CounterCrime - Using Counterfactual Explanations to Explore Crime Reduction Scenarios

Marcos M. Raimundo, Germain Garcia-Zanabria, Luis Gustavo Nonato, *Member, IEEE*, and Jorge Poco, *Senior Member, IEEE*

***Abstract*—Analyzing the impact of socioeconomic and urban variables on crime is a complex data analysis problem. Exploring synthetic, correlation-based scenarios using changes in a set of variables could alter a region's definition from unsafe to safe (known counterfactual explanation), which can aid decision-makers in interpreting crime in that region and define public policies to mitigate criminal activity. We propose CounterCrime, a visual analytics tool for crime analysis that uses counterfactual explanations to add insights for this problem. This tool employs various interactive visual metaphors to explore the counterfactual explorations generated in each region. To facilitate exploration, we organize our analysis at three levels: the whole city, the region group, and the regional level. This work proposes a new perspective in crime analysis by creating "what-if" scenarios and allowing decision-makers to anticipate changes that would make a region safer. The tool guides the user in selecting variables with the most significant effect in all city regions. Using a greedy strategy, the system recommends the best variables that may influence crime in unsafe regions as the user explores. Our tool allows for identifying the most appropriate counterfactual explorations at the regional level by grouping them by similarity and determining their feasibility by comparing them with existing examples in other regions. Using crime data from São Paulo, Brazil, we validated our results with case studies. These case studies reveal interesting findings; for example, scenarios that influence crime in a particular unsafe region (or set of regions) might not influence crime in other unsafe regions.**

***Index Terms*—Counterfactual Explanations, Crime Analysis, Visual Analytics Tools, Machine Learning.**

Received 10 October 2024; revised 16 June 2025; accepted 19 June 2025. Date of publication 11 July 2025; date of current version 5 September 2025. This work was supported in part by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grant 2022/09091-8 and Grant 2021/07012-0, in part by the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) under Grant E-26/204.593/2024, in part by the Fundação Getulio Vargas, in part by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) under Grant 307184/2021-8 and Grant 311144/2022-5, and in part by PROCIENCIA under Grant PE501087483-2024. Recommended for acceptance by Y. Zhao. *(Corresponding author: Marcos M. Raimundo.)*

Marcos M. Raimundo is with Fundação Getúlio Vargas, Rio de Janeiro 22250-145, Brazil, and also with the University of Campinas, Campinas 13083-970, Brazil (e-mail: mraimundo@ic.unicamp.br).

Germain Garcia-Zanabria is with the Department of Data Science, University of Engineering and Technology - UTEC, Lima 15063, Peru (e-mail: ggarciaz@utec.edu.pe).

Luis Gustavo Nonato is with ICMC-USP, São Carlos 13566-590, Brazil, and also with New York University, New York, NY 10012 USA (e-mail: gnonato@icmc.usp.br).

Jorge Poco is with Fundação Getúlio Vargas, Rio de Janeiro 22250-145, Brazil (e-mail: jorge.poco@fgv.br).

This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2025.3586202, provided by the authors.

Digital Object Identifier 10.1109/TVCG.2025.3586202

I. Introduction

SEVERAL studies have shown that variables such as population density [1], [2], [3], unemployment rate [4], [5], socioeconomic indicators [6], [7], [8], and even the concentration of bars and bus stops [9], [10], [11], [12] directly affect the crime dynamics in specific locations of a city. Most of these studies focus on analyzing the impact of a few variables on the increase or decrease of crime rates [6], [7], [8] or on how a given variable is associated with the emergence of crime hotspots [13], [14], [15], [16], [17]. However, accounting for only a few variables is insufficient to capture the full complexity of the crime dynamics. Therefore, analytical tools capable of handling multiple data attributes are essential in this context.

Machine learning models are tools designed to capture patterns and dependencies among variables present in data. This is accomplished during the learning process, where a model is trained on a dataset and adjusts its parameters to minimize the discrepancy between its predictions and the actual observed outcomes. As a result, the trained model serves as a proxy for the underlying data distribution, enabling researchers to investigate and interpret the learned relationships. For example, such models can be used to explore the connections between urban physical environments and socio-economic variables [18] or to compare algorithmic predictions with human decisions in domains like criminal justice [19]. This capacity to represent and query complex data structures is fundamental for generating deeper insights into multifaceted phenomena.

To uncover the relationships captured by machine learning models, Explainable AI (XAI) methods aim to make the predictions of "black-box" systems understandable to humans. A variety of XAI techniques have been developed, including popular feature-attribution methods such as LIME [20] and SHAP [21], which provide valuable insights into machine learning models, providing explanations as weights of regression models [22] for imprisonment sentences in assault cases, and using SHAP values [23] for crime prediction. To give more insights into crime analysis, counterfactual explanations [24], instead of identifying influential features as SHAP and LIME, reveal specific changes within the urban and socioeconomic feature space that would result in a shift in the model's classification [25], [26]. Fig. 2Ⓐ illustrates this idea by showing the decision boundary of a classifier, where instances (regions of a city) on the red side are classified as unsafe and those on the blue side as safe. Counterfactual explanations (CFs) correspond to perturbations

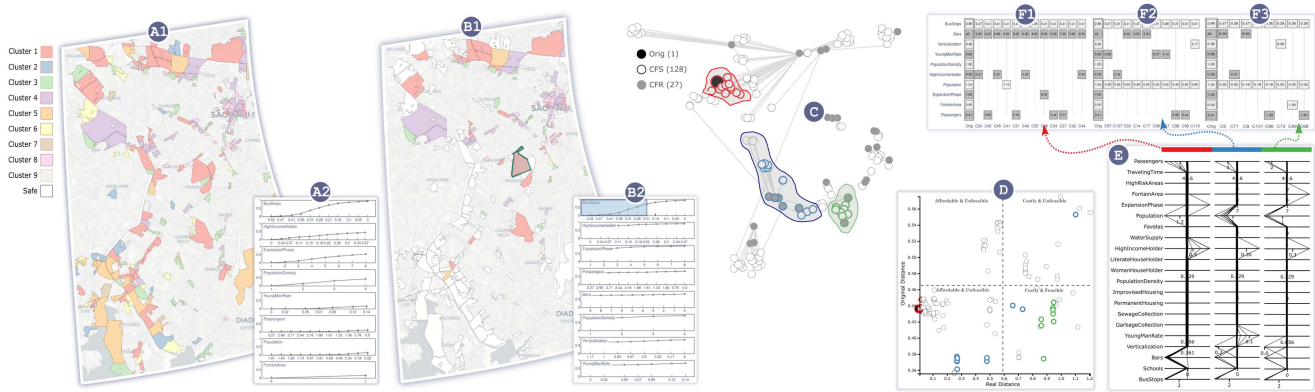

Fig. 1. The proposed counterfactual explanation crime analysis tool, called CounterCrime, is composed of two main parts: Global analysis involving a map (A1, and B1) showing the impact of clustering (in colors) and in white, the effect of filtering (A2, and B2). Local analysis clusters similar CFs for a single region in C and evaluate their costs in D; CFs are inspected in E and F.

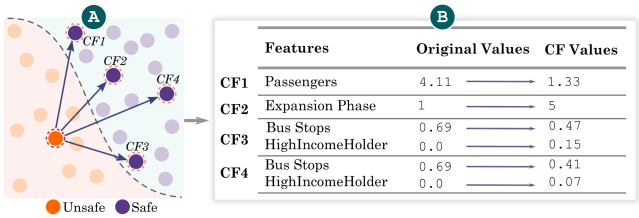

Fig. 2. Illustrative representation of a decision boundary and counterfactual explanations (purple) of a given instance (orange).

in the instances' attributes to move them across the border. The purple points in Fig. 2Ⓐ are four CFs generated from a given sample (orange point). Fig. 2Ⓑ shows the original and perturbed values of the instance's attributes. Note that the same variables can be perturbed differently, resulting in distinct CFs, such as CF3 and CF4, which correspond to distinct perturbations in `BusStops` and `HighIncomeHolder`. In other words, counterfactual explanations identify critical variables and generate diverse hypothetical scenarios that enhance decision-makers' understanding of the factors that could jointly influence criminality.

The example above illustrates the usefulness of counterfactuals as an analytical mechanism. Despite their potential, crime analysis methods based on machine learning [27], [28] have not yet fully explored crime factors (CFs) to investigate critical variables and their relationships with crimes in specific locations within a city. However, it is important to clarify that CFs alone do not establish causality [29], and they should be primarily used as supplementary resources to support experts in decision-making. Our research team has been working on crime analysis for several years, developing visual analytic tools to explore various aspects of crime hotspots [13] and identify crime patterns over a city [30]. Our research examines the relationship between urban and socioeconomic variables and crime, utilizing a predictive model as a proxy for the real-world mechanisms underlying criminal behavior. Counterfactual explanations are employed to emulate diverse scenarios, enabling experts to explore how specific changes to urban features could hypothetically reduce crime rates in particular city locations. By building on a predictive model (Logistic Regression), we can uncover complex correlations between variables and crime, offering a tool for deeper analysis — while explicitly acknowledging that this does not imply a causal relationship.

The counterfactual-based methodology, depicted in Fig. 1, is supported by a visual analytics tool named CounterCrime. This tool facilitates interactive exploration of counterfactual scenarios in crime, as each region can have a set of possible CFs; the proposed tool highlights the variables or groups of variables that may contribute to reducing crime rates. CounterCrime provides recommendations to guide users in selecting and exploring variables. It reveals spatial patterns associated with cluster regions affected by the same set of variables. The system automatically clusters regions with similar behavior, reducing the analytical burden. CounterCrime enables interactive resources, allowing users to select the most appropriate counterfactual explanations.

This work's main contributions are (1) the usage of counterfactual explanations to simulate scenarios and understand which variables may influence crimes in unsafe regions; (2) an exploratory framework that guides users in identifying critical variables linked to unsafe regions; (3) a methodology to explore counterfactual explanations in specific regions, grouping CFs by similarity and determining their feasibility based on comparisons with safe regions; (4) a visualization-assisted analytical tool called CounterCrime that integrates CFs and interactive resources to explore simulated scenarios and understand what may influence crime in unsafe regions; (5) two case studies investigating crime-related phenomena in São Paulo (the largest city in South America) were validated by criminology experts with positive feedback.

II. Related Work

The proposed methodology relates to three main subjects: (i) counterfactual explanations and explainable machine learning, (ii) visualization tools to explainable machine learning and counterfactual explanations, and (iii) crime data analysis.

A. Counterfactual Explanations and Explainable Machine Learning

The literature on counterfactual explanations is vast and includes various applications and methodologies. The use of counterfactual explanations to identify patterns in specific machine learning models is commonly referred to as actionable knowledge. Examples of actionable knowledge methods include heuristics [31], [32], [33], kNN [34], and A∗-like methods [35], [36] to extract a single counterfactual explanation in tree-based classifiers. Additionally, linear-integer optimization methods have been proposed to find single [37], [38] and multiple [39] counterfactual explanations. Furthermore, actionable knowledge has been utilized to answer why-not questions, seeking to understand why particular systems fail to produce adequate query results [40], [41], [42].

Counterfactual explanations are a critical element of explainable artificial intelligence (XAI) methods for both model explanation [43] and recourse actions [39]. The flexibility of CFs allows them to be employed with various types of models [43], [44], [45]. They have proven effective in linear-integer formulations [39], convex optimization [45], [46], and iterative procedures to compute multiple CFs for model explanation [47].

