Huang *et al. Computational Urban Science* (2025) 5:72
https://doi.org/10.1007/s43762-025-00233-6

Computational Urban Science

ORIGINAL PAPER

Open Access

Spatio-temporal cokriging crime predictions using social media data: a multi-type case study in San Jose, California

Yanhong Huang[1], Bo Yang[2], Xiangyu Ren[2], Yujian Lu[1], Minxuan Lan[3] and Xi Gong[4*]

Abstract

Crime prevention requires accurate prediction of the spatial and temporal distribution of criminal activities to effectively allocate law enforcement resources. However, many trending crime prediction algorithms lack comprehensive spatio-temporal structures and often consider only single input variables. This study innovatively using in ST-Cokriging method integrated both historical crime records as the primary variable and crime-related geo-tagged Twitter data as the co-variable for crime prediction. The predictive method has been specifically developed to assess crime risk across three major crime types—street crime, property crime, and vehicle crime—and applied in the San Francisco Bay Area (SFBA), California, a region characterized by high development and heightened crime sensitivity, for both prediction and validation. The results indicate that incorporating social media data into a spatio-temporal statistical method improves the associations between predicted and actual crime risk, reduced the Root Mean Squared Error (RMSE), and enhanced the identification of crime risk areas for both weekdays and weekends across three crime types compared to the method without the co-variable. This study presents a new multi-variable approach to more accurately predict crime, enabling law enforcement proactively address crime of varying nature in urban areas.

Keywords Public Safety, GIS, Spatio-temporal Analysis, ST-Cokriging, Social Media, Urban Systems

1 Introduction

Accurate crime prediction enables a strategic and proactive approach to public safety, allowing law enforcement agencies to allocate their resources efficiently, anticipate criminal activities, and potentially prevent them before they occur (Braga et al., 2014; Chainey et al., 2008; Mohamad Zamri et al., 2021). However, accurate crime prediction is challenging due to the complex and variable nature of spatial and temporal crime patterns. Crime incidents often exhibit significant spatial clustering, making it difficult to generalize predictions across different neighborhoods or cities. Also, temporal patterns can fluctuate during specific times, weekdays or weekends, and special events. Variability in data quality, reporting practices, and the influence of socio-economic factors further exacerbate these challenges, requiring sophisticated computational models to accurately capture and predict crime trends (Ferreira et al., 2012; Wang et al., 2019). Incorporating both spatial and temporal patterns in crime prediction is imperative, as it provides a comprehensive understanding of crime dynamics, significantly increasing the accuracy and effectiveness of predictive models (Du & Ding, 2023; Hu et al., 2018).

*Correspondence:
Xi Gong
xigong@psu.edu
[1] Department of Geography & Environmental Studies, UNM Center for the Advancement of Spatial Informatics Research and Education (ASPIRE), University of New Mexico, Albuquerque, NM 87131, USA
[2] Department of Environmental Studies, University of California, Santa Cruz, CA 95064, USA
[3] Department of Geography and Planning, University of Toledo, Toledo, OH 43606, USA
[4] Department of Biobehavioral Health, Institute for Computational and Data Sciences (ICDS), The Pennsylvania State University, University Park, PA 16802, USA

Springer

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

A variety of crime prediction methodologies have been developed over the past decades. Traditionally, crime statistics and hotspot analyses have relied on single input crime data to identify crime-prone areas (Amerio & Roccato, 2005; Braga et al., 2014). Then, crime prediction methodologies have evolved to integrate multiple variables, recognizing that crime activity is influenced by a complex interplay of different factors (Tasnim et al., 2022). For example, machine learning models that consider crime data and ten urban indicators can increase the accuracy of homicide prediction in Brazil from 2001 to 2020 (Alves et al., 2018). It has been found that the spatio-temporal graph neural network framework could achieve both high predictive accuracy and strong interpretability, outperforming existing models in crime prediction while effectively handling data sparsity and missing information (Tang et al., 2023). Moreover, the transformer-based model could enhance the telecommunication network crime predictions in China and various crimes such as street crime, property crime, and personal offense in the United States (Butt et al., 2025; Shi et al., 2023). Additional studies have explored using geo-tagged Twitter (now known as X) data as additional variables to enhance the modeling of criminal activity distributions (Bendler et al., 2014; Corso et al., 2016; Liu et al., 2022). The results demonstrated that this approach outperformed traditional methods by incorporating these additional variables. However, traditional crime prediction methods often fail to integrate spatial and temporal dimensions, limiting their ability to accurately capture patterns and forecast future crime activities. Recent advancements in methodologies further incorporate both spatial and temporal information into the multivariable crime prediction model to further improve performance. Despite substantial advancements in crime prediction methodologies, there exists a literature gap in using the multivariable spatio-temporal crime prediction model that combines both historical crime data and social media data for crime prediction.

In this study, we developed a new spatio-temporal Cokriging (ST-Cokriging) method to jointly incorporate historical crime call data and Twitter data for enhanced crime prediction in large urban areas such as San Jose, California. By utilizing both traditional crime reports and real-time social media data, our approach aims to overcome limitations of single-source crime modeling and improve spatial coverage and predictive accuracy. The crime call data were used as the primary variable in the ST-Cokriging, while Twitter data—filtered using crime-related keywords—served as the secondary co-variable. We then designed an ST-Cokriging algorithm (detailed in the Methods section) that rigorously accounts for the spatio-temporal statistical structure of both datasets to generate improved crime predictions. Previous research suggests that integrating secondary variables can enhance crime prediction through multi-source data fusion (Yang et al., 2020; Yu et al., 2020).

The integration of social media into predictive models of social dynamics has emerged as a transformative approach, as volunteered geographic information provides real-time insights and a more nuanced understanding of underlying social factors (Ahn & Spangler, 2014; Rousidis et al., 2020; Schoen et al., 2013; Wang et al., 2020). Human activities, as reflected in social media interactions and community behaviors, show a significant correlation with crime patterns, affecting both the prevalence and nature of various criminal activities (Kadar & Pletikosa, 2018; Vomfell et al., 2018). Research has shown that social media can provide real-time insights into human behavior and societal trends, which, in turn, influence crime patterns (Vomfell et al., 2018). Twitter, with its real-time data and rich user content, provides valuable insights into social dynamics, making it ideal for predictive models (Gayo-Avello, 2013; Zhang et al., 2014; Zheng et al., 2018). For instance, Twitter has proven effective in mapping motor vehicle thefts in Mexico City, while property crime like burglaries and larcenies in Montreal were found to correlate with geo-tagged Twitter sentiment from 2011 to 2017 (Da Silva et al., 2019; Piña-García & Ramírez-Ramírez, 2019). Integrating social media data with historical crime data meaningfully improves hotspot prediction—e.g., CrimeTelescope achieved ~5.2% higher accuracy in New York city, while enabling richer, timely situational awareness via interactive maps (Yang et al., 2018). Additionally, Twitter data can augment traditional community channels in South Africa by enabling rapid, transparent crime reporting and data collection that helps detect patterns, support prediction, and guide enforcement where constant policing is impractical (Featherstone, 2013). Collectively, these studies underscore social media as a complementary, real-time data source that strengthens crime monitoring, hotspot prediction, and operational decision-making across diverse contexts.

While various studies have focused on predicting specific crime such as assaults (Liu et al., 2022; Uittenbogaard & Ceccato, 2012), robberies (Chainey, 2013), or burglaries (Piña-García & Ramírez-Ramírez, 2019)), few have explored predictions of a wide range of crime types. Integrating multivariable spatio-temporal methods can provide a more holistic understanding of crime patterns and improve prediction accuracy across different criminal activities. Given the differing crime patterns between weekdays and weekends (Piña-García & Ramírez-Ramírez, 2019; Yang et al., 2020), this study separately tested predictions and validations for three

crime types (street crime, property crime, and vehicle crime) for weekdays and weekends. The large volume of Twitter data from the study area was filtered using keywords tailored to each specific crime type. The algorithm explicitly calculated and considered the unique spatial and temporal auto-dependencies within the spatial and temporal domains, marking the first effort to estimate these differences among street crime, property crime, and vehicle crime in this context. The primary aim is to deepen the understanding of crime dynamics, as different crime types follow distinct patterns influenced by factors such as urban layout, social behavior, and law enforcement practices. Moreover, tailoring predictive models to specific crime types enables law enforcement agencies to implement more targeted interventions and allocate resources more effectively by accurately identifying where and when particular types of crime are most likely to occur.

2 Data and method

2.1 Study area

San Jose, located in the heart of Silicon Valley, California, has emerged as a global hub for high-tech and internet industries, propelling it to become California's fastest-growing economy since the 1990s (Zandiatashbar & Kayanan, 2020). It is the third-largest city in the state and the 12th largest city by population in the United States, boasting a population of approximately 1.01 million (Berry-James et al., 2020). Despite its reputation as a prosperous tech hub, the city's economic growth and population accumulation have brought urban challenges, including an escalating crime rate that correlates with a widening wealth gap (Yuan et al., 2022, 2024). From 2013 to 2022, the violent crime rate in San Jose increased from 326.6 to 516.8 per 100,000 people, while its property crime rate reached 2597.5 per 100,000 population in 2022, surpassing the national level of 1954.4 (SJPD, 2023b; U.S. Department of Justice—Federal Bureau of Investigation, 2023). This juxtaposition of rising prosperity alongside increasing crime rates underscores the complex nature of urban crime dynamics. Factors such as socio-economic changes, neighborhood characteristics, and evolving social trends intricately weave together, influencing the city's crime patterns. This context presents a rich and multifaceted backdrop for examining the spatial and temporal dimensions of crime in San Jose, offering insights into how economic progress and urban development intersect with public safety challenges.

We applied the Urban Growth Boundary of San Jose, provided by the Bureau of Land Management, County of Santa Clara, as the study boundary to extract the San Jose urban region (Fig. 1). Established in response to rapid urban expansion between 1950 and 1970, this boundary aims to regulate sprawl and mitigate environmental impacts. The San Jose urban region is also an important planning framework in the San Jose 2040 General Plan, to which this research may provide valuable urban development insights. This boundary covers an area of 370.3 km^2, encompassing most of the population activities and economic areas, and encapsulates 99.3% of all crime data in San Jose.

2.2 Crime data and preprocessing

The crime data for San Jose was sourced directly from the San Jose Police Department's phone call records, with records in 2014 selected for the case study (SJPD, 2023a). Dataset in 2014 was chosen for the study due to the early stages of Silicon Valley's economic growth, which resulted in significant increases in crime during this period. The dataset includes a catalog of phone calls reporting various crime across the city, with each crime's address digitized. Each record is associated with a specific crime incident and includes key attributes such as call type, crime location, crime activity type, weapon involvement, and timestamp information. The crime locations reported in phone call records were geocoded using Geoapify (Geoapify, 2024), providing precise geographical coordinates along with temporal and crime reporting information for further spatio-temporal analysis. In 2014, San Jose recorded a total of 313,817 crime-related phone calls.

