Accident Analysis and Prevention 121 (2018) 223–230

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Traffic crash analysis with point-of-interest spatial clustering

Ruo Jia[a,b], Anish Khadka[b], Inhi Kim[c,*]

[a] School of Transportation, Southeast University, Southeast University, Si Pai Lou #2, Nanjing, 210096, China
[b] Southeast University-Monash University Joint Graduate School, Southeast University, Suzhou, 215123, China
[c] Monash Institute of Transport Studies, Department of Civil Engineering, Monash University, Clayton, Victoria, 3800, Australia

ARTICLE INFO

Keywords:
Spatial clustering
Kernel density estimation
Traffic crashes
Spatial regression

ABSTRACT

This paper presents a spatial clustering method for macro-level traffic crash analysis based on open source point-of-interest (POI) data. Traffic crashes are discrete and non-negative events for short-time evaluation but can be spatially correlated with long-term macro-level estimation. Thus, the method requires the evaluation of parameters that reflect spatial properties and correlation to identify the distribution of traffic crash frequency. A POI database from an open source website is used to describe the specific land use factors which spatially correlate to macro level traffic crash distribution. This paper proposes a method using kernel density estimation (KDE) with spatial clustering to evaluate POI data for land use features and estimates a simple regression model and two spatial regression models for Suzhou Industrial Park (SIP), China. The performance of spatial regression models proves that the spatial clustering method can explain the macro distribution of traffic crashes effectively using POI data. The results show that residential density, and bank and hospital POIs have significant positive impacts on traffic crashes, whereas, stores, restaurants, and entertainment venues are found to be irrelevant for traffic crashes, which indicate densely populated areas for public services may enhance traffic risks.

1. Introduction

In recent times, there has been a significant surge in the economic and human loss resulting from road accidents. The externalities associated with injuries and fatalities like medical costs produce additional financial burdens and even deprive those injured the ability to live a normal and productive life. It is estimated that about 1.25 million people die in road crashes each year (WHO, 2015). Moreover, it is the leading cause of death among those aged between 15 and 29 years, thus providing crucial insight about the loss of productivity in terms of the national interest. In the past, safety aspects were often neglected by policymakers when designing transportation infrastructure. However, safety now plays a key role in the planning and designing of road and public facilities infrastructures.

The increase in auto ownership as well as economic growth has resulted in a tremendous increase in travel demand over the years. It may be easily concluded that the increase in the number of vehicles on existing infrastructure is one of the major causes of accidents. However, it has been reported that 90% of the world's fatalities occur in under-developed and developing countries, which account for approximately 54% of the world's vehicles (WHO, 2016). This highlights the importance of analyzing the key parameters and indicators including spatial variations that contribute to road accidents in those countries. However, for under-developed countries, the main obstacle for accident detection and related factors evaluation are poor quality and inefficient data sources. The only assessment that has been performed is the accident hotspots analysis, which is determined in the zone or the segment of the road where the accident frequency is found to exceed certain threshold limits. These threshold limits differ from place to place with spatial difference while the categories of data available may not support the rationality of traffic crash assessment and response measures. The diversity in land use, travel behavior, road features, traffic volume and socio-demographic characteristics presents a broad scope for detailed accident studies. The difficulty in acquiring accurate and reliable data from the government and city level authorities hinders traffic spatial feature analysis. Thus, due to the unavailability of reliable data, studies on accident analysis have been limited especially in under-developing countries where the traffic problems are generally more severe than in developing countries.

However, with the help of open-source data, reliable point-of-interest (POI) data can be collected from anywhere in the world. Although they may not be the typical factors used in traditional traffic accident analysis, these POI data are specific data of land use factors with precise location information that are expected to be highly related to user characteristics and traffic crashes in macro and micro aspects. This paper focuses on a spatial clustering method for macro-level traffic

[*] Corresponding author.
E-mail addresses: ruo_jia@126.com (R. Jia), anishkhadka04@yahoo.com (A. Khadka), Inhi.kim@monashe.edu (I. Kim).

https://doi.org/10.1016/j.aap.2018.09.018
Received 11 May 2018; Received in revised form 17 September 2018; Accepted 18 September 2018
Available online 25 September 2018

crash analysis based on POI data to reflect spatial properties and correlations to identify the distribution of traffic crash frequency for areas where traditional traffic crash and traffic data are not reliable.

Our study aims to address the following two questions for POI based crash spatial analysis: (1) How can POI influencing factors for land use be quantified to evaluate traffic crashes? (2) How does the spatial regression model perform? Further, what is the relationship between correlated spatial characteristics of POI data and traffic crashes?

This paper is structured as follows. Section 2 presents a review of previous research on crash analysis and related measurement methods. Section 3 describes the difficulty in acquiring accurate and reliable data from the government in Suzhou and proposes an open-source method of acquiring POI data to identify land use factors. Section 4 focuses on the methodology used in the study from 3 aspects: 1) Using the kernel density estimation (KDE) method to transform the POI and crash data into density function for estimation; 2) Using the natural breaks clustering algorithm to identify and quantify the POI density with regards to attributing and comparing data; 3) Introducing the ordinary least squares regression (OLS) model, spatial error model (SEM) and the spatial lag model (SLM) to estimate the macro distribution of traffic crashes with POI data. Section 5 presents the regression results and analyzes the spatial characteristics and differences of the spatial regression models. Section 6 provides the conclusions and the recommendations for policies and policymakers.