B. Visualization Tools for Explainable Machine Learning and Counterfactual Explanations

Visualization is a practical tool for understanding machine learning models [48], [49], as it provides insightful interpretations of various models, including convolutional neural networks [50] and tree-based models [51], [52]. Additionally, specific tools have been developed to assist in analyzing activation patterns in neural networks [53], [54] and to facilitate comprehension of the learning process of ranking mechanisms [55]. Model-agnostic methods that employ simplification have also been proposed to interpret machine learning models. For example, surrogate models [20] enables the visual inspection of decision boundaries [56], [57]. Some approaches rely on partial dependence plots [58], [59] or Shapley values [21] to explore the significance of features. Furthermore, methods aim to extract and visualize rules from models to understand predictions [60]. Visualization methods for comparing multiple models' predictions can also identify discrepancies [61] and anomalies among models [62].

Counterfactual explanations can also be combined with visualization techniques to interpret machine learning models. These methods aim to answer the question of *which features or groups of features should be adjusted to change a prediction outcome*. Some methods use greedy schemes to change binary or sparse features, such as those used in text classification, to generate counterfactuals [63]. However, most counterfactual-based visualization methods focus on tweaking a single feature [58], finding the closest sample with a different outcome [59], or modifying features to improve a prediction [64]. More closely related approaches also enable the visualization of counterfactual explanations resulting from multiple attribute changes. These include interactive systems like ViCE that facilitate user exploration of multi-attribute adjustments [28], SDA-Vis which uses constrains adjustments to generate counterfactuals across multiple features [65], and methods like DECE designed for investigating specific hypotheses by constraining feature modifications within user-defined intervals [27], More recently, manual, counterfactual modifications were applied to graph neural networks to understand better patient-specific networks, as well as relevance values for genes and interactions [66]; and manual interventions in the projected space of time-series were also performed to achieve counterfactual explanations [67].

C. Crime Data Analysis

Machine learning techniques are increasingly used for crime analysis [68], [69]. These techniques are used to identify high crime rate regions [13] and to understand their relationship with various urban factors [16], [30]. Machine learning methodologies have also been developed for crime forecasting [17]. To conduct these analyses, most techniques rely on identifying crime hotspots [70] and the study of the relationship between crimes and external variables such as socioeconomic and infrastructure factors. According to environmental criminology, the concentration and persistence of crimes in certain locations are not random but, instead, result from the characteristics of those locations [71], [72]. Studies have shown that crime is closely related to population density [1], [2], [3], socioeconomic factors [6], [6], [7], [8], [73], unemployment rate [4], [5], and even the concentration of bars and bus stops [9], [10], [11], [12]. By utilizing machine learning techniques to analyze these factors, it becomes possible to identify patterns and trends that can help predict and prevent crime in high-risk areas.

Crime analysis visualization methods enable the investigation of crime incidence at detailed street-level granularity [15], [74], [75] or coarser scales such as census regions [13]. By illuminating crime dynamics over time, these methods enhance the understanding of crime patterns and trends [30], [72]. Analytical capabilities of visualization-assisted analysis methods range from simple color map tools [76], [77], [78], [79], [80], [81], [82], [83] to more advanced solutions that facilitate linked views and interactive exploratory resources [13], [30], [52], [84], [85], [86], [87], [88], [89].

Our methodology differs from previous work by using a predictive model as a proxy for crime mechanisms to emulate actionable scenarios through counterfactual explanations. Rather than identifying static factors that define unsafe regions, our approach explores specific, hypothetical changes in those factors that could transform a high-crime area into a safer one. The goal is not to prescribe solutions but to enhance understanding of the complex dynamics that influence criminality. Moreover, our method identifies clusters of regions that share similar counterfactual explanations, allowing for the analysis of variables that impact entire groups. This feature supports a richer analytical framework by enabling investigations at both the individual and cluster levels. Consequently, our approach offers a distinct and complementary perspective to techniques such as LIME [20] and SHAP [21]. While those methods primarily rank feature importance, our focus on feature-level change scenarios shifts

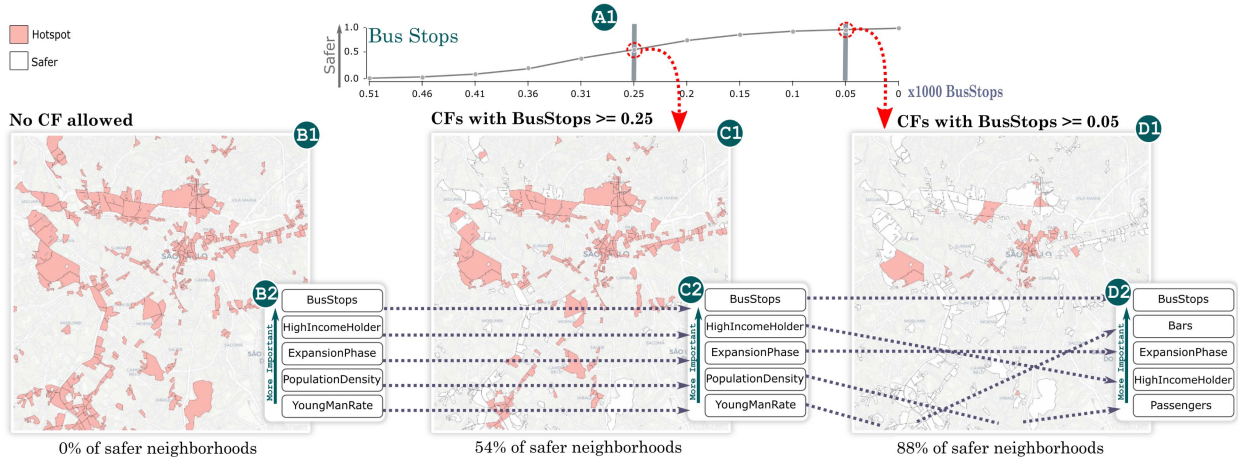

Fig. 3. A case study investigating the impact of allowing counterfactual explanations that reduce the number of `BusStops` in criminality. (A1) indicates the proportion of regions safer with that filtering for each value. (B1), (C1), and (D1) show the regions are already safe in white and the crime hotspots in red. (B2), (C2), and (D2) show the most important variables for finding counterfactual explanations for the remaining regions.

the analytical question from *Which features matter most?* to *What kind of change in which features could lead to a different outcome?*

III. Motivating Case Study

Numerous methods exist to analyze data and determine the factors contributing to high crime rates in a given region [76], [77], [78], [79], [80], [81], [82], [83], [90], [91]. While one popular approach involves identifying critical variables by grouping hotspots with similar behavior, a more complex and underexplored issue is understanding how modifying these variables influences crime in unsafe regions. To address this challenge, consider a scenario using urban and socioeconomic variables to train a model that classifies areas as safe or unsafe. With this model, we can examine how certain variables must change to transform regions classified as dangerous into regions classified as secure. This is precisely where counterfactual explanations become relevant.

Take a look at the scenario presented in Fig. 3. The line chart in (A1) illustrates the proportion of regions (y-axis) that will be classified as safe if the number of bus stops decreases (x-axis) relative to the regions initially categorized as dangerous. This curve is computed based on counterfactual explanations. The red regions in (B1) correspond to originally unsafe areas. By reducing the number of `BusStops` to at most 250 and 50 in (C1) and (D1), respectively, 54% and 88% of the regions become classified as safe (white polygons on the map indicate the regions that become safe). It is worth noting that the sign "$>=$" above the maps denotes "at most," implying that for some regions, a slight decrease in bus stops could classify them as safe. In contrast, others may require a more drastic reduction.

Fig. 3(B2), (C2), and (D2) present a ranking of variables based on their impact on changing the classification to safe. The results show that BusStops consistently has the most significant impact on changing the classification of regions from unsafe to safe. The ranking of variables can be derived from counterfactuals, as explained in the following sections. Additionally, the order can change when adjusting the CF threshold, as shown in Fig. 3(D2), where `Bars` becomes the second most important variable after considering CFs greater than 50 for `BusStops`. Therefore, if reducing the number of `BusStops` beyond 50 is not feasible, the next option would be to decrease the number of `Bars`. It is worth noting that these emulated scenarios allow for observing the actual numerical impact of variables on the model's classification. Any real-world changes should be made in collaboration with experts to validate their effect on crime and consider other factors, such as transit.

This case study highlights the value of counterfactual explanations as an analytical resource. In this context, CFs provide insights into the relationships the model has learned between specific variables and crime, as well as how changes to these variables can help understand what, according to the model, may influence crime in unsafe regions.

IV. Design Requirements

Our research team has valuable experience working with crime analytics professionals, which has helped us understand the main difficulties faced in this field. We conducted a comprehensive survey of the literature on crime analysis and explainable machine learning to properly design our visual analytics system. Our findings revealed that many authors emphasized linking urban and socioeconomic variables with crime [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12]. Furthermore, we identified the need to analyze crimes at three different levels of detail: local (specific locations) [13], [92], group (clusters of regions with similar characteristics) [93], and global (the entire city) [16]. Interestingly, many authors also stressed the importance of providing model explanations at these same three levels of detail [27], [52], [59], [94], [95]. Therefore, our visual analytic system, *CounterCrime*, was designed to address our analytical needs and meet the demands of several other authors. Considering the diverse challenges faced in crime analytics, we believe that *CounterCrime* will be a valuable tool for professionals in this field.

Given that regions can be classified as safe or unsafe based on a classification model and that there are multiple and diverse counterfactual explanations for each region (for further information on how these are computed, please refer to the following section), we aim to develop an analytical tool that can explore and analyze the counterfactuals associated with either a single region or a cluster of regions that share similar counterfactuals. The primary objective of this tool is to identify the variables that have the most significant impact on classifying regions as unsafe. Furthermore, the tool must enable users to modify these variables and observe the resulting changes in the classification of regions, transitioning from unsafe to safe. We have established a set of primary requirements that the tool must meet to achieve these objectives.

R1. Analysis of Multiple Counterfactual explanations: Each region has many associated counterfactual explanations, making it essential to have enabling resources for visualizing and analyzing sets of counterfactuals linked to unsafe regions.

R2. Identification and Modification of Relevant Variables: The system must provide mechanisms to identify the variables that most frequently appear in effective counterfactual scenarios and allow users to modify their values. For example, it should be able to answer questions such as: Which regions experience a change in crime classification to "safe" when specific variables are modified? How many regions undergo such a classification change when the values of certain variables are altered?

R3. Clustering Regions with Similar Counterfactual explanations: Analyzing counterfactual explanations for each region can be challenging, especially in large cities. Therefore, it is highly desirable to cluster regions with similar counterfactual explanations to explore them together, revealing "global" patterns. Additionally, the system must evaluate whether certain counterfactuals are realistic.

R4. Seamless Data Science Workflow Integration The analytical framework should integrate seamlessly with standard data science workflows, offering a user-friendly environment that supports intuitive exploration and experimentation.

Based on the requirements outlined above, we have identified a set of tasks that the analytical tool needs to be able to perform:

T1. Visualize and explore counterfactual explanations: A key requirement is the ability to analyze counterfactual explanations (R1). The tool should enable users to visualize and explore counterfactual explanations for single and clustered regions, displaying which regions are impacted by particular CFs and how (R2 and R3).