Three major categories of crime have been selected from the dataset: street crime, property crime, and vehicle crime. The street crime included 2,010 records encompassing "strong arm robbery", "strong arm robbery (combined event)", "armed robbery", "robbery", "armed robbery (combined event)", "purse snatch robbery", "robbery, gang related", "assault with deadly weapon", "assault with deadly weapon (combined event)", "assault", "assault with deadly weapon, gang", "assault and battery", and "assault on an officer". Property crime incorporated 11,563 calls, with categories ranging from "burglary report", "burglary, "vehicle burglary", "theft", "grand theft", "petty theft prior conviction", "theft of recyclables", "petty theft", and "theft, gang related". Finally, vehicle crime, with 12,778 records, included "misdemeanor hit and run", "felony hit and run", "stolen vehicle", and "stolen vehicle gang related".

Spatial and temporal patterns of criminal activities vary between weekdays and weekends as the have different spatial and temporal patterns (Newton et al., 2008; Uittenbogaard & Ceccato, 2012; B. Yang et al., 2020). A higher frequency of the three identified crime types is observed on weekends—specifically Saturday (day 6 in a week) and Sunday (day 7 in a week)—compared with weekdays, as reflected in their spatial distributions. Given

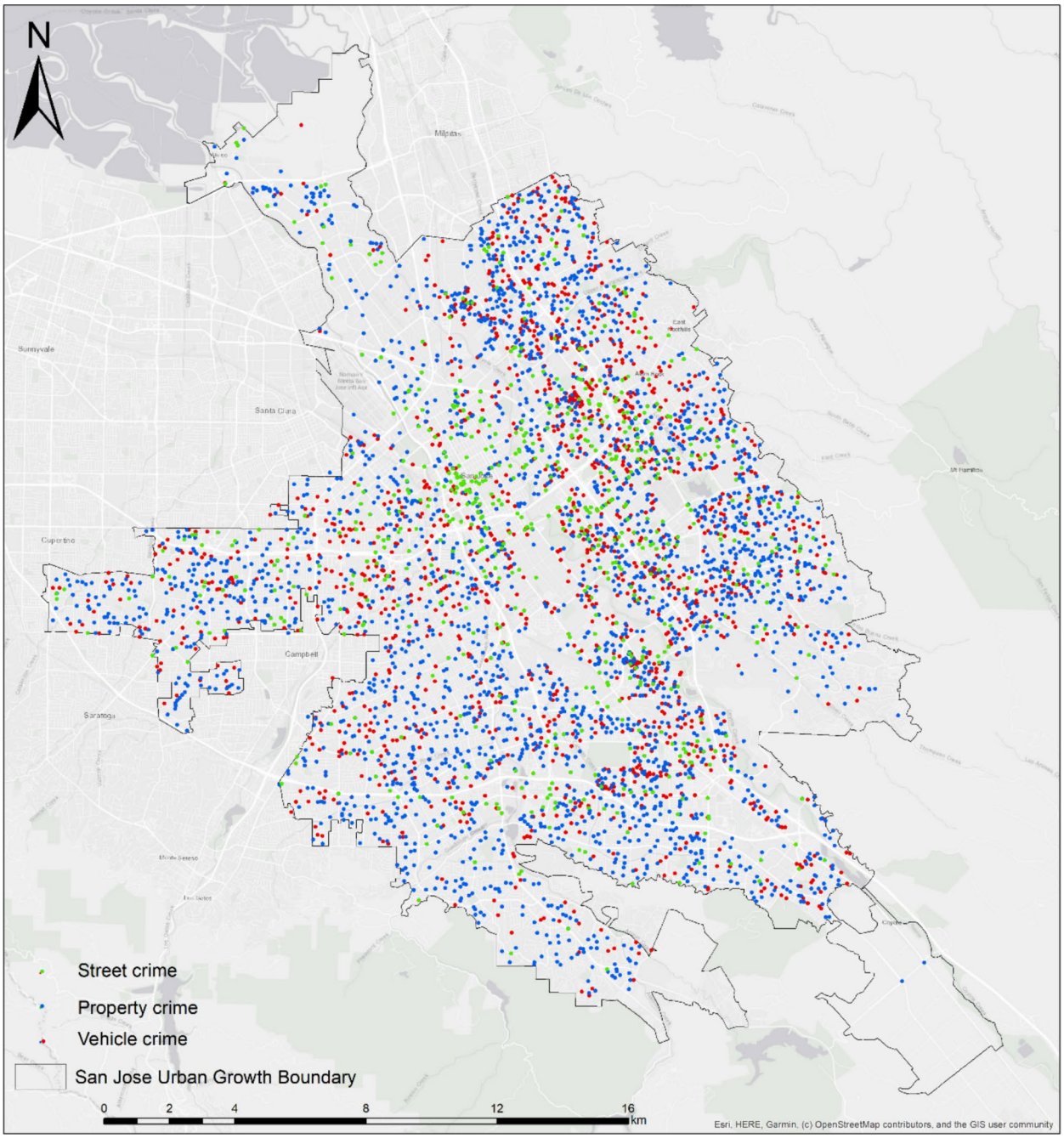

Fig. 1 Street crime, property crime, and vehicle crime distributions overlay with city map in San Jose in 2014

these variations, it is crucial to categorize crime data distinctly into weekday and weekend groups for more accurate analysis. Following the division adopted in previous studies, the weekday category spans from Monday at 12:00 AM to Friday at 11:59 PM local time, enabling a comprehensive analysis of crime patterns during regular workdays. The weekend category covers the period from Saturday at 12:00 AM to Sunday at 11:59 PM local time, capturing the distinct dynamics of crime incidents that typically occur during these days. The data for the three crime types was aggregated to generate a crime risk map using the kernel density function (DeVeaux et al., 1999). A cell size of 100 m was applied for the kernel density estimation (KDE) to optimize both the predictive performance and practical applicability in policing and crime prevention strategies (Chainey, 2013; Du & Ding, 2023).

A fixed search radius (bandwidth) of 2 km was applied to ensure appropriate smoothing and capture broader spatial patterns in crime activities, covering approximately 1,257 grid cells (100 m × 100 m) within the KDE surface. This approach transformed discrete crime points into continuous risk maps, with each pixel indicating crime risk across the study area (Chainey, 2013).

2.3 Social media data collection and filtering

In this study we take consideration of the Twitter data as the co-variable in ST-Cokriging to enhance the crime prediction results, as the Twitter data has been widely used for crime activities predicting and social behavior modeling (Gayo-Avello, 2013; Lan et al., 2019; Vomfell et al., 2018). The Twitter data employed in this study was collected using the Twitter Academic API, a resource offered by Twitter that allows for extensive and granular data gathering. For this study, our focus was on geotagged tweets originating from San Jose in 2014. Geotagged tweets are those where the user has opted to include their geographic locations (longitude/latitude) at the time of posting, enabling us to capture spatial information tied to each tweet. There were 1,048,575 geotagged tweets collected via the API in San Jose, California in 2014. Each geo-tagged tweet was logged with its tweet ID, username, creation time, and full text.

Since the historical crime data is associated with specific crime types, a keyword-based strategy was employed to filter tweets corresponding to each type of criminal activity. We meticulously developed a set of keywords for each crime category to accurately capture crime-related tweets. The design of our dataset ensures a close reflection of the unique characteristics of each crime type studied, along with their corresponding spatio-temporal patterns. For instance, to identify tweets related to street crime, the dataset was filtered using keywords such as "assault," "robbery," "robbed," and "assaulting", etc. The comprehensive list of the keywords used is presented in Table 1 below. Following the keyword filtering, a manual review was conducted to further refine the selection of relevant tweets. These combined processes helped pinpoint 190 tweets related to street crime, 720 tweets related to property crime, and 408 tweets related to vehicle crime (Lal et al., 2020).

2.4 ST-Cokriging method

The Kriging method is a traditional geostatistical interpolation technique that models spatial autocorrelation to minimize estimation variance. Cokriging is an extension introduced by Journel & Huijbregts (1978) improves interpolation by incorporating spatially correlated secondary co-variable, enhancing predictions in environmental and resource modeling (Goovaerts, 1997). ST-Cokriging algorithm adopts a spatio-temporal statistical model to consider multi-sources. The historical crime data are considered as the primary variable of the prediction, while the auxiliary data that correlated with crime were modeled as the co-variable. We innovatively developed the ST-Cokriging to incorporate filtered Twitter data as the co-variable by adding the secondary co-variable with spatiotemporal structure, which consider social behavior patterns in the spatial and temporal domain. The spatio-temporal structure for ST-Cokriging considering the both space and time aspects were modeled using the mathematical framework (Eq. 1):

Table 1 Keywords and tweet filtering process for the three crime types

Crime type	Keywords	Filtered tweet Count (Keywords filtering)	Re-filtered Tweet Count (Manual Review)
Street crime	"assault", "assaults", "assaulting", "assaulted", "rob", "robbed", "robs", "robbing", "robbery", "robberies", "robber", "robbers	3229	720
Property crime	"theft", "thief", "thieves", "thefts", "stole", "steal", "stolen", "break in", "break-in", "breaking-and-entering", "forced entry", "unlawful entry", "intruder", "invade", "invasion", "broke in", "broken in", "breaks in", "burglary", "larceny", "larcenies", "burglarize", "burglaries", "burglarized", "burgled", "Burglary-in-Progress", "burglarizing"	1035	408
Vehicle crime	"hijack", "Joyriding", "Carjacking", "joyride", "carjack", "hijacking", "hijacked", "hit and run", "hit & run", "hit and killed", "hit and runs", "hit-and-run", "hit&run", "car accident", "car injury", "car injuries", "car victim", "crash", "driving drunk", "drunk driver", "over speeding", "speeding", "rear ending", "road accident", "road rash", "vehicle accident", "vehicle crash", "vehicle injury", "vehicle injuries", "vehicle victim", "hit a car", "car collision", "vehicle collision", "hit a person", "hit a parked car", "speeding", "break", "pedestrian crash" ["hit", "hitting", "hits"] and* ["run", "running", "ran", "escape", "escaped", "flee", "fleeing"] ["car", "cars", "vehicle", "vehicles", "motor", "auto", "motorcycle"] and ["theft", "thief", "thieves", "thefts", "stole", "steal", "stolen", "gone", "GPS tracking"]	588	190

