Crime prediction and prevention using police patrolling data: challenges and prospects

Thales Vieira[1], Tiago Paulino[1], João Matheus Siqueira Souza[2], and Edival Lima[3]

[1]Institute of Computing, Federal University of Alagoas (UFAL), Maceió, Brazil
[2]Institute of Mathematics and Computer Sciences, University of São Paulo (ICMC-USP), São Carlos, Brazil
[3]Military Police of the State of Alagoas, Maceió, Brazil
{thales, tps2}@ic.ufal.br, joaomatheus@usp.br, edival.lima@pm.al.gov.br

***Abstract*—Spatiotemporal crime analysis and prediction aim at identifying criminal patterns in space and time. In previous work, crime prediction has been performed by identifying hotspots from data, which means areas of high criminal activity on the streets. By focusing efforts on such sites, police patrolling is expected to be more efficient, thus reducing criminal activity. However, not many studies focus on investigating how police patrolling affects crime, and whether it can be a predictor of crime activity. In this paper we discuss the main challenges of this problem, and describe some work in progress towards developing a robust methodology to represent, visually analyze, and build predictors for criminal activity, considering both criminal and police patrolling spatiotemporal data. As a case study, we use real datasets from the Military Police of the state of Alagoas, Brazil (PM-AL).**

I. Introduction

Crime analysis is a valuable resource employed by crime analysts to acquire knowledge and support the decision-making of public security agencies [1]. Specifically, spatiotemporal crime analysis and prediction aim at identifying criminal patterns in space and time. For instance, spatial crime prediction may be performed through hotspot mapping, which is the identification of the areas of high concentrations of crime [2]. This definition was later modified to include not only locations with a high concentration of crimes but also sites where crime is frequent but does not occur in high volume [3]. The accurate identification of crime hotspots is an essential step to efficiently planning police patrolling and assisting other crime reduction actions [4]. Nevertheless, several associated computational problems are still not well solved.

One of the major questions when tackling spatiotemporal crime data is how to discretize and aggregate, in space and time, crime occurrences. For instance, spatial discretization approaches for crime analysis include census units [5], police districts [6], regular grids [7] and corners of a city street graph [3]. This choice restricts the methods that can be employed to detect hotspots, and also their accuracy and consequently the effectiveness of police patrol planning strategies.

Another relevant aspect is the (non-consensual) definition of a crime hotspot, which may include sites with a high concentration of crimes, or with regular frequency. More generally, one may argue that a spatial unit is a hotspot if criminal activity is predictable there, *i.e.* it is not random. Depending on the definition, different techniques for hotspot identification may be employed. Nevertheless, crime data is generally locally scarce, unbalanced and inaccurate, making it challenging to employ predictive modelling techniques and develop visualization tools. By contrast, georeferenced police patrolling data is generally prohibitive in size: the geolocation of a patrol vehicle is generally tracked every few seconds the whole day long, resulting in a massive amount of data.

Although several works suggest planning police patrolling based on hotspots [8]–[10], there is a lack of investigation on the correlations and effectiveness of such an approach. In particular, it is not clear in which situations police patrolling may effectively inhibit crime. This is a challenging problem that involves examining not only huge volumes of crime and georeferenced police patrolling data but also socio-demographic aspects.

In this paper, we present an investigation and/or discussion of the following research questions:

1) how to efficiently represent, in space and time, criminal and police patrolling data?
2) what is the most appropriate spatial aggregation unit for criminal hotspots representation?
3) which classes of models are appropriate to build crime predictors, considering the characteristics of criminal datasets?
4) under which circumstances does police patrolling inhibit criminal occurrence?
5) is police patrolling data a predictor of criminal occurrence?

We propose answers to some of the aforementioned questions, discuss the main challenges of the problem, and describe some work in progress towards developing a robust methodology to represent, visually analyze, and build predictors for criminal activity, considering both criminal and police patrolling spatiotemporal data. As a case study, we use a database from the Military Police of the state of Alagoas, Brazil (PM-AL), which is comprised of both pedestrian robbery and police patrolling georeferenced data from a limited period.

In Section II, we formally describe the problem to be addressed (Section II). Then, the dataset employed in this work is described in Section III, and the proposed approach

is presented in Section IV. Finally, we discuss the major challenges and difficulties to be tackled (Section V).

