Information Sciences 646 (2023) 119414

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Crime risk prediction incorporating geographical spatiotemporal dependency into machine learning models

Yue Deng [a,b], Rixing He [a,b], Yang Liu [c,*]

[a] College of Resources Environment and Tourism, Capital Normal University, 105 West Third Ring Road North, Haidian District, Beijing 100048, China

[b] Key Laboratory of Three-Dimensional Information Acquisition and Application, Ministry of Education, Capital Normal University, 105 West Third Ring Road North, Haidian District, Beijing 100048, China

[c] Beijing Academy of Science and Technology, No.27, West 3rd Ring Rd North, Beike Building, Beijing 100089, China

ARTICLE INFO

Keywords:
Crime risk prediction
Spatiotemporal dependency
Inverse distance weighting
Spatiotemporal lag variable
Machine learning

ABSTRACT

The spatiotemporal distribution of crime is closely related to the environment, exhibiting a typical characteristic of "spatiotemporal autocorrelation". However, most of the existing machine learning-based crime prediction methods have difficulty in simulate the spatiotemporal dependence of crime. In this study, we mitigate the spatiotemporal dependence embedded in crime data by introducing a spatiotemporal lag variable. To verify the feasibility of the proposed methods, four machine learning methods were used to determine whether considering spatiotemporal dependency could improve model prediction accuracy and explore the impact of various factors (i.e., environmental factors and demographical factors) on crime risk intensity in different locations using crime data collected from June 2014 to May 2018 in Dallas. The results indicated the following: (1) incorporating spatiotemporal lag variables can effectively improve the prediction accuracy of machine learning models; (2) variables predicting crime are highly nonlinear over time and space, and tree-based nonlinear models greatly outperform linear models in predicting crime; and (3) interpretable machine learning models can reveal the unique contribution of each variable to researchers and practitioners. These findings contribute to our understanding of the mechanism of crime occurrence and may guide the development of crime prevention strategies.

1. Introduction

Accurate predictions crime risk in specific scenarios can not only be used to help residents select safe places live to protect their lives and property but also enable police departments to increase patrols in high-risk areas to prevent and reduce crimes by reducing response time [1,2]. Fig. 1 illustrates the implicit link between spatial and temporal crime patterns by showing the results of Anselin local Moran's I calculations for three consecutive years of robberies in the city of Dallas. The spatial autocorrelation index indicates that the occurrence of crime is not random but shows some spatial dependence. And the crime hot and cold spots in Dallas seem to remain relatively stable over a three-year period, which indicates the time-dependent of crime events. Overall, the occurrence of crime events exhibits typical spatiotemporal dependencies [3], regardless of the spatiotemporal units of research [4]. However, existing

* Corresponding author.
E-mail address: liuyang.bjast@gmail.com (Y. Liu).

https://doi.org/10.1016/j.ins.2023.119414
Received 8 July 2023; Accepted 18 July 2023
Available online 26 July 2023

crime prediction models, such as random forest and eXtreme Gradient Boosting, are nonspatial and do not require calibration with spatial parameters, making it difficult to simulate criminal spatiotemporal dependency [5], which decreases the accuracy of crime prediction to a certain extent.

To address these deficiencies, we use spatiotemporal lag variables to test the hypothesis that taking into account spatiotemporal dependence might greatly increase the accuracy of crime prediction. This article's remaining sections are arranged as follows. The methods used to forecast crimes are reviewed and critically analyzed in detail in Section 2. Several machine learning-based crime prediction models and evaluation metrics used in this study are presents in Section 3. The selection of the case study area, data sources, and influencing factors are covered in Section 4. In Section 5, we assess the proposed method on a real-world crime dataset and provide a clear discussion of the results. Finally, in Section 6, a sensitivity analysis of the construction parameters of spatiotemporal lag variables is presented. The conclusions of the study and a discussion of possible future studies are also included.

2. Review of related work

2.1. The methods of predicting crime

Crime prediction is a difficult issue in criminology and urban security research [6] that involves analyzing and modeling existing crime data and various related factors (environmental and demographic) that may affect crime and making judgments about the crime situation, structure, and trends in a specific time and space in the future. The most used methods can be grouped into the following four categories: (1) hotspot detection, (2) near-repeat prediction, (3) regression-based, and (4) machine learning.

Hotspot detection technology is a traditional and effective strategy for identifying spatiotemporal distribution patterns of crime, which usually identifies high-risk areas of crime by simply using the geographic location of crime events [1]. Commonly used methods include cluster analysis [7], kernel density estimation (KDE) [8], geostatistics [9] and spatiotemporal scan statistics [10]. It can swiftly locate crime hotspots and depict how crime is distributed both spatially and temporally, assisting law enforcement in take the appropriate measures to avoid and lessen crime incidents [11,12]. However, the hotspot identification strategy also has several disadvantages. First, the predictions of future crime only assume that the locations of past hotspots are the same as the locations of future hotspots, without developing a theoretical explanation for the generation of crime hotspots. Second, the hotspot identification usually relies on historical data, the model's accuracy suffers due to the spatial hotspots of crime occurrences fluctuate over time.

The near-repeat prediction model describes the empirical observation that neighborhoods where crime has previously occurred have a higher probability of committing crime in the future. The most commonly used models include the Knox test [13], self-exciting point process (SEPP) [14], and space–time epidemic-type aftershock sequence (ETAS) [15]. Most previous "near repeat" studies used only the time, location, and type of crime to make predictions, without fully accounting for the heterogeneity of the environment surrounding the crime events. These methods can determine on which spatial extent and temporal scale crime events show distinctive aggregation characteristics, but it is difficult to reveal the crime occurrence mechanism and the spatial propagation pattern of crime events in detail.

The core idea of the regression-based model is to substantially model crime risk using independent variable factors to achieve long-term crime risk prediction. Among them, the risk terrain model (RTM) [16] and the negative binomial regression model[17] are two of the most effective regression-based models. Since the parameter estimates of regression models quantify the strength of the relationship between independent variables and crime risk, it is straightforward to explain why a place is considered high-risk for crime. In addition, when there is not enough historical crime data for crime pattern analysis in a region, the regression-based model can predict crime risk from regional environmental factors. However, it has the disadvantage that the model almost always assumes a linear relationship between the predictor and dependent variables, ignoring the spatiotemporal dependence and spatiotemporal non-stationarity characteristics of crime data.

Machine learning is widely used in crime prediction because it can process high-dimensional data quickly and efficiently and provides a better fit than traditional regression-based models even when the data are incomplete. Representative approaches include neural network models [18], random forest models [1,19], and graph convolution models [20]. Although most machine learning methods outperform other models in prediction tasks, they are usually nonspatial and do not require the calibration of spatial

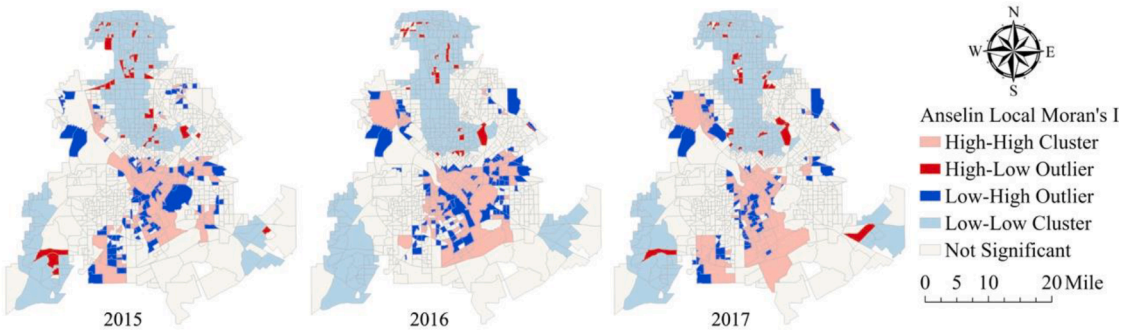

Fig. 1. Spatial autocorrelation results for the distribution of robbery crimes in Dallas from 2015 to 2017.

parameters, making it difficult to model the spatiotemporal dependence present in crime data [21]. Relevant studies [22–24] concluded that using spatiotemporal lag variables as explanatory variables in prediction tasks can effectively alleviate the spatiotemporal dependence present in data and significantly improve the predicted results. However, whether considering spatiotemporal dependence in a crime prediction task can yield the same significant performance improvement has not been determined thus far.

