San Francisco Crime Data Analysis

Deepika Sulakhe, Vasavya Sri, Snehal Maske, Pooja Raju

Abstract—**This project analyzes and visualizes violent crime data in San Francisco from 2018 to 2024, using geographic visualization techniques to identify spatial and temporal crime patterns. Based on 2014 NeighborhoodScout data, San Francisco ranks among the least safe 3% of U.S. cities, with a recent increase in violent crime post-COVID in 2023. Focusing on four police districts, the analysis employs heatmaps and hotspot analysis to pinpoint high-crime areas. The cleaned dataset, comprising 7641 records, includes attributes like incident datetime, location, crime category, and police district. Key visualization tools, including Folium and Flourish, were used alongside statistical tests for clustering. Heatmaps show crime density with color gradients, while hotspot analysis aggregates crimes within a 300-meter radius to indicate incident density. Results highlight regions with high crime rates through interactive visualizations such as marker clusters, heatmaps, dot plots, and bar graphs, revealing trends over time and across different areas. The study's primary goal is to identify cyclical patterns of violent crime, with secondary objectives of finding prevalent crimes in specific regions and determining safe and dangerous areas. The findings enhance the understanding of criminal behavior patterns, potentially informing law enforcement strategies and public safety measures. The methodologies can be adapted for crime analysis in other cities, contributing to broader crime prevention efforts.**

I. Introduction

According to NeighborhoodScout data from 2014, San Francisco ranks among the least safe 3% of U.S. cities, with a 1 in 139 chance of being a victim of violent crime compared to 1 in 225 for the entire state of California [1]. Violent crime in San Francisco saw a minor 3 percent increase in 2023 post -Covid [2]. This leads to San Francisco's violent crime issue warrants further exploration. This analysis utilizes open-source crime data from San Francisco to perform spatial and temporal analysis of violent crime using geographic visualization software. Nine regions with violent crimes were selected based on preliminary data analysis. The analysis employs hotspot analysis and tracking analysis for geographic visualization. While the scope is data visualization and interpretation, the findings could aid other crime analysis tasks like classification by revealing notable trends and patterns in the data. The methods can also be applied to other cities or countries. The primary aim is to determine if there are meaningful cyclical patterns of violent crime across space and time within different San Francisco areas. Secondary objectives include identifying if certain crimes are more prevalent in specific geographic regions and pinpointing safe/dangerous areas and high-crime days. By focusing on San Francisco's violent crime problem through spatial and temporal analysis using visualization techniques, this study seeks to enhance the understanding of criminal behavior patterns in the city. The findings inform further analysis and potential interventions.

II. Motivation

Violent crime poses a significant threat to the safety and well-being of residents in urban areas like San Francisco. With its complex social dynamics and diverse population, the city faces ongoing challenges in addressing crime rates that impact communities, businesses, and the overall perception of safety. Recent trends indicate persistent issues with violent crime, necessitating proactive measures to understand and combat its root causes.

In recent years, San Francisco has grappled with fluctuations in crime rates, with certain neighborhoods experiencing disproportionate levels of violence. This underscores the urgency of implementing targeted strategies to address crime hotspots and prevent further escalation. The need for innovative approaches to crime analysis and public safety planning has never been more pressing.

By conducting a comprehensive analysis of violent crime data spanning from 2018 to 2024, this study seeks to shed light on the spatial and temporal patterns of criminal activity in San Francisco. Understanding where and when violent crimes occur can provide invaluable insights for law enforcement agencies, policymakers, and community stakeholders to develop effective intervention strategies.

Moreover, the implications of this research extend beyond San Francisco. As cities worldwide grapple with similar challenges related to crime and public safety, the methodologies and visualizations developed in this study offer a replicable framework for analyzing crime data and informing evidence-based decision-making. By sharing our findings and best practices, we aim to contribute to a broader conversation on crime prevention and urban safety management, ultimately fostering safer and more resilient communities.

III. Related Work

In 2016, Venturini and Baralis utilized crime data from 2015 and found that incidents of vehicle theft in San Francisco varied across different geographic areas depending on the day of the week [3] While vehicle theft itself is not classified as a violent crime, it stands to reason that similar findings might be observed across other types of crimes, including those of a violent nature. McDowal, Loftin, and Pate (2012) explored a widely-held theory suggesting that higher temperatures correlate positively with the number of violent crimes, indicating a greater likelihood of such crimes occurring during warmer months nationwide [4] They provided evidence supporting this trend generally but noted variations in crime patterns between different geographic areas. These findings, alongside those of Venturini and Baralis, underscore the importance of continuing spatial-temporal analysis of crime.