T2. Rank CFs and variables: The tool must rank variables based on their prevalence and impact within the generated counterfactual scenarios. This guide highlights the variables most often involved in successful re-classifications from unsafe to safe (R2).

T3. Cluster regions: An important requirement is the ability to analyze sets of regions with similar CFs (R3). To facilitate cluster-level analysis, the tool should be able to cluster unsafe regions with similar CFs.

T4. Show spatial distribution of regions: Depicting the spatial distribution of unsafe regions is important for analyzing the impact of counterfactual explanations (R1, R2, and R3).

T5. Cluster similar counterfactual explanations: Since each region has multiple associated CFs, organizing the CFs based on similarity can facilitate the exploratory process (R1).

T6. Assess the feasibility of CFs: Since CFs generate simulated scenarios, it is essential to determine whether certain CFs are realistic. To accomplish this, the tool must compare CFs with the real attributes of safe classified regions and perform this task (R3).

T7. Provide the capability of integration with data science tools: The tool will be packaged for seamless integration into Jupyter Notebooks, ensuring compatibility with standard Python-based data science workflows.

Our team's experience analyzing crime-related phenomena inspired the requirements and tasks described above. To fulfill these tasks, we developed CounterCrime, a Jupyter Notebook toolbox (described in Section VI). Specifically, CounterCrime analyzes counterfactual explanations associated with unsafe regions identified by a machine learning model.

V. Counterfactual Explanations

This section discusses the mathematical and computational foundations underlying the proposed visual analytic tool. In our context, each data instance corresponds to a spatial region in São Paulo and is given a classification model. We will associate each region classified as unsafe with counterfactual explanations, which are the basis of our analysis. Before detailing how we compute the CFs, we describe the dataset used to train the model and how the classification model has been settled.

A. Data Set

The 18,953 São Paulo census tracts are the spatial regions classified as safe or unsafe. Socioeconomic, urban, and historical crime data are aggregated in each region. The Center provided crime data from 2006 to 2017 for the Study of Violence - the University of São Paulo (nev.prp.usp.br), which carefully assembled, curated, and cleaned the datasets. We have a deep partnership with them to analyze and comprehend the origins and characteristics of data, aiming at a less biased analysis. The total number of crimes from 2006 to 2017 is assigned to each region. Urban infrastructure data, such as the location of schools, bus stops, and bars, were provided by the Center for Metropolitan Studies (centrodametropole.fflch.usp.br); housing, sanitary conditions, and population profile are obtained from the 2010 Brazilian census IBGE (ibge.gov.br) Finally, São Paulo's subway system provided the urban mobility data (transparencia.metrosp.com.br). All the data are georeferenced according to the census tracts.

B. Identifying Hotspots Using Urban and Socioeconomic Features

Each census tract corresponds to an instance of data, whose attributes include urban and socioeconomic variables. Crime data serve as the dependent variable used for training the classifier. Specifically, for the case studies in this paper, we selected the 500 regions (out of 18,953) with the highest number of crime events,

labeling these regions as zero and the remaining ones as one. This threshold (which can be adjusted by the user) corresponds to approximately 2.5% of the total and was chosen because these regions are sufficiently diverse in terms of associated attributes, thus exhibiting different patterns. At the same time, selecting 500 regions avoids extreme class imbalance, making it feasible to train the model directly — i.e., without the need for techniques to handle imbalanced data.

Creating the model: We use a Logistic Regression as the classification model. The model was trained to hold 20% of the data for testing, relying on 5-fold cross-validation to select the parameters and l_1 regularization. The model's performance was 0.90in AUC. It is worth mentioning that despite the Logistic Regression being chosen thanks to good performance in preliminary experiments, the methodology is model agnostic.

Defining hotspots: Logistic Regression was used to select the 500 regions with the lowest probability of being classified as safe, computing counterfactual explanations for these regions.

C. Counterfactual Explanation Computation

Given a sample $\mathbf{x}$, counterfactual explanations consist of creating synthetic samples whose classification differs from the classification of $\mathbf{x}$. Since we aim to explore scenarios of change for unsafe regions, we compute counterfactual explanations that classify a region as safe.

Mathematically, given a sample $\mathbf{x}$ and a decision function $r(\mathbf{x})$, we want to find a new sample $\overline{\mathbf{x}}$ such that $r(\mathbf{x}) < \tau$ and $r(\overline{\mathbf{x}}) \geq \tau$, where τ is a given threshold, and the new sample $\overline{\mathbf{x}}$ must be as close as possible to the original sample, that is, $\overline{\mathbf{x}} \approx \mathbf{x}$. There are a variety of counterfactual explanations that can lead to the desired result. Therefore, our approach seeks to compute multiple and diverse counterfactual explanations [43]. Specifically, we rely on MAPOCAM [96], an *a posteriori* multi-objective optimization algorithm, to find a set of counterfactual explanations.

Let $f_i(\overline{\mathbf{x}}) = |\overline{x_i} - x_i|, \forall i \in \{1, \ldots, m\}$ be a cost function associated with ith variable, where m is the number of variables and $\overline{\mathbf{x}_i}, \mathbf{x}_i$ are the ith component of $\overline{\mathbf{x}}$ and $\mathbf{x}$, respectively. Given two CFs $\overline{\mathbf{x}}^{(1)}$ and $\overline{\mathbf{x}}^{(2)}$ if there exists $f_j(\bullet)$ and $f_k(\bullet)$ such that $f_j(\overline{\mathbf{x}}^{(1)}) < f_j(\overline{\mathbf{x}}^{(2)})$ and $f_k(\overline{\mathbf{x}}^{(1)}) > f_k(\overline{\mathbf{x}}^{(2)})$ it is impossible to assign an order relation between $\overline{\mathbf{x}}^{(1)}$ and $\overline{\mathbf{x}}^{(2)}$. The order relation is only feasible if a solution is better (or worse) for all cost functions. An *a posteriori* multi-objective optimization method aims to find a good representation of all solutions such that no other solution is better for all objectives; these solutions are called Pareto-optimal solutions. The MAPOCAM [96] algorithm, detailed in Appendix A, available online, is a model-agnostic scheme capable of finding multiple Pareto-optimal counterfactual explanations.

More importantly, when the change in feature is used as an objective in MAPOCAM, any other user preference can be satisfied within the generated set of counterfactual explanations [96]: this is a stepping stone to CounterCrime since we can trust that any advantageous counterfactual explanation is already accessible by the methodology. No other method surveyed has these properties. The main limitation of MAPOCAM is to find CFs in a single direction, increasing or decreasing each variable's value.

D. Ranking Variables Based on Their Importance

Considering the set of regions $\mathcal{H}$ classified as unsafe, for each $h \in \mathcal{H}$ we have a set of counterfactual explanations $\mathcal{C}_h$ associated with h. The counterfactual explanations have a sparse representation in the sense $|a_i \neq 0, \forall i \in \{1, \ldots, m\}| \ll m$, where $a_i = |x_i - \overline{x}_i|; \overline{\mathbf{x}} \in \mathcal{C}_h$. In other words, just a few attributes of $\overline{\mathbf{x}}$ differ from those in $\mathbf{x}$. Therefore, it is essential to identify the variables that jointly are more likely to achieve a counterfactual. To determine the variables that operate "together" to achieve a CF, we build a stochastic matrix whose entries correspond to the probability of selecting a new variable given that we already picked another one. The stationary state of the matrix (Perron eigenvector) point out the important variables.

In mathematical terms, given the counterfactual explanations $\mathcal{C}_h$ of a region h, let c_{ij} be the co-occurrence index indicating the number of times that $a_i \neq 0$ and $a_j \neq 0$. The entries in each stochastic matrix P^h are given by $P^h_{ij} = \frac{c_{ij}}{\sum_{j=1}^{d} c_{ij}}$, what ensures that $\sum_{j=0}^{d} P^h_{ij} = 1$. The stationary eigenvector π of P^h satisfies $\pi P^h = \pi$, and each entry π_i indicates the importance of the variable i when computing CFs for the region h. Sorting the entries of π, we get a ranked list of the most relevant variables considering the CFs. In other words, the variables tend to be concomitantly present in the CFs. To compute the variables' importance of a set of regions, we average the stochastic matrices of the regions and compute the stationary eigenvector of the averaged matrix (ensuring the sum of each row is one).

E. Clustering Regions Based on Their Counterfactual Explanations

One of the main goals of this work is to find counterfactual explanations for a set of regions. The idea is to find clusters of regions that share similar counterfactual patterns.

Assuming that similar regions (in terms of their counterfactual explanations) tend to have similar stochastic matrices, we employed k-means using the Frobenius Norm of the difference between Stochastic Matrices (Section V-D). We empirically set $k = 9$ in our implementation, observing the Clustering Comparison View. This number of clusters enabled a large diversity of patterns, preserving coherence among the regions in the same cluster.

VI. CounterCrime

Based on the requirements outlined in Section IV, we have created *CounterCrime*. This powerful visual analytic tool thoroughly explores counterfactual explanations. Fig. 4 showcases the CounterCrime system, which is comprised of seven essential components: Ⓐ *CFs' Recommendation Filter* enables users to explore counterfactual explanations on each variable. Ⓑ *Map View* visualizes the regions and clusters of regions. Ⓒ Cluster Comparison displays the importance of variables for each cluster of regions. Ⓓ *CFs' Projection*, a 2D visual representation of

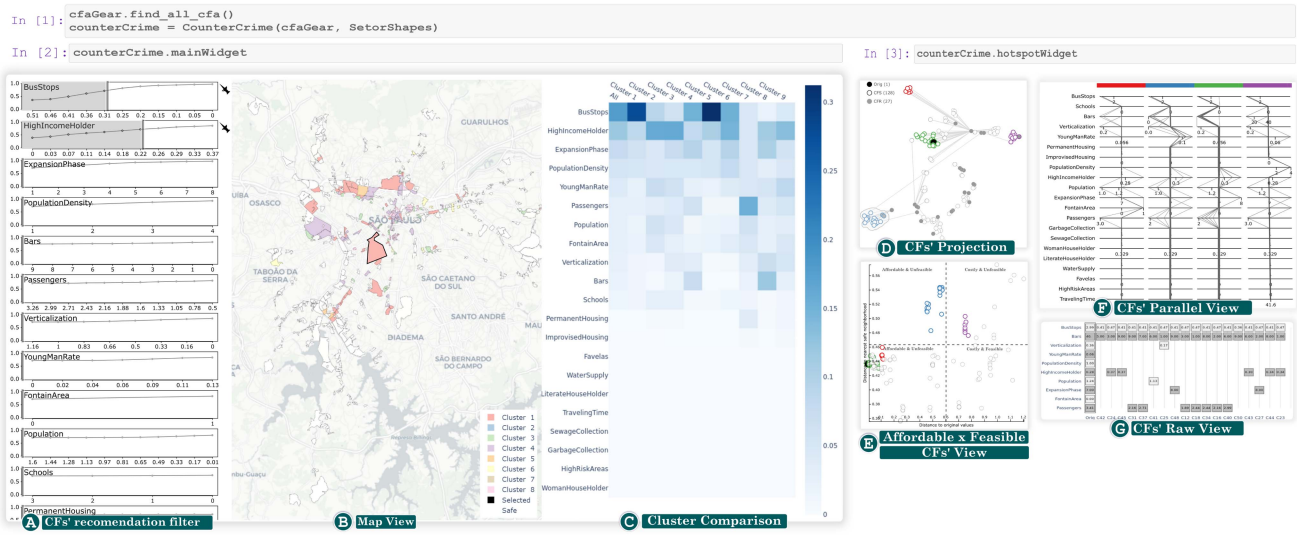

Fig. 4. The proposed counterfactual explanation-based crime analysis tool, called CounterCrime, is composed of 7 views: Ⓐ CFs' Recommender Filtering, Ⓑ Map View, Ⓒ Cluster Comparison. We can see that Ⓐ allows counterfactual explanations that reduce `BusStops` to at most 0.31 (thousands) and increases `HighIncomeHolder` to at most 0.22, the regions in Ⓑ that are now safer are labeled as white. After that, we can observe CFs from a specific region by clicking on the map. Ⓓ shows the CFs' Projection that can be selected to show overview values in Ⓕ the CFs' Parallel View, and it can be more thoroughly investigated in Ⓖ CFs' Raw View. The Affordable × Feasible CFs' View in Ⓓ shows how easy and reasonable it is to make a change that creates the CF.

