annual, allowing for a more detailed or more generalized view of the data;
- **Day of the week**: Allows users to select specific days of the week for analysis, helping to identify patterns of events in different periods, for example, only weekends, only weekdays, or only a specific day of the week.;
- **Time of day**: By dividing the day into four periods (early morning, morning, afternoon, and evening), it is possible to analyze more accurately the difference in crime rates at different times of the day;
- **Neighborhoods**: We standardized the data based on latitude and longitude, allowing users to select specific neighborhoods within each municipality and providing more localized analyses. For an adequate neighborhood analysis, it is necessary to integrate this filter with the official map of the cities, where the boundaries between the neighborhoods are highlighted;
- **Address**: Facilitates the more detailed selection of the location, including street information for a more fine-grained analysis;

- **Crime types**: Covers all types of incidents found in the data, allowing for a detailed analysis of the different types of crimes registered;
- **Department**: Allows the choice of the responsible department, such as police stations or police posts, contributing to an analysis by area of responsibility. For an adequate department analysis, it is necessary to integrate this filter with the official distribution of police stations/departments of each city;
- **Unit**: Allows the selection of the responsible unit, following the pre-established hierarchy by the public security agency. For an adequate unit analysis, it is necessary to integrate this filter with the pre-established hierarchy of a given country or state;

Based on these filters, various scenarios can be constructed for analysis. For example, one could investigate the distribution of robberies in neighborhood "X" in the early morning hours on weekends from January to April, contrasting it with a similar scenario but evaluating weekdays. Alternatively, one could perform an analysis with the same configuration evaluating different neighborhoods. Another interesting example would be to analyze entire and nearby neighborhoods, contrasting the information found with and without a specific event, such as a soccer game or a concert. In short, through these filters, we have infinite scenarios that can be created and contrasted in various analyses.

3.2 Trend Analysis

The first analysis module of the tool focuses on three key elements to provide a comprehensive understanding of the temporal trends of events. The first field (item 2 in Fig. 1) displays the total number of incidents based on the filters used, allowing you to see the number immediately. The second graphical component (item 4 in Fig. 1) visually highlights the behavior of the event, allowing cumulative analyses over time-based on the defined granularity, including day, month, or year.

This section highlights the fusion of trend lines obtained by applying a regression algorithm to the graph points. We decided to use linear regression as our model for its simplicity and efficiency in simple cases and online systems. This technique provides functions that model the relationships between points in a graph, allowing us to obtain a mathematical expression for the defined time flow. We use the derivative of the regression function at the last point of the graph to determine if the trend is increasing or decreasing (item 3 in Fig. 1). The arrows next to the numerical value indicate the rate of change in the number of cases, highlighting the derived values. It is important to note that this value does not directly correspond to the rate of change but provides a relative measure of the slope of the trend. This approach combines visual interpretation with powerful mathematical analysis to provide valuable insights into the temporal patterns of events.

3.3 Distributions Analysis

The distribution analysis module is designed to provide a detailed and comprehensive view of criminal events' spatial and temporal aspects, including updated distribution by neighborhood and time of day. Each of these distributions represents critically important aspects of the data, allowing for a deeper analysis of the spatial and temporal nuances of the events.

In our neighborhood analysis (item 5 in Fig. 1), we highlight the neighborhoods with the highest incident rates, providing a clear visual view of the most impacted geographic areas, where the redder the neighborhood, the higher its crime rate. This spatial analysis is important for identifying concentration patterns in certain areas and guiding specific public safety strategies.

The assignment by the time of day (item 6 in Fig. 1) completes this analysis by examining the variation in occurrences during the different times of day. Dividing the day into specific periods, such as early morning, morning, afternoon, and evening, this distribution allows for the identification of temporal patterns that can influence the allocation of resources and the effectiveness of preventive measures.

It is important to note that the flexibility of this module allows for the inclusion of multiple distributions for different attributes present in the data. This adaptability adjusts the tool to meet specific cases, ensuring a personalized and in-depth approach to criminal data analysis. This adaptability reinforces the tool's versatility, making it a valuable ally in various public safety scenarios.