II. Problem Statement

In this paper, we investigate several problems that require jointly analyzing a spatiotemporal crime dataset ($\mathcal{C}$) and a police patrolling dataset ($\mathcal{P}$), over the same period.

We consider $\mathcal{C}$ to be very sparsely distributed both in space and time since crime occurrences are expected to be concentrated in a few locations of a city. Also, one may not expect a large number of occurrences per location, even in hotspots, thus making the training of predictive models difficult. This dataset is also prone to spatiotemporal inaccuracies [11] because it is a compilation of reports of victims.

Formally, let $\mathcal{C} = \{(c_1, t_1), (c_2, t_2), \ldots, (c_n, t_n)\}$, where c_i is the geolocation (given in latitude/longitude pairs) and t_i is the time (including day, hour and minute) of the i-th crime occurrence. Here, we consider $\mathcal{C}$ to be comprised of a specific type of crime.

Differently, police patrolling datasets are expected to be huge, since their data is automatically collected by sensors placed in police vehicles. Other difficulties include missing data from one or many cars during a specific period; and data collected from idling cars. We define a police patrolling dataset as $\mathcal{P} = \{(p_1, s_1, t_1, v_1), (p_2, s_2, t_2, v_2), \ldots, (p_n, s_n, t_n, v_n)\}$, where p_i is the geolocation of vehicle v_i at time t_i, and s_i is its speed at that moment.

We also consider exploiting external datasets $\mathcal{T}$, which may include, for instance, socioeconomic indicators, georeferenced points of interest and climate data.

Given a set $\mathcal{X}$ of spatial aggregation units of a city, the first relevant problem is to identify, using $\mathcal{C}$, a subset $\mathcal{H} \subset \mathcal{X}$ of crime hotspots, which is expected to be much smaller than $\mathcal{X}$. Then, we restrict crime analysis to $\mathcal{H}$.

The second problem is about training crime predictors $\hat{f}(h, t)$, where $h \in \mathcal{H}$ is a hotspot and t is a time unit representing a time window, *e.g.* a specific day or hour. Such a predictor is expected to be trained using data from $\mathcal{C}$, $\mathcal{P}$ and $\mathcal{T}$, and shall estimate if and/or how many crimes will occur in h during t. As aforementioned, this is not straightforward due to the characteristics of the data, and thus classical predictive modelling approaches do not apply.

The third problem is aimed at developing methods and visualization tools to comprehend the correlations between crime and police patrolling data collected from the same city and period. We consider that these correlations may be spatially independent, in which case an analyst may investigate whether police patrolling is effective to inhibit crime in general (at least for the studied region); or spatially dependent. The latter means that police patrolling could have distinct impacts in different regions of a city, which may be valuable information to assist police patrolling planning. Applications of these correlation studies include finding answers to research questions 4 and 5 (Section I).

It is also worth mentioning that, to make these problems computationally tractable, efficient computational representations and algorithms must be developed. In Section IV we propose compact representations for both crime and police patrolling data that rely on a street corners-based spatial discretization.

III. The Military Police of the State of Alagoas (PM-AL) Dataset

Two datasets were provided by the Military Police of the State of Alagoas (PM-AL) for research purposes, through a cooperation agreement between PM-AL and the Federal University of Alagoas (UFAL): the Passerby Robbery Occurrences in Alagoas Dataset (PROD-AL) and the PM-AL Police Patrolling Dataset (PMAL-PPD).

PROD-AL is a spatiotemporal crime dataset comprised of passerby robbery occurrences in the state of Alagoas. It holds 8,572 occurrences from January to October 2021. Occurrences locations are represented as latitude/longitude pairs, and their times are provided as a day of the year and an hour of the day (minutes are not available).

PMAL-PPD is comprised of roughly 50 million entries per month, where each entry represents spatiotemporal data of a specific police vehicle, *i.e.* its geographical position, the collection day and time of day, and a vehicle id. The geolocation of each car is collected through GPS; transmitted through radio-frequency communication in real-time every few seconds; and stored in a PM-AL database. We noted that, for each car vehicle, geolocations are not uniformly collected over time, but every 10 to 30 seconds. In this work, we consider only data collected from January to October 2021 to match the period of occurrences available in PROD-AL.

