CriPAV: Street-Level Crime Patterns Analysis and Visualization

Germain García-Zanabria, Marcos M. Raimundo, Jorge Poco *Member, IEEE*, Marcelo Batista Nery, Cláudio T. Silva *Fellow, IEEE,* Sergio Adorno, Luis Gustavo Nonato *Member, IEEE*

Abstract—Extracting and analyzing crime patterns in big cities is a challenging spatiotemporal problem. The hardness of the problem is linked to two main factors, the sparse nature of the crime activity and its spread in large spatial areas. Sparseness hampers most time series (crime time series) comparison methods from working properly, while the handling of large urban areas tends to render the computational costs of such methods impractical. Visualizing different patterns hidden in crime time series data is another issue in this context, mainly due to the number of patterns that can show up in the time series analysis. In this paper, we present a new methodology to deal with the issues above, enabling the analysis of spatiotemporal crime patterns in a street-level of detail. Our approach is made up of two main components designed to handle the spatial sparsity and spreading of crimes in large areas of the city. The first component relies on a stochastic mechanism from which one can visually analyze probable×intensive crime hotspots. Such analysis reveals important patterns that can not be observed in the typical intensity-based hotspot visualization. The second component builds upon a deep learning mechanism to embed crime time series in Cartesian space. From the embedding, one can identify spatial locations where the crime time series have similar behavior. The two components have been integrated into a web-based analytical tool called CriPAV (Crime Pattern Analysis and Visualization), which enables global as well as a street-level view of crime patterns. Developed in close collaboration with domain experts, CriPAV has been validated through a set of case studies with real crime data in São Paulo - Brazil. The provided experiments and case studies reveal the effectiveness of CriPAV in identifying patterns such as locations where crimes are not intense but highly probable to occur as well as locations that are far apart from each other but bear similar crime patterns.

Index Terms—Crime Data, Spatio-Temporal Data, Visual Analytics, Crime Hotspots, Stochastic Matrix

1 Introduction

Crime hotspot analysis has been one of the main resources employed by public security agencies to plan police patrolling and design preventive actions [**?**]. Hotspot detection methods typically account for the absolute number of crime events in each specific location, neglecting sites where crimes are likely but do not occur in large numbers, mainly when compared to their surroundings. Areas with a high probability of crimes can be more harmful to the community than a place where the crime wave occur in a short period of time [**?**].

The issue above derives from the fact that there is no consensus about a spatial hotspot definition. Distinct definitions can lead to different hotspot configurations. Moreover, hotspot computation strongly depends, among other factors, on the discretization applied to the spatial domain. The most common spatial discretization is a regular grid with cell granularity varying according to the scale on which the analysis should be performed, ranging from dozen meters to large areas covering entire neighborhoods. However, crimes are mostly concentrated in "micro" places that are relatively stable over time [**?**]. Therefore, fine-grained crime analysis demands a level of discretization that should reach the scale of streets, which is difficult to be obtained with regular grids [**?**], as the density and arrangement of streets tend to vary considerably across a city.

Another important aspect related to hotspot analysis is the reasons that lead to the appearance of a hotspot in a given location. According to environmental criminology, the concentration and persistence of crimes in certain locations are not random; that is, they occur due to prevalent characteristics present in those locations [**?**]. Early studies demonstrate that crimes in a particular location are related to demography [**?**], [**?**], population [**?**], [**?**], socioeconomic factors [**?**], [**?**], [**?**], and unemployment [**?**], [**?**], [**?**] in that location. Consequently, changes in those properties over time may impact crime activities, making the temporal analysis of hotspots a fundamental task [**?**], [**?**]. Nevertheless, most hotspot-based analytic tools, mainly the ones in use by security agencies, do not enable resources to identify and group hotspots, according to their behavior over time, hampering the identification of factors that can make crime viable or not.

In collaboration with two sociologists with extensive experience in the study of violence and crime, we have designed a visual analytic tool to scrutinize crime activities in a street-level of detail. Considering that crimes occur in the streets, our approach relies on street networks as the spatial discretization. Specifically, the discretization domain is a network where edges correspond to the streets and nodes represent streets' intersections, thus avoiding the issue of finding a proper level of refinement commonly present in grid-based methods. The proposed methodology relies on mathematical and computational mechanisms to identify hotspots based not only on the intensity of crimes but also on their probability. Moreover, we rely on a deep learning model to embed crime time series in high-dimensional space to make possible the identification of hotspots with similar behavior over time, a task

- *Germain García-Zanabria is with ICMC-USP, São Carlos, Brazil. E-mail: germaingarcia@usp.br*
- *Marcos M. Raimundo and Jorge Poco are with Fundação Getúlio Vargas, Brazil. E-mail: {marcos.raimundo,jorge.poco}@fgv.br*
- *Marcelo Batista Nery is with RIDC -FAPESP and Institute of Advanced Studies – Global Cities Program. E-mail: mbnery@gmail.com*
- *Sergio Adorno is with NEV-CEPID/USP, São Paulo, Brazil. E-mail: sadorno@usp.br*
- *Cláudio Silva is with New York University, USA. E-mail: csilva@nyu.edu*
- *Luis Gustavo Nonato is with ICMC-USP, São Carlos, Brazil and New York University, USA. E-mail: gnonato@icmc.usp.br*

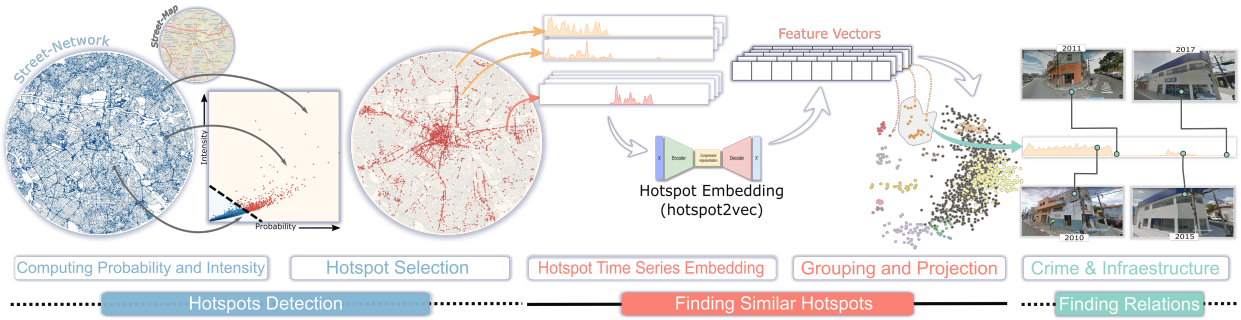


Fig. 1. The proposed street-level crime visualization methodology, *CriPAV*, comprises three main steps. Hotspot Detection: identifying hotspots based on crime intensity and crime probability. Finding Similar Hotspots: hotspot time series embedding (Hotspot2Vec), clustering, and projection into a visual space. Relating Crime & Urban Infrastructure: finding the relation between urban infrastructure and crimes.

difficult to perform with conventional hotspot analysis tools. The proposed methodology, illustrated in Fig. 1, has been assembled in a visualization system called *CriPAV*, which, besides enabling a more general characterization of hotspots, provides visual resources to identify hotspots with similar dynamics over time. As we show in the provided use cases, identifying crime hotspots with similar time behavior helps the understanding of how changes in urban infrastructure impact crime activity over time.

In summary, the main contributions of this work are:

- A new method to identify crime hotspots based not only on the number of crimes but also on the probability of them occurring. By combining the probability and intensity of crimes in a scatter plot, our methodology enables visual identification of locations where crimes are concentrated as well as sites where crimes are frequent but do not necessarily occur in large amounts.
- A method to create groups of hotspots with similar time dynamics despite their spatial location. The method relies on a deep learning autoencoder embedding mechanism called Hotspot2Vec.
- CriPAV, a visualization-assisted analytical tool that integrates a set of linked views to uncover relevant information about hotspots. CriPAV enables mechanisms to identify, explore, and analyze crime hotspots in a street-level of detail, enabling users to explore the possible causes for the observed crime patterns.
- A set of case studies that attest the effectiveness and usefulness of the proposed methodology to reveal interesting phenomena about the dynamics of crime in São Paulo - Brazil.

2 Related Work

In order to better contextualize our methodology, in this section, we discuss previous work related to crime prevention through environmental design, hotspot identification, and time-oriented data visualization.

Crime Prevention Through Environmental Design - CPTED Since socio-demographic, spatial, and temporal attributes are essential information for understanding the dynamics of crimes, several studies focus on analyzing the relation between those factors and individual crime types. Caplan et al. [?] propose Risk Terrain Modeling (RTM) as a methodology to analyze crime based on the dynamic interaction among social, physical, and behavioral factors that occurs in particular places. Some works search for correlations between crime and liquor stores [?], [?], [?], subway stations [?], and parks [?]. Cozen et al. [?] divide physical characteristics related to crime into six categories: surveillance, territoriality, activity support, access control, target hardening, and image management. Image management promotes the maintenance of the environment as a way to contain crime, as vacant, poorly maintained places, damaged buildings, graffiti, and garbage are commonly associated with high crime rates [?].

As discussed above, CPTED methods aim at detecting correlations between characteristics of the environment and crimes, but they can hardly figure out how such correlations evolve. In contrast, our approach enables the analysis of the surroundings of specific locations to assist users in understanding how the environment affects crime over time. Following an approach similar to image management, our methodology relies on a temporal photo gallery to assist in the analysis of how environmental factors and crime patterns evolve.

