e-ISSN: 2637-0395

Available online at
http://journal.uitm.edu.my/ojs/index.php/BEJ

Built Environment Journal 22(2) 2025, 318 – 329.

Built Environment Journal

Mapping the Property Crime Spatial Pattern in Selangor using Social Media Data Mining and GIS

Siti Hawa Mat Sapuan[1], Nafisah Khalid[2*], Maisarah Abdul Halim[2], Nabilah Naharudin[2], Ainon Nisa Othman[2],

[1]*SWM Environment Sdn. Bhd., No. 3&3A, Jalan Kencana 1 A/25, Taman Pura Kencana, 83300 Batu Pahat, Johor, Malaysia*
[2]*Studies for Surveying Science and Geomatics, Faculty of Built Environment, Universiti Teknologi MARA,40450 Shah Alam, Selangor, Malaysia*

ARTICLE INFO

Article history:
Received 14 April 2025
Revised 09 June 2025
Accepted 19 June 2025
Online first
Published 01 July 2025

Keywords:
property crime
data mining
hotspot analysis
GIS

DOI:
10.24191/bej.v22i2.5935

ABSTRACT

The previous studies had shown that the information extracted from social media can be utilised in locating offenders, establishing probable cause for warrants, and identifying potential witnesses. By integrating social media data mining with Geographic Information System (GIS) techniques, it is possible to map property crime hotspots based on user-generated content. This approach can provide valuable insights to complement officially reported crime data. Through GIS-based spatial analysis, patterns and distributions of crime can be identified, enabling the detection of high-crime areas or hotspots. The ability of hotspot analysis to present crime concentration across a geographical landscape makes it a powerful and practical tool for law enforcement and urban planning. From this study, the total property crime geocoded data that managed to collect is 488 cases consisting of the snatch, burglary, theft, and car theft crime in Selangor. Each crime case has been analysed for its spatial pattern and distribution where for both snatch and burglary crime exhibits clustering while theft crime gives pattern and distribution of random across the study area. The spatial pattern of each crime show that the southern part of Selangor has more crime cases as compared to northern part of Selangor. The red zones show the areas with a very high value of z score indicate significant spatial clustering The findings would be beneficial to the relevant authorities regarding the underreporting of crime cases.

INTRODUCTION

Crime is a global issue that affects both developing and developed nations. Urbanisation, often accompanied by population growth, contributes to the complexity of this problem. As a developing country,

[1*] Corresponding author. *E-mail address*: nafisahkhalid@uitm.edu.com
https://doi.org/10.24191/bej.v22i2.5935

Malaysia is experiencing rapid urban development, which has led to the expansion of urban communities. According to the Department of Statistics Malaysia (DOSM) 2018, the total population grew by 1.1% in the first quarter of 2019, reaching 32.66 million, comprising 29.31 million citizens and 3.35 million non-citizens, an increase from 32.29 million the previous year. This rising trend continued with the population increasing to 32.75 million in 2020, 32.78 million in 2021, and reaching 32.82 million in 2022.

The population growth, particularly in urban areas, often has been linked to rising crime rates due to increased population density and socio-economic disproportions (Castell-Britton, Sigifredo. (2024). A crime must be prevented as it presents a significant challenge for the social, economic, and development of urbanisation all over the world (Canter & Youngs, 2016; Kusuma, Hariyani, & Hidayat, 2019).

Index crimes are defined as crimes that occur frequently and have significant impact, making them reliable indicators for assessing the overall crime situation. Index crimes can be categorised into two (2) types, which are violent crimes and property crimes. Violent crimes include murder, rape, robbery with or without a firearm, gang robbery with or without a firearm, and causing intentional harm or assault. While property crime constitutes house break-in and theft, snatch theft, vehicle theft, and other theft (Ghani, 2017; Sugiharti et al., 2023).