3.4 Temporal Heat Map

The Temporal Heatmap module was developed to provide an efficient and intuitive visualization of the distribution of occurrences over a given period. This approach combines two important and complementary concepts: a heat map (item 7 in Fig. 1) that identifies and highlights areas of high concentration and a calendar (item 8 in Fig. 1) that displays the days of the month. Combining these factors provides a dynamic visual representation of the concentration of incidents on each day of the month, allowing for a comprehensive and detailed analysis. The integration provides visibility of all months within the selected interval, allowing for comparative analysis and the identification of recurring patterns over the months. This broader temporal perspective contributes to a comprehensive understanding of the distribution of incidents over time.

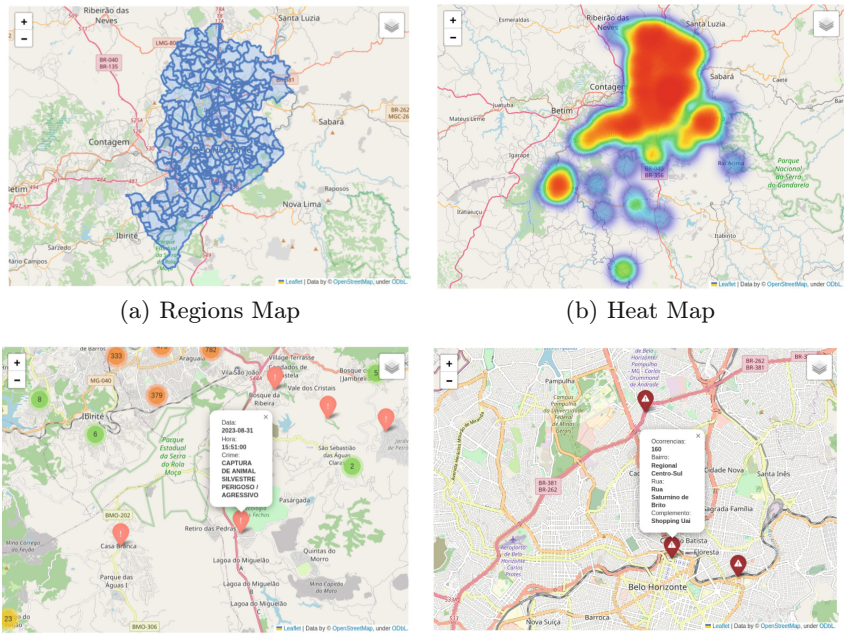
3.5 Layered Event Map

One of the fundamental pillars of our tool is the ability to analyze data based on its georeferenced properties. We adopt a cartographic presentation through maps (item 7 in Fig. 1) to do this. We implemented four distinct layers to provide information about the geography of crime, aiming to provide a rich and complete experience in spatial event analysis.

The first layer demarcates the city's neighborhoods, Fig. 2 a), providing a clear geographic context and facilitating the identification of areas with higher incident rates. The second layer is a heatmap (Fig. 2 b) that highlights areas with the highest density of incidents. It varies color hue and uses the distance between points to provide a clear view of areas with large concentrations of incidents.

The third layer marks the points on the map, Fig. 2 c). Initially, these points are grouped based on distance, but with user interaction, as they are zoomed in, they move apart, allowing for more detailed access to each occurrence. Each point aggregates three essential pieces of information: date, time, and type of event, enriching the analysis with details about the time and classification of the event.

The fourth and final layer highlights critical points, Fig. 2 d), identified through a clustering process based on a maximum distance limit between points. In a simplified way, we define a radius in meters, for which we want to perform the neighborhood grouping, and a minimum value of grouped occurrences to be considered a critical point. From these parameters, a clustering algorithm is applied; in our case, we consider the k-means [7], which will do the grouping and define the critical points highlighted on the map. We add details such as date, time, neighborhood, and complements for these points.



(a) Regions Map (b) Heat Map

(c) Crime Points Map (d) Critical Points Map

Fig. 2. Description of Event Map Layers. In (b), concentrations increase on a scale from blue to red. (Color figure online)

It is important to note that all these layers are integrated into a single map, allowing them to be modified and combined as needed. This approach provides a comprehensive view of the geography of crime, allowing users to customize their analysis and explore different aspects of crime in an efficient and organized way. This flexibility also extends to the ability to add new layers, maintain the organizational structure of the panel, and expand the analytical capabilities of the tool.