The main contributions of this study are as follows: 1) Four supervised machine learning methods are used to demonstrate that it is necessary to consider spatiotemporal dependencies in the predictive modeling of crimes. 2) Through the introduction of Shapley additive explanations (SHAP) values, we demonstrate that each feature is spatially different in its importance to the predictive outcome of the model. The results provide targeted guidance for prediction-based policing.

3. Methodology

Supervised machine learning can be used to infer a predictive model from a set of labeled training instances that are applied to new unseen examples, and is not a new method for crime prediction. However, few studies have considered spatiotemporal dependencies when modeling the relationship between crime risk and geospatial data. Given the above research background, four supervised machine learning algorithms (linear regression, k nearest neighbors, random forest and eXtreme gradient boosting) were selected for this study to illustrate the impact of considering the spatiotemporal dependency of dependent variables on crime prediction accuracy. In addition, various evaluation metrics, such as PAI, PEI RRI and ROC, were utilized to examine the prediction accuracy of the model.

3.1. Linear regression

Linear regression (LR) is a typical representative of supervised machine learning methods, which assumes a linear relationship between the crime risk y and the independent features x. The goal of the methods is to identify the optimal linear equation that can predict the risk of the crime event based on the independent geospatial features. Given a dataset $D = \{x_i, y_i\}$, the LR model can be expressed as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, i = 1, 2, \cdots, n \tag{1}$$

where y_i represents the risk of crime, β_0 is the constant value, β_k is the coefficient of the $k-th$ independent feature, and x_{ik} is the $k-th$ independent feature. ε_i represents the residual term. An important requirement of LR method is that there should be no multicollinearity among these independent features. In addition, because LR method is based on the assumption that there is a linear relationship between the independent and dependent features, the results can more easily explain why a region has a higher crime risk than the results produced by "black box" machine learning algorithms.

3.2. K nearest neighbors

K nearest neighbors (KNN) is a basic and simple supervised machine learning methods that can be used for both regression and classification tasks. In a regression task, its core step is to find the k nearest training samples to the predicted sample, and the outcome of the predicted sample is the average of the labels of the k nearest training samples[25]. The KNN method is inefficient due to the needs to calculate the distance between all training and prediction samples separately and to find the top k nearest training samples to the prediction samples. To improve the training efficiency, this study uses the K-D tree to find the nearest neighbor samples.

3.3. Random forest

Random forest (RF) is a representative integrated machine learning approach that uses decision trees as a base predictor [26]. There are two ways to interpret the term 'random'. First, the sample features are selected randomly. That is, some, but not all, features are chosen randomly to serve as training features for the decision tree. Second, the training samples count is determined randomly. In other words, training samples for constructing each decision tree are chosen randomly from the training dataset. Eventually, the term "forest" refers to generating multiple regression trees for the samples and combining the predictions of these trees to obtain the result by voting. In addition, RF is insensitive to multicollinearity, relatively robust to missing and unbalanced data, and yields reasonable predictions [1].

3.4. eXtreme gradient boosting

eXtreme gradient boosting (XGBoost) is a generalized gradient boosting methods that is a common technique in ensemble learning, mainly through algorithm-based innovation and hyperparameters optimization to improve learning efficiency and prevent overfitting [27]. XGBoost is a CART-based classifier consisting of several correlated decision trees jointly, where the input samples of the next decision tree are correlated with the training and prediction results of the previous decision tree, which can be used to solve most regression problems. Given a dataset $D = \{x_i, y_i\}, i = 1,2,\cdots,n$, where x_i is the training dataset, y_i is the target variable. n is the number of samples. f represents a regression tree, and F represents the set of regression trees. The model can be expressed as follows:

$$\overline{y_i} = \sum_{k=1}^{K} f_k(x_i), f \in F \tag{2}$$

The objective function is defined as follows:

$$obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{k} (\Omega f_k) \tag{3}$$

$$\Omega f_k = \gamma T + 1 \Big/ 2\lambda \sum_{j=1}^{T} \omega_j^2 \tag{4}$$

In the formula, $\hat{y}_i$ and y_i represent the predicted and observed value, respectively, l is a loss function, mainly used to measure the difference between $\hat{y}_i$ and y_i. Ωf_k represents the regularization term of the k tree, which is used to penalize the model complexity to avoid overfitting. T represent the total number of leaf nodes in the decision tree, ω_i is the leaf weight value, γ is the penalty coefficient, and λ is the weight penalty coefficient.

3.5. Model construction and evaluation

3.5.1. Model construction

In this study, LR, KNN, RF, and XGBoost were used to evaluate the accuracy of crime prediction with/without considering the spatiotemporal lag variables. The models and independent variables were selected as shown in Table 1.

3.5.2. Model evaluation

We employed four established measurement indicators to assess the performance of the prediction models: the predictive accuracy index (PAI), predictive efficiency index (PEI), recapture rate index (RRI) and receiver operating characteristic curve (ROC).

$$PAI_t = n_t/N/a_t/A \tag{5}$$

where n_t is the number of crime events successfully predicted with a fixed number of study areas of t, and N is the total number of crime events in the predicted period. α_t is the number of predicted areas of t, and A is the total number of study areas.

$$PEI = PAI_t/PAI_m \tag{6}$$

where PEI is the ratio of actual PAI_t over the maximum PAI_m for a fixed number of study areas of t. The value range of PEI is between 0 and 1.

$$RRI = P_t/O_t \tag{7}$$

where P_t is the number of crime events predicted for a fixed number of study areas of t, and O_t is the number of crime events observed in practice. An RRI value greater than 1 indicates that the model used overpredicts the crime risk. An RRI value smaller than 1 suggests that the model underpredicts the crime risk.

A simplified example is given to explain these three metrics. If there are a total of 100 crimes in 500 grid cells in a study area and 50 crimes are captured in the top 1% of grid cells (5 grids) using historical crime data, the PAI is (50/100)/(5/500) = 0.5/0.01 = 50. If the top 5 performing grid cells in the current data capture exactly 60 crimes in total, then the PEI is 50/60 = 5/6. If the actual number of crime events in the top 1% of grid cells is 50 and the predicted number of crimes is 60, then the RRI is equal to 60/50 = 1.2. Thus, the value of each indicator changes according to the set regional threshold (from 0 to 100%). The purpose of introducing these indicators is to help police departments allocate limited resources to some of the top ranked areas (e.g., 1% or 5%) that are identified as having the highest crime risk.