Based on the Department of Statistics Malaysia (DOSM) (2024), the crime index increased 3.2 per cent in 2023 to 52,444 cases as compared to 50,813 cases in 2022. Violent and property crimes recorded an increase of 1.0% and 3.8% respectively in 2023. Property crime has more recorded cases as compared to violent crime, with 40,465 cases in 2022 and 41,991 cases in 2023. The reported cases in 2023 show the decreasing trend as compared to the cases reported in 2019 to 2021.

Accurate crime cases reporting, and management poses a significant challenge for the relevant authorities. Analysing crime data across various geographic locations and crime types is essential for identifying effective, long-term solutions. Nowadays, social media has gained considerable attention due to its real-time and interactive nature. Recent studies have highlighted the potential of platforms such as Twitter and Facebook in capturing current social trends (Vivek & Prathap, 2023; Kaur & Saini, 2024). These platforms generate vast amounts of raw data, much of which is geographically tagged. As a result, social media has emerged as a valuable and highly suitable source for geographically annotated data in crime analysis (Bernabeu-Bautista et al., 2021). Twitter data does not represent all members of the public, nor does it serve all Internet users. However, based on previous studies, Twitter has become a popular platform for research related to crime and has been used in several studies to track crime trends, as well as predict patterns of crime (Wang, Yu, & Liu, 2019; Cano-Marin et al., 2023).

Data obtained from social media platforms can be analysed using a Geographic Information System (GIS). GIS is a computerised tool designed to capture, store, query, analyse, and visualise geospatial data. One of its key applications is in crime mapping and analysis, which begins with the process of geocoding. Geocoding is a process which assigns geographic coordinates to specific locations, effectively mapping them in space. Once crime locations are geocoded, this spatial data serves as the foundation for hotspot analysis, allowing researchers to identify areas with high concentrations of criminal activity. Previous studies have used the hotspot analysis using global and local Moran analysis for various crime and accidents cases, such as drug hotspots analysis and traffic accident cases (Ristea & Leitner, 2020; Hazaymeh et al., 2022; Saraiva & Teixeira, 2023).

https://doi.org/10.24191/bej.v22i2.5935

STUDY AREA

The study area covers the state of Selangor. As shown in Figure 1, Selangor covers a total area of 8,104 km² is one of the most urbanised states in Malaysia, and it has consistently been ranked among the states with the highest number of reported crime cases. This high crime rate may be attributed to its dense population, rapid urban development, and socio-economic diversity.

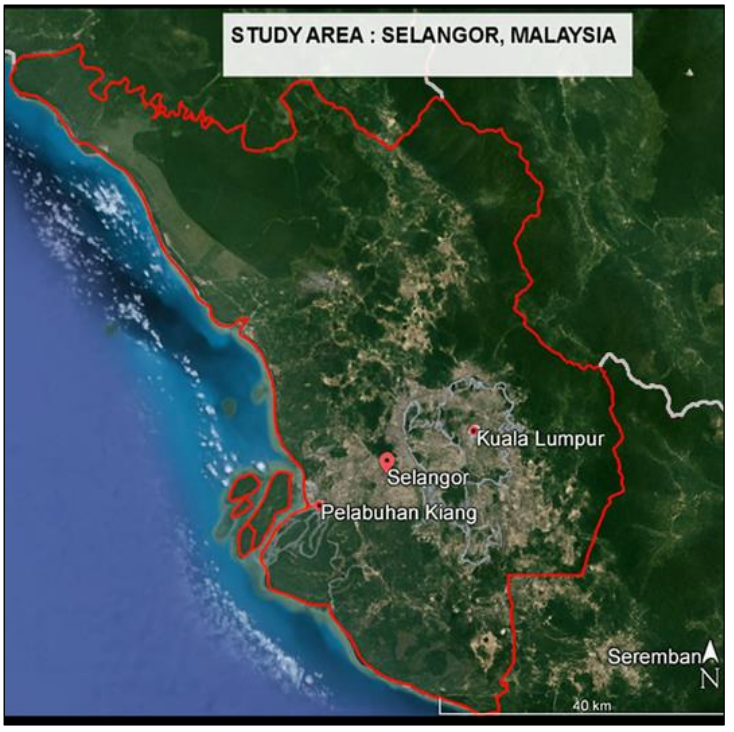


Fig. 1. Study Area

Source: Authors (2025)

