CrimAnalyzer: Understanding Crime Patterns in São Paulo

Germain Garcia, Jaqueline Silveira, Jorge Poco *Member, IEEE,* Afonso Paiva, Marcelo Batista Nery, Claudio T. Silva *Fellow, IEEE,* Sergio Adorno, Luis Gustavo Nonato *Member, IEEE*

Abstract—São Paulo is the largest city in South America, with crime rates that reflect its size. The number and type of crimes vary considerably around the city, assuming different patterns depending on urban and social characteristics of each particular location. Previous works have mostly focused on the analysis of crimes with the intent of uncovering patterns associated to social factors, seasonality, and urban routine activities. Therefore, those studies and tools are more global in the sense that they are not designed to investigate specific regions of the city such as particular neighborhoods, avenues, or public areas. Tools able to explore specific locations of the city are essential for domain experts to accomplish their analysis in a bottom-up fashion, Revealing how urban features related to mobility, passersby behavior, and presence of public infrastructures (e.g., terminals of public transportation and schools) can influence the quantity and type of crimes. In this paper, we present CrimAnalyzer, a visual analytic tool that allows users to study the behavior of crimes in specific regions of a city. The system allows users to identify local hotspots and the pattern of crimes associated to them, while still showing how hotspots and corresponding crime patterns change over time. CrimAnalyzer has been developed from the needs of a team of experts in criminology and deals with three major challenges: i) flexibility to explore local regions and understand their crime patterns, ii) identification of spatial crime hotspots that might not be the most prevalent ones in terms of the number of crimes but that are important enough to be investigated, and iii) understand the dynamic of crime patterns over time. The effectiveness and usefulness of the proposed system are demonstrated by qualitative and quantitative comparisons as well as by case studies run by domain experts involving real data. The experiments show the capability of CrimAnalyzer in identifying crime-related phenomena.

Index Terms—Crime Data, Spatio-Temporal Data, Visual Analytics, Non-Negative Matrix Factorization

1 Introduction

Since the mid-1970s, Brazilian society has experienced a transition process from military dictatorship to democracy. With this political transition, it was expected that conflicts would increasingly be solved, reducing the prevalence of violence. That has not happened. In fact, the transition has been accompanied by an explosion of conflicts, many of which associated with urban crimes. There is still no consensus among social scientists about the reasons that explain these trends in the evolution of crime and violence in Brazilian society, in particular in the big cities [1]. Among the explanations that arise more frequently is the exhaustion of traditional security policy models. Concerning this last aspect, it is undeniable that crimes have not only grown, but also become more violent and modernized. In contrast, agencies in charge of law and order (e.g, police and criminal justice system) have not kept up with these trends. The gap between the dynamics of crime and violence and the state's ability to contain them within the rule of law has widened. Therefore, introducing modern instruments for the management of public order and crime containment is imperative to make public security policies more efficient, not only in São Paulo [1] but also in any big city in under-development countries.

Crime Mapping, a branch of Geographic Information Systems (GIS) devoted to explaining spatio-temporal behavior of crimes, has emerged as a research field to support criminologists in their analytical process, leveraging the importance of local geography as a determinant of crime types and the rate of which they occur in a particular region [48]. The capability of identifying and visualizing crime hotspots and the ability to filter crime-related attributes to reveal particular information such as burglary in commercial areas or the seasonality of auto theft in certain neighborhoods is among the key components of a crime mapping approach [38]. Most existing tools developed for crime mapping focused on the detection of *hotspots*, that is, areas with a high number of criminal incidents [14]. Although sophisticated mechanisms have been proposed to detect hotspots [15], the search for a high prevalence of crimes ends up neglecting sites where certain types of crimes are frequent but not sufficiently intense to be considered statistically significant [51]. Moreover, most techniques enable only rudimentary mechanisms to analyze an important component of unlawful activities, the temporal evolution of crimes and their patterns. In fact, visualization resources for temporal analysis available in the majority of crime mapping systems are very restrictive, impairing users from performing elaborated queries and data exploration [3].

There is yet another important aspect to be considered in the context of crime mapping, the specificities of urban areas under analysis. São Paulo, for example, bears one of the highest crime

- *Germain Garcia, Jaqueline Silveira, and Afonso Paiva are with ICMC-USP, São Carlos, Brazil. E-mail: {germaingarcia,alva.jaque}@usp.br, apneto@icmc.usp.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-USP, São Paulo, Brazil. E-mail: marsadorno@usp.br*
- *Jorge Poco is with Fundação Getúlio Vargas, Brazil and Universidad Católica San Pablo. E-mail: jorge.poco@fgv.br*
- *Claudio 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*

1. São Paulo is both a state and a city. In this paper, any time that we do not explicitly specify, São Paulo will refer to the city.

rates in the world, at least one order of magnitude higher than cities such as New York and San Francisco, making glyph-based crime mapping solutions such as *LexisNexis* [2], *NYC Crime Map* [3], and *CrimeMapping* [4] completely unsuitable for analyzing crimes in São Paulo. Nevertheless, the pattern of crimes changes considerably around São Paulo, even between regions that are geographically close to each other, demanding analytical solutions tailored to reveal local hotspots and corresponding crime patterns. Such local solutions should also be able to uncover the dynamic of hotspots over time. Those capabilities are not currently available in most crime mapping tools.

This work presents CrimAnalyzer, a new visual analytic tool customized to support the analysis of criminal activities in urban areas with the characteristics of São Paulo, that is, high criminality rates with great variability in the pattern of crimes, even in geographically close regions. CrimAnalyzer enables a number of linked views tailored to reveal patterns of crimes and their evolution over time, assisting domain experts in their decision-making process and providing guidelines not only for repressive but above all preventive actions, strengthening the planning and implementation of institutional actions, especially from the police.

In collaboration with a team of domain experts, we have designed visual analytic functionalities that allow users to select and analyze regions of interest in terms of their hotspots, crime patterns, and temporal dynamics. Moreover, the proposed system enables resources for users to dig deeper in particular sites to understand its prevalent crimes and their behavior over time. Furthermore, CrimAnalyzer implements a methodology based on Non-Negative Matrix Factorization [27] to identify hotspots based not only on the number of crimes but also on the rate they occur.

In summary, the main contributions of this work are:

- A new methodology to identify crime hotspots based not only on the number of crimes but also on their variation and recurrence rate.
- A visual analytics machinery that allows users to visually perform spatial and temporal queries towards understanding patterns and temporal dynamics of crimes.
- CrimAnalyzer, a visualization-assisted tool that integrates the analytical machinery in a set of linked views. CrimAnalyzer operates on target spatial regions to uncover relevant information of the region as a whole and also from its individual sites.
- A set of case studies revealing interesting phenomena about the dynamics of crime in São Paulo, supporting hypotheses and theories raised by domain experts and described in the literature.

2 Related Work

The literature about crime analysis is extensive, ranging from statistics and data science to visualization and GIS. Broadly speaking, crime analysis methods can be grouped into two major categories, geo-referenced and non-geo-referenced approaches. The latter, non-geo-referenced approaches, rely on mathematical and computational mechanisms such as data mining [12], [55], optimization [51], machine learning [52], [54], statistics [35], and data visualization [16], [53], to identify crime patterns, criminal behavior, and also the consistency of criminal justice. In the following, though, we focus on geo-referenced techniques developed for crime mapping that are more closely related to our approach. In order to better contextualize our methodology, we divide geo-referenced techniques into two groups, hotspot centered and spatio-temporal criminal pattern identification. It must be clear that there is considerable overlap between those groups, meaning that a hotspot centered technique can also rely on spatio-temporal patterns to leverage its analysis, but the main focus of such technique is, in fact, hotspot identification.

Hotspot centered. Identifying crime hotspots is a major task in the context of crime mapping [8], [14], [15]. Although some works rely on Kriging [33], the most common approach for hotspot identification is a combination of Spatial Scan Statistics [25] and Kernel Density Estimation (KDE) [10], using point clouds, density map, or grid-based approaches as visualization resources [22], [32], [45]. As pointed out by Hart and Zandbergen [21], properly setting the parameters of a KDE is not easy and a loose choice of parameters can lead to erroneous or inaccurate results that overestimate or disregard hotspot locations [9]. Another issue with KDE based techniques is that locations presenting regular, but not intense, criminal activities are hardly pointed out as hotspots. To avoid the issues above, our approach relies on Non-negative Matrix Factorization to detect hotspots, thus avoiding parameter tuning while being able to "capture" locations difficult to be identified as hotspots by KDE based methods.

Spatio-temporal pattern identification. Besides crime hotspot identification, the analysis and visualization of temporal and spatial crime patterns are also of great importance in crime mapping [7], [13], [29], [32], [42]. Ratcliffe [41] and Townsley [47], for instance, incorporates *aoristic analysis* [39], [40] in their hotspot visualization systems in order to identify important spatio-temporal patterns of crimes. The aoristic analysis takes into account the uncertainty of the exact moment that an event occurred when examining the overall incidence of crimes over time. Lukasczyk et al. [30] provide a topological perspective of the temporal evolution of hotspots based on a spatio-temporal Reeb graph built from a KDE mapping. Although interesting, techniques described above are still incipient in clearly revealing spatio-temporal crime patterns and their dynamics. Our approach, in contrast, combines a number of intuitive visual resources from which one can clearly identify crime patterns and their temporal evolution in specific locations.

There is a number of spatio-temporal techniques that rely on clustering methods to group spatially and/or temporally similar crime events in order to identify patterns. Those methods can be organized into two categories, the ones that build upon conventional clustering algorithms and the ones that rely on Self-Organizing Map (SOM) to identify patterns. Clustering-based methods extract feature vectors from spatial and temporal crime attributes and cluster those attributes via k-means [2], [46] or nearest neighbor clustering [23], [28].