* The tweets need to have at least one keyword from each of the 2 groups connected by "AND"

$$\widehat{Z}_1(x_0,t_0)=\sum_{i=1}^{T}\sum_{j=1}^{N_i}\alpha_{ij}Z_1(s_{ij},t_i)+\sum_{k=1}^{M}\beta_k Z_2(u_k,t\prime) \quad (1)$$

where $\widehat{Z}_1(x_0,t_0)$ is the predictor of the criminal activities at pixel location x_0 and at time t_0; The time point t_0 is the predicting time point in the temporal domain. $Z_1(s_{ij},t_i)$ denotes the primary variable of the historical crime data at location s_{ij} at time series $t_i, j=1,\ldots,N_i, i=1,\ldots,T$; $Z_2\left(u_k,t^{'}\right)$ denotes the co-variable of the crime related Twitter data at location u_k at time $t^{'}$ and $k=1,\ldots,M$. Two sets of weights $\left\{\alpha_{ij}:j=1,\ldots,N_i;\ i=1,\ldots,T\right\}$ and $\left\{\beta_k:k=1,\ldots,M\right\}$ were obtained by solving the spatio-temporal covariance matrix for the best unbiased linear predictor. In this ST-Cokriging framework, the input data consist of: (1) Primary variable – KDE-generated raster surfaces representing the spatial intensity of historical crime data over time, and (2) Secondary co-variable – KDE surfaces derived from crime-related Twitter data. The output of the ST-Cokriging model is a spatiotemporal prediction surface—a raster grid estimating the crime intensity at each location x_0 and time t_0. For different types of crime activities, the primary variable was derived from the historical crime maps at specific time points, while the co-variable represents the spatial–temporal pattern of crime-related tweets at a predicting time. The spatio-temporal semi-variogram are calculated first to estimate the different spatio-temporal patterns of the street, property, and vehicle crime, respectively.

The linear system in ST-Cokriging is solved as a linear system to get the weights for the prediction. In particular, the Cokriging variance for the spatio-temporal prediction predictor $\widehat{Z}_1(x_0,t_0)$ can be computed as stated in previous literature (Cressie & Huang, 1999; Kyriakidis & Journel, 1999; Snepvangers et al., 2003). Also, the uncertainty associated with the prediction is also calculated based on the spatial–temporal distribution of the data. This methodology allows for the prediction of crime occurrences in a given spatial and temporal context considering the relationship between crime data and auxiliary tweet variables. This study models the different types of crime data, including street crime, property crime, and vehicle crime.

To estimate the spatio-temporal structures for different types of crime vary spatial and temporal in dependence, the study employs separate spatial and temporal semi-variograms to estimate the spatio-temporal structures for crime types within and between a primary variable and the co-variable. A semi-variogram is a statistical tool used in spatial statistical analysis to quantify the spatial and temporal correlation or dependence between data points at varying distances. Spatial semi-variograms are calculated using Eq. 2:

$$\gamma(h_s)=\frac{1}{2N(h_s)}\sum_{i=1}^{N(h_s)}\left[Z(s_1,t_1)-Z(s_1+h_s,t_1)\right]^2 \quad (2)$$

where $N(h_s)$ is the number of randomly chosen pairs for same type of the crime within the fixed spatial distance of h_s, measures the average spatial variation of crime data points based on spatial distance, while temporal semi-variograms are calculated using Eq. 3:

$$\gamma_t(h_t)=\frac{1}{2N(h_t)}\sum_{i=1}^{N(h_t)}\left[Z(s_1,t_1)-Z(s_1,t_1+h_t)\right]^2 \quad (3)$$

where $N(h_t)$ is the number of data pairs of same crime type which are located at the same location while separated by h_t period, consider the time difference between data points at the same location. Based on equations above, spatial and temporal semi-variograms were derived for the primary variable for three types of criminal activities. Thus, for each type of crime—street crime, property crime, and vehicle crime—spatial and temporal semi-variograms were calculated separately for weekdays and weekends. A least square fitting method was employed to determine the best fitting models for both spatial and temporal semi-variograms, including Gaussian, exponential, spheric, or linear models. Because previous study has shown that crime incidents typically exhibit no consistent directional preference, as offenders tend to operate or move within localized areas without a predominant spatial orientation (Gilmour & Higham, 2022). Once the spatial and temporal semi-variograms were estimated and fitted, they were combined to the spatio-temporal structure to measure how the variance between data points changes as the distance and time lag between them increases. Covariance matrices were derived for both the primary and co-variable, considering spatial and temporal distances. This matrix aids in the understanding of how primary and co-variable relate over space and time, ensuring a consistent temporal dimension and positive definiteness in their relationship. Spatio-temporal covariances have the property that they can be written as a product or the sum of a valid spatial covariance and a valid temporal covariance. To ensure an optimal balance between efficiency and effectiveness, the valid spatial covariance model and valid temporal covariance model were combined in product form (Yang et al., 2020).

2.5 Accuracy evaluation

The effectiveness of the ST-Cokriging prediction was evaluated using the Pearson Correlation Coefficient (r), Root Mean Squared Error (RMSE), as well as PAI (Prediction Accuracy Index) and PEI (Prediction

Efficiency Index) which are commonly used evaluation metrics to assess the effectiveness of spatial crime forecasting models (Chainey et al., 2008). r (Pearson) measures the correlation between predicted and actual data, while RMSE measures the differences between predicted and actual data. PAI assesses how effectively the prediction captures crime hot-spots, considering the ratio of crime successfully predicted within a hotspot area to the total area of interest. PAI can be calculated using Eq. 4,

$$PAI = (\frac{n}{N})/(\frac{a}{A}) \tag{4}$$

where n represents the number of crimes accurately predicted within the identified hotspot area, N is the total number of crimes during the prediction period, a denotes the area of the crime hotspots, and A is the total area of the region under study. For example, if the model can accurately predict 80% of all crime activities in 40% of the overall area, the PAI would be 2. Hence, successfully forecasting a higher percentage of crime activities in smaller hotspots would yield higher PAI values (Chainey et al., 2008).

PEI, ranging from 0 to 1, compares the actual PAI to the maximum possible PAI, indicating how well the prediction captures hotspots relative to the best possible outcome. PEI can be calculated using Eq. 5,

$$PEI = \frac{PAI}{PAI_{max}} \tag{5}$$

where PAI_{max} denotes the maximum value of possible PAI (Chainey et al., 2008).

We evaluated the prediction accuracy of each calibrated ST-Cokriging model using Correlation, RMSE, PAI & PEI by comparing the predicted values with real crime data at a bi-weekly temporal scale. To better predict and validate the performance of the prediction, we create the validating scenario that crime risk for each bi-week was predicted using prior three bi-weeks crime primary data and tweeter secondary data, then the fourth week data were saved as reference data for validation. For example, to predict crime risk for the 10th bi-week, we used crime data from the 7th to 9th, and data from the 10th bi-week was reserved for subsequent validation. We assessed the models' performance using predictions for the 10th, 16th, and 22nd bi-weeks starting on May 5, July 28, and October 20 of 2014. These time periods were strategically selected to capture seasonal variations while ensuring a consistent temporal interval. Crime predictions for weekdays and weekends were modeled separately for each crime type.

3 Results

3.1 Spatio-temporal structure of crime types

KDE helps to visualize the intensity of events, making it particularly useful for identifying hotspots in tweets. We analyzed the filtered tweets for each crime type to map their spatial patterns using the kernel density function for estimating the probability density of spatial events across a surface (Okabe et al., 2009). Cell size of 100 m (Chainey, 2013) were selected for the density map with the same search radius of 2 km as KDE of crime data (Fig. 2) to balance the need for detailed resolution and computational efficiency; this scale is fine enough to capture local variations while maintaining a manageable data size for analysis. Both the primary variable (crime records) and the co-variable (filtered tweets) were processed using the kernel density function to ensure consistency in the spatial resolution of the variables used for further predictive modeling. Hotspots for street crime-related tweets are primarily located in the eastern part of San Jose downtown and the northern San Jose, while hotspots for tweets related to property crime and vehicle crime are concentrated in the eastern part of San Jose downtown (Fig. 2).

Using crime density as input, the spatial and temporal semi-variograms for both weekday and weekend groups regarding the three crime types were estimated separately and then combined to spatial–temporal variances. Figure 3 depicts spatial and temporal semi-variograms and 3-D plots of the spatio-temporal covariance models for each scenario. Notably, the semi-variogram patterns differ between the spatial and temporal domains. The spatial semi-variogram was fitted with a Gaussian function based on the likelihood, whereas the temporal semi-variogram was fitted with an exponential function determined by the lowest residual value of the OLS fitting method and the shape of the semi-variogram (Eqs. 6 and 7). This tailored model selection approach ensures an optimal fit for our data. The OLS-fitted functions for the spatial and temporal semi-variograms for three crime types are (Table 2):

$$\gamma_s(h_s) = n + s \bullet [1 - exp(-\frac{h_s^2}{p^2})] \tag{6}$$

$$\gamma_t(h_t) = n + s \bullet [1 - exp(-\frac{h_t}{p}) \tag{7}$$

where the n is the nugget (unexplained randomness), s is the sill (maximum intensity), $p(meter)$ is range (influence zone),$h_s(meter)$ is spatial distance, and the $h_t(day)$ is the temporal distance.