MATERIAL AND METHODS

This study consists of five (5) main steps which is data collection includes base map and mining crime data from social media using R software, data pre-processing, data processing, data analysis and map production as shown in Figure 2.

https://doi.org/10.24191/bej.v22i2.5935

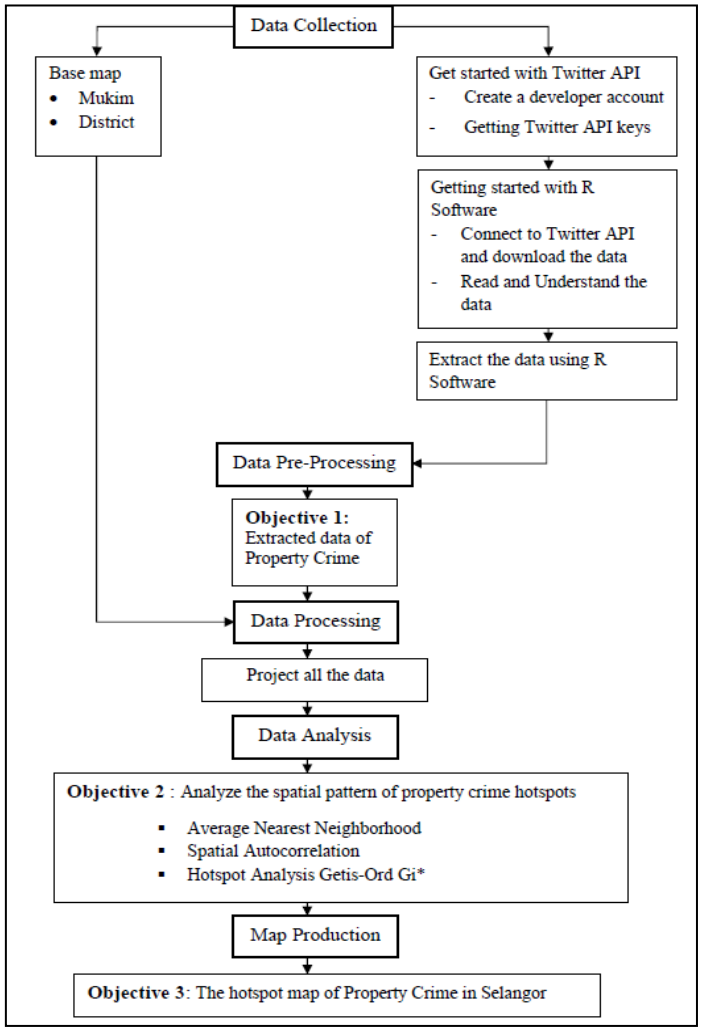


Fig. 2. Overall Methodology

Source: Authors (2025)

Data Collection

In this study, there are two (2) types of data were collected, which are extracted crime data consisting of tweets about crime, timestamp, and location of the tweets themselves. The data used in this study has been retrieved using Twitter API with some procedure to be carried out. Then, the next database map is a map obtained from Subang Jaya Town Council (MPSJ).

Twitter was selected as the principal data source for this study due to the availability of geotagged postings, which allow for the extraction of precise location information from user-generated content. This is essential for mapping crime incidents and conducting spatial analysis, especially when integrated with official geographic. Furthermore, Twitter has been successfully used in previous studies related to crime

detection, disaster response, and urban safety, thereby demonstrating its reliability and effectiveness in similar contexts (Cano-Marin et al., 2023; Vivek & Prathap, 2023).

(i) Property Crime Cases

Property crime cases, the detail of a crime location, and time happened are essential in this study. Since the data is from social media Twitter, there was a limitation while extracting the data where Twitter only returns data from the past six until nine days form the current date and cannot retrieve the previous data. So, a schedule for collecting data was created where every seventh day, starting from the data was collected, there was data extracted. This way can prevent the gap between the week's data. Here the characteristics of extracted data including duration of the tweet, radius of the study area and keywords used

(ii) Getting the Twitter Data

For getting the Twitter data, there was a command assigned that gave instructions to the software to carry out the process of extracting raw Twitter data. The command should be correct, or else the process cannot be run. The command used in this study is shown in Table 1. Figure 3 shows the raw data for some part of the snatch crime only.