The main goal of techniques that rely on SOM is the identification of similarities among crime attributes. Chen et al. [11], in collaboration with the Tucson Police Department, proposed a spatio-temporal visualization system called COPLINK, which combines hyperbolic trees, GIS, and SOM in a unified analytical tool. Andrienko et al. [4] rely on a SOM matrix display to leverage a visual analytic framework to explore spatio-temporal similarities between events. Hagenauer et al. [20] extended the previous approaches to explore the space-time evolution of the patterns, in addition to their demographic and socio-economic

2. communitycrimemap.com
3. maps.nyc.gov/crime/
4. crimemapping.com

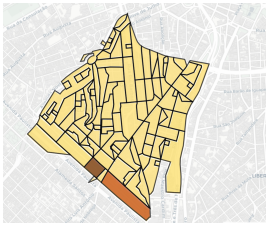
attributes. In order to understand patterns between crime types, SOM has also been the basis for the spatio-temporal crime analysis system proposed by Guo and Wu [19], which builds upon a visualization infra-structure called VIS-STAMP [18] that integrates dimensionality reduction and parallel coordinates in the analysis of crime patterns. SOM has well-known issues such as the proper setting of weights, number of nodes, and overfitting [26]. Moreover, SOM-based techniques described above do not integrate hotspot detection as part of the system, leaving aside an important component of analysis in the context of crime mapping.

3 Challenges, Data Set, and Analytical Tasks

For eighteen months, we interacted with two experts in social sciences whose research focusses on criminal analysis. One of the sociologists is a well-known researcher in the study of violence in South America. The other sociologist is an expert in public safety and social sciences applied to urban issues, with a background in GIS and large experience in spatio-temporal analysis of crime. In partnership with the police department of state of São Paulo, the team of experts built a data set (detailed in Sec. 3.2) containing seven years of criminal records in São Paulo. They approached us to develop a visual analytics tool to assist the understanding and analysis of the data.

Nomenclature. Before further detailing the problem, the requirements raised from the interaction with the domain experts, and the data set, let's first settle some nomenclature that will be employed throughout the manuscript.

– *Site* is the smallest territorial unity given in the spatial discretization. In our context, the sites are defined as the census units of São Paulo, each containing from 250 to 350 residences and/or commercial establishments.



– *Region* is a set of sites, which can correspond to a whole neighborhood, a particular portion of a neighborhood, or even a group of sites adjacent to a street or avenue. The inset on the right shows an example of a region and its corresponding sites.

– *Hotspots* are sets of sites within a region with relevant criminal activity. The exact meaning of "relevant" will be clear when we present the mechanism we designed for hotspot detection. The reddish sites in the inset image correspond to hotspot sites in the given region.

– *Crime type* refers to the type of criminal activity, ranging from burglary to bodily injury (death, sexual, and drug-related crimes are not included in our study).

– *Crime pattern* accounts for the prevalence of a group of crime types in a given region or sites. In other words, if we say that the crime pattern in a set of sites is robbery, car theft, and commercial establishment attack, we mean that the three crime types are the most prevalent ones in those sites.

3.1 Problem Analysis

We had several rounds of meetings and interviews with the experts to identify the main challenges involved in the analysis of crime data. After several interactions, we came up with the following issues:

- **Analyzing the characteristics and dynamics of crimes in particular regions of the city.** From their experience and interaction with officers from the police department, the experts conjecture that the type and dynamic of crimes have been changing over the years, mainly in particular regions of the city. Moreover, the type of crimes can change even in regions located close to each other depending on the urban characteristics of each region. The main difficulty to perform this analysis without a visual analytics tool is to properly query the data set, which can be a time-consuming and exhausting job. Many times a large number of images are generated as results, and the work of analyzing them becomes impossible. Moreover, highly prevalent crimes overshadow the presence of less frequent ones, which might also be of interest, demanding specific tools to enable a proper analysis. Given the difficulties, the experts have been performing their analysis focusing on just one or two types of crime, considering the city as a whole or analyzing large areas that serve as administrative units within the city. Such broad analysis hampers the validation (or denial) of hypothesis and conjectures that have a local nature.
- **Identifying crime hotspots within a particular region.** The identification of crime hotspots is among the most important tasks when analyzing crimes and their dynamics. Hotspots are usually identified as locations that have a greater than the average number of criminal records [14]. However, criminal sites that are not so prevalent in terms of the number of criminal events, but bears criminal activities that deserve special analysis, tend not to be detected when a "frequentist" approach is employed to identify crime hotspots. Due to the lack of more sophisticated mechanisms, the number of criminal records has been the main mechanism employed by the experts in their identification of hotspots. Because of this, it was necessary to propose a new method for hotspot detection that meets the described restrictions. This requirement was emphasized by the domain experts.
- **Understanding and comparing crime patterns.** Domain experts believe that sites and hotspots within the same region can present different crime patterns. An issue in this context is to know whether the pattern of crime varies from a site (or hotspot) to another in the same region. In affirmative case, experts would like to understand how crime types are distributed and how they evolve along time. The experts were looking for a solution that would intuitively allow them to make such comparisons.

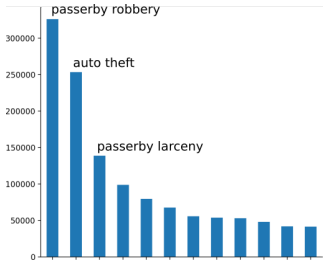
Challenges above point to a visual analytic solution endowed with functionalities to easily select regions of interest while enabling resources to assist the analysis of crime location, patterns, and temporal evolution. We followed a design process that involved the experts in most stages of the development [31], redesigning procedures, components, and functionalities according to experts feedback and demands.

3.2 São Paulo Robbery, Burglary, and Larceny Data

The data set assembled by the experts consists of criminal records provided by the police department of São Paulo. Only criminal acts as to robbery, burglary, and larceny were provided, leaving out murder, homicide, drug-related felony, and sexual assault.

Each record contains the identification number of the census unit (site) where the crime happened, the type of crime,

and the date and time of the crime. The data set contains crime records from 2000 to 2006. In the very beginning of our studies, we noticed that the information as to 2005 and 2006 was not consistent with previous years and a sanity check needed be performed by the experts. Since the sanity check turned out more complex than expected, we opt to include only information from 2000 to 2004 in our studies, in a total of 1,574,920 records.



Crime types range in 121 categories, and the 10% most frequent crime types correspond to about 80% of the total crimes. The inset on the right shows the histogram of the 10% most prevalent crime types, labeling the three most frequent ones, *passerby robbery*, *auto theft*, and *passerby larceny*. To facilitate the analysis, experts split the original data in three independent categories, *vehicle robbery* (includes cars, motorcycles, trucks, etc.) with 295,081 instances, *larceny* in general, with 587,885 instances, and a third category with all the other types of robbery and burglary, with 691,954 instances.

Although the number of crime types is quite large, the crimes that domain experts are interested in are not that large, ranging from 3 to 5. Other crime types are sparse enough to be analyzed individually, and do not require a sophisticated visualization tool to interpret them. Moreover, some crimes can be grouped into categories, an alternative suggested by the expert and incorporated into CrimAnalyzer. In other words, in each of the three sub-datasets, the experts ranked and grouped the crime types according to their importance.

3.3 Analytical Tasks

After identifying the main challenges faced by the experts and understanding how the data was structured, we conducted a new series of interviews to raise questions to be investigated. It has become clear that the experts are interested in understanding the dynamics of crimes over the city by analyzing the variation of crime patterns over space and time. From the iterative processes with the experts, we compiled the following list of analytical tasks:

- **Interactive selections (T1)**: How can spatial regions of interest be interactively selected? Is it possible to make the interactive selection of regions flexible enough to pick from single spots to whole neighborhoods and particular avenues?
- **Crime patterns over the city (T2)**: Which are the crime patterns in particular regions and sites? How do criminal patterns change from the center to residential areas and outskirts? What about the patterns along the main avenues, streets, and highways?
- **Dynamic of crimes over time (T3)**: How have crime types evolved, over time, in particular regions of the city?. More specifically, have crime patterns changed in particular regions over the years? Are crime types seasonal?
- **Crime patterns and hotspots over space (T4)**: Which are the hotspots in a given region? Which are their crime patterns? How different (if the difference exists) are the crime patterns in distinct hotspots within the same region?

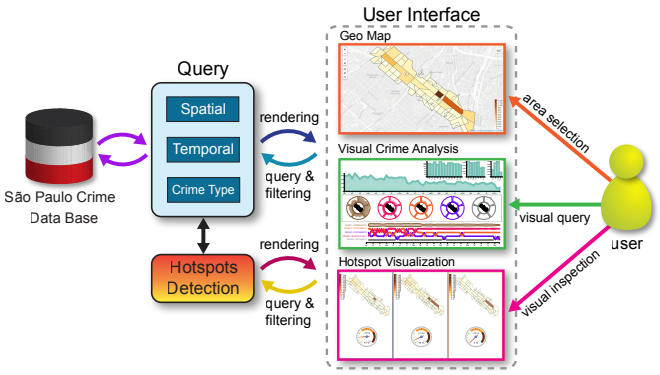


Fig. 1. Pipeline overview of the CrimAnalyzer System.

- **Crime patterns and hotspots over time (T5)**: Have crime hotspots changed over time in a given region? Have crime patterns changed over time in a given hotspot?

As mentioned before, the lack of interactive mechanisms to select regions of interest combined with general-purpose analysis and visualization techniques have prevented domain experts from freely exploring the data to verify hypotheses and conjectures. The first step to enable more powerful analytic resources is the design of a proper interactive selection tool, being this the goal of T1.