To further illustrate the significance of this analytical approach, Wu (2016) demonstrated that it is possible to classify crime types using spatial-temporal data from San Francisco [5]. Although Wu achieved only a 28.51% classification accuracy, it is presumed that improvements in methodology could enhance this accuracy. The dataset used by Wu was also employed in the current project. Wu visualized the occurrence of the most common crimes over 12 months (aggregated from 2003 to 2015) with line charts, and utilized heat maps to show particular crimes over days of the week and hours of the day. Geographic visualizations were also used to highlight regions with higher instances of specific crimes .

Herrmann (2015) examined violent crime data from New York City and found that crime hotspots shifted spatially over time [6]. This adds complexity to crime analysis and highlights the necessity for visual analysis across both space and time. Herrmann mentioned that while it is often possible to identify patterns within a hotspot over time, a thorough understanding of crime in a region requires a detailed analysis of each hotspot, in contrast to viewing the region as a whole. The tendency of hotspots to shift might suggest the aggregated analysis by Wu (2016) was insufficient, potentially contributing to the relatively low classification accuracy.

Wheeler (2016) assessed visualization methods for analyzing crime trends in New York, contending that the non-Poisson distribution of crime occurrences complicates data visualization using tables and metrics like percent change [7]. Wheeler suggested that time series visualizations such as line graphs are more effective for capturing meaningful trends within the data. He argued that enhancing geographic representation over time, as proposed for San Francisco crime data, could improve understanding if modern visualization techniques are employed.

To support the focus on specific regions within San Francisco, Weisburd, Bernasco, and Bruinsma (2009) argued that analyzing crime over large geographic areas is flawed [8]. They posted that the social and environmental factors within any narrow geographic area significantly affect criminal behavior in that region, making smaller-scale analyses more insightful. Complementing this, Bernasco and Block (2011) studied crime patterns in Chicago and suggested that crime analysis should avoid a per-block basis [9]. They argued that individual blocks often have small populations, providing an incomplete view of crime trends. Instead, examining groups of blocks or regions offers a more comprehensive understanding, aligning with the approach reported in this paper.

IV. Dataset Exploration

The original dataset consisted of 10000 records spanning from 2018 to 2024. After the data cleaning process the dataset has 7641 records. The attributes included are incident_datetime, incident_date, incident_time, incident_year, incident_day_of_week, report_datetime, row_id, incident_id, incident_number,report_type_code, report_type_description, incident_code, incident_category, incident_subcategory, incident_description, resolution, police_district, filed_online, cad_number, intersection, cnn, analysis_neighborhood, supervisor_district, supervisor_district_2012, latitude, longitude, point. A detailed description of each attribute is mentioned in Table 1. For our analysis, the relevant attributes are incident_date,incident_time, incident_year, police_district, intersection and the geospatial coordinate attributes, latitude and longitude. This project focuses on visualizing crime data in San Francisco using various data visualization techniques. The methodology consists of several steps, including data collection, preprocessing, and visualization. Below is a detailed description of each step:

1. Data Collection and analysis : Data source is from the crime data obtained from the San Francisco government's public dataset using the Socrata API. The dataset contains information about crime incidents in San Francisco.

Accessing Data: Using the Socrata API, data is fetched and stored in a pandas Data Frame for further processing.

Cleaning Data: Removing unnecessary columns to streamline the dataset. Handling missing values by dropping rows with null longitude and latitude values as most of the heatmaps designing data is acquired based on the point (which is a String that has a combination of latitude and longitude), incident category, incident year, incident year.

Did not process or use the data like rowid, cnn , intersection, supervisor_district_2012,supervisor_district.

The below shown figures are the visualizations(graphs) based on categories of data analysis, which thefts are happening most based on the crime description and which day has the most crimes happening in the past years and which have the highest crimes.