3.6. Interpretable XGBoost

Although machine learning models have higher predictive accuracy and better model fits than traditional regression-based models for most prediction tasks, explaining their prediction processes transparently is more difficult. The feature importance process in traditional machine learning packages only indicates which features are important but does not explain how those features affect the

Table 1
List of models and independent variables used in the experiment.

Models	Environmental variables	Demographic variables	Spatiotemporal lag variable
Linear regression	✓	✓	
Linear regression + ST_lag[a]	✓	✓	✓
KNN	✓	✓	
KNN + ST_lag	✓	✓	✓
Random forest	✓	✓	
Random forest + ST_lag	✓	✓	✓
XGBoost	✓	✓	
XGBoost + ST_lag	✓	✓	✓

[a] represents the model considering the spatiotemporal lag variable.

prediction results. Thus, to improve the interpretability and transparency of crime prediction models, we introduce SHAP [28] to explain which variables are most important for learning and making decisions. How it works is shown in the following equation.

$$\varnothing_i(f,x) = \sum_{S \subseteq N \setminus \{i\}} (|S|!(n-|S|-1)!/n!)(f_x(S \bigcup \{i\}) - f_x(s)) \tag{8}$$

where $\varnothing_i(f,x)$ denotes the value for each variable i, S denotes the set of features used in the model, and f_x is the prediction obtained by considering the indicated set of features. $|S|!(n-|S|-1)!/n!$ is the weight, and $(f_x(S \bigcup \{i\}) - f_x(s))$ represents the difference between the values acquired before and after the new feature i was added. The feature importance levels can be sorted by comparing the attribution values of different features. The greatest advantage of the Shapley method is that the SHAP values can not only reflect the influences of the features in each sample but also present whether these impacts are positive or negative.

4. Data and preprocessing

4.1. Data source and the units of analysis

In this study, we focused on robbery, which is a violent crime in which someone take something valuable from another by using force or threatening to use force [6,29]. Federal Bureau of Investigation (FBI) research has shown that robberies occur frequently and mostly on the street, which makes them ideal crime events for spatially oriented research [30]. The data used in this study was downloaded from the Dallas Open Data portal (https://www.dallasopendata.com/), and includes information on interpersonal robberies reported in Dallas from June 2014 to May 2018. We predicted the risk of robbery in Dallas using a 200 by 200 feet grid cell, which facilitates the spatial correlations of crime, demographic and environmental factors across the research region. A total of 240,481 grid cells were generated for analysis using the Dallas administrative boundary. Since crime patterns at the micro level tend to exhibit strong spatiotemporal stability over time [31,32] (Fig. 1), we evaluated the accuracy of the method proposed in this study on a yearly cycle and further revealed important factors that may affect crime risk, which is of great significance for crime-reduction oriented urban planning and police patrols.

4.2. Impact factors

The factors influencing the distribution of crime risk are complex and variable. According to their attributes, these factors are divided into three categories: environmental variables, demographic variables and spatiotemporal lag variables.

4.2.1. Environmental variables

Crime risk are highly correlated with the surrounding environment, that is, crime tends to be concentrated around some facilities (such as bus stations and bars). In this study, 22 environmental factors (as shown in Table 2) were used to generate a total of 36 independent variables that were incorporated into crime prediction models. Specifically, for 18 types of point of interest (POI) data (ID 1-ID 18 in Table 2), such as larger business retailers, banks, hotels, and hospitals, we use the density of each POI within a 1000-foot

Table 2
Environmental variables.

ID	Factors	Total number	Area (Sq. Miles)	Distance	Density
1	Smaller food stores	8014		✓	✓
2	Eating and drinking places	4313		✓	✓
3	Apartments	3258		✓	✓
4	Larger business retailers	2572		✓	✓
5	Gyms and hair salons	2209		✓	✓
6	Movie, amusement and entertainment	795		✓	✓
7	Gasoline stations	759		✓	✓
8	Liquor stores	498		✓	✓
9	Banks	367		✓	✓
10	Check-cashing stores	267		✓	✓
11	Middle or high schools	240		✓	✓
12	Motels	230		✓	✓
13	Hotels	230		✓	✓
14	Mobile home parks	139		✓	✓
15	DART stations	49		✓	
16	Hospitals	44		✓	
17	Libraries	26		✓	
18	Shopping malls	10		✓	
19	Coordinates (x)	30,120			
20	Coordinates (y)	30,120 [a]			
21	Parks		25.3[b]	✓	
22	Street line buffer proportion		117.8[b]		

[a] represents the total number of factors, and [b] is the total area of the factor.

bandwidth and the distance to the nearest POI as inputs to estimate the number of crimes within grid cells. Since the numbers of libraries, DART stations, hospitals and shopping malls in Dallas are relatively small and dispersed, the density was not calculated, and only the nearest distance was considered. At the same time, to capture the potential spatial relationships among the coordinates and particular crimes, the coordinates of the grid centroid were used as covariates in subsequent models (ID 19 and ID 20). For the park (ID 21), we used the distance as an independent variable. Another independent variable (ID 22) for crime modeling was the percentage of grid cells occupied by the road buffer (72-foot). The rationale for factor selection and a more detailed description can be found in the study by Andrew [1] et al.

4.2.2. Demographic variables

Theories about crime, such as social disorganization theory [33,34], suggest that demographic factors can explain the spatial distribution characteristics of urban crime to some extent. Therefore, demographic data should be fully considered in microscale crime prediction to improve the prediction ability of the model. The 5-year Community Survey (2014) estimated at the block level were utilized to calculate the eight demographic parameters used in this analysis (Table 3).

4.2.3. Spatiotemporal lag variable

We observed in Fig. 1 that the crime data have spatiotemporal autocorrelation, which violates the assumption of mutual independence of observations in ordinary linear regression models and may affect the accuracy of the model [35,36]. For each unit of analysis, we wanted to know whether the historical crime information within a certain distance from the unit could help predict crime risk in the next period. Therefore, we included historical crime data within a certain neighborhood of each unit as explanatory variables in the model. Specifically, as shown in Fig. 2, a temporal search bandwidth of 24 months and a spatial search bandwidth of 200 m were used to calculate the spatiotemporal lag variables for each unit using the inverse distance weighting (IDW) method (see Section 6.1 for the basis of bandwidth selection) [37].

where ST_lag_i denote the spatiotemporal lag for analysis unit i; d_{ij} and $d_{t_{ij}}$ denote the spatial and temporal distance between i and j, respectively; r_s and r_t denote the spatial and temporal search bandwidth, respectively; and nc_i denotes the number of crimes in grid j.

All the above variables used for modeling were obtained by integrating and cleaning open-source data, commercial data and local government data, which were geocoded to the grid level. To accurately validate the fit of the model, we divided the data into training sets and verification sets at a ratio of 7:3.

5. Results evaluation

Four machine learning algorithms (LR, KNN, RF, and XGBoost) were applied, in the experimental environment, M1. The software R was used, including the mlr3, caret and SHAPforxgboost packages. The optimal combination of parameters for the machine learning model was determined via 5-fold cross validation.

5.1. VIF feature selection

A high correlation between independent variables reduces the accuracy and interpretability of the model, especially the linear regression model. To mitigate the effect of multicollinearity on modeling results, we used the variance inflation factor (VIF) for multicollinearity testing and eliminating features with large covariance. When the VIF value of an independent variable is greater than 10, it is considered to have considerable covariance. According to the rule, the variable "distance to the nearest shopping malls (VIF = 11.58)" was excluded from the next modeling, and 44 features without collinearity were retained, as shown in Table 4.