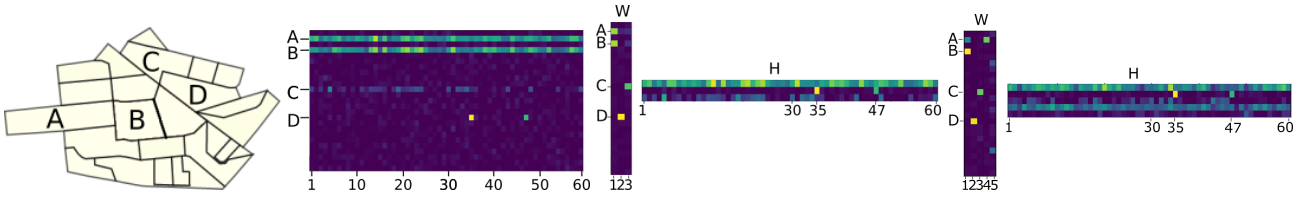
It has also become clear during the interviews that it is important to drill down from high-level summaries to individual analysis of sites and hotspots. Analyzing data in different scales is also essential to understand how patterns vary across space and time. For example, the pattern of crimes and hotspots can change throughout months and over different days of the week. This fact is related to T3, and requires particular data aggregation and filtering to be properly addressed.

Analytical tasks T2 and T3 are related to the problem of understanding the different patterns of crimes around the city and their evolution over time, as discussed in Sec. 3.1, while tasks T4 and T5 are associated to the problem of analyzing hotspots, also discussed in Sec. 3.1. To be properly addressed, those tasks demand specific mechanisms to detect hotspots and also visual resources to explore and understand them.

Among our goals is the integration of interactive selection methods and dedicated visual analysis tools towards allowing domain experts to accomplish both confirmatory and exploratory analysis. Moreover, some domain experts are not trained in computer science, thus, the system should be as simple and intuitive as possible. However, simplicity and expressiveness must be balanced to render the system capable of supporting spatio-temporal analysis at different scales, while being able to uncover non-trivial hotspots and crime patterns.

3.4 The CrimAnalyzer System

Based on the requirements and analytical tasks outlined in Section 3.3, we have developed CrimAnalyzer, a system for exploring spatio-temporal crime data in specific locations. CrimAnalyzer enables simple, yet compelling, visual resources to query, filter, and visualize crime data. The visual resources are supported by a mathematical and computational machinery tailored to extract and polish information so as to visually present it in an intuitive and meaningful way. The modules and system architecture are illustrated in Fig. 1. Users visually query the data set by interacting with a map and selecting a region of interest as well as by interacting with the different linked views that make up the system.



(a) Region of Interest (b) Data matrix X (c) Hotspots (columns of W) and their occurrence (rows of H) (d) The same as c) but with rank 5

Fig. 2. a) Region of interest. b) Data matrix containing crime information from the regions in a). Rows correspond to sites while columns are time slices. The darker the color, the closer to zero the number of crimes is. c) Rank 3 NMF decomposition from X. d) Rank 5 NMF from X.

4 Hotspot Identification Model

As discussed in Sec 3, hotspot identification is one of the most important tasks for crime analysis. Here, hotspots have a more general connotation than in previous work, corresponding to sites where criminal activity is high but also to locations where the number of crimes is not large, but frequent enough to deserve a detailed analysis. For example, in a given region, sites whose number of crimes is much larger than in any other sites are clearly important hotspots. However, the region can also contain a particular site where crimes are frequent, but happening in much smaller magnitude if compared against the prominent ones. The region can also contain sites where crimes are not frequent at all, but present spikes in particular time frames. We consider the three different phenomena as hotspots, seeking to identify sites where crimes are frequent and in large number, sites where crimes are frequent but do not in large number, and sites where crimes are not frequent, but happen in large numbers in particular time frames. The different crime behavior will be further discussed and illustrated in Sec. 4.2.

Analysis of individual sites. There are many alternatives to identify hotspots, as discussed in Sec. 2. Although varying in terms of complexity, existing techniques typically rely on the comparison of statistical information to identify hotspots. Hotspots can be identified as particular points or as area units, depending on how the data is organized, delegating to the visualization the task of properly revealing the hotspots. The problem with this approach is that crimes happening in small magnitude or in isolated time frames tend not to be statistically significant, hardly being pointed out as hotspots.

Another issue is that several sites might be identified as hotspots, but their temporal relation remains unclear. For example, sites can be timely correlated, meaning that crimes are committed in those sites in the same time slices. It makes sense to group timely correlated sites in a single hotspot, but computing hotspots individually and group them according to temporal matches is not easy and involves the use of thresholds to decide which sites should be grouped.

Analysis of groups of sites. Instead of analyzing sites individually, one can resort to techniques that directly identifies groups of sites as hotspots. A straightforward alternative is to extract features from the sites and apply a clustering scheme to group similar sites in hotspots (see Sec. 2). However, the problem of extracting meaningful features that characterize sites spatially and temporally is quite involved, mainly due to the sparsity of the crime data. In the course of our development, we tried different alternatives to define spatio-temporal crime feature vectors, ranging from simple cumulative time windows to more sophisticated methods based on graph wavelet coefficients [49], but we could not obtain results that complied with our requirements.

To get around the difficulties pointed above, we opted to an approach based on Non-Negative Matrix Factorization (NMF) [27], which worked pretty well for us in identifying hotspots according to our needs.

4.1 Non-Negative Matrix Factorization

Before presenting the details on how we have adapted NMF to operate for our work, lets shortly review the main concepts and ideas involved in an NMF analysis. An $m \times n$ matrix X is said *non-negative* if all entries in X are greater or equal to zero ($X \geq 0$). The goal of NMF is to decompose X as a product $W \cdot H$, where W and H are non-negative matrices with dimensions $m \times k$ and $k \times n$, respectively (the roles of m, n, and k will be clear in Sec. 4.2). In mathematical terms, the problem can be stated as follows:

$$\arg\min_{W,H} \|X - WH\|^2 \quad \text{subject to} \quad W, H \geq 0 \tag{1}$$

A solution for the minimization problem (Equation 1) provides a set of basis vector w_i, corresponding to the columns of W, and a set of coefficients h_j, corresponding to the columns of H, such that each column x_j of X is written as the linear combination $x_j = \sum_i h_{ij} w_i$, (or $x_j = W h_j$). In other words, for each column in X we have a corresponding column in H whose entries are coefficients associated to the columns (basis vectors) of W. The matrix representation below (Equation 2) illustrates the relation between columns of X and H as well as coefficients in H and basis vector in W.

$$\begin{bmatrix} | & & | & & | \\ x_1 & \dots & x_j & \cdots & x_n \\ | & & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ w_1 & w_2 & \cdots & x_k \\ | & | & & | \end{bmatrix} \begin{bmatrix} | & & h_{1j} & & | \\ h_1 & \cdots & h_{2j} & \cdots & h_n \\ | & & \vdots & & | \\ & & h_{kj} & & \end{bmatrix} \tag{2}$$

There are two important aspects in an NMF decomposition that will be largely exploited in the context of hotspot detection, namely, low rank approximation and sparsity. Low rank approximation accounts for the fact that the basis matrix W usually has a much lower rank than the original matrix X, meaning that (the columns of) X is represented using just a few basis vectors. As detailed in the next subsection, we rely on low rank approximation to define the number of hotspots, that is, by setting the rank of W we also set the number of hotspots. Sparsity means the basis and coefficient matrices contain many entries equal (or close) to zero, which naturally enforces only relevant information from X to be kept in W and H. This fact is important to identify particular sites within a hotspot and the time slices where each hotspot shows up.

4.2 Identifying Hotspots with NMF

We rely on NMF to identify hotspots, their rate of occurrence and "intensity". The matrix X to be decomposed as the product

$W \cdot H$ comprises crime information in a particular region of interest. Specifically, each row in X corresponds to a site of the region and each column to a time slice. In order to facilitate discussion, we present the proposed approach using a synthetic example. Fig. 2(a) shows a region made up of 25 sites, and we generated synthetic crime data in 60 time slices, representing months over five years. For sites denoted as A and B, we draw 60 samples from a normal distribution with mean 8 and variance 4, ensuring that A and B are correlated, that is, when the number of crimes in A is large the same happens with B (the number of crimes in B is generated by perturbing the values of A using a uniform random distribution with values between -3 and 3). This construction is simulating two regions with a high prevalence of crimes over time. Crimes in the site denoted as C in Fig. 2(a) follows a normal distribution with mean 1 and variance 4, corresponding to a location where crimes are not large in number, but happening quite frequently. Finally, for site D we draw 60 samples from a normal distribution with mean 0 and variance 0.25, except for time slices 35 and 47, where we set the number of crimes equal to 15 and 10 respectively, simulating a site where crimes are no frequent but happen in large numbers in particular time slices. For all the other sites, we associated 60 samples drawn from a normal distribution with mean 0 and variance 0.25. Values for all sites are rounded to the closest integer and negative values set to zero. Fig. 2(b) illustrates the matrix X built from the synthetic data described above. Notice that the simulated crime dynamics are clearly seen in X.

Given an $m \times n$ matrix $X \geq 0$, an NMF decomposition of X results in matrices $W \geq 0$ and $H \geq 0$. In practice, the rank of W is significantly smaller than both m and n, i.e., $k = \text{rank}(W) \ll m, n$. Here, the columns of W correspond to hotspots while entries in the rows of H indicate the "intensity" of the hotspot in each time slice. Fig. 2(c) illustrates matrices W and H obtained from matrix X in Fig. 2(b) using a NMF decomposition with rank $k = 3$. Notice that the entries in the first (left most) column of W have values close to zero almost everywhere, except in the entries corresponding to the sites A and B. Therefore, the hotspot derived from the first column of W highlights sites A and B as the relevant ones. The high prevalence of crimes on those regions can clearly be seen from the first (top) row of matrix H, which has most of its entries with non-zero values. The second column of W is mostly null, except in the entry corresponding to site D, where crimes are not frequent but happen with high intensity in particular time slices. Notice that the second row of H has basically two entries different from zero, corresponding exactly to the time slices 35 and 47, when the site D faces a large number of crimes. Finally, the last column of W gives rise to a hotspot that highlights site C, where crimes are frequent, but in smaller magnitude when compared to A and B. The incidence and intensity of crimes in C are clearly seen in the third (bottom) row of H.