3.6 Regression Models

The application of regression models for forecasting has proven to be a promising tool. ARIMA (AutoRegressive Integrated Moving Average) and Facebook (FB) Prophet models are relevant in various areas. Their versatility lies in capturing trends, seasonality, and temporal patterns in various domains, significantly contributing to predictive analytics in different contexts. In our case, it has proven to be a promising tool in helping police forces to assist strategic planning and efficient resource allocation [4]. In order to facilitate the analysis and understanding of the data, as well as to make predictions that generate relevant insights for the user, we created two sections in our tool, one for each model mentioned below.

The ARIMA model is a powerful tool for forecasting time series, composed of three key elements: AR (AutoRegressive), I (Integrated), and MA (Moving Average), represented by the parameters 'p', 'd', and 'q'. The AR term uses past values to predict the next, with 'p' determined by the PACF (Partial Autocorrelation Function). The MA term uses past prediction errors, with 'q' identified by the ACF (Autocorrelation Function). The differentiation order, 'd', ensures stationarity and is defined by tests such as ADF (Augmented Dickey-Fuller) and KPSS (Kwiatkowski-Phillips-Schmidt-Shin). These components are crucial to maximize the effectiveness of ARIMA in forecasting a variety of time series.

On the other hand, FBProphet, developed by Facebook and released in 2017, is a robust and effective tool for analyzing and forecasting time series. It is straightforward and robust, and it handles time series that exhibit seasonal patterns and holidays very well, being especially suitable for datasets with gaps or missing values. It is an additive model with three main components: Trend, which is modeled using a piecewise linear regression model, allowing flexibility in adjusting the trend to the data; Seasonality, modeled using Fourier series, allowing flexible modeling of different seasonal patterns; and Noise, where the model captures the random fluctuations in the data that the trend or seasonality components cannot explain. Finally, the posterior prediction is made via Bayesian inference, where the posterior distribution of the model parameters is estimated rather than just point estimates.

4 Experimental Evaluation

To evaluate the applicability of the crime analysis tool proposed in this article, we instantiated it using criminal data from Brazil, one of the countries with the

highest crime rates in the world, according to the Global Organized Crime Index (See Footnote 1). More specifically, we used the criminal records stored by the Military Police of the State of Minas Gerais (PMMG), one of the most violent states in Brazil. The PMMG is subordinate to the Public Security Secretariat of Minas Gerais (SSPMG), which currently maintains a Research, Development, and Innovation Project with the universities of the authors of this article.

Our evaluation is divided into three complementary parts: the characterization of the data, aimed at the treatment and understanding of the same; the practical application of the tool, allowing a dynamic interaction of different scenarios; and the application and testing of predictive models. The objective is to iterate over all the modules of the tool, previously described in the previous section.

4.1 Data Characterization

The Military Police of Minas Gerais (PMMG) has had a system for recording criminal occurrences and police actions since 2005. The system has over 40 million cases related to the 853 municipalities. Each instance in the database represents a police report and is characterized by 224 attributes that provide a wide range of information. These attributes cover several aspects, such as temporal data, including the date and time of the recording; spatial aspects, such as latitude, longitude, and municipality; and descriptive attributes, such as the specific nature of the incident, the responsible battalion, and the police units involved.

To ensure the effectiveness of subsequent analyses, we treated the data by manually inspecting each attribute. In this inspection, we identified a significant challenge related to the large volume of null values, which would affect the usefulness of the attributes in data analysis algorithms. Another fundamental criterion for attribute selection was identifying a high correlation between them. Eliminating redundant attributes was based on the premise that highly correlated attributes can offer similar information. The exclusion of attributes unique to each instance, such as the unique identifier, was also performed to simplify the dataset's structure. In total, 164 attributes were discarded, providing efficient data cleaning. The set of attributes that we will consider in our analysis are already consolidated in the literature, as observed in the systematic review [6]. These attributes cover spatial, temporal, and categorical factors and are considered essential for defining scenarios through the tool proposed in this work. After data treatment and selection, we analyzed its behavior in different aspects.

