Accident Analysis and Prevention 120 (2018) 281–294

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Understanding the effects of trip patterns on spatially aggregated crashes with large-scale taxi GPS data

Jie Bao[a,b], Pan Liu[a,b,*], Xiao Qin[c], Huaguo Zhou[d]

[a] Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing, 210096, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou #2, Nanjing, 210096, China
[c] Department of Civil and Environmental Engineering, University of Wisconsin-Milwaukee, NWQ4414, P.O. Box 784, Milwaukee, WI 53201, United States
[d] Department of Civil Engineering, Auburn University, 238 Harbert Engineering Center, Auburn, AL 36849-5337, United States

ARTICLE INFO

Keywords:
Big data
Trip pattern
Taxi GPS data
Spatial analysis
Crashes

ABSTRACT

The primary objective of this study was to investigate how trip pattern variables extracted from large-scale taxi GPS data contribute to the spatially aggregated crashes in urban areas. The following five types of data were collected: crash data, large-scale taxi GPS data, road network attributes, land use features and social-demographic data. A data-driven modeling approach based on Latent Dirichlet Allocation (LDA) was proposed for discovering hidden trip patterns from a taxi GPS dataset, and a total of fifty trip patterns were identified. The collected data and the identified trip patterns were further aggregated into167 ZIP Code Tabulation Areas (ZCTA). Random forest technique was used to identify the factors that contributed to total, PDO and fatal-plus-injury crashes in the selected ZCTAs during the study period. Geographically weighted Poisson regression (GWPR) models were then developed to establish a relationship between the crashes and the contributing factors selected by the random forest technique. Comparative analyses were conducted to compare the performance of the GWPR models that considered traditional traffic exposure variables only, trip pattern variables only, and both traditional exposure and trip pattern variables. The model specification results suggest that the trip pattern variables significantly affected the crash counts in the selected ZCTAs, and the models that considered both the traditional traffic exposure and the trip pattern variables had the best goodness-of-fit in terms of the lowest MAD and AICc values.

1. Introduction

During the past decade, increased attention has been given to understanding the spatial pattern of crashes (Noland, 2003; Hadayeghi et al., 2006; Lovegrove et al., 2008; Huang et al., 2010; Li et al., 2013; Rhee et al., 2016; Siddiqui et al., 2012; Lee et al., 2014). Researchers have proposed numerous methods for analyzing crash data at various spatially aggregated levels, such as states (Noland, 2003), counties (Huang et al., 2010; Li et al., 2013), traffic analysis zones (TAZ) (Rhee et al., 2016; Siddiqui et al., 2012), and ZIP code tabulation areas (ZCTA) (Lee et al., 2014), etc. The spatial analysis of crashes has become more and more prevalent because researchers have come to believe that traffic safety is an essential component of urban transportation planning (FHWA, 2005; NCHRP, 2010). Accordingly, research is needed to establish a relationship between crashes in a particular geographic area and zone-level contributing factors such as land use, socio-demographic, and road network characteristics. In addition, with a better understanding of the spatial pattern of crashes, transportation professionals can identify the areas with greater-than-expected crashes, and apply proactive countermeasures to the areas with higher crash risks to enhance safety more efficiently.

In past, numerous studies have investigated the factors that contribute to crashes at spatially aggregated levels. The influence factors considered by previous studies fall under three general categories: traffic exposures (Tarko et al., 1996; Aguero-Valverde and Jovanis, 2006), road network attributes (Rhee et al., 2016; Siddiqui et al., 2012), and socio-demographic characteristics (Rhee et al., 2016; Lee et al., 2014). Traffic exposures are probably the most important factors for modeling spatially aggregated crash data. In the traditional crash frequency models that focus on roadway sections or intersections, annual average daily traffic (AADT) and total entering volume (TEV) have been widely used for measuring traffic exposures (Yu et al., 2014; Lee et al., 2017). However, when modeling spatially aggregated crash data, the focus of crash models is the entire road network in a particular

* Corresponding author at: Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing, 210096, China.
E-mail addresses: baojie@seu.edu.cn (J. Bao), pan_liu@hotmail.com (P. Liu), qinx@uwm.edu (X. Qin), hhz0001@auburn.edu (H. Zhou).

https://doi.org/10.1016/j.aap.2018.08.014
Received 24 August 2017; Received in revised form 26 July 2018; Accepted 14 August 2018
Available online 01 September 2018
0001-4575/

geographic area. In this condition, defining traffic exposures is not easy, and data collection is even more difficult. Theoretically, the total number of trips and trip purposes directly affect crash counts. Previous studies have used the estimated total number of trips in a TAZ for measuring traffic exposures when modeling spatially aggregated crash data (Siddiqui et al., 2012; Naderan and Shahi, 2010). The total number of trips were estimated with trip generation models and household travel survey data, which usually suffer from low-quality data problems, and hence large errors.

Recently, the rapid rise and prevalence of mobile technologies have enabled the collection of a large amount of data associated with human activities, resulting in a surge of studies on human mobility (González et al., 2008; Hasan et al., 2013; Bao et al., 2017). Transportation systems can greatly benefit from big data in the areas such as traffic flow prediction and travel demand estimation. Theoretically, big data also has potential to be incorporated in traffic safety studies to help transportation professionals better understand the mechanism and contributing factors of crashes.

In a recent study, the authors of the paper investigated how to incorporate the human activity information obtained from social media data in the spatial analysis of crashes (Bao et al., 2017). More specifically, we classified human activities into seven categories by the venue type information extracted from Twitter check-in data, and developed geographically weighted regression (GWR) models to establish a relationship between the crash counts reported in a TAZ and various contributing factors. The results suggested that human activity variables significantly affected the crash counts in a TAZ.

One of the limitations of our previous study is that social media users may come from specific groups, and may not, therefore, be representative of the whole population (Bao et al., 2017; Chen and Schintler, 2015). Accordingly, using social media data for safety analyses could produce biased results. To address this concern, additional research is needed to employ other data sources to account for the biases of social media data, and to generate a better understanding of trip patterns. In fact, recent studies have started using large-scale taxi GPS data for understanding the trip patterns in urban areas (Liu et al., 2015; Tang et al., 2015). Compared with social media data, taxis GPS dataset has a larger sample size, and covers more age groups of travelers (Chen et al., 2014). In addition, unlike household travel survey data, taxi GPS data are publicly available in many cities, providing researchers with great convenience and opportunities.

The primary objective of this study was to investigate how the trip pattern variables extracted from large-scale taxi GPS data contribute to the spatially aggregated crashes in urban areas. More specifically, this paper sought answers to the following questions: (*a*) how to discover hidden trip patterns from large-scale taxi GPS data; and (*b*) how trip pattern variables affect the number of property-damage-only (PDO) and fatal-plus-injury crashes at spatially aggregated levels.

2. Data sources

Data were collected from the City of New York in the United States, and the study period was from January 1 st to December 31th, 2015. The study area included Manhattan, the Bronx, Brooklyn and Queens, covering the majority of the metropolitan area of New York. The authors excluded Staten Island from consideration because this area had very few taxi trip observations. In the present study, the ZIP Code Tabulation Area (ZCTA) was considered the basic unit of analysis. ZCTAs are built from census blocks that are aggregated based on common postal addresses assigned to streets. Previous studies have suggested that ZCTA is a reasonable zoning scale for spatial analysis of crashes and human activities (Lee et al., 2014; Qian and Ukkusuri, 2015). The final dataset included 167 ZCTAs in the City of New York, and the boundaries of the selected ZCTAs are depicted in Fig. 1.

The following five types of data were collected: crash data, taxi trip data, road network attributes, land use features and social-demographic data. The crash data were collected from the New York City Police Department (NYPD). The information obtained from the crash data included the date, time, severity, collision type, and geo-location of each crash. A total of 173,606 crashes, including 142,849 PDO and 30,757 fatal-plus-injury crashes, were reported during the selected time period in the study area. Fig. 1 also depicts the distribution of crashes across different ZCTAs.

The taxi GPS data were collected from the New York City Taxi & Limousine Commission (NYCTLC). The taxicabs of New York have two varieties: yellow and green. The taxis painted yellow can pick up passengers anywhere in the City of New York, while the taxis painted green are allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island. To ensure that the taxi GPS data fully covered the whole study area, we collected the GPS data for both yellow and green taxis.

For each taxi trip the following information was extracted from the taxi GPS dataset: pick-up timestamp, pick-up geo-location, drop-off timestamp, drop-off geo-location, trip distance, and the payment information. More specifically, we followed the following three steps to extract trip information from the taxi GPS dataset. First, the taxi trips with pick-up and drop-off points within the study area were selected. Second, unreasonable trips, which mainly arose due to the failure of taxi meters, were removed. Note that a trip was considered unreasonable if: (*a*) the travel distance was zero; (*b*) the fare was less than the starting price (2.5 dollars); (*c*) the duration was less than one-minute, or (*d*) the average speed was more than 80 miles per hour. Finally, only the trip records with both pick-up and drop-off timestamps were considered for further analyses because the timestamp information is critical for exploring trip patterns. The unreasonable trips account for nearly 4.45% of the total observations. By removing the unreasonable trips, the final dataset, which consisted of 156,079,000 taxi trips recorded during the selected time period in the study area, was created.

The road-network-attribute data were collected from the New York City Department of Transportation (NYCDOT) and the TIGER files of U.S. Census Bureau. ArcGIS shape files depicting the road network attributes were obtained, and the information provided by the ArcGIS shape files included the length, road type and the posted speed limit of each road segment. Traffic volume data were collected from the New York State Department of Transportation (NYSDOT). Note that the NYSDOT only provided the average annual daily traffic (AADT) on freeways and major arterials. The daily vehicle kilometers traveled (DVKT) on freeways and major arterials were then computed for each ZCTA on the basis of the AADT and the road network attributes. More specifically, we split the freeways and major arterials by the boundaries of the selected ZCTAs with the spatial tools provided by ArcGIS and calculated the length of each segment of the freeways and major arterials in each ZCTA. The DVKT was then calculated by summarizing the products of road lengths and the AADT for different road segments.