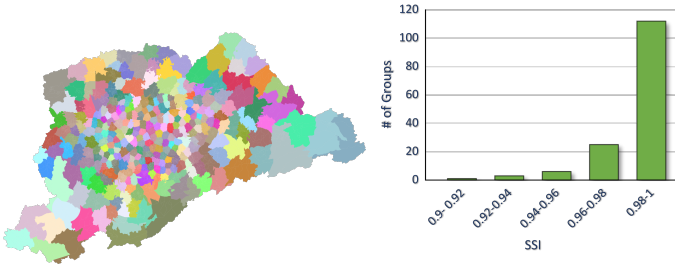
One can argue that the results presented in Fig. 2(c) worked so well because we wisely set the rank of W equal $k = 3$ and that in practice it is difficult to find a proper value for the rank. To answer this question, Fig. 2(d) shows the result of factorizing matrix X setting the rank of W equal $k = 5$. Notice that the main difference between the rank $k = 3$ and rank $k = 5$ factorizations is that the first column of W in Fig. 2(c) was split into two columns in the rank $k = 5$ factorization, giving rise to columns 1 and 4 of W in Fig. 2(d). Nevertheless, the first column still indicates the correlation between A and B, which thus is not completely missed due to the presence of column 4. The right most column of W in Fig. 2(d) is mostly noise, and it represents sites with a little criminal activity, what is attested by the bottom row of H in Fig. 2(d), which is almost null. Therefore, increasing k tends to split meaningful hotspots while creating some noisy, not so important ones, which can easily be identified from almost zero rows in H.

Improving the identification of hotspots. Most entries in matrix H are close to but are not zero, demanding a threshold to decide whether or not a hotspot takes place in a given time slice. Playing with thresholds is always inconvenient, mainly for non-experienced users. In order to avoid the use of thresholds, we binarize the matrix H using the Otsu's algorithm [36], considering that a hotspot appears in a given time slice if its corresponding entry in the binarized version of matrix H is 1.

The synthetic example discussed above shows that hotspots generated from NMF attend the requirements of our problem, justifying our choice of NMF as the mathematical model for tackling the problem. Among the different versions of NMF, we opt to the sparse non-negative matrix factorization proposed by Kim and Park [24], which allows for enforcing sparsity in both W and H simultaneously.

We conclude this section saying that, as far as we know, this is the first time that NMF is used as a mechanism to detect hotspots in crime mapping.

Comparison with spatial statistics The *Getis-Ord* G_i^* *statistics* [17], [34] is a well-known hotspot detection method available in the toolbox *Local Indicator of Spatial Association* (LISA) [5] G_i^* operates by measuring the local spatial autocorrelation variation over a region of interest. G_i^* reports a p-value and a z-score for each location in the region of interest, marking as hotspots those with statistically significant (low p-values) large z-scores.

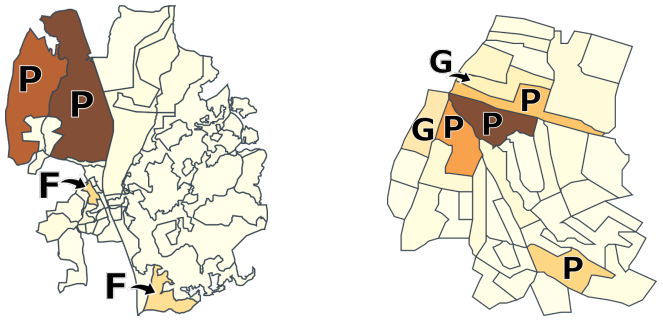


(a) São Paulo clustering (b) *SSI* distribution for $k = 3$

Fig. 3. (a) Division of São Paulo into 300 groups and (b) SSI distribution in those regions. NMF and G_i^* detect the same hotspots in most of the cases.

In order to perform a quantitative comparison between NMF and G_i^*, we grouped the census units into 300 regions as shown in Fig. 3(a). The regions have been computed by applying k-means to the coordinates of the centroid of the census units. Since there are about 30,815 census units, setting the number of clusters equal to 300 tends to generate groups with about 100 units in the denser areas of the city. The *Sokal-Sneath index* (SSI), a well known binary data classification similarity measure [44], is employed to compare the hotspots resulting from NMF with $k = 3$ (the default rank value in our system) against the ones obtained by G_i^* with a 99% confidence level (we relied on the G_i^* implementation available in the *PySAL* Python library [43]). Specifically, we assign each site to one of the four categories (labels):

- P: if the site is a hotspot for both NMF and G_i^* (positive match);
- F: if the site is a hotspot detected by NMF, but not by G_i^*;



(a) $|\mathsf{F}| = 2$ and $|\mathsf{G}| = 0$ (b) $|\mathsf{F}| = 0$ and $|\mathsf{G}| = 2$

Fig. 4. Qualitative comparison between NMF and G_i^*. (a) Region where NMF detects more hotspots and (b) region where G_i^* detects more hotspots.

- G: if the site is a hotspot detected by G_i^*, but not by NMF;
- N: if the site is not a hotspot for both methods (negative match).

The SSI similarity measure is then computed as:

$$SSI = \frac{2|\mathsf{P}| + 2|\mathsf{N}|}{2|\mathsf{P}| + |\mathsf{F}| + |\mathsf{G}| + 2|\mathsf{N}|},$$

where $|\cdot|$ denotes the cardinality. A $SSI = 1$ means that the hotspots detected by both methods in a given region match exactly.

Histogram depicted in Fig. 3(b) gathers SSI values from all the 300 regions. Note that *SSI* values are larger than 0.90, most of them lying in the range $[0.98, 1.00]$, showing the good match between NMF and G_i^*. In fact, most of the locations pointed out as hotspot by G_i^* are also captured by NMF. However, in about 200 regions, NMF detected a few more hotspots than G_i^*.

Fig. 4 illustrates typical situations where NMF and G_i^* differ in a few places. In Fig. 4(a), NMF and G_i^* have both found the hotspots labeled as P (the labels in Fig. 4 are according to the classification used by SSI, the darker the site is, the more crimes it has), but NMF has captured two extra hotspots, labeled as F (unlabeled units belong to the category N). Notice that the color code indicates that the sites F are indeed regions with a prevalence of criminality, although not captured by G_i^*. In Fig. 4(b), in contrast, G_i^* detects two more hotspots than NMF (G sites). Notice that the color code of the units marked as G in Fig. 4(b) indicates that crimes in those regions are not so intense as in the P hotspots. The reason why G_i^* points the G sites as hotspots is that those sites are neighbors of units where the number of crimes is high ("real" hotspots), so the kernel integration employed by G_i^* ends up being contaminated by the neighbor sites where crimes are prevalent. In other words, the G sites are pointed out as hotspots due to their proximity with P sites. Sites pointed as F in Fig. 4(a) have not been captured by G_i^* because they are isolated in the middle of units with no crimes. Therefore, besides not demanding a grid discretization, NMF tends to capture hotspots in a more consistent manner, being an attractive alternative to conventional statistical approaches.

The value k (NMF rank) impacts the *SSI* measure. We have run the comparisons ranging $k = 3, \ldots, 10$, getting an average *SSI* greater than 0.98 for $k = 3, 4, 5$, but slightly better for $k = 3$. This result motivated us to set $k = 3$ as the default value in CrimAnalyzer.

	Facet			Filter			Task				
	Space	Time	Type	Space	Time	Type	T1	T2	T3	T4	T5
Map View	✓			✓			✓	✓		✓	
Hotspot View	✓	✓			✓		✓			✓	✓
Cumulative Temporal View		✓			✓		✓		✓		✓
Global Temporal View		✓			✓		✓		✓		✓
Ranking Type View		✓	✓				✓	✓	✓	✓	✓
Radial Type View		✓	✓			✓	✓	✓	✓	✓	✓

TABLE 1
View properties and their analytical tasks (Sec. 3.3).

5 Visual Design

This section describes the visual components of CrimAnalyzer. Fig. 5 illustrates the web-based system, which comprises a Control Menu (a) six interactive views (b-g), and a filter widget (h). Table 1 shows the properties of each view. For instance, *Map View* shows the space facet and *Ranking Type View* the temporal and crime type facets. Columns under *filter*'s category show how to interact with each view. Some views allow to constrain space, time, and crime types. The design of visual resources was driven by the analytical tasks described in Sec. 3.3. In Table 1, Columns under *task's* category indicate the relation of each view and the analytical tasks. For instance, *Ranking Type View* and *Radial Type View* account for all analytical tasks.

In the area of crime analysis, visualizations have always been used to display the data; however, improvements or new designs over existing visualizations are needed. For instance, the *Ranking Type View* is a novel alternative visualization in this context, which turns out to be effective to elucidate the dynamics of different types of crime over time in specific locations of the city. Although this visual metaphor is well known by the visualization community, it has never been used for crime analysis.

In the following, we describe each visual components, starting with the *Control Menu* (see Fig. 5).

5.1 Control Menu

The *control menu* has three options: dataset, time discretization (*i.e.*, months or days) and the number of hotspots (rank of the NMF decomposition). As shown in Fig. 5(a) we are using the dataset "Roubo" with monthly discretization and three hotspots ($k = 3$) in most of our analysis. The NMF decomposition is performed using the *Nimfa* Python library [56], which is able to evaluate and automatically choose the rank value k. However, such an automated process is computationally costly, limiting its use in an interactive visual analytics application such as CrimAnalyzer. Therefore, we have opted to allow the user manually set k.

5.2 Map View

This is the first component used by users to start the analytical process, where the user can define the region of interest. This view is comprised of a geographical map and a choropleth map to encode the number of crimes committed at each *site* in the region. Also, users can zoom and pan the map.

Region selection: Users can define a region of interest by 1) clicking on the map (to select a site), 2) drawing a polyline (to select avenues or streets for example), 3) drawing a polygon (to select a whole neighborhood), or 4) provide the address of a location. Drawings can be expanded to include other sites in the neighborhood. Finally, CrimAnalyzer defines the *region* by computing the *sites* that intersect the drawing. In Fig. 5(b) we can see how a region is represented.

