Journal of Safety Research xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Journal of Safety Research

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

Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data

Feng Chen, [a] Suren Chen, [b,*] Xiaoxiang Ma [a]

[a] College of Traffic Engineering and Key Laboratory of Road & Traffic Engineering of the Ministry of Education, Tongji University, 4800 Cao'an Road, Shanghai 201804, China
[b] Department of Civil & Environmental Engineering, Colorado State University, Fort Collins, CO 80523, United States

ARTICLE INFO

Article history:
Received 2 May 2017
Received in revised form 8 December 2017
Accepted 27 February 2018
Available online xxxx

Keywords:
Real-time driving environment
Mixed logit model
Refined temporal scale
Random parameter
Big data

ABSTRACT

Driving environment, including road surface conditions and traffic states, often changes over time and influences crash probability considerably. It becomes stretched for traditional crash frequency models developed in large temporal scales to capture the time-varying characteristics of these factors, which may cause substantial loss of critical driving environmental information on crash prediction. Crash prediction models with refined temporal data (hourly records) are developed to characterize the time-varying nature of these contributing factors. Unbalanced panel data mixed logit models are developed to analyze hourly crash likelihood of highway segments. The refined temporal driving environmental data, including road surface and traffic condition, obtained from the Road Weather Information System (RWIS), are incorporated into the models. Model estimation results indicate that the traffic speed, traffic volume, curvature and chemically wet road surface indicator are better modeled as random parameters. The estimation results of the mixed logit models based on unbalanced panel data show that there are a number of factors related to crash likelihood on I-25. Specifically, weekend indicator, November indicator, low speed limit and long remaining service life of rutting indicator are found to increase crash likelihood, while 5-am indicator and Number of merging ramps per lane per mile are found to decrease crash likelihood. The study underscores and confirms the unique and significant impacts on crash imposed by the real-time weather, road surface and traffic conditions. With the unbalanced panel data structure, the rich information from real-time driving environmental big data can be well incorporated.

1. Introduction

Traffic crashes under adverse driving environments cause a lot of social and economic loss in most countries. To develop various prevention strategies, it is critical to first understand the impact of contributing factors on crash risk. Most traditional crash frequency studies, however, were conducted over extended temporal units with aggregated information (e.g., yearly, monthly). Traffic safety studies that focus on time-varying driving environmental data in fine temporal units (e.g., hourly or daily) are still rare. Real-time driving environmental data including weather conditions and traffic characteristics may have great influence on the crash occurrence, especially for some adverse driving conditions where weather may vary drastically over time.

As a result of adopting extended time scales, it is obvious that some crucial time-varying driving environmental information, such as weather and traffic data, is therefore lost due to data aggregation (Lord & Mannering, 2010; Mannering & Bhat, 2014). Besides, the importance of certain time-varying explanatory environmental variables might not be discovered unless data in more refined temporal scales are adopted in the model, resulting in ecological fallacy (Freedman, 1999). It becomes even crucial for traffic facilities that undergo substantial variations regarding driving environments (e.g., inclement weather in mountainous areas, frequent traffic state transformation in urban areas). Moreover, the crash frequency prediction models developed based on averaged or cumulative data over extended time periods may result in estimation error due to unobserved effects (Mannering & Bhat, 2014; Mannering, Shankar, & Bhat, 2016; Washington, Karlaftis, & Mannering, 2011).

As ITS applications become more popular around the world, real-time driving environmental records collected continuously become more obtainable in many major transportation systems. These driving environmental big data bring rich information and also great opportunities for carrying out more advanced crash prediction than ever. Many researchers have endeavored to develop crash prediction models with the detailed monitoring data, primarily focusing on real-time relative risk or likelihood of crashes mostly based on the case–control data structure, which may not sufficiently utilize the abundant information that the real-time big data can offer.

* Corresponding author.
E-mail addresses: fengchen@tongji.edu.cn, (F. Chen), suren.chen@colostate.edu, (S. Chen), xiaoxiang.ma@tongji.edu.cn. (X. Ma).

https://doi.org/10.1016/j.jsr.2018.02.010
0022-4375/

When adopting crash prediction models with refined temporal data, it is methodologically challenging to develop appropriate models for driving environmental data with both time-varying and spatial-varying information. Multiple observations are processed for the same road segment by using more refined temporal units. These multiple observations over the same roadway unit would be somehow correlated with each other by sharing the same geographical location, setting up serial correlations within the data (Mannering & Bhat, 2014). These potential serial correlations bring methodological challenges in building proper crash models. The present study focuses on developing crash likelihood prediction models considering driving environmental big data with refined temporal scales and adopting disaggregated panel-data structure. Mixed logit models, which can consider the random nature of some parameters, are developed using panel data to deal with temporal correlation in the present study. This study explores different types of contributing factors including real-time driving environmental characteristics comprehensively. Crash data on highway I-25 in Colorado will be analyzed to provide some valuable findings of contributing factors, especially time-varying variables.

1.1. Real-time crash models

In the last few years, there have been many studies primarily focusing on calibrating real-time crash risk models that study the relative crash risk with real-time traffic and environmental conditions prior to crashes (e.g., Abdel-Aty & Pande, 2005; Abdel-Aty, Pande, Lee, Gayah, & Santos, 2007; Abdel-Aty, Uddin, Pande, Abdalla, & Hsia, 2004; Ahmed & Abdel-Aty, 2012; Chen, Ma, & Chen, 2014; Golob & Recker, 2003, 2004; Golob, Recker, & Pavlis, 2008; Hassan & Abdel-Aty, 2013; Lee, Hellinga, & Saccomanno, 2003; Lee, Saccomanno, & Hellinga, 2002; Shi, Abdel-Aty, & Yu, 2016; Xu, Wang, & Liu, 2013a, 2013b; Yu & Abdel-Aty, 2013a, 2013b; Yu, Abdel-Aty, & Ahmed, 2013). In the majority of these studies, the relative crash probability was often analyzed by comparing conditions with and without crashes, rather than direct crash likelihood (e.g., Yu & Abdel-Aty, 2013b). Most of these crash probability studies adopted the matched case–control design (e.g., Abdel-Aty, Hassan, Ahmed, & Al-Ghamdi, 2012; Ahmed & Abdel-Aty, 2012; Xu et al., 2013a, 2013b; Yu & Abdel-Aty, 2013b), in which a preselected number (e.g., four) of non-crash cases were produced to match each specific crash case. In the studies summarized above, the data structure was established on the base of case–control crash records, rather than the rich information of full driving environmental data containing varying information in spatial and time domains for road segments. Important factors such as location and geometry were matched out to enable observation control. Moreover, selection bias can become a serious problem for case–control studies (Hernan, Hernandez-Diaz, & Robins, 2004; Paik, 2004). Therefore, unlike these existing studies, the present study develops direct crash likelihood models for road segments using driving environmental big data, which can take advantage of the entire informative panel-data without data selection.

1.2. Panel data crash frequency models

Crash frequency prediction model is a fundamental tool to analyze crash risks on highways by directly quantifying crash counts. The basic models include Poisson and Negative Binomial models. Panel data models have frequently been used for spatial and temporal varying data while still considering the heterogeneity of individual observations in social science. Owing to the cross-sectional and time-serial characteristics of some crash data, crash frequency analysis in the last decade or so has utilized panel data models, including but not limited to random effects Poisson models and Negative Binomial models. For example, Noland (2003) and Noland and Oh (2004) developed the fixed effects Negative Binomial models to study fatal and injured traffic crash frequency the influence of renovation on roadway infrastructure. To deal with the limitation of fixed effects Poisson or Negative Binomial models including its inability to consider time-specific or site-specific variations, random effects Negative Binomial models can be developed (Shankar, Albin, Milton, & Mannering, 1998). In addition, other random effect or random parameter crash frequency models were also explored (e.g., Aguero-Valverde, 2013; Anastasopoulos & Mannering, 2009; Chin & Quddus, 2003; Kweon & Kockelmam, 2005; Miaou, Song, & Mallick, 2003). For example, Anastasopoulos and Mannering (2009) predicted annual crash frequency using a random parameter Negative Binomial model with 9-year data. These panel data crash frequency studies mainly focused on modeling longitudinal data resulted from yearly repeated observations (multi-year crash frequency), thus are unable to capture the effects of those contributing factors that vary within a year. For instance, when it comes to traffic flow and weather information, these crash studies usually formulate long-term aggregated and/or averaged variables to represent their effects, such as annual average daily traffic volume and number of days with rainfall over a year (e.g., Aguero-Valverde & Jovanis, 2006).

When a smaller temporal unit is used, the resulting crash dataset is inevitably characterized with excess zeros, which bring about another methodological difficulty. The excessive zeroes of the records need to be taken care of for a refined-scale panel data model to be properly established. In light of that, zero-inflated Poisson and Zero-inflated Negative Binomial models were adopted in some studies (e.g., Anjana & Anjaneyulu, 2015; Miaou, 1994), which are extensions of standard Poisson and Negative Binomial regression models. Note that these zero-inflated models also face criticism from some researchers despite the fact that they usually perform better than corresponding standard models (e.g., Lord, Washington, & Ivan, 2005, 2007; Vangala, Lord, & Geedipally, 2015). Random effect or random parameter zero-inflated models were also attempted to analyze annual crash frequency (Huang & Chin, 2010) using multi-year data. However, to the authors' knowledge, studies that investigate panel-data crash likelihood models with refined temporal scales are still scarce.

1.3. Discrete outcome models

Over the past decades, various discrete outcome models have been widely used to study crash injury severity due to the fact that different models bear different merits as well as limitations. Ordered logit and ordered probit models were applied to examine numerous risk contributing factors related to injury severity in previous studies (Abdel-Aty, 2003; Duncan, Khattak, & Council, 1998). Other studies investigated the application of Multinomial logit models (Islam & Mannering, 2006) and nested logit models (Chang & Mannering, 1999) to establish the relationship between different risk contributing factors and different injury severity levels. Despite the fact that multinomial logit model has been extensively employed in injury severity studies given its advantage over ordered probability models, it was found to suffer from irrelevant independence alternative (IIA) restriction (Jones & Hensher, 2007). To relax the IIA restriction and also account for unobserved heterogeneity, mixed logit models were proposed and then have been widely adopted in the studies on crash injury (e.g., Behnood & Mannering, 2015; Chen & Chen, 2011; Kim, Ulfarsson, Shankar, & Mannering, 2010; Ma, Chen, & Chen, 2015; Milton, Shankar, & Mannering, 2008). For example, Behnood and Mannering (2015) applied mixed logit model to study the temporal stability of factors affecting driver-injury severities in single-vehicle crashes.

For crash likelihood studies using discrete outcome models instead of crash injury severity modeling, Qi, Smith, and Guo (2007) have studied freeway crash likelihood using a random effect ordered probit model. Given the relative virtue as discussed above, mixed logit models will be adopted for the first time in the present study to investigate the hourly crash likelihood for road segments. In addition, random parameter models, rather than fixed effect models that have been commonly applied, will be adopted to account for unobserved heterogeneity. By