5.2. Accuracy of the models

The general consensus among modern researchers and practitioners is that setting thresholds helps allocate limited police resources to manageable regions with the highest risk of future crime to effectively maximize public safety gains [38,39]. Therefore, in this study, we identified the top n areas with the highest crime rates by setting a fixed threshold. These areas with high crime risk are of most concern and are conducive to rational police force allocation. Fig. 3 shows the PAIs PEIs, RRIs and ROCs of robberies captured by different models in the top 1% regions (more than 700 grids). Fig. 3(a) shows that the model that considered spatiotemporal lag (solid

Table 3
Demographic variables.

ID	Factors	Mean	SD	Connotation
1	Poverty Percentage	0.21	0.16	Percentage of poor families
2	Percentage Female	0.17	0.14	Percentage of families headed by women with children under 18
3	Percentage Moved	0.16	0.14	The percentage of people who migrated within the last year
4	Percentage Unemployed	0.09	0.08	The percentage of those over 16 who are unemployed
5	Percentage Hispanic	0.27	0.23	The proportion of the population who are Hispanic
6	Percentage Non-Hispanic Black	0.28	0.29	The percentage of black people who are non-Hispanic
7	Ethnic Heterogeneity Index	0.48	0.22	Simpson's Index is used to calculate an ethnic heterogeneity index
8	Population Density	5.58	6.43	The density of the residential population

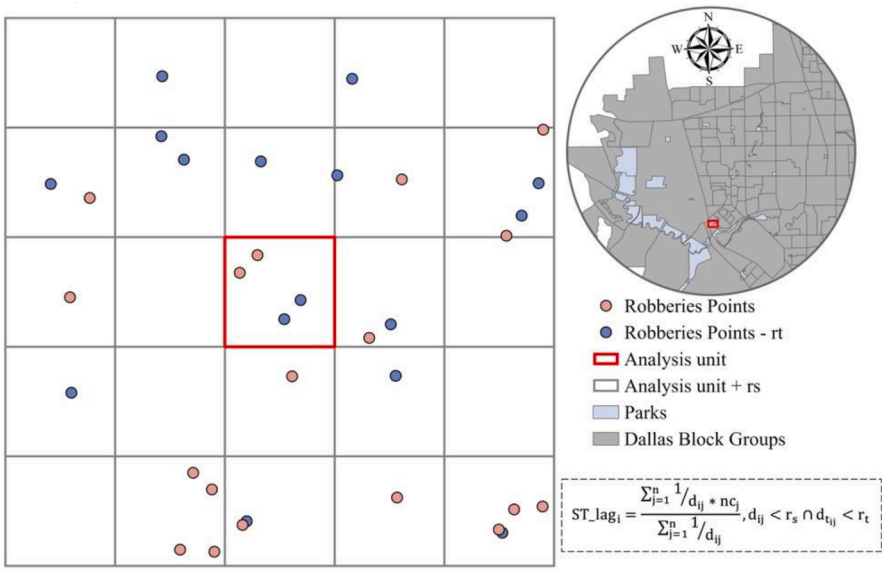

Fig. 2. Spatiotemporal search bandwidths in IDW.

Table 4
VIF feature selection results.

ID	Features	VIF	ID	Features	VIF
1	X	2.155	23	Dist_Hospital	5.147
2	Y	8.097	24	Buffer_Motel	1.299
3	Buffer_Larger business retailers	2.320	25	Dist_Motel	2.699
4	Dist_Larger business retailers	6.063	26	Buffer_Hotel	1.447
5	Buffer_Smaller food clothing stores	2.297	27	Dist_Hotel	2.336
6	Dist_Smaller food stores	7.8396	28	Buffer_Mobile home park	5.228
7	Buffer_Gasoline stations	1.956	29	Dist_Mobile home park	1.258
8	Dist_Gasoline stations	3.365	30	Buffer_Bank	4.524
9	Buffer_Eating and drinking places	3.330	31	Dist_Bank	1.453
10	Dist_Eating and drinking places	6.821	32	Buffer_Check cashing	4.227
11	Buffer_Liquor stores	1.591	33	Dist_Check cashing	1.845
12	Dist_Liquor stores	7.587	34	Dist_Park	1.647
13	Buffer_Movie theaters services	1.634	35	Street line buffer proportion	1.183
14	Dist_Movie theaters services	3.494	36	Percentage Poverty	2.733
15	Buffer_Gyms and hair salons	2.303	37	Percentage Female	1.665
16	Dist_Gyms and hair salons	4.871	38	Percentage Unemployed	1.656
17	Dist_Libraries	2.069	39	Percentage Moved	1.542
18	Buffer_Middle or high school	1.327	40	Ethnic Heterogeneity Index	4.289
19	Dist_Middle or high school	2.638	41	Percentage Non-Hispanic black	2.830
20	Dist_Dart stations	2.784	42	Percentage Hispanic	2.167
21	Buffer_Apartment	1.869	43	Population Density	1.780
22	Dist_Apartment	2.733	44	ST_lag	2.121

line) outperformed the model that did not consider spatiotemporal lag (dotted line), regardless of the choice of threshold, models, and evaluation metrics. In addition, the PAI values decreased slightly as the number of regions increased, and each model had the highest PAI only in the top n areas.

In Fig. 3(b), only the top 1% of the research area are included in the PEI metrics. In comparison to the best forecast, the findings demonstrate that XGBoost + ST_lag successfully captured more than 55% of the robberies. This is a much better result than XGBoost, which tended to capture only 10% − 30% of the crimes. In addition, KNN performed the worst out of each method and captured less than 20% of the robberies. The closeness of the RRI to 1 represents the accuracy of the prediction results. The RRI metric depicted in Fig. 3(c) shows that the model that considered spatiotemporal lag only slightly underestimated the total number of robberies, with ratios typically greater than 0.8, while models that did not consider spatiotemporal lag overestimated these values, with ratios often greater than 1 and even close to 5. Also, among the eight models, XGBoost + ST_lag and linear regression + ST_lag had the most accurate predictive values (closest to 1), while the linear regression model had the worst accuracy. The Y-axis of the ROC curve depicted in Fig. 3(d) again demonstrates that considering spatiotemporal lag variables can greatly increase the model accuracy, which also confirms that the distribution of crime events is characterized by spatiotemporal aggregation [40,41].