Site selection: During the exploration, when a region has already been defined, this view might be used for spatial filtering (*e.g.*, to

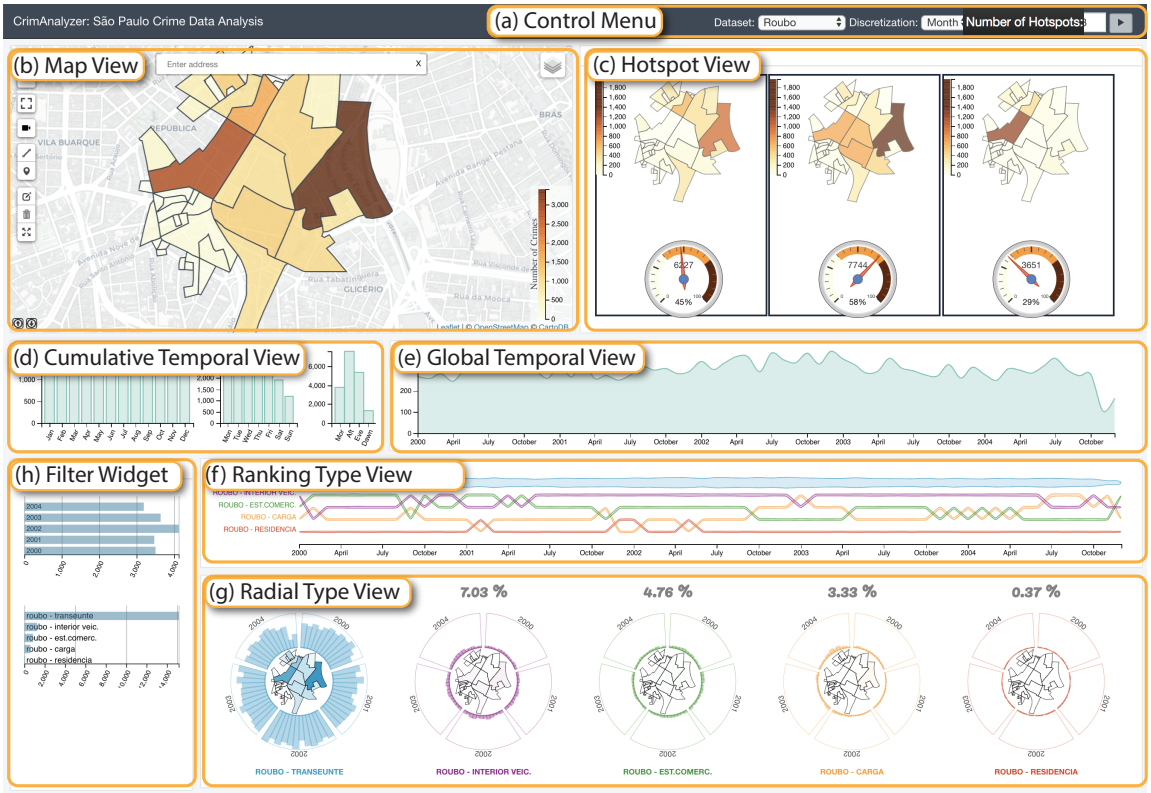


Fig. 5. CrimAnalyzer system: the spatial and temporal interactive views enable the exploration of local regions while revealing their criminal patterns over time.

focus on a particular site). This operation is performed by clicking a site, which is highlighted by mapping a texture to the corresponding area.

Filtering: When other views make spatial filtering (*i.e.*, selecting a site), the corresponding site is highlighted by changing its texture. When a time or type filter is activated by other views, our choropleth map is recalculated using the filtered data.

5.3 Hotspots View

An important component of our approach is the hotspots identification. In Sec. 4.2, we explained how Non-Negative Matrix Factorization has been used to reveal hotspots. In this view, we use multiple maps to represent the spatial distribution of each hotspot. Users can specify the number of hotspots in the *Control Menu*. Below each hotspot (see Fig. 5(c)), there is a gauge widget that depicts the number of crimes in the hotspot (the top number in the gauge), the temporal rate of occurrence of the hotspot (the bottom percentage in the gauge), and how relevant is that hotspot in the whole set of crimes (the gauge pointer). The importance of the hotspot is computed by a function $f : [0,1] \times [0,1] \rightarrow [0,1]$ that assign a value to each pair (`rate_of_crimes`, `frequency_of_crimes`), where `rate_of_crimes` denotes the number of crimes in the hotspot divided by the total of crimes and `frequency_of_crimes` is the temporal number of occurrences of the hotspot (computed for the binarized matrix H) divided by total number of time slices. In fact, f is simply a bilinear interpolation in the unit square where $f(0,0) = 0$, $f(0,1) = 0.5$, $f(1,0) = 0.7$, $f(1,1) = 1$. With this distribution of values, we give more relevance to hotspots where the number of crimes is larger.

Selection: A hotspot selection filters the crimes in space and type. All the other views are recomputed to match the selected hotspot.

Filtering: Filtering the crimes using other views (*i.e.*, space, time, or type) does not affect this view. If we want to recompute hotspots based on filtered data, for example, a particular crime type, we have to click the "Hotspots" button after performing the data filtering.

5.4 Global Temporal View

This view gives an overview of the number of crimes committed over the whole time period, relying on a line chart with a filled area between the data value and the base zero line (see Fig. 5(e)).

Time selection: In this view, we can constraint the analysis at a particular time interval, which can be defined by brushing a rectangle on the *Global Temporal View*. Only continuous time period can be selected. Next view will allow us to select multiple time intervals. All views (except the hotspot that need to be recomputed) are affected and automatically adjusted accordingly to the time selection.

5.5 Cumulative Temporal View

This view uses a bar chart to present the number of crimes accumulated by month, day, and period of the day (see Fig. 5(d)). In this view, we can see some patterns from non-continuous time intervals. This is also very useful to compare weekends or weekdays.

Filtering: When other views are used to filter the dataset, the filtered data is also overlaid on the global *Cumulative Temporal View*, thus enabling a comparative analysis (see Fig. 9).

5.6 Ranking Type View

This view depicts three relevant pieces of information in a single metaphor: crime type evolution, crime type ranking, and number of crimes in each time slice. As shown in Fig. 5(f), each crime type

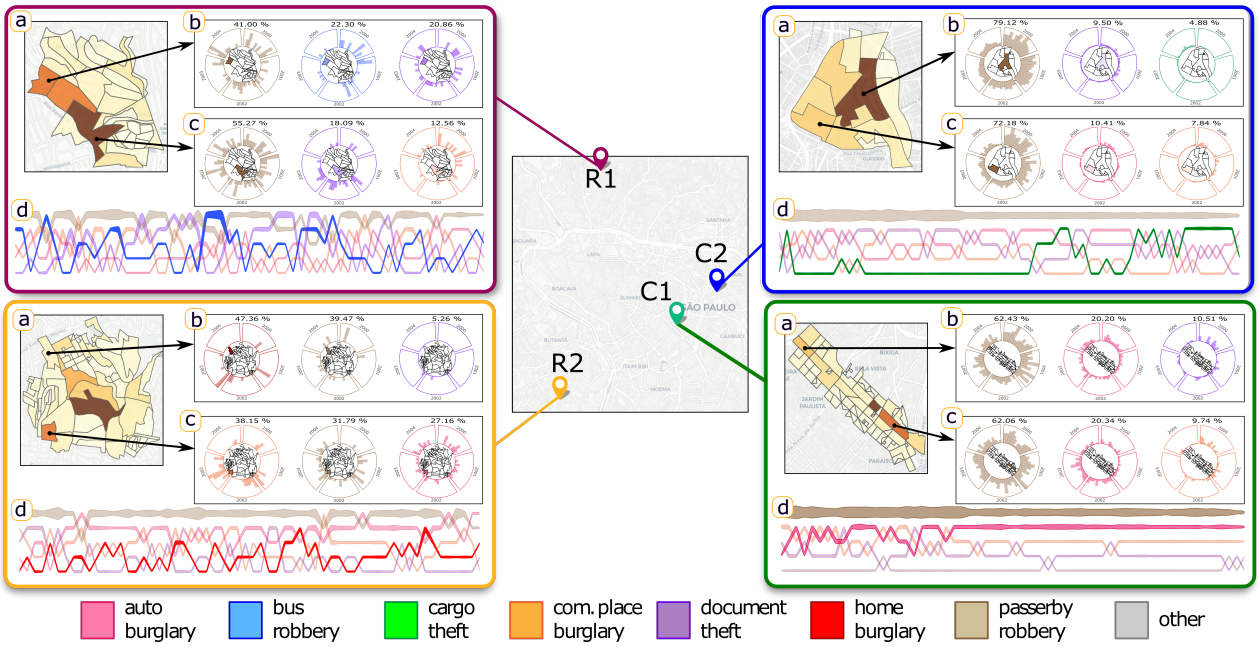


Fig. 6. Summary of criminal activities and corresponding patterns in four different regions of São Paulo. Crime patterns might change substantially among the regions and also along the time.

is represented by a polyline. The vertical position, on each time step, encodes the relevance compared to others. Moreover, the line width is proportional to the number of crimes belonging to it.

Filtering: When a filter is activated in other views, the ranking view is recomputed using the filtered data.

5.7 Radial Type View

In this view, we are using multiple bar charts with a radial layout. Each chart represents a different crime type, for instance, in Fig. 5(g) we have five crime types. In addition, the number on top of each chart shows the percentage for each crime type. Each chart is divided into sectors, where each sector is comprised of 12 bars depicting the months each year.

Crime Type selection: Clicking a chart filters the data to a specific crime type. In this way, users can focus their analysis on the most crime-prevalent types. Selected crime types are represented by a dashed borderline.

Time selection: We provide interactivity features on each chart to enable comparison among the same month on different years and same month across different crime types.

Filtering: When the dataset is filtered, each chart is recomputed to represented the filtered data.