The spatio-temporal statistical structure demonstrated distinct patterns across varied crime types, highlighting importance for developing refined crime prediction

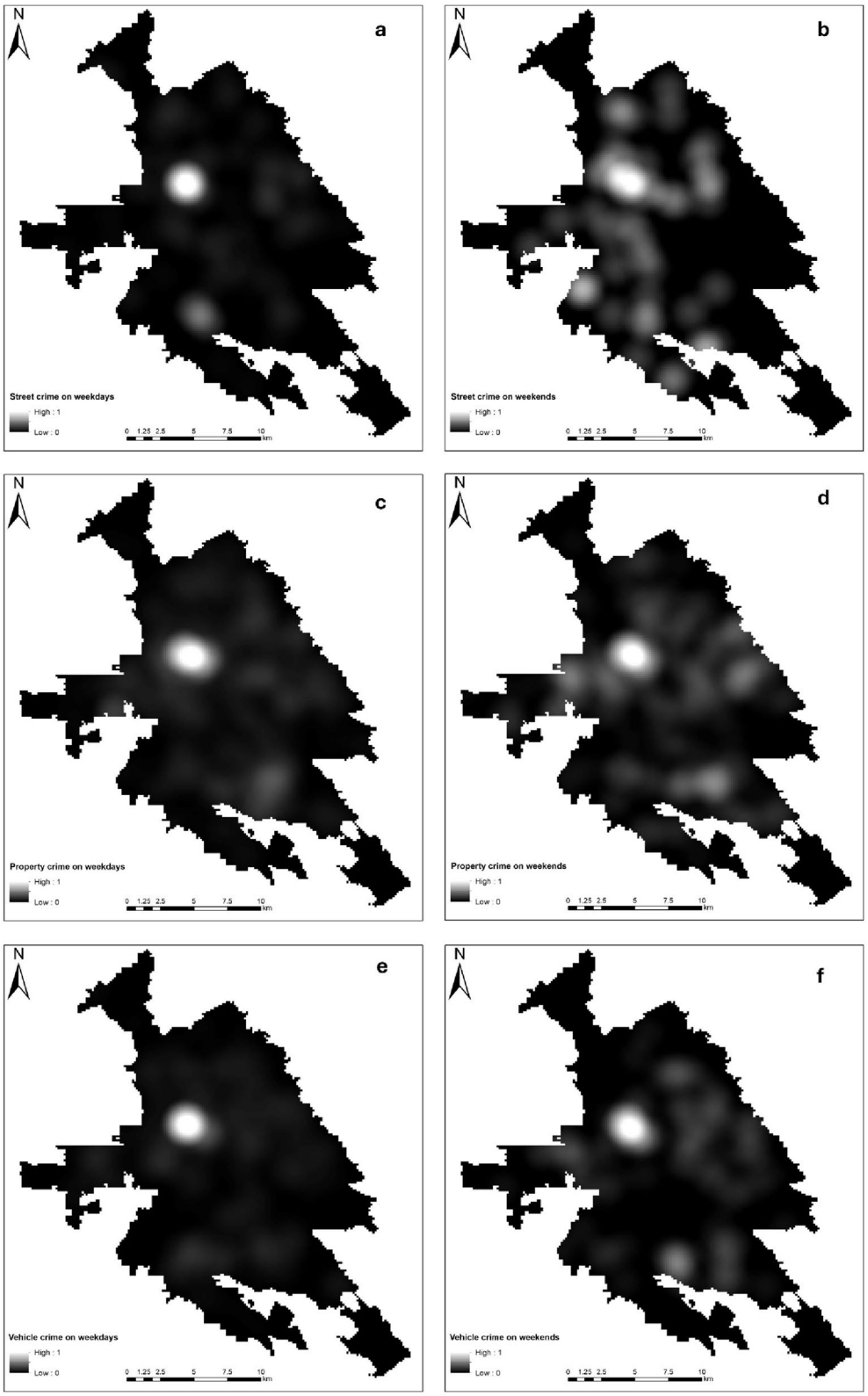

Fig. 2 Kernel density patterns of crime-related tweets of 2014 in San Jose, California, USA, **a** street crime on weekdays; **b** street crime on weekends; **c** property crime on property crime; **d** property crime on weekends; **e** vehicle crime on weekdays; **f** vehicle crime on weekends

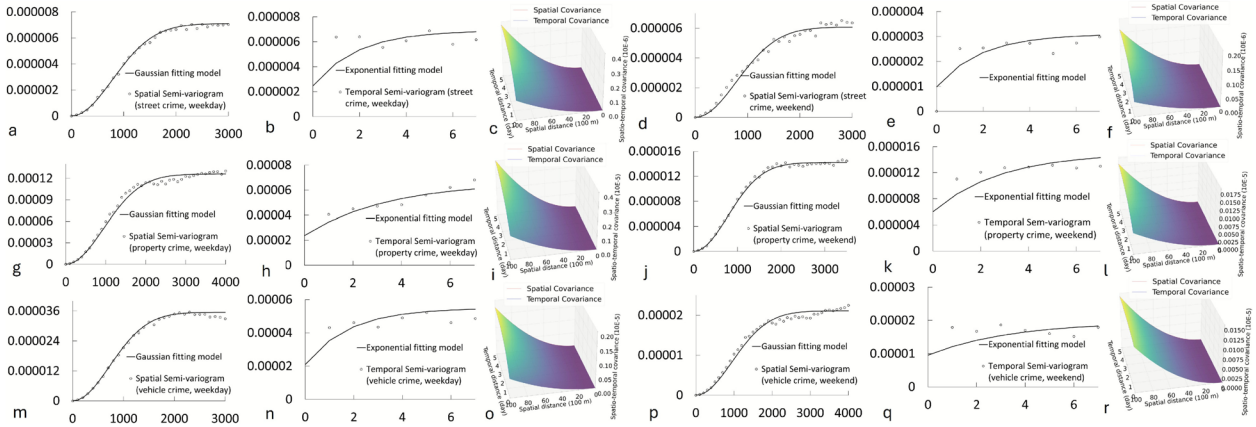

Fig. 3 Spatial and temporal semi–variograms and fitting models for weekly-based periods and 3D plots of spatio-temporal covariance model for ST-Cokriging: **a**-**f** street crime: **a** spatial semi–variogram for weekdays, **b** temporal semi–variogram for weekdays, **c** 3D plots of spatio-temporal covariance model for ST-Cokriging for weekdays, **d** spatial semi–variogram for weekends, **e** temporal semi–variogram for weekends, **f** 3D plots of spatio-temporal covariance model for ST-Cokriging for weekends; **g**-**l** property crime: **g** spatial semi–variogram for weekdays, **h** temporal semi–variogram for weekdays, **i** 3D plots of spatio-temporal covariance model for ST-Cokriging for weekdays, **j** spatial semi–variogram for weekends, **k** temporal semi–variogram for weekends, **l** 3D plots of spatio-temporal covariance model for ST-Cokriging for weekends; **m**-**r** vehicle crime: **m** spatial semi–variogram for weekdays, **n** temporal semi–variogram for weekdays, **o** 3D plots of spatio-temporal covariance model for ST-Cokriging for weekdays, **p** spatial semi–variogram for weekends, **q** temporal semi–variogram for weekends, **r** 3D plots of spatio-temporal covariance model for ST-Cokriging for weekends

Table 2 Parameters of fitting functions for the spatial and temporal semi-variograms for three crime types

		Spatial fitting			Temporal fitting		
Crime type	**Time period**	***n***	***s***	***p***	***n***	***s***	***p***
Street crime	weekday	0	7.10×10^{-6}	1110.0	2.46×10^{-6}	4.45×10^{-6}	1.9
	weekend	0	6.75×10^{-6}	1110.0	0.98×10^{-6}	2.12×10^{-6}	1.9
Property crime	weekday	0	12.60×10^{-5}	1332.0	2.37×10^{-5}	4.27×10^{-5}	3.4
	weekend	0	1.42×10^{-5}	1100.0	0.59×10^{-5}	0.90×10^{-5}	2.8
Vehicle crime	weekday	0	3.55×10^{-5}	1032.0	2.09×10^{-5}	3.39×10^{-5}	1.8
	weekend	0	2.11×10^{-5}	1338.0	1.01×10^{-5}	0.25×10^{-5}	3.4

models. Firstly, our analysis reveals a significant difference between spatial and temporal variograms. Spatial semi-variograms consistently exhibited an absence of the nugget effect (with the nugget value consistently equating to zero), suggesting minimal measurement error within the spatial domain. This can be primarily attributed to the high resolution of the spatial measurements. Conversely, temporal variograms displayed notable nugget effects, ranging from 0.98×10^{-6} to 2.37×10^{-5}, as delineated in Fig. 3. This suggests a greater degree of randomness within the temporal domain, which can be attributed to the choice of using day-level temporal intervals and not further dividing them into hours and minutes.

In addition, the analysis distinguishes the different types of data in terms of their range (p) and sill (s) effects. Notably, house crime categorized under property crime exhibited higher spatial and temporal ranges compared to vehicle crime and street crime (such as assault and robbery) on weekdays. This implies property crime exerts a more extended temporal influence are likely due to the increased preparation time associated with these crime on weekdays. Furthermore, it is noteworthy that the spatio-temporal structures vary between weekdays and weekends. This variance is particularly pronounced in the case of property crime, with significant differences observed between these temporal groupings—12.60×10^{-5} and 1.42×10^{-5} for spatial sills, and 2.37×10^{-5} and 0.59×10^{-5} for temporal sills, respectively. This variation can be correlated to the differing profiles of criminal types involved in property crime during weekdays and weekends, potentially influenced by the routine presence or absence of inhabitants due to workday schedules.

For street crime, the spatial range on weekdays (1110) is significantly lower than that of property crime (1332), implying that the impact zone of street crime is more localized and possibly linked to specific hotspots prone to such incidents. Temporally, the range is identical on both weekdays and weekends, suggesting that the temporal impact of this crime is stable regardless of the day of a week. However, the temporal sill shows a slight increase on weekdays (4.45×10^{-6} vs 2.12×10^{-6}), which implies a higher variability in the timing of this crime during weekdays, possibly due to changes in social activity patterns. The spatial sill values for street crime are lower in comparison to other two crime types, which could be attributed to the more impulsive nature of such crime, leading to a smaller spread of influence over spatial scale.

Vehicle crime demonstrates the smallest spatial range on weekdays (1032) and the largest on weekends (1338) among three crime types, indicative of the mobility inherent to this crime. The reduced spatial range on weekdays could be a reflection of routine commuting paths and concentrated parking areas, while the expanded range on weekends might point to a wider dispersion of vehicles as people travel to varied destinations or leave cars in less secure locations. The temporal sill, however, is higher on weekdays than weekends (3.39×10^{-5} vs 0.25×10^{-5}), which could be attributed to the higher volume of vehicles in use and therefore a greater opportunity for this crime. Interestingly, vehicle crime exhibits higher spatial sill values on weekdays (3.55×10^{-5}), likely reflecting the routine of individuals commuting and using their vehicles more during the week, thus increasing the opportunity for such crime. Conversely, the weekend sees a significant drop in the sill (2.11×10^{-5}), perhaps due to the decreased routine activity, with vehicles less frequently left in vulnerable public spaces.

3.2 Crime prediction and validation

Based on the estimated spatio-temporal structure, the ST-Cokriging model predict the crime risk (crime density with 100 m grid) by incorporating both the historical crime risks as primary variable and filtered Twitter data as the co-variable. Figure 4 visually contrasts the predicted weekday crime risk during the 22nd bi-week for three crime types. Predicted crime risk maps through ST-Cokriging are illustrated in Fig. 4b, d, and f, and the actual referencing crime risk maps for these weeks is illustrated in Fig. a, c, and e. Figure 5 showcases the predicted (Fig. 5b, d, and f) and actual referencing (Fig. 5a, c, and e) weekend crime risk distribution during the 22nd bi-weeks for three crime types. For a fair comparison, all images use a consistent color scale. A close similarity is observed between the predicted and referencing crime risk for all three crime types.