TABLE I
VISUALIZATION COMPONENTS AND REQUIREMENTS ATTENDED (SEE SECTION IV)

Visual Component	T1	T2	T3	T4	T5	T6	T7
CFs′ recommendation filter	✓	✓					✓
Map view			✓	✓			✓
Cluster Comp.			✓	✓			✓
CFs′ Projection					✓	✓	✓
Affordable×Feasible CFs′ View					✓	✓	✓
CFs′ Parallel View							✓
CFs′ Raw View	✓						✓

CFs. Ⓔ *Affordable × Feasible CFs' View* is a scatter plot presenting the distribution of counterfactual explanations based on the distance to their corresponding original instance and closest safe region. Ⓕ *CFs' Parallel View* showcases the changes of a cluster of counterfactual explanations. Ⓖ *CFs' Raw View* displays CFs and corresponding original values for a given region. Table I demonstrates the relationship between each visual component (listed in the first column) and the associated tasks (T1-T7 columns).

A. CFs' Recommendation Filter

This component is a selector that allows users to choose a counterfactual value s_i for a variable i in question, as depicted in Fig. 4Ⓐ. To simplify the concept, let us assume that we want to decrease the value of a variable x_i to a desired value s_i. Fig. 4Ⓐ shows counterfactual explanations with `BusStops` ≥ 0.31 and `HighIncomeHolder` ≤ 0.22. Fig. 5 shows the impact of these changes in three unsafe regions. While CF CF1 and CF5 in Region ① and C1 in Region ③ satisfy the chosen threshold, no CF in Region ② fulfills the filter (although CF5 satisfies the condition of `HighIncomeHolder` $= 0.22$, the `BusStops` $= 0.16$, which violates the filter values).

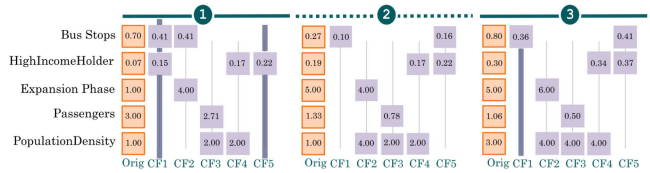

Fig. 5. *CFs' Recommendation Filtering* filters CFs with `BusStops` ≥ 0.31 and `HighIncomeHolder` ≤ 0.22. Three unsafe regions with their original values (columns "Orig") and 5 corresponding CFs (columns CF1 to CF5) which offers changes that would make these regions safe. Based on the filtering thresholds (0.31 and 0.22, respectively), counterfactuals CF1 and CF5 in region ① and CF1 in region ③ attend the filter; thus, they were selected for posterior analysis.

Counterfactual brushing: This feature selects and presents counterfactual explanations based on certain thresholds for each variable. The threshold selection is restricted to changes in a single direction since the MAPOCAM only generates CFs with this characteristic. Both the *CF Recommendation Filter* and the *Map View* are influenced by the threshold selection and automatically adjusted to reflect the filtering. The *CF Recommendation Filter* reorganizes the variables according to their importance. It also updates the internal curves representing the number of regions they will impact. *Map View* is also updated by changing the colors of the affected clusters to white, indicating that those regions are now classified as safe.

B. Map View

In Fig. 4Ⓑ, the visual component displays unsafe regions across the city based on a discretization using census units. The color legend on the right indicates the classification of each cluster of regions. When one or more CFs are identified through the *CF Recommendation Filter*, the affected regions are changed to be rendered in white. This choice aims to visually blend these

counterfactually altered regions with the map's other, inherently safe, non-outlined regions.

Region Selection: Click on that region to explore CFs in a specific region. This action updates all views (except for the *CF Recommendation Filter*, *Map View*, and *Cluster Comparison*, which are global views) to reflect the data associated with the selected region. The selected region is highlighted with larger black borders.

C. Cluster Comparison

This heatmap (see Fig. 4Ⓒ) visually represents the impact of each variable in each cluster. The darker cells indicate the most important variables in the cluster. The heatmap is derived from the stochastic matrix associated with each cluster. For more details, please refer to Section V-D.

D. CFs' Projection

After computing counterfactual explanations for an individual region, a multidimensional projection technique is employed to project them onto a 2D visual space, as shown in Fig. 4Ⓓ. Each counterfactual is represented as a circle, and its position is calculated using t-SNE projection. The black point corresponds to the original variables of the region, while the CFs are represented as white points. Additionally, gray points represent safe regions that are closest to a CF. Lines connect each counterfactual to its nearest safe region, with longer lines indicating less feasible counterfactuals in real-world scenarios. The length of these lines represents a feature distance — specifically, the maximal percentile distance between the original region and its counterfactual. For each feature, this distance is calculated as the proportion of regions with values lying between those of the original and counterfactual regions. The maximal percentile distance is then defined as the largest such value across all features. The legend on the left displays the label of the CFs and the number of elements in each cluster. At the same time, the colors in the borders correspond to the selection explained below.

Cluster Selection: By lasso selection, filtering a cluster of CFs is possible. The selected elements are clustered and can be analyzed using *CFs' Parallel View*.

E. Affordable × Feasible CFs' View

Fig. 4Ⓔ displays the *Affordable* × *Feasible* scatter plot, where each counterfactual is represented as a circle. The position of each circle represents the counterfactual's feasibility, with the x-axis indicating the normalized distance between the counterfactual instance and its associated original instance and the y-axis representing the distance from the CF to the closest safe region in the original data.

Each quadrant of the *Affordable* × *Feasible CFs' View* has a distinct interpretation. The desired CFs lie in the lower-left quadrant since they are close to the original values and a region classified as safe. The lower-right quadrant represents CFs that are more difficult to achieve (farther from the original value) but still feasible in the real world since they are close to a region classified as safe. CFs in the upper-left quadrant are close to their host instance. However, their feasibility in the real world is uncertain since they are far from the nearest region classified as safe. The upper-right quadrant contains CFs that are not desirable, hard to achieve, and far from a region classified as safe, making them the least favorable.

Cluster Selection: Lasso selection allows filtering a subset of CFs that meet specific criteria. The selected elements can then be clustered and analyzed using the *CFs' Parallel View*.

F. CFs' Parallel View

To more effectively analyze clusters that have been interactively selected using the *CFs' Projection* and *Affordable* × *Feasible View*, we have created a visual representation demonstrating each variable's changes (see Fig. 4Ⓕ) for the selected cluster. Each cluster is depicted using vertical parallel coordinates, with the polylines representing each counterfactual. The position of each line in the coordinate system reflects the magnitude of the change for that particular variable. The center of the coordinates represents the original value, with the orientation to the left indicating a decrease from the actual value and the direction to the right showing an increase.

G. CFs' Raw View

The view shown in Fig. 4Ⓖ provides detailed information about the counterfactual explanations of a particular region. The first column displays the original values for each variable, while the remaining columns correspond to the counterfactual explanations. Each cell within the table indicates the value needed for that variable to achieve the desired outcome.

H. Implementation Details

CounterCrime is a Jupyter-Notebook system that utilizes the Widget framework and a Python3 visual library. The system includes modules for computing and visualizing CFs. Sklearn [97], Pandas [98], and Numpy [99] Python libraries were employed to calculate Logistic Regression, CFs, stochastic matrix, census block clusterization, and filtering. Visualization resources were developed using Plotly with Python interface widgets for geo-map representations, cluster heatmaps, *CFs Raw View*, and choropleth maps. D3.js [100] was used to create the projection scatter plots, parallel coordinates, and line charts. Each visualization metaphor was implemented as a class communicating with other classes through callback functions. Finally, we have developed a bi-directional communication channel between Jupyter-Notebook and CounterCrime to manage the combination of Python libraries and visualization tools. To facilitate testing of the system, a Docker container has been made available in the Supplemental Material.

VII. Evaluation

In this section, we present two case studies that, informed by our team's substantial experience in criminology, employ the proposed methodology to analyze scenarios of crime reduction

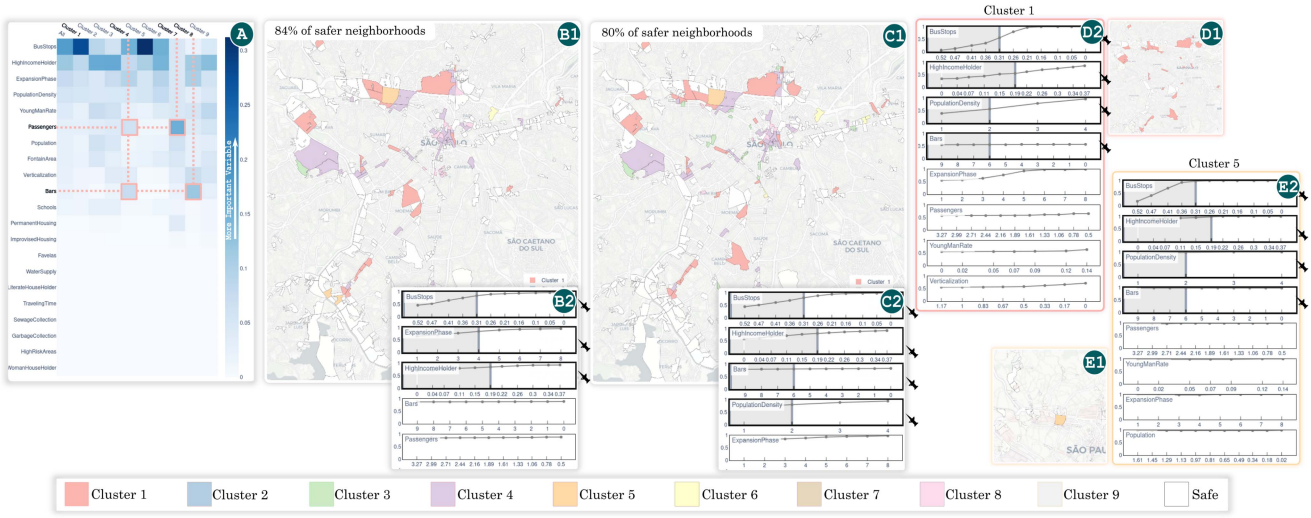

Fig. 6. Case study investigating crime in the whole city. (A) shows the Hotspots' Cluster Comparison that guides the investigation in a general and local manner. (B1) and (C1) show the result in Map View of different filtering selections of the CFs' Recommender Filtering in (B2) and (C2). (D2) and (E2) show the CFs' Recommender Filtering constrained to Clusters 1 and 5 (depicted in Map View in (D1) and (E1)).

using real crime data from São Paulo, Brazil. The first case study focuses on a global analysis of CFs for multiple regions and addresses analytical tasks T1, T2, T3, and T4. The second case study, on the other hand, focuses on a single region and its CFs, addressing analytical tasks T5 and T6.