5.8 Filter Widget

This widget is comprised of a time and crime type histogram. For instance, Fig. 5(h) summarizes our data in five years (2000-2004) and five crime types. Moreover, we use this histograms to filter our data. Clicking a bar, we can remove a year or a crime type. This filtering affects the whole interface.

Although most of the presented visual resources are not novel individually, many of them (such as *hotspot view*, *ranking type view*, and *radial type view*) are nontrivial in the context of crime mapping. Even more important, the combination of all of them allows multiple analysis simultaneously, revealing interesting crime patterns, as shown in the next section.

6 Case Studies

This section presents three case studies that show the effectiveness of CrimAnalyzer in addressing the analytical tasks presented in Sec. 3.3. The first case study addresses analytical tasks T1, T2, and T3, while the second focusses on hotspots analysis and it is related to T4 and T5. The third case study is aimed to make a parallel between criminal activity in São Paulo and some crime related phenomena reported in the literature (related to T3). In all case studies, except if explicitly stated, we used the robbery and burglary chunk of the dataset as described in Sec 3.2, with a monthly discretization.

6.1 Comparing Crime Patterns over the City (T1, T2, T3)

The goal of this case study is to analyze pattern of crimes in different regions of the city in order to understand how they change according to urban characteristics. Moreover, we also investigate the temporal evolution of crime patterns in different regions.

To perform the study we selected four areas in São Paulo, two in the center of the city, denoted as C1 and C2 in Fig. 6, and two in residential areas, pointed as R1 and R2 in Fig. 6. C1 is a financial district, hosting the headquarter of important banks and financial institutions, while C2 is a commercial area with many stores, an important metro terminal, and also several touristic places. Both C1 and C2 have a huge flow of people during the whole year. Residential areas R1 and R2 differ in terms of the economic level of residents, R1 is a middle-class neighborhood while R2 is a richer area, with luxurious buildings and houses.

Fig. 6 bottom right depicts region C1, selected by drawing a polyline along the main avenue of the financial district (T1), and highlights the *radial type view* (C1-c) of the three most prevalent crime types of two sites in C1 (indicated by the arrows). The *ranking type view* (C1-d) on the bottom shows how the incidence of the five most frequent crimes varies along the time. Two crime types lead the ranking along the years (the beige and pinkish curves on top), `passerby robbery` and `auto burglary`. By analyzing the *radial type view* (C1-b and C1-c)

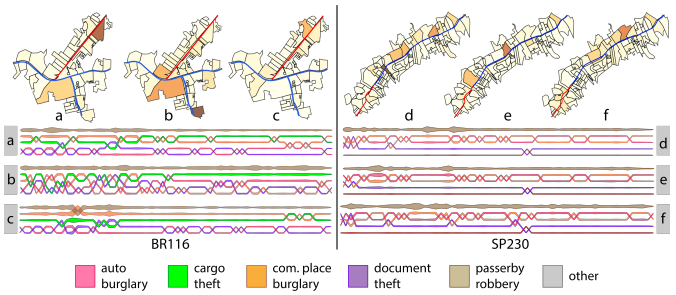


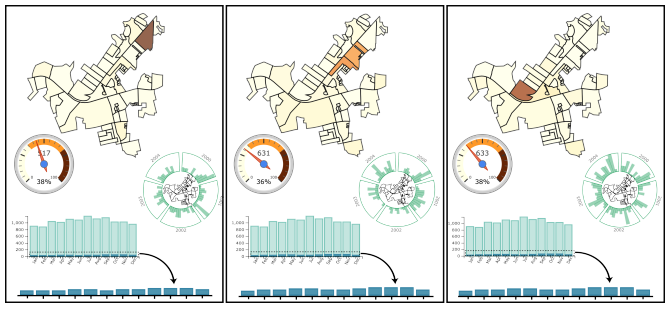
Fig. 7. Hotspots around the BR116 and SP230 highways considering all crime types.

of the highlighted sites, one can notice that those two crime types are indeed the prevalent ones in those regions (encoded by the color). Inspecting other sites by simply clicking on them on the map view, we concluded that `passerby robbery` and `auto burglary` are the prevalent crime types in almost all sites in C1. It is important, however, that these findings be interpreted in the context of the hypothesis that the spatial distribution of passerby robbery and auto burglary is shaped by the configuration of the street network. Note that the road network in C1 is the most regular; we can identify this by the number of well-defined city blocks in the analysis area. A city block with this characteristic is common in more consolidated and central urban areas, which leads us to conjecture a relationship between urban infrastructure and burglary risk.

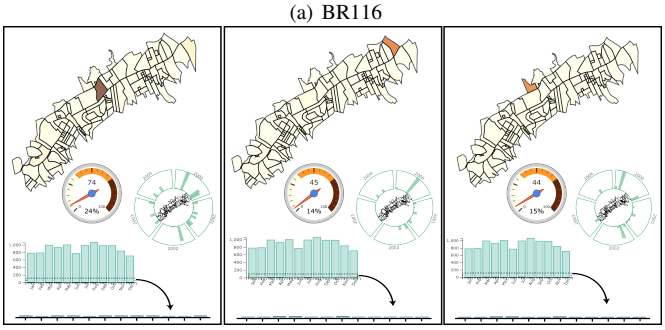
Performing the same analysis in region C2 (top right in Fig. 6), which was selected by clicking and expanding the central site of the region (the brownish one), we observe a different behavior. The *ranking type view* (C2-d) shows that there is one crime type that has been grown over the years (green curve), `cargo theft`. Selecting `cargo theft` from the *radial type view* in the CrimAnalyzer interface (Fig. 5(h)), the *map view* (Fig. 5(b)) reveals that `cargo theft` is not prevalent in the whole region, but it is concentrated in just a few sites, being the dark brown site in the center of the region. Notice that `cargo theft` became the third most common crime type in that region over time, being behind only of `passerby robbery` and `document theft`. Other sites present a more uniform behavior, having `passerby robbery`, `auto burglary`, and `commercial establishment burglary` as the main crime types.

Moving from the city center to more residential areas, the analysis reveals a substantial change in crime patterns, as one can observe on the left of Fig. 6, where the crime pattern in R1 and R2 is summarized. In the residential region R1, for example (top left in Fig. 6), `passerby robbery` remains the most common crime type, followed by `document theft`. However, some sites in R1 have `bus robbery` (passengers and/or drivers of public bus service are robbed) as the second most common criminal activity (R1-d). The orange site pointed out by the top arrow is an example (R1-b). Site-by-site crime pattern analysis is easy to perform with CrimAnalyzer, in this case, since the number of sites is mild and users need only to select the site on the map to make its crime pattern revealed. The importance of `bus robbery` in R1 is easily noticed in the *ranking type view* (R1-d) depicted on the bottom, where the blue curve (`bus robbery`) reaches high-rank levels in several opportunities.

Similarly to what happen in C1, C2, and R1 (and also in most of the city), region R2 (bottom left in Fig. 6) does have `passerby robbery` as the predominant crime type, what can clearly be



(a) BR116



(b) SP230

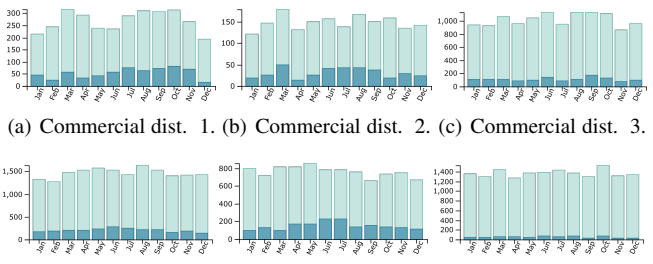
Fig. 8. `Cargo theft` hotspots along two important highways, BR116 and SP230. BR116 presents a much larger and more frequent number of *cargo theft* than SP230.

seen from the *ranking type view* (R2-d). However, crime patterns vary considerably among the sites, and some of them do not even have `passerby robbery` as the prevalent crime, as the two highlighted sites, which have `passerby robbery` as second in importance (R2-c and R2-d). Moreover, `home burglary` is the most relevant crime in one of those regions. In fact, home burglary is a relevant crime in R2 as a whole, as indicated by the reddish curve in the *ranking type view* (R2-d). Notice that `home burglary` has increased in importance over the years.

The discussion above shows that the visual analytic functionalities implemented in CrimAnalyzer are able to sort out analytical tasks T1, T2, and T3 in a simple, intuitive, and effective way. The flexibility to handle spatially complex neighborhoods with different shapes allows users to scrutinize set of blocks as well as regions along avenues and streets (analytical task T1). The combination of the *ranking type view* and the *radial type view* allows users to understand crime pattern in each region and in particular sites, making evident how crime patterns change around the city and even from site to site in a particular region (analytical task T2), a task difficult to be performed without the our visualization infrastructure. In particular, *ranking type view* and *radial type view* turn out to be effective in revealing the temporal behavior of crime patterns, making clear that patterns have changed along the years (analytical task T3). With the provided visual resources, this analysis would be an arduous process, demanding the implementation of multiple filters and sophisticated numerical and computational tools. In fact, the difficulty in performing a similar analysis with existing analytical systems is partly due to limitations on their visual resources and partly to the inadequacy of existing tools to reveal gist information hidden in the data.

6.2 Hotspot Analysis and Cargo Theft (T1, T4, T5)

This case study has been driven by the domain experts, and they were interested in a particular type of crime, `cargo theft`. Although `cargo theft` does not figure among the most prominent



(a) Commercial dist. 1. (b) Commercial dist. 2. (c) Commercial dist. 3.

(d) Commercial street 1. (e) Commercial street 2. (f) Commercial street 3.

Fig. 9. Commercial establishment burglary tends to increase during the winter (winter in South America goes from mid June to mid September).

crime types in São Paulo, it is of great interest due to its spatial characteristic, the high values involved, and the engagement of violent gangs in this type of criminal activity. It is well known that robbery (or theft) of high valuable cargo commodities tends to happen close to the main highways connecting São Paulo to other regions of Brazil. Therefore, domain experts focused their analysis in two important highways, SP230, which connects São Paulo to states in the south of Brazil, and BR116, which connects São Paulo to Rio de Janeiro.