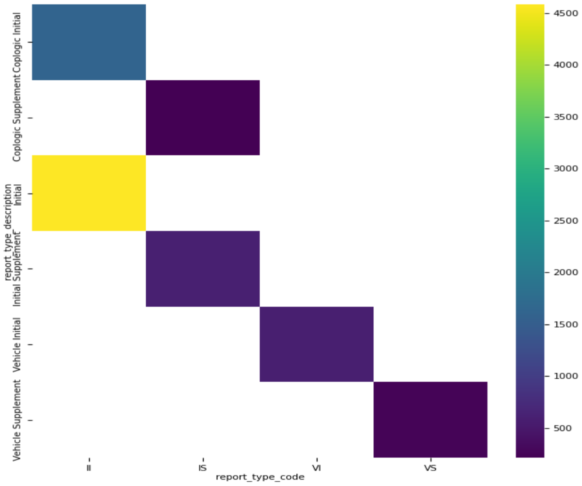


Fig1 : Crime happened mostly as per recorded description

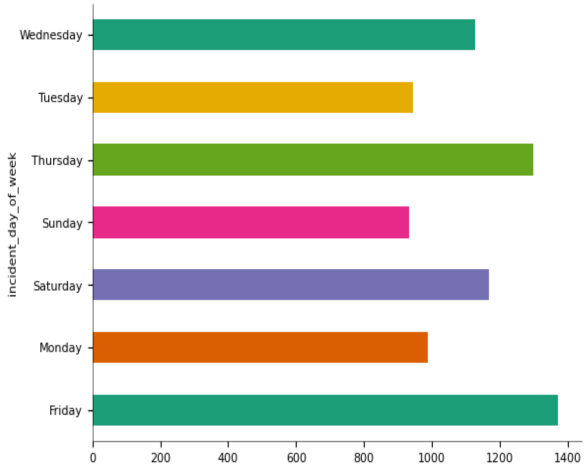


Fig2: Crime happened mostly a per day

2. Data Visualization: Data visualization is a critical component of this project, facilitating the interpretation and understanding of crime data in San Francisco. Several visualization techniques and tools were employed to present the data in an informative and engaging manner. This section outlines the key visualizations created, the tools used, and the insights derived from these visualizations.

Setting Up the Map: Using the Folium library to create a map centered around San Francisco.**Added Marker Clusters**, also Heatmaps Implementing Marker Clusters to visualize the distribution of crime incidents on the map

Heat Maps and Other Layers: Adding different layers, such as Heat Maps, to visualize the density of crime incidents across various regions as per year (2018-2014).

Enhancing Visuals with Flourish: Embedding Flourish visualizations into the Folium map to provide additional insights. Adding custom HTML and CSS to enhance the presentation of the map. The final interactive map to an HTML file, ensuring all elements are correctly embedded and styled .

Below shown is the table indicating the data columns and datatype of that particular column.

Column Name	Type
incident_datetime	Date/time
incident_date	Date/time
incident_time	String
incident_year	String
incident_day_of_week	Date/time
report_datetime	Date/time
row_id	Number
incident_id	Number
incident_number	Number
report_type_code	Number
report_type_description	String
incident_code	Number
incident_category	String
incident_subcategory	String
incident_description	String
resolution	String
police_district	String
filed_online	Boolean
cad_number	Number
intersection	String
cnn	Number
analysis_neighborhood	String
supervisor_district	String
supervisor_district_2012	String
latitude	Decimal
incident_datetime	Date/time
incident_date	Date/time
incident_time	String
incident_year	String
incident_day_of_week	Date/time
longitude	Decimal
point	String

Table 5. Methodology

V.(i) Heatmaps

Heatmaps are a powerful visualization tool used to represent the density of data points over a geographical area. In the context of crime data, heatmaps are utilized to display areas with varying levels of crime intensity, providing an intuitive visual summary of crime distribution. How Heatmaps Work:Color Gradient: Heatmaps use a color gradient to indicate the intensity of data points in a particular area. Typically, warmer colors (e.g., red, orange) indicate higher densities, while cooler colors (e.g., blue, green) indicate lower densities.Implementation in the Project:Data Preparation: The dataset containing crime incidents, including their geographic coordinates (latitude and longitude), is cleaned and preprocessed.

Heatmap Generation: Using the Folium library in Python, a heatmap layer is added to the map. This layer visually represents the density of crime incidents, allowing users to easily identify hotspots.The heatmap highlights specific areas in San Francisco with high crime rates, providing a quick overview of crime distribution. This visualization is crucial for identifying regions that may require increased law enforcement presence or further investigation.