To facilitate the analysis in our studies, we also filter the datasets in space to a rectangular area of approximately 40km^2 in the city of Maceió, which includes 4,618 street corners, as shown in Fig. 1. This resulted in a smaller crime dataset with 818 crime occurrences, but the police patrolling dataset is still huge, with tens of millions of entries, which we better discuss in the following section.

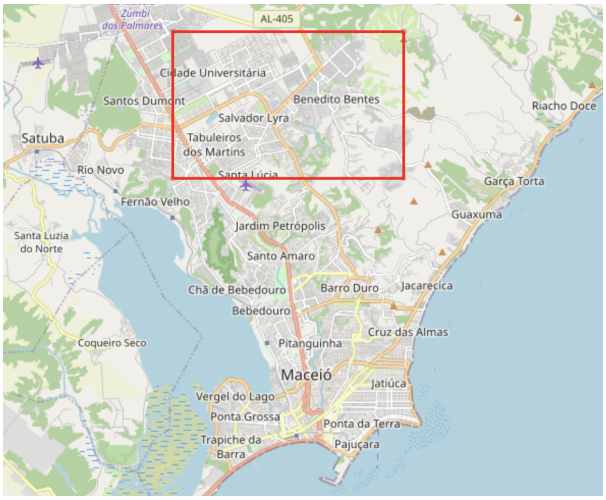

Fig. 1. Rectangular study case area in the city of Maceió, in red.

IV. Proposed approach

To make these challenging problems tractable, we propose an approach made up of computationally efficient data representations, preprocessing algorithms and a hotspot identification technique. These components were validated through simple visualizations on the datasets described in Section III.

A. Spatial representation and preprocessing

As in [3], we adopt a spatial representation based on the city street graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node $v \in \mathcal{V}$ represents a georeferenced city street corner and each edge $e \in \mathcal{E}$ represents a street segment.

Initially, a city street graph is loaded from the OpenStreetMap database. Then, two binary fixed-size time series κ_v and ρ_v are initialized and filled with zeros, for each corner v of the city. These time series represent, respectively, if crimes occurred and if police patrolling vehicles passed nearby. Each bin of the time series represents a time window of one hour (due to the time resolution of PROD-AL). As a result, each time series has a size of 7,296 for the studied period.

Spatiotemporal data from $\mathcal{C}$ are projected to their nearest corner in $\mathcal{V}$. Each crime occurrence $(c_i, t_i) \in \mathcal{C}$ is first projected to its nearest edge $e_n \in \mathcal{E}$ using c_i. Then, the occurrence is projected to its nearest node v_n of e_n. Finally, the bin of the time series κ_{v_n} that corresponds to time t_i is set to 1. Analogously, spatiotemporal data from $\mathcal{P}$ are projected to a node v_n, and the corresponding bins of the time series ρ_{v_n} are set to 1.

It is worth mentioning that this procedure improves the accuracy of the spatial aggregation of the occurrences, in comparison to a naive projection to the nearest node, which may be far from the occurrence location (see Fig. 2). Also, the size of the resulting time series is constant, *i.e.* it does not depend on the size of $\mathcal{C}$ and $\mathcal{P}$. For instance, considering the dataset described in Section III, only 9,236 time series of size 7,296 are required, resulting in approximately 64 megabytes of memory (if 1 byte per bin is used).

In this work, we adopt the CityHub Library [12] to load the city street graph and to project georeferenced data to corners. As a result, in our case study of 818 crime occurrences, 506 of 4,618 corners were associated with crimes.

Fig. 2. Corner projection algorithm: a point of interest p is first projected to the nearest edge e_n (in pink), and then to the nearest extremity v_n of e_n (in pink). If a straightforward projection algorithm was used, point p would be assigned to a corner in the wrong street (Rua Armando Faria Lobo).

B. Hotspot identification

In our preliminary case studies, we followed the hotspot definition of García-Zanabria *et al.* [3]: "micro" places where crimes are relatively stable (or predictable), but not necessarily occur with high intensity. To identify corners with a high probability of crime occurrence, the stationary vector π of a stochastic matrix P is computed, where P_{ij} is an integer representing the number of days that the corners v_i and v_j faced crime events.

A simple thresholding method was applied to detect the hotspots that covered more than a percentual T of the number of occurrences. First, the corners are sorted in descending order according to their probability given by π. Then, corners with higher probability are iteratively inserted into the hotspot set $\mathcal{H}$, until $\mathcal{H}$ covers more than $T\%$ of the occurrences.