Table 1. Example of Command Used

Keyword	Command
Ragut	ragut9.4<-search_tweets("ragut OR diragut OR kes ragut",n=18000,geocode="3.273118,101.527369,100mi", retryonratelimit = TRUE, include_rts = FALSE)

Source: Authors (2025)

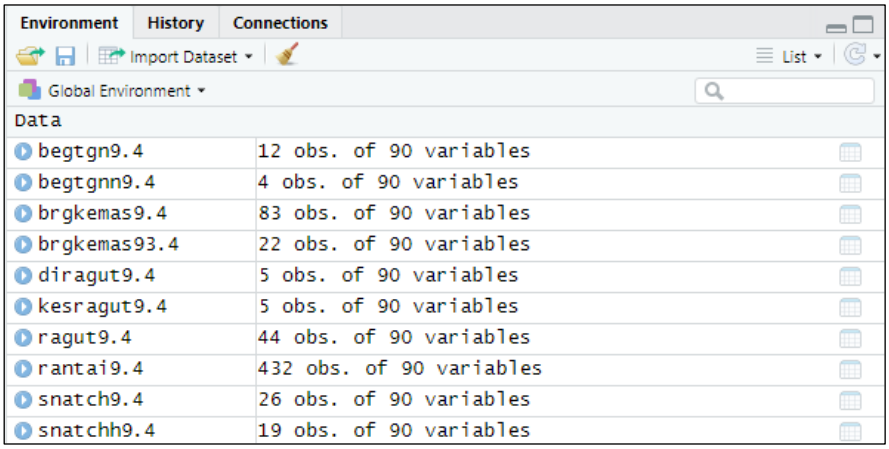


Fig. 3. Raw Data Extracted from Social Media

Source: Authors (2025)

(iii) Base Map Data

The base map of Selangor was obtained from secondary sources. The data was available in shapefile format, and the coordinate system of the map is in Kertau RSO. For this study, coordinate conversion from

Kertau RSO to WGS 84 was necessary for further processing. The spatial information contained in the map was the mukim and the district of Selangor.

Data Preprocessing

Data pre-processing is a method for cleaning the data extracted from Twitter, as it helps eliminate unnecessary or unrelated data that is often collected when extracting Twitter data. The purpose of this phase in this study is to eliminate bad data, such as redundant, incomplete, or incorrect data and begin to tabulate the usable data. So, this step cleaned and filtered the data by removing columns, modifying columns, removing duplicate values, dealing with missing values, and filtering the coordinates.

Data Processing

Data processing requires the collection and manipulation of data to produce useful information and data analysis. In this study, this phase includes the steps to prepare a database to place the processed feature dataset, assign the coordinate systems. For this study, the property crime data were transformed from Excel into shapefile format by displaying the XY data and projecting the data using the Project tools. The dataset was then modified based on the coincidence of distance data using the Integrate tools and then made as weighted point data using the Collect Event tools before it could be used for further analysis. All this step makes use of the software ArcGIS application. The steps that had been taken in this study were: (i) creating a database, (ii) assign the coordinate system of the framework, (iii) registration of the base map, and (iv) plot the cleaned crime data.

Data Analysis

This study used the tools in the ArcGIS application, which are the Average Nearest Neighbour and Spatial Autocorrelation tools. The average nearest neighbour tool produces an analysis that measures the average distance between each feature and its nearest neighbours. If the average distance is less than the average for a hypothetical random distribution, the distribution of the dataset being analysed is considered to be clustered. If the average distance is greater than that of a hypothetical random distribution, the distribution of datasets is considered to be dispersed. The average nearest neighbour index (ANN) can be used to identify whether the distribution is clustered or dispersed.