Our first analysis, presented in Fig. 3, shows the evolution of the number of records over the years. There is an initial adaptation period in the early years, marked by relatively low numbers of records. This phenomenon can be attributed to implementation and adjustment processes in data collection, reflecting a period of database construction and consolidation. In the last ten years, there has been stability in the volume of annual occurrences, with modest variations and no sharp trends of increase or decrease. Even during the pandemic, we can

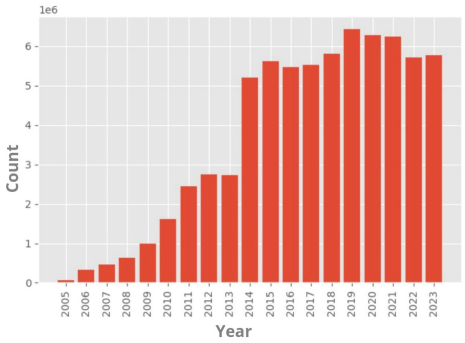
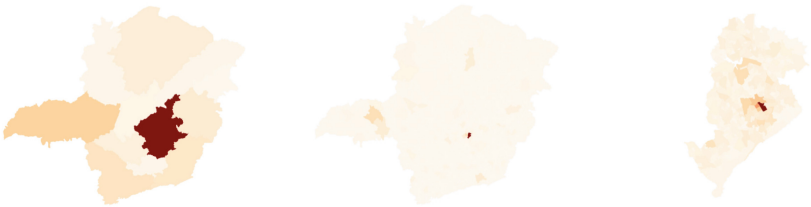


Fig. 3. Occurrence frequency per year in MG

observe that the total number of criminal occurrences in Minas Gerais in Brazil did not decrease.

The importance of analyzing data at different spatial levels is highlighted in the study [1], in which the analyses conducted in Vancouver are divided into three perspectives: census tracts, dissemination areas, and street segments (blocks). These perspectives provide different levels of detail and scope. Census tracts are the most comprehensive perspective, providing information about large geographic areas. Dissemination areas provide an intermediate level of detail, providing information about smaller areas. Street segments provide the most detailed information about tiny areas. Considering the importance of analyzing data at different spatial levels, for our study, we will consider three regions: the most comprehensive, the state of Minas Gerais; the intermediate, the most incident municipality; and the most detailed, the most incident neighborhood.

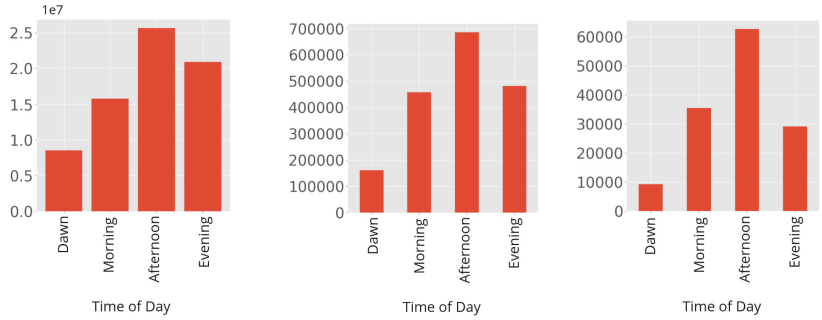


(a) Mesoregion of MG (b) Municipality of MG (c) Neighborhood of BH

Fig. 4. Temporal distributions of occurrences in MG and BH by color intensity, ranging from light to dark.

To identify the regions to be analyzed, we performed the spatial distribution of occurrences, as illustrated in Fig. 4. In Fig. 4 (a), we can see a large concentration of crimes in the Metropolitan Region of Belo Horizonte (RMBH), which

houses the state capital, Belo Horizonte (BH). This city has the highest number of occurrences, as shown in Fig. 4 (b). When we analyze the city of BH, we can see a concentration of occurrences in the Seção Urbana Primeira (SUP), as shown in Fig. 4 (c), a densely populated region characterized by large commercial centers. In addition, the SUP has the most significant number of tourist attractions compared to other neighborhoods.