A. *Case Study 1: Reducing Crime in the Whole City*

This study evaluates the impact of allowing counterfactual explanations with multiple variables to analyze what may influence crime on hotspots (unsafe regions) across the city. The study also aims to understand how different variables affect other parts of the city or region's clusters.

To conduct this case study, we use the system to analyze changes that influence crime on hotspots of the entire city (as shown in Fig. 6). First, we analyze the most critical variables across the city in *Cluster Comparison* (A). The analysis reveals that `BusStops`, `HighIncomeHolder`, and `ExpansionPhase` are the most critical global variables (top variables). However, despite not significantly impacting the whole city, we observe that `Passengers`and `Bars` are essential for Cluster 7 and 8, respectively, and both are important for Cluster 4. As described in Section VI-A, the *CF Recommendation Filter* ranks the variables by their importance by allowing counterfactual explanations using brushing. Brushing variables allows the counterfactual explanations in that range to re-rank the remaining variables using dangerous regions. With this mechanism, (B2) shows brushing in the variables `BusStops`, `ExpansionPhase`, and `HighIncomeHolder` using the *CF Recommendation Filter*. We chose to allow CFs that would not change too aggressively. The white regions in (B1) indicate that `BusStops`, `ExpansionPhase`, and `HighIncomeHolder` could reduce the number of regions classified as unsafe in 84% of the city. Notably, `Bars` and `Passengers` are the most critical non-selected variables for the remaining hotspots (fourth and fifth line charts in (B2)). As mentioned earlier, these variables are essential to Clusters 4 (purple) and 8 (pink), which remain with many regions classified as unsafe.

However, the `ExpansionPhase` variable cannot be changed as it indicates the urbanization period of a region. As a result, we have decided not to allow counterfactual explanations with this variable. After applying the *CF Recommendation Filter*, we selected `Bars` and `PopulationDensity` as two significant variables. The white regions in (C1) indicate similar regions classified as safe due to CFs. However, a majority of the unsafe regions still belong to Cluster 1 (red), 4 (purple), and 8 (pink). *Cluster Comparison* (A) reveals that `BusStops` and `HighIncomeHolder` are not essential for Clusters 4 and 8. In contrast, `BusStops` has high importance in Cluster 1 (darker blue indicates higher significance).

The *CF Recommendation Filter* displays lines for each variable that show the proportion of regions classified as safe (y-axis) when a variable is brushed (x-axis). The increase or decrease in safety is determined by the extent of CFs allowed from other variables. Two distinct scenarios are analyzed - the *CF Recommendation Filter* and *Map View* for Clusters 1 (the red one, with most of the hotspots) and Cluster 5 (the orange one, with almost all regions classified as safe). For Cluster 5, it is observed from the lines in (E2) that safety can be improved by reducing `BusStops` to 0.31, resulting in almost all regions being classified as safe (the line approaches 1 after 0.31, with a rapid decrease before that). No other variable shows a significant loss (all variables remain at the same level regardless of the x-axis value). In contrast, for Cluster 1, the lines in (D2) show that the number of regions classified as safe (y-axis) changes similarly by moderately allowing CFs in any of the three most relevant variables. Except `Bars` (x-axis change does not affect y-axis), the top three ranked variables significantly impact when their values are moderately brushed. Therefore, these three variables must be considered to determine an optimal safety configuration for this cluster. This conclusion is supported by previous research on the impact of income [6], [73], [101], bars

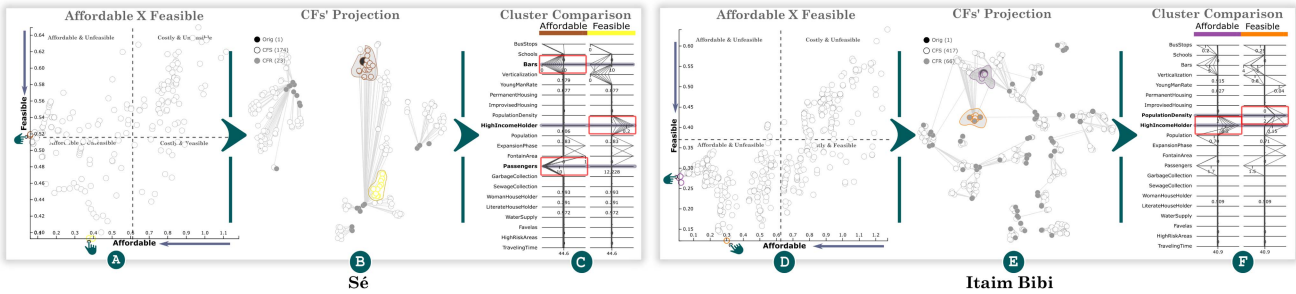

Fig. 7. Case study investigating single regions. Ⓐ and Ⓓ show CFs' Projection with two selections that represent the most affordable and feasible counterfactual explanations on Ⓑ and Ⓔ. Ⓒ and Ⓕ overview the selection using CFs' Parallel View.

or liquor stores [11], [12], [102], [103], bus stops [104], [105], [106], and population [1], [2], [107]. The proposed tool helps emulate, according to the model, the combination of factors that are most influential in the scenarios associated with each region, cluster, and city under study.

Despite the *CF Recommendation Filter* presenting the most critical variables for the remaining hotspots, specific regional cases require different strategies. Moreover, using a *CF Recommendation Filter* for specific clusters can drastically reduce the exploration time.

B. Case Study 2: Investigating Counterfactual Explanations on Specific Regions

This study aims to analyze the counterfactual explanations in specific regions that could not change the classification to safe when exploring counterfactual scenarios by clusters. Here we will evaluate the affordability (i.e., how easy it is to change from the current state) and the feasibility (if there are any records of these changes in another region) of the found counterfactual explanations. From now on, we refer to Fig. 7 for this case study.

Analyzing the Sé region in São Paulo: We selected the first region using the selection tool in *Map view*. It consists of a region called Sé, which corresponds to cluster 4 in Fig. 6 (first row in Fig. 7). Sé is located in the central area of São Paulo with an intense flow of people, a reduced number of local population, and high crime rates.

We use the *Affordable* × *Feasible CFs' View* Ⓐ to select the most affordable (brown points) and the most feasible (yellow points) counterfactual explanations for the Sé region are the points with the lowest value on the X and Y axes, respectively. Then, we use the *CFs' Projection* Ⓑ to find other counterfactual explanations similar to those selected. We do this so that *CFs' Parallel View* Ⓒ shows how the variables vary in the selected counterfactual explanations.

Analyzing the *CFs' Parallel View* Ⓒ, we can see that (i) the most affordable counterfactual explanations (brown selection) are defined by a decrease in the number of `Bars` and `Passengers`; and (ii) the most feasible counterfactual explanations (yellow selection) are defined by an increase in the `HighIncomeHolder` variable.

Analyzing those counterfactual explanations, we find that: The first cluster of counterfactual explanations is not feasible because, in downtown areas of a big city like São Paulo, people pass through going to work and other activities, including going to bars at night. Then, applying this kind of public policy, reducing `Bars` and `Passengers` would not have a basis in reality. The second cluster is a better counterfactual scenario because Sé has good infrastructure, a common trait in safe areas. However, Sé is also a poor region; thus, increasing the income of local people who live there would also increase the probability of that region being classified as safe.

Analyzing the Itaim Bibi region in São Paulo:

We selected the second region using the selection tool in *Map view*. It consists of a region called Itaim Bibi from cluster 8 in Fig. 6 (Second-row in Fig. 7). Itaim Bibi is a rich and sophisticated region known for its corporate headquarters and intense nightlife.

We proceed in the same way as the previous region, using the *Affordable* × *Feasible CFs' View* Ⓓ and then the *CFs' Projection* Ⓔ to select clusters with the most affordable (purple selection) and most feasible (orange selection) counterfactual explanations. Analyzing the *CFs' Parallel View* Ⓕ, we can see the following: (i) an increase in the `HighIncomeHolder` defines the most affordable counterfactual explanations (purple selection); and (ii) an increase in the `PopulationDensity` variable defines the most feasible counterfactual explanations (orange selection).

Analyzing the selected clusters, we can conclude that: We cannot ensure whether it is a good or bad set of counterfactual explanations. It would require a deeper analysis to understand why increasing income in a rich region would be affordable. However, we do not find this variable in the feasible counterfactual explanations because it is hard to find wealthier regions with the same characteristics as this region. The second group of counterfactual explanations effectively addresses the issue, as Itaim Bibi, a wealthy region with low population density, would attract criminals. Thus, improving the density of inhabitants would be a good factor that might positively influence criminality. The system reinforces its found regions with similar characteristics to Itaim Bibi but with increased population density.

Sé and Itaim Bibi are regions with different characteristics but with a high level of crime. Sé is a depopulated region in downtown São Paulo with a high flow of people that attracts criminal events. On the other hand, Itaim Bibi is a highly populated region with a low flow of people with high incomes (corporate headquarters and intense nightlife) that attracts

criminals. Therefore, the feasible counterfactual explanations in both regions are very different because both regions differ in nature.

These examples do not mean that we should use these counterfactual explanations because our system does not intend to implicate causal relations. We aim to demonstrate how using our system can create different counterfactual scenarios with other characteristics. It is up to public policy decision-makers to determine which would be more appropriate to put into practice.

VIII. Evaluation

To evaluate CounterCrime's effectiveness, we employed a dual-method approach that examined its technical performance and practical relevance in real-world applications. This evaluation involved a user study with computer science experts and interviews with professionals in the crime domain.

A. User Study

CounterCrime has been designed to help professionals analyze how different attributes relate to a model's classification of crime. To evaluate the effectiveness of the proposed tool, we gathered the experts' opinions about the methodology, functionalities, and visual components. They also examined the system and library modules, including machine learning and visualization tools. Here we present a summary of results, a detailed report of the users' responses is presented in Appendix B.1, available online.

Participants: We recruited 12 professionals from distinct fields such as programming, computer science, data science, mechatronics, mathematics, and physics. These participants work as data scientists, researchers, and machine learning practitioners with 1 to 12 years of experience applying their expertise to crime analysis ($mean = 3.33,\ SD = 2.8$).

Procedure: The evaluation took place in one-on-one, face-to-face sessions. Initially, we presented the methodology with working examples. Using these examples, participants performed **Task 1**, which involved reproducing the examples provided earlier. Next, we introduced the components detailed in Sections VI-A to VI-G, structured as classes joined by callback functions and the dataset. Participants then performed **Task 2**, which involved interacting with the notebook containing these modules to assess their practically in implementing additional visualizations.

After completing the tasks, we collected feedback on the CounterCrime system, methodology, case studies, and functionalities. Participants responded to quantitative questions (QT) on a Likert scale (1 to 5) and qualitative questions (QL).