2. Literature review

Various studies have been conducted on a macro (zonal) and micro (segments, intersections) level to examine the relationship between road accidents and various parameters like road features, land use and socio-demographic and environmental characteristics. Poisson regression models are broadly used in accident studies due to their ability to handle non-negative count data. However, their limitations to cater for over and under-dispersion in data due to the assumption of equal mean and variance led to the development of models such as the negative binomial (NB) and Poisson lognormal (PLN) models (Lord and Mannering, 2010). It has been found that the adjacent locations share common contributing risk factors and have significant influence on the crash (Fawcett et al., 2017). The NB and PLN models assume fixed parameters throughout the observations and thus ignore spatial correlations resulting in biased results (Barua et al., 2015; Li et al., 2013). Aguero-Valverde and Jovanis (2006) compared a full Bayes(FB) hierarchical model with the NB model and found that the inclusion of spatial correlations in FB models enhanced the prediction accuracy of the model. Generally, different models are compared using performance criterion, such as deviance information criterion (DIC) (Barua et al., 2015; Cai et al., 2017). Barua et al. (2015) concluded that 83.8% of the variability in their dataset is explained by spatial correlation which is vital to crash analysis. The spatial correlations also act as a proxy variable accounting for unobserved heterogeneity in the model. It is found that traditional crash analysis methods mostly depend on time-series data while the crash count distribution lacks spatial indicators and features.

Macro-level analysis is found to be suitable for conducting area-wide studies with an added advantage of less detailed data required compared to micro-level analysis (Fawcett et al., 2017). Moreover, as the macro-level crash analysis that consider spatial factors is found to be more reliable and meaningful to both researchers and policymakers, it has received more attention in recent years. It has been found that geographically weighted Poisson regression (GWPR) provides more spatial randomness compared to the generalized linear model (GLM), as it allows coefficients to vary spatially throughout the observations (Matkan et al., 2011, Hadayeghi et al., 2010). Prasannakumar et al. (2011) conducted both spatial (religious location and educational institution) and temporal studies (monsoon and non-monsoon periods) using GIS-based methods. With computational advancement, spatial conditional autoregressive and Bayesian spatial models have been used extensively in analyzing hotspots related to pedestrian accidents (Wang et al., 2016), injury severity levels (Barua et al., 2016) and vehicle crashes (Fawcett et al., 2017; Mitra, 2009) due to their ability to address spatial, temporal correlations and unobserved heterogeneity. Moran’s I is an index that is often used to verify the spatial association in a dataset (cluster or dispersion) (Cai et al., 2017; Mitra, 2009).

Macro-level spatial analysis is generally carried out on traffic analysis zones (TAZs) and uses a geometric matrix to reflect the spatial weight in TAZs. It has been found that geometric centroid-distance-order weight performs better compared to other spatial weight features (Wang et al., 2016). Researchers argue that the TAZs are relatively smaller in size resulting in increased movement of vehicles in and out of the zones. Therefore, the inclusion of unobserved factors beyond the zone is bound to have an adverse effect on the results (Montella, 2010). It is suggested that traffic analysis districts (TADs), formed by the aggregation of several TAZs, might be a better alternative to TAZs as movement is restricted inside the zone due to its relatively larger area. Many researchers have conducted crash hotspot identification using various methods such as crash frequency, crash rate, potential for improvement, empirical Bayes (EB) and KDE. These methods are compared using quantitative tests such as consistency test, rank difference test and false identification test (Montella, 2010; Cheng and Washington, 2005, 2008; Yu et al., 2014). Studies have reported that the EB method outperforms all other methods due to its consistency and reliability. The EB method is extensively used in before-after studies as it accounts for past crash records of the treated site as well as expected collision frequency of a similar reference site by using various collision prediction models (CPMs) (Hauer, 1992). In addition, by transferring the kernel density function into a form that is analogous to the form of the EB function, Yu et al. (2014) further proved that the KDE method can eventually be considered as a simplified version of the EB method in which crashes reported at neighboring spatial units are used as the reference population for estimating the EB-adjusted crashes. Theoretically, the KDE method may outperform the EB method when the neighboring spatial units provide more useful information about the expected crash frequency than a safety performance function. For TAZ-based spatial analysis, the spatial weighted matrix promises to provide reliable features that capture the relationship between neighboring spatial units. Research also has been conducted on pedestrian crashes using urban design and land use characteristics as a proxy variable for pedestrian activity (Harwood et al., 2008). Quistberg et al. (2015) reported that locations like restaurants and high-density employment and residential areas have higher collision rates. To the authors’ knowledge, no past research has used the location POI data with the KDE method to evaluate the crash spatial performance even though it is expected to be theoretically efficient.

This paper focuses on the macro-level traffic crash spatial analysis by undertaking a case study of Suzhou Industrial Park (SIP), China. Due to the difficulty in acquiring accurate and reliable data from the government and city level authorities regarding parameters like traffic volume, vehicle kilometers of travel (VKT), etc., the conditional autoregressive regression and Bayesian spatial models are deemed inappropriate. Therefore, this paper aims to distinguish the POI features from traffic analysis zones by hotspot estimation and evaluate their influence on traffic crashes using spatial regression models.

