Accident Analysis and Prevention 106 (2017) 358–369

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Full length article

Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas

Jie Bao[a,b], Pan Liu[a,b,*], Hao Yu[a,b], Chengcheng Xu[a,b]

[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

ARTICLE INFO

Keywords:
Big data
Human activity
Twitter
Safety
Spatial analysis

ABSTRACT

The primary objective of this study was to investigate how to incorporate human activity information in spatial analysis of crashes in urban areas using Twitter check-in data. This study used the data collected from the City of Los Angeles in the United States to illustrate the procedure. The following five types of data were collected: crash data, human activity data, traditional traffic exposure variables, road network attributes and social-demographic data. A web crawler by Python was developed to collect the venue type information from the Twitter check-in data automatically. The human activities were classified into seven categories by the obtained venue types. The collected data were aggregated into 896 Traffic Analysis Zones (TAZ). Geographically weighted regression (GWR) models were developed to establish a relationship between the crash counts reported in a TAZ and various contributing factors. Comparative analyses were conducted to compare the performance of GWR models which considered traditional traffic exposure variables only, Twitter-based human activity variables only, and both traditional traffic exposure and Twitter-based human activity variables. The model specification results suggested that human activity variables significantly affected the crash counts in a TAZ. The results of comparative analyses suggested that the models which considered both traditional traffic exposure and human activity variables had the best goodness-of-fit in terms of the highest R^2 and lowest AICc values. The finding seems to confirm the benefits of incorporating human activity information in spatial analysis of crashes using Twitter check-in data.

1. Introduction

During the past two decades, considerable efforts have been devoted to understanding the spatial pattern of crashes (Noland, 2003; Huang et al., 2010; Li et al., 2013; Rhee et al., 2016; Aguero-Valverde and Jovanis, 2006; Siddiqui et al., 2012; Abdel-Aty et al., 2013; Quddus, 2008; Tarko et al., 1996; Noland and Oh, 2004; Noland and Quddus, 2004; Hadayeghi et al., 2003; Ukkusuri et al., 2011; Naderan and Shahi, 2010). 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), and traffic analysis zones (TAZ) (Rhee et al., 2016; Aguero-Valverde and Jovanis, 2006; Siddiqui et al., 2012). The TAZ is a statistical entity delineated by the Departments of Transportation and metropolitan planning organizations for tabulating traffic-related census data, such as journey-to-work and place-of-work (Abdel-Aty et al., 2013). Traditionally, TAZ is a basic unit of analysis in urban transportation planning. In recent years, the spatial analysis of crash pattern based on TAZs has become more and more prevalent because researchers have come to believe that traffic safety should be considered an essential component of urban transportation planning (FHWA, 2005; NCHRP, 2010). In addition, with the spatial analysis results, transportation authorities can clearly identify the areas with greater-than-expected crashes and to apply proactive countermeasures to reduce crashes in these areas.

Numerous studies have investigated the factors contributing to crashes at spatially aggregated levels. Noland et al. collected 14-year crash data from fifty states in the United States to investigate how various road infrastructure improvements affected the number of traffic-related fatalities and injuries (Noland, 2003). They found that demographic changes in age cohort, increased seat-belt use, reduced alcohol consumption and improvements in medical technology significantly affected the reduction in fatalities and injuries at state level. Huang et al. (2010) conducted a county-level road safety analysis using data collected from 67 counties in the state of Florida in the United States. They used the Bayesian spatial model to establish a relationship between the crash counts reported in a county in each year and various contributing factors, such as aggregated road density, truck traffic volume and population density. The study suggested that young drivers,

* 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), seudarwin@gmail.com (H. Yu), iamxcc1@gmail.com (C. Xu).

http://dx.doi.org/10.1016/j.aap.2017.06.012
Received 14 December 2016; Received in revised form 15 June 2017; Accepted 16 June 2017

deprived areas with low income and high truck traffic volume were more likely to be involved with crashes (Huang et al., 2010). Rhee et al. (2016) conducted a spatial regression analysis of crash data reported at 423 TAZs in Seoul. They applied geographically weighted regression (GWR) analysis to model the crash data with a set of explanatory variables. The result showed that the number of subway stations and bus stops, the length ratio of a central bus-only lane and the number of assess points had significant effects on crash counts and severity in a TAZ (Rhee et al., 2016).

A set of factors may influence the crash counts reported in different spatial units. The influence factors can be classified into three general categories: traffic exposure (Aguero-Valverde and Jovanis, 2006; Quddus, 2008; Tarko et al., 1996), road network attributes (Noland and Oh, 2004; Noland and Quddus, 2004) and socio-demographic characteristics (Rhee et al., 2016; Hadayeghi et al., 2003; Ukkusuri et al., 2011). Traffic exposure is probably the most important factor for modeling spatially aggregated crash data. So far, the most commonly used traffic exposure variables for modeling spatially aggregated crash data include the average annual daily traffic (AADT) and the daily vehicle miles travelled (DVMT) (Aguero-Valverde and Jovanis, 2006; Siddiqui et al., 2012; Abdel-Aty et al., 2013). In fact, both AADT and DVMT are surrogate measures for traffic exposure, and do not directly measure detailed human activities, such as trip length and trip purposes. More recent studies have started considering the number of trips made in a TAZ for modeling spatially aggregated crash data. The number of trips with different purposes was estimated using trip generation models and household travel survey data. Previous studies have suggested that the total number of trips was directly related to crash counts in a TAZ, and the trips with different purposes may influence both crash counts and crash severity in different ways (Siddiqui et al., 2012; Naderan and Shahi, 2010).