Hotspot Identification Techniques Hotspot identification techniques aim to identify locations with high risk of crimes, relying on a wide range of methodologies such as Kernel Density Estimation (KDE) [?], [?], point and glyph based methods [?], choropleth mapping [?], local auto-correlation [?], [?], clustering [?], [?], and statistical spatio-temporal analysis [?]. Some of those methodologies have several variants, for example, KDE, which has a family of variants such as Marching Squares KDE (MSKDE) [?], Network-constrained KDE (NKDE) [?], and Network-constrained Getis-Ord Gi* [?].

Hotspot detection methods, mainly based on aggregation schemes, rely on regular grids to perform the computation, facing the issue related to the scale (grid refinement) on which the analysis should be performed. To get around the grid refinement problem, some works propose to compute hotspots based on multiple resolutions [?], aggregating crime events in scales that range from states and cities to census regions and neighborhoods. A growing body of research suggests that crime hotspot should be computed in micro-places [?], as crime activities tend to concentrate in particular street locations [?], [?].

Hotspot detection methods search for regions with a high prevalence of crimes, neglecting locations where crimes are not so intense but highly probable. The methodology presented in this work proposes a new scheme to identify hotspots in a street-level of detail, which accounts for not only the intensity of crimes, but also its frequency over time.

Time data visualization The literature about time-varying visualization methods has extensively been reviewed and organized in different surveys [?], [?], [?], [?]. Aigner et al. [?], [?], for instance, presents a systematic review that categorizes existing approaches into three main groups: time, data, and representation. Bach et al. [?] classify time-varying data visualization techniques based on space-time cube operators. Some approaches unify temporal and spatial analysis in joint visualizations [?], [?]. To better

contextualize our contribution, we focus on techniques that rely on feature extraction or data transformation mechanisms to leverage visualizations of phenomena presented in (spatial) temporal data.

Data transformation is the main resource to uncover hidden patterns in time-dependent data. For instance, Zhang et al. [**?**] propose the VizStruct system, a visualization tool to map time series to a visual space using discrete-time Fourier transform (DTF). More specifically, time-dependent gene data is projected to a two-dimensional visual space such that the projected positions reflect the relation between gene structures in the original data set. Woodring et al. [**?**] rely on wavelet transformation to represent and visualize time series in a multi-resolution manner. Lekschas et al. [**?**] propose a feature-based technique for interactive visual pattern search in sequential data such as time series. For that, they used a convolutional autoencoder for unsupervised representation, active learning, and interactive feedback-driven adjustment.

In the context of crime analysis, Garcia-Zanabria et al. [**?**], [**?**] propose systems with multiple linked views to represent and visualize the dynamics of crime over time. Malik et al. [**?**] propose the Visual Analytics Law Enforcement Toolkit (VALET), which enables linked views such as calendar and line charts to assist users understanding the time evolution of crimes. Godwin et al. [**?**] propose a cumulative temporal view as part of the HotSketch dashboard system. Most of the methods described above rely on aggregation schemes (periods of the day, days of the week, months, etc.) to visualize time patterns. However, aggregation works as a low-pass filter, hidden patterns with small magnitude and smoothing out abrupt changes.

In contrast to the methods described above, our approach relies on a machine learning model to automatically embed time series to a Cartesian space, rendering the comparison of crime time series easier and more accurate. Moreover, the embedding scheme identifies groups of time series of crimes with similar behavior. Thus, making it possible to uncover patterns of crimes even when they occur in locations far apart from each other, a trait not present in most of the methods described above.

3 Challenges in Crime Analysis and Analytical Tasks

This research is mainly focused on addressing the spatio-temporal aspects of crime analysis. To understand our study's motivation, we first present the challenges that experts face on this subject, then thoroughly define nomenclature and finally present the analytical tasks that guided our system's design.

3.1 Challenges

A long-term collaboration (almost four years) with two sociologists with a lifelong carrier in the study of violence and crime enabled us to understand the particularities of crime events that, without proper care, might hurt or undermine the analysis of crime. We had several meetings, including some seminars, with the experts and their students to discuss the main issues they faced to perform their analysis, which we summarize in the following.

Crime events occur in time and space, so it is necessary to aggregate crime events in time and space. Spatial aggregation can be made in census units, police districts, or other units of interest, such as cells of a regular grid, which is one of the most popular choices. Top image in Fig. 2 illustrates a KDE-based hotspot visualization in a regular grid discretization. Notice that the hotspot

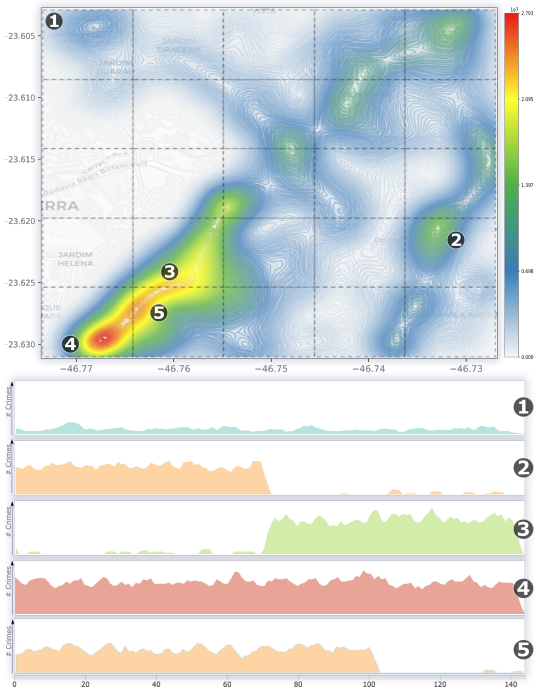


Fig. 2. Challenges in crime analysis. (Top) a heatmap of crime occurrences in São Paulo is aggregated by a simple squared grid. (Bottom) a time series of crime occurrences in different locations. Both representations show the challenges of properly identifying critical regions to study criminality and time patterns that emerge with crime occurrences.

identification depends on the level of refinement of the grid. For instance, in Fig. 2, if we consider the dashed grid as the basis for the discretization, the whole-cell marked with the number 4 would be considered a hotspot, even though crimes are concentrated in a small portion of the cell. Finding a proper resolution to define the grid cell size is not an easy task, and an inappropriate choice can hamper the analysis, leading to inaccurate conclusions.

Another issue when computing hotspots is that locations with a high concentration of crimes tend to be highlighted as the important hotspots, dimming sites where crimes are frequent but do not occur in high volume. For instance, in Fig. 2, the cells marked as $2,3,4$, and 5 could be identified as hotspots, as they present a high number of crime events (see the associated time series on the bottom in Fig. 2. In contrast, the region marked as 1 in Fig. 2 would hardly be identified as a hotspot since the volume of crimes in that region is not comparable to those marked as $2,3,4$, and 5. However, as shown in the corresponding time series (first time series on the bottom), crimes are quite frequent in the region 1, that is, there are crime events registered during the whole period, thus making region 1 worth of attention. Therefore, it is important to define hotspots taking into account not only to volume of crimes but also the frequency they happen in each particular location.

The temporal behavior of crime events in each location is also an important phenomenon to be analyzed. Specifically, the dynamic of crimes in particular locations can be associated with socioeconomic and urban factors; thus, identifying such temporal patterns can assist in understanding the interplay between crime and those factors. The time series on the bottom of Fig. 2 illustrate different temporal crime patterns related to locations where crimes were intense for a while and then vanish (region 2 and 5), locations that face a sudden increase in crime activity (region 3), and regions where crime are fairly frequent, changing only on the intensity (regions 1 and 4). Extracting the temporal crime patterns

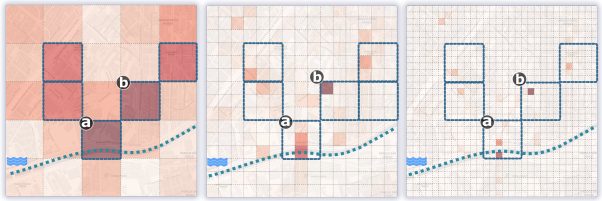


Fig. 3. Grid thematic mapping with different resolution parameters (one of the most common hotspots mapping techniques): (left) 5×5, (center) 15×15, and (right) 30×30 cells.

automatically, grouping them according to their similarity, and enabling analytical tools to assist the experts in their investigation are of paramount importance.

In summary, Fig. 2 illustrates the main challenges we faced when designing and developing the proposed tool, namely, a proper spatial discretization, a flexible mechanism to hotspot detection that accounts for volume as well as the frequency of crimes, and the identification of locations with similar temporal crime behavior. Moreover, designing a visual analytic tool that enables a versatile exploration of hotspots with similar behavior was another challenge we faced during the development of CriPAV. Before detailing how we tackled the issues above, we settle the nomenclature and clearly define the analytical that guided our tool's design.

3.2 Nomenclature

The following terminology will be employed throughout the manuscript.

Anchor Point is the smallest unit of study in a street network. In our context, anchor points are the intersections of streets. Fig. 4-right shows an example of a region modeled as a network, where the black nodes are the anchor points.

Time unit is the time scale that the temporal data is aggregated, for example, weekly or monthly.

Crime time series is the temporal evolution of crimes in each particular location. More specifically, it is a time series discretized in time units associated with each anchor point representing the dynamics of crimes over time.

Crime Type is the type of crime activity. In this work, we focus on three categories: passerby, commercial establishment, and vehicle robbery.

