Accident Analysis and Prevention 102 (2017) 213–226

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Intersection crash prediction modeling with macro-level data from various geographic units

Jaeyoung Lee*, Mohamed Abdel-Aty, Qing Cai

*Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816-2450, United States*

ARTICLE INFO

*Article history:*
Received 2 December 2016
Received in revised form 15 February 2017
Accepted 11 March 2017

*Keywords:*
Crash prediction model
Micro-level traffic safety
Macro-level traffic safety
Zonal effect
Random-effects model
Safety performance function

ABSTRACT

There have been great efforts to develop traffic crash prediction models for various types of facilities. The crash models have played a key role to identify crash hotspots and evaluate safety countermeasures. In recent, many macro-level crash prediction models have been developed to incorporate highway safety considerations in the long-term transportation planning process. Although the numerous macro-level studies have found that a variety of demographic and socioeconomic zonal characteristics have substantial effects on traffic safety, few studies have attempted to coalesce micro-level with macro-level data from existing geographic units for estimating crash models. In this study, the authors have developed a series of intersection crash models for total, severe, pedestrian, and bicycle crashes with macro-level data for seven spatial units. The study revealed that the total, severe, and bicycle crash models with ZIP-code tabulation area data performs the best, and the pedestrian crash models with census tract-based data outperforms the competing models. Furthermore, it was uncovered that intersection crash models can be drastically improved by only including random-effects for macro-level entities. Besides, the intersection crash models are even further enhanced by including other macro-level variables. Lastly, the pedestrian and bicycle crash modeling results imply that several macro-level variables (e.g., population density, proportions of specific age group, commuters who walk, or commuters using bicycle, etc.) can be a good surrogate exposure for those crashes.



## 1. Introduction

Numerous traffic crash prediction models have been developed for different facilities for various crash types. Highway Safety Manual (HSM) (AASHTO, 2010) provides a range of segment and intersection-based crash models (also known as safety performance functions) for many facility types including but not limited to rural two-lane two-way roads, rural multilane highways, and urban and suburban arterials. According to the HSM (AASHTO, 2010), the crash models play a key role in identifying crash hotspots (i.e., network screening), and evaluating safety countermeasures using the empirical Bayes method. Recently, many studies have been conducted to adopt the screening method using the crash models in multiple jurisdictions in the United States including but not limited to Alabama (Turner et al., 2012), Florida (Abdel-Aty et al., 2014; Abdel-Aty et al., 2016), Kansas (Schrock, 2011), Louisiana (Sun et al., 2011), Oregon (Dixon and Monsere, 2011), and Utah (UDOT, 2011). Also, there are some efforts to apply hotspot screening using crash prediction model in other countries including Canada (Persaud et al., 2012), Australia (Wemple et al., 2010), and Italy (Cafiso et al., 2012). A majority of the crash prediction models have been built at the micro-level, such as intersection, segment, or corridor levels. On the other hand, some researchers have estimated crash prediction models at the macro-level (e.g., traffic analysis zones) to incorporate highway safety considerations in the long-term transportation planning process.

Although the numerous macro-level studies have found that a variety of demographic and socioeconomic zonal characteristics have substantial effects on traffic safety, few studies have attempted to coalesce micro-level with macro-level data for estimating crash models. Abdel-Aty et al. (2016) and Lee (2014) proposed a methodology to integrate macro-level and micro-level data to provide a comprehensive perspective by balancing the two-levels. Still, their methodology is based on the macro-level crash models. Park et al. (2015) estimated segment-level crash models to evaluate the effectiveness of bicycle facilities. The authors included block-group based macro-level data including population density and income and found that they are statistically significant in the segment-level crash models. Recently, Huang et al. (2016) estimated crash prediction models separately at the micro-

* Corresponding author.
*E-mail address:* Jaeyoung.Lee@ucf.edu (J. Lee).

http://dx.doi.org/10.1016/j.aap.2017.03.009

level and macro-level and compared the model performance. The results indicated that the micro-level model has a better fit and performance. The authors claimed that the micro-level approach is able to provide better insights on microscopic factors that directly contributes to traffic crashes while the macro-level approach is beneficial when monitoring regional safety and relating it with socio-demographic factors.

Huang and Abdel-Aty (2010) discussed the multi-level data in traffic safety. The multi-level data includes occupant, driver/vehicle, crash, site, and geographic region and the extra temporal dimension. The authors suggested many ideas to explore crashes at the multi-level. For instance, analyzing traffic crash counts 1) at intersection and time-level; 2) at county and corridor-level; 3) at county level with spatial effect; and so on. Guo et al. (2010) developed several crash prediction models for signalized intersections with corridor-level spatial correlation. The authors found that the Poisson spatial model with the corridor-level spatial effects provides the best model fit. This study is inspired by the studies by Huang and Abdel-Aty (2010) and Guo et al. (2010), but is different as it applies macro-level variables and random-effects along with micro-level variables to developed micro-level crash models.

