Accident Analysis and Prevention 101 (2017) 37–43

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A Hybrid Latent Class Analysis Modeling Approach to Analyze Urban Expressway Crash Risk

Rongjie Yu[a,b], Xuesong Wang[a,b,*], Mohamed Abdel-Aty[c]

[a] Road and Traffic Key Laboratory, Ministry of Education, Shanghai 201804, China
[b] College of Transportation Engineering, Tongji University,4800 Cao'an Road, Shanghai 201804, China
[c] Department of Civil, Environmental and Construction Engineering, University of Central Florida Orlando, FL 32826-2450, United States

ARTICLE INFO

Article history:
Received 17 February 2016
Received in revised form
14 November 2016
Accepted 3 February 2017

Keywords:
Crash risk analysis
Latent class analysis
Bayesian random parameter model
Unobserved heterogeneity

ABSTRACT

Crash risk analysis is rising as a hot research topic as it could reveal the relationships between traffic flow characteristics and crash occurrence risk, which is beneficial to understand crash mechanisms which would further refine the design of Active Traffic Management System (ATMS). However, the majority of the current crash risk analysis studies have ignored the impact of geometric characteristics on crash risk estimation while recent studies proved that crash occurrence risk was affected by the various alignment features. In this study, a hybrid Latent Class Analysis (LCA) modeling approach was proposed to account for the heterogeneous effects of geometric characteristics. Crashes were first segmented into homogenous subgroups, where the optimal number of latent classes was identified based on bootstrap likelihood ratio tests. Then, separate crash risk analysis models were developed using Bayesian random parameter logistic regression technique; data from Shanghai urban expressway system were employed to conduct the empirical study. Different crash risk contributing factors were unveiled by the hybrid LCA approach and better model goodness-of-fit was obtained while comparing to an overall total crash model. Finally, benefits of the proposed hybrid LCA approach were discussed.

1. INTRODUCTION

Crash risk analyses have emerged in recent years as the development of traffic sensing technology and their potential usage in Active Traffic Management System (ATMS) (Oh et al., 2001). Crash risk analysis differs from developing safety performance functions (SPFs) as it was carried out to identify crash-prone traffic statuses (Roshandel et al., 2015). To reach this goal, a conventional approach is to extract traffic flow data for both crash cases and non-crash cases using the "matched-case control" data structure (Abdel-Aty et al., 2004); whereas crash risk analysis models were developed under the assumption that the same traffic statuses would hold identical crash probability for different roadway sections. Geometric characteristics' effects on crash occurrence risk have been neglected.

However, a previous study (Pande and Abdel-Aty, 2006) revealed that geometric characteristics have significant impact on rear-end crash risk in addition to the traffic flow parameters; and the unobserved heterogeneity brought by geometric characteristics need to be addressed in crash risk analyses (Yu and Abdel-Aty, 2013a). In addition, based on a meta-analysis study, Roshandel et al. (2015) claimed that the location confounding effects have substantial impact on the relationships between traffic flow characteristics and crash occurrence. Failing to account for these unobserved heterogeneity would lead to biased estimated parameters and incorrect inferences (Mannering and Bhat, 2013). Therefore, an issue of interest arises here is how to address the heterogeneous effects of geometric characteristics when analyzing crash risk based on traffic data.

In this study, a hybrid modeling approach that combines latent class analysis (LCA) and Bayesian random parameter logistic regression model was proposed. The LCA technique was used to segment crash data into homogenous subgroups based on geometric characteristics while Bayesian random parameter logistic regression models were utilized to develop crash risk analysis models for each subgroup of crash data.

* Corresponding Author November 2016 Revision submitted for possible publication in Accident Analysis and Prevention.
E-mail address: wangxs@tongji.edu.cn (X. Wang).

http://dx.doi.org/10.1016/j.aap.2017.02.002
0001-4575/

2. BACKGROUND

2.1. Latent Class Analysis

Data heterogeneity has been recognized as a key issue that needs to be addressed in transportation studies. Researchers attempted to resolve it through segmenting the data based on expert domain knowledge; however, homogenous subset datasets could not be guaranteed (Depaire et al., 2008). Therefore, in order to reduce the unobserved heterogeneity more effectively, clustering and latent class analysis techniques were introduced. Bhat (1997) first used an endogenous segmentation approach to analyze intercity travel mode choice. Individuals that have identical preferences and sensitivities to level-of-service variables were grouped together. Through comparing the proposed analysis method to traditional analysis methods, it was concluded that the endogenous segmentation model provided better fit and more reasonable results.