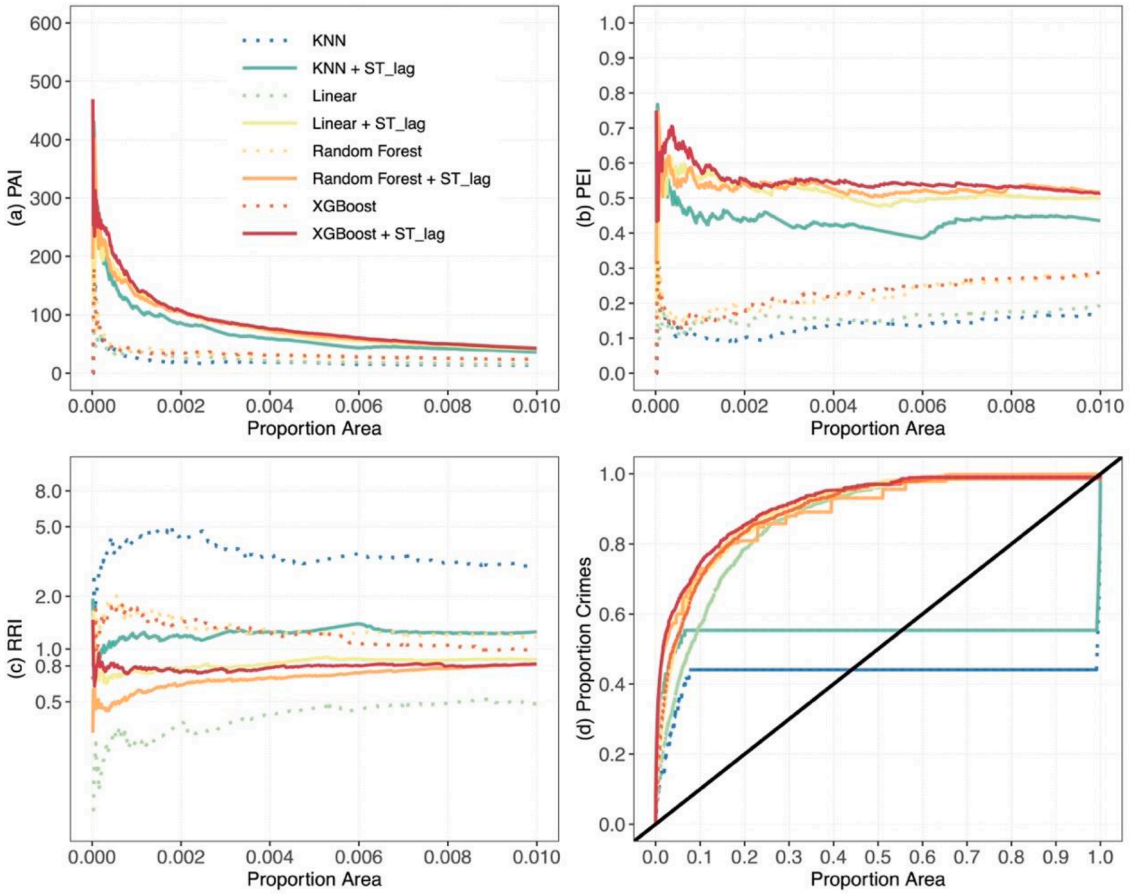

Fig. 3. Accuracy metrics obtained under fixed thresholds for each model.

5.3. *Interpreting XGBoost*

5.3.1. *Global interpretability*

The optimal XGBoost model was utilized in this study to describe the effect of the independent variables on predictive model by plotting the SHAP value of features in analysis grid. In Fig. 4(a), the y-axis represents different independent features, and the x-axis depicts the effects of the independent features on the model outputs. When the independent and dependent variables have a positive correlation, the SHAP value is positive; when it is negative, the correlation is negative. The magnitudes of the SHAP values are indicated by different colors: the color changes from yellow to purple to represent the change of SHAP value from low to high. For example, Fig. 4(a) shows that small food stores had the greatest influence on robbery than distance to the apartment. Other variables, such as the distance to middle or high school, the distance to hospital and the percentage of unemployed, had a smaller impact on the model.

Fig. 4(b) to 4(g) further show the interrelationship between crime and the variable by plotting the trend between specific variables and SHAP values. For example, Fig. 4(b) illustrates the effect of distance to the nearest apartment on model output. Smaller values of the variables were associated with higher SHAP values, indicating that they were associated with robbery. That is, the grids closest to the apartment (or even the apartment itself) were more likely to have robberies than other grids. Fig. 4(d) illustrates the impact of the density of apartments on the model output. Larger values of variables were associated with higher SHAP values, indicating that they were associated with robbery. That is, grids with more apartments were more likely to have robberies than other grids. Fig. 4(e) and Fig. 4(g) illustrate the effect of eating and drinking place density and motel density on the model output, respectively. The higher the value of the variable, the higher the SHAP value, indicating that grids with higher eating and drinking place density or motel density had a higher risk of robberies. The inverse relationship between bank density and SHAP values in Fig. 4(f) illustrates that higher bank density was associated with a lower chance of robbery, which is contrary to our intuition. This suggests that the more banks there are, the stronger the security measures and the less conducive to criminals committing robberies. In addition, a higher spatiotemporal lag variable (Fig. 4(c)) was associated with larger SHAP values, implying a positive association with robbery. This further confirms that considering spatiotemporal lag variables can increase the model accuracy to a certain extent.

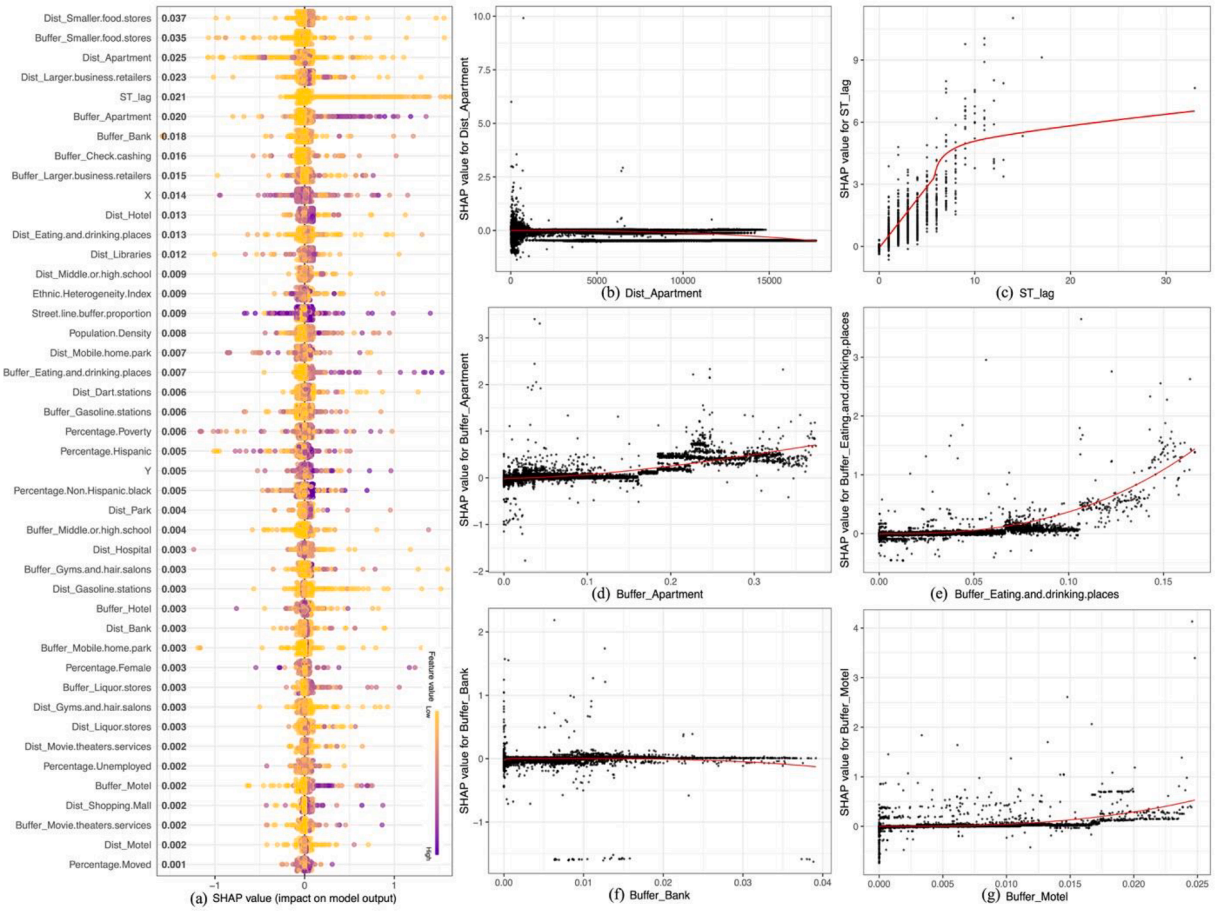

Fig. 4. Distributions of SHAP values.