3. Data description

This study conducts a case study of the SIP district. Suzhou is ranked as a top 10 city by Gross Domestic Product (GDP) index in China. It has a land area of 278 km^2 and a population of around 803,000. It is located very close to Shanghai, the economic center of China, as shown in Fig. 1. SIP is one of the major industrial development zones in Suzhou which was established in 1995 by an agreement between China’s central government and the Singapore government. Therefore, the SIP land

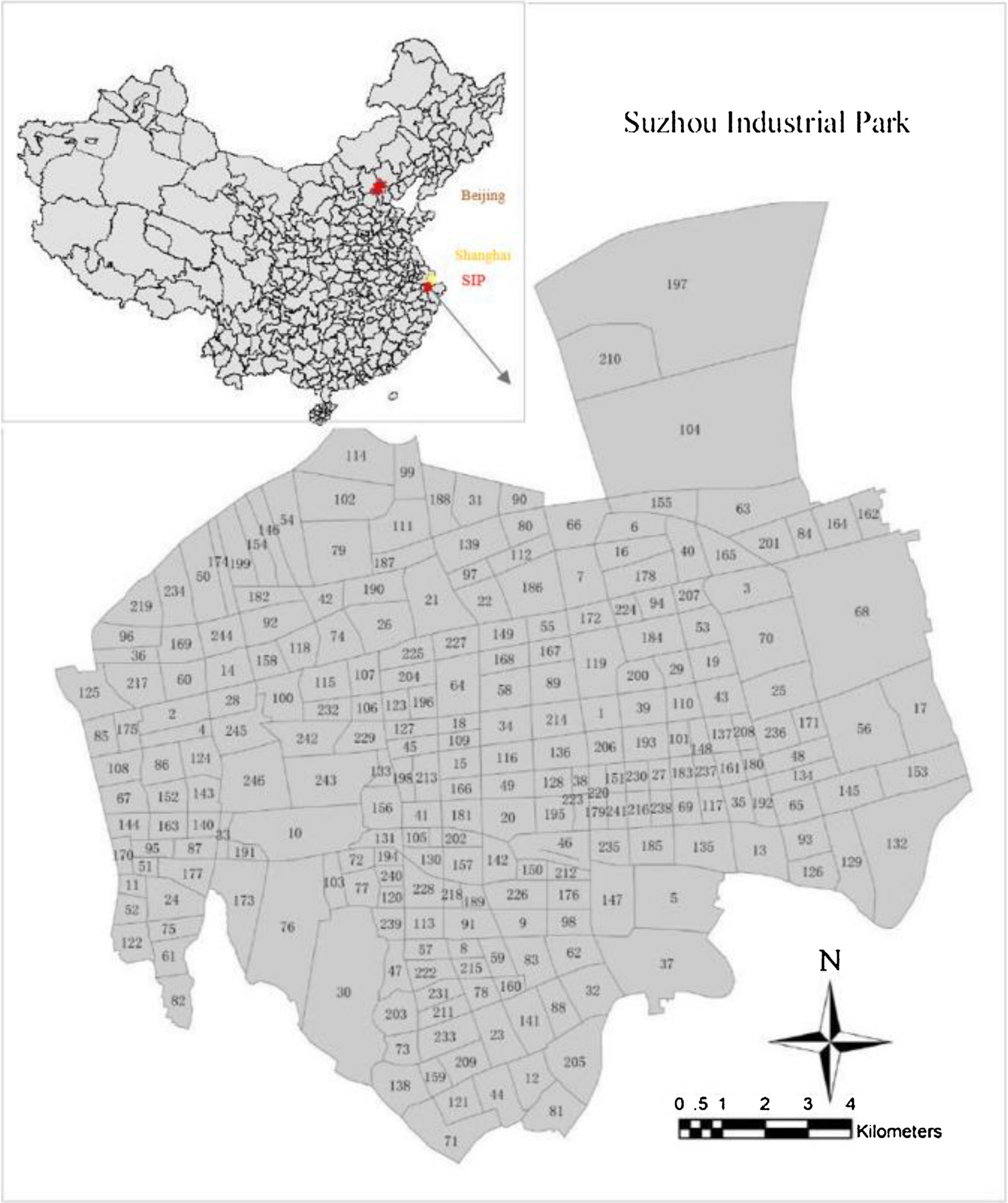


Fig. 1. The location of SIP and the TAZ division.

development differs from conventional Chinese districts by having a western influenced land use (neighborhood centers) rather than a road network pattern (vehicle-oriented), which could be a typical case in China.

This study aims to identify various spatial factors associated with traffic crashes in SIP which can contribute to the practical analytic process and further guide planning agencies in developing effective measures to reduce accidents. Therefore, to obtain meaningful data 246 SIP TAZs were used in this study as shown in Fig. 1. Twenty-four thousand crash records were collected from the Traffic Police Brigade for 2016. Considering the universality of the relationship between crashes and land use factors in a city from the macro perspective, the data used were no-injury and slight-injury vehicle crash events only. The analysis tool in this study is based on ArcGIS 10.2, GeoDa 1.8.16 beta (Nakaya, 2016) software and R software.

Due to unreliable traffic volume and unavailability of land use data, this paper focuses on the POI based spatial analysis and excludes vital parameters such as traffic volume and vehicle kilometers of travel (VKT). Instead, the traffic crash data is integrated with the POI data to identify varying land use and its impact on traffic safety. The POI data was obtained from Google Map Application Programming Interface (API) and Baidu (China's biggest search engine) Map API. It includes

Table 1
POI Data Description based on TAZs.

Variable	Min.	1 st Quartile.	Median	Mean	3rd Quartile	Max.
Number of crashes	0.00	13.00	43.00	96.72	109.00	859.00
Number of banks	0.00	0.00	1.00	4.31	3.00	127.00
Number of entertainment places	0.00	0.00	1.00	4.02	4.00	76.00
Number of hotels	0.00	0.00	0.00	1.98	2.00	31.00
Number of hospitals	0.00	0.00	0.00	2.05	2.00	33.00
Number of restaurants	0.00	0.00	3.00	25.74	19.75	485.00
Number of residences	0.00	1.00	3.00	5.70	7.00	51.00
Number of schools	0.00	0.00	2.00	6.10	6.00	126.00
Number of stores	0.00	1.00	5.00	38.63	25.00	945.00
Number of parks & scenic spots	0.00	0.00	0.00	0.71	1.00	28.00

nine categories of POI data such as schools, restaurants, hospitals, hotels, stores, banks, entertainment venues, residential zones, parks and scenic spots. These kinds of POI data are specific forms of land use with precise location information which are expected to be highly related to user characteristics and traffic crashes in macro and micro aspects. The POI data is aggregated to TAZs in SIP. Their description is shown in Table 1 below. It is noted that both the mean value and variance of each parameter differs significantly, showing that the data distribution is extremely unbalanced while the spatial characteristics of crashes and POI data are missing for TAZ aggregated counts.