For Spatial Autocorrelation tools, it was used to determine the correlation between neighbouring observations, to identify patterns and spatial clustering rates between neighbouring districts. It produces an interpretation based on the z-score and p-value, indicating whether the difference is statistically significant or not, based on the number of features and variance for the values of the data. The index value then is interpreted in terms of the null hypothesis, stating whether the attribute or observed pattern is randomly distributed or not. The null hypothesis is not rejected if the p-value is not statistically significant shows that the attribute is randomly distributed. In contrast, the null may be rejected if the p-value is statistically significant, either the z-score positive or negative shows that the spatial distribution is clustered or dispersed, respectively.

Map Production

For the data visualisation, the hotspot map was generated using Getis-Ord Gi* analysis. The map produced contains the information of the north arrow, the coordinate system used, the legend, and the grid of coordinates. The legend comprises the GiZScore value results from Hot Spot Analysis tools with five (5) classifications and the administrative boundary of Selangor.

https://doi.org/10.24191/bej.v22i2.5935

RESULTS AND ANALYSIS

Property Crime Data Extracted from Social Media

Data were collected over fourteen (14) weeks in March 2020, as shown in Table 2. It can be seen that week one (1) recorded the highest number of cases. This suggests a significant concentration of cases at the beginning of the observation period.

Table 2. Summary of Data Collected

Week	Snatch	Burglary	Car Theft	Theft	Total
1	28	11	6	8	53
2	10	14	6	7	37
3	12	12	2	9	35
4	16	12	0	4	32
5	20	8	1	5	34
6	15	4	4	2	25
7	11	4	0	6	21
8	20	10	0	9	39
9	23	5	1	6	35
10	23	6	2	6	37
11	21	5	2	8	36
12	20	4	0	4	28
13	20	9	1	6	36
14	24	8	1	7	40
Total	263	112	26	87	488

Source: Authors (2025)

Analysing Spatial Pattern

In this section, the Spatial Autocorrelation for each type of crime was analysed as shown in Table 3 to Table 6. The analysis was conducted based on the report generated by the software after processing was completed. The report determined whether the distribution of crime is clustered, dispersed, or random. Then the null hypothesis was also discussed, whether it was rejected or not, based on the z-score and p-value.

The spatial autocorrelation report indicates that burglary, snatch theft, and car theft are clustered in their distribution. These findings are consistent with those reported by Ahmad et al. (2024), who revealed that the Moran's Index analysis, conducted from 2015 to 2020 across regions in Selangor, showed positive spatial autocorrelation. This suggests a tendency for property crimes to cluster within specific areas.

https://doi.org/10.24191/bej.v22i2.5935

Table 3. Burglary Spatial Autocorrelation Report

Spatial Autocorrelation		
Maximum Correlation at 2nd iteration: 24,300m		
Moran's Index	0.024	p-value is statistically significant with positive z-score: Clustered
z-score	2.075	Null hypothesis
p-value	0.038	Rejected

Source: Authors (2025)

Table 4. Snatch Spatial Autocorrelation Report

Heading Spatial Autocorrelation		
Maximum Correlation at 2nd iteration: 24,300m		
Moran's Index	0.032	p-value is statistically significant with positive z-score: Clustered
z-score	1.637	Null hypothesis
p-value	0.094	Rejected

Source: Authors (2025)

Table 5. Theft Spatial Autocorrelation Report

Heading Spatial Autocorrelation		
Maximum Correlation at 6th iteration: 59,000m		
Moran's Index	0.168	p-value is not statistically significant : Random
z-score	1.197	Null hypothesis
p-value	0.231	Not rejected

Source: Authors (2025)

