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

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