We carried out further validation against reference by calculating the correlation r and RMSE for bi-weeks 10, 16, and 22, as shown in Tables 3 and 4. For all types of crime (street crime, property crime, and vehicle crime), the prediction model demonstrates higher accuracy when co-variables are incorporated. This is evident on both weekdays and weekends, and across all the bi-weeks observed (bi-weeks 10, 16, and 22). For instance, for street crime during weekdays of bi-weeks 22, the prediction with the co-variable has a correlation coefficient of 0.5219 and an RMSE of 0.0377, whereas the prediction without the co-variable has a correlation coefficient of 0.4230 and an RMSE of 0.1000. The property crime category during weekdays of the same bi-weeks, when modeled with tweets as the co-variable, produces a correlation coefficient of 0.8803 and an RMSE of 0.0393. Without the co-variable, the correlation coefficient drops to 0.8524 and the RMSE increases to 0.0977. The correlation coefficient for vehicle crime prediction during weekdays is 0.8939 when including the co-variable and 0.8905 when excluding the co-variable. The RMSE is 0.0323 with the co-variable and 0.0993 without the co-variable. For street crime on the weekend of bi-week 22, the correlation is 0.3661 (with co-variable) and 0.2859 (without co-variable), while the RMSEs are 0.0443 and 0.0970, respectively. For the property crime category during weekends of this bi-weeks, the prediction with the co-variable results in a correlation coefficient of 0.7336 and an RMSE of 0.0421. In contrast, without the co-variable, the correlation coefficient is 0.7373 with an RMSE of 0.1089. In the weekend of bi-weeks 22, the vehicle crime model with the co-variable produces a correlation of 0.7874 and an RMSE of 0.0339. Without the co-variable, the correlation coefficient is 0.7802 with an RMSE of 0.0994.

To conclude, integrating crime-related tweets as a co-variable significantly improved predictive accuracy for all crime types. Specifically, during weekdays, the

(See figure on next page.)

Fig. 4 Biweekly crime prediction results (weekdays) from ST-Cokriging and validation against actual reference crime risk map: **a** actual crime risk map during weekdays of bi-weeks 22 for street crime; **b** ST-Cokriging predicted crime risk map during weekdays of bi-weeks 22 for street crime; **c** actual crime risk map during weekdays of bi-weeks 22 for property crime; **d** ST-Cokriging predicted crime risk map during weekdays of bi-weeks 22 for property crime; **e** actual crime risk map during weekdays of bi-weeks 22 for vehicle crime; **f** ST-Cokriging predicted crime risk map during weekdays of bi-weeks 22 for vehicle crime

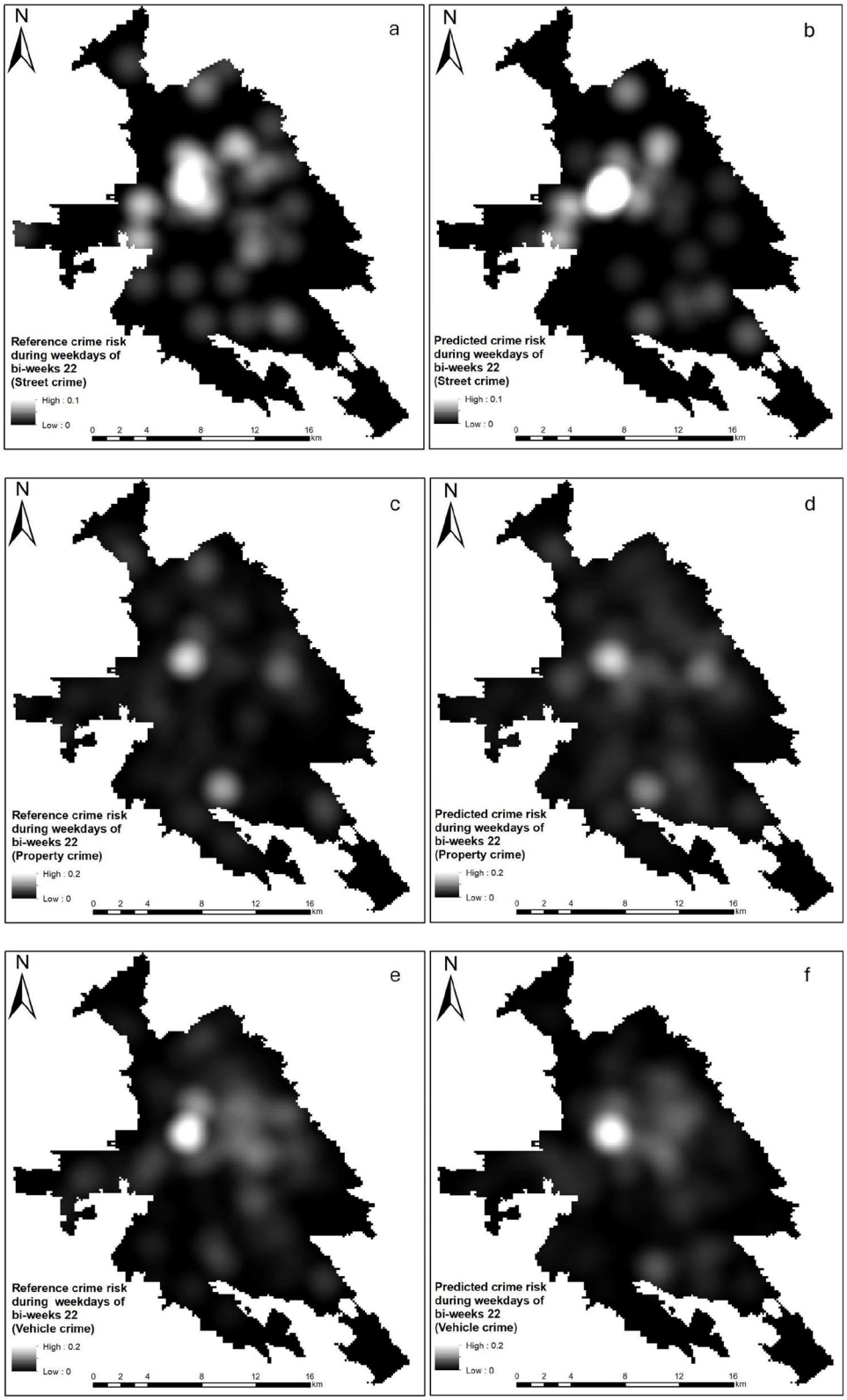

Fig. 4 (See legend on previous page.)

average increase of correlation coefficient was 15.8% for street crime, 1.9% for property crime, and 0.4% for vehicle crime. Meanwhile, on weekends, the correlation increased by 13.9%, 12.4%, and 6.3% for street crime, property crime, and vehicle crime, respectively. The RMSE significantly decreased when incorporating crime-related tweets as a co-variable compared to the model without co-variables.

3.3 Crime hotspots analysis

PAIs and PEIs were computed based on the crime risk map to generate spatial distributions of crime hotspots for optimized law enforcement resource allocation. For each threshold, a map can be generated for crime risk hotspots and a highlighted thresholding area were generated as the predicted police patrolling area. Then hit points were calculated as the detected criminal activities within these highlighted areas. These hit points were subsequently utilized to derive the PAI metrics at the given threshold. By comparing the PAIs at different thresholds, the optimal threshold was chosen and the corresponding PAI and PEI can be calculated.

PAI curves for ST-Cokriging predictions with and without co-variable were compared during weekdays and weekends of bi-weeks 22 for street crime, property crime, and vehicle crime (Fig. 6). PAI represents the percentage of all actual crime events during each bi-weekly period that occurred within predicted hotspots, typically increasing as the threshold value rises. The inclusion of crime-related tweets significantly improved crime prediction performance, as evidenced by the substantial increase in PAI values across both weekdays and weekends of the bi-weeks. Additionally, Table 5 indicates that incorporating crime-related tweets as a co-variable in predictions results in a higher PAImax compared to the control group where this co-variable is excluded.

Larger hotspots can capture more criminal activities but pose challenges for efficient police deployment. It is essential to identify hotspots of an optimal size to enable effective resource allocation. The inflection point in the PAI curves serves as a guideline for selecting the ideal PAI/PEI, ensuring optimal hotspot identification (Chainey et al., 2008) (Table 6). For example, for weekdays of bi-weeks 22 (vehicle crime), an optimal threshold value of 0.0616 was selected corresponding to a PAI inflection point of 1.85 (Fig. 6e). Based on this approach, we determined the best thresholds during weekdays and weekends in bi-weeks 22 for street crime, property crime, and vehicle crime with their associated PEI and PAI values detailed in Table 6. The highest recorded PAI was 1.85 during weekdays in bi-weeks 22 for vehicle crime. Figure 7 depicts hotspot maps during weekdays in bi-weeks 22 for street crime, property crime, and vehicle crime using optimal PAI thresholds of 0.0112, 0.0392, and 0.0616, respectively. For street crime, two small hotspots are shown in red, with successful predictions (hit points) marked in red and non-hit points in green. The hotspots accurately predicted 8 out of 47 street crime incidents, resulting in a hit rate of 17.02% and covering 3.49% of the total area. Property crime hotspots are represented by two large and two small areas in red. The model successfully predicted 88 out of 304 incidents, with a hit rate of 28.94%, and the hotspots spanned 6.80% of the study area. Vehicle crime analysis identified two large and one small hotspot, with 141 out of 378 crime incidents successfully predicted, yielding a hit rate of 37.30% and hit area of 13.68%.

4 Discussion and conclusions

In this study, we used the spatio-temporal Cokriging crime prediction method to incorporate historical crime data and voluntary and geotagged social media posts to predict different crime in three major categories: street crime, property crime, and vehicle crime, within the larger metropolitan area of San Jose, located in the SFBA. The police historical crime calls were utilized as the primary variable for the prediction, and we utilized the ST-Cokriging algorithm to incorporate spatio-temporal structure of social media Twitter data as the co-variable for predictions of different crime categories. The results indicate that including geotagged crime-related tweets as the co-variable alongside historical crime data significantly improved crime prediction accuracy in San Jose for case study year, demonstrating the value of combining social media content with traditional crime data for enhanced forecasting. This study provides new findings and methods that explore the integration of multi-source digital data to further refine the predictive capabilities of crime forecasting models. By leveraging the voluntary, geotagged, and dynamic

(See figure on next page.)