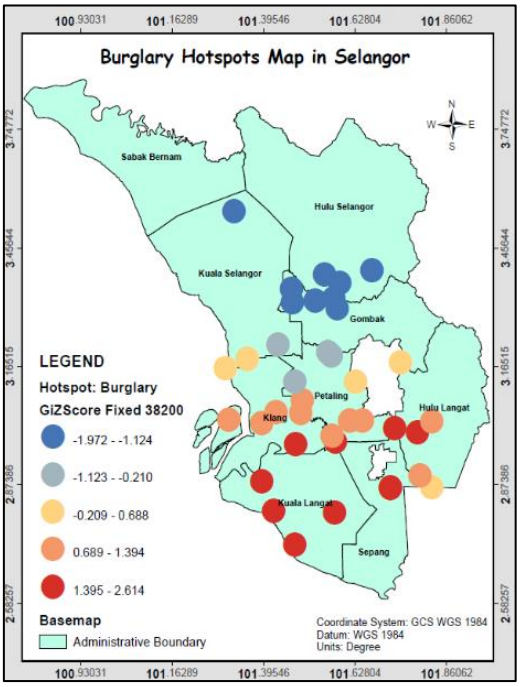
Table 6. Car Theft Spatial Autocorrelation Report

Heading Spatial Autocorrelation		
Maximum Correlation at 2nd iteration: 24,300m		
Moran's Index	-0.084	p-value is statistically significant with positive z-score: Clustered
z-score	- 0.335	Null hypothesis
p-value	0.737	Rejected

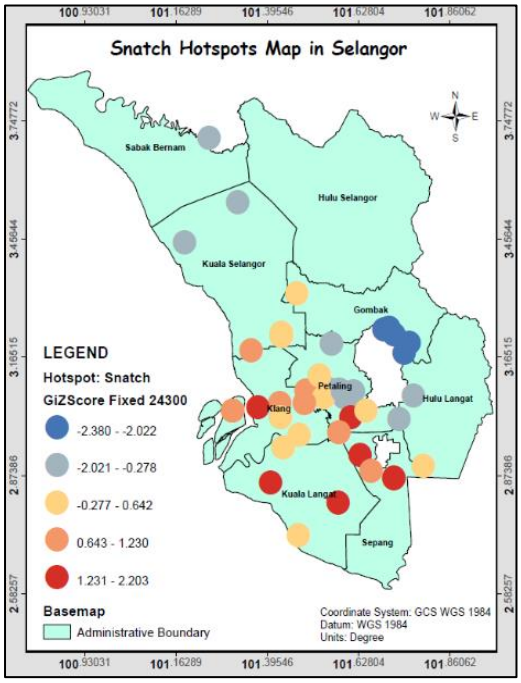
Source: Authors (2025)

Hotspot Map of Property Crime

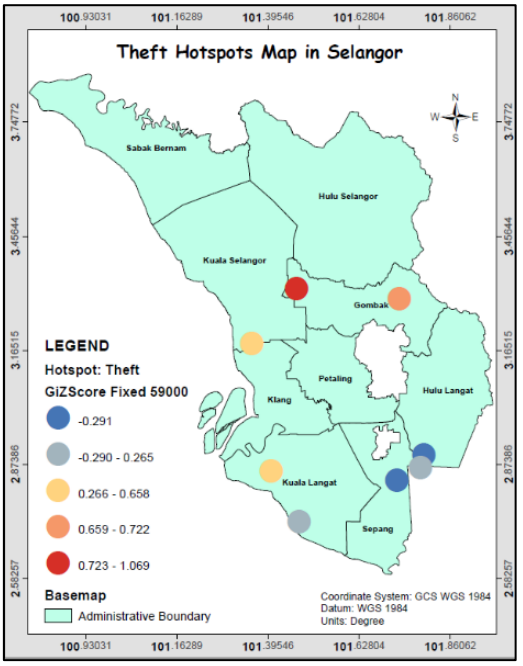
Figure 4 shows the property crime hotspot map in Selangor. Figure 4 (a) shows a burglary hotspot map. It can be seen that the hot spots, displayed in red and the cold spots, displayed in blue, are presented by points where the values were $2.614>z>0.689$ and $-1.972<z<-0.210$ standard deviations, respectively. Burglary crime is frequently observed in the southern part of Selangor, with the area having the most clustered hotspots being Kuala Langat, followed by Petaling, Klang, Hulu Langat, and Sepang.