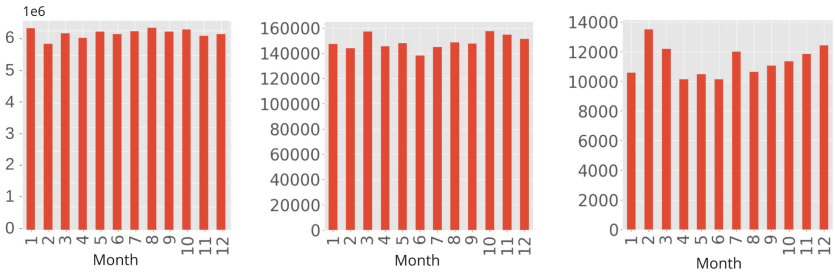
(a) MG (b) BH (c) SUP

Fig. 5. Distributions of occurrences by time of day in MG, BH, and the SUP

In our second analysis, we evaluated the distribution of occurrences throughout the day, considering the same levels of vision established previously. As shown in Fig. 5, the behaviors are similar in the data. We can see a higher crime rate in the afternoon shift and a lower rate in the early morning, regardless of the spatial level analyzed.

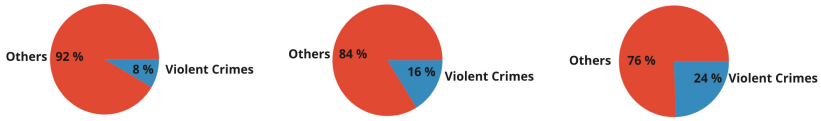
In our third analysis, we present the total number of occurrences each month for each spatial level established, which reveals distinct behaviors, as shown in Fig. 6. Figures 6(a) and (b) show that considering the state and the city as a whole, there is a balance of occurrences during the months. However, when considering smaller areas, such as neighborhoods, Fig. 6(c), there is a higher rate in February, March, July, and December. This trend may be related to the events related to each month. In the first two months, we have the occurrence of Carnival, a popular festival in Brazil, which will be analyzed separately in the case study. In July and December, we have the school vacation period. Both periods are marked by the concentration of many people celebrating or walking around the major commercial centers. Thus, we observe that smaller and more populous regions tend to be more sensitive to such events.

In our fourth analysis, we compare the percentage of violent crimes to other occurrences, Fig. 7. We observe a difference in behavior with a growing increase in the representation of violent crimes as we move from the most comprehensive region to the most specific, representing a higher concentration of crimes in smaller areas.



(a) MG (b) BH (c) SUP

Fig. 6. Distributions of occurrences by month in MG, BH, and SUP.



(a) MG (b) BH (c) SUP

Fig. 7. Violent crimes (blue) and other crimes (red) in MG, BH, and SUP. (Color figure online)

4.2 Case Study

After the characterization of the data, in this section, we aim to present case studies that demonstrate the tool's applicability. In our first case study, we chose to analyze the Brazilian Carnival. Carnival is one of the largest popular festivals in Brazil, attracting millions of people to the streets annually in February and/or March. The event moves billions of reais and is recognized for its joy, cultural diversity, and grandeur. However, this grandeur is also a factor that contributes to the increase in crime rates. Criminals take advantage of the crowds to commit crimes, such as theft, robbery, and even murder. This reality worries the authorities, who seek strategies to combat crime during the event.

To assist in decision-making in this context, we will demonstrate the application of our tool that can be used to understand the behavior of the data during this period. The tool allows us to identify patterns and trends, which can be used to define priority points for allocating security resources, such as patrol schedules and route definition. In this study, we will analyze the data for theft related to the 2023 Carnival, specifically in the district *Seção Urbana Primeira*, in Belo Horizonte, Minas Gerais. This region was shown to be a region of high crime concentration, as illustrated in Fig. 4(c). The official Carnival period in the region lasted from February 4 to 26. All filtering, such as date range, crime type, and location, are defined interactively and simplified through the filters in the tool, as discussed in Sect. 3.1.