4. Methodologies

In past studies, crash rate or frequency and land use have been considered as the response and explanatory variables, respectively. However, in this study, the data is aggregated by a count model and a probability distribution function is obtained for the POI features using KDE. The purpose of natural break clustering is to reclassify and find the best arrangement of the resulting POI density values from KDE. Finally, the geographically weighted average is verified, and a simple linear regression and two spatial models are used for statistical modeling and making inferences. In this section, the regression models were tested to estimate the properties and correlation of POI distribution with traffic crash frequency.

4.1. Kernel density estimation

KDE is considered the most promising method to describe the spatial patterns that exist in various parameters (Chainey and Ratcliffe, 2013). It is found to be superior and advantageous compared to the statistical hotspot and clustering techniques. By using the density method, an arbitrary spatial unit of analysis can be defined that is homogenous for the entire area, which makes the comparison and ultimately classification possible. KDE involves placing a symmetrical surface over each point and evaluating the distance from the point to a reference location based on a mathematical function and then summing the value for all the surfaces points for that reference location. It involves placing a kernel over each observation and summing these individual kernels to obtain the density estimate for the distribution of traffic crashes (Fotheringham et al., 2000). The equation is shown as follows:

$$f(x, y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{d_i}{h}\right) \tag{1}$$

where $f(x, y)$ is the density estimate at the location (x, y), n is the number of observations, h is the bandwidth or kernel size, K is the kernel function, and d_i is the distance between the location (x, y) and the location of the i th observation. The objective of placing these humps or kernels over a point is to create a smooth and continuous surface. The density estimate is obtained by summing these values at every observation within the bandwidth, including those at which no incidences of the indicator variable were recorded.

In KDE, bandwidth selection is considered one of the most crucial tasks as it can alter the results significantly. The selection of optimum bandwidth varies according to the data and study area. A large bandwidth leads to a smooth density pattern which makes it difficult to separate local hotspots, whereas, a small bandwidth leads to a sharp density pattern highlighting only individual hotspot locations (Shariatmohaymany et al., 2013). However, there is no hard and fast rule to obtain the optimum bandwidth; the selection is carried out using iterative techniques (trial and error method) (Xie and Yan, 2008; Shariatmohaymany et al., 2013). In this study, bandwidth in the range from 50 to 1000 m is tested. The optimum bandwidth is found to be 800 m as it resulted in a density pattern that is neither too sharp nor too smooth.

4.2. Natural break cluster

Based on the KDE results, the levels of density areas are divided into different classes. It must be noted that natural break cluster is used to determine the best arrangement of the density values obtained from KDE that result in clusters of density. To maximize the variance between classes and minimize the variance resulting from the cluster method, the natural break cluster method was applied in this study. It involves various iterative process steps where calculations are repeated using different breaks in the dataset to determine the set of breaks that result in the smallest-in-class variance while minimizing the squared deviation within each class. The process involves dividing the ordered data into groups. Initial group divisions can be arbitrary. The flowchart of the natural break cluster method is shown in the Fig. 2 below.

The following four steps are repeated until the required criteria are met:

Step 1. Calculate the sum of squared deviations between classes (SDBC).

Step 2. Calculate the sum of squared deviations from the array mean (SDAM).

Step 3. Subtract the SDBC from the SDAM (SDAM - SDBC), which equals the sum of the squared deviations from the class mean (SDCM).

Step 4. After inspecting each of the SDBC, a decision is made to move one unit from the class with the largest SDBC to the class with the lowest SDBC.

New class deviations are then calculated while the process is repeated until the sum of the within class deviations reaches a minimum value. Subsequently, all break combinations may be examined. SDCM is calculated for each combination while the combination with the lowest SDCM is selected. Since all break combinations are examined, this process guarantees that the one with the lowest SDCM is obtained.

4.3. Spatial autocorrelation

Spatial autocorrelation occurs when events occurring at different but nearby locations are correlated with each other. This phenomenon is quite likely to be observed for traffic crashes that occur within a city. Spatially autocorrelated data should not be analyzed by normal regression analysis as the correlation violates the basic assumption of ordinary least squares (OLS) regression. Straightforward spatial regression analysis uses a spatial weight matrix and maximum likelihood estimation to minimize the possible bias resulting from spatially autocorrelated data (Griffith, 1988; LeSage, 1998, 1999; LeSage and Pace, 2009; Rhee et al., 2016). Transportation and planning agencies generally use OLS regression to make investigations and inferences on the