Quantitative questions: *(QT1) "How relevant do you consider the proposed crime analysis tool?."*; *(QT2)"How easy is it to perform crime analysis on the system?."*; *(QT3) "Given a dataset containing urban, socioeconomic, and crime data, how easy is it to run crime analysis in other localities?."*; *(QT4) "To what extent does the proposed approach simplify the integration of new visualization and data analysis modules for crime analysis?."*; *(QT5) "How easy is it to modify parts of CounterCrime to perform data analysis in other contexts?."*;

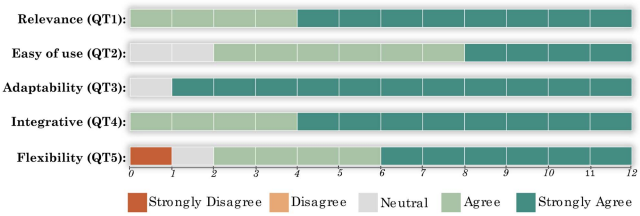

Fig. 8. Responses from twelve computer experts not involved in the tool's development. The bars are colored according to the expert's Likert scale answers.

Qualitative questions: *(QL1) "How do you perceive the integration with Jupyter Notebook in terms of enhancing the tool's versatility?"*; *(QL2) "Describe potential applications for this tool beyond crime analysis?"*; *(QL3) "What are your views on the effectiveness of CounterCrime in analyzing crime-related phenomena?."*; *(QL4) "Are you familiar with other visualization tools that perform similar analyses? If so, what aspects of CounterCrime make it superior to those methods?"*; *(QL5) "What are the limitations or disadvantages of CounterCrime when compared to other methods?"*.

Results: Fig. 8 summarizes the result of the quantitative user evaluation, showing that the proposed method has been positively evaluated. The statistics from the twelve participants are : Relevance ($Mean = 4.67, SD = 0.47$), Ease of Use ($Mean = 4.17, SD = 0.68$), Adaptability ($Mean = 4.83, SD = 0.55$), Integrative Capability ($Mean = 4.67, SD = 0.47$), and Flexibility ($Mean = 4.16, SD = 1.14$).

The outcome of the qualitative evaluation can be summarized as follows:

Module Flexibility (QL1 and QL2): Participants considered the modularization of the analysis and visualization tools in CounterCrime quite adequate. They noted that the flexibility the proposed methodology provides facilitates the application of counterfactual explanations to other domains such as COVID-19, air pollution, agriculture, and education. Some participants pointed out that the system might not be scalable but suggested that exporting the counterfactual results to a dashboard could address this issue. A participant commented: "*The integration is beneficial, as it enhances the tool's versatility by enabling the creation of interactive and dynamic charts directly within Jupyter. This makes data analysis more efficient and facilitates the communication of insights in an engaging manner.*"

Usefulness (QL3 and QL4): Most computer experts affirmed that CounterCrime is a valid tool to understand crime and propose changes. They noted that while the proposed changes may not be direct due to the set of variables, the tool can be used to find alternative ways to address the issues. An expert commented: "*Generating counterfactual explanations and allowing interaction with the modified attributes is a strong novelty of this framework. Furthermore, the portability of the notebook eases sharing with distinct stakeholders in urban data analysis.*"

Limitations (QL5): The main limitations identified by the experts include the constraints of the data (only a few variables could be effectively changed), the lack of identification of the closest safe region, and the need for a comprehensive tutorial to understand the system. An expert remarked: "*To improve*

the capability of the framework, it would be necessary to find data about policing (allocation of police personnel and police vehicles) and use data already available in the system at a higher resolution (yearly, for example). It could help to figure out less costly changes to fighting crime."

B. Domain Expert Review

We conducted interviews with three crime experts with backgrounds in economics. These experts are researchers and authorities who have been actively working on real-world crime phenomena for 2, 9, and 13 years, respectively. We demonstrated CounterCrime's functionalities by presenting the methodology, including some working examples, as well as its functionalities and visual components. We then addressed questions about the system and basic concepts such as the dataset, clustering, the machine learning model, and counterfactual explanations. Finally, we collected their feedback. Here we present a summary of results, a detailed report of the users' responses is presented in Appendix B.2, available online.

Methodology – global analysis: The analysis using the *CFs' recommendation filter* was appreciated for explaining important variables, and the interactive filtering enables possible planning interventions. The clustering of regions was considered fundamental for understanding similarities among areas. The shortcoming was the dataset, which included only a few variables. One of the experts commented: *(1) "The significance of bus stops (the most important variable) is a good indication of what we would expect in passerby crimes. The differences among clusters can reveal the dynamics of each locality."*; *(2) "The interactive updates can help us understand the most relevant factors for each location. It assists in planning focused interventions and comprehending criminality in detail.*"

Methodology – local analysis: The local analysis evaluated whether a policymaker can use the (synthetic) counterfactual explanations to reduce regional criminality. One expert raised a concern about the method's lack of guarantees regarding causality. Another expert noted: *"An example of a real region can facilitate the proposition and understanding of interventions by policymakers. My only concern is if the most similar safe region is too different from the analyzed localities. But this is a secondary concern"*.

Usability: A domain expert mentioned *(1) "This framework helps diagnose criminogenic factors and propose changes to reduce crimes in particular regions." (2) "The suggested changes (reduction of bus stops, increase in income, and decrease in population) might involve other costs that might not be feasible to a policymaker. However, it is important to know which regions are similar and the relative importance of those counterfactual explanations that reduce criminality. The variable importance and counterfactual explanations might be useful in making other decisions not covered by the framework, for example, increasing safety at bus stops and police presence in some places.*"

This work concluded successfully, opening new research avenues in crime analysis. One of the crime experts commented: *"I would use this framework daily to gather criminogenic and protective factors for a locality and use it as a starting point to propose local interventions."* Additionally, a computer expert remarked: *"In my opinion, this framework contributes to research in many other fields"*.

IX. Discussion and Future Work

CounterCrime satisfies the requirements of analysis in the current research stage of counterfactual explanations by enabling the exploration and validation of multiple counterfactual explanations in various regions. However, there are limitations and future work that would expand the research on counterfactual explanations and crime analysis.

Distinct and multiple machine learning models: As discussed in Section V, CounterCrime is not limited to any specific classifier, but this early approach uses Logistic Regression, which already showed a rich analysis. However, enriching the system with multiple, diverse classifiers to validate the counterfactual explanations would increase the analysis's robustness and reliability.

Counterfactual explanations and multiple crime types (or other social goals): This work focuses on Passerby robbery, but changing to another crime type would be simple. However, using counterfactual explanations to change one type may change other relevant goals. To address this, the Map View can show the effect of filtered counterfactuals on the model's classifications for other crime types or other social goals; for example, reducing bus stops may force people to use cars more frequently, causing traffic jams and pollution. In regional analysis, other views, such as parallel coordinates, can evaluate the impact of each counterfactual on other social goals, indicating the probability of a region being safe and other impacts. Finally, the generation of counterfactuals can be adapted to changes that make a region better to all social goals simultaneously.

Scalability issues: The research relies on a small set of variables to avoid issues that arise when increasing the number of variables. For example, increasing the number of variables can create problems in accurately identifying critical variables for clusters in *Cluster Comparison* and in the extensive list of variables for *CFs' Recommender Filtering*. The recommendation mechanism partially solves this issue, and two solutions could be grouping variables by similar impact or identifying and highlighting possible pairs of variables.

Dataset limitations: This research relied on a small set of variables available at official public institutions in Brazil, which did not include the most appropriate variable that policymakers could easily change. In future research, we aim to close deeper partnerships with public institutions to have access to more changeable variables such as police patrol, urban maintenance, graffiti, public illumination, and street vendors. We state this step as future work due to the necessity of long and bureaucratic talks, but we think it would result in a reliable tool to plan policies for crime reduction.

Quality of dataset and fairness aspects: The data gathering and curation of this paper were done closely with professionals of NEV (Center of Study of the Violence) with vast experience in crime analysis in São Paulo and developing different projects concerned with Democratic Policing, Human Rights, Race

victimization, and society data bias (see NEV Publications). The crime records were provided by the Police department to NEV experts, and experts assembled, curated, and cleaned the datasets. Even with this care, we know that this type of dataset has its shortcomings in ethical aspects, mainly because of the lack of documented police occurrences affecting marginalized people and a high incidence of documented occurrences committed by marginalized people. Because of this, the usage of fairness models was considered in this work, but it was beyond the scope of this work. Fairness modeling demands the definition of desired outcomes and sensible variables. Still, increasing police in marginalized people might be good or bad depending on the optics; the sensible variable is also unclear. This definition's subjectivity and the lack of fairness for our problem made this contribution out of our scope.

Limitations of crime analysis on policing: Integrating crime analysis into policing has significantly influenced strategies by enhancing predictive capabilities and optimizing resource allocation. Technological tools now utilize vast amounts of data to forecast potential crime hotspots and identify individuals who might be at risk of engaging in criminal activities or becoming victims. For instance, the works of Lum and Isaac (2016) and Richardson et al. (2019) discuss how predictive policing systems leverage historical crime data and sophisticated algorithms to anticipate criminal activities [108], [109]. However, reliance on biased data — often referred to as "dirty data" — can perpetuate existing social inequalities, reinforce problematic policing practices, and introduce new biases into law enforcement [109], [110], [111].

Furthermore, studies by Amiruzzaman et al. (2022) highlight how AI can analyze urban environments and social behaviors to correlate visual diversity with crime, thereby informing police deployment and urban planning [111], [112]. These tools, however, must be employed with caution to avoid reinforcing discriminatory practices. Several studies emphasize the critical need for transparency, accountability, and eliminating biases in data used for predictive models [108], [109], [110], [111]. Implementing ethical considerations and ensuring data accuracy in AI-driven crime analysis is paramount for maintaining public trust and achieving equitable crime reduction [110], [111].

Generalization to other datasets: We built a methodology for analyzing crime patterns that could be directly employed to analyze counterfactual explanations in any data. Maybe the *Map View* may not be applicable in some contexts; many datasets contain a spatial component. We plan to use our system with datasets such as COVID-19 infection risk, car crash incidents, or traffic jams to generalize our results in future work. Scatter plots with user-selected relevant variables could also benefit other fields, such as medical diagnosis and loans.

Relation with explainability methods: Feature attribution explanation methods [113], such as SHAP, aim to determine the importance of each feature for the model decision. While SHAP gives additive feature attribution (how much a specific variable contributes to the prediction of a sample), and LIME provides local importance to each variable, counterfactual explanations provide information using concrete examples in the original feature space, therefore avoiding adding a new level of abstraction, as feature attribution does. This makes the outcome of counterfactual fully interpretable and useful for identifying specific adjustments to produce a desired result. However, not all exploratory scenarios may be feasible in practice. Therefore, experts need to review these scenarios and assess their practical viability.