Crime Intensity is the number of crimes in each anchor point in a given time unit. The intensity can be aggregated in a time unit interval.

Hotspots are anchor points with relevant crime activity. In our context, they are anchor points where crime events are of high intensity, high probability, or both.

Hotspot temporal groups are groups of anchor points whose time series share similar behavior. This concept enables the analysis over time, making it possible to identify groups of similar hotspots.

Crime Pattern accounts for the prevalence of criminal events in a given location over time. In our context, a crime pattern in an anchor point refers to the temporal behavior of the corresponding crime time series.

3.3 Analytical Tasks

After understanding the challenges that the experts have faced to accomplish their studies, we compiled the tasks that the proposed analytical tool must enable to tackle those challenges.

T1. Analyze street-level crime hotspots Define and depict crime hotspots on a street level of detail, aiming to identify locations with similar crime patterns.

T2. Probability $\times$ Intensity crime hotspots. Enable a mechanism capable of identifying hotspots based not only on the absolute number of crimes but also on the probability of their occurrence.

T3. Crime Patterns Analysis Enable the analysis of crime patterns in a particular hotspot or a group of hotspots.

T4. Analyze the surroundings of a hotspot Scrutinize urban characteristics around a hotspot.

T5. Group similar hotspots. Group hotspots according to the similarity of their patterns to visualize the spatial distribution of similar hotspots and possible causes for the observed pattern.

T6. Compare hotspots' groups. Support the comparison of hotspot groups according, making it possible to analyze the patterns of different groups and the dispersion of hotspots within a group.

The proposed visual analytic tool, CriPAV, has been designed to integrate several linked views devoted to making the tasks above doable. The working flow underlying CriPAV consists of three major components (see Fig. 1): i) hotspots detection, ii) grouping of hotspots with similar crime patterns, and iii) analysis of the surroundings of hotspots. Before detailing each component implemented in CriPAV, we motivate our decision of selecting the street map as the basis for the spatial discretization.

4 Street-level Domain Discretization

Establishing a proper spatial discretization is crucial to define the crime patterns. As already discussed, most techniques rely on regular grids where each cell can cover dozens or hundreds of square meters, thus impacting the aggregation of crime events. Fig. 3 shows examples of grid discretization using three different levels of refinement. It could be noted that the grid size impacts not only the precise location of crime events but also on which cell is highlighted as a hotspot. For instance, in the courser resolution (Fig. 3-left) cells that are neighbors of (a) and (b) are also pointed out as hotspots, even though crimes are mostly concentrated in (a) and (b). Moreover, identifying the factors that might be generating the crime opportunity is also not straightforward when the level of refinement is not properly chosen. An example is depicted in Fig. 3, where the high intensity of crimes in (a) might be associated with the river that crosses the region, a hypothesis that can only be raised after refining the grid (Fig. 3-middle and right). Also, the use of regular grids can limit the temporal analysis of crimes. Suppose that crime events are continuously registered in a street corner during a period and, after a while, it moves to a nearby corner. In a grid representation, such a temporal behavior can hardly be caught if both corners lie on the same grid cell. Therefore, since urban crimes tend to take place in micro-places such as street segments and corners, accomplishing a spatial discretization in a street-level of detail is a meaningful choice.

Given the issues related to regular grids and recent studies, crimes rarely concentrate on regions larger than a street segment or intersection [**?**], [**?**], [**?**], [**?**], we decided to adopt street corner as the spatial 'discretization unit.' This decision has been further supported by the domain experts that have collaborated with this project. According to them, police officers tend to report the crime location as the intersections [**?**], as the exact location of a crime is rarely known precisely. Moreover, studies suggest that criminals

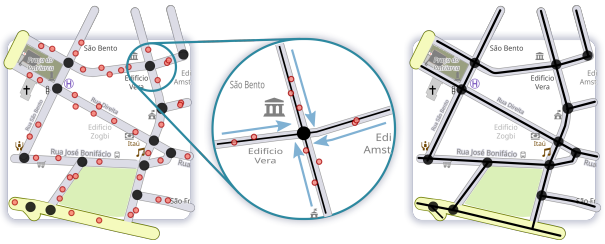


Fig. 4. Data modeling process. (left) Region of study with anchor points (black dots) and crime events (red dots). (center) Approximation of each crime event to the nearest anchor point. (right) Final street network, each anchor point contains a temporal series.

tend to act on street corners [**?**], [**?**] since there is a tendency of the population to meet and socialize on street corners [**?**].

Specifically, we model the spatial domain as a graph derived from the street network map, where street intersections are the nodes (*i.e.*, anchor points) and street segments correspond to edges of the graph. Crimes are spatially aggregated in the graph nodes, as illustrated in Fig. 4. On the left, in Fig. 4 we show a region of interest with anchor points (black points) and crime events (red dots). Each crime event is aggregated to the nearest anchor point, as illustrated in the middle image of Fig. 4, resulting in a network with crime events associated the each node (Fig. 4-right). Crime events are then aggregated temporally according to the time discretization (weeks or months) in each node, given rising to each anchor point's crime time series.

5 Hotspots Detection

As illustrated on the left of Fig. 1, hotspot identification is a primary component of CriPAV. Hotspots are visually defined from a 'Probability $\times$ Intensity' scatter plot, where each dot corresponds to an anchor point. The intensity axis of each anchor point is the temporally aggregated number of crime events in the anchor point divide by the maximum number of crime events among all anchor points. The computation of how likely crimes are in each anchor point is more intricate, and it will be detailed in the following subsection.

5.1 Probability of Crimes

The probability of crimes in each particular anchor point is computed based on the stationary state of a stochastic matrix built from the crime time series. In order to better guide the reader, we need to settle some mathematical concepts.

Probability vector. A distribution vector p is called a *probability vector* if all its elements are non-negative real numbers whose sum is equal 1, *i.e.*, $p_i \geq 0, \sum_{i=1}^{n} p_i = 1$.

Stochastic Matrix. An $n \times n$ square matrix $P = (P_{ij})$ is called *stochastic* if each row (or column) is a probability vector. Stochastic matrices are used to describe the transitions of a Markov chain [**?**] where each entry P_{ij} is the probability of changing from state i to state j.

Under mild conditions, it can be shown that a stochastic matrix has an eigenvalue equal to 1 whose corresponding (left) eigenvector is a probability vector called the stationary vector of the stochastic matrix P. In mathematical terms, the stationary vector is given by the solution of the following equation:

$$\pi P = \pi, \quad \sum_{i=1}^{n} \pi_i = 1 \tag{1}$$

In our context, the probability of crimes in each anchor point is given by the stationary vector of a stochastic matrix built from the crime time series in each anchor point. The construction of such a stationary matrix is detailed in the following.

Computing the Stochastic Matrix Suppose a spatial discretization (street network) with n anchor points $V = \{\tau_1, \tau_2, \ldots, \tau_n\}$ each associated to a time series $T = \{t_1, t_2, \ldots, t_m\}$ describing crime events aggregated into m time instants. We can define a function $f : V \times T \to \mathbb{R}$ that associates the number of crime events $f(\tau_i, t_j)$ in the anchor point τ_i in the time slice t_j. We denote by D the $n \times m$ matrix where each entry D_{ij} corresponds to $f(\tau_i, t_j)$. From $f(\tau_i, t_j)$ we define an occurrence matrix $\widehat{D}$ where $\widehat{D}_{ij} = 1$ if $f(\tau_i, t_j) > 0$ and $\widehat{D}_{ij} = 0$ if $f(\tau_i, t_j) = 0$. $\widehat{D}$ is a binary matrix where each entry $\widehat{D}_{ij}$ indicates whether crimes took place in the anchor point τ_i in the time slice t_j.

From $\widehat{D}$ we define the $n \times n$ co-occurrence matrix $\widehat{P}$:

$$\widehat{P} = \widehat{D} \cdot \widehat{D}^T \tag{2}$$

Each entry $\hat{P}_{ij}$ of $\widehat{P}$ corresponds to the number of times that the anchor points τ_i and τ_j faced crime events in the same time slice, that is, the number of times that crimes took place simultaneously in τ_i and τ_j. A large value of $\hat{P}_{ij}$ indicates that τ_i and τ_j present similar crime activity over time. Dividing each row of $\widehat{P}$ by the sum of its values, we end up with a stochastic matrix P, that is, $P_{ij} = \hat{P}_{ij} / \sum_{k=1}^{n} \hat{P}_{ik}$. The entry P_{ij} corresponds to the probability $Pr(\tau_i, \tau_j)$ of a crimes take place simultaneously in τ_i and τ_j.

The reasoning behind the construction of the stochastic matrix P is that certain crime types are seasonal, occurring concurrently in different city locations depending on the day of the week, the fortnight of the month, and the month of the year. Matrix P, as defined above, captures such a seasonality, being able to point as likely of crimes anchor points where crime activity is not intense, but occurs concurrently with other anchor points.

Given the stochastic matrix P, the *probability* of crime occurrence in each anchor point is given by the stationary vector π of P. That is, the probability of a crime event to occur in τ_i is the value in the i-th entry in π.

5.2 Selecting Hotspots

The *probability* and *intensity* values summarizing crime activities in each anchor point enable the use of a *Probability vs. Intensity* scatter plot to visually identify anchor points based on their intensity, probability, or both, as illustrated in Fig. 6(a).