Several researchers explored the effects of geographic units on crash modeling at the macro-level. Abdel-Aty et al. (2013) investigated the effect of different zonal systems. The authors compared crash models based on three different areal units: BGs (block groups), CTs (census tracts) and TAZs. The authors discovered that the BG based model had the larger number of significant variables for total and severe crashes compared to models based on other geographical units. Lee (2014) developed TSAZ (traffic safety analysis zones) by aggregating existing TAZs with comparable crash characteristics, and compared TAZ-based and TSAZ-based models and claimed that the TSAZ-based model outperforms the TAZ model in terms of goodness-of-fit. The authors argued that if a zone size is small it cannot capture global crash patterns; on the contrary, we may lose many local features if the zone is too large. Similarly, it is necessary to find the data from the optimal sized spatial unit that can provide the best modeling results for intersection crash models (Xu et al., 2014, 2016).

There have been several efforts to attempt macro-level factors in micro-level crash prediction modeling. Mitra and Washington (2012) investigated the role of various candidate variables other than annual average daily traffic (AADT). Some of the variables were collected from the area near intersections such as the presence of school by type, the number of pubs, weather, total population, and population by age group within a specific range from the intersection. Among the variables, the presence of college within half mile, the number of bars within quarter mile, population between age 0 and 15 (as a random parameter), population between age 16 and 64, average annual precipitation, and average annual number of rainy days were found significant. The authors compared the model with traffic parameters only with the full model, and found that the variable exclusion overstates the effect of minor AADT by 40% and major AADT by 14%. Thus, the authors concluded that the exclusion of key variables caused omitted variable bias in modeling. Wang and Huang (2016) related crash counts of road network to roadway and TAZ variables. The authors developed a Bayesian hierarchical joint model and found the relationship between road network crash risk and micro-level variables (i.e., traffic volume) along with macro-level variables (i.e., socioeconomic, trip generation, and network density variables). Wang et al. (2017) investigated the effects of zonal factors associated with crash occurrence on intersections by different transportation modes: motor vehicle, bicycle, and pedestrian. The authors revealed several important findings: (1) the significant variable sets differ by transportation mode; (2) the omission of zonal variables resulted in biased parameters; (3) zonal factors played a more important role for bicycle and pedestrian crashes; and (4) a smaller buffer size to extract zonal factors resulted in better estimations. In the study of Park et al. (2015), the authors estimated crash modification functions for bike lanes using a before-and-after study with empirical Bayes and cross-sectional methods. The authors developed safety performance functions with some socioeconomic variables and found that both population density and median household income have a significant effect on bicycle crashes on segments.

Although there are several studies suggesting ideas to link macro-level and micro-level data, no studies have tried to analyze the effects of macro-level variables from existing geographic units on micro-level crash models. Also, it is worth to investigate which geographic unit provides the optimal data for micro-level crash prediction models. Therefore, this paper aims at answering the three research questions: (1) Can intersection crash prediction models be improved by considering macro-level geographic units? (2) What would be the best spatial unit for the crash prediction models? and (3) what macro-level factors do have significant effects on intersection crashes?

## 2. Methodology

Generally, there are two methods to obtain macro-level factors. First method is to collect zonal factors from existing geographic units (e.g., CT, TAZ, county). It assumes that zonal factors aggregated in a zone have influences on the intersection crash counts within the zone. This approach has two possible issues: (1) modeling results may be largely affected of the selection of geographic units. Thus, the authors aim at finding the best geographic units for intersection crash prediction modeling by crash type in this study; and (2) some intersections may be located in zone boundaries. In this case, the intersections may be simultaneously affected by the factors of the multiple adjacent zones. Table 1 summarizes the number of intersections located within 50 feet from zone boundaries.

The percentage of intersections near zone boundaries is the highest in TAZs (85%). This may be because one of the zoning criteria for TAZs is to recognize physical boundaries such as arterials (Lee et al., 2014b). In general, the percentage of intersections near zone boundaries goes down as the zone size increases. In the case of CCD and County, only 7.8% and 0.6% of intersections were located near zone boundaries, respectively.

Another method to collect zonal factors is creating a buffer surrounding the intersections, and extracting zonal data from the buffers. The extracted data from the buffer zone are then incorporated in to micro-level crash prediction models. This method also has an issue that the determination of buffer size may be subjective and arbitrary. In this study, the former approach was adopted since the latter one has already been attempted by Miranda-Moreno et al. (2011), Pulugurtha and Sambhara (2011), Mitra and Washington (2012), and Shah et al. (2017).

**Table 1**
The number of intersections located near zone boundaries and the percentages.

| Geographic Units | No of intersections near zone boundaries | No of intersections not near zone boundaries | % of intersections near zone boundaries |
|---|---|---|---|
| Block Group (BG) | 6272 | 2075 | 75.1% |
| Traffic Analysis Zone (TAZ) | 7111 | 1236 | 85.2% |
| Census Tract (CT) | 4687 | 3660 | 56.2% |
| ZIP-Code Tabulation Area (ZCTA) | 1789 | 6558 | 21.4% |
| Traffic Analysis District (TAD) | 2152 | 6195 | 25.8% |
| Census County Division (CCD) | 648 | 7699 | 7.8% |
| County | 51 | 8296 | 0.6% |