In our case study, we set $T = 30\%$. As a result, out of 506 corners with occurrences, 179 corners were identified as hotspots, as revealed in Fig. 3.

C. Visualization

We propose to investigate research questions 3, 4 and 5 by exploring the time series of crime occurrences and police patrolling restricted to the identified hotspots, and here denoted by $\kappa_{\mathcal{H}}$ and $\rho_{\mathcal{H}}$. To this aim, straightforward visualizations have been developed to validate the previous steps and to gain insights.

For example, a scatter plot describing the average patrolling time (in hours) and the number of crimes of each hotspot is revealed in Fig. 4. There is no clear pattern indicating that hotspots with a higher number of crimes are more patrolled. This may indicate that the effectiveness of police patrolling is spatially dependent. Also, this plot does not allow a visual analysis of temporal patterns considering both crime occurrences and police patrolling nearby.

Alternatively, the time series may be concurrently visualized for a specific hotspot, as demonstrated in Fig. 5. By zooming into a time window around a crime occurrence, a clear patrolling time pattern may be observed. However, this

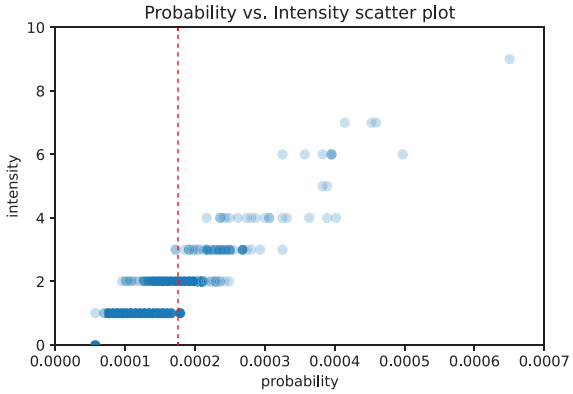

Fig. 3. Scatter plot of the corners associated to crime occurrences: corners to the right of the red dotted vertical are considered hotspots. Low opacity settings were used to enhance the visualization of clusters of corners.

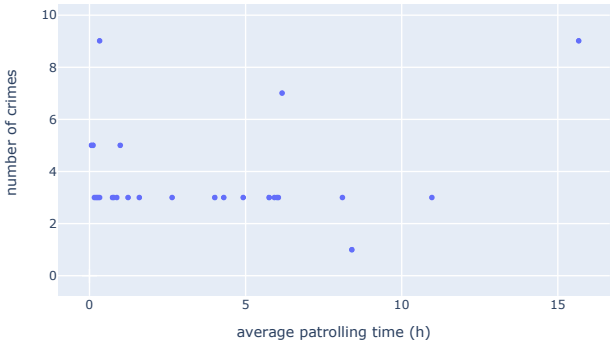

Fig. 4. Scatter plot describing the average patrolling time vs. the number of crimes: each hotspot is represented by a point.

visualization does not allow global analysis of the temporal correlation between crime occurrences and police patrolling.

V. Challenges and Future Work

A. Hotspot identification improvements

Although crime tends to accumulate in micro places [13] such as street corners, we noticed that, in some areas of the city of Maceió, crime is scattered in many neighbouring corners. Besides, crime series for single corners are extremely sparse, with no more than 9 occurrences in a 10 months period (see Figs. 3, 4 and 5). Thus, an adaptive spatial discretization approach to represent hotspots could mitigate the sparseness problem, while still appropriately representing areas where crime occurs in a similar fashion.

Another relevant investigation is related to the detection of hotspots with a high probability of crimes, but low intensity. As revealed in Fig. 3, the approach of [3] essentially resulted in corners with a high intensity of crimes. Our hypothesis is that the sparseness of the $\kappa_{\mathcal{H}}$ time series resulted in very few crimes occurring on the same day. Thus, we aim to investigate more appropriate definitions for the transitions of the stochastic matrix P. Alternatively, different definitions of sites where crime is predictable could be researched.