In order to filter out relevant anchor points (*i.e.*, high probability and/or high intensity), we use a function $g = [0,1] \times [0,1] \to [0,1]$ that assigns a value to each anchor point, as for example $g(probability, intensity) = ((1-\alpha) * probability + \alpha * intensity)$. The value of α is the weight one wants to give to intensity of probability when filtering the hotspots. The parameter α also controls the slope of straight lines in the *Probability vs. Intensity* plane that corresponds to the level sets of g. To select a certain percentage k of the anchor points as hotspots, we can choose a value of g such that $k\%$ of the anchor points are on the positive side of the straight line. Fig. 6 illustrates the hotspot selection methodology when $\alpha = 1$ and $\alpha = 0.5$, respectively. The dotted line represents the straight line chosen such that 5% of the most relevant anchor points (on the top right) are chosen as hotspots.

We have implemented the linear hotspot selection mechanism as an alternative to the interactive brush-based interactive tool, as the users (domain experts) deemed the linear approach more comfortable to use than a brush-based mechanism.

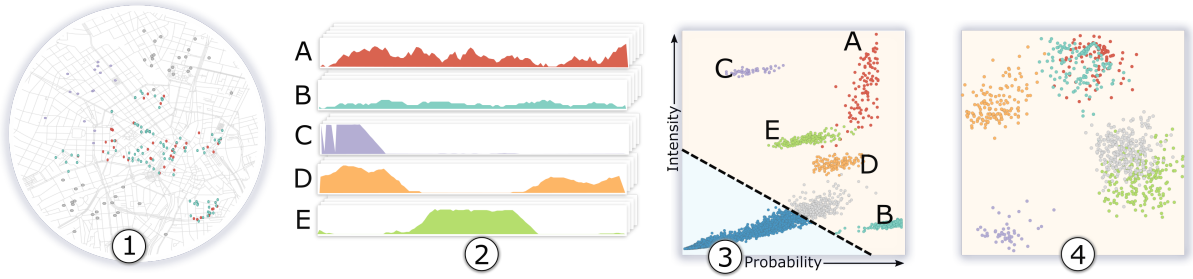
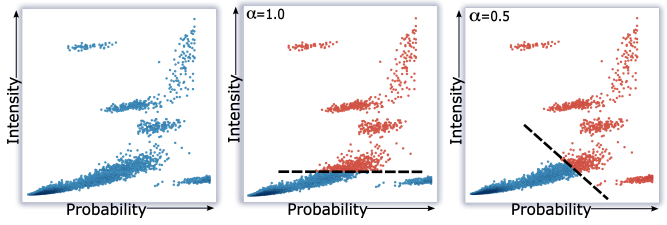


Fig. 5. (1) Region of interest. (2) Time series containing information of interest anchor points in (1). (3) Hotspot filtering in *Probability versus Intensity*. (4) Hotspots clusterization based on Hotspot2Vec.



(a) Prob. & Inte. (b) Filtered Prob. & Inte. with $\alpha = 1$ (c) Filtered Prob. & Inte. with $\alpha = 0.5$

Fig. 6. (a) *Probability* versus *Intensity* scatter plot, (b) filtering based on intensity and probability with $\alpha = 1$, and (c) filtering based on intensity and probability with $\alpha = 0.5$, the red anchor points are the hotspots.

5.3 Validation

We created a synthetic data set with 14,000 anchor points and 120 time-slices representing months over ten years to evaluate the hotspot selection approach. Fig. 5(1) shows the region of study where anchor points are labeled based on five criteria: high intensity & high probability (Fig. 5(2-A)), a normal distribution with mean 1.5 and standard deviation 3; low intensity & high probability (Fig. 5(2-B)), a normal distribution with mean 0.5 and standard deviation of 0.4; high-intensity & low probability (Fig. 5(2-C)) with mean 10 and standard deviation 2 until time-slice 20 and mean 0 and standard deviation of 0.2 for the remaining time. Crimes that only occurs in the beginning and at the end of the time interval (Fig. 5(2-D)) has a normal distribution with mean 2 and standard deviation 0.5 from time-slice 0 to 30 and 80 to 120, with mean 0 and standard deviation 0.3 from 30 to 80 time-slice. Crimes occurring only in the middle of the time interval (Fig. 5(2-E)) has been generated from a normal distribution with a mean of 3 and standard deviation of 0.2 from time slices from 40 to 80, and mean 0 and standard deviation of 0.2 for the remaining time slices. Values for all sites are rounded to the closest integer, and negative values are set to zero. The anchor points with low intensity and low probability are not of interest.

Given the time series on each anchor point, we calculate the *probability* and *intensity* using the methodology described above. Fig. 5(3) illustrates the probability vs. intensity scatter plot, where points on the top-right of the dotted line (settled with $\alpha = 0.5$) correspond to the 5% most relevant anchor points. Notice that points on the top-right correspond to group A (red points), presenting high probability and intensity values. Bottom-right points correspond to group B, anchor points with frequent but not intense crime rates compared to A and C. Finally, top-left points correspond to group C, where crimes are not frequent (low probability) but have a high intensity in a certain period. The remainder points (orange and light green) show a moderate level of probability and intensity. Worth mentioning that by pushing the parameter to $\alpha = 1$ (makes the line horizontal), we miss group B. In this group, crime activity is presented overall period of time but is not so intense as the other ones. Computing the intense hotspots is done in the most existing hotspot analysis tools, which is only a particular case of our approach.

6 Finding Similar Hotspots

Another essential task that our methodology must accomplish is identifying hotspots with similar temporal behavior (see Fig. 1-Finding Similar Hotspots). Finding the temporally similar hotspot means searching for a similar time series, which is a difficult problem. Methods such as Discrete-Time Wrapping can be used to this end but with the price a high computational cost and instability to noise [**?**]. Instead, we opt for a deep learning embedding technique we called Hotspot2Vec.

Hotspot2Vec. We use an autoencoder to map each time-series $TS = \{ts_1, ts_2, \ldots, ts_m\}$ to a feature space. The autoencoder model is trained with a set of time-series $\widehat{TS} = \{\widehat{ts}_1, \widehat{ts}_2, \ldots, \widehat{ts}_m\}$, where $\widehat{ts}_i = 1$ if $ts_i > 0$ and $\widehat{ts}_i = 0$ otherwise, for all $i \in \{1, \ldots, m\}$. The idea is to train the deep learning model to capture the temporal behavior of crimes without considering the intensity of crimes. Therefore, anchor points with crimes happening at the same time interval will be considered similar, no matter the intensity of crimes in each location.

Autoencoder is a well-known neural network model in which the input and output are the same. The middle layer of the network has a bottleneck that creates a compressed representation, aiming to reduce the data's dimensionality.

The autoencoder architecture consists of an encoder with three pairs of 1D convolution and max-pooling layers. The first two convolution layers have 16 kernels of size 3, and the last convolution has a filter of size 3. In our case, the time series contains 144 values, and in each convolution, the time series is reduced by half, reaching a size of 18. We use convolutional layers to capture local information around each step, matching time series that differ by small shifts.

The decoder has a similar architecture to the encoder but changing the max-pooling layers to up-sampling layers. All the convolution layers use the Relu activation function except for the last one, which uses a sigmoid function.

Grouping Similar Hotspots. The encoder's output is used as feature vectors, and a clustering algorithm is applied to group hotspots based on their proximity in the feature space. We choose a hierarchical variant of BDSCAN [**?**] called HDBDSCAN [**?**]. The choice for HDBSCAN is because it can automatically find the number of clusters (as DBSCAN) without tuning several parameters, relieving users of this task, which is essential for domain experts with little training in machine learning.

Projection. Empirical tests showed that reducing the dimensionality of time series to 18 still preserved good properties in capturing

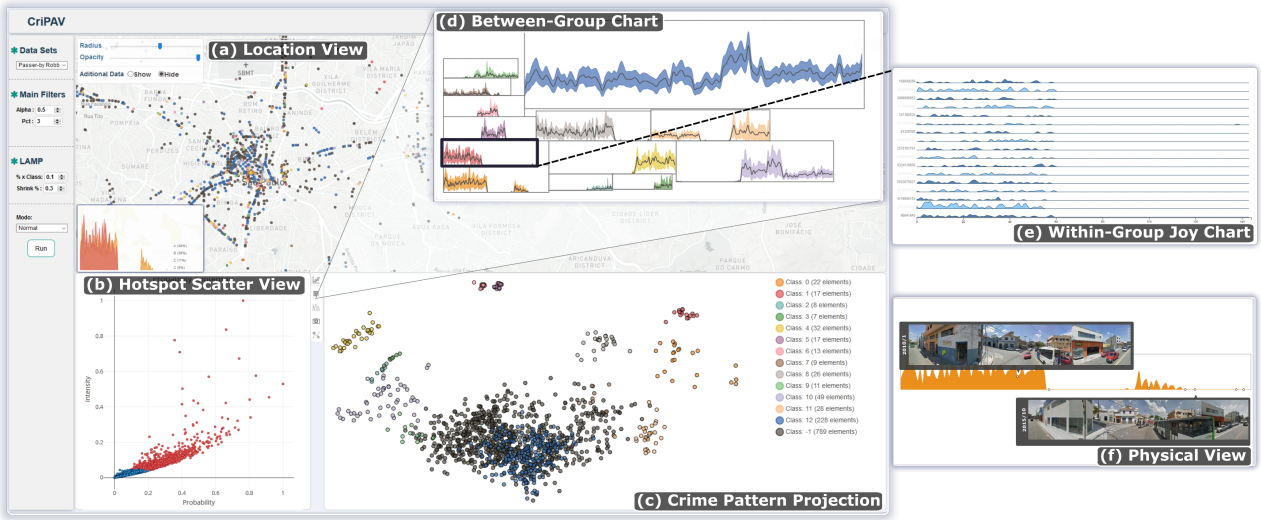


Fig. 7. CriPAV system: Hotspot (b), spatial (a and c), and temporal interactive views (d, e, and f) enabling the exploration of local regions while revealing their criminal patterns over time.