Similar analyses approaches were also adopted in traffic safety research. Depaire et al. (2008) used a latent class clustering technique to identify homogenous crash types; the analysis results indicated that through the clustering analysis, hidden relationships within the crash data were revealed. Similarly, de Oña et al. (2013) employed latent class clustering to segment the crash data; comparisons between the cluster-based analyses and full-data analysis concluded that it is beneficial to conduct the segmentation before analyzing the dataset. Recently, Shaheed and Gkritza (2014) estimated a latent class multinomial logit model to deal with the unobserved heterogeneity that lies in crash injury severity analysis. In addition to the abovementioned studies that used latent class analysis techniques to segment crash data, Ma and Kockelman (2006) and Eluru et al. (2012) employed latent segmentation model to classify roadway sections, while Ng et al. (2002) clustered homogenous traffic analysis zones. In this study, latent class analysis technique was employed to obtain crashes that share identical impacts from geometric characteristics.

2.2. Random Parameter Logit Models

Random parameter logit model (or called mixed logit model) was introduced to traffic safety analysis with its advantage of accounting for data heterogeneity effects (Train, 2009). Firstly, random parameter logit model was utilized to analyze crash injury severity for the purpose of allowing certain variables to vary across different roadway segments (Milton et al., 2008). Then the modeling technique became popular and was employed in several crash injury severity studies (Anastasopoulos and Mannering, 2011; Savolainen et al., 2011; Kim et al., 2012). Recently, random parameter logistic regression models have been employed in crash risk analysis to deal with the distinct traffic status' effects (Xu et al., 2014) and the heterogeneity effects at the crash unit level (Yu and Abdel-Aty, 2013b).

2.3. Hybrid Analysis Models with Bayesian Inference

Recently, several studies have proposed hybrid analysis models with Bayesian inference technique to benefit from the advantages of multiple distinct techniques. Liang and Lee (2014) proposed a hybrid algorithm that integrated two data mining methods—Dynamic Bayesian Network (DBN) and supervised clustering—to analyze driver cognitive distraction with eye movement and driving performance measures. While Chen et al. (2015) proposed a multinomial logit model and Bayesian network hybrid approach to analyze driver injury severity, where the multinomial logit model was used to identify significant contributing factors and the Bayesian network was employed to establish the statistical associations. In addition, a hybrid classifier of combining Decision Table (DT) and Naïve Bayes (NB) was utilized to predict driver injury severity in rear-end crashes (Chen et al., 2016).

3. DATA PREPARATION

The proposed hybrid modeling approach was tested using data from Shanghai's urban expressway system. A total of three datasets were used: (1) crash data in September, 2013; (2) roadway section geometric characteristics data; and (3) roadway section traffic data collected by loop detectors.

The Shanghai urban expressway system was split into 206 roadway sections using on-ramps and off-ramps as dividing points. And the obtained loop detector traffic data were aggregated into 2-minute interval data based on roadway sections. Roadway section geometric characteristics were obtained from on-line street-view

Table 1
Summary statistics of roadway section geometric characteristics.

Variable	Description	Summary Statistics
Length	Roadway section length	Mean: 944.5 (m) Standard Deviation: 585.8 (m)
Lane Number	Number of lanes	2: 59 (count number) 3: 59 4: 70 5: 18
Ramp type	Ramp combination type: 1. On-ramp and On-ramp 2. On-ramp and Off-ramp 3. Off-ramp and On-ramp 4. Off-ramp and Off-ramp	1: 79 2: 21 3: 71 4: 35
Speed Limit Sign	If there is speed limit sign within the roadway section: 1. Has speed limit sign 0. No speed limit sign	1: 83 0: 123
Curve	If it is a curve section: 1. Curve section 0. Straight section	1: 96 0: 110
Speed Limit Value	Speed limit for the roadway section	50 (km/h): 6 60 (km/h): 71 80 (km/h): 129
Upstream Auxiliary Lane	Access lane numbers for upstream ramp	1: 26 2: 174 3: 6
Downstream Auxiliary Lane	Access lane numbers for downstream ramp	1: 26 2: 173 3: 7