First, to highlight the impact of theft rates on this festive event, we can observe Fig. 8, which presents the trend module of the tool encompassing the

first quarter of 2023. In this graph, it is evident that there was a significant increase in crime in February. Another point to note is the similarity of the occurrences in January and March, periods that precede and follow Carnival, highlighting that in February when Carnival took place, the behavior is quite atypical and requires a differentiated strategy by the public security agents.

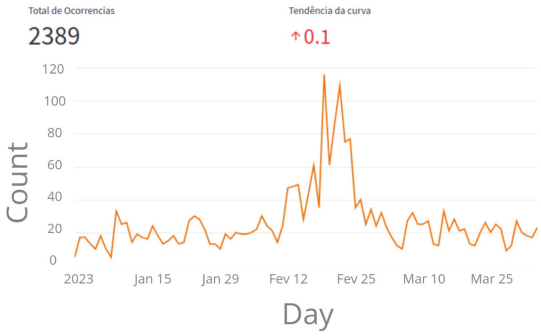
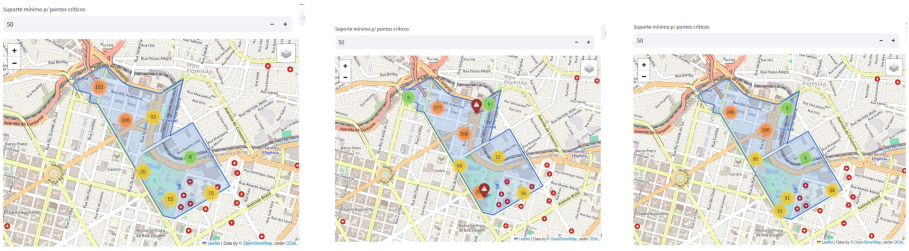


Fig. 8. Tool trend analysis module for the first quarter of 2023

By analyzing the map using the layers of neighborhoods, occurrence points, and critical points, it is possible to contrast these data with the location of carnival blocks, which indicate large concentrations of people and are generally spread throughout the city. Figure 9 illustrates the location of the blocks through markers, and Fig. 10 represents three distinct moments: before, during, and after Carnival.



Fig. 9. Locations of the 2023 Carnival blocks, SUP neighborhood demarcated



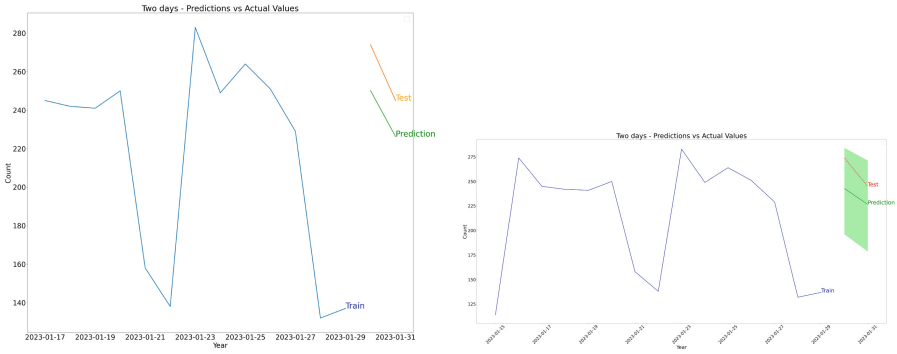
(a) Pre-Carnival (b) Carnival (c) Post-Carnival

Fig. 10. Comparison of the spatial distribution of occurrences in the SUP neighborhood, before, during, and after the 2023 Carnival

A significant increase in crime is noted in the downtown region during Carnival, Fig. 10 (b), highlighting the presence of critical points during this period. These critical points are highly correlated with the location of the blocks, thus highlighting areas of greater danger that should receive greater attention, along with police reinforcement. Another point to note is that, as with the number of occurrences, crime points before and after Carnival, Fig. 10(a and c), tend to maintain a pattern. This reinforces the importance of using the tool not only during special events but also in general and daily use to discover insights about crime incidence. The tool allows resources to be distributed and patrols to be scaled more efficiently in places that demand attention, all based on filters that can define multiple scenarios for analysis.