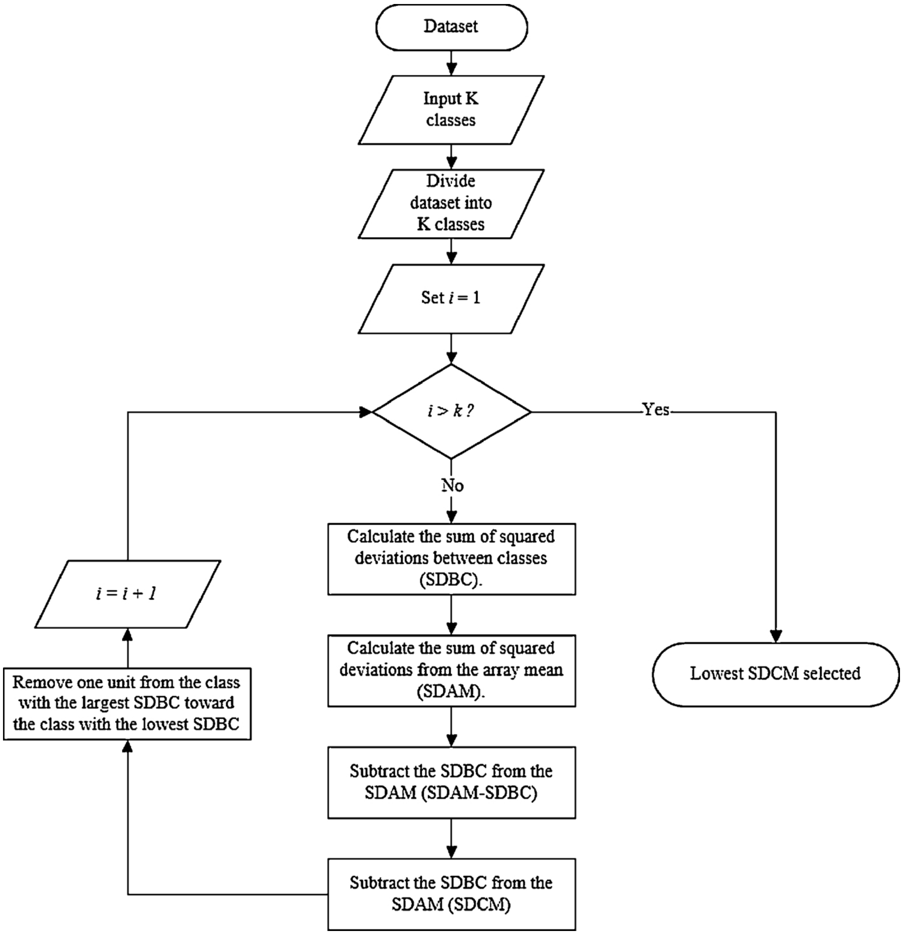

Dataset
Input K classes
Divide dataset into K classes
Set i = 1
i > k ?
Yes
No
Calculate the sum of squared deviations between classes (SDBC).
Calculate the sum of squared deviations from the array mean (SDAM).
Subtract the SDBC from the SDAM (SDAM-SDBC)
Subtract the SDBC from the SDAM (SDCM)
Remove one unit from the class with the largest SDBC toward the class with the lowest SDBC
i = i + 1
Lowest SDCM selected

Fig. 2. Natural break cluster flow chart.

number of crashes or the crash rate. Crashes are discrete and non-negative integer values so it is more appropriate to use a count model for traffic accident studies. However, when analyzing large areas on an annual basis, with dozens of crashes, the practical benefits of count models become less and even negligible.

Spatial models were introduced to address the spatial dependence existing in the data. The form of the general spatial model is (Griffith, 1988):

$$Y = \rho W_1 Y + X\beta + \mu \tag{2}$$

$$\mu = \lambda W_2 \mu + \varepsilon \text{ where } \varepsilon \sim MVN(0, \sigma^2 I_n), \tag{3}$$

where Y, the dependent variable, $(n \times 1)$ is a vector of the natural logarithm of crashes in the n TAZs in a year, X is a $(n \times n)$ matrix of k explanatory variables, ρ and λ are spatial autoregressive coefficients, W_1 and W_2 are $(n \times n)$ spatial weight matrices, μ is the unobserved error term that incorporates spatial correlation through its first term, ε is a $(n \times 1)$ vector of unobserved error terms that are identically and independently distributed, MVN denotes the multivariate normal distribution, and I_n is the $(n \times n)$ identity matrix. In Eq. (3), the model becomes a spatial lag model (SLM) when W_2 is 0 or a spatial error model (SEM) when W_1 is 0.