Fig. 5 Biweekly crime prediction results (weekends) from ST-Cokriging and validation against actual reference crime risk map: **a** actual crime risk map during weekends of bi-weeks 22 for street crime; **b** ST-Cokriging predicted crime risk map during weekends of bi-weeks 22 for street crime; **c** actual crime risk map during weekends of bi-weeks 22 for property crime; **d** ST-Cokriging predicted crime risk map during weekends of bi-weeks 22 for property crime; **e** actual crime risk map during weekends of bi-weeks 22 for vehicle crime; **f** ST-Cokriging predicted crime risk map during weekends of bi-weeks 22 for vehicle crime

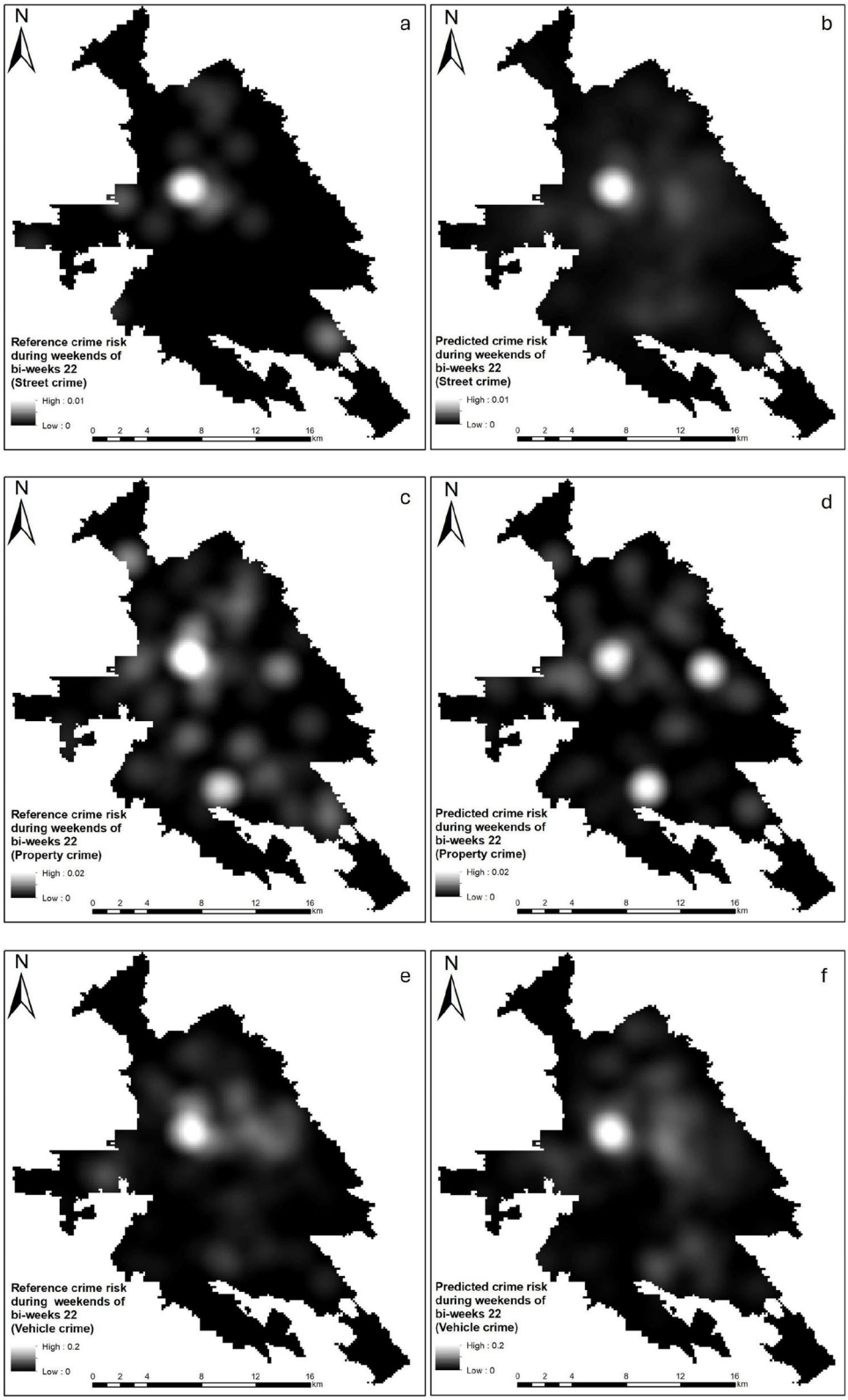

Fig. 5 (See legend on previous page.)

Table 3 Statistical tests comparing ST-Cokriging predictions against reference data for three crime types during weekdays in bi-weeks

	Street crime				Property crime				Vehicle crime			
	w/co-variable		w/o co-variable		w/co-variable		w/o co-variable		w/co-variable		w/o co-variable	
	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE
Weekdays in bi-weeks 10	0.5012	0.0377	0.3381	0.1003	0.8240	0.0358	0.8088	0.1007	0.8616	0.0372	0.8613	0.1000
Weekdays in bi-weeks 16	0.6038	0.0459	0.3916	0.1001	0.8321	0.0351	0.8184	0.1007	0.8976	0.0312	0.8886	0.1016
Weekdays in bi-weeks 22	0.5219	0.0378	0.4230	0.1000	0.8802	0.0393	0.8524	0.0977	0.8939	0.0323	0.8905	0.0993

Table 4 Statistical tests comparing ST-Cokriging predictions against reference data for three crime types during weekends in bi-weeks

	Street crime				Property crime				Vehicle crime			
	w/co-variable		w/o co-variable		w/co-variable		w/o co-variable		w/co-variable		w/o co-variable	
	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE	r	RMSE
Weekends in bi-weeks 10	0.4151	0.0448	0.2482	0.0993	0.7219	0.0346	0.5910	0.0986	0.7599	0.0382	0.7252	0.1008
Weekends in bi-weeks 16	0.4479	0.0440	0.2772	0.1000	0.7621	0.0411	0.5174	0.0988	0.7966	0.0378	0.6472	0.1007
Weekends in bi-weeks 22	0.3661	0.0443	0.2859	0.0970	0.7336	0.0421	0.7373	0.1089	0.7847	0.0339	0.7802	0.0994

nature of tweets and identifying different keywords to model social responses to the three categories of crime, our model offers a more comprehensive understanding of crime patterns, outperforming traditional methodologies. It demonstrates that social media information can effectively aid in modeling crime patterns in both spatial and temporal contexts. This methodology holds great potential for broader applications, such as optimizing police patrol routes, identifying high-risk areas for targeted interventions, and supporting urban planning efforts to design safer communities.

This study extends the analysis of a previous case study by investigating three major crime categories across a larger metropolitan area with diverse crime activities. This additional modeling of spatial and temporal dependency and autocorrelation has revealed distinct spatial–temporal patterns in the semivariograms for the three crime types. The results suggest that property crime has a longer temporal impact, likely due to the increased preparation time required during the week. Vehicle crime initially displayed the highest prediction accuracy using historical crime data alone, suggesting a strong reliability in traditional methods for this crime category. However, the incremental improvement in accuracy with the inclusion of crime-related tweets for this category was minimum. Despite the fact that social media data offers some enhancement, its impact is relatively limited when considered in the context of vehicle crime, which is already well-predicted. In contrast, street crime had the lowest prediction accuracy when based solely on historical crime data; however, the integration of social media data yielded a substantial increase in predictive accuracy. The significant improvement features the value of real-time and volunteered geographic information in understanding and forecasting more complex and less predictable crime types. The predictive accuracy of property crime was moderate both with and without tweet data. This suggests a certain level of predictability inherent to these crimes, which is not substantially enhanced nor diminished by the addition of social media data.

An additional innovative aspect of this study is the use of a dual approach for filtering geotagged crime-related tweets, combining keyword-based methods with manual review to enhance the accuracy of crime predictions. The keyword filtering process involved selecting tweets based on predefined keywords for various crime categories, such as "assault," "robbery," and "theft." However, relying solely on keyword filtering can introduce noise into the dataset due to the broad nature of some terms. To address this, a manual review step was implemented to refine the dataset, ensuring that only the most relevant tweets were included in the analysis. We employed a supervised method to further filter out tweets containing crime-related keywords which were related to crime events. The filtered dataset could serve as training data for an AI model, enabling future integration of the Large Language Model (LLM). This would allow for real-time fine-tuning of crime-related keywords and automatically filtering social media posts for the next step in prediction. Such an approach would enable crime prediction to be more automated and in real-time, utilizing separate models for street crime, vehicle crime, and property crime, each designed to account for distinct spatial and