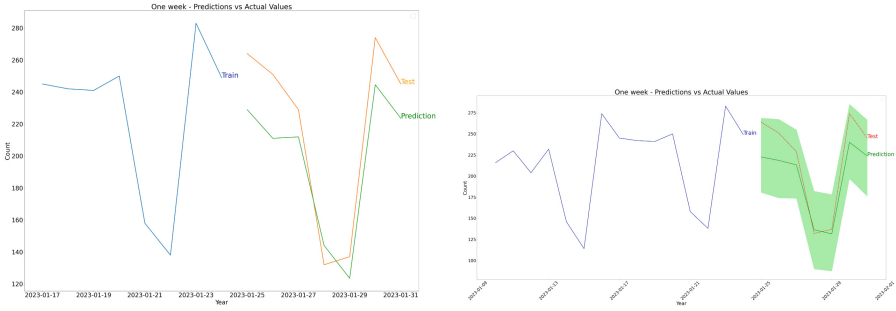
As discussed in Sect. 3.6, in addition to descriptive analyses, which allow the identification of patterns and trends, the tool also has predictive modules that provide valuable information for public safety. Based on the study of [13], these modules do not generate predictions of where and when the next crime will be committed, but rather the relative risk level that a crime will be committed at a given time and location. The tool allows users to define the "where" and "when" questions through filters. The crime level is provided through regression models. This functionality allows public safety to identify with greater precision the areas and periods of greatest risk, which can assist in the planning of preventive actions and the allocation of resources more efficiently.

To illustrate the functioning of this module, we tested and evaluated the two algorithms implemented by the tool (ARIMA and Prophet), considering two distinct granularities for prediction: day and week. First, we selected the data from 01/01/2022 to 01/29/2023, i.e., one year and 29 days, related to theft to perform the training. From the generated model, we predicted crimes for the next two days (01/30/2023 to 01/31/2023), contrasting with the actual occurrences in this period. At the second granularity, we performed the training with data from 01/01/2022 to 01/24/2023. However, the model was evaluated with predictions for the entire following week (01/25/2023 to 01/31/2023). In Figs. 11 and 12, we present the results of this analysis for two days and a week, respectively, considering the two algorithms implemented in the tool.



(a) ARIMA (b) Prophet

Fig. 11. Two-day prediction. In blue, is the last training interval, in green is the prediction (in light green, the confidence interval), and in orange/red, is the actual value (Color figure online)



(a) ARIMA (b) Prophet

Fig. 12. A week-long prediction. The last training interval is in blue, the prediction is in green (the confidence interval is in light green), and the actual value is in orange/red. (Color figure online)

We can see that both models performed well in the prediction task, with values close to the real ones for both granularities. When evaluating the performance of the two models, we observed promising results. Both models performed satisfactorily, presenting predicted values close to the actual data. These studies provide valuable insights into how to quickly apply these models in similar contexts, highlighting both their advantages and specific limitations when predicting crimes. The integration of these models provides a comprehensive approach to predictive analytics, contributing more effectively to the improvement of public safety.

5 Conclusions

In this work, we present a tool for criminal data analysis that is able to characterize the data and transform it into information that helps develop assertive crime-fighting and prevention tactics. Our tool understands the dynamics of crime occurrence from two different perspectives: temporal and spatial, considering two types of analysis: descriptive and predictive, presenting a simple and intuitive interface to be used by public agents responsible for security. To assess the applicability of the criminal analysis tool proposed in this article, we instantiated it using criminal data from Brazil, one of the countries with the highest crime rates in the world, using the criminal records stored by the Military Police of the State of Minas Gerais (PMMG). We perform two descriptive and predictive analyses, considering data from Belo Horizonte, the capital of Minas Gerais. In the descriptive analysis, we compared the crime map before, during, and after Carnival, the world's most famous Brazilian festival, demonstrating a significant change in criminal behavior during the festival, with an increase of crimes concentrated in specific tourist points, which can help to allocate public security agents more efficiently. In predictive analysis, we trained and evaluated two crime prediction models that present predicted values close to the actual data, consequently allowing public security to develop preventive actions more assertively. In sum, we demonstrate that the proposed tool offers an approach with complete descriptive and predictive analyses, contributing more effectively to improving public safety.