In order to perform their analysis, domain experts relied on the polyline selection tool to select a considerable number of sites along the highways and avenues that connect the city to the highways. The number of regions involved in these analysis renders a site-by-site investigation tedious, making hotspots a better alternative. Fig. 7 shows three *hotspots* obtained from the regions selected along BR116 and SP230 and nearby avenues. The highways are highlighted in red and the nearby avenues in blue in the hotspot maps depicted in Fig. 7. The *ranking type view* reveals the crime patterns in each hotspot (considering only the five most relevant crime types). Notice that in BR116, `cargo theft` figures among the most relevant crimes (green lines), becoming the second most relevant crime at multiple times. In SP230, though, `cargo theft` is not predominant, not appearing among the five most relevant crimes in the *ranking type view* in any hotspot. In SP230, the predominant pattern is `passersby robbery`, `vehicle burglary`, and `commercial establishment burglary`. CrimAnalyzer makes clear which sites are relevant in each hotspot, their crime patterns, and how crime patterns evolve, thus properly addressing analytical tasks T4 and T5.

However, the experts are interested in `cargo theft`. To center the analysis in a single crime type users only need to select that type in the *radial type view*, filtering the data such that hotspots and all the views are updated to depict only information related to the selected crime type. Fig. 8 shows the hotspots associated to `cargo theft` only. The *gauge widgets* show that the number of `cargo theft` in the BR116 is one order of magnitude larger than in SP230, also presenting a higher rate of occurrence. The temporal evolution (*radial type view*) on the center-right of each *grid* shows the temporal behavior of `cargo theft` in each hotspot. It is clear that the number of `cargo theft` in SP230 has lessened over the years, while in BR116 no reduction is observed. The histograms below the gauge widget show the intensity of `cargo theft` (the short dark bars) in each month, comparing them against the total number of crimes in the region.

The CrimAnalyzer viewing tools also make clear that, in BR116, `cargo theft` takes place mainly along the highway (red curves in BR116 maps in Fig. 7), while in SP230 the relevant sites of each hotspot are located in the avenue that connects the

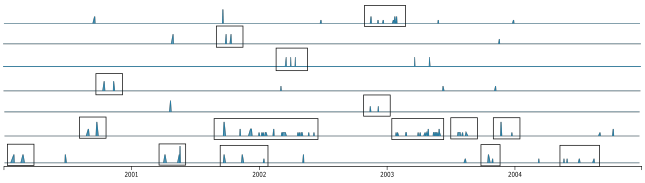


Fig. 10. The *Near Repeat Victimization* phenomena. When a home is burgled, the risk of recidivism in a short period of time is not only higher for the targeted home, but also for the nearby homes.

highway to the city (blue curves in SP230 maps in Fig. 7). Domain experts considered this an important finding because it is known that the `modus operandi` of criminal offenders and, hence, the location of `Cargo Theft` change according to the transported product. So, the possibility of identifying these roads should make public security policies more efficient. Another interesting aspect pointed out by the experts is the capability of revealing hotspots associated with sparse criminal activities, as the one depicted in Fig. 8(b) (see the spikes in the radial view). Sparse hotspots are relevant and deserve to be investigated, as they may be associated with local characteristics that would likely increase the chance of crimes being committed. Notice that these findings could hardly be made without the visual resources enabled by systems such as CrimAnalyzer.

6.3 Seasonality and the Temporal Element of Crime (T3)

This case study corroborates whether some criminal behaviors described and validated in previous works also take place in São Paulo.

Seasonality An important aspect related to criminal activities is seasonality. There is a number of studies in the literature that support the hypothesis that certain crime types are seasonal while others are not. For instance, van Koppen and Jansen [50] argue that, in Netherlands, commercial establishment burglary (robbery) increases during the winter due to the increased number of dark hours during the day. In South America, winter usually starts in mid-June and last until mid-September, during this period, especially in July and August, the number of dark hours is higher than in the rest of the year. An interesting question related to task T3 is whether the findings of van Koppen and Jansen is valid in São Paulo. To look for an answer, we relied on CrimAnalyzer to explore six major commercial areas in São Paulo city, three commercial districts and three popular commercial streets. Fig. 9 shows the *cumulative temporal view* of each of the analyzed regions. The overlaid darker histograms correspond to the number of `commercial establishment burglary` and `robbery` in each month. The overlaid histogram is generated by simply selecting `commercial establishment burglary` in the *temporal type view*.

From Fig. 9 one clearly sees that five out of six regions present an increase in the number of `commercial establishment burglary` and `robbery` during the winter (a-e), thus supporting the findings of van Koppen and Jansen. Although we can not claim with certainty that the hypothesis is true, the analysis enabled by CrimAnalyzer provides evidence about the seasonality of this type of crime, thus helping to answer one of the questions associated to task T3.

Near Repeat Victimization Near repeat victimization theory claims that when a home is burgled, the risk of recidivism is

not only higher for the targeted home, but also for the nearby homes, with risk period that seems to decay after some weeks or months [37]. The near repeat victimization theory has found evidence of its veracity in a number of countries, but we could find no report about it in São Paulo.

Using CrimAnalyzer, we scrutinized two regions in São Paulo where `home burglary` is a recurrent crime, including region R2 discussed in the case study presented in Sec. 6.1. Fig. 10 shows the time series, in a daily temporal scale, of seven sites in the analyzed regions, which varies in terms of the frequency of crimes and the number of `home burglary`. The boxed spikes point `home burglary` events that occur less than thirty days apart from each other. Notice that even in sites where `home burglary` is really occasional (rows 2 to 5 in Fig. 10), the near repeat victimization phenomena can clearly be observed.

Seasonality and *Near Repeat Victimization* are straightforward to be observed with CrimAnalyzer, enabling a number of analytical possibilities. For instance, in warmer seasons, day light lasts longer, encouraging a larger number of people to stay on the streets, increasing their exposure to illicit acts and criminal activities. During holiday season, it is common people to travel to countryside, leaving their property unprotected, facilitating burglary and other forms of crime. Those phenomena can also be analyzed with CrimAnalyzer.

7 Evaluation from the Experts

After using CrimAnalyzer and running a variety of experiments, including the case study reported in Sec. 6.2, the domain experts have given us the following feedback.

"*Despite its limitations, CrimAnalyzer has allowed us to better understanding challenges not yet elucidated by conventional crime analysis tools. First, by using solid mathematical and computational resources to reveal geo-referenced criminal activities, CrimAnalyzer incites the search for plausible explanations for the observed criminal patterns, what would be impossible with conventional analysis. Second, CrimAnalyzer motivates reflection about the relationship among the different crime types and about topological, directional, and relational connections that might affect the number of crimes in specific locations and time intervals. Third, an analytical tool that enables the analysis of crimes in specific locations leads to thinking the city in its complexity and, at the same time, guides the investigation of urban characteristics (administrative, demographic, physical, and social) and their interaction from which the observed local patterns result. Fourth, CrimAnalyzer uncovers the heterogeneity of the city as to its urban infrastructure, the differences among commercial, financial, and residential areas, the flow of people, public and private transportation, as well as the need for improvements, not only in terms of policing in specific locations and according to the type of crimes, but also, and mainly, in terms of tools to assist criminal investigation towards reducing the high rates of impunity. Finally, in contrast to more simplistic statistical methodology, the deterministic approach for hotspot identification turns out fundamental to emphasize the dynamics of spatio-temporal processes and to capture typical social manifestations such as crimes.*"

The experts were quite enthusiastic about CrimAnalyzer, as it allowed them to understand and raise hypotheses about a number of phenomena, as in the `cargo theft` case, that would be hard otherwise. Specifically, one of the experts said: "*Analyzing the vast amount of information enabled by CrimAnalyzer, we could detect spatio-temporal patterns and trends that will allow us to improve public policies...*".

8 Discussion and Limitations

CrimAnalyzer was developed in close cooperation with domain experts. The current version satisfies their requirements, however, some limitations and future work have been identified as part of our collaboration.

NMF stability. Our approach for identifying hotspots is not stable, this is because the Non-Negative Matrix Factorization technique depends on the initial conditions of the optimization procedure. To counteract this effect, some implementations, like the one we are using in our system, enables us to run the method a number of times, keeping the solution with the smallest error. Although the results get quite stable after enabling the multiple run alternative, a more robust approach could be sought to mitigate possible effects.

Space Discretization. The space discretization used in CrimAnalyzer is the census units in São Paulo, we adopted this measure because our collaborators had an interest in seeing the analysis in this level of detail. However, we are aware of the modifiable areal unit problem (MAUP), census units do not represent "natural units" of analysis and the result of certain analysis can change by modifying the aggregation unit [8]. An immediate future work would be to extend and make more flexible our space discretization. In this way, we should be able to apply our tool in other scenarios.

Multiple data sources. Crime events by their own rarely tell the whole story. Additional data that can be used to enhance the understanding of the crime layer. For example, the presence of bars and pubs, distance to parks, vacant land and buildings, weather, among other information might have a relation with certain criminal activities. Given the increasing number of initiatives to make data publicly available, we are considering to combine that information to further understanding crimes in urban areas. An interesting mathematical tool in this context is tensor decomposition, a generalization of matrix decomposition able to extract patterns from multiple data sources. Developing visual analytical tools to map tensor decomposition information into visual content is an important problem [6] that has barely been approached in the context of crime analysis.

Global vs Local approach. CrimAnalizer uses a local-based approach to explore and analyze crime patterns. Even though this was a requirement from the domain experts, and we agree that it was the correct approach to this problem, mainly because domain experts have prior knowledge and hypothesis regarding crime behaviors in particular locations, in some of our interviews with domain experts we discussed the option of having a global-based technique that might process the whole space and propose interesting locations to be explored. This alternative was accepted by the experts but as a complementary technique. As future work, we are also interested in tackling this problem from both perspectives (global and local).