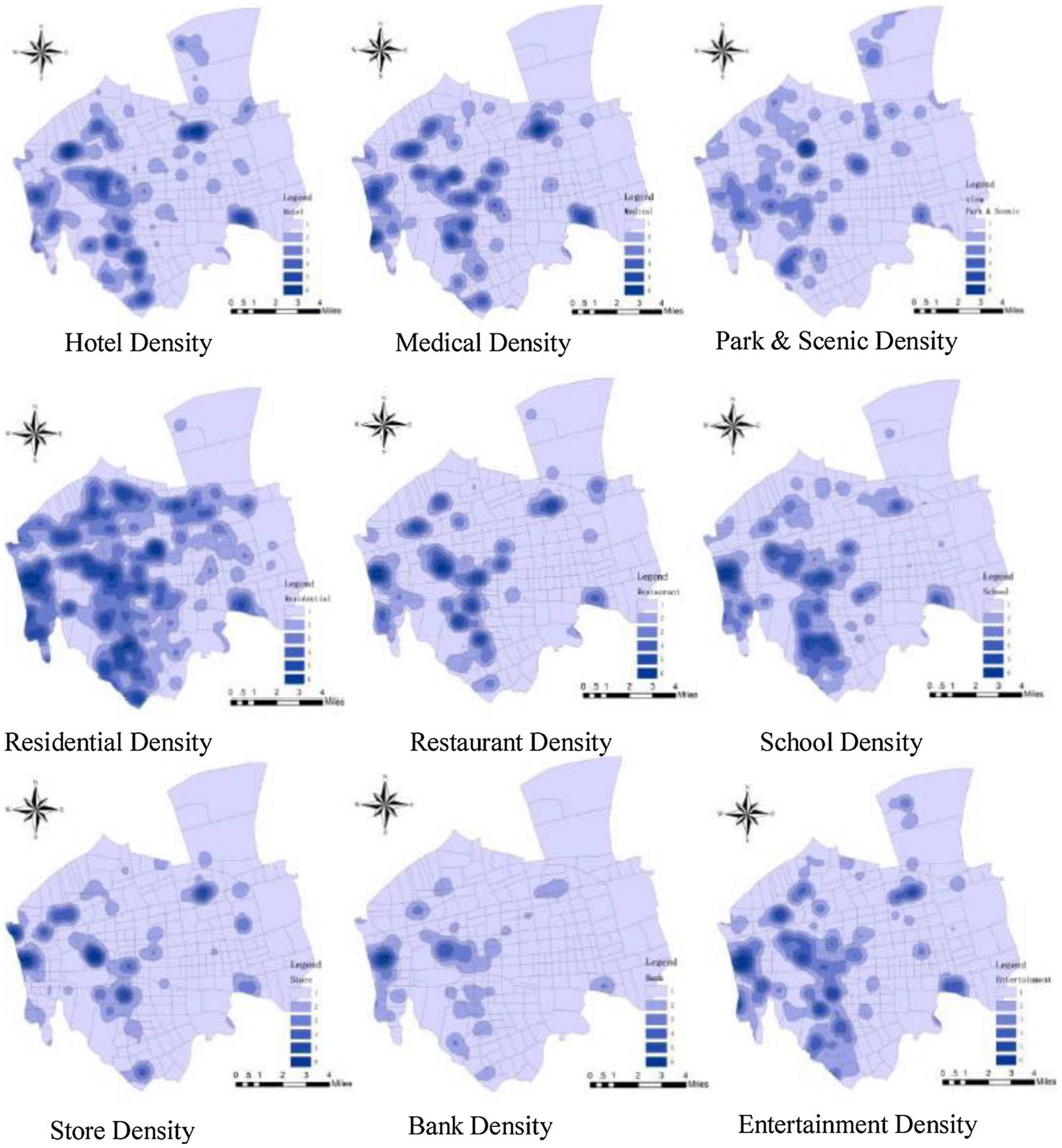
In the SLM, a dependent variable in a region is subject to a spill-over effect from the dependent variable in the neighboring regions. This effect is accounted for by the spatial weight matrix, (W_1). Similarly, in the SEM model, the error in one region is dependent on the error in neighboring regions through W_2. Here, the spatial weight matrices W_1 and W_2 are defined using Rook's method of the spatial contiguity matrix proposed by Griffith (1988). Rook's method defines neighbors such that if a portion of the boundary between two regions is shared, the corresponding element of the spatial weight matrix, W_{ij} is 1 and 0 otherwise.

5. Results

5.1. Kernel density estimation cluster result

The natural break cluster-based KDE method was applied to quantify the POI influencing factors of land use to evaluate traffic crashes in SIP. As mentioned in the above sections, the initial selection of the k-value in a natural break cluster is carried out arbitrarily. Then, an iterative process as shown in Fig. 2 is carried out until the criteria of lowest SDCM is met. In this study, the k-value in a natural break cluster is selected as 6 to ensure that the dense and sparse areas are clearly divided into different classes. The POI KDE with natural break cluster reclassification results are presented in Fig. 3 below. The hotspots coded in deep blue represent high concentrations and are observed in areas that are densely populated. Different POI data show different features which reflect the spatial diversity of POI and land use characteristics for non-hot areas. It is noted that the city scale and land use spatial features can be clearly distinguished by the POI hotspot estimation method.

Based on density clustering, the POI count data are transformed into the cluster level of kernel density by aggregating the TAZ count with the weighted average method. The traffic crash data also have been converted into a clustering dataset by following the previous steps. The clustering attributes of TAZs are calculated and presented in Table 2. In this step, the unbalanced distribution is converted to the clustering result by minimizing the squared deviation within each class. The clustering results were found to perform better for spatial regression



Hotel Density

Medical Density

Park & Scenic Density

Residential Density

Restaurant Density

School Density

Store Density

Bank Density

Entertainment Density

Fig. 3. POI hotspots feature used in SIP case.

Table 2
Clustering Data Description based on TAZs.

Variable	Min.	1 st Quartile	Median	Mean	3rd Quartile	Max.
Crash Cluster	1.00	1.27	1.79	2.07	2.64	5.58
Bank Density	1.00	1.72	2.39	2.67	3.48	6.00
Entertainment Venue Density	1.00	1.00	1.22	1.59	1.96	5.26
Hotel Density	1.00	1.00	1.21	1.59	1.99	4.59
Hospital Density	1.00	1.00	1.25	1.60	1.92	5.23
Restaurant Density	1.00	1.00	1.09	1.50	1.82	5.00
Residential Density	1.00	1.45	2.14	2.35	3.03	5.56
School Density	1.00	1.00	1.22	1.61	2.00	5.28
Store Density	1.00	1.00	1.04	1.34	1.40	5.02
Park & Scenic Spot Density	1.00	1.00	1.15	1.38	1.58	4.30

analysis as the land use information in neighboring spatial units has been converted to a uniform clustering standard for each TAZ.

5.2. Spatial regression results

Based on the quantified clustering and non-clustering count data, the dataset was evaluated using a simple, ordinary regression model and two spatial models. The macro-level spatial analysis results of the OLS, SEM, and SLM regressions are presented in Table 3. It is noted that the standard deviation errors of the dataset have been eliminated because of the clustering pre-classification. The spatial lag correlation coefficient of the clustering dataset in SLM and SEM is 0.498 and 0.594, respectively, which is much larger than the non-clustering count (0.054 and 0.019, respectively). It shows that the dependent variable in neighboring locations influences the crash frequency significantly. It was also verified that it is vital to consider the spatially influenced crash analysis on a macro level. The Log-likelihood and Akaike information criterion which indicate the goodness of fit of the models is lower in the case of the clustering data approach compared to the non-clustering

Table 3
Model performance.