Our tool is not limited to static analysis. Its adaptability to future models and the ability to integrate new layers of information emphasizes the dynamic and evolutionary nature of the system. This flexibility ensures the tool can evolve in response to changing criminal dynamics, maintaining its relevance over time. In summary, this work highlights the achievements made in crime data analysis and points to untapped potential and ongoing possibilities for improvement. The strategic partnership between research and security institutions is essential to drive innovation and face complex challenges related to public security. In this scenario, the developed tool emerges as a valuable contribution, offering valuable insights and a model for future approaches to crime management.

References

1. Andresen, M.A., Malleson, N.: Testing the stability of crime patterns: implications for theory and policy. J. Res. Crime Delinquency **48**(1), 58–82 (2011)
2. Bappee, F.K., Soares Júnior, A., Matwin, S.: Predicting crime using spatial features. In: Bagheri, E., Cheung, J.C.K. (eds.) Canadian AI 2018. LNCS (LNAI), vol. 10832, pp. 367–373. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-89656-4_42
3. Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., Pentland, A.: Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of 16th International Conference on Multimodal Interaction (2014)
4. Borowik, G., Wawrzyniak, Z.M., Cichosz, P.: Time series analysis for crime forecasting. In: 26th International Conference on Systems Engineering (2018)

5. Catlett, C., Cesario, E., Talia, D., Vinci, A.: A data-driven approach for spatio-temporal crime predictions in smart cities. In: 2018 IEEE International Conference on Smart Computing (2018)
6. Du, Y., Ding, N.: A systematic review of multi-scale spatio-temporal crime prediction methods. ISPRS Int. J. Geo-Inf. **12**, 209 (2023)
7. Hamerly, G., Elkan, C.: Learning the k in k-means. In: Advances in Neural Information Processing Systems, vol. 16 (2003)
8. Han, J., Pei, J., Tong, H.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2022)
9. Kumar, M., et al.: Forecasting of annual crime rate in India: a case study. In: International Conference on Advances in Computing, Communications and Informatics (2018)
10. Mandalapu, V., Elluri, L., Vyas, P., Roy, N.: Crime prediction using machine learning and deep learning: a systematic review. IEEE Access **11**, 60153–60170 (2023)
11. Marzan, C.S., Baculo, M.J.C., de Dios Bulos, R., Ruiz Jr., C.: Time series analysis and crime pattern forecasting of city crime data. In: Proceedings of the 1st International Conference on Algorithms, Computing and Systems, pp. 113–118 (2017)
12. McCue, C.: Data Mining and Predictive Analysis: Intelligence Gathering and Crime Analysis. Butterworth-Heinemann, Boston (2014)
13. Perry, W.L.: Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations. Rand Corporation, Washington, DC (2013)
14. Pina-Sánchez, J., Buil-Gil, D., Brunton-Smith, I., Cernat, A.: The impact of measurement error in regression models using police recorded crime rates. J. Quant. Criminol. **39**(4), 975–1002 (2023)
15. Prieto Curiel, R.: Weekly crime concentration. J. Quant. Criminol. **39**(1), 97–124 (2023)
16. Walter, R.J., Tillyer, M.S., Acolin, A.: Spatiotemporal crime patterns across six us cities: analyzing stability and change in clusters and outliers. J. Quant. Criminol. **39**(4), 951–974 (2023)
17. Wang, H., Bah, M.J., Hammad, M.: Progress in outlier detection techniques: a survey. IEEE Access **7**, 107964–108000 (2019)
18. Wooditch, A.: The benefits of patrol officers using unallocated time for everyday crime prevention. J. Quant. Criminol. **39**, 161–185 (2021)
19. Xu, R., Wunsch, D.: Survey of clustering algorithms. IEEE Trans. Neural Networks **16**(3), 645–678 (2005)
20. Yan, W.: Toward automatic time-series forecasting using neural networks. IEEE Trans. Neural Networks Learn. Syst. **23**(7), 1028–1039 (2012)
21. Zhang, G.P.: Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing **50**, 159–175 (2003)