V.(ii) Hotspot Analysis

The initial step was to project all the data onto the map, including data from all years and instances of selected crimes. Given the large volume of points, many overlapping, the map initially appeared cluttered and unclear. To address this, integration was used to aggregate all crimes that occurred within a radius of 300 meters into single points. This aggregation made it easier to visualize clusters of crime incidents. However, the aggregated points alone did not indicate the number of original crime points they represented. Recognizing the limitations caused by the dataset's size and proximity of data points, an optimized hotspot analysis was conducted. The data was split into separate files for each of the nine **police districts: Southern, Out of SF, Ingleside, Park, Tenderloin, Central, Mission, RichMond, Taravel.** A geodatabase was created to store nine feature classes, each representing one police district. These files were then used to generate individual hotspots, with each district's crime data visualized separately using a **dot plot**.

The goal of the hotspot analysis was to identify locations with the highest frequency of crime occurrences. This aggregated visualization, encompassing data from all selected years and crimes, allowed for the identification of crime-prone areas. While the data aggregation presents a historical view, it provides a useful hypothesis: crimes today are more likely to occur in the previously identified hotspots.

Hotspot analysis goes a step beyond heatmaps by statistically identifying areas with significantly higher occurrences of incidents compared to other areas. This method helps in pinpointing precise locations that experience unusually high crime rates.

How Hotspot Analysis Works:

Data Aggregation: Crime incidents within a certain radius (e.g., 300 meters) are aggregated into single points. This helps in reducing data clutter and highlights areas with dense crime occurrences.

Weighting Points: The size of each aggregated point is weighted to reflect the number of incidents within the radius. Larger points indicate higher densities. Statistical tests (e.g., Moran's I) are conducted to determine if there is significant spatial autocorrelation, meaning the data points are not randomly distributed but instead show significant clustering.

VI. Results

After performing the data processing and data cleaning and analysis, the results and insights obtained were divided into three major categories. The first category consisted of the results obtained from statistical analysis obtained from folium Heatmaps. The second category consisted of the results obtained from the Crime incidents per year based on category analysis followed by the number of active crime reports registered as per that police district. As shown in the below figures.,

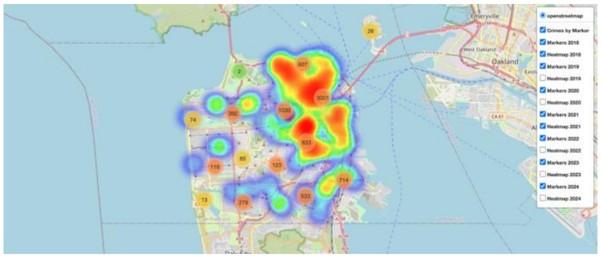


Fig3: Showing heat maps and markers distribution of   crimes in that area as per year.

Fig3 shows the Marker clusters(Markers 2018, 2019,..2024) added to the map to group nearby crime incidents. This feature reduced cluster and made it easier to explore specific areas of interest. Clicking on a cluster revealed detailed information about the aggregated incidents where one data point in the cluster contains the data like which you can see in fig.4

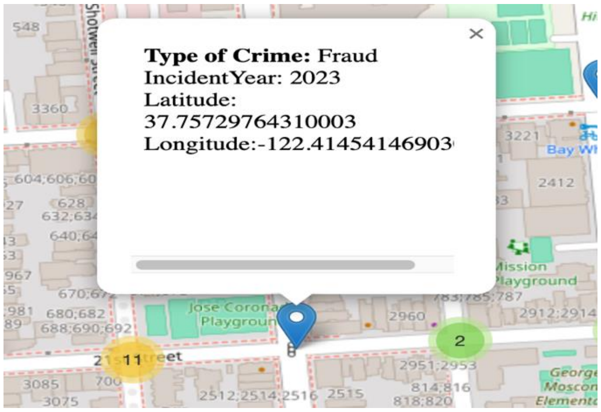

Type of Crime: Fraud
IncidentYear: 2023
Latitude:
37.75729764310003
Longitude:-122.41454146903

Fig. 4:Crime Data as per a single cluster in a region

Heat Map: A heat map layer was included to visualize the density of crime incidents across San Francisco. The heat map provided an intuitive representation of crime hotspots, with warmer colors indicating higher crime densities and the graph was analyzed to reflect results based on the year that you select from 2018 to 2024

The Second category consists of the results analyzed/ concentrated mainly on the Crime incidents and Active cases as per that police district region the data is plotted in an **interactive visualization** using dot plot and bar charts. **Fig 3** shows the dot plot created in Flourish is embedded directly into the Folium map. This integration allowed users to explore spatial and temporal aspects of crime data simultaneously. Custom HTML and CSS were added to enhance the presentation, including a header and footer for context and attribution. Custom CSS was applied to improve the visual appeal of the map. The header provided a title for the visualization, while the footer included data source information and credit to the visualization team