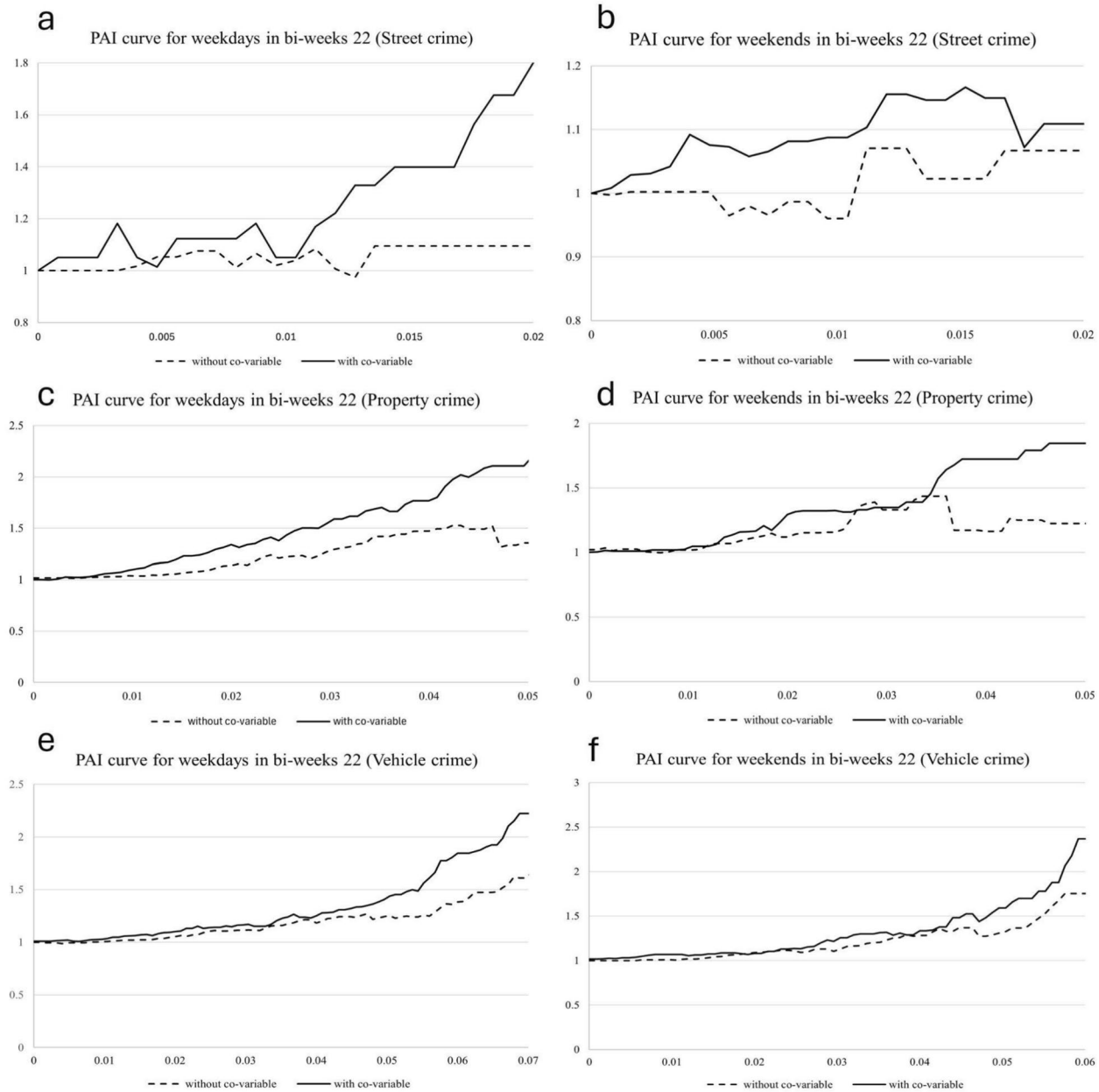

Fig. 6 PAI curves for ST-Cokriging predictions of crime risks in bi-weeks 22 (predictions with and without co-variable, X-axis: distance; Y-axis: PAI): **a** weekdays for street crime; **b** weekends for street crime; **c** weekdays for property crime; **d** weekends for property crime; **e** weekdays for vehicle crime; **f** weekends for vehicle crime

Table 5 PAI_{max} for ST-Cokriging predictions of crime risks with/without co-variable in bi-weeks 22 for three crime types

	PAI_{max} (weekdays of bi-weeks 22)		**PAI_{max} (weekend of bi-weeks 22)**	
	w/co-variable	**w/o co-variable**	**w/co-variable**	**w/o co-variable**
Street crime	1.80	1.09	1.16	1.07
Property crime	2.19	1.52	1.85	1.38
Vehicle crime	2.27	1.65	2.47	1.78

Table 6 The chosen threshold and corresponding PEI of bi-week 22 for three crime types

	Optimal threshold (bi-weekday)			Optimal threshold (bi-weekend)		
	Threshold	PEI	PAI	Threshold	PEI	PAI
Street crime	0.0112	0.73	1.32	0.0160	0.91	1.06
Property crime	0.0392	0.80	1.76	0.0360	0.88	1.63
Vehicle crime	0.0616	0.81	1.85	0.0544	0.72	1.78

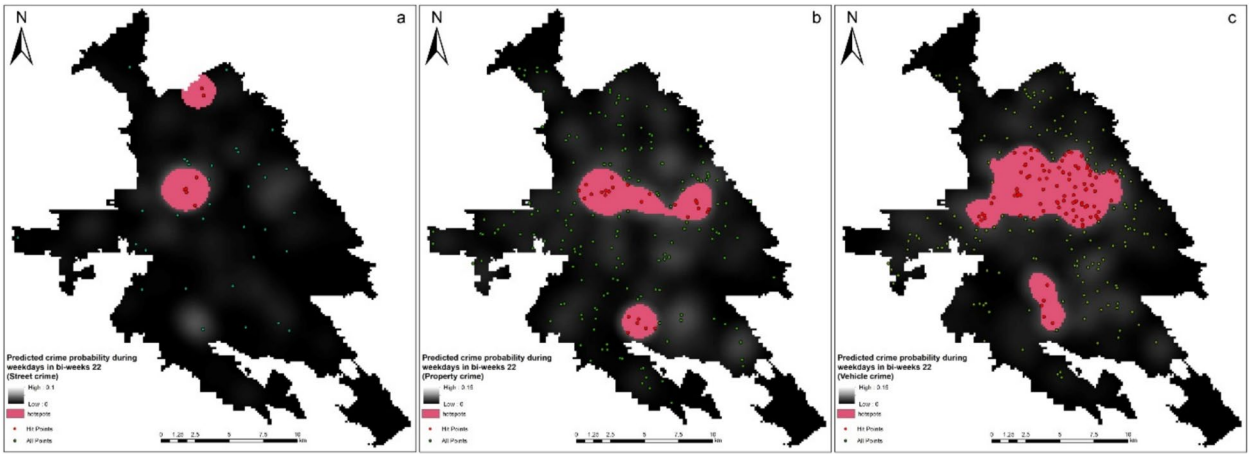

Fig. 7 Predicted crime hotspots and actual location of crime incidents during weekdays in bi-weeks 22, 2014. **a** street crime, **b** property crime, **c** vehicle crime

temporal dependency and autocorrelations. This study represents a pioneering method that enhances the differentiation and modeling of various crime types.

Police phone call records were utilized as a proxy for crime data as the primary variable in the study. While these records do not represent actual crime data, they are closely aligned with crime incidents as many calls involve emergency situations or reports of ongoing criminal activities. The police call data captures real-time or near-real-time responses to incidents that often lead to formal crime reports. This makes them a valuable source of information, offering immediate insights into criminal behaviors. In addition, our findings revealed that the predictive accuracy for the three crime types was lower on weekends compared to weekdays. This discrepancy may be due to the increased variability in social behaviors during weekends, such as larger gatherings, nightlife activities, and alcohol consumption, which can lead to more spontaneous or unpredictable crime patterns. Weekdays follow routine activities and schedules that create consistent patterns, while weekends are more unpredictable, making crime prediction models less effective. Weekends consist of only two days, so there are fewer crime incidents recorded than on weekdays when there is a greater amount of data available.

This study has several limitations. First, using Twitter data may introduce multiple data quality issues. Sampling bias arises because the platform captures only a subset of the population, often overrepresenting certain demographic or interest groups while underrepresenting others. In addition, geolocation bias further limits representativeness, as only about 1% of tweets are geotagged. This lack of spatial information can distort spatial analyses, given that users who share their location may differ systematically in age, socioeconomic status, or online behavior from those who do not. Moreover, relying solely on Twitter may overlook complementary insights available from other social media platforms. Future research could enhance prediction accuracy by integrating multi-platform social media data. In addition to our bi-weekly prediction scheme (using the prior three bi-weekly periods to predict the fourth), we conducted a sensitivity analysis with weekly and monthly windows—specifically, three prior weeks to predict the fourth week, and three prior months to predict the fourth month—using the same model specification and hyperparameters. Across

these alternatives, we observed similar predictive accuracy, indicating that the model's performance is robust to the temporal window choice and that the way we capture temporal autocorrelation is not materially affected by shifting the window length.

In conclusion, this study employed crime prediction modelling by integrating geotagged crime-related tweets as a co-variable alongside historical crime data within ST-Cokriging framework. The method was applied to crime risk predictions of three crime categories during both weekdays and weekends in a major metropolitan area of concern (San Francisco Bay Area). Our findings offer valuable insights into the distinct dynamics of criminal activities over time and space across different crime types. The integration of social media data as a co-variable significantly improved the accuracy of crime predictions, outperforming models based solely on historical crime data. This demonstrates the value of crowd-sourced geotagged information in refining predictive models and strengthening crime alert mechanisms. The findings highlight the potential of social media responses to improve real-time crime detection and inform evidence-based police enforcement. The strong predictive performance underscores the potential of this integrated approach, and emphasizes the need to integrate additional machine learning, large language models (LLMs), and AI techniques to more accurately capture the complex and multifaceted nature of criminal behavior. This study contributes to the advancement of crime prediction by revealing distinct spatial and temporal dynamics among street, property, and vehicle crimes, thereby providing actionable insights for law enforcement and urban planners to design targeted prevention strategies—such as allocating patrol resources based on crime type, improving environmental design in recurrent hotspots, and strengthening community-based interventions to enhance public safety.

Acknowledgements
This work was supported by the University of New Mexico Office of the Vice President for Research under WeR1 Faculty Success Program, Research Allocations Committee (RAC) awards (#8oh6a4x35h, #gvvrxwyj64), and FRESSH Pilot Program; and the University of New Mexico, A&S Interdisciplinary Science Cooperative through the Office of Research (Faculty Team Research Concept Competition Award #TA-1003).

Authors' contributions
Yanhong Huang: Conceptualization, Software, Formal analysis, Writing - Original Draft, Writing – review and editing. Bo Yang: Conceptualization, Software, Methodology, Validation, Writing – review and editing, Supervision. Xiangyu Ren: Validation, Writing - Review & Editing. Yujian Lu: Data curation. Minxuan Lan: Writing - Review & Editing. Xi Gong: Conceptualization, Validation, Data Curation, Writing – review and editing, Supervision, Funding acquisition, Project administration.

