Accident Analysis and Prevention 121 (2018) 223–230

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Traffic crash analysis with point-of-interest spatial clustering

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

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

ARTICLE INFO

Keywords:
Spatial clustering
Kernel density estimation
Traffic crashes
Spatial regression

ABSTRACT

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

1. Introduction

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

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

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

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

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

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

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

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

2. Literature review

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

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

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

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

3. Data description

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