The land use data were collected from the New York City Department of City Planning (NYCDCP). The land use falls under six categories: Residential (R), Commercial (C), Industrial & Manufacturing (I), Transportation (T), Public Institutions (P), and Open Space & Outdoor Recreation (O). For each taxi trip, the land use features were assigned to both pick-up and drop-off locations. More specifically, the pick-up and drop-off locations of each taxi trip were assigned with the corresponding census tracts. The land use feature of each pick-up/drop-off location was then determined by the dominant land use feature of the corresponding census tract. The social-demographic data were obtained from the U.S. Census Bureau. The obtained information included the number of people segregated by age cohorts, poverty level, the population with a bachelor's degree, median household income, unemployment population, and the average travel time to work. Finally, the crash data, the road network attributes, and the social-demographic data were aggregated into corresponding ZCTAs with the software PostGIS. The descriptive statistics of the variables are summarized in

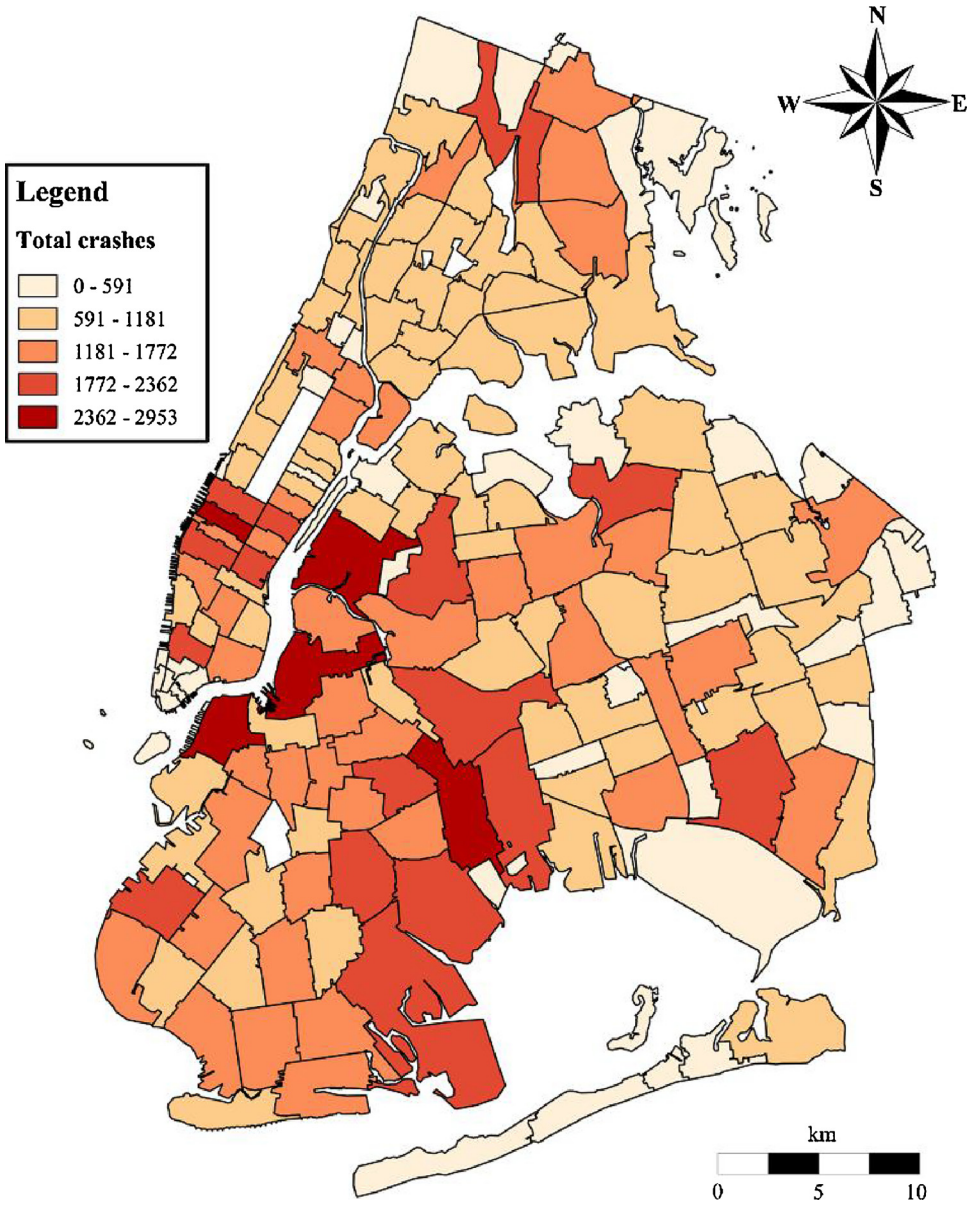


Fig. 1. The spatial distribution of crashes across different ZCTAs in the study area.

Table 1.

3. Methodology

Latent Dirichlet Allocation (LDA) method was proposed for discovering the hidden trip patterns from the large-scale taxi GPS dataset. Random forest technique was then used to identify the factors that contribute to both the PDO and the fatal-plus-injury crashes reported at the selected ZCTAs during the study period. The Moran's I statistic was used to verify whether the spatial correlation existed among explanatory variables. Geographically Weighted Poisson Regression (GWPR) models were then developed to establish a relationship between crash counts and the contributing factors selected by the random forest technique. The methods used in the present study are briefly discussed in this section.

3.1. Latent dirichlet allocation (LDA)

The LDA method was initially proposed to discover hidden topics from large corpus of documents (Blei et al., 2003). Recent studies have started applying the LDA method to human activity classification (Hasan and Ukkusuri, 2014), traffic incident duration prediction (Pereira et al., 2013) and dynamic origin-destination analysis (Côme et al., 2014). The LDA method has several advantages. First, many supervised learning methods, such as support vector machine, Bayes classifier, and decision tree classifier, need to predefine the number of trip patterns. This requirement limits the possible trip patterns that can be identified. By contrast, the LDA is a data-driven approach that allows possible trip patterns to be generated based on data. Second, compared with traditional unsupervised learning methods, the LDA assigns a probability value to each item in a trip pattern, providing a probabilistic explanation of the identified trip patterns. Third, the LDA provides a generating mechanism to discover hidden trip patterns. The generating mechanism can be further extended to simulate individual trips based on input data. Finally, the LDA accounts for the problem of missing values and outliers in large-scale datasets (Hasan and Ukkusuri, 2014).

The basic unit of analysis of the LDA method is a "word", and the inputs are all categorical variables. In our case, each taxi trip record was attached with a land use feature, and was further transformed into

Table 1
Descriptive statistics of variables.

Variable	Description	Min	Max	Mean	S.D.
Crash related variables					
Total crash	Total number of Crashes in each ZCTA	0	2953	1039.56	611.25
PDO crash	PDO crash count in each ZCTA	0	2309	855.38	509.87
Fatal-plus-injury crash	Fatal-plus-injury crash count in each ZCTA	0	644	184.17	115.64
Road and traffic related variables					
DVKT	Daily vehicle kilometers traveled in each ZCTA (10^6 vehicle. km)	0	1.573	0.398	0.294
Freeway density	Freeway length / Area of ZCTA (km/km^2)	0	6.665	0.764	1.114
Arterial density	Arterial length / Area of ZCTA (km/km^2)	0	18.119	1.699	2.594
Local road density	Local road length / Area of ZCTA (km/km^2)	0.774	48.268	20.644	7.971
Speed Limit (20 mph)	Ratio of road length with posted speed 20 mph to total length	0	0.41	0.02	0.06
Speed Limit (25 mph)	Ratio of road length with posted speed 25 mph to total length	0.19	1	0.72	0.17
Speed Limit (30 mph)	Ratio of road length with posted speed 30 mph to total length	0	0.24	0.02	0.03
Speed Limit (35 mph)	Ratio of road length with posted speed 35 mph to total length	0	0.23	0.01	0.03
Speed Limit (40 mph)	Ratio of road length with posted speed 40 mph to total length	0	0.15	0.01	0.02
Speed Limit (45 mph)	Ratio of road length with posted speed 45 mph to total length	0	0.1	0.01	0.02
Speed Limit (50 mph)	Ratio of road length with posted speed 50 mph to total length	0	0.48	0.04	0.06
Commercial area	The ratio of the area allocated for commercial purpose	0.08	1	0.31	0.21
Residential area	The ratio of the area allocated for residential purpose	0	0.92	0.69	0.21
Industrial area	The ratio of the area allocated for industrial purpose	0	0.23	0.08	0.11
Transportation area	The ratio of the area allocated for transportation purpose	0	0.67	0.03	0.46
Public institution area	The ratio of the area allocated for public institutions purpose	0	0.34	0.04	0.19
Outdoor recreation area	The ratio of the area allocated for outdoor recreation purpose	0	0.45	0.02	0.35
Demographic and socio-economic variables					
Unemployment	The number of unemployed people per km^2 in each ZCTA	0	5740	1310	1060
MIC	Median household income in each ZCTA(dollars)	0	232031	59967	31156
TTTW	Average travel time to work in each ZCTA (minutes)	0	56.6	38.21	7.62
Population density	The number of people per km^2 in each ZCTA	0	55477	17155	11681
Bachelor	The number of people with bachelor degree per km^2 in each ZCTA	0	36039	5194	6795
Younger	The number of people under 18 years old per km^2 in each ZCTA	0	10198	3320	2331
Elder	The number of people over 65 years old per km^2 in each ZCTA	0	10185	2087	1769

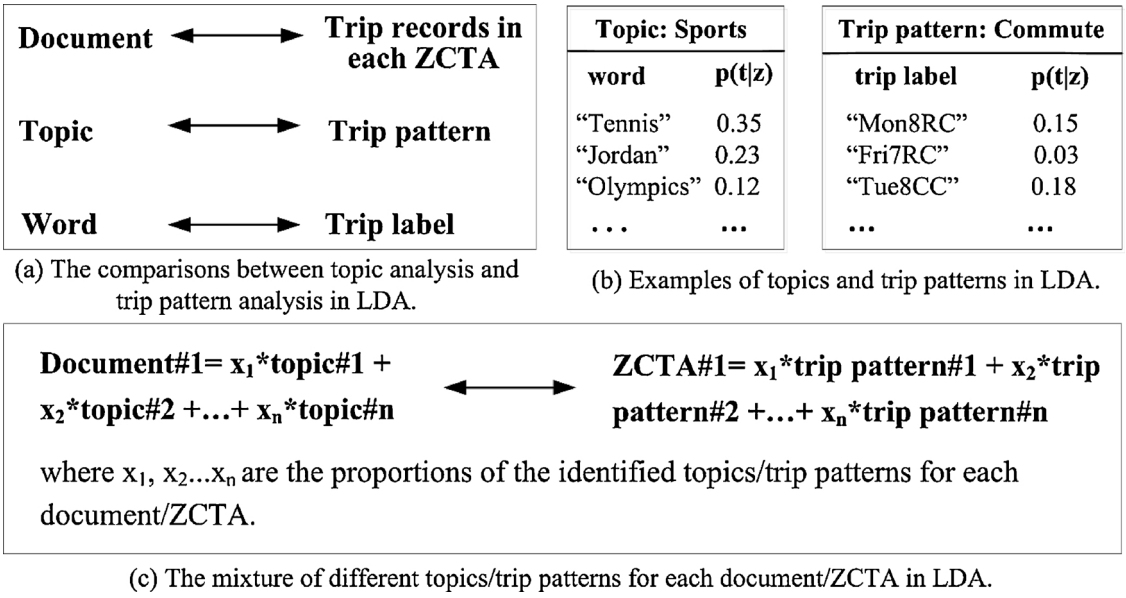


Fig. 2. Illustration of topics and trip patterns in LDA.

a trip label. For example, a taxi trip from a commercial area to a residential area at 6:00 pm on Monday was considered a trip record, and the trip record was transformed into a trip label Mon18CR. In the present study, trip labels were basic units of analysis for exploring trip patterns with the LDA method. As is shown in Fig. 2(a), each trip label was treated as a word of a topic, and each trip pattern was treated as a topic of a document. The trip records in each ZCTA were treated as the documents of a corpus. In addition, the total trip records in the whole dataset were assigned to each ZCTA by their origins in accordance with the definition of trip labels.