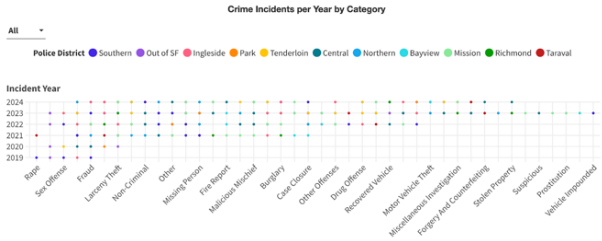


Fig 5: Dot plot that indices crime per year based on category in that police district.

The above **Fig5** displays the distribution of crime incidents over time. Each dot represented an individual crime, providing a granular view of how crime rates fluctuate across different years and those particular police district. This visualization highlighted specific periods with higher crime activity and allowed for easy identification of outliers and trends.

Fig 6 shows the Bar graph illustrates the frequency of different crime categories over time. This comparative analysis showed the prevalence of active crime records and identified periods with notable increases or decreases in crime categories. The bar graph was instrumental in understanding the relative distribution of crime as per area and shows which area is unsafe for the tourists while traveling.

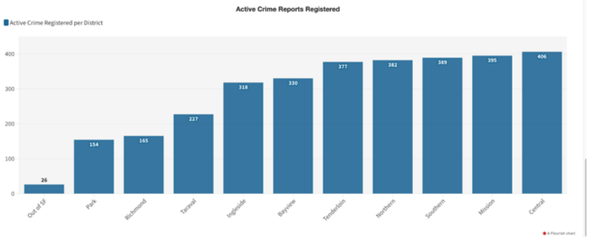


Fig6: Active Crime records registered

VII. Limitations

The visualization techniques employed in this project—primarily heatmaps and hotspot analysis—have several limitations. Heatmaps, while effective for showing general areas of high crime density, can obscure specific details due to the overlapping of incident points in areas with high crime density. This can make it difficult to identify precise hotspots or differentiate between closely situated incidents. Additionally, heatmaps provide a static view, failing to capture the temporal evolution of crime patterns, which is crucial for understanding how hotspots shift over time.

Hotspot analysis, though useful for identifying high-crime areas, relies heavily on the chosen parameters, such as the distance threshold for clustering incidents. Incorrect parameter settings can either oversimplify the data by merging too many points or create an overabundance of insignificant hotspots. The method also tends to aggregate incidents, potentially masking the nuances of specific crime types or their temporal fluctuations.

Moreover, both visualization methods lack integration with other relevant data sources, such as socioeconomic and environmental factors, which are important for a comprehensive understanding of crime patterns. Without these contextual layers, the visualizations can offer a limited perspective, focusing solely on where crimes occur without providing insights into why they occur in those locations.

The accuracy and completeness of the underlying data are also critical. Missing or erroneous records can significantly skew the visualizations, leading to misleading conclusions. Ensuring high-quality data collection and preprocessing is essential to mitigate these issues.

Future enhancements could involve incorporating dynamic visualizations to capture temporal changes, refining parameter selection methods for hotspot analysis, and integrating additional data layers to provide a more comprehensive view of crime patterns. By addressing these limitations, the analysis can yield more actionable insights, aiding in more effective crime prevention and intervention strategies.

VII. Future Work

Integration of Real-time Safety Assessment:

Inspiration from "Safe Spot" App: Implement a feature similar to the "Safe Spot" app, which provides real-time safety assessments of specific areas based on current crime data. Users could receive immediate feedback on the safety of their current or intended location, categorized by the type of crime prevalent in the area.

Real-time Data Integration: Implementing real-time data feeds to ensure the most current crime data is available.

Categorical Safety Indicators: Developing algorithms to classify areas as safe or unsafe based on various crime categories, such as burglary, assault, etc.

User Interface Enhancements: Creating an intuitive and accessible interface where users can easily check the safety of an area, similar to the "Safe Spot" app interface shown in the image.

Crime Reporting Feature:

User-Generated Reports: Allow users to report crimes directly through the app. This feature would enable crowd-sourced data collection, improving the real-time accuracy of crime data.

Verification Mechanisms: Implement verification mechanisms to ensure the authenticity of reported crimes, such as requiring photographic evidence or corroboration from multiple users.