5.3.2. *Local interpretability*

Analyzing the spatial contributions of independent variables to the robbery prediction model helps explain the occurrence of crime in an area, which can provide meaningful information to the police to develop target crime prevention strategies. Fig. 5 shows the spatial variation in the contribution of several selected variables to the model. Each feature has a unique spatial distribution pattern of importance. That is, a feature shows different crime prediction effects in different regions. In general, variables such as apartments, gas stations and streets were important in predicting crime in Dallas, while the density of large commercial retailers was a weak predictor of crime. Specifically, Fig. 5(a) demonstrates that the majority of the contribution of population density to the crime prediction model was located in the northern parts. Fig. 5(b) shows that the contribution of the distance to the nearest gas station was mostly limited to the southern part of the city. Fig. 5(c) and 5(d) show that the density of apartments and the distance to the nearest apartment were two strong predictors, and their effects were evident in all areas of Dallas. Fig. 5(e) shows that the contribution of the density of large business retailers was a weak predictor of crime in Dallas, with values below 0.5. The main reason may be that large business retailers are more dispersed in Dallas, so their density values are low, and their contribution to crime prediction is lower. Fig. 5(f) shows that the contribution of the street line buffer proportion was also a strong predictor for crime prediction in Dallas and shows a stronger spatial characteristic that had a stronger effect outside the city than inside the city, indicating that robberies were more likely to occur on streets and that suburban streets were more prone to robberies than streets in urban centers.

To further clarify the importance of local features on the XGBoost model in each grid, we randomly selected two grid samples for local interpretability analysis. In general, the SHAP value of the sample grid represents the correlation or causal relationship that exists between the independent feature and dependent feature, so it can guide the police department in developing robbery prevention strategies in specific places. The SHAP value of the two arbitrarily selected samples in Fig. 6 represents the contribution of each dimensional feature of each individual sample to the model output. Positive SHAP values indicate the positive effect of the corresponding feature on the model, while negative values indicate a negative effect. For example, at location 6000 (Fig. 6(a)), the total SHAP value was 0.468, which was greater than the base value of 0.415. The features with positive SHAP values were the distance to the nearest gasoline station, distance to a large business retailer and distance to apartments, while the variables with negative SHAP values included spatiotemporal lag, percentage female and the remaining features. At location 28,600 (Fig. 6(b)), the total SHAP value was 0.381, which was smaller than the base value of 0.381. The features with positive SHAP values were the percentage females and the

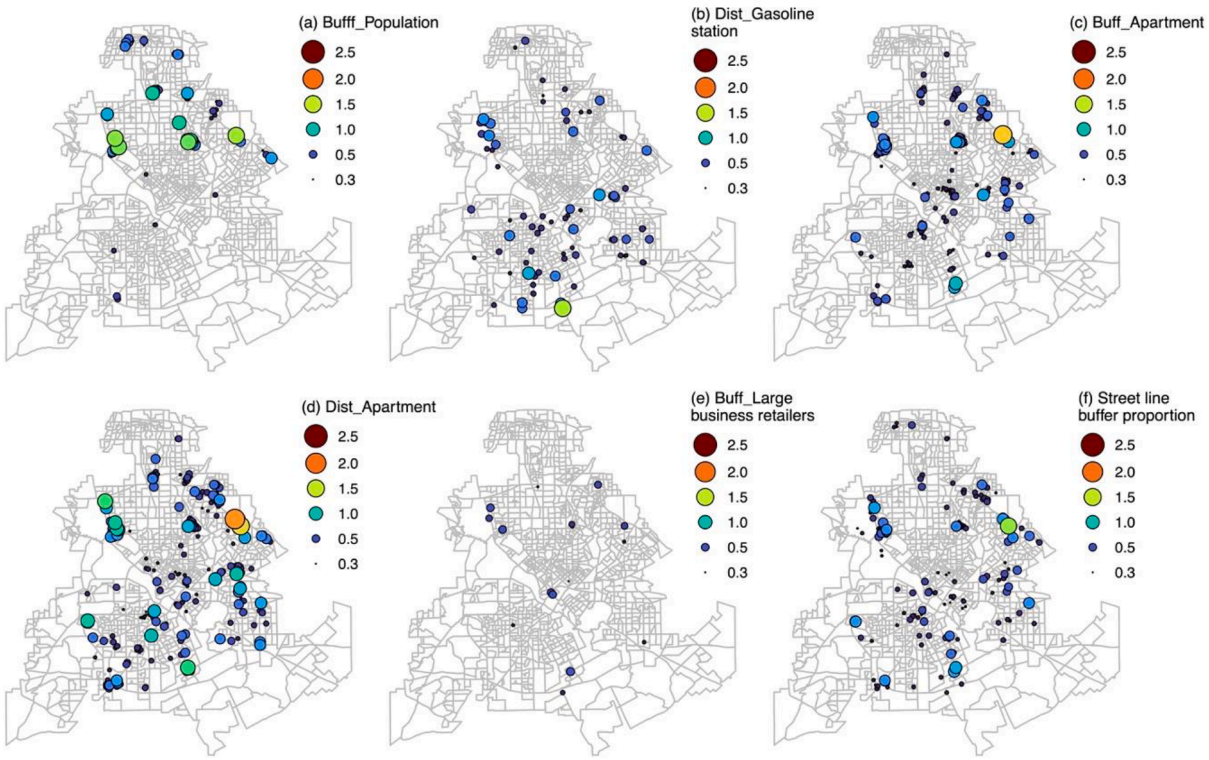

Fig. 5. Contribution of different independent variables to predicted crime counts over space.

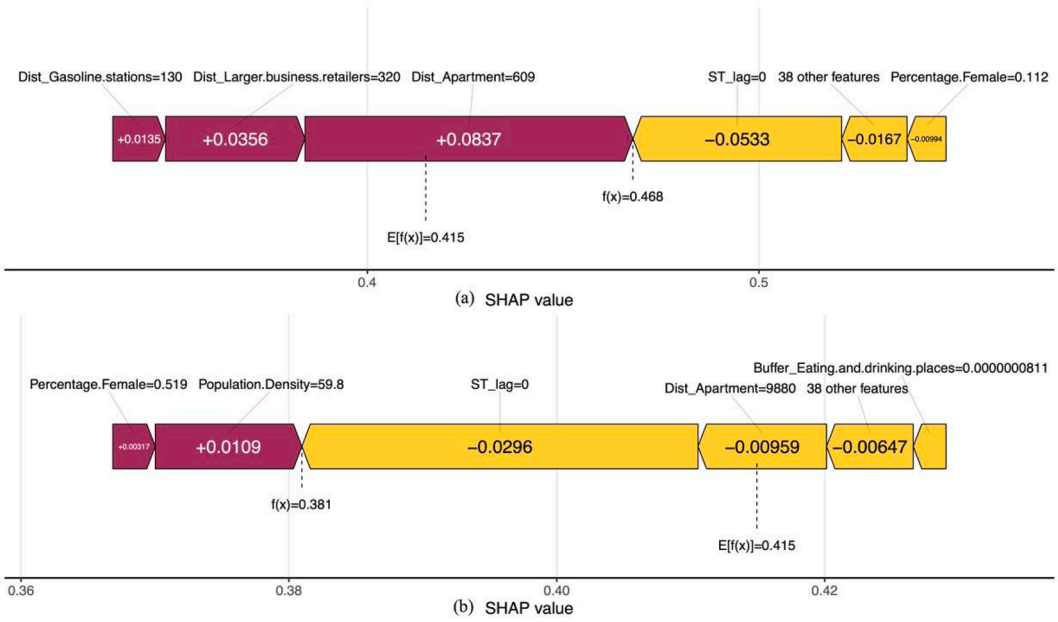

Fig. 6. The local SHAP values of the variables at two randomly selected grid samples.

population density, while the variables with negative SHAP values were the spatiotemporal lag, the distance to the nearest apartment, and the remaining variables. These two instances demonstrate that the local effects of independent variables on robbery may operate in different directions than the global model and may vary by region.