	Sec.	*T1*	*T2*	*T3*	*T4*	*T5*	*T6*
Location view	7.1		✓				
Hotspot scat. view	7.2	✓					
Temporal Patt. Ploj.	7.3				✓		✓
B-Group chart	7.4				✓	✓	✓
W-G Joy Chart	7.5				✓	✓	✓
Physical view	7.6			✓			

TABLE 1
Methodological and visualization properties and their related tools.

their similarity. During our tests, we evaluated four possibilities for the embedding space dimension: 72, 36, 18, and 9 (the temporal series have 144 bins and we reduce the dimension by half in each trial). The dimension of 18 turned out to be more stable and consistent over the runs. To visualize the resulting embedding, we relied on a modified version of the LAMP projection technique [**?**] (github.com/lgnonato/LAMP), which maps the embedded time series to a 2D visual space. LAMP is a computationally efficient projection method that can be tuned to preserve labeled clusters during the mapping [**?**]. Fig. 1-(Finding Similar Hotspots) shows an example of the HDBSCAN clusterization and LAMP projection (Grouping and Projection).

6.1 Validation

We use the same data created to validate the Probability×Intensity hotspot selection to assess the proposed time series grouping mechanism. After selecting the hotspots using the scatter plot, we applied Hotspot2Vec and HDBSCAN to estimate the groups. Fig. 5(4) shows the 2D projection resulting from LAMP. Notice that three time series groups (C, D, and E) out of the five original ones were properly clustered, while groups A and B were, as expected projected on the top of each other, since the embedding does not account for the intensity of crimes, they are considered similar.

7 Visual Components of CriPAV

In collaboration with domain experts, we designed visual components to integrate the hotspot detection and grouping mechanism in a linked view interactive tool. This system, depicted in Fig. 7, provides six components: (a) location view to visualize anchor points and hotspots; (b) scatter plot showing the distribution of anchor points according to their intensity and probability property; (c) hotspots projection view in a two-dimensional space; (d) a visual representation of hotspot groups; (e) visualization of time series from each hotspot group; (f) a photo panel, showing images of urban infrastructure over time for a selected anchor point. We design these visual components in close collaboration with domain experts and their requirements. Each view has been designed to address at least one analytical task described in Sec. 3. Table 1 indicates the relation between the visual widget and the tasks (columns). For instance, *Within-Group Joy Chart* and *Between-Group Chart* account for three tasks. We detail each visual component in the following subsections.

7.1 Location view

This view, depicted in Fig. 7(a), enables the visualization of anchor points' geographical location. Each node's color depends on the group it belongs to, which is computed as described in Sec. 6. The location view is particularly useful to overview the spatial distribution of hotspots and their similarity. Besides, it is possible to show additional information in the background coloring the census units according to a given property, for example, socioeconomic and social vulnerability index (this resource will be exploited in the case studies). Location view provides 2D and 3D visualization of the hotspots, being possible to change visualization properties such as size and opacity of the hotspots. Color opacity and elevation, the latter only available in the 3D view, can be set to correspond to the intensity or probability of the hotspots.

Anchor Point Selection. By a simple clicking in an anchor point, *Physical View* shows the Google Street View photos in the surroundings of the selected hotspots.

7.2 Hotspot Scatter View (Probability vs Intensity)

Illustrated in Fig. 7(b), the hotspot scatter view shows the Probability×Intensity scatter plot. To identify anchor points as hotspots, one can use the linear discriminant filter function described in Sec. 5 by tuning the parameters α and *pct* (percentage of anchor points to be considered hotspots), although an interactive brushing mechanism is also available.

7.3 Crime Pattern Projection

Once hotspots have been selected with *Hotspot Scatter View*, they are grouped according to similarity, and the multidimensional projection is performed to reveal their distribution in the feature space. As shown in Fig. 7(c), hotspots are colored according to their groups in the projection. Moreover, the legend on the right encodes the label and the number of elements in each group.

Group Selection. By clicking in the label of a group, *Location View*, *Hotspot Scatter View*, *Between-Group Chart*, and *Within-Group Joy Chart* are updated to highlight only the data in the selected group, making it easier for users to focus their analysis on the selected group of hotspots.

Filtering. It is possible to select hotspots using a lasso selection. The filtered hotspots are highlighted in the *Location View*, *Hotspot Scatter View*, *Between-Group Chart*, and *Within-Group Joy Chart*. Besides, one can analyze individual time series of selected hotspots using the *Within-Group Joy Chart*.

7.4 Between-Group Chart

To better visualize the intra-cluster patterns, we create a visual representation that summarizes the crime pattern in each group (see Fig. 7(d)). This visualization shows the average time series of each group and the standard deviation of the group's time series. The rectangular glyphs, whose size reflects the number of hotspots in the group, are arranged to keep the most similar groups closer to each other in the layout, following the proximity relation observed in the *Crime Pattern Projection View*. The rectangular glyphs arrangement is computed from an optimization procedure similar to the method described in [**?**]. The Between-Group Chart is useful to understand the crime patterns present in the data.

Group Selection. By clicking in a rectangular glyph, *Location View* shows additional information encoded on the geo-map, such as the socioeconomic and social vulnerability index.

7.5 Within-Group Joy Chart

This view relies on filled line plots (see Fig. 7(e)) to present, in each line of the chart, the crime time series of the hotspots of a specific group. This visualization aims to provide a detailed visualization of the crime pattern in the hotspot group.

Time Series Selection. By clicking in a particular time series, *Physical View* shows the photos of their surroundings.

7.6 Physical View

This view uses *Google Street View* to extract and organize photos of the surroundings of selected hotspots over time (see Fig. 7(f)). Each photo is a collage of many photos extracted during spatial padding. This padding is accomplished for each time slice. *Physical View* helps domain experts to understand the relationship between crime patterns and the urban infrastructure over time.

8 IMPLEMENTATION

CriPAV is a web-based system over a python Flask server. The system's core is divided into data/space modeling, computation of hotspots, and visualization modules. For the data/space modeling, we used OSMnx [**?**] and NetworkX (networkx.github.io) python libraries to extract the street network and to aggregate crime events, respectively. To achieve interactive rates, we assign to each crime record the nearest edge, vertex node. We also perform a street network simplification to remove unnecessary vertices from the street network. The computation of the stochastic matrix, node probability, and intensity is performed using Pandas and Numpy python libraries for hotspot identification. Finally, all visualization resources have been developed based on JavaScript libraries: Deck-GL (deck.gl) for the geo-map representations and choropleth maps; PlotlyJs (plotly.com/javascript) for *Probability*$\times$*Intensity* representation; D3.js (d3js.org/) for the projection scatter-plot, line, and area charts. We have developed an extra component to deal with Google Street View photos and SKImage [**?**] python libraries.

9 CASE STUDIES

This section presents three cases involving real crime data from São Paulo (the city of São Paulo) - Brazil. The case studies show how CriPAV addresses the analytical tasks described in Sec. 3.3 in three different scenarios. The first case study addresses T1, T2, and T5, highlighting the importance of accounting for probability when analyzing hotspots. The second case study focuses on the relation between hotspots and socioeconomic factors, addressing T2 and T4. The third case study aims to show the potential of CriPAV to assist the experts in their search for possible explanations for crime patterns, relating crime patterns to urban infrastructure (T3 and T6). In all case studies, except when explicitly stated, we focus on passerby robbery as the crime type.

All case studies use crime records assembled by domain experts and provided by the Police Department of São Paulo (the largest South American city with around 12 million inhabitants). The data set consists of information on about $1,650,000$ crime incidents recorded from 2006 to 2017. We have worked with three crime types: passerby, commercial establishment, and vehicle robbery. Each record contains the census unit's identification ID where the crime took place, the date and time of the event, type, and geocode location (*i.e.*, latitude and longitude) information.

9.1 Intensity versus Intensity & Probability

The main goal is to use the proposed methodology to analyze the impact of selecting hotspots based only on crime intensity versus select them from the Probability vs. Intensity scatter plot.

Fig. 8 shows the *Hotspot Scatter View* (top-left graphic) where red dots are anchor points selected as hotspots in both the Intensity only and Intensity & Probability. The blue dots correspond to anchor points considered in the Intensity & Probability only, while the green dots are anchor points considered in the Intensity only group. The gray dots are not considered hotspots. The *Within-Group Joy Charts* (top-right) shows crime patterns from the blue and green set of hotspots. The *Location View* (bottom map) shows the geolocation of the blue and green hotspots in a specific region of the city.

Notice that the set of green hotspots left out in the Probability x Intensity selection correspond to crimes that take place in a short period, while the blue hotspots present patterns of crimes spread out over the whole period. An interesting aspect pointed out in the *Location View* (Fig. 8-bottom) is that the green (probable) hotspots tend to show up along main streets, avenues, and highways (dashed lines). In contrast, the blue (intense) hotspots are more spread, also appearing in secondary streets. Domain experts considered this an exciting finding because it shows that crime patterns seem to change according to urban infrastructure in a way they have not observed before. The possibility of identifying those patterns can make public security policies more efficient.