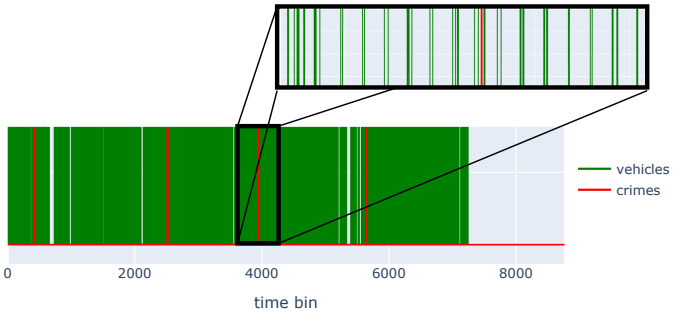

Fig. 5. Simultaneous visualization, for a specific hotspot, of the time series of crime occurrences and police patrolling, including a zoom to finely analyze a specific time window.

B. Crime prediction

Given the time series $\kappa_{\mathcal{H}}$ and $\rho_{\mathcal{H}}$, we aim to develop crime predictors $\hat{f}$, as described in Section II. However, the sparseness of the crime time series makes it a difficult task, where most predictive techniques may not be employed. Variations of Poisson models may be appropriate alternatives that are currently under investigation. Police patrolling data in $\rho_{\mathcal{H}}$ may also be exploited to train better crime predictors and is a manner to answer research questions 4 and 5.

C. Correlation analysis

The uncomplicated visualizations described in Section IV-C are insufficient to solve visual analytics tasks related to research questions 4 and 5. Appropriate correlation analysis techniques and sophisticated visualization tools may be studied. In particular, they must consider the sparsity of the crime time series, temporal information of both κ and ρ, and spatial autocorrelation.

Acknowledgment

The authors would like to thank the Military Police of the State of Alagoas (PM-AL) for providing the datasets.

References

[1] T. C. O'Shea, K. Nicholls, J. Archer, E. Hughes, and J. Tatum, *Crime analysis in America: Findings and recommendations*. US Department of Justice, Office of Community Oriented Policing Services ..., 2003.

[2] S. Chainey, L. Tompson, and S. Uhlig, "The utility of hotspot mapping for predicting spatial patterns of crime," *Security journal*, vol. 21, no. 1, pp. 4–28, 2008.

[3] G. Garcia-ZANABRIA, M. M. M. Raimundo, J. Poco, M. Batista Nery, C. T. Silva, S. F. Adorno de Abreu, and L. G. Nonato, "Cripav: Street-level crime patterns analysis and visualization," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021.

[4] S. Samanta, G. Sen, and S. K. Ghosh, "A literature review on police patrolling problems," *Annals of Operations Research*, pp. 1–44, 2021.

[5] G. Garcıa, J. Silveira, J. Poco, A. Paiva, M. B. Nery, C. T. Silva, S. Adorno, and L. G. Nonato, "Crimanalyzer: Understanding crime patterns in sao paulo," *IEEE transactions on visualization and computer graphics*, vol. 27, no. 4, pp. 2313–2328, 2019.

[6] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," in *International Conference on Computing Science, Communication and Security*. Springer, 2020, pp. 277–289.

[7] S. D. Johnson, K. J. Bowers *et al.*, "Stable and fluid hotspots of crime: differentiation and identification," *Built Environment*, vol. 34, no. 1, pp. 32–45, 2008.

[8] J. Leigh, S. Dunnett, and L. Jackson, "Predictive police patrolling to target hotspots and cover response demand," *Annals of Operations Research*, vol. 283, no. 1, pp. 395–410, 2019.

[9] L. Li, Z. Jiang, N. Duan, W. Dong, K. Hu, and W. Sun, "Police patrol service optimization based on the spatial pattern of hotspots," in *Proceedings of 2011 IEEE international conference on service operations, logistics and informatics*. IEEE, 2011, pp. 45–50.

[10] C. S. Koper, "Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots," *Justice quarterly*, vol. 12, no. 4, pp. 649–672, 1995.

[11] L. Tompson, S. Johnson, M. Ashby, C. Perkins, and P. Edwards, "Uk open source crime data: accuracy and possibilities for research," *Cartography and Geographic Information Science*, vol. 42, no. 2, pp. 97–111, 2015.

[12] K. Salinas, T. Gonçalves, V. Barella, T. Vieira, and L. G. Nonato, "Cityhub: A library for urban data integration," in *2022 35nd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2022.

[13] A. A. Braga, A. V. Papachristos, and D. M. Hureau, "The concentration and stability of gun violence at micro places in boston, 1980–2008," *Journal of Quantitative Criminology*, vol. 26, no. 1, pp. 33–53, 2010.