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: 30 July 2025 Revised: 3 November 2025 Accepted: 25 December 2025
Published online: 31 December 2025

References

Ahn, H. Il, & Spangler, W. S. (2014). Sales prediction with social media analysis. *Annual SRII Global Conference, SRII, Annual SRII Global Conference (SRII)*, 213–222. https://doi.org/10.1109/SRII.2014.37

Alves, L. G. A., Ribeiro, H. V., & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica a: Statistical Mechanics and Its Applications, 505*, 435–443. https://doi.org/10.1016/j.physa.2018.03.084

Amerio, P., & Roccato, M. (2005). A predictive model for psychological reactions to crime in Italy: An analysis of fear of crime and concern about crime as a social problem. *Journal of Community & Applied Social Psychology, 15*(1), 17–28. https://doi.org/10.1002/casp.806

Bendler, J., Brandt, T., Wagner, S., & Neumann, D. (2014). Investigating crime-to-twitter relationships in urban environments - Facilitating a virtual neighborhood watch. *ECIS 2014 Proceedings - 22nd European Conference on Information Systems, July 2017*. https://www.wi.uni-muenster.de/research/publications/169230

Berry-James, R. J. M., Gooden, S. T., & Johnson, R. G. (2020). Civil rights, social equity, and census 2020. *Public Administration Review, 80*(6), 1100–1108. https://doi.org/10.1111/puar.13285

Braga, A. A., Papachristos, A. V., & Hureau, D. M. (2014). The effects of hot spots policing on crime: An updated systematic review and meta-analysis. *Justice Quarterly, 31*(4), 633–663. https://doi.org/10.1080/07418825.2012.673632

Butt, U. M., Letchmunan, S., Ali, M., & Sherazi, H. H. R. (2025). START: A Spatiotemporal Autoregressive Transformer for Enhancing Crime Prediction Accuracy. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2025.3550196

Chainey, S. (2013). Examining the influence of cell size and bandwidth size on kernel density estimation crime hotspot maps for predicting spatial patterns of crime. *Bsglg, 60*(1), 7–19. https://popups.uliege.be/0770-7576/index.php?id=422&utm_source=chatgpt.com

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal, 21*(1–2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

Corso, A. J., Alsudais, A., & Hilton, B. (2016). Big social data and GIS: Visualize predictive crime. *AMCIS 2016: Surfing the IT Innovation Wave - 22nd Americas Conference on Information Systems*, 1–10. https://aisel.aisnet.org/ecis2016_rp/157/

Cressie, N., & Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association, 94*(448), 1330–1339. https://doi.org/10.1080/01621459.1999.10473885

Da Silva, S., Boivin, R., & Fortin, F. (2019). Social media as a predictor of urban crime. *Criminologie, 52*(2), 83–109. https://doi.org/10.7202/1065857ar

DeVeaux, R. D., Bowman, A. W., & Azzalini, A. (1999). Applied Smoothing Techniques for Data Analysis. In *Technometrics* (Vol. 41, Issue 3). https://doi.org/10.2307/1270572

Du, Y., & Ding, N. (2023). A Systematic Review of Multi-Scale Spatio-Temporal Crime Prediction Methods. In *ISPRS International Journal of Geo-Information* (Vol. 12, Issue 6). https://doi.org/10.3390/ijgi12060209

Featherstone, C. (2013). The relevance of social media as it applies in South Africa to crime prediction. In *2013 IST-Africa Conference and Exhibition, IST-Africa 2013* (Issue IST-Africa Conference and Exhibition). https://ieeexplore.ieee.org/abstract/document/6701724

Ferreira, J., Joao, P., & Martins, J. (2012). *GIS for Crime Analysis:Geography for Predictive Models*. https://www.routledge.com/Spatial-Analysis-and-GIS/Fotheringham-Rogerson/p/book/9780849339337

Gayo-Avello, D. (2013). A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Social Science Computer Review, 31*(6), 649–679. https://doi.org/10.1177/0894439313493979

Geoapify. (2024). *Geoapify*. https://www.geoapify.com/

Gilmour, C., & Higham, D. J. (2022). Modelling burglary in Chicago using a self-exciting point process with isotropic triggering. *European Journal of Applied Mathematics, 33*(2), 369–391. https://doi.org/10.1017/S0956792521000048

Goovaerts, P. (1997). Geostatistics for Natural Resources Evaluation. https://global.oup.com/academic/content/series/a/applied-geostatistics-age/?lang=en&cc=us

Hu, Y., Wang, F., Guin, C., & Zhu, H. (2018). A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation. *Applied Geography, 99*, 89–97. https://doi.org/10.1016/j.apgeog.2018.08.001

Journel, A. G., & Huijbregts, C. J. (1978). Mining Geostatistics. https://search.worldcat.org/title/800035147

Kadar, C., & Pletikosa, I. (2018). Mining large-scale human mobility data for long-term crime prediction. *EPJ Data Science, 7*(1). https://doi.org/10.1140/epjds/s13688-018-0150-z

Kyriakidis, P. C., & Journel, A. G. (1999). Geostatistical space-time models: A review. *Mathematical Geology, 31*(6), 651–684. https://doi.org/10.1023/A:1007528426688

Lal, S., Tiwari, L., Ranjan, R., Verma, A., Sardana, N., & Mourya, R. (2020). Analysis and Classification of Crime Tweets. *Procedia Computer Science, 167*(2019), 1911–1919. https://doi.org/10.1016/j.procs.2020.03.211

Lan, M., Liu, L., Hernandez, A., Liu, W., Zhou, H., & Wang, Z. (2019). The spillover effect of geotagged tweets as a measure of ambient population for theft crime. *Sustainability (Switzerland), 11*(23), 1–17. https://doi.org/10.3390/su11236748

Liu, L., Lan, M., Eck, J. E., Yang, B., & Zhou, H. (2022). Assessing the intraday variation of the spillover effect of tweets-derived ambient population on crime. *Social Science Computer Review, 40*(2), 512–533. https://doi.org/10.1177/0894439320983825

Mohamad Zamri, N. F., Md Tahir, N., Megat Ali, M. S. A., Khirul Ashar, N. D., & Al-misreb, A. A. (2021). Mini-review of street crime prediction and classification methods. *Jurnal Kejuruteraan, 33*(3), 391–401. https://doi.org/10.17576/jkukm-2021-33(3)-02

Newton, A. D., Hirschfield, A., Armitage, R., Rogerson, M., Monchuk, L., & Wilcox, A. (2008). *Evaluation of Licensing Act: Measuring Crime and Disorder in and around Licensed Premises, Research Study SRG/05/007 Annex 2: Birmingham, prepared for the Home Office. July 2007*. https://eprints.hud.ac.uk/id/eprint/9546/1/Licensing_Final_Report_March_2008_Supplementary_Annex.pdf?utm_source=chatgpt.com

Okabe, A., Satoh, T., & Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science, 23*(1), 7–32. https://doi.org/10.1080/13658810802475491

Piña-García, C. A., & Ramírez-Ramírez, L. (2019). Exploring crime patterns in Mexico City. *Journal of Big Data, 6*(1), 65. https://doi.org/10.1186/s40537-019-0228-x

Rousidis, D., Koukaras, P., & Tjortjis, C. (2020). Social media prediction: A literature review. *Multimedia Tools and Applications, 79*(9–10), 6279–6311. https://doi.org/10.1007/s11042-019-08291-9

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research, 23*(5), 528–543. https://doi.org/10.1108/IntR-06-2013-0115

Shi, T., Fu, J., & Hu, X. (2023). TSE-tran: Prediction method of telecommunication-network fraud crime based on time series representation and transformer. *Journal of Safety Science and Resilience, 4*(4), 340–347. https://doi.org/10.1016/j.jnlssr.2023.07.001

SJPD. (2023a). *Crime Statistics - Annual*. https://www.sjpd.org/records/crime-stats-maps/crime-statistics-annual

SJPD. (2023b). *San Jose Police Department*. https://www.sjpd.org/records/documents-policies

Snepvangers, J. J. J. C., Heuvelink, G. B. M., & Huisman, J. A. (2003). Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma, 112*(3–4), 253–271. https://doi.org/10.1016/S0016-7061(02)00310-5

Tang, J., Xia, L., & Huang, C. (2023). Explainable Spatio-Temporal Graph Neural Networks. *International Conference on Information and Knowledge Management, Proceedings*, 2432–2441. https://doi.org/10.1145/3583780.3614871

Tasnim, N., Imam, I. T., & Hashem, M. M. A. (2022). A novel multi-module approach to predict crime based on multivariate spatio-temporal data using attention and sequential fusion model. *IEEE Access, 10*, 48009–48030. https://doi.org/10.1109/ACCESS.2022.3171843

U.S. Department of Justice—Federal Bureau of Investigation. (2023). *Crime data explorer. NIBRS Estimates*. UCR Publications. https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/explorer/crime/nibrs-estimates

Uittenbogaard, A., & Ceccato, V. (2012). Space-time clusters of crime in Stockholm, Sweden. *Review of European Studies, 4*(5), 148–156. https://doi.org/10.5539/res.v4n5p148

Vomfell, L., Härdle, W. K., & Lessmann, S. (2018). Improving crime count forecasts using Twitter and taxi data. *Decision Support Systems, 113*, 73–85. https://doi.org/10.1016/j.dss.2018.07.003

Wang, K., Wang, P., Chen, X., Huang, Q., Mao, Z., & Zhang, Y. (2020). A Feature Generalization Framework for Social Media Popularity Prediction. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, *28th ACM International Conference on Multimedia (MM)*, 4570–4574. https://doi.org/10.1145/3394171.3416294

Wang, Z., Liu, L., Zhou, H., & Lan, M. (2019). Crime geographical displacement: Testing its potential contribution to crime prediction. *ISPRS International Journal of Geo-Information*, *8*(9). https://doi.org/10.3390/ijgi8090383

Yang, B., Liu, L., Lan, M., Wang, Z., Zhou, H., & Yu, H. (2020). A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery. *International Journal of Geographical Information Science, 34*(9), 1740–1764. https://doi.org/10.1080/13658816.2020.1737701

Yang, D., Heaney, T., Tonon, A., Wang, L., & Cudré-Mauroux, P. (2018). CrimeTelescope: Crime hotspot prediction based on urban and social media data fusion. *World Wide Web, 21*(5), 1323–1347. https://doi.org/10.1007/s11280-017-0515-4

Yu, H., Liu, L., Yang, B., & Lan, M. (2020). Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method. *ISPRS International Journal of Geo-Information*, 9–11. https://www.mdpi.com/2220-9964/9/12/732

Yuan, Y., McNeeley, S., & Melde, C. (2024). Understanding the fear of crime and perceived risk across immigrant generations: Does the quality of social ties matter? *Crime and Delinquency, 70*(3), 812–843. https://doi.org/10.1177/00111287221113306

Yuan, Y., Sanchez, C. V., & Punla, C. (2022). Procedural justice, neighborhood context, and domestic violence reporting intention among subgroups of immigrants. *Policing and Society, 32*(10), 1180–1192. https://doi.org/10.1080/10439463.2022.2029437

Zandiatashbar, A., & Kayanan, C. M. (2020). Negative consequences of innovation-igniting urban developments: Empirical evidence from three US cities. *Urban Planning, 5*(3), 378–391. https://doi.org/10.17645/up.v5i3.3067

Zhang, P., Wang, X., & Li, B. (2014). Evaluating Important Factors and Effective Models for Twitter Trend Prediction. In J. Kawash (Ed.), *Online Social Media Analysis and Visualization* (pp. 81–98). https://doi.org/10.1007/978-3-319-13590-8_5

Zheng, X., Han, J., & Sun, A. (2018). A survey of location prediction on Twitter. *IEEE Transactions on Knowledge and Data Engineering, 30*(9), 1652–1671. https://doi.org/10.1109/TKDE.2018.2807840

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.