X. Conclusion

In this work, we present a visual framework for evaluating counterfactual explanations in crime analysis, which has two key benefits: (1) it provides a framework using counterfactual explanations (built over a machine learning model acting as proxy of real data) for emulating crime scenarios across the city, and (2) enables the exploration of hypothetical change scenarios within the feature space, offering insights into what changes the model associates with influencing crime classifications in unsafe regions. The proposed visualization mechanics provide practical guidance for understanding the different variables' role in these counterfactual scenarios and how those scenarios affect the model's classifications across the city. The clustering procedure and its related visualization aid in identifying clusters of regions with similar counterfactual explanations. Visualization tools at the local level help analyze clusters of counterfactual explanations and their costs compared to actual regions classified as safe. The work shows that crime is not uniform throughout the city, and the same counterfactuals can similarly affect clusters of regions. The understanding gained from these clusters can inform decision-makers in considering targeted approaches, potentially optimized for each cluster based on the model's sensitivities. Our work represents a relevant step towards supporting the decision-making process for public policies related to crime. Additionally, implementing CounterCrime as a Python library that can be used directly in Jupyter Notebooks simplifies the analysis process, allowing analysts to use their standard frameworks and only call our tool when necessary.

Acknowledgment

The opinions, hypotheses, conclusions, and recommendations expressed in this material are the responsibility of the authors and do not necessarily reflect the views of FAPESP, FAPERJ, PROCIENCIA, the *Fundação Getulio Vargas*, and CNPq.

References

[1] M. Oliveira, C. Bastos-Filho, and R. Menezes, "The scaling of crime concentration in cities," *PLoS One*, vol. 12, no. 8, 2017, Art. no. e0183110.
[2] L. G. Alves, H. V. Ribeiro, E. K. Lenzi, and R. S. Mendes, "Distance to the scaling law: A useful approach for unveiling relationships between crime and urban metrics," *PLoS One*, vol. 8, no. 8, pp. 1–8, 2013.
[3] A. Gomez-Lievano, H. Youn, and L. M. A. Bettencourt, "The statistics of urban scaling and their connection to Zipf's law," *PLoS One*, vol. 7, pp. 1–11, Jul. 2012.
[4] D. E. Hojman, "Inequality, unemployment and crime in Latin American cities," *Crime Law Social Change*, vol. 41, no. 1, pp. 33–51, 2004.
[5] S. D. Levitt, "Alternative strategies for identifying the link between unemployment and crime," *J. Quantitative Criminol.*, vol. 17, no. 4, pp. 377–390, 2001.
[6] A. Cotte Poveda, "Violence and economic development in colombian cities: A dynamic panel data analysis," *J. Int. Develop.*, vol. 24, no. 7, pp. 809–827, 2012.

[7] B. Atems, "Identifying the dynamic effects of income inequality on crime," *Oxford Bull. Econ. Statist.*, vol. 82, no. 4, pp. 751–782, 2020.

[8] M. Kelly, "Inequality and crime," *Rev. Econ. Statist.*, vol. 82, no. 4, pp. 530–539, 2000.

[9] P. Day, G. Breetzke, S. Kingham, and M. Campbell, "Close proximity to alcohol outlets is associated with increased serious violent crime in New Zealand," *Australian New Zealand J. Public Health*, vol. 36, no. 1, pp. 48–54, 2012.

[10] E. S. McCord and J. H. Ratcliffe, "Intensity value analysis and the criminogenic effects of land use features on local crime patterns," *Crime Patterns Anal.*, vol. 2, no. 1, pp. 17–30, 2009.

[11] T. H. Grubesic and W. A. Pridemore, "Alcohol outlets and clusters of violence," *Int. J. Health Geographics*, vol. 10, pp. 1–12, 2011.

[12] M. Livingston, "A longitudinal analysis of alcohol outlet density and domestic violence," *Addiction*, vol. 106, pp. 919–25, 2010.

[13] G. Garcia-Zanabria et al., "CrimAnalyzer: Understanding crime patterns in São Paulo," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 4, pp. 2313–2328, Apr. 2021.

[14] T. C. Bailey et al., *Interactive Spatial Data Analysis*, vol. 413. Essex, U.K.: Longman Scientific & Technical, 1995.

[15] S. Shiode, N. Shiode, R. Block, and C. R. Block, "Space-time characteristics of micro-scale crime occurrences: An application of a network-based space-time search window technique for crime incidents in Chicago," *Int. J. Geographical Inf. Sci.*, vol. 29, no. 5, pp. 697–719, 2015.

[16] J. Borges et al., "Feature engineering for crime hotspot detection," in *Proc. 2017 IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Computed, Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, 2017, pp. 1–8.

[17] J. Borges et al., "Time-series features for predictive policing," in *Proc. IEEE Int. Smart Cities Conf.*, 2018, pp. 1–8.

[18] N. Naik, S. D. Kominers, R. Raskar, E. L. Glaeser, and C. A. Hidalgo, "Computer vision uncovers predictors of physical urban change," in *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 29, pp. 7571–7576, 2017.

[19] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human decisions and machine predictions," *Quart. J. Econ.*, vol. 133, no. 1, pp. 237–293, 2018.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 1135–1144.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, Curran Associates, Inc., 2017, pp. 4765–4774.

[22] H. Rodger, A. Lensen, and M. Betkier, "Explainable artificial intelligence for assault sentence prediction in New Zealand," *J. Roy. Soc. New Zealand*, vol. 53, no. 1, pp. 133–147, 2023.

[23] X. Zhang, L. Liu, M. Lan, G. Song, L. Xiao, and J. Chen, "Interpretable machine learning models for crime prediction," *Comput. Environ. Urban Syst.*, vol. 94, 2022, Art. no. 101789.

[24] J. S. Levy, "Counterfactuals, causal inference, and historical analysis," *Secur. Stud.*, vol. 24, no. 3, pp. 378–402, 2015.

[25] U. Kuhl, A. Artelt, and B. Hammer, "For better or worse: The impact of counterfactual explanations' directionality on user behavior in XAI," 2023, *arXiv:2306.07637*.

[26] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Conv. Inf. Commun. Technol. Electron. Microelectronics*, 2018, pp. 210–215.

[27] F. Cheng, Y. Ming, and H. Qu, "DECE: Decision explorer with counterfactual explanations for machine learning models," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1438–1447, Feb. 2021.

[28] O. Gomez, S. Holter, J. Yuan, and E. Bertini, "ViCE: Visual counterfactual explanations for machine learning models," in *Proc. Int. Conf. Intell. User Interfaces*, 2020, pp. 531–535.

[29] P. Judea, "An introduction to causal inference," *Int. J. Biostatist.*, vol. 6, no. 2, pp. 1–62, 2010.

[30] G. Garcia-Zanabria et al., "CriPAV: Street-level crime patterns analysis and visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 12, pp. 4000–4015, Dec. 2022.

[31] Q. Yang, J. Yin, C. Ling, and R. Pan, "Extracting actionable knowledge from decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 43–55, Jan. 2007.

[32] S. Subramani et al., "Mining actionable knowledge using reordering based diversified actionable decision trees," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, Springer, 2016, pp. 553–560.

[33] G. Tolomei, F. Silvestri, A. Haines, and M. Lalmas, "Interpretable predictions of tree-based ensembles via actionable feature tweaking," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2017, pp. 465–474.

[34] C. Yang, W. N. Street, and J. G. Robinson, "10-year CVD risk prediction and minimization via inverse classification," in *Proc. 2nd ACM SIGHIT Int. Health Informat. Symp.*, Association for Computing Machinery, 2012, pp. 603–609.

[35] Q. Lu, Z. Cui, Y. Chen, and X. Chen, "Extracting optimal actionable plans from additive tree models," *Front. Comput. Sci.*, vol. 11, no. 1, pp. 160–173, 2017.

[36] Q. Lv et al., "Achieving data-driven actionability by combining learning and planning," *Front. Comput. Sci.*, vol. 12, no. 5, pp. 939–949, 2018.

[37] Z. Cui, W. Chen, Y. He, and Y. Chen, "Optimal action extraction for random forests and boosted trees," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, Association for Computing Machinery, 2015, pp. 179–188.

[38] A. Parmentier and T. Vidal, "Optimal counterfactual explanations in tree ensembles," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8422–8431.

[39] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proc. Conf. Fairness Accountability Transparency*, 2019, pp. 10–19.

[40] Y. Gao, Q. Liu, G. Chen, B. Zheng, and L. Zhou, "Answering why-not questions on reverse top-k queries," in *Proc. VLDB Endowment*, vol. 8, no. 7, pp. 738–749, 2015.

[41] Z. He and E. Lo, "Answering why-not questions on top-k queries," in *Proc. Int. Conf. Data Eng.*, 2012, pp. 750–761.

[42] L. Chen, X. Lin, H. Hu, C. S. Jensen, and J. Xu, "Answering why-not questions on spatial keyword top-k queries," in *Proc. Int. Conf. Data Eng.*, 2015, pp. 279–290.

[43] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.

[44] A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke, "Actionable interpretability through optimizable counterfactual explanations for tree ensembles," *arXiv*, vol. 1, pp. 1–8, 2019.

[45] A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke, "FOCUS: Flexible optimizable counterfactual explanations for tree ensembles," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 5313–5322.

[46] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. Conf. Fairness Accountability Transparency*, 2020, pp. 607–617.

[47] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera, "Model-agnostic counterfactual explanations for consequential decisions," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 895–905.

[48] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren, "A survey of surveys on the use of visualization for interpreting machine learning models," *Inf. Vis.*, vol. 19, no. 3, pp. 207–233, 2020.

[49] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Vis. Informat.*, vol. 1, no. 1, pp. 48–56, 2017.

[50] Z. J. Wang et al., "CNN explainer: Learning convolutional neural networks with interactive visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1396–1406, Feb. 2021.

[51] X. Zhao, Y. Wu, D. L. Lee, and W. Cui, "iForest: Interpreting random forests via visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 407–416, Jan. 2019.

[52] M. P. Neto and F. V. Paulovich, "Explainable matrix-visualization for global and local interpretability of random forest classification ensembles," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1427–1437, Feb. 2021.

[53] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. P. Chau, "ActiVis: Visual exploration of industry-scale deep neural network models," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 88–97, Jan. 2018.

[54] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 8, pp. 2674–2693, Aug. 2019.

[55] M. M. Pereira and F. V. Paulovich, "RankViz: A visualization framework to assist interpretation of learning to rank algorithms," *Comput. Graph.*, vol. 93, pp. 25–38, 2020.

[56] D. Collaris and J. J. Van Wijk, "ExplainExplore: Visual exploration of machine learning explanations," in *Proc. IEEE Pacific Visual. Symp.*, 2020, pp. 26–35.

[57] S. Sawada, "Model-agnostic visual explanation of machine learning models based on heat map," in *Proc. Eurographics Conf. Vis.*, 2019, pp. 1–3.

[58] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2016, pp. 5686–5697.

[59] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 56–65, Jan. 2020.

[60] Y. Ming, H. Qu, and E. Bertini, "RuleMatrix: Visualizing and understanding classifiers with rules," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 342–352, Jan. 2019.

[61] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, "Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 364–373, Jan. 2019.

[62] K. Xu, M. Xia, X. Mu, Y. Wang, and N. Cao, "EnsembleLens: Ensemble-based visual exploration of anomaly detection algorithms with multi-dimensional data," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 1, pp. 109–119, Jan. 2019.

[63] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini, "Interpreting black-box classifiers using instance-level visual explanations," in *Proc. 2nd Workshop Hum.-In-the-Loop Data Analytics*, 2017, pp. 1–6.

[64] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush, "LSTMVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 667–676, Jan. 2018.