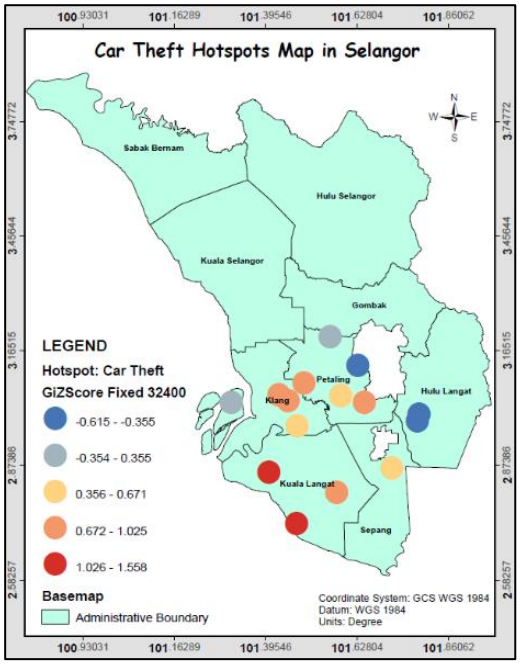
(a)



(b)



(c)



(d)

Fig. 4. Property Crime Hotspot Map in Selangor. Burglary Hotspot Map (a), Snatch Hotspot Map (b), Theft Hotspot Map (c) and Car Theft Hotspot Map (d)

Source: Authors (2025)

https://doi.org/10.24191/bej.v22i2.5935

From the map in Figure 4(b), the snatch hot spots are presented by points with value 2.203>z>0.643 standard deviations, indicating statistically significant clustering of high crime occurrences. These hotspots are primarily located in the southern part of Selangor, specifically in the districts of Kuala Langat, Sepang, Petaling, and Klang. This suggests that snatch thefts are more frequent and concentrated in these areas compared to others.

On the other hand, the cold spots, where snatch thefts are significantly less frequent, are represented by z-scores ranging from -2.380 to -0.278. These areas include the districts of Gombak, Kuala Selangor, and Sabak Bernam. The negative z-scores in these districts indicate a lower-than-average occurrence of snatch thefts, suggesting relatively safer zones in terms of this specific crime.

The theft crime hotspot map in Figure 4(c) shows that the area that requires more attention from the local government and Royal Malaysia Police in managing the crime event was at the range z-score of 0.659<z<1.069, which is in the district of Gombak. While based on the range z-score of -0.265<z<-0.291, the area that required the least attention was at Sepang and Hulu Langat district.

The map in Figure 4(d), indicates that the car theft crime event hotspots are concentrated in the southern part of Selangor, which is Kuala Langat, followed by Petaling and Klang. The red zones show the areas with a very high value of z-score, indicating significant spatial clustering at the range of 0.672<z<1.558. These findings are consistent with those reported by Ahmad et al. (2024),where the property crimes were clustered within specific areas.

CONCLUSION

The determination of crime location from social media helps to analyse the current trend of crime in an area. For this study, the study area was the Selangor state. For data collection, Twitter API and R Studio were used to collect the location of property crime data in the Selangor area. There are five phases in this section, excluding the phase of preliminary studies. The phases are data collection, data pre-processing, data processing, data analysis, and lastly, the production of hotspot maps. In summary, the total data collected and being used for processing is 488 cases consisting of snatch, burglary, car theft, and theft. For average nearest neighbours, each crime was determined by their pattern and distribution based on the average distance between neighbourhood features, locational and crime rate similarity for Global Moran's, respectively. From the output of the analysis tool, there is an underlying spatial process at work, presented by the indicators for the statistical significance based on the z-score.

ACKNOWLEDGEMENTS/FUNDING

The authors would like to acknowledge the support of the College of Built Environment and Research Management Centre, Universiti Teknologi Mara (UiTM) Shah Alam, Selangor, Malaysia and Subang Jaya Town Council (MPSJ) for providing the facilities and financial support for this research.

CONFLICT OF INTEREST STATEMENT

The authors agree that this research was conducted in the absence of any self-benefits, commercial or financial conflicts and declare the absence of conflicting interests with the funders.