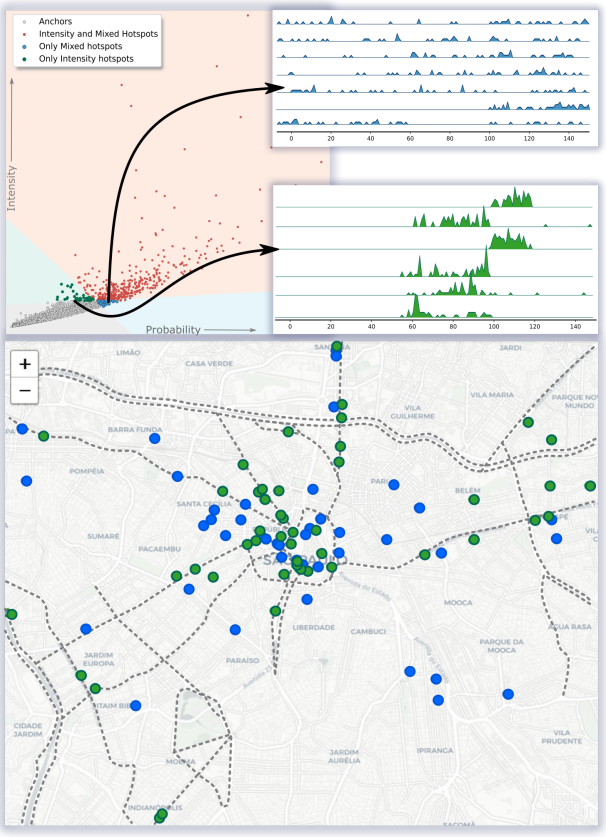


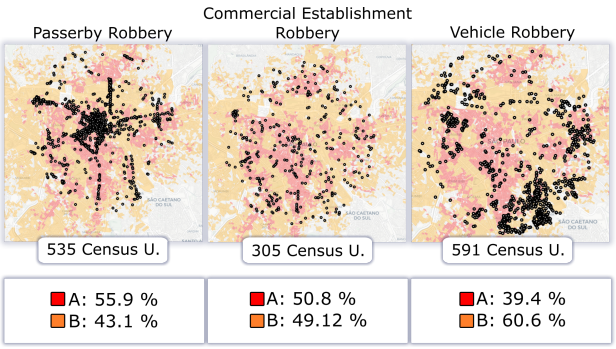
Fig. 8. Analysis of hotspot identification technique. Comparison between missed hotspots with different cut variables. (top-left) Hotspot Scatter View: red dots are hotspots identified by Intensity and Mixed definitions; blue dots are hotspots identified just by Mixed definition; and green dots are points identified only by Intensity Definition. (top-right) Time series of blue and green dots. (bottom) Location View with plotted points.

CriPAV enables the selection of both types of hotspots, and more importantly, it can show the importance of using the probability to characterize hotspots. Therefore, the case study shows that the proposed methodology can found robust hotspots based not only on the Intensity but also on the Probability (T1) while showing the geographical localization of the hotspots (T2). Finally, it is possible to analyze different crime patterns (T5) and their relation with urban characteristics.

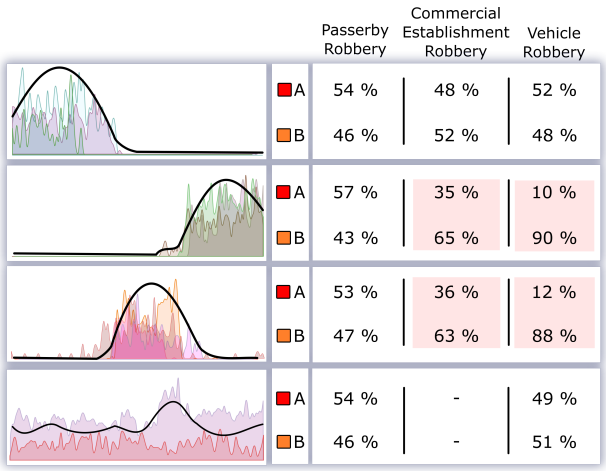
9.2 Understanding Crime Patterns and Socioeconomic Factors

The main goal is this case study is to analyze spatial and temporal crime patterns over the city, relating the patterns with the socioeconomic characteristics of each location.

Specifically, domain experts aimed to address the following questions: (i) how are hotspots concentrated for each crime type? (ii) how are crime patterns spatially distributed?, and (iii) how socioeconomic factors relate to crime hotspots? Domain experts provided socioeconomic indicators that account for population, housing, mobility, and urban expansion in census unit of São Paulo, labeling those units according to eight categories ranging from A (better socioeconomic level) to H (worst level). We selected a region of study with 5,890 census units spread around downtown São Paulo. First, the *location view* was used to visualize the concentration of hotspots. Fig. 9(a) shows the hotspot distribution for each crime type. The colored background corresponds to census units of level A (red) and B (orange). Notice that there



(a) First row: Location View of hotspots distribution of each crime type and number of census units involved. Second row: the relation between hotspots and socioeconomic variables for each crime type.



(b) Relation between *Between-Group Chart* temporal patterns and socioeconomic variables.

Fig. 9. Comparison of Passerby, Commercial Establishment, and Vehicle robberies patterns over the city. (a-first row) Concentration and dispersion tendency of hotspots; (a-second row) Relation between hotspots and socioeconomic variables, and (b) Spatial behavior of hotspots clusters and socioeconomic variables.

are three prominent patterns in terms of hotspot distribution. Passerby robbery are concentrated in the center of the city, Commercial Establishment robbery is spread throughout the city, while Vehicle robbery hotspots are concentrated in the periphery. The socioeconomic indicators are also related to crime types. The bottom legend in Fig. 9(a) shows the percentage of hotspots in each social class. We can see that passerby robbery is more prevalent in class A, commercial establishment robbery is equally distributed, and vehicle robbery takes place mostly in class B. Moreover, passerby, commercial establishment, and vehicle robberies hotspots are concentrated in only a small portion of the census unit, 535 (9%), 305 (5%), and 591 (10%), respectively.

The relation between socioeconomic factors and crime types is even more pronounced when we consider temporal behavior. During the exploration of the three crime types using CriPAV, domain experts perceived four prominent temporal patterns. Relying on *Between-Group Chart*, it is possible the notice the patterns depicted in Fig. 9(b) left, where the leading four patterns are clearly seen, namely: high crime rates in the past, high crime rates in the recent past, recent high crime rates, and crime events spread over the whole time interval. Socioeconomic factors seem not to affect passerby robbery over time. It might happen because this type of crime is positively related to people's flow; it does not matter if in class A or B. In contrast, commercial establishment and vehicle

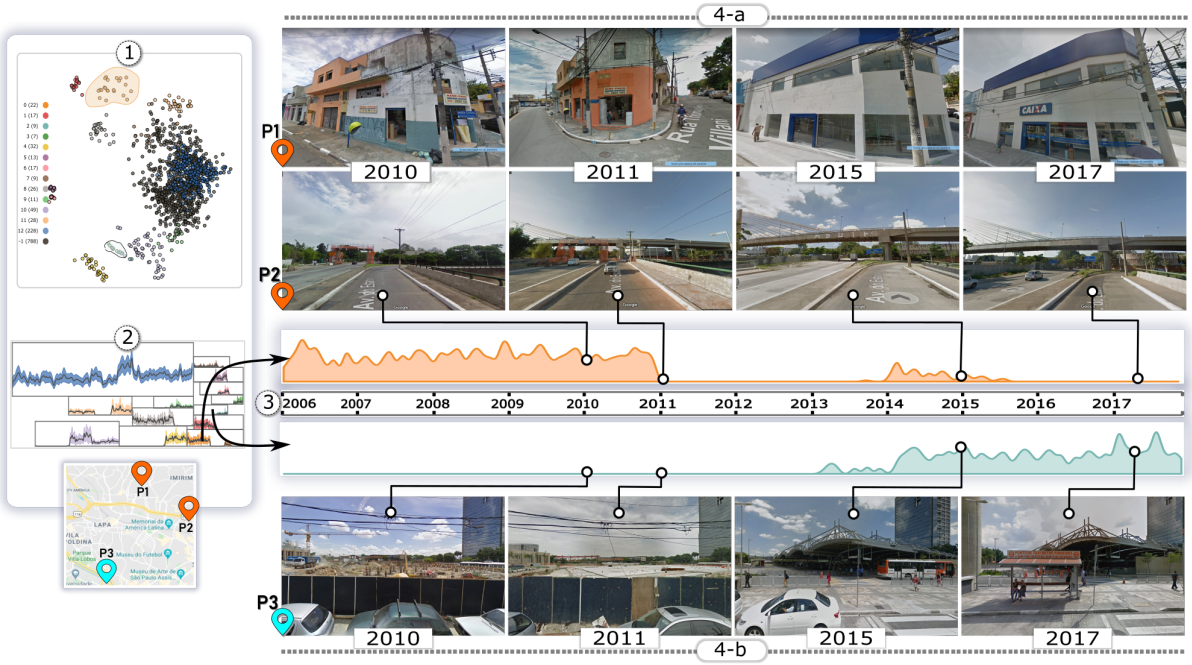


Fig. 10. Hotspots infrastructure comparisons over time. (1) Clusterization and projection of hotspots using Hotspot2Vec, (2) Between-Group Chart showing temporal behavior of each group, (3) time series of the selected group of hotspots (Group 0 and Group 2), and (4) temporal images of points P1, P2, and P3.

robberies present a change over the years, migrating from class A to class B, making class B neighborhoods more vulnerable to those two types of crimes in more recent years. The reason for that shift is an interesting aspect to be further investigated, calling the expert's attention to a phenomenon they have never noticed before.