In a LDA model, each topic can be considered a mixture of words. Accordingly, each trip pattern in the present study can be considered a mixture of trip labels. Fig. 2(b) illustrates an example for a specific topic and trip pattern in a LDA model. Each document consists of a mixture of different topics. Accordingly, the taxi trip records in each ZCTA consist of a mixture of different trip patterns (See Fig. 2(c)). The outputs of the LDA model include the identified trip patterns from the total trip records, as well as the proportions of the identified trip patterns for each ZCTA.

Let us assume that K represents the number of the hidden trip patterns; M represents the number of ZCTAs; V represents the number of trip labels for each trip pattern; θ is a $M \times K$ matrix that represents

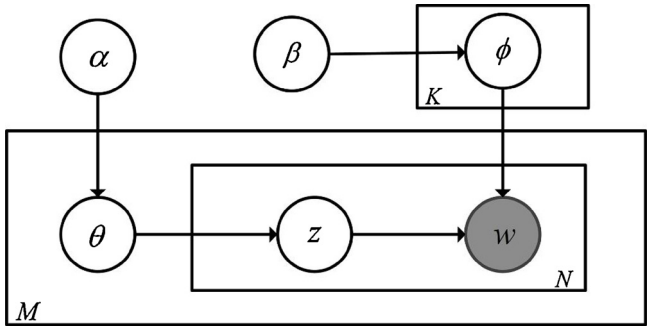


Fig. 3. The generating mechanism of the LDA method.

the mixture weights for the K trip patterns in each ZCTA; and ϕ is a $V \times K$ matrix that represents the mixture weights for the V trip labels in each trip pattern. θ and ϕ follow Dirichlet distributions with hyper-parameters α and β, respectively. The data generating mechanism of a LDA model can be illustrated by the flow chart given in Fig. 3 and is discussed as follows:

(1) For each trip pattern k∈{1,…,K}, choose a trip label distribution vector ϕ(k) ~ Dirichlet (β); and
(2) For each ZCTA m∈{1,…,M} in the whole dataset:
  (a) Choose a trip pattern distribution vector: θ(m) ~ Dirichlet (α); and
  (b) For each trip label i in ZCTA m:

a) Draw a trip pattern zi ~ Multinomial (θ(m)); zi∈{1,…, K}; and
b) Given the selected trip pattern zi, draw a trip label ti ~ Multinomial (ϕ(zi)); ti∈ {1,…,V}.

The present study used a Bayesian MCMC simulation based method for specifying the LDA models. The Dirichlet distribution is given as follows:

$$P(\theta \mid \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta^{\alpha_k - 1} \tag{1}$$

$$P(\phi \mid \beta) = \frac{\Gamma(\sum_v \beta_v)}{\prod_v \Gamma(\beta_v)} \prod_v \phi^{\beta_v - 1} \tag{2}$$

The conditional posterior distribution for z_i is given by

$$P(z_i = k \mid z_{-i}, t) \propto P(t_i \mid z_i = k, z_{-i}, t_{-i}) P(z_i = k \mid z_{-i}) \tag{3}$$

where $\mathbf{z}_{-i}$ is the assignment of all z_i, not including the current one. The conditional probability in Eq. (3) is not affected by θ and ϕ, because by integrating over the parameters θ and ϕ the conditional probability of each z_i depends only on $\mathbf{z}_{-i}$ and $\mathbf{t}$. The first term of Eq. (3) can be derived as:

$$P(t_i \mid z_i = k, z_{-i}, t_{-i}) = \int P(t_i \mid z_i = k, \phi^{(k)}) P(\phi^{(k)} \mid z_{-i}, t_{-i}) d\phi^{(k)} \tag{4}$$

where $\phi^{(k)}$ represent the multinomial distribution over trip labels associated with trip pattern k. According to the Bayes' rule, $P(\phi^{(k)} \mid \mathbf{z}_{-i}, \mathbf{t}_{-i})$ is conditional upon the production between $P(\mathbf{t}_{-i} \mid \phi^{(k)}, \mathbf{z}_{-i})$ and $P(\phi^{(k)})$. Since $P(\phi^{(k)})$ conjugates to $P(\mathbf{t}_{-i} \mid \phi^{(k)}, \mathbf{z}_{-i})$, the posterior distribution $P(\phi^{(k)} \mid \mathbf{z}_{-i}, \mathbf{t}_{-i})$ will be Dirichlet$(n_{-i,k}^{(t)} + \beta)$, where $n_{-i,k}^{(t)}$ is the number of instances of trip label t assigned to trip pattern k, not including the current trip label. Accordingly, the following results can be derived:

$$P(t_i \mid z_i = k, z_{-i}, t_{-i}) = \frac{n_{-i,k}^{(t_i)} + \beta}{n_{-i,k}^{(\cdot)} + V\beta} \tag{5}$$

where $n_{-i,k}^{(\cdot)}$ is the total number of trip labels assigned to trip pattern k, not including the current one; $n_{-i,k}^{(t_i)}$ represents the number of trip label t_i assigned to trip pattern k, not including the current one. With the same method, $P(z_i = k \mid \mathbf{z}_{-i})$ on the right hand side of Eq. (3) can be estimated as:

$$P(z_i = k \mid z_{-i}) = \frac{n_{-i,k}^{(m)} + \alpha}{n_{-i,\cdot}^{(m)} + K\alpha} \tag{6}$$

where $n_{-i,k}^{(m)}$ is the number of trip labels assigned to trip pattern k in ZCTA m, not including the current one, and $n_{-i,\cdot}^{(m)}$ is the total number of trip labels in ZCTA m, not including the current one. Based on Eqs. (5) and (6), the conditional posterior distribution for z_i can be calculated as:

$$P(z_i = k \mid z_{-i}, t) \propto \frac{n_{-i,k}^{(t_i)} + \beta}{n_{-i,k}^{(\cdot)} + V\beta} \frac{n_{-i,k}^{(m)} + \alpha}{n_{-i,\cdot}^{(m)} + K\alpha} \tag{7}$$

The MCMC algorithm was used to sample each z_i iteratively from the conditional distribution specified by Eq. (7) until the predetermined maximum iteration times had been reached. The inference was made based on the remaining draws after discarding the draws during the burn-in period. The final estimation results can be computed as:

$$\phi_k^t = \frac{n_k^{(t)} + \beta}{n_k^{(\cdot)} + V\beta} \tag{8}$$

3.2. Random forest

The random forest method is one of the most efficient methods for evaluating the importance of variables. Compared with traditional variable selection methods, such as classification trees and stepwise regression method, the random forest method has the capability of handling the multi-collinearity problem associated with candidate variables. Previous studies have indicated that the random forest method has high capability in obtaining unbiased and stable results without a separate cross-validation test dataset. In addition, the random forest method runs efficiently on large-scale datasets even with thousands of factors, making it a suitable technique for big data analysis (Jiang et al., 2016).

The random forest method was used in the present study to identify the contributing factors that had a great influence on the crash counts in each ZCTA. The random forest technique consists of an ensemble of randomized classification and regression trees (Breiman, 2001). A predetermined number of classification and regression trees are generated randomly and aggregated to give one single prediction. When solving the classification problems, the random forest model chooses the classification with the most votes from all the trees in the forest. During the training procedure, each tree is grown on a bootstrap sample from the original training dataset. Then, at each node of a tree, the best split is searched from a randomly selected subset of the whole predictors.

In a random forest model, two measures are commonly used measures for evaluating variable importance. They are: Out-of-bag (OOB) error rate and Gini index. For each of the bootstrap sample in a particular tree, about two thirds of the data points are used for training, while the remaining data points are used for testing, and these data are called the "out of bag" (OOB) data. The OOB error rate is calculated by the proportion of times that the voted class label is not equal to the true class, averaged over all cases in the OOB data. The importance of a particular variable can be measured by the average difference between the OOB error rate after and before permuting the variable over all trees in the forest. The Gini index measures the impurity of each node in a particular tree. The Gini index for a node t in a particular tree can be calculated using the following equation:

$$G(t) = 1 - \sum_{i=1}^{m} p^2(i \mid t) \tag{9}$$

where $G(t)$ denotes the Gini index for the node t; i represents the number of classes; and $p(i|t)$ represents the estimated class

probabilities. Then, the importance measure for variable X_i is defined as the average decrease in the Gini index over all trees in the forest for the nodes, whose splitting variable is X_i. Generally speaking, the larger the average decrease in the Gini index is the more important is the candidate variable. In the present study, the Gini index was used to select the contributing variables to the crash counts in each ZCTA.

3.3. Spatial correlations test

The Moran's I is a spatial autocorrelation statistic that can be used for measuring spatial dependency. The Moran's I statistic was used in this study to help identify whether the explanatory variables of the ZCTA-level crash model were spatially correlated. The Moran's I statistic for an unstandardized spatial weight matrix $\boldsymbol{C}$ takes the following forms (Lawson, 2009):

$$Moran'sI = \frac{N}{\sum_i \sum_j c_{ij}} \frac{\sum_i \sum_j c_{ij}(E_i - \overline{E})(E_j - \overline{E})}{\sum_i (E_i - \overline{E})} \tag{10}$$

where N is the number of spatial units (ZCTA in the present study); E_i and E_j is the proportion of explanatory variables in the i^{th} and j^{th} ZCTA; $\overline{E}$ represents the average proportion of explanatory variables across different ZCTAs; c_{ij} denotes an element of an unstandardized spatial weight matrix $\boldsymbol{C}$, which measures the connection between areas i and j. c_{ij} equals 1 if ZCTA i and j are neighbors, and 0 otherwise. The range of Moran's I statistic is between -1 and +1. Higher positive value indicates a greater degree of spatial dependence while negative value indicates spatial dispersion. A value near zero indicates a spatially random pattern.

The statistical significance of the Moran's I index is usually tested using the Z-score. The null hypothesis is that the proportions of explanatory variables are spatially independent in the study area. The Z-score of Moran's I can be calculated by:

$$Z_I = \frac{I - E(I)}{SD(I)} \tag{11}$$

where $E(I)$ and $SD(I)$ are the expectation and standard deviation of the Moran's I. A positive Z-score implies that the neighboring ZCTAs tend to have similar proportions of explanatory variables, whereas a negative Z-score indicates that the proportions of explanatory variables tend to be more dissimilar among neighboring ZCTAs.

3.4. Geographically weighted Poisson regression (GWPR)

Previous studies have proposed numerous methods to address the spatial heterogeneity problem when modeling spatially aggregated crash data. Some early studies have suggested using random-effect Poisson or negative binomial models (Tarko et al., 1996; Aguero-Valverde and Jovanis, 2006), while more recent studies have suggested using Bayesian spatial models to account for the unobserved spatial heterogeneity in crash data (Huang et al., 2010). Both random-parameter and Bayesian spatial models are relatively complex, and are not easy to put into general use at agencies. Recently, Hadayeghi et al. introduced the geographically weighted Poisson regression technique for modeling spatially aggregated crash data (Hadayeghi et al., 2010). Xu and Huang compared the GWPR model with the random-parameter model, and concluded that the GWPR was superior to the random-parameter model in fitting performance (Xu and Huang, 2015). In addition, compared with random-parameter models, the spatial distribution of the coefficients of variables in a GWPR model can be displayed in an easily identifiable manner. This property is very desirable for planners (Xu and Huang, 2015).

