Accident Analysis and Prevention 85 (2015) 207–218

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Calibration of crash risk models on freeways with limited real-time traffic data using Bayesian meta-analysis and Bayesian inference approach

Chengcheng Xu[a,b,*], Wei Wang[a,b], Pan Liu[a,b], Zhibin Li[a,b]

[a] Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing 210096, China
[b] Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Si Pai Lou #2, Nanjing 210096, China

ARTICLE INFO

Article history:
Received 13 August 2014
Received in revised form 9 September 2015
Accepted 23 September 2015

Keywords:
Bayesian meta-analysis
Meta-regression
Real-time crash risk
Bayesian inference
Bayesian predictive densities

ABSTRACT

This study aimed to develop a real-time crash risk model with limited data in China by using Bayesian meta-analysis and Bayesian inference approach. A systematic review was first conducted by using three different Bayesian meta-analyses, including the fixed effect meta-analysis, the random effect meta-analysis, and the meta-regression. The meta-analyses provided a numerical summary of the effects of traffic variables on crash risks by quantitatively synthesizing results from previous studies. The random effect meta-analysis and the meta-regression produced a more conservative estimate for the effects of traffic variables compared with the fixed effect meta-analysis. Then, the meta-analyses results were used as informative priors for developing crash risk models with limited data. Three different meta-analyses significantly affect model fit and prediction accuracy. The model based on meta-regression can increase the prediction accuracy by about 15% as compared to the model that was directly developed with limited data. Finally, the Bayesian predictive densities analysis was used to identify the outliers in the limited data. It can further improve the prediction accuracy by 5.0%.

1. Introduction

Dynamic safety management systems provide a proactive way to improve traffic safety on freeway mainlines. The first and most important step in dynamic safety management is the identification of hazardous traffic conditions with high crash likelihood. To this end, numerous studies have developed crash risk models to investigate the relationship between crash risks and real-time traffic variables (Abdel-Aty et al., 2004, 2005, 2007, 2012a,b; Abdel-Aty and Pande, 2005, 2006; Abdel-Aty and Pemmanaboina, 2006; Lee and Abdel-Aty, 2008; Pande and Abdel-Aty, 2005; Hassan and Abdel-Aty, 2011, 2013; Ahmed and Abdel-Aty, 2011; Hossain and Muromachi, 2010; Zheng et al., 2010; Ahmed et al., 2012; Li et al., 2013, 2014; Xu et al., 2012a,b, 2013a,b, 2014, 2015; Yu and Abdel-Aty, 2013a). These models are used to predict the crash risks on freeway segments over a short time period, such as 5 min. The development of crash risk models usually requires high quality and quantity of real-time traffic data. The used traffic data are collected from freeway surveillance system for quite a long time, such as one or two years. The crash risk models that are directly developed with limited data cannot well capture the relationship between crash risks and traffic flow characteristics. Previous studies suggested that this kind of data can result in biased model estimation results and inadequate predictive performance (Lord, 2006; Lord and