The discussion above shows that the hotspot detection and grouping functionalities implemented in CriPAV are able to sort out the three analytic questions raised by domain experts. In particular, *Location View* and *Between-Group Chart* turn out to be effective in showing the spatial distribution of the hotspots (T2) and the relation between hotspots and the characteristic of their surroundings (T4).

9.3 Relating Urban Infrastructure and Crime Patterns

This case study assesses the effectiveness of CriPAV in assisting experts in analyzing the relation between crime patterns and urban infrastructure. Such an analysis demands a street-level spatial representation of crime events (T2), a mechanism to investigate physical urban constructions in specific locations (T3), and comparing different hotspots groups (T6).

To start the analysis, we used *Hotspot Scatter View* to select 3 percent of the most important hotspots considering probability and intensity. Fig. 10(1) shows the groups and projection of the selected hotspots. The color encodes the different groups, and the legend on the left represents the number of elements in each group.

The Between-Group Chart (Fig. 10(2)) presents a visualization of the groups and their summary pattern. To perform a detailed analysis, we selected three hotspots denoted as P1, P2, and P3 on the map, as depicted on the left bottom of Fig. 10. P1 is located in the north São Paulo, in an industrial area of the city. P2 is located in the east part of the city, and it is surrounded by recreational places such as sports clubs and Samba schools. P3 is located in a district with high-quality infrastructures in terms of transport, health, culture, and education. It is surrounded by metro stations, a university, and a bus terminal. P1 and P2 belong to the orange group highlighted in Fig. 10(1) top while P3 comes from the blue group on the bottom of Fig. 10(1). Fig. 10(3) shows the average time series of the orange (top) and blue (bottom) groups. These time series were extracted using the *Between-Group Chart*. Notice that the time series from the orange group presents relatively high criminality rates in the early years of the time interval, facing a quick drop in 2011. On the other hand, the blue group's time series shows that the hotspots belonging to that group do not have relevant crime activity until 2013, presenting relatively high crime rates since then.

To raise the hypothesis as to why such patterns happen in those locations, we used *Physical View* to inspect images of the urban infrastructure around those hotspots. Fig. 10(4-a) shows some pictures of the urban infrastructure around P1 over the years. Notice that until 2011 there was a store with a poorly preserved facade, graffiti, trash bags, and broken windows. In 2011, the store was replaced by a bank branch, improving cleanliness and possibly the security in this location. The change might have triggered the radical drop in criminality in P1. Regarding P2 (second row of Fig. 10(4-a)), a bridge was under construction before 2011. Once done, the bridge certainly changed people's flow in that neighborhood, impacting the criminality rate. Performing the same analysis for P3, we observe a different behavior. There was no crime activity in P3 for several years, and after 2013, crime rates increased substantially. Using the *Physical View*, we extract images of the surroundings of P3 over the years. P3 is in front of a bus terminal. The images in Fig. 10(4-b) show that before 2013 the bus terminal was under construction, as shown in the first two photos. The bus terminal was inaugurated in 2013, which coincides with the beginning of rising crime rates in that location. In contrast to P1 and P2, the urban infrastructure might have triggered the crime in that location.

This case study shows that urban infrastructure can certainly impact crime dynamics, thus dictating crimes' emergence and disappearance in particular locations.

10 EVALUATION FROM EXPERTS

We designed an expert evaluation in which each expert performed some tasks using the CriPAV system. After using our methodology and performing experiments, we collected experts' opinions about the method, functionalities, and visual components. In this section, we describe the details of the evaluation and the obtained results.

10.1 Participants

We selected four domain experts from the Center of Study of Violence of São Paulo University (NEV-USP). One of them directly involved in the tool's design; the other three had no participation in the development. The experience of the experts in crime data analysis and the usage of computational tools ranged from 5 to 20 years, being a technology chair, a physical geographer, a sociologist, and a politics researcher. All participants completed three stages in our study: an introduction to CriPAV, a presentations of some applications (case studies), and the accomplishment of some tasks.

10.2 The Evaluation Process

Before setting the tasks, we presented the methodology (including the case studies), the tool, and its visual components to experts. This presentation built confidence and aimed to bring participants to a similar level of knowledge about the instrument, as they are scientists of different areas and domains.

After that, using the second case study as a starting point, we designed tasks that induced the experts to use most of the visual resources available in CriPAV.

Task 1: Participants were asked to extract two hotspots for any crime type and their corresponding photos in the first task. The experts had to select two interesting corners based on their experience and knowledge of the city.

Task 2: In the second task, participants were asked to analyze hotspot's temporal patterns seeking explanations for the temporal crime behavior. This task is subjective because each expert could find different reasons based on their expertise.

Although Tasks 1 and 2 are similar, Task 2 demands a more intensive use of the visual resources. The system was re-started after each task to push further the users to interact with the system. Moreover, in task 2, we asked the expert to explore temporal patterns with different behavior (high crime rates in the past, high crime rates in the recent past, current high crime rates, and crime events spread over the whole time interval).

After the presentation and performing the tasks, participants commented about the methodology, case studies, and CriPAV system. They answered qualitative (QL) and quantitative questions (QT).

The qualitative questions are: *"What is your impression about the proposed hotspot detection technique?" (QL1)*; *"What do you highlight as interesting in case study 2 (QL2) and 3" (QL3)?*; *"If you have already accomplished similar analysis to the ones performed in the case studies, which tool/methodology have you used?"(QL4)*; *"Which are the challenges when performing analysis similar to the ones presented in the case studies?"(QL5)*; *"Do you consider CriPAV a useful tool, why?"(QL6)*; *"In your opinion, which are the most helpful visual components?" (QL7)*; *"Besides the existing visual resources, which other components do you think would be useful to incorporate in CriPAV?" (QL8)*; *"Which are the advantage of CriPAV when compared to other tools you are used to?"(QL9)*; *"Which are the disadvantages?" (QL10)*; *"Please, provide your final comments" (QL11).*

Fig. 11. Summary of quantitative questions.

The quantitative questions are: *"Do you consider case studies 2 (QT1) and 3 (QT2) relevant? (Yes, Partially, No)"*; *"Have you tried to do analysis similar to the ones described in the case studies before? (Yes, No)" (QT3)*; *"Are you currently able to do those analyses with another tool? (Yes, No)" (QT4)*; *"Have you got new knowledge or conclusions from your exploration? (Yes, No)" (QT5)*; *"Do you consider CriPAV a useful tool for crime analysis? (Yes, Partially, No)" (QT6).*

Sec. 10.3 presents the qualitative answers grouped in different topics. Fig. 11 depicts a summary of the quantitative answers, summarized as i) the experts considered attractive the case studies 2 and 3 (QT1 and QT2); ii) half of them tried to do a similar analysis before (QT3), iii) only one expert could accomplish analysis similar to the case studies with another tool (see details in the "Usefulness" paragraph in the following subsection), and iv) all of them have got new knowledge by playing with CriPAV (QT5). Moreover, most of the experts considered CriPAV a helpful tool for crime analysis (QT6).

10.3 Results

We collected the experts' opinions about our hotspot detection technique, usefulness, and usability of our system, described in the following.

Methodology (QL1 and QL11): The probability vs. intensity method for hotspot detection was deemed relevant for the users, being considered strong, promising, and creative. One of the experts commented: *"Using robust mathematical and computational mechanisms helps to identify crime hotspot not only by the most likely dangerous locations but also places that might not receive enough attention from public safety policies, despite frequent criminal activities"*. Regarding the time series clustering mechanism, one of the experts commented: *"Identifying similar temporal behavior of hotspots located in different locations leads to thinking about the city's complexity. At the same time, it guides the investigation of plausible explanations (demographic profile and urban pattern) for violence dynamics, which would be difficult with conventional tools"*. Moreover, *"The identification of specific places with similar criminal behavior is important to apply similar public policies..."*.

Case studies (QL2 and QL3): We collected the experts' impressions on Case Study 2 and 3 to validate our analysis. Case Study 2 took longer to be understood by the experts, but they appreciated the way hotspots are defined and the pattern distribution stratified by socioeconomic variables. *"The match of crime events, temporal behavior, and socioeconomic variables helps a lot understand the complexity of crimes. I emphasize the importance of detecting spatio-temporal patterns and presenting different approaches and explanations for the dynamics of crime records"*. On the other hand, experts were delighted with Case Study 3. Their explanations were in line with our analysis. *"This study is so valid; identifying*

patterns and the possibility of visualizing the place and the changes over time presents itself as a tool that dramatically increases the agility of the analysis and, I believe, the planning of actions from public agencies".