In a GWPR model, the crash counts are predicted by a set of explanatory variables of which the coefficients are allowed to vary over space (Fotheringham et al., 2002). After a review on the model specifications for zone-level crash frequency modeling in previous studies (Xu and Huang, 2015; Li et al., 2013; Hadayeghi et al., 2010), the following model form is considered in the present study:

$$Y_i \sim \text{Poisson}(\lambda_i) \tag{12}$$

$$\ln(\lambda_i) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)\ln(DVKT_i) + \beta_2(u_i, v_i)x_{2i} + \ldots + \beta_k(u_i, v_i)x_{ki} + \varepsilon_i \tag{13}$$

where Y_i denotes the observed crash frequency at the i^{th} ZCTA; λ_i denotes the expected crash frequency at the i^{th} ZCTA; $DVKT_i$ denotes the daily vehicle kilometers traveled at the i^{th} ZCTA; x_{ki} denotes the k^{th} explanatory variable at the i^{th} ZCTA ($k = 2, 3, \ldots, K$); and (u_i, v_i) denotes the coordinates of the i^{th} point in space (the centroid of the i^{th} ZCTA). Accordingly, the GWPR model allows parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ to be different across various ZCTAs such that spatial heterogeneity can be addressed. The local coefficients $\boldsymbol{\beta}$ can be denoted as the following matrix form:

$$\beta = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \ldots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \ldots & \beta_k(u_2, v_2) \\ \ldots & \ldots & \ldots & \ldots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \ldots & \beta_k(u_n, v_n) \end{bmatrix} \tag{14}$$

where n is the number of ZCTAs in the study area. The coefficients for each ZCTA in each row of the above matrix are estimated as follows (Fotheringham et al., 2002):

$$\hat{\beta}(i) = (\mathrm{X}^T \mathrm{W}(i) \mathrm{X})^{-1} \mathrm{X}^T W(i) Y \tag{15}$$

where $\mathbf{W}(i)$ is the weighting matrix, which can be denoted as:

$$\mathrm{W}(i) = \begin{bmatrix} w_{i1} & 0 & \ldots & 0 \\ 0 & w_{i2} & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & \ldots & \ldots & w_{in} \end{bmatrix} \tag{16}$$

where w_{in} is the allocated weight for ZCTA n at the regression point of ZCTA i. Generally, Gaussian and bi-square functions are the most commonly used weight functions, which can be expressed as:

$$\text{Gaussian weight function:} w_{ij} = \exp[-\frac{1}{2} \times (d_{ij}/b)^2] \tag{17}$$

$$\text{Bi-square weight function:} w_{ij} = \begin{cases} [1 - (d_{ij}/b)^2]^2 & \text{if } d_{ij} < b \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

where d_{ij} is the distance between the centroids of two ZCTAs, and the parameter b represents a bandwidth. In the present study, the Gaussian weight function with an adaptive bandwidth was adopted. More specifically, the areas with high density data points were assigned with a relatively low bandwidth, while those with low density data points were assigned with a relatively high bandwidth. The reason for using the Gaussian weight function is that the data away from the inflection point on the Gaussian curve can still get some fairly large weights, which helps ensure sufficient local information for calibrating a given local regression model (Wang and Chen, 2017). In addition, the adaptive bandwidth was adopted in the Gaussian weight function because the adaptive bandwidth accounts for the high standard errors of the estimated coefficients when the data are scarce (Li et al., 2013). The Corrected Akaike Information Criterion (AICc) was used for selecting the bandwidth and determining the best model. Note that the model with the lowest AICc was considered the best.

4. Results of data analysis

4.1. Spatiotemporal pattern of taxi trips and crashes

The drop-off locations of taxi trips and reported crashes during the selected time period in the study area are plotted in Fig. 4(a) and (b) to illustrate the spatial distributions of the taxi trips and crashes. The density of the taxi trips and the crashes are depicted with different

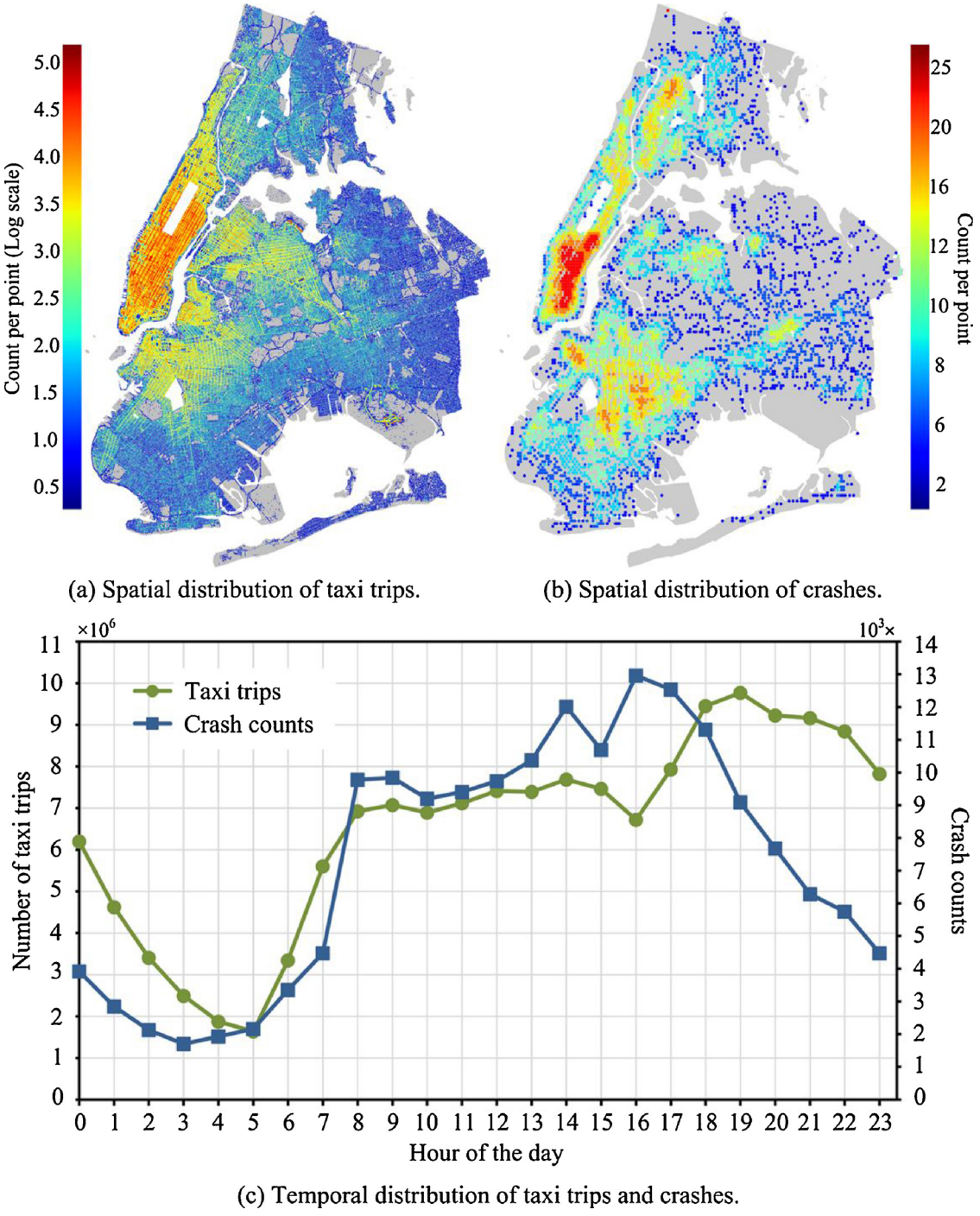


Fig. 4. The spatiotemporal distribution of taxi trips and crashes.

colors in which red color generally indicates high density, while blue generally indicates low density. By inspecting Fig. 4 visually we found that the spatial and temporal distribution of the taxi trips exhibited somewhat similar patterns with those of the crash counts. However, in certain time periods and spatial units they exhibited different patterns. For example, there is a peak for crashes around 4 pm while during that time the taxi trips have a significant drop, mainly due to the shift of taxi drivers. The findings that can be obtained from Fig. 4 are twofold. First, there is a strong relationship between trip trips and crashes. Second, the relationship between taxi trips and crashes is very complex and may also be interacted with other external factors. Thus, it may not be appropriate to directly include taxi trips as an explanatory variable in crash models. Additional efforts are needed to extract trip pattern information that are more closely related with crashes from the taxi GPS data.

4.2. Results of LDA analyses and trip patterns

The LDA analysis was conducted with the software R. Two R packages can be used to conduct the LDA analysis: the *lda* package and the *topicmodels* package. In this study, the *lda* package was used. The hyper-parameters of α and β and the number of trip patterns (K) need to be pre-specified before a LDA model is specified. The hyper-parameter α controls the shape and sparsity of the distribution of parameter $\theta^{(m)}$. A larger α indicates the distributions that are more uniform over topics, while a smaller α indicates sparser distributions. Usually the values of α

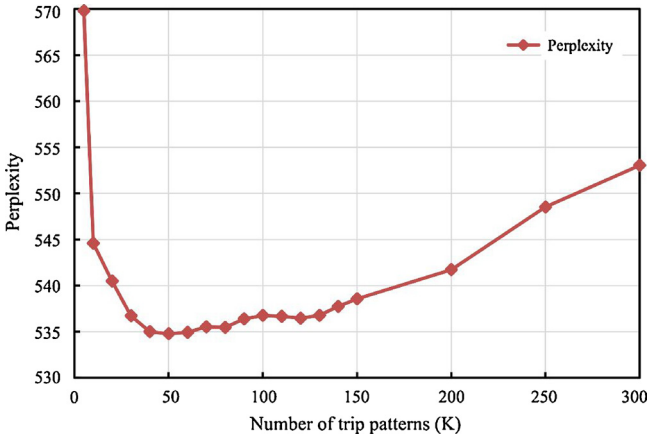


Fig. 5. Perplexity versus the number of trip patterns.

and β are selected subjectively by analysing the classification results, and the most commonly applied procedure sets $\alpha = 50/K$ and $\beta = 0.1$ (Hasan and Ukkusuri, 2014; Farrahi and Gatica-Perez, 2011). In this study, we tested various combinations of α and β, and found that when $\alpha = 5/K$ and $\beta = 0.1$ the LDA analysis reached the best results. In this condition, the identified trip patterns were intuitive and easy to understand, and distinguished from each other very well. Different values of K were also tested based on perplexity, which measured the fitness of the LDA model to training data. The lower the perplexity score is, the better predictive performance will the LDA model achieve. The perplexity of LDA can be computed as:

$$perplexity = \exp\left\{-\frac{\sum_{m=1}^{M} \log p(t_m)}{\sum_{m=1}^{M} N_m}\right\} \quad (19)$$

where N_m represents the number of trip labels in ZCTA m. The results suggest that when $K = 50$ the perplexity reached the minimum (See Fig. 5).