The booming use of ubiquitous mobile computing has led to a dramatic increase in big data resources that capture the real-time movements of vehicles and individuals. 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 to crashes. Unlike traditional data sources which are mainly composed of the data obtained from the physical world, big data may provide useful information about human activities, which has been considered the most important category of contributing factors to crashes. In fact, the information about human activities has been largely ignored by previous traffic safety studies, mainly because such data are usually not available.

Recent studies have started using social media data to investigate the pattern of human activities in urban areas (Lenormand et al., 2014; Hasan and Ukkusuri, 2014; Luo et al., 2016; Cheng et al., 2011; Lee et al., 2016). Social media represents a set of online location-aware applications that allow users to create and exchange contents. The information that can be collected from social media includes individual activities, personal interests and public events. Previous studies have suggested that a strong positive correlation exists between AADT and the tweets posted on roads (Lenormand et al., 2014). In addition, the origin-destination matrix inferred from Twitter was closely related to that estimated using traditional travel survey data (Lee et al., 2016).

Compared with household travel survey data, the human activity data collected from social media have several advantages: (a) most social media tools, such as Twitter and Facebook, allow users to check in their current locations. Activity purposes can be classified by the location category of these check-ins; (b) the multiday check-ins can further be applied to analyze individuals' mobility, which is closely related to individuals' daily travel pattern and demographic information; and (c) the social media data can be collected nearly instantaneously with large samples at low cost, while the household travel survey is usually time consuming with low response efficiency and accuracy (Ma et al., 2013).

The primary objective of this study is to investigate how to incorporate Twitter-based human activity information in spatial analysis of crashes in urban areas. More specifically, this paper sought answers to the following questions: (a) how to gather human activity information from Twitter check-in data; (b) how to incorporate human activity variables in spatial analysis of crash data; (c) whether the incorporation of human activity variables estimated from Twitter check-in data improves the performance of crash models at TAZ-level; and (d) how different types of human activities contribute to crashes with different collision types at TAZ-level. The rest of the paper is organized as follows. Section 2 discusses the procedures for gathering various types of data from different sources. Section 3 describes the methodology of this paper. The results of data analysis are presented in Section 4. Section 5 is a summarization of this study.

2. Data sources

This study used the data collected from the City of Los Angeles in the United States to illustrate the procedure for incorporating human-activity data in spatial analysis of crashes. The City of Los Angeles has been at the forefront of implementing the open data policy. Numerous types of data sources, such as real-time traffic flow data, census data for transportation and crash data, are publicly available, providing big data researchers with great convenience and opportunities. The study period was from February 2010 to January 2011. In the present study, TAZ was considered the basic spatial unit of analysis. There are 896 TAZs in the City of Los Angeles (See Fig. 1). The shape file for TAZ boundaries was obtained from the TIGER files of U.S. Census Bureau. The following five types of data were collected: crash data, human activity data, traditional traffic exposure variables, road network attributes, and social-demographic data. The data were collected from different sources. More specifically, crash data were collected from the Statewide Integrated Traffic Records System (SWITRS) which was maintained by the California Department of Transportation (Caltrans). The information obtained from crash records included the date, time, crash severity, collision type, and the geo-location. In total, 21,673 crashes were reported in the City of Los Angeles during the study period.

The road network data were collected from Caltrans and Los Angeles County websites. The highway and freeway network data were collected from the Caltrans's GIS data library. The local street and road network were collected from the GIS data portal of the Los Angeles County. The social-demographic data for the study area were obtained from the AASHTO's CTTP data library. The information obtained from the social-demographic data set included the number of population segregated by ethnicity and age cohorts, poverty level, the number of population with a bachelor's degree, median household income, the number of unemployment population, and average travel time to work.

Four traditional traffic exposure variables were collected. The annual average daily traffic (AADT) and truck AADT on freeways were collected from the GIS data library of the Caltrans, which provided the GIS format files of all count locations on the California freeway network. The number of trip productions and attractions made by automobiles in each TAZ were collected from the census transportation planning products (CTTP) maintained by the American Association of State Highway and Transportation Officials (AASHTO).

The human activity data were collected from the large-scale check-in data of Twitter. Twitter is a widely used social media application, where individuals can post short messages, namely tweets for sharing common ideas, activities, or interests (Boyd and Ellison, 2007). Recently, Twitter has become a prevalent tool for people in urban areas, generating nearly 135,000 new users every day. Tweets are labelled with geo-location information after getting users' approval. In addition, Twitter also allows users to post status from third-party check-in services, such as Foursquare. When the users of Foursquare check-in to a place, the status will be shared to their Twitter pages. A typical tweet