Usefulness (QL4, QL5, and QL9): Domain experts considered the CriPAV a helpful tool for crime analysis, mainly when compared against other tools: "*A similar analysis would demand the use of several GIS-based tools (Terraview, SPRING, Mapinfo, QGIS, etc.) with specific functionalities to assist users in a particular analysis.*" The mentioned GIS-based alternatives are general-purpose tools that can be used for crime analysis. The generality of those tools brings advantages and also weaknesses. For instance, Terraview [**?**] has excellent flexibility to organize information in data sets, and it has the capability of assessing geocoding quality. However, it does not allow a direct street-level analysis of the data, demanding users to implement (or acquire) routines for this purpose. SPRING [**?**] can handle numerical and thematic geo-fields (geo-objects) through the so-called LEGAL (Spatial Language for Algebraic Geoprocessing) map algebra, introducing flexibility to take multiple data sources. However, learning map algebra is not so straightforward, requiring some training from the users. Mapinfo [**?**] is quite helpful to handle vector maps and alphanumeric data, but, similarly to Terraview, street-level analysis can only be performed through dedicated packages. QGIS [**?**] is free software that allows users to view and overlay vector and raster data, being more versatile than ArcGIS [**?**]. However, QGIS also demands training to be used properly. In fact, some experts reported that QGIS can become confusing for non-experinced users, hampering the analytical process.

In contrast to the general-purpose GIS tools, CriPAV has been designed for the specific task of crime analysis in a street-level of detail. Therefore, CriPAV does not demand several hours of training to be used, enabling an intuitive mechanism to assist users in understanding spatio-temporal crime patterns.

Usability (QL6 and QL9): The domain expert involved in the design of CriPAV acknowledged that the system does integrate all the aspects discussed in our meetings. *"The tool is very dynamic and easy to use, capable of presenting very clearly the time series, corresponding photos, and other additional information. For urban administrators and security officers, it offers meaningful information on territorial occupation and crime occurrences".*

Familiarity vs. Preference (QL7): Regarding familiarity with the resources available in CriPAV, we observed that most experts were already used to some visual tools such as *Hotspot Scatter View*, *Location View*, and *Within-Group Joy Chart*. However, although they were not familiar with the *Between-Group Chart*, most of them preferred this visual resource to select hotspots due to the facility identifying patterns.

Scope for future Research (QL8 and QL10): We also investigate resources the experts deem relevant during the user study but are not available in CriPAV yet. We identified that the main requirements are related to additional information to analyze the hotspots, such as the flow of people and information related to violent crimes. It is worth mentioning that CriPAV can naturally incorporate that information once we have access to it.

This multidisciplinary work ended successfully, creating new research fronts in crime analysis. One of the experts commented: *"The results demonstrate the importance of analyzing crime patterns at small scales and have important implications for theoretical development and empirical research".*

11 DISCUSSION AND LIMITATIONS

As detailed in Sec. 3.3, CriPAV was designed to meet the experts' demands. However, there are some limitations and future work that will be addressed in follow-up work.

Automatic crime perception on photos. The *Physical View* relies on google street view to build a temporal photo collage in the surroundings of a particular corner. However, the relation between physical characteristics and crime activity is performed manually. This analysis can be time-consuming and tedious, mainly when many images are available. A solution for this problem is to build upon *Image Emotion Recognition* [**?**] to automatically detect parts of the image related to danger, helping users to quickly focus on those images.

Spatial Discretization. We used the network generated by the Python library *osmnx* as a spatial representation. However, there are other representations, some of which with a more refined level of detail. Enabling the tool to switch between street network representations is essential, and we will address this issue in the future.

Multiple Data Sources and Scenarios. Combine different data sources would be helpful to analyze the whole crime story in each location. CriPAV enables information about infrastructure and socioeconomic variables, but given the increasing number of initiatives to make urban data publicly available, incorporating other sources of information would considerably enrich CriPAV's analytic power. Among the information that could be handled are the number of bars and bus stops near each anchor point, which correlate with certain types of crimes. Moreover, although domain experts are primarily interested in analyzing only three crime types, we could extend the number of crime types. CriPAV is scalable to handle many crime types and a large number of crime instances. The user can choose the type of crime from a list in the main menu. We believe that the additional level of information will open new avenues for interpreting crime patterns.

12 CONCLUSION

In this work, we proposed a visualization-assisted methodology for crime analytic. The proposed tool brings two main technical contributions: a mechanism based on a stochastic matrix to identify hotspots based not only on the intensity but also on the probability of crime events and a technique to find similar hotspots embedding the time series in a cartesian space (Hotspot2Vec). Our system, called CriPAV, turned out to be effective to assist experts to figure out the relation between crime patterns and urban characteristics, revealing phenomena and patterns that were previously unknown by experts in crime analysis.

ACKNOWLEDGMENTS

This work was supported by CNPq-Brazil (grants #302643/2013-3, #303552/2017-4, #301642/2017-63, and #312483/2018-0), CAPES-Brazil (grants #10242771), NEV-CEPID (grants #2013/07923-7) São Paulo Research Foundation (FAPESP)-Brazil (grants #2013/07375-0, #2014/12236-1, #2016/04391-2, #2017/05416-1, and #2019/04434-1), and Getulio Vargas Foundation. The views expressed are those of the authors and do not reflect the official policy or position of the São Paulo Research Foundation. We also thanks Intel for making available part of the computational resources we use in the development of this work. Silva is funded in part by: the Moore-Sloan Data Science Environment at NYU; NASA; NSF awards CNS-1229185, CCF-1533564, CNS-1544753,

CNS-1730396, CNS-1828576, and DARPA. Finally, we would thank NEV experts, especially Alcides Eduardo dos Reis Peron, André Rodrigues de Oliveira, Marcelo Batista Nery, and Letícia Simões Gomes, for their help and support in the user study.

Germain García-Zanabria is a postdoctoral researcher at San Pablo Catholic University - Peru. He received his Ph.D. in Computer Science in 2021 from the University of São Paulo (ICMC-USP), São Carlos, Brazil. He holds his M.Sc. degree in Computer Science in 2016 from the San Pablo Catholic University, Arequipa, Peru. He holds a B.E. in System Engineering in 2012 from Universidad Nacional San Antonio Abad del Cusco, Cusco, Peru. His research interests comprise data science, data visualization, visual analytics, and visual learning models.

Marcos M. Raimundo received the B.S. degree in computer engineering, the M.S. and Ph.D. degree in electrical engineering from University of Campinas, Campinas, Brazil, in 2011, 2014, and 2018 respectively. He is a postdoctoral researcher at the School of Applied Mathematics at Fundação Getúlio Vargas. His research interests include machine learning, multiobjective optimization, mathematical programming, and operations research.

Jorge Poco is an Associate Professor at the School of Applied Mathematics at Fundação Getúlio Vargas in Brazil. He received his Ph.D. in Computer Science in 2015 from New York University, his M.Sc. in Computer Science in 2010 from the University of São Paulo (Brazil), and his B.E. in System Engineering in 2008 from National University of San Agustín (Peru). His research interests are data visualization, visual analytics, machine learning and data science. He has served in several program committees, including IEEE SciVis, IEEE InfoVis, and EuroVis.

Marcelo Batista Nery is Technology Transfer Coordinator at Center for the Study of Violence (RIDC -FAPESP) and Research Collaborator at the Institute of Advanced Studies – Global Cities Program, both from the University of São Paulo (USP), São Paulo, Brazil. PhD in Sociology at USP, with Split PhD at City and Regional Planning at the University of California, Berkeley. Master in Remote Sensing by the INPE (Brazil's National Institute for Space Research). He has a background in Geo-Information and Sociology areas, with emphasis on spatial analysis, geoprocessing, urban planning, public security, homicide, criminal dynamics, and urban spatial distribution.

Cláudio T. Silva is Professor of Computer Science and Engineering and Data Science at New York University. His research has focused on data science, visualization, graphics, and geometry processing. Recently he has been particularly interested in urban and sports applications. He received his BS in mathematics from Universidade Federal do Ceará (Brazil), and his MS and PhD in computer science at SUNY-Stony Brook. He has published over 250 peer reviewed journal and conference papers, and he has been an inventor on 12 US patents. He has advised or co-advised 15 post-docs, 20 PhD and 9 MS students. Claudio is a Fellow of the IEEE and has received the IEEE Visualization Technical Achievement Award. Silva's work has been covered in The New York Times, The Economist, ESPN, and other major news media.

Sergio Adorno is the Director of the Center for the Study of Violence of the University of São Paulo (NEV/USP). Currently a Full Professor at the Department of Sociology, University of São Paulo (USP), he was also Dean of the Faculty of Philosophy, Languages and Human Sciences at USP, 2012-2016, Executive Secretary of the National Association for Research in Social Sciences ANPOCS (1997-2000) and President of Brazilian Society of Sociology (1991-1995). He did a post-doctoral internship at the Centre de Recherches Sociologiques sur le Droit et les Institutions Pénales, CESDIP, França (1994-95). He teaches Sociological Theory and Political Sociology. His main fields of research are violence, human rights, crime and social control, theory of justice and democracy, social public policies.

Luis Gustavo Nonato received the PhD degree in applied mathematics from the Pontificia Universidade Catolica do Rio de Janeiro, Rio de Janeiro-Brazil, in 1998. His research interests include visualization, visual analytics, machine learning, and data science. He is full professor with the Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, Brazil. He was a visiting professor in the Center for Data Science, New York University, New York, from 2017 to 2018. From 2008 to 2010 he was a visiting scholar in the Scientific Computing and Imaging Institute, University of Utah, Salt Lake City. Besides having served in several program committees, including IEEE SciVis, IEEE InfoVis, and EuroVis, he was associate editor of the Computer Graphics Forum and currently he is associate editor of the IEEE TVCG. He is also the editor-in-chief of the SBMAC SpringerBriefs in Applied Mathematics and Computational Sciences.