In total, fifty trip patterns were identified with the LDA model. For each trip pattern the LDA model provided the probability values of all the trip labels. A particular trip pattern can be defined based on the top ten trip labels ranked by the probability values (Hasan and Ukkusuri, 2014). Table 2 presents the top ten trip labels and their probability values for twelve selected trip patterns. Further analyses suggest that the twelve trip patterns greatly affected the crash in the selected ZCTAs. From the trip labels the following information can be obtained for a particular trip: the time and the day of the week of the trip; and the land use features associated with the pick-up and the drop-off locations. Fig. 6 depicts the temporal distributions of the twelve selected trip patterns based on the information provided by the trip labels. From Table 2 and Fig. 6, the properties of the selected twelve trip patterns can be clearly identified, and are discussed below.

- Trip pattern #1: commuting trips during morning peak periods;
- Trip pattern #8: weekend trips from outdoor recreation places to commercial areas;
- Trip pattern #10: weekend trips from residential areas to commercial areas at late-night;
- Trip pattern #13: commuting trips during morning peak periods;
- Trip pattern #17: trips from commercial areas to airport/train station;
- Trip pattern #21: trips from airport/train station to commercial/residential areas;
- Trip pattern #26: weekend trips from residential areas to commercial areas at late-night;
- Trip pattern #30: weekend trips around the commercial areas after dinner time;
- Trip pattern #37: commuting trips during evening peak periods;
- Trip pattern #40: weekend trips around the commercial areas in the afternoon;
- Trip pattern #45: commuting trips during evening peak periods; and
- Trip pattern #49: weekday trips around the commercial areas after dinner time.

4.3. Results of random forest analyses

With the identified fifty trip patterns, fifty trip pattern variables were defined. Each trip pattern variable was defined as the proportion of a particular trip pattern in the selected ZCTAs. The random forest analyses were conducted to identify the trip pattern variables that greatly affected the crashes reported in the selected ZCTAs. Previous studies have suggested that the contributing factors may vary across different crash severities (Rhee et al., 2016; Siddiqui et al., 2012). Thus, the random forest analyses were conducted separately for the total, the

Table 2
Results of the LDA model and identified trip patterns.

Trip label	P(t|z)	Trip label	P(t|z)	Trip label	P(t|z)	Trip label	P(t|z)	Trip label	P(t|z)	Trip label	P(t|z)
Trip pattern #1		**Trip pattern #8**		**Trip pattern #10**		**Trip pattern #13**		**Trip pattern #17**		**Trip pattern #21**	
Thu7CC	0. 0084	Wed9OC	0. 0089	Sat1RC	0. 0236	Thu8RC	0. 022	Fri13CT	0. 0189	Sun22TR	0. 0061
Wed7CC	0. 0084	Sat17OC	0. 0089	Sun1RC	0. 0222	Wed8RC	0. 0217	Fri0CR	0. 0182	Sun23TR	0. 0061
Thu8CC	0. 0082	Sat13OC	0. 0089	Sat0RC	0. 0205	Tue8RC	0. 0213	Thu15CT	0. 017	Sun21TR	0. 006
Tue8CC	0. 0082	Sat12OC	0. 0087	Sun0RC	0. 0202	Fri8RC	0. 0211	Thu14CT	0. 0166	Mon10TC	0. 0055
Wed8CC	0. 0081	Tue9OC	0. 0087	Sat2RC	0. 0196	Mon8RC	0. 0194	Fri14CT	0. 0166	Sun20TR	0. 0054
Thu9CC	0. 008	Thu9OC	0. 0086	Sat23RC	0. 0193	Thu7RC	0. 0193	Fri12CT	0. 016	Sun19TC	0. 0054
Wed9CC	0. 0079	Sat15OC	0. 0085	Sun2RC	0. 019	Wed7RC	0. 0187	Thu13CT	0. 0157	Mon15TC	0. 0054
Tue9CC	0. 0078	Sat14OC	0. 0085	Fri23RC	0. 0178	Tue7RC	0. 0186	Tue14CT	0. 0144	Mon11TC	0. 0053
Tue7CC	0. 0077	Sat18OC	0. 0083	Sat22RC	0. 0161	Fri7RC	0. 0176	Sun17CC	0. 0142	Mon23TR	0. 0053
Fri8CC	0. 0075	Sun15OC	0. 0082	Sun3RC	0. 0152	Mon7RC	0. 0162	Fri15CT	0. 0137	Sun17TC	0. 0052
Trip pattern #26		**Trip pattern #30**		**Trip pattern #37**		**Trip pattern #40**		**Trip pattern #45**		**Trip pattern #49**	
Sun1RC	0.0415	Sat19CC	0.0238	Wed18CC	0.0373	Sat15CC	0.0293	Wed18CR	0. 0133	Tue19CC	0.0084
Sun0RC	0.0411	Sat21CC	0.0234	Wed19CC	0.0351	Sat14CC	0.0292	Thu18CR	0. 0133	Tue20CC	0.0084
Sat1RC	0.0361	Sat20CC	0.0232	Thu18CC	0.0334	Sat13CC	0.0276	Wed17CR	0. 0128	Wed19CC	0.0083
Sun2RC	0.0361	Sat22CC	0.0227	Thu19CC	0.0322	Sun14CC	0.0271	Thu17CR	0. 0124	Wed20CC	0.0083
Sat2RC	0.0314	Sat18CC	0.0201	Tue19CC	0.0312	Sun13CC	0.0242	Tue18CR	0. 0122	Mon19CC	0.0083
Sat0RC	0.0311	Sat11CC	0.0197	Thu20CC	0.0287	Sat17CC	0.0237	Fri18CR	0. 0122	Mon18CC	0.0082
Sun3RC	0.0268	Sun12CC	0.0197	Wed20CC	0.0284	Sun15CC	0.0237	Tue17CR	0. 0117	Wed21CC	0.008
Sun23RR	0.026	Sat23CC	0.0192	Thu17CC	0.0265	Sat16CC	0.0223	Fri17CR	0. 0117	Tue18CC	0.008
Sat23RC	0.0243	Fri23CC	0.0189	Fri18CC	0.0254	Sun12CC	0.0208	Mon17CR	0. 0105	Tue21CC	0.0079
Sat22RR	0.0241	Sat12CC	0.0189	Fri17CC	0.0244	Sun11CC	0.0203	Wed19CR	0. 0105	Mon20CC	0.0077

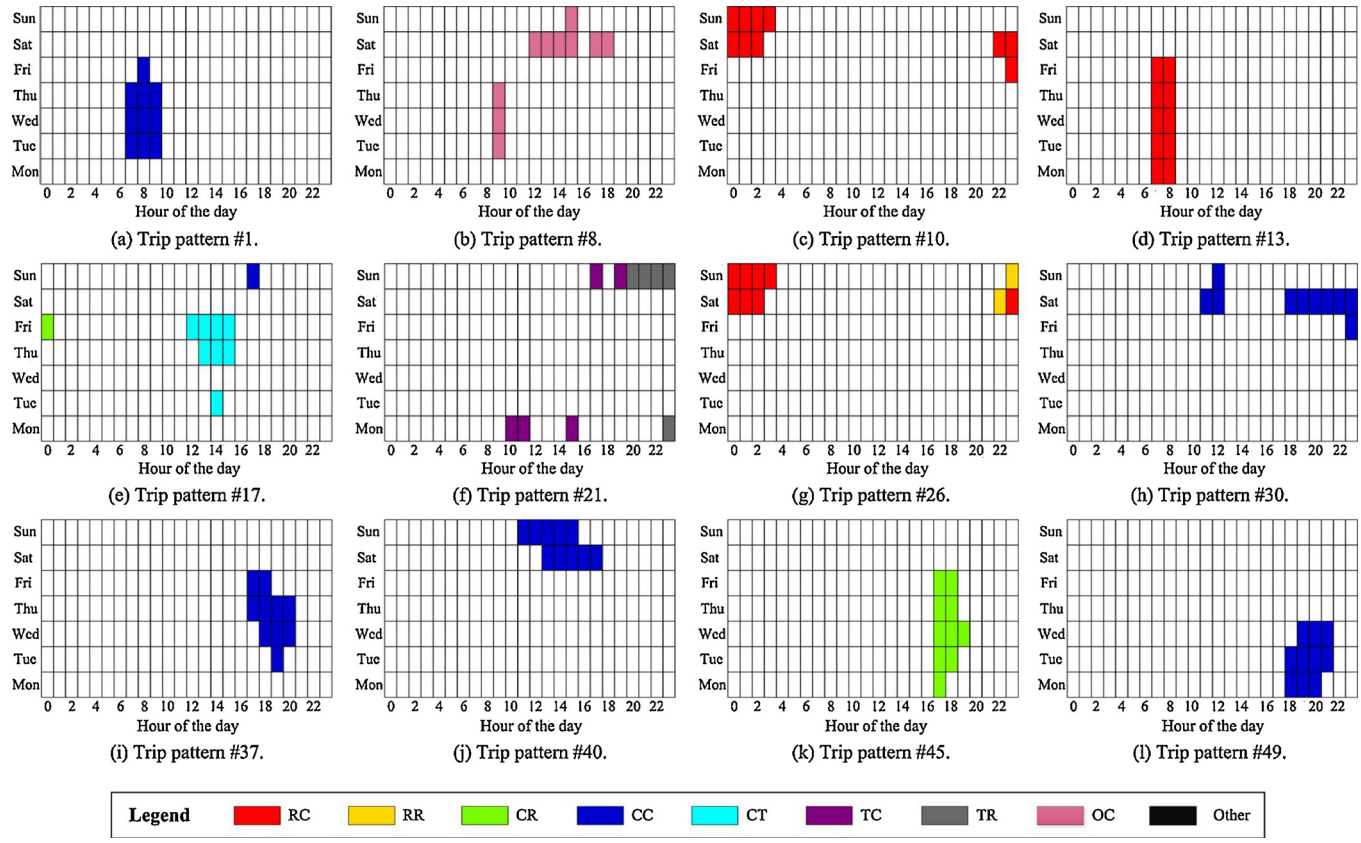


Fig. 6. Visualization of the temporal distributions of the identified trip patterns.

PDO, and the fatal-plus-injury crashes. To achieve stable estimates of variable importance, the forests were trained with the whole dataset for fifty times. The average OOB error rates for different numbers of trees in the forest were calculated. With 500 trees the constant minimum error rates were reached and the number of trees was then set to be 500.

A total of 74 candidate variables, including the fifty trip pattern variables and the 24 road network and social demographic variables (see Table 1), were evaluated using the random forest technique. The random forest was trained fifty times for each type of crashes, and the average normalized variable importance was calculated. To select the number of contributing factors, the random forest analyses were conducted in a successive phase in which the number of input variables was set from 1 to 74. Fig. 7 depicts the selected contributing factors for the total, the PDO and the fatal-plus-injury crashes at the ZCTA-level. The results suggest that the trip patterns that contributed to the PDO crashes were different from those contributed to the fatal-plus-injury crashes.