AUTHORS' CONTRIBUTIONS

Siti Hawa Mat Sapuan carried out the research together with her supervisor, Nafisah Khalid. Nafisah Khalid provided the idea, revised the article and supervised the research progress. Maisarah Abdul Halim, Nabilah Naharudin and Ainon Nisa Othman providing support in extracting data from social media and analysing the crime hotspot pattern.

REFERENCES

Ahmad, N., Lim, K. Y., & Rahman, A. A. (2024). Criminological Insights: A Comprehensive Spatial Analysis of Crime Hot Spots of Property Offenses in Malaysia's Urban Centers. *Forum Geografi, 38*(1), 2024, 94–109. https://doi.org/10.23917/forgeo.v38i1.4306

Bernabeu-Bautista, Á., Serrano-Estrada, L., Perez-Sanchez, V. R., & Martí, P. (2021). The Geography of Social Media Data in Urban Areas: Representativeness and Complementarity. ISPRS International Journal of Geo-Information, 10(11), 747. https://doi.org/10.3390/ijgi10110747

Cano-Marin, E., Mora-Cantallops, M., Sánchez-Alonso, S. (2023). Twitter as a Predictive System: A Systematic Literature Review, Journal of Business Research, 157 (2023) 113561, https://doi.org/10.1016/j.jbusres.2022.113561

Canter, D., & Youngs, D. (2016). Crime and society. Journal of the Academy of Social Science. http://doi.org/10.1080/21582041.2016.1259495

Castell-Britton, Sigifredo. (2024). The Relationship Between Overpopulation and Crime Rates in San Andres Island. Qeios.

Department of Statistics Malaysia. (2024). Crime Statistics, Malaysia, 2024.

Hazaymeh, K., Almagbile, A., & Alomari, A. H. (2022). Spatiotemporal Analysis of Traffic Accidents Hotspots Based on Geospatial Techniques. ISPRS International Journal of Geo-Information, 11(4), 260. https://doi.org/10.3390/ijgi11040260

Kaur, M., Saini, M. Role of Artificial Intelligence in The Crime Prediction And Pattern Analysis Studies Published Over The Last Decade: A Scientometric Analysis. *Artif Intell Rev* **57**, 202 (2024). https://doi.org/10.1007/s10462-024-10823-1

Kusuma, H., Hariyani, H. F., & Hidayat, W. (2019) The Relationship Between Crime and Economics Growth in Indonesia. 2nd ICIEBP The 2nd International Conference on Islamic Economics, Business, and Philanthropy (ICIEBP) Theme: "Sustainability and Socio-Economic Growth" Volume 2019. https://doi.org/10.18502/kss.v3i13.4271

Ristea, A., & Leitner, M. (2020). Urban Crime Mapping and Analysis Using GIS. ISPRS International Journal of Geo-Information, 9(9), 511. https://doi.org/10.3390/ijgi9090511

Saraiva, M., & Teixeira, B. (2023). Exploring the Spatial Relationship between Street Crime Events and the Distribution of Urban Greenspace: The Case of Porto, Portugal. ISPRS International Journal of Geo-Information, 12(12), 492. https://doi.org/10.3390/ijgi12120492

Sugiharti, L., Purwono, R., Esquivias, M. A., & Rohmawati, H. (2023). The Nexus between Crime Rates, Poverty, and Income Inequality: A Case Study of Indonesia. Economies, 11(2), 62. https://doi.org/10.3390/economies11020062

Twitter Inc. (2019). Twitter Brand Resources. Retrieved from Twitter.com: https://about.twitter.com/en_us/company/brand- resources.html

Vivek, M., Prathap, B.R. Spatio-temporal Crime Analysis and Forecasting on Twitter Data Using Machine Learning Algorithms. *SN COMPUT. SCI.* 4, 383 (2023). https://doi.org/10.1007/s42979-023-01816-y

Wang, Y., Yu, W., & Liu, S. (2019). The Relationship Between Social Media Data and Crime Rates in the United States. SAGE Journal. https://doi.org/10.1177/2056305119834585