Multiple cities and different scenarios Finally, we intend to apply and validate our system in other cities and countries. Currently, we are in the process of collecting crime data from multiple locations, and in a short time, we expect to release the system to analyze multiples cities in Brazil. In addition, our approach can be extended to other scenarios than crime analysis. For instance, one can use the system to analyze the dynamics of traffic-accidents in particular locations of the city, making possible

to uncover how the number of car-car crashes, car-bus crashes, run overs, etc. evolve over time.

9 Conclusion

We introduced a visual analytics tool to support the analysis of crimes in local regions. We developed CrimAnalyzer in close collaboration with domain experts and translated their analytical into the visualization system. We also propose a technique based on NMF to identify hotspots. Our system was validated by qualitative and quantitative comparisons, and case studies using real data and with feedback from the domain experts. Moreover, we verified two crime behavior (*i.e.*, *seasonality* and *near repeat victimization*) using São Paulo crime data.

Acknowledgments

This work was supported by CNPq-Brazil (grants #302643/2013-3 and #301642/2017-63), CAPES-Brazil (grants #10242771), and São Paulo Research Foundation (FAPESP)-Brazil (grant#2014/12236-1, #2016/04391-2 and #2017/05416-1). 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. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] S. Adorno. Democracy in progress in contemporary brazil: Corruption, organized crime, violence and new paths to the rule of law. *International Journal of Criminology and Sociology*, 2:409–425, 2013.

[2] T. Aljrees, D. Shi, D. Windridge, and W. Wong. Criminal pattern identification based on modified k-means clustering. In *2016 International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, pp. 799–806, 2016.

[3] M. A. Andresen. Mapping crime prevention: What we do and where we need to go. In *Crime Prevention in the 21st Century*, pp. 113–126. Springer, 2017.

[4] G. L. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. v. Landesberger, P. Bak, and D. Keim. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. *Computer Graphics Forum*, 29(3):913–922, 2010.

[5] L. Anselin. Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27:93 – 115, 1995.

[6] R. Ballester-Ripoll, P. Lindstrom, and R. Pajarola. Tthresh: Tensor compression for multidimensional visual data. *IEEE TVCG*, 2019.

[7] C. Brunsdon, J. Corcoran, and G. Higgs. Visualising space and time in crime patterns: comparison of methods. *Computers, Environment and Urban Systems*, 31(1):52–75, 2007.

[8] S. Chainey and J. Ratcliffe. *GIS and Crime Mapping*. Wiley, 2005.

[9] S. Chainey, S. Reid, and N. Stuart. When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime. In *Socio-Economic Applications of Geographic Information Science*, pp. 21–36. Taylor & Francis, 2002.

[10] S. Chainey, L. Tompson, and S. Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

[11] H. Chen, H. Atabakhsh, T. Petersen, J. Schroeder, T. Buetow, L. Chaboya, C. O'Toole, M. Chau, T. Cushna, D. Casey, and Z. Huang. COPLINK: Visualization for crime analysis. In *National Conf. Digital Government Research*, 2003.

[12] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau. Crime data mining: A general framework and some examples. *Computer*, 37(4):50–56, 2004.

[13] S. N. de Melo, D. V. S. Pereira, M. A. Andresen, and L. F. Matias. Spatial/temporal variations of crime: A routine activity theory perspective. *International Journal of Offender Therapy and Comparative Criminology*, pp. 1–26, 2017.

[14] J. Eck, S. Chainey, J. Cameron, and R. Wilson. Mapping crime: Understanding hotspots. Technical report, National Institute of Justice, 2005.

[15] E. Eftelioglu, S. Shekhar, and X. Tang. *Crime hotspot detection: A computational perspective*, pp. 82–111. IGI Global, 2016.

[16] R. Gao, H. Tao, H. Chen, W. Wang, and J. Zhang. Multi-view display coordinated visualization design for crime solving analysis: Vast challenge 2014: Honorable mention for effective use of coordinated visualizations. In *Conf. VAST*, pp. 321–322, 2014.

[17] A. Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3):189–206, 1992.

[18] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns. *IEEE TVCG*, 12(6):1461–1474, 2006.

[19] D. Guo and J. Wu. Understanding spatiotemporal patterns of multiple crime types with a geovisual analytics approach. In *Crime Modeling and Mapping Using Geospatial Technologies*, pp. 367–385. Springer, 2013.

[20] J. Hagenauer, M. Helbich, and M. Leitner. Visualization of crime trajectories with self-organizing maps: a case study on evaluating the impact of hurricanes on spatio-temporal crime hotspots. In *Proceedings of the 25th conference of the International Cartographic Association*, 2011.

[21] T. Hart and P. Zandbergen. Kernel density estimation and hotspot mapping: Examining the influence of interpolation method, grid cell size, and bandwidth on crime forecasting. *Policing: An International Journal*, 37(2):305–323, 2014.

[22] E. Johansson, C. Gåhlin, and A. Borg. Crime hotspots: An evaluation of the kde spatial mapping technique. In *2015 European Intelligence and Security Informatics Conference*, pp. 69–74, 2015.

[23] R. Kerry, P. Goovaerts, R. P. Haining, and V. Ceccato. Applying geostatistical analysis to crime data: Car-related thefts in the baltic states. *Geographical Analysis*, 42(1):53–77, 2010.

[24] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[25] M. Kulldorff. A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496, 1997.

[26] J. Lampinen and T. Kostiainen. Generative probability density model in the self-organizing map. In *Self-Organizing Neural Networks: Recent Advances and Applications*, pp. 75–94. Springer, 2002.

[27] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 556–562, 2001.

[28] N. Levine. CrimeStat IV: A spatial statistics program for the analysis of crime incident locations. Technical report, National Institute of Justice, 2013.

[29] D. Li, Y. Wang, S. Wu, J. Qi, and T. Wang. An visual analytics approach to explore criminal patterns based on multidimensional data. In *2017 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5563–5566, 2017.

[30] J. Lukasczyk, R. Maciejewski, C. Garth, and H. Hagen. Understanding hotspots: A topological visual analytics approach. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 36:1–36:10, 2015.

[31] T. Munzner. A nested model for visualization design and validation. *IEEE TVCG*, 15(6):921–928, 2009.

[32] T. Nakaya and K. Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS*, 14(3):223–239, 2010.

[33] M. A. Oliver and R. Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

[34] J. K. Ord and A. Getis. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical analysis*, 27(4):286–306, 1995.

[35] D. W. Osgood. Statistical models of life events and criminal behavior. In *Handbook of Quantitative Criminology*, pp. 375–396. Springer, 2010.

[36] N. Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[37] N. Polvi, T. Looman, C. Humphries, and K. Pease. The time course of repeat burglary victimization. *The British Journal of Criminology*, 31(4):411–414, 1991.

[38] J. Ratcliffe. Crime mapping: Spatial and temporal challenges. In *Handbook of Quantitative Criminology*, pp. 5–24. Springer, 2010.

[39] J. H. Ratcliffe. Aoristic analysis: the spatial interpretation of unspecific temporal events. *International Journal of Geographical Information Science*, 14(7):669–679, 2000.
[40] J. H. Ratcliffe. Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *Journal of Quantitative Criminology*, 18(1):23–43, 2002.
[41] J. H. Ratcliffe. The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research*, 5(1):4–28, 2004.
[42] J. H. Ratcliffe. A temporal constraint theory to explain opportunity-based spatial offending patterns. *Journal of Research in Crime and Delinquency*, 43(3):261–291, 2006.
[43] S. J. Rey and L. Anselin. PySAL: A Python Library of Spatial Analytical Methods. *The Review of Regional Studies*, 37(1):5–27, 2007.
[44] C. S and S. Cha. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, pp. 43–48, 2010.
[45] R. B. Santos. *Crime Analysis with Crime Mapping*. SAGE, 4th ed., 2016.
[46] L. J. S. Silva, S. Fiol-González, C. F. P. Almeida, S. D. J. Barbosa, and H. Lopes. CrimeVis: An interactive visualization system for analyzing crime data in the state of rio de janeiro. In *19th International Conference on Enterprise Information Systems (ICEIS)*, pp. 193–200, 2017.
[47] M. Townsley. Visualising space time patterns in crime: the hotspot plot. *Crime Patterns and Analysis*, 1(1):61–74, 2008.
[48] M. Townsley. Crime mapping and spatial analysis. In *Crime Prevention in the 21st Century*, pp. 5–23. Springer, 2017.
[49] P. Valdivia, F. Dias, F. Petronetto, C. T. Silva, and L. G. Nonato. Wavelet-based visualization of time-varying data on graphs. In *IEEE VAST*, pp. 1–8, 2015.
[50] P. van Koppen and R. Jansen. The time to rob: variations in time of number of commercial robberies. *Journal of Research in Crime and Delinquency*, 36(1):7–29, 1999.
[51] D. Wang, W. Ding, H. Lo, M. Morabito, P. Chen, J. Salazar, and T. Stepinski. Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems*, 39:93–106, 2013.
[52] T. Wang, C. Rudin, D. Wagner, and R. Sevieri. Learning to detect patterns of crime. In *Machine Learning and Knowledge Discovery in Databases*, pp. 515–530, 2013.
[53] J. Xu and H. Chen. Criminal network analysis and visualization. *Commun. ACM*, 48(6):100–107, 2005.
[54] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav. Crime pattern detection, analysis prediction. In *Int. Conf. Electr., Communic. Aero. Tech. (ICECA)*, vol. 1, pp. 225–230, 2017.
[55] Z. Ying. Analysis of crime factors correlation based on data mining technology. In *Int. Conf. Robots Intel. Sys. (ICRIS)*, pp. 103–106, 2016.
[56] M. Zitnik and B. Zupan. Nimfa: A Python Library for Nonnegative Matrix Factorization. *Journal of Machine Learning Research*, 13:849–853, 2012.