The proportions of trip patterns #10, #17, #21, #26 and #37 greatly affected the fatal-plus-injury crashes in the selected ZCTAs. The trip patterns #17 and #21 represent the trips from airport/train station to commercial/residential areas. Siddiqui et al. suggested that airport-related trips are associated high crash risks because of the hasty attitude to reach the destination. In addition, the airport-related trips are usually related to non-familiar travelers with rental cars and on high-speed roads, leading to crashes with high severity levels (Siddiqui et al., 2012). The trip patterns #10 and #26 represent the late-night trips around commercial areas on weekends. These trip patterns could be related to the leisure activities of young people, and therefore, are more likely to be involved with drunk and careless driving behaviors, leading to increased risks of fatal-plus-injury crashes (Gregersen and Berg, 1994; Twisk and Stacey, 2007). Moreover, the police reports also confirmed this explanation. According to the collision reports collected from the New York Police Department, nearly 29.87% of the fatal-plus-injury crashes occurred during late-night were associated with drunk, unsafe and careless driving behaviors.

4.4. Specification of the GWPR models

The candidate variables included the 24 road network and social demographic variables (see Table 1) and the trip pattern variables selected with the random forest analysis. The results of the Moran's I test are given in Table 3. All the explanatory variables are statistically significant at 0.05 level, indicating strong spatial correlation. Moreover, the z-score values are all positive, suggesting that the spatial distribution of variables is more likely to be spatially clustered. Thus, it is highly desirable to use the GWPR model to explore the spatial heterogeneity in the data.

Three general categories of models were developed: the total crash model, the PDO crash model and the fatal-plus-injury crash model. For each category three different types of models were developed, and their performance was compared: the model with the traditional traffic exposure variables only, the model with the trip pattern variables only, and the model with both the traditional traffic exposure and the trip pattern variables. Thus, in total the research team developed nine GWPR models. The models were specified with the software GWR 4.0. The specification results of the GWPR models are given in Tables 4–6.

A stepwise procedure was followed to select the explanatory variables that should be included in the GWPR models. The variables were added one by one by checking their significance and the AICc of the model. Only the variables that were statistically significant with at least 90% level of confidence were selected. In the GWPR models, one variable may be significant in some ZCTAs while insignificant in others. To address this concern, the variables that were significant in over 80% of the ZCTAs were selected. The variable selection procedure was repeated iteratively until the AICc value reached the minimum. The GWPR models with the smallest AICc were considered the best models. To verify possible multicollinearity between the selected explanatory

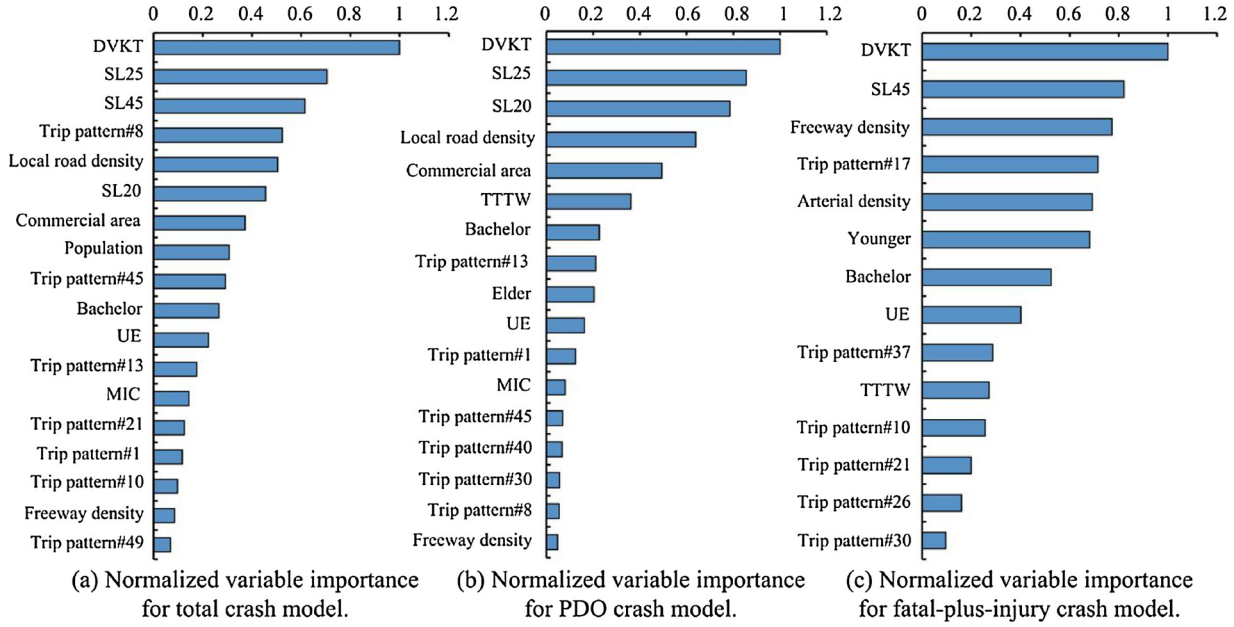


Fig. 7. Normalized variable importance ranking for total, PDO, and fatal-plus-injury crashes.

Table 3
Moran's I test for candidate independent variables.

Variables	Moran's index	Expected index	z-score	p-value
DVKT	0.134	−0.006	4.679	0.000
Freeway density	0.163	−0.006	4.786	0.000
Arterial density	0.38	−0.006	11.3	0.000
Local road density	0.512	−0.006	14.507	0.000
Speed Limit (20 mph)	0.159	−0.006	4.859	0.000
Speed Limit (25 mph)	0.426	−0.006	12.05	0.000
Speed Limit (30 mph)	0.168	−0.006	5.052	0.000
Speed Limit (35 mph)	0.262	−0.006	8.555	0.000
Speed Limit (40 mph)	0.268	−0.006	7.845	0.000
Speed Limit (45 mph)	0.128	−0.006	3.893	0.000
Speed Limit (50 mph)	0.152	−0.006	4.675	0.000
Commercial area	0.356	−0.006	10.158	0.000
Residential area	0.341	−0.006	9.731	0.000
Industrial area	0.243	−0.006	7.331	0.000
Transportation area	0.353	−0.006	10.232	0.000
Public institution area	0.254	−0.006	7.714	0.000
Outdoor recreation area	0.224	−0.006	6.88	0.000
Unemployment	0.569	−0.006	16.191	0.000
MIC	0.592	−0.006	17.157	0.000
TTTW	0.699	−0.006	19.852	0.000
Population density	0.555	−0.006	15.696	0.000
Bachelor	0.681	−0.006	19.495	0.000
Younger	0.54	−0.006	15.227	0.000
Elder	0.554	−0.006	15.889	0.000
Trip pattern #1	0.232	−0.006	7.357	0.000
Trip pattern #8	0.066	−0.006	2.515	0.012
Trip pattern #10	0.145	−0.006	4.965	0.000
Trip pattern #13	0.072	−0.006	2.354	0.019
Trip pattern #17	0.248	−0.006	7.571	0.000
Trip pattern #21	0.089	−0.006	3.444	0.001
Trip pattern #26	0.201	−0.006	5.908	0.000
Trip pattern #30	0.321	−0.006	10.071	0.000
Trip pattern #37	0.314	−0.006	9.56	0.000
Trip pattern #40	0.367	−0.006	13.01	0.000
Trip pattern #45	0.174	−0.006	5.271	0.000
Trip pattern #49	0.592	−0.006	17.15	0.000

variables, the variance inflation factors were calculated following the ordinary least square analyses. The VIF values for all the selected variables were less than 10, suggesting that the problem of multicollinearity did not exist.

Comparative analyses were conducted to compare the models with different types of trip pattern variables. For all three categories (total, PDO, and fatal-plus-injury crashes), the models with both traditional traffic exposure and trip pattern variables always performed the best in terms of the lowest mean absolute deviation (MAD) values and AICc values. The finding suggests that incorporating trip pattern variables improved the goodness-of-fit of the GWPR models. In addition, the coefficients of the traditional traffic exposure variables were quite similar in both the $GWPR_T$ and the $GWPR_B$ models, implying that including trip pattern variables in the GWPR models did not produce biased estimates for the influence of traditional traffic exposure variables.

Moreover, the likelihood ratio tests were conducted to compare the GWPR models and traditional fixed-parameter Poisson regression (FPR) models. As is shown in Table 7, the test results suggest that the GWPR models are significantly better than the FPR models at the 95% confidence level for all the developed crash models. The results of likelihood ratio test confirmed the superiority of GWPR method in modeling spatially aggregated crash data.

The model specification results suggest that the trip pattern variables significantly affected the crash counts in a ZCTA. In addition, the selected trip pattern variables were quite different in the PDO and the fatal-plus-injury crash models. More specifically, the proportions of trip pattern #13 and #45 significantly affected the PDO crashes, while the proportions of trip pattern #10 and #17 significantly affected the fatal-plus-injury crashes. Trip pattern #13 and #45 represent the commuting trips during morning peak and evening peak, respectively. Commuting trips usually constitute a large proportion of the total number of daily trips, and accordingly have a great influence on crash occurrence. In addition, during the peak periods the average operating speed on road networks is relatively low due to the traffic congestion, leading to crashes with low severity levels.

In addition, trip pattern #10 represents the late-night trips around commercial areas on weekends. This trip pattern is usually related with leisure activities, especially for young people. Many previous studies have suggested that young drivers after late-night entertainment are usually involved with drunk, unsafe and careless driving behaviors, thus leading to increased risks of fatal-plus-injury crashes (Gregersen and Berg, 1994; Twisk and Stacey, 2007). In addition, according to the collision reports collected from the NYPD, nearly 29.87% of the fatal-plus-injury crashes occurred during late-night are caused by drunk, unsafe and careless driving behaviors. Trip pattern #17 represents trips

Table 4
Results of three different total crash models.

Total crash model ($GWPR_T$[a])			Total crash model ($GWPR_P$[a])			Total crash model ($GWPR_B$[a])		
Variable	Coefficient (mean)	VIF (OLS)	Variable	Coefficient (mean)	VIF (OLS)	Variable	Coefficient (mean)	VIF (OLS)
Intercept	−1.166		Intercept	2.945		Intercept	−1.712	
Ln(DVKT)	0.377	1.164	Trip pattern #10	1.664	1.093	Ln(DVKT)	0.397	1.409
Commercial area	0.464	1.23	Trip pattern #45	3.001	1.193	Trip pattern #10	1.771	1.301
Freeway density	0.103	1.384	Commercial area	0.178	1.249	Trip pattern #45	2.043	1.202
Local road density	0.037	2.726	Freeway density	0.209	1.32	Commercial area	0.562	1.259
Speed Limit (20 mph)	2.066	1.288	Local road density	0.043	2.735	Freeway density	0.105	1.39
Speed Limit (25 mph)	2.353	1.842	Speed Limit (20 mph)	2.921	1.334	Local road density	0.039	2.744
Speed Limit (45 mph)	7.535	1.14	Speed Limit (25 mph)	3.083	1.974	Speed Limit (20 mph)	2.241	1.344
Population density	2.2E-05	3.801	Speed Limit (45 mph)	11.702	1.146	Speed Limit (25 mph)	2.607	2.009
Bachelor	−1.3E-05	3.695	Population density	2.0E-05	3.825	Speed Limit (45 mph)	6.792	1.18
			Bachelor	−2.1E-05	3.679	Population density	2.1E-05	3.828
						Bachelor	−1.3E-05	3.721
Goodness of fit								
MAD	298.353			449.376			290.891	
AICc	357.461			428.218			340.886	

P = with the trip pattern variables only.
B = with both the traditional traffic exposure and the trip pattern variables.
[a] T = with the traditional traffic exposure variables only.

from airport/train stations to commercial/residential areas. This trip pattern is associated with the fatal-plus-injury crash counts mainly because that airport-related trips are usually related to non-familiar travelers with rental cars on high-speed roads and also with hasty attitude to reach the destination, thus leading to crashes with high severity levels (Siddiqui et al., 2012).