Dependent Variable	Clustering			Non-clustering Count		
Models	OLS	SLM	SEM	OLS	SLM	SEM
Number of Observations	246					
Mean dependent variable	2.706			96.72		
S.D. dependent variable	1.451			141.124		
Degrees of Freedom	236	235	236	236	235	236
Lag coeff. (Rho)	–	0.498	0.594	–	0.054	0.019
Sum squared residual	360.039	–	–	3,928,230	–	–
Log-likelihood	−395.9	−361.1	−361.3	−1539.5	−1539.2	−1539.5
Akaike information criterion	811.8	744.2	742.6	3099	3100.4	3098.9

Table 4
Regression results based on clustering data.

Model	Variable	Coefficient	Std.Error	t-Stat	Probability
OLS	Constant	0.644 **	0.297	2.167 **	0.031 **
	Bank Density	0.341 ***	0.107	3.177 ***	0.002 ***
	Entertainment Venue Density	0.235	0.317	0.742	0.459
	Hotel Density	0.003	0.203	0.016	0.987
	Hospital Density	0.381 *	0.212	1.800 *	0.073 *
	Restaurant Density	−0.088	0.316	−0.279	0.78
	Residential Density	0.288 **	0.142	2.030 **	0.048 **
	School Density	−0.227	0.236	−0.961	0.337
	Store Density	−0.087	0.225	−0.386	0.7
	Park & Scenic Spot Density	0.208	0.175	1.191	0.235
SEM	W_{ij}	0.498 ***	0.049	10.185 ***	0.000 ***
	Constant	0.215	0.248	0.868	0.385
	Bank Density	0.191 **	0.088	2.180 **	0.029 **
	Entertainment Venue Density	0.009	0.257	0.035	0.972
	Hotel Density	0.036	0.164	0.222	0.825
	Hospital Density	0.355 **	0.171	2.071 **	0.038 **
	Restaurant Density	−0.104	0.256	−0.407	0.684
	Residential Density	0.127	0.115	1.105	0.269
	School Density	−0.105	0.191	−0.547	0.584
	Store Density	−0.092	0.182	−0.506	0.613
	Park & Scenic Spot Density	0.129	0.141	0.911	0.362
SLM	Constant	1.400 ***	0.345	4.058 ***	0.000 ***
	Bank Density	0.241 **	0.11	2.192 **	0.028 **
	Entertainment Venue Density	0.158	0.276	0.571	0.568
	Hotel Density	0.244	0.184	1.33	0.184
	Hospital Density	0.282	0.186	1.519	0.129
	Restaurant Density	−0.288	0.289	−0.995	0.32
	Residential Density	0.052	0.127	0.407	0.684
	School Density	−0.024	0.218	−0.11	0.912
	Store Density	−0.156	0.213	−0.735	0.462
	Park & Scenic Spot Density	0.074	0.15	0.496	0.62
	LAMBDA	0.594 ***	0.048	12.376 ***	0.000 ***

*** Represents the significance levels of 1%. ** Represents the significance levels of 5%. * Represents the significance levels of 10%.

count approach. Thus, the clustering-based spatial regression analysis improves the effectiveness of the models.

The performance results of clustering data for the three models are presented in Table 4. In the OLS model, as shown in Table 4, bank density, hospital density, and residential density were found to have a strong positive correlation (p-value = 0.002, 0.073, 0.048 and t-stat = 3.177, 1.800, 2.030) with traffic crashes. It is an interesting observation that a unit increase in bank cluster increases the traffic crash cluster level by 0.34. One reason to explain this might be that the traffic situation around banks is dominated by temporary parking causing distractions to both drivers and pedestrians. The risk is increased because while banks attract many customers and employees daily, the demand is stochastic (random). Furthermore, the uncertainty in waiting time along with the risk associated with large sums of money transactions enhances anxiety. A similar trend of traffic crashes can be noted for hospitals and residential densities. However, unlike most studies, school density is found to have a negative correlation with traffic crashes. The probable reason might be because most vehicles tend to reduce their speed due to school zone signage and low speed limits. Moreover, the section of roadway near the main entrance of schools is crowded with the people and cars waiting for children, making drivers more cautious. The parameters associated with stores, restaurants and entertainment venues were found to be irrelevant (p-value > 0.1, t-stat < 2) for traffic crashes.

The spatial correlation violates the basic assumption of OLS models. SLM and SEM consider the spatial matrix to minimize the possible bias resulting from spatially autocorrelated data as shown in Table 4. The regression results of clustering data show that only banks and hospitals have a spatial influence on traffic crashes (p-value = 0.029, 0.038, and t-stat = 2.180, 2.071 in SLM and p-value = 0.028, and t-stat = 2.192 for bank density in SEM). SEM performs the best in this case. The SLM shows that banks and hospitals have a positive correlation with traffic crashes because public services attract people, which enhances their exposure and with it crash risk.

It is noted that more sophisticated count modeling frameworks using Bayesian estimation have received considerable popularity in these research areas. Such models may soon become more practical, but they require a large set of data over a long period of time. The Bayesian estimation methods require posterior and prior distributions which are sometimes impractical. The models presented in this paper are already practice-ready and perform well using typical data and software available to cities and planning agencies.

6. Conclusions

This study proposed a natural break cluster method to reclassify the kernel density and a spatial analysis model to estimate the correlation of POI data with traffic crashes. By using a simple, ordinary model and two spatial models, it has shown that the performance of KDE with natural break cluster is much better than the traditional count method. A comparison of the performance of spatial regression models for POI density cluster found different models reflect different features in the SIP case but all confirm that crash data is spatially correlated. It is therefore recommended to use the spatial error model for such applications.