6. Discussion and conclusion

6.1. The effects of the spatial and temporal search bandwidths

We observed that the spatiotemporal lag variable can increase the model accuracy in Section 5.2. Temporal search bandwidths and spatial search bandwidths were two important decisive parameters in the process of constructing spatiotemporal lag variables. Therefore, we designed a series of different temporal search bandwidths and spatial search bandwidths and selected the optimal parameters through several combination experiments.

Table 5 shows the cumulative number of crimes captured by the model and the PAI values of crime predictions on a fixed number of grids when different temporal search bandwidths (3 months to 24 months) were chosen. The results show that the PAI values increased with increasing temporal search bandwidth, regardless of the number of fixed grids. That is, the more historical crime data there are, the higher the prediction accuracy of the model. This is also consistent with the "spatially approximate repetition" phenomenon of crime events.

Based on the determination of the optimal temporal research bandwidth of 24 months, Table 6 further shows the cumulative number of crimes captured by the model and the PAI value of crime prediction in the fixed grid for different spatial search bandwidths (200 feet to 2000 feet). The results show that the PAI value gradually decreased with increasing spatial search radius, and the best results were obtained when the 200-foot grid was chosen. This implies that extending the range of explanatory variables to the surrounding rid may reduce the model accuracy. Overall, the effect of the spatiotemporal search radius on the crime prediction model confirms that crime events have the characteristics of "spatiotemporal aggregation" and "near repeat" phenomena and further indicates that considering the spatiotemporal lag variable can improve the accuracy when modeling crime risk.

6.2. Conclusion and future research prospects

Crime incidents seriously affect the health and safety of residents. Accurate crime prediction improves our understanding of the mechanism of crime occurrence and can guide public security departments for the reasonable allocation of resources to prevent and reduce crime. However, traditional crime prediction rarely considers spatiotemporal dependence under large data volumes. In this study, four machine learning methods, LR, KNN, RF, and XGBoost, were used to predict crimes for long time series based on the introduction of spatiotemporal lag variables for robbery crime records in Dallas from 2014 to 2018 to evaluate whether considering spatiotemporal dependence could significantly improve the accuracy of crime prediction. The results showed that the inclusion of spatiotemporal lag variables alleviated the spatiotemporal dependence and improved the accuracy of the models to some extent. In addition, the performance of integrated machine learning based on decision trees such as RF and XGBoost was relatively good and accurate in terms of crime risk modeling, compared with models such as KNN and linear regression. Another contribution of this study was the introduction of SHAP values to provide interpretability of variables in crime risk prediction models. For example, small food stores, apartments, and time-lag variables were three important factors that influenced robberies in the study area. In addition, local SHAP values revealed the local microenvironment within each grid, and this information may provide clues to find potential crime risk points to guide police departments in developing data-driven crime prediction strategies or to inform the construction of future safe cities [42].

In addition, there are still some issues that require further research. First, in this study, we used a grid-based approach, and different types of analysis cells (e.g., road segments) are likely to yield more accurate results [43,44]. Therefore, in future work, street segments will be introduced to generate models for predicting crime. Second, as with most crime prediction studies, this study focused only on the environmental spatial variables that lead to crime and ignored the impact of human activities on crime risk. In future research, the accuracy of crime prediction may be improved by introducing big data on human behavior, such as cell phone data and trajectory data. Third, this model was only tested in Dallas. Even if it has high accuracy in this data setting, more research is necessary to determine whether it can be used to detect other types of crimes in different regions.

In conclusion, our aim was to develop a useful crime prediction model and improve the accuracy of crime prediction by introducing spatiotemporal lag variables. Our research also provided insights into crime risk prediction in Dallas. In addition, we generated useful ideas for other spatiotemporal prediction problems.

7. Disclosure statement

No potential conflict of interest was reported by the authors.

8. Data and code availability statement

The data and codes that support the findings of this study are available with the identifier(s) at the following link (https://figshare.com/s/3704d1b845cca9d34d09).

Declaration of Competing Interest

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.

Table 5
Accuracy metrics for different temporal search bandwidths in a fixed number of areas.

ID[a]	rt[b] = 3 months		rt = 6 months		rt = 12 months		rt = 24 months	
	CCrime[c]	PAI[d]	CCrime	PAI	CCrime	PAI	CCrime	PAI
2	6	78.21	7	91.24	10	130.34	17	332.38
20	18	33.52	26	48.41	25	46.55	91	177.92
200	255	39.89	234	36.60	247	38.63	460	89.94
500	410	30.25	427	31.50	442	32.61	670	52.40
1000	603	22.14	638	23.43	650	23.87	875	34.22
2000	851	15.28	915	16.43	877	15.75	1000	19.55
5000	1243	8.31	1261	8.43	1243	8.31	1109	8.67

ID[a] represents the number of areas; rt[b] represents the temporal search bandwidth; CCrime[c] represents the cumulative number of robberies captured within each of the thresholds; PAI[d] is reserved for two decimal places only.

Table 6
Accuracy metrics for different spatial search bandwidths in a fixed number of areas.

ID[a]	rs[b] = 200 ft		rs = 600 ft		rs = 1000 ft		rs = 1400 ft		rs = 1800 ft		rs = 2000 ft	
	CCrime[c]	PAI[d]	CCrime	PAI	CCrime	PAI	CCrime	PAI	CCrime	PAI	CCrime	PAI
2	17	332.38	1	19.55	7	136.86	10	195.51	0	0.00	6	117.31
20	91	177.92	31	60.61	25	48.88	22	43.01	22	43.01	17	33.24
200	460	89.94	232	45.36	187	36.56	214	41.84	218	42.62	186	36.37
500	670	52.40	475	37.15	411	32.14	432	33.78	415	32.46	384	30.03
1000	875	34.22	722	28.23	620	24.24	637	24.91	558	21.82	590	23.07
2000	1000	19.55	929	18.16	893	17.46	861	16.83	784	15.33	833	16.29
5000	1109	8.67	1259	9.85	1228	9.60	1201	9.39	1192	9.32	1206	9.43

ID[a] represents the number of areas; rsb represents the spatial search bandwidth; CCrime[c] represents the cumulative number of robberies captured within each of the thresholds; PAI[d] is reserved for two decimal places only.

Data availability

Data will be made available on request.