The social demographic factors in the PDO crash models were also different from those in the fatal-plus-injury crash models. The number of elder people in a ZCTA was positively correlated with the number of PDO crashes. The finding is intuitive because elder people usually have longer perception-reaction time and are more likely to be involved in crashes. However, elder people usually drive at lower speeds, leading to lower crash severity levels. By contrast, the number of younger people in a ZCTA was positively correlated with the number of fatal-plus-injury crashes. The finding is consistent with many previous studies, which have suggested that youngers tend to drive at higher speeds and are more likely to be engaged in aggressive driving behaviors (Huang et al., 2010; Aguero-Valverde and Jovanis, 2006). In addition, the population with a bachelor's degree in a ZCTA was negatively correlated with both the PDO and the fatal-plus-injury crashes. The explanation is that people with higher education levels are more likely to follow traffic rules, and are more apt to have safer driving behaviors. The ratio of the area allocated for commercial purpose were positively correlated with both PDO and the fatal-plus-injury crashes counts. The finding is intuitive and consistent with the results of previous studies (Aziz et al., 2013; Rhee et al., 2016). Moreover, the density of freeway and arterial in a ZCTA were positively correlated with the fatal-plus-injury crash counts. The finding is also consistent with many previous studies which suggested that these road types are usually involved with high-speed vehicles, thus leading to increased crash severity levels (Hadayeghi et al., 2010).

With the results of the GWPR models, figures can be developed to illustrate the spatial patterns of the coefficients of the trip pattern variables across the selected ZCTAs. As shown in Fig. 8(a), the coefficients of the proportions of trip pattern #13 and #45 reach a peak around the Manhattan area. The finding is intuitive because Manhattan is the most populated commercial area with the greatest number of commuting trips in the City of New York. In the Bronx, however, the coefficient of the proportion of trip pattern #13 is relatively small, while the coefficient of the proportion of trip pattern #45 is quite large. The finding suggests that in the Bronx, the commuting trips during evening peak periods had a greater influence on the PDO crashes than

Table 5
Results of three different PDO crash models.

PDO crash model ($GWPR_T$[a])			PDO crash model ($GWPR_P$[a])			PDO crash model ($GWPR_B$[a])		
Variable	Coefficient (mean)	VIF (OLS)	Variable	Coefficient (mean)	VIF (OLS)	Variable	Coefficient (mean)	VIF (OLS)
Intercept	−1.384		Intercept	3.402		Intercept	−1.908	
Ln(DVKT)	0.416	1.05	Trip pattern #13	1.274	1.082	Ln(DVKT)	0.432	1.32
Commercial area	0.517	1.224	Trip pattern #45	2.608	1.204	Trip pattern #13	2.254	1.247
Local road density	0.04	1.898	Commercial area	0.726	1.341	Trip pattern #45	1.753	1.207
Speed Limit (20 mph)	2.453	1.206	Local road density	0.045	2.18	Commercial area	0.656	1.343
Speed Limit (25 mph)	2.052	1.548	Speed Limit (20 mph)	3.122	1.253	Local road density	0.044	2.18
Elder	4.3E-05	2.088	Speed Limit (25 mph)	2.43	1.749	Speed Limit (20 mph)	2.475	1.263
			Elder	1.54E-04	3.66	Speed Limit (25 mph)	2.252	1.755
			Bachelor	−4.3E-05	4.137	Elder	6.1E-05	3.806
						Bachelor	−9.0E-06	4.427
Goodness of fit								
MAD	260.018			366.135			247.812	
AICc	353.942			441.846			339.111	

P = with the trip pattern variables only.
B = with both the traditional traffic exposure and the trip pattern variables.
[a] T = with the traditional traffic exposure variables only.

Table 6
Results of three different fatal-plus-injury crash models.

Fatal-plus-injury crash model ($GWPR_T$[a])			Fatal-plus-injury crash model ($GWPR_P$[a])			Fatal-plus-injury crash model ($GWPR_B$[a])		
Variable	Coefficient (mean)	VIF (OLS)	Variable	Coefficient (mean)	VIF (OLS)	Variable	Coefficient (mean)	VIF (OLS)
Intercept	−0.131		Intercept	4.791		Intercept	−0.08	
Ln(DVKT)	0.371	1.405	Trip pattern #10	1.878	1.046	Ln(DVKT)	0.372	1.413
Commercial area	0.425	1.225	Trip pattern #17	4.629	1.393	Trip pattern #10	0.939	1.049
Freeway density	0.037	1.235	Commercial area	0.395	1.24	Trip pattern #17	4.91	1.397
Arterial density	0.065	1.344	Freeway density	0.037	1.18	Commercial area	0.54	1.244
Speed Limit (45 mph)	7.334	1.101	Arterial density	0.112	1.292	Freeway density	0.049	1.241
Younger	1.42E-04	1.294	Speed Limit (45 mph)	8.664	1.116	Arterial density	0.031	1.529
Bachelor	−1.5E-05	1.272	Younger	1.0E-04	1.39	Speed Limit (45 mph)	5.947	1.129
			Bachelor	−1.9E-05	1.276	Younger	1.24E-04	1.403
						Bachelor	−1.4E-05	1.281
Goodness of fit								
MAD	54.825			69.874			52.334	
AICc	337.998			387.254			320.834	

P = with the trip pattern variables only.
B = with both the traditional traffic exposure and the trip pattern variables.
[a] T = with the traditional traffic exposure variables only.

did the commuting trips during morning peak periods. One possible explanation is that many people living in the Bronx may work in Manhattan. They are more likely to take public transits during morning peak periods while have more chances for hailing a taxi after work (Qian and Ukkusuri, 2015). In addition, as shown in Fig. 8(b), the coefficient of the proportion of trip pattern #10 takes the highest value in downtown Manhattan, mainly because most of the famous recreational and entertainment places in the City of New York are located in this area.

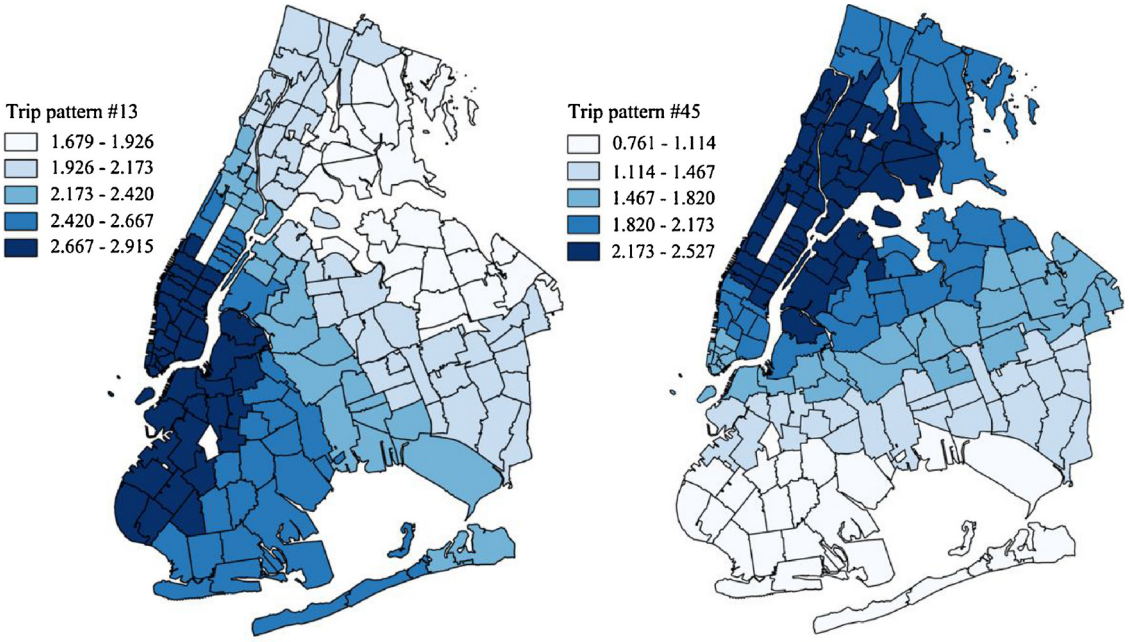
5. Conclusions and discussions

The present study investigated how the trip patterns variables extracted from large-scale taxi GPS data contributed to the crashes at spatially-aggregated levels in urban areas. A data-driven modeling approach based on Latent Dirichlet Allocation was proposed for discovering the hidden trip patterns from a large-scale taxi GPS dataset. Random forest technique was used to identify the factors that contributed to the total, PDO and fatal-plus-injury crashes in the selected ZCTAs during the study period. GWPR models were then developed to establish a relationship between the crash counts and the contributing factors selected by the random forest technique. Comparative analyses were conducted to compare the performance of the GWPR models that considered the traditional traffic exposure variables only, the trip pattern variables only, and both the traditional exposure and the trip pattern variables. The model specification results suggest that the trip pattern variables significantly affected the crash counts in the selected ZCTAs. The results of the comparative analyses suggest that the models that considered both the traditional traffic exposure and the trip pattern variables had the best goodness-of-fit in terms of the lowest MAD values and AICc values. The finding seems to confirm the benefits of incorporating the trip pattern information extracted from large-scale taxi GPS data in the spatial analysis of crashes.

