G Model
AAP 3845 No. of Pages 9

ARTICLE IN PRESS

Accident Analysis and Prevention xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Bootstrap resampling approach to disaggregate analysis of road crashes in Hong Kong

Xin Pei[a,*], N.N. Sze[b,1], S.C. Wong[c,2], Danya Yao[d,3]

[a] Department of Automation, Tsinghua University, Beijing, China
[b] Department of Civil and Natural Resources Engineering, University of Canterbury, Christchurch, New Zealand
[c] Department of Civil Engineering, The University of Hong Kong, Hong Kong, China
[d] Department of Automation, Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Received 11 November 2014
Received in revised form 30 April 2015
Accepted 6 June 2015
Available online xxx

Keywords:
Single-vehicle crash
Multiple-vehicle crash
Injury severity
Count data model
Bootstrap resampling method

ABSTRACT

Road safety affects health and development worldwide; thus, it is essential to examine the factors that influence crashes and injuries. As the relationships between crashes, crash severity, and possible risk factors can vary depending on the type of collision, we attempt to develop separate prediction models for different crash types (i.e., single- versus multi-vehicle crashes and slight injury versus killed and serious injury crashes). Taking advantage of the availability of crash and traffic data disaggregated by time and space, it is possible to identify the factors that may contribute to crash risks in Hong Kong, including traffic flow, road design, and weather conditions. To remove the effects of excess zeros on prediction performance in a highly disaggregated crash prediction model, a bootstrap resampling method is applied. The results indicate that more accurate and reliable parameter estimates, with reduced standard errors, can be obtained with the use of a bootstrap resampling method. Results revealed that factors including rainfall, geometric design, traffic control, and temporal variations all determined the crash risk and crash severity. This helps to shed light on the development of remedial engineering and traffic management and control measures.

1. Introduction

Road safety is a global issue linked to health and development. With more than 1.3 million road deaths and 50 million injuries occurring every year, road crashes are expected to become the fifth leading cause of death worldwide by 2030 (WHO, 2009). Fatalities and injuries result in losses of life and property and decreased quality of life. A better understanding of the factors contributing to road crashes, injuries, and deaths is critical to the development of appropriate road safety measures.

Considering the differences in crash circumstances and collision mechanisms, the factors contributing to, and their effects on injury severities of, single-vehicle (SV) and multi-vehicle (MV) crashes can be differentiated. Therefore, the prediction performances of separate crash prediction models for different collision types are superior to that of a combined crash prediction model. Separate crash prediction models for SV and MV crashes also have the capability to reveal the distinctive relationships between risks of crash and various contributory factors (Mensah and Hauer, 1998). In particular, the associations between crash frequencies and possible risk factors, such as geometric design, weather, seasonal variations (Shankar et al., 1995), and day and night conditions on rural roads (Persaud and Mucsi, 1995), for both SV and MV crashes have been revealed.

The effects of possible risk factors on injury severity of SV and MV crashes can be differentiated. In particular, the different risk factors associated with fatality risks of SV and MV crashes in Northern Sweden have been significantly distinguished (Öström and Eriksson, 1993). Differences in the associations between injury risk and possible factors have also been identified for truck drivers on rural roads (Chen and Chen, 2011), motorways (Bham et al., 2012; Qu et al., 2014; Kuang et al., 2015), and urban roadways (Yau, 2004; Yau et al., 2006; Jung et al., 2010, 2012; Fréchède et al., 2011; Xie et al., 2011; Kim et al., 2013). In this study, we attempt to identify the differences in the relationship between possible factors and risk of crashes of different types with respect to collision types (i.e., SV versus MV crashes) and crash severity (i.e., killed and severe injury versus slight injury crashes). Due to

* Corresponding author. Fax: +86 10 6279 5043.
E-mail addresses: peixin@mail.tsinghua.edu.cn, happyqiezi@gmail.com (X. Pei), tony.sze@canterbury.ac.nz (N.N. Sze), hhecwsc@hku.hk (S.C. Wong), yaody@mail.tsinghua.edu.cn (D. Yao).
1 Fax: +64 3364 2758.
2 Fax: +852 2559 5337.
3 Fax: +86 10 62786911.

http://dx.doi.org/10.1016/j.aap.2015.06.007

their non-negative and random nature, count data models, including Poisson regression and negative binomial regression models, have long been used to model crash frequencies (Ivan et al., 1999; Ivan, 2004). Advanced modeling approaches have been developed to cope with the complicated natures of crash distributions, residual distributions, variations in parameter estimates, heterogeneous effects of risk factors, and the effects of different exposure measures (Qin et al., 2004, 2006; Geedipally and Lord, 2010; Lord and Mannering, 2010; Xiong and Mannering, 2013).

Crash frequencies are often aggregated to daily, monthly, or yearly levels. However, aggregate crash prediction models may suffer from the ecological fallacy problem (Robinson, 1950; Golob et al., 2004), in which inferences of individual attributes may be blurred after aggregation. In contrast, the disaggregated approach using a finer sample can reduce the degree of ecological fallacy (Sullivan, 1990; Abdel-Aty and Pande, 2007). To remove the above concern and to take advantage of the availability of hourly-based crash, traffic, and weather data on 112 urban roadway segments that are evenly distributed in the Hong Kong territory over a 5-year period from 2002 to 2006, we develop disaggregated crash prediction models to measure the association between possible factors, such as road geometrics, traffic control, temporal variation, and weather, and risks of SV and MV crashes in Hong Kong. In particular, vehicle kilometer (VKM) is used as a proxy of exposure in the proposed crash prediction models (Pei et al., 2012). However, the predominance of zero counts in such a disaggregated crash prediction model based on hourly crash data is of concern, the estimates may be biased.