References

[1] A.P. Wheeler, W. Steenbeek, Mapping the Risk Terrain for Crime Using Machine Learning, J. Quant. Criminol. 37 (2) (2021) 445–480.
[2] A.A. Braga, A.V. Papachristos, D.M. Hureau, The Effects of Hot Spots Policing on Crime: An Updated Systematic Review and Meta-Analysis, Justice Q. 31 (4) (2014) 633–663.
[3] B.W. Reyns, Environmental criminology: Evolution, theory and practice, Secur. J. 29 (3) (2016) e1–e3.
[4] Z. He, L. Tao, Z. Xie, C. Xu, Discovering spatial interaction patterns of near repeat crime by spatial association rules mining, Sci. Rep. 10 (1) (2020).
[5] S. Georganos, T. Grippa, A. Niang Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff, S. Kalogirou, Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling, Geocarto Int. 36 (2) (2021) 121–136.
[6] H. Zhu, F. Wang, An agent-based model for simulating urban crime with improved daily routines, Comput. Environ. Urban Syst. 89 (2021), 101680.
[7] G. Hajela, M. Chawla, A. Rasool, A Clustering Based Hotspot Identification Approach For Crime Prediction, Procedia Comput. Sci. 167 (2020) 1462–1470.
[8] M.S. Gerber, Predicting crime using Twitter and kernel density estimation, Decis. Support Syst. 61 (2014) 115–125.
[9] H. Yu, et al., Crime Prediction with Historical Crime and Movement Data of Potential Offenders Using a Spatio-Temporal Cokriging Method, Int. J. Geo-Informat. 9 (2020) 732.
[10] S. Shiode, Street-level Spatial Scan Statistic and STAC for Analysing Street Crime Concentrations, Trans. GIS 15 (3) (2011) 365–383.
[11] S. Chainey, L. Tompson, S. Uhlig, The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime, Secur. J. 21 (1) (2008) 4–28.
[12] S. Mondal, D. Singh, R. Kumar, Crime hotspot detection using statistical and geospatial methods: a case study of, GeoJournal 87 (6) (2022) 5287–5303.
[13] J.H. Ratcliffe, G.F. Rengert, Near-Repeat Patterns in Philadelphia Shootings, Secur. J. 21 (1) (2008) 58–76.
[14] K. Bowers, S. Johnson, K. Pease, Prospective Hot-Spotting: The Future of Crime Mapping? Br. J. Criminol. 44 (2004) 641–658.
[15] G. Mohler, Marked point process hotspot maps for homicide and gun crime prediction in Chicago, Int. J. Forecast. 30 (3) (2014) 491–497.
[16] J.M. Caplan, L.W. Kennedy, J. Miller, Risk Terrain Modeling: Brokering Criminological Theory and GIS Methods for Crime Forecasting, Justice Q. 28 (2) (2011) 360–381.
[17] M. Lan, L. Liu, J.E. Eck, A spatial analytical approach to assess the impact of a casino on crime: An example of JACK Casino in downtown Cincinnati, Cities 111 (2021), 103003.
[18] A. Rummens, W. Hardyns, L. Pauwels, The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context, Appl. Geogr. 86 (2017) 255–261.
[19] S. Yao, et al., Prediction of Crime Hotspots based on Spatial Factors of Random Forest. In *2020 15th International Conference on Computer Science & Education (ICCSE)*, 2020.
[20] Wang, Y., et al. *Deep Temporal Multi-Graph Convolutional Network for Crime Prediction*. in *Conceptual Modeling*. 20Cham: Springer International Publishing.
[21] Sun, J., et al., *CrimeForecaster: Crime Prediction by Exploiting the Geographical Neighborhoods' Spatiotemporal Dependencies*, in *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*. 2020, Springer-Verlag: Ghent, Belgium. p. 52–67.
[22] A. Soltani, et al., Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms, Cities 131 (2022), 103941.
[23] T.J. Kiely, N.D. Bastian, The spatially conscious machine learning model. Statistical Analysis and Data Mining, ASA Data Sci. J. 13 (1) (2020) 31–49.
[24] T.D. Phan, Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. in *2018 International Conference on Machine Learning and Data Engineering (iCMLDE)*, 2018.
[25] Guo, G., et al. *KNN Model-Based Approach in Classification*. 2003. Berlin, Heidelberg: Springer Berlin Heidelberg.

[26] L. Breiman, Random Forests, Mach. Learn. 45 (1) (2001) 5–32.
[27] Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[28] Shapley, L.S., *17. A Value for n-Person Games*, in *Contributions to the Theory of Games (AM-28), Volume II*, K. Harold William and T. Albert William, Editors. 2016, Princeton University Press. p. 307-318.
[29] R. Wright, S.H. Decker, N. Shover, Armed Robbers in Action: Stickups and Street, Culture (1997).
[30] Investigation., F.B.O., *Crime in the United States.* Washington, DC: United States Department of Justice., 2014.
[31] A.P. Wheeler, R.E. Worden, S.J. McLean, Replicating Group-Based Trajectory Models of Crime at Micro-Places in Albany, NY, J. Quantit. Criminol. 32 (4) (2016) 589–612.
[32] A.S.N. Curman, M.A. Andresen, P.J. Brantingham, Crime and Place: A Longitudinal Examination of Street Segment Patterns in Vancouver, BC, J. Quantitat. Criminol. 31 (1) (2015) 127–147.
[33] Sampson, R.J., *Great American City: Chicago and the Enduring Neighborhood Effect.* 2012, Chicago, IL: The University of Chicago Press.
[34] Sutherland, E.H., Juvenile Delinquency and Urban Areas: A Study of Rates of Delinquents in Relation to Differential Characteristics of Local Communities in American Cities. Clifford R. Shaw , Henry D. McKay , Norman S. Hayner , Paul G. Cressey , Clarence W. Schroeder , T. Earl Sullenger , Earl R. Moses , Calvin F. Schmid. Am. J. Sociol., 1943. **49**(1): p. 100-101.
[35] A.S. Fotheringham, H. Yue, Z. Li, Examining the influences of air quality in China's cities using multi-scale geographically weighted regression, Trans. GIS 23 (6) (2019) 1444–1464.
[36] L. Anselin, et al., Handbook of applied economic statistics, Marcel Dekkes (1998).
[37] C. Daly, Guidelines for assessing the suitability of spatial climate data sets, Int. J. Climatol. 26 (6) (2006) 707–721.
[38] M. Anderson, H. Giles, Fairness and Effectiveness in Policing: The Evidence, J. Commun. 55 (2005) 872–874.
[39] J.M. Caplan, L.W. Kennedy, E.L. Piza, J.D. Barnum, Using Vulnerability and Exposure to Improve Robbery Prediction and Target Area Selection, Appl. Spat. Anal. Policy 13 (1) (2020) 113–136.
[40] T. Ohyama, M. Amemiya, Applying Crime Prediction Techniques to Japan: A Comparison Between Risk Terrain Modeling and Other Methods, Eur. J. Crim. Policy Res. 24 (4) (2018) 469–487.
[41] DAVID Weisburd, The law of crime concentration and the criminology of place, Criminol. Interdiscipl. J. 53 (2) (2015) 133–157.
[42] J. Eck, R. Clarke, R. Guerette, Risky Facilities: Crime Concentration in Homogeneous Sets of Establishments and Facilities, Crime Prevent. Stud. 21 (2007) 225–264.
[43] G. Drawve, A. Wooditch, A research note on the methodological and theoretical considerations for assessing crime forecasting accuracy with the predictive accuracy index, J. Crim. Just 64 (2019), 101625.
[44] G. Rosser, et al., Predictive Crime Mapping: Arbitrary Grids or Street Networks? J. Quant. Criminol. 33 (3) (2017) 569–594.