[65] G. Garcia-Zanabria, D. A. Gutierrez-Pachas, G. Camara-Chavez, J. Poco, and E. Gomez-Nieto, "SDA-Vis: A visualization system for student dropout analysis based on counterfactual exploration," *Appl. Sci.*, vol. 12, no. 12, pp. 1–20, 2022.

[66] J. M. Metsch et al., "CLARUS: An interactive explainable AI platform for manual counterfactuals in graph neural networks," *J. Biomed. Informat.*, vol. 150, 2024, Art. no. 104600.

[67] U. Schlegel, J. Rauscher, and D. A. Keim, "Interactive counterfactual generation for univariate time series," 2024, *arXiv:2408.10633*.

[68] T. Wang, C. Rudin, D. Wagner, and R. Sevieri, "Learning to detect patterns of crime," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, Springer, 2013, pp. 515–530.

[69] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, "Crime pattern detection, analysis & prediction," in *Proc. 2017 Int. Conf. Electron. Commun. Aerosp. Technol.*, 2017, pp. 225–230.

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

[71] T. Newburn and R. Sparks, "Criminal justice and political cultures: National and international dimensions of crime control," *Criminal Justice Political Cultures: Nat. Int. Dimens. Crime Control*, vol. 17, pp. 1–276, 2012.

[72] J. Eck and D. L. Weisburd, "Crime places in crime theory," *Crime Place: Crime Prevention Stud.*, vol. 4, pp. 1–33, 2015.

[73] J. L. Lauritsen, M. L. Rezey, and K. Heimer, "Violence and economic conditions in the United States, 1973–2011: Gender, race, and ethnicity patterns in the national crime victimization survey," *J. Contemporary Criminal Justice*, vol. 30, no. 1, pp. 7–28, 2014.

[74] S. Shiode and N. Shiode, "Network-based space-time search-window technique for hotspot detection of street-level crime incidents," *Int. J. Geographical Inf. Sci.*, vol. 27, no. 5, pp. 866–882, 2013.

[75] K. Salinas, T. Gonçalves, V. Barella, T. Vieira, and L. G. Nonato, "CityHub: A library for urban data integration," in *Proc. 35th SIBGRAPI Conf. Graph. Patterns Images*, 2022, pp. 43–48.

[76] A. I. Robinson, F. Carnes, and N. M. Oreskovic, "Spatial analysis of crime incidence and adolescent physical activity," *Prev. Med.*, vol. 85, pp. 74–77, 2016.

[77] S. N. de Melo, L. F. Matias, and M. A. Andresen, "Crime concentrations and similarities in spatial crime patterns in a brazilian context," *Appl. Geogr.*, vol. 62, pp. 314–324, 2015.

[78] X. Ye, X. Xu, J. Lee, X. Zhu, and L. Wu, "Space–time interaction of residential burglaries in Wuhan, China," *Appl. Geogr.*, vol. 60, pp. 210–216, 2015.

[79] S. V. Nath, "Crime pattern detection using data mining," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol. Workshops*, 2006, pp. 41–44.

[80] R. Deryol, P. Wilcox, M. Logan, and J. Wooldredge, "Crime places in context: An illustration of the multilevel nature of hot spot development," *J. Quantitative Criminol.*, vol. 32, no. 2, pp. 305–325, 2016.

[81] 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 *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2014, pp. 321–322.

[82] S. Kim, S. Jeong, I. Woo, Y. Jang, R. Maciejewski, and D. S. Ebert, "Data flow analysis and visualization for spatiotemporal statistical data without trajectory information," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 3, pp. 1287–1300, Mar. 2018.

[83] V. Spicer, J. Song, P. Brantingham, A. Park, and M. A. Andresen, "Street profile analysis: A new method for mapping crime on major roadways," *Appl. Geogr.*, vol. 69, pp. 65–74, 2016.

[84] A. Malik, R. Maciejewski, T. F. Collins, and D. Ebert, "Visual analytics law enforcement toolkit," in *Proc. IEEE Int. Conf. Technol. Homeland Secur.*, 2010, pp. 222–228.

[85] A. M. M. Razip et al., "A mobile visual analytics approach for law enforcement situation awareness," in *Proc. IEEE Pacific Visual. Symp.*, 2014, pp. 169–176.

[86] A. Godwin and J. T. Stasko, "HotSketch: Drawing police patrol routes among spatiotemporal crime hotspots," in *Proc. 50th Hawaii Int. Conf. Syst. Sci.*, 2017, pp. 1–9.

[87] C. Calhoun, C. E. Stobbart, D. M. Thomas, J. A. Villarrubia, D. E. Brown, and J. H. Conklin, "Improving crime data sharing and analysis tools for a web-based crime analysis toolkit: WebCAT 2.2," in *Proc. IEEE Syst. Inf. Eng. Des. Symp.*, 2008, pp. 40–45.

[88] L. J. S. Silva, S. Fiol-González, C. F. Almeida, S. D. Barbosa, and H. Lopes, "CrimeVis: An interactive visualization system for analyzing crime data in the state of Rio De Janeiro," in *Proc. Int. Conf. Enterprise Inf. Syst.*, 2017, pp. 193–200.

[89] G. Garcia-Zanabria et al., "Mirante: A visualization tool for analyzing urban crimes," in *Proc. Conf. Graph. Patterns Images*, 2020, pp. 148–155.

[90] D. Wang et al., "Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns," *Comput. Environ. Urban Syst.*, vol. 39, pp. 93–106, 2013.

[91] G. D. Breetzke and A. L. Pearson, "The fear factor: Examining the spatial variability of recorded crime on the fear of crime," *Appl. Geogr.*, vol. 46, pp. 45–52, 2014.

[92] M. Craglia, R. Haining, and P. Wiles, "A comparative evaluation of approaches to urban crime pattern analysis," *Urban Stud.*, vol. 37, no. 4, pp. 711–729, 2000.

[93] M. B. Short, P. J. Brantingham, A. L. Bertozzi, and G. E. Tita, "Dissipation and displacement of hotspots in reaction-diffusion models of crime," in *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 9, pp. 3961–3965, 2010.

[94] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.

[95] Q. Wang, Z. Xu, Z. Chen, Y. Wang, S. Liu, and H. Qu, "Visual analysis of discrimination in machine learning," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 1470–1480, Feb. 2021.

[96] M. M. Raimundo, L. G. Nonato, and J. Poco, "Mining pareto-optimal counterfactual antecedents with a branch-and-bound model-agnostic algorithm," *Data Mining Knowl. Discov.*, vol. 38, pp. 2942–2974, 2024.

[97] F. Pedregosa et al., "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[98] M. Wes et al., "Data structures for statistical computing in Python. scipy," vol. 445, no. 1, pp. 51–56, 2010.

[99] C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, Sep. 2020.

[100] B. Michael, O. Vadim and H. Jeffrey "D³ data-driven documents", *IEEE trans. visualization and comput. graph.*, vol. 17, no. 12, pp. 2301–2309, 2011.

[101] B. Boldis, M. San Sebastián, and P. E. Gustafsson, "Unsafe and unequal: A decomposition analysis of income inequalities in fear of crime in Northern Sweden," *Int. J. Equity Health*, vol. 17, no. 1, pp. 1–13, 2018.

[102] T. Hu, X. Zhu, L. Duan, and W. Guo, "Urban crime prediction based on spatio-temporal Bayesian model," *PLoS One*, vol. 13, no. 10, pp. 206–215, 2018.

[103] M. Hobbs et al., "Close proximity to alcohol outlets is associated with increased crime and hazardous drinking: Pooled nationally representative data from New Zealand," *Health Place*, vol. 65, pp. 1–7, 2020.

[104] R. Zahnow and J. Corcoran, "Crime and bus stops: An examination using transit smart card and crime data," *Environ. Plan. B: Urban Analytics City Sci.*, vol. 48, no. 4, pp. 706–723, 2021.

[105] T. D. Stucky and S. L. Smith, "Exploring the conditional effects of bus stops on crime," *Secur. J.*, vol. 30, no. 1, pp. 290–309, 2017.
[106] R. F. Abenoza, V. Ceccato, Y. O. Susilo, and O. Cats, "Individual, travel, and bus stop characteristics influencing travelers' safety perceptions," *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2672, no. 8, pp. 19–28, 2018.
[107] L. G. Alves, H. V. Ribeiro, and R. S. Mendes, "Scaling laws in the dynamics of crime growth rate," *Physica A: Statist. Mechanics Appl.*, vol. 392, no. 11, pp. 2672–2679, 2013.
[108] K. Lum and W. Isaac, "To predict and serve?," *Significance*, vol. 13, no. 5, pp. 14–19, 2016.
[109] R. Richardson, J. M. Schultz, and K. Crawford, "Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice," *NYUL Rev. Online*, vol. 94, p. 15-55, 2019.
[110] B. Taylor, A. Kowalyk, and R. Boba, "The integration of crime analysis into law enforcement agencies: An exploratory study into the perceptions of crime analysts," *Police Quart.*, vol. 10, no. 2, pp. 154–169, 2007.
[111] C. Sanders and C. Condon, "Crime analysis and cognitive effects: The practice of policing through flows of data," in *The Policing of Flows*, Evanston, IL, USA: Routledge, 2020, pp. 73–91.
[112] M. Amiruzzaman, Y. Zhao, S. Amiruzzaman, A. C. Karpinski, and T. H. Wu, "An AI-based framework for studying visual diversity of urban neighborhoods and its relationship with socio-demographic variables," *J. Comput. Social Sci.*, vol. 6, no. 1, pp. 315–337, 2023.
[113] E. S. Ortigossa, T. Gonçalves, and L. G. Nonato, "Explainable artificial intelligence (XAI)–From theory to methods and applications," *IEEE Access*, vol. 12, pp. 80799–80846, 2024.

Marcos M. Raimundo received the BS degree in computer engineering, and the MS and PhD degrees in electrical and computer engineering from the University of Campinas, in 2011, 2014, and 2018, respectively. He is an assistant professor with the Institute of Computing (IC), University of Campinas (UNICAMP). Before joining the faculty, he expanded his expertise as a postdoctoral researcher with the School of Applied Mathematics, Fundação Getulio Vargas (FGV). His current work applies principles from operations research and optimization to address challenges in trustworthy AI, including fairness, explainability, robustness and the development of ethical artificial intelligence systems.

Germain Garcia-Zanabria received the BE degree in system engineering from the Universidad Nacional de San Antonio Abad del Cusco, Cusco, Peru, in 2012, the MSc degree in computer science from San Pablo Catholic University, Arequipa, Peru, in 2016, and the PhD degree in computer science from the University of São Paulo (ICMC-USP), São Carlos, Brazil, in 2021. He is a assistant professor with the Department of Data Science, University of Engineering and Technology, Lima, Peru. His research interests include data science, visualization, analytics, and learning models.

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

Jorge Poco (Senior Member, IEEE) received the BE degree in systems engineering from the National University of San Agustín, Peru, in 2008, the MSc degree in computer science from the University of São Paulo, Brazil, in 2010, and the PhD degree in computer science from New York University, in 2015. He is an associate professor with the School of Applied Mathematics, Fundação Getúlio Vargas in Brazil. His research interests are data visualization, visual analytics, machine learning, and data science. He has served in several program committees, including IEEE SciVis, IEEE InfoVis, and EuroVis.

Open Access provided by 'Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614' within the CRUI CARE Agreement