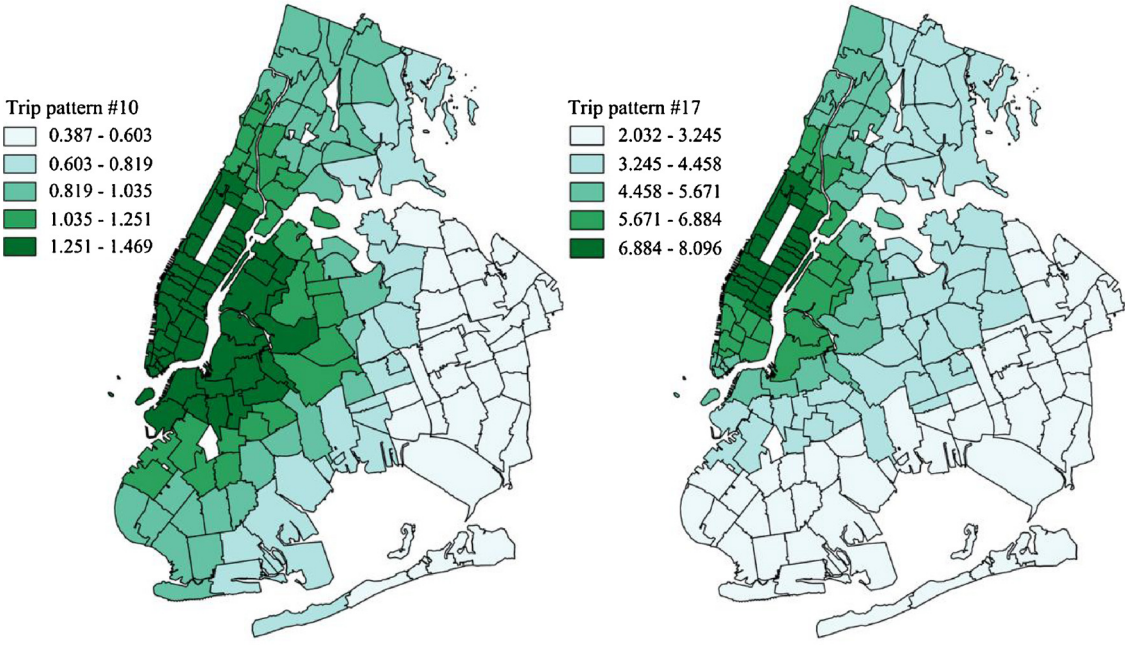
Traditionally, trip pattern information was estimated with the trip generation models developed based on household travel survey data. Taxi GPS data provided a promising way for estimating trip patterns and traffic conditions for a road network. In addition, from taxi GPS data researchers can identify the trip patterns that cannot be easily discovered with traditional household travel survey data. For example, in the present study fifty trip patterns were identified. The identified trip patterns included late-night trips and airport/train station-related trips. The present study demonstrates that the inclusion of the trip pattern information extracted from large-scale taxi GPS data greatly improve the fitness performance of crash models. Thus, taxi GPS data

Table 7
Likelihood ratio test for the GWPR and FPR models.

	Total crash model (T^*)	Total crash model (P^*)	Total crash model (B^*)
$LL(\beta_{FPR})$	−175.519	−201.02	−173.608
$LL(\beta_{GWPR})$	−158.982	−189.882	−147.699
$LR = -2[LL(\beta_{FPR}) - LL(\beta_{GWPR})]$	33.074	22.276	51.818
Critical χ^2 (95% confidence)	18.307	19.675	21.026
	PDO crash model (T^*)	PDO crash model (P^*)	PDO crash model (B^*)
$LL(\beta_{FPR})$	−181.698	−211.078	−178.747
$LL(\beta_{GWPR})$	−162.283	−200.741	−155.452
$LR = -2[LL(\beta_{FPR}) - LL(\beta_{GWPR})]$	38.83	20.674	46.59
Critical χ^2 (95% confidence)	14.067	16.919	18.307
	Fatal-plus-injury crash model (T^*)	Fatal-plus-injury crash model (P^*)	Fatal-plus-injury crash model (B^*)
$LL(\beta_{FPR})$	−177.123	−192.606	−167.333
$LL(\beta_{GWPR})$	−152.329	−175.658	−140.733
$LR = -2[LL(\beta_{FPR}) - LL(\beta_{GWPR})]$	49.588	33.896	53.2
Critical χ^2 (95% confidence)	15.507	16.919	18.307



(a) The spatial distribution of the coefficients of trip pattern variables: PDO crash model.



(b) The spatial distribution of the coefficients of trip pattern variables: fatal-plus-injury crash model.

Fig. 8. The spatial pattern of the coefficients of the trip pattern variables in the GWPR models.

have potential to serve as new data sources for roadway safety analyses, and are particularly useful for analyzing spatially aggregated crash data.

One possible concern is that the trip pattern extracted from the taxi GPS dataset may not be able to represent the general trip making of the whole city. The sample bias problem is a prevalent problem in many big data related studies, such as travel behavior analysis and human mobility studies that rests on taxi trip data. The inherent biases/limitations of taxi trip data may potentially affect the results of crash models, introducing biases in model specification. In the present study, the results of comparative analyses suggested that the coefficients of traditional traffic exposure variables were quite similar in both the models with and without trip pattern variables. However, it is still very difficult to fully demonstrate that the incorporation of trip pattern variables extracted from large-scale taxi GPS data does not introduce biases in crash models. This is particularly true for spatially aggregated crash models which focus on a large transportation network, and most of the exposure variables, such as AADT, DVMT and the total number of trips in a ZCTA, cannot be directly measured. For example, in most cases the AADT on the entire network is not available, and researchers use AADT on freeways as a surrogate measure for the AADT on the entire network. By doing so biases will also be introduced, and it is very difficult to identify which is the perfect model without any biases.

In the author's opinion, even though taxi trip data may suffer from various biases and limitations, it does provide new information to help better understand the variability in crash counts at zone-level. The

finding of the comparative analyses suggested that the best models considered both traditional traffic exposure and trip pattern variables. In other words, even though the trip pattern variables cannot fully replace traditional traffic exposure variables, they can be included in crash models as supplementary information to help improve model performance. In addition, more and more on-board GPS data of private vehicles and other modes can be collected, especially in urban areas. In the future, other data sources, such as the mobile phone data could be combined with taxi GPS data for generating a better understanding of the human mobility patterns in urban areas. The authors recommend that future studies could focus on this issue.

Acknowledgements

This researchwas supported by the National Natural Science Foundation of China (Grant No. 51561135003) and the Scientific Research Foundation of Graduate School of Southeast University (No. YBJJ1790).

References

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accid. Anal. Prev. 38, 618–625.
Aziz, H.M.A., Ukkusuri, S.V., Hasan, S., 2013. Exploring the determinants of pedestrian-vehicle crash severity in New York City. Accid. Anal. Prev. 50, 1298–1309.
Bao, J., Liu, P., Yu, H., Xu, C., 2017. Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas. Accid. Anal. Prev. 106, 358–369.
Blei, D.M., Andrew, Ng.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Chen, Z., Schintler, L.A., 2015. Sensitivity of location-sharing services data: evidence from american travel pattern. Transportation 42 (4), 669–682.
Chen, M., Mao, S., Liu, Y., 2014. Big data: a survey. Mob. Netw. Appl. 19, 171–209.
Côme, E., Randriamanamihaga, A., Oukhellou, L., Aknin, P., 2014. Spatio-temporal analysis of dynamic origin-destination data using latent dirichlet allocation. Application to Vélib' Bikesharing system of Paris. Proceedings of the 93th TRB Annual Meeting.
Farrahi, K., Gatica-Perez, D., 2011. Discovering routines from large-scale human locations using probabilistic topic models. ACM Trans. Intell. Syst. Technol. 2, 1–27.
Federal Highway Administration, 2005. Safetea-LU: Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users. U.S. Department of Transportation.
Fotheringham, A.S., Brunsdon, C., Charlton, M.E., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, Chichester.
González, M.C., Hidalgo, C.A., Barabási, A.L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782.
Gregersen, N.P., Berg, H.Y., 1994. Lifestyle and accidents among young drivers. Accid. Anal. Prev. 26 (3), 297–303.
Hadayeghi, A., Shalaby, A.S., Persaud, B.N., Cheung, C., 2006. Temporal transferability and updating of zonal level accident prediction models. Accid. Anal. Prev. 38 (3), 579–589.
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, 676–688.
Hasan, S., Ukkusuri, S.V., 2014. Urban activity pattern classification using topic models from online geo-location data. Trans. Res. Part C 44, 363–381.
Hasan, S., Schneider, C.M., Ukkusuri, S.V., González, M.C., 2013. Spatiotemporal patterns of urban human mobility. J. Stat. Phys. 151, 304–318.
Huang, H.L., Abdel-Aty, M.A., Darwiche, A.L., 2010. County-level crash risk analysis in Florida Bayesian spatial modeling. Trans. Res. Rec.: J. Trans. Res. Board 2148, 27–37.
Jiang, X., Abdel-Aty, M., Hu, J., Lee, J., 2016. Investigating macro-level hotzone identification and variable importance using big data: a random forest models approach. Neurocomputing 181, 53–63.
Lawson, A., 2009. Bayesian Disease Mapping Hierarchical Modeling in Spatial Epidemiology. CRC Press.
Lee, J., Abdel-Aty, M.A., Choi, K., 2014. Analysis of residence characteristics of at-fault drivers in traffic crashes. Saf. Sci. 68, 6–13.
Lee, J., Abdel-Aty, M.A., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accid. Anal. Prev. 102, 213–226.
Li, Z.B., 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, 89–97.
Liu, X., Gong, L., Gong, Y., Liu, Y., 2015. Revealing travel patterns and city structure with taxi trip data. J. Transp. Geogr. 43, 78–90.
Lovegrove, G.R., Lim, C., Sayed, T.A., 2008. Using macrolevel collision prediction models to conduct Road safety evaluation of regional transportation plan. Proceedings of the 87th TRB Annual Meeting.
Naderan, A., Shahi, J., 2010. Aggregate crash prediction models: introducing crash generation concept. Accid. Anal. Prev. 42, 339–346.
NCHRP (National Cooperative Highway Research Program), 2010. PLANSAFE: Forecasting the Safety Impacts of Socio-demographic Changes and Safety Countermeasures. Transportation Research Board. NCHRP, pp. 8–44.
Noland, R.B., 2003. Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. Accid. Anal. Prev. 35, 599–611.
Pereira, F.C., Rodrigues, F., Akiva, M.B., 2013. Text analysis in incident duration prediction. Trans. Res. Part C 37, 177–192.
Qian, X.W., Ukkusuri, S.V., 2015. Exploring spatial variation of Urban taxi ridership using geographically weighted regression. Proceedings of the 94th TRB Annual Meeting.
Rhee, K., Kim, J., Lee, Y., Ulfarsson, G.F., 2016. Spatial regression analysis of traffic crashes in Seoul. Accid. Anal. Prev. 91, 190–199.
Siddiqui, C., Abdel-Aty, M., Huang, H., 2012. Aggregate nonparametric safety analysis of traffic zones. Accid. Anal. Prev. 45, 317–325.
Tang, J., Liu, F., Wang, Y., Wang, H., 2015. Uncovering urban human mobility from large scale taxi GPS data. Physica A 438, 140–153.
Tarko, A.P., Sinha, K.C., Farooq, O.A., 1996. Methodology for identifying highway safety problem areas. Trans. Res. Rec.: J. Trans. Res. Board 1542, 49–53.
Twisk, D.A.M., Stacey, C., 2007. Trends in young driver risk and countermeasures in European countries. J. Safety Res. 38, 245–257.
Wang, C., Chen, N., 2017. A geographically weighted regression approach to investigating the spatially varied built-environment effects on community opportunity. J. Transp. Geogr. 62, 136–147.
Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accid. Anal. Prev. 75, 16–25.
Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prev. 66, 80–88.