Three density cluster parameters were found to be positively correlated with traffic crashes namely, hospitals, banks, and residential areas in the OLS model. The common attributes of these three parameters have greatly increased points of conflict between vehicles, bicycles, and pedestrians. The results suggest that areas with mixed-land use require additional local-level research to develop effective target-oriented treatments to improve safety. The school density cluster had a negative effect on accident risk. Considering the differences between spatial areas in three parameters, banks and hospital are positively correlated in SEM, and only banks are positively correlated in SLM.

This study has important and direct implications for transport policies. This work has analyzed traffic crashes at the TAZ level, which is the spatial unit used in transportation planning. Also, the data collection method shows strong adaptability and repeatability, and is an effective way to handle crash analysis problems of poor quality and inefficient data. Moreover, the POI data used in this paper was limited to nine categories, which may be increased in future studies. The approach using POI density to conduct spatial regression and find correlation with traffic crashes is very meaningful. The application of spatial analysis on traffic safety is still in its early stages at planning agencies. It will be more actively applied with further developments in data collection, computer technology and implementation of new models. As methods advance, it will be helpful to achieve institutional standardization of the unit of analysis to harmonize traffic safety analysis and transportation planning. Further study is required to identify the optimal aggregation unit for analyzing the characteristics of traffic crashes in SIP and elsewhere, and how this unit can be integrated with transportation planning.

Acknowledgment

This work was supported by Jiangsu Industrial Technology Research Institute.

References

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38 (3), 618.

Barua, S., El-Basyouny, K., Islam, M.T., 2015. Effects of spatial correlation in random parameters collision count-data models. Anal. Methods Accid. Res. 5–6, 28–42.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Methods Accid. Res. 9, 1–15.

Cai, Q., Abdelaty, M., Lee, J., Eluru, N., 2017. Comparative analysis of zonal systems for macro-level crash modeling. J. Safety Res.

Chainey, S., Ratcliffe, J., 2013. GIS and Crime Mapping.

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prev. 37 (5), 870–881.

Cheng, W., Washington, S., 2008. New criteria for evaluating methods of identifying hot spots. Transp. Res. Rec. J. Transp. Res. Board 2083 (2083), 76–85.

Fawcett, L., Thorpe, N., Matthews, J., Kremer, K., 2017. A novel Bayesian hierarchical model for road safety hotspot prediction. Accid. Anal. Prev. 99 (Pt A), 262–271.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. In: In: Isaaks, E., Mohan Srivastava, R. (Eds.), Quantitative Geography : Perspectives on Spatial Data Analysis, vol. 50. Sage Publications, pp. 143–163 1 ppXI - XII.

Griffith, D.A., 1988. Spatial econometrics: methods and models. Econ. Geogr. 65 (2), 160.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using geographically Weighted Poisson Regression. Accid. Anal. Prev. 42 (2), 676–688.

Harwood, D.W., Bauer, K.M., Richard, K.R., et al., 2008. Pedestrian safety prediction methodology. Nchrp Web Document.

Hauer, E., 1992. Empirical bayes approach to the estimation of “unsafety”: the multivariate regression method. Accid. Anal. Prev. 24 (5), 457–477.

LeSage, J.P., 1998. Spatial Econometrics. University of Toledo.

LeSage, J.P., 1999. The Theory and Practice of Spatial Econometrics. University of Toledo.

LeSage, J.P., Pace, R.K., 2009. Introduction to Spatial Econometrics. Chapman & Hall/CRC, New York.

Li, Z., Wang, W., Liu, P., Bigham, J.M., Ragland, D.R., 2013. Using geographically weighted poisson regression for county-level crash modeling in California. Saf. Sci. 58 (10), 89–97.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44 (5), 291–305.

Matkan, A.A., Mohaymany, A.S., Mirbagheri, B., Shahri, M., 2011. Explorative Spatial Analysis of Traffic Accidents Using GWPR Model for Urban Safety Planning. In.

Mitra, S., 2009. Spatial autocorrelation and bayesian spatial statistical method for analyzing intersections prone to injury crashes. Transp. Res. Rec. J. Transp. Res. Board 2136 (2136), 92–100.

Mohaymany, A., Kashani, A.T., Nosrati, H., et al., 2013. Development of head-on conflict model for overtaking maneuvers on two-lane rural roads using inductive loop detectors. J. Transp. Saf. Secur. 5 (4), 273–284.

Montella, A., 2010. A comparative analysis of hotspot identification methods. Accid. Anal. Prev. 42 (2), 571–581.

Nakaya, T., 2016. GWR4 User Manual.

Prasannakumar, V., Vijith, H., Charutha, R., Geetha, N., 2011. Spatio-temporal clustering of road accidents: GIS based analysis and assessment. Procedia - Soc. Behav. Sci. 21 (2), 317–325.

Quistberg, D.A., Howard, E.J., Ebel, B.E., Moudon, A.V., Saelens, B.E., Hurvitz, P.M., Curtin, J.E., Rivara, F.P., 2015. Multilevel models for evaluating the risk of pedestrian–motor vehicle collisions at intersections and mid-blocks. Accid. Anal. Prev. 84, 99–111.

Rhee, K.A., Kim, J.K., Lee, Y.I., et al., 2016. Spatial regression analysis of traffic crashes in Seoul. Accid. Anal. Prev. 91, 190–199.

Wang, X., Yang, J., Lee, C., Ji, Z., You, S., 2016. Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accid. Anal. Prev. 96, 12–21.

WHO, 2015. Road Traffic Injuries. World Health Organization.

Xie, Z., Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Comput. Environ. Urban Syst. 32 (5), 396–406.

Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prev. 66 (3), 80–88.