Crime Alert Notifications:

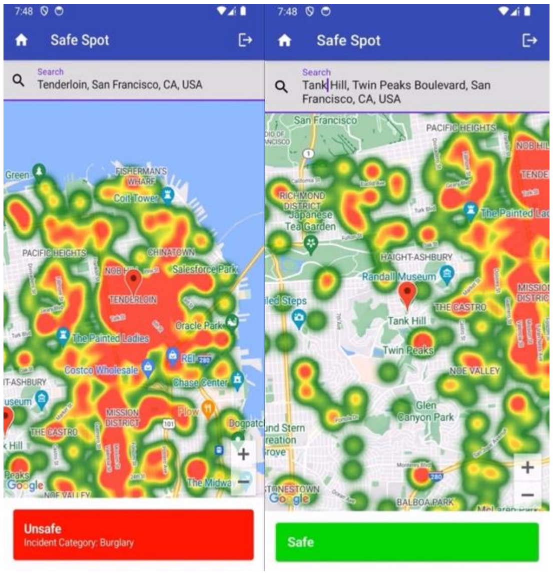
Personalized Alerts: Enable users to set up alerts for specific types of crimes or for particular areas of interest. Users would receive notifications when crimes occur in those specified areas or categories.

Geofencing Technology: Utilize geofencing to alert users when they enter areas with high crime rates or recent crime activity.

Community Engagement and Resources:

Safety Tips and Resources: Provide users with safety tips and resources based on the types of crimes prevalent in their area.

Community Collaboration: Encourage community collaboration through forums or social features where users can share information and strategies for staying safe.



Search
Tenderloin, San Francisco, CA, USA

Unsafe
Incident Category: Burglary

Search
Tank Hill, Twin Peaks Boulevard, San Francisco, CA, USA

Safe

Fig 7: Safe Spot App prototype

By incorporating these features, the visualization tool can become a comprehensive resource for both analyzing crime patterns and providing practical, real-time safety information to the public, enhancing both awareness and proactive safety measures.

.

IX. Conclusion

This project effectively used geographic visualization to analyze violent crime in San Francisco (2018-2024). Results show crime clusters in specific areas. Folium and Flourish were used for interactive visualizations, displaying trends and hotspot analysis. The study identified high-crime zones in targeted police districts, aiding in law enforcement strategies. Cyclical crime patterns were observed, with certain areas consistently high in crime. Different crime types showed varying prevalence across regions. The findings suggest targeted interventions in hotspots for better results. The methods and visuals can be adapted for other cities, aiding in crime prevention efforts. Overall, the project showcases the value of geographic visualization in understanding and combating violent crime, offering a framework for future research and practical applications in public safety planning.

References

[1] Neighborhood Scout. (2014). San Francisco, CA crime analytics. Retrieved March 15, 2018, from https://www.neighborhoodscout.com/ca/sanfrancisco/crime

[2]https://missionlocal.org/2024/01/explore-sf-crime-fell-ever-so-slightly-in-2023/

[3] Venturini, L., & Baralis, E. (2016). A spectral analysis of crimes in San Francisco. Urban GIS '16 Proceedings of the 2nd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics. Retrieved https://doi.org/10.1145/3007540.3007544 from

[4] Mcdowall, D., Loftin, C., & Pate, M. (2012). Seasonal cycles in crime, and their variability. Journal of Quantitative Criminology, 28(3), 389-410 Retrieved from https://doi.org/10.1007/s10940-011-9145-7

[5] Wu, X. (2016). An informative and predictive analysis of the San Francisco Police Department Crime Data. ProQuest Dissertations Publishing. Retrieved from ProQuest database.

[6] Herrmann, C. R. (2015). The dynamics of robbery and violence hot spots. Crime Science, 4(1), 1-14. https://doi.org/10.1186/s40163-015-0042-5

[7] Wheeler, A. P. (2016). Tables and graphs for monitoring temporal crime trends. International Journal of Police Science & Management, 18(3), 159-172. https://doi.org/10.1177/1461355716642781

[8] Weisburd, D., Bernasco, W., & Bruinsma, G. (2009). Putting crime in its place: Units of analysis in geographic criminology. Retrieved http://www.springer.com/gp/book/9780387096872 from

[9] Bernasco, W., & Block, R. (2011). Robberies in Chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. Journal of Research in Crime and Delinquency, 48(1), 33-57. Retrieved https://doi.org/10.1177/0022427810384135