The problem related to the issue of excess zeros in a traditional Poisson process was recognized in previous research (Shankar et al., 1997; Washington et al., 2011). To address this issue in crash counts, alternate model formulations, such as zero-inflated count data models (Miaou, 1994; Lee and Mannering, 2002; Shankar et al., 2003; Huang et al., 2008) were proposed and adopted for crash prediction modeling. These models assumed that zero counts were derived from the dual-state process in normal- and zero-count states, which means that there are two safety states for road entities. These models often outperformed traditional count data models with a better goodness-of-fit. However, the validity of the zero-inflated model and its application in crash prediction models were criticized by Lord et al. (2005, 2007), considering the zero-generating process of zero crash count. They argued that a road link should never be judged as being in an inherently safe state and that a zero crash count could be avoided by developing a suitable and manageable database with reasonable space and time scales. In concern of the safety variation over time for each road entity, Malyshkina et al. (2009) proposed Markov switching count data models, which allow the safety state of roadway to switch between two states. Their model achieved a superior statistical fit in contrast to traditional models. However, the fundamental assumption of this modeling approach is still based on two safety states of roads.

Apart from the improvement in statistical fit, the recognition of significant contributory factors to crash risk is essential and useful in practice. Some possible risk factors may not be recognized by traditional statistical methods when the crash counts are subject to excess zeros. Bootstrap resampling approach is capable of reducing the bias in parameter estimates and standard errors of crash prediction models with excess zeros. The bootstrap method is widely used in classification trees for road safety analysis (Harb et al., 2009; Chung, 2013) but rarely in regression models. It is expected that the standard error and confidence intervals of parameters of crash regression models obtained from bootstrap resampling approaches can be improved (Efron, 1979).

The remainder of this paper is organized as follows. We first describe the study design and data collection method in Section 2. Then, we discuss the methods of analysis in Section 3. The results are presented in Section 4 and their implications are discussed in Section 5. Section 6 presents the concluding remarks and recommendations for future research.

2. Study design and data

We first establish a comprehensive crash database containing traffic volume, road geometrics and traffic control factors, weather conditions, and temporal distribution on 112 urban roadway segments in Hong Kong using geographical information system (GIS) techniques.

Extensive traffic count data are obtained from the Hong Kong Annual Traffic Census system (Transport Department, 2002–2006), which consists of over 1500 stations and covers 86.8% of all motorways in Hong Kong (Tong et al., 2003; Lam et al., 2003). In particular, directional traffic flows are measured continuously at 112 core stations throughout the study period. These 112 core stations are evenly and widely distributed across the territory and cover 164.6 km (i.e., 8.0%) of all of the motorways in Hong Kong. As these locations are selected for transport planning purposes, there is unlikely to be any safety-related bias. The roadway segments that are considered in this research are defined according to the standards of ATC, of which the traffic flow and geometric design characteristics are consistent throughout the segments. With respect to the time interval, as mentioned by Rothrock and Keefer (1957), it is difficult to distinguish different traffic states (free-flow or congestion regime) if more detailed traffic flow data, e.g., less than 1 h interval is used. In concern of the traffic volume variation along time, we derive directional traffic volumes for all of the road segments adjacent to the core stations for every 4-h period [07:00–11:00 (morning), 11:00–15:00 (noon), 15:00–19:00 (afternoon), 19:00–23:00 (evening), 23:00–03:00 (middle of the night), and 03:00–07:00 (dawn)] every day 2002–2006. This gives 10,956 4-h time units and yields a sample comprising 2,230,314 observations. The VKM is obtained by multiplying road segment length by traffic volume and is used as the proxy of exposure measure for every road segment in each time period.

Crash data are obtained from the Traffic Information System (TIS) maintained by the Transport Department, which captures precise information on crash circumstances, road environment, and vehicles and casualties involved in every road crash that involves personal injury. Using the GIS technique, 7790 crashes that occur during 2002–2006 are accordingly mapped onto 210 corresponding spatial units. As the TIS database captures information on collision types and numbers of vehicles involved, it reveals that 3393 crashes are SV crashes (43.6%) and 4397 are MV crashes (56.4%).

In Hong Kong, crashes are categorized into three types with respect to crash severity based on the injury severity of the most seriously injured person in a crash: fatal, serious, and slight. A fatal crash refers to a crash in which at least one person is killed immediately or is injured and subsequently dies within 30 days of the crash. A serious injury crash refers to a crash in which one or more persons are injured and detained in hospital for more than 12 h. A slight injury crash is one in which one or more persons are injured but not to the extent that a hospital stay of more than 12 h is required. In this study, we group fatal and serious injury crashes together as killed and seriously injured (KSI) crashes in the subsequent analysis. We segregate the dataset into two with respect to crash severity levels: KSI and slight injury crashes. Of the 7790 crashes, 1634 (21.0%) are KSI crashes, of which 859 are SV crashes and 775 are MV crashes, and 6156 (79.0%) are slight injury crashes, of which 2534 are SV crashes and 3622 are MV crashes.

Road geometric designs and traffic controls are also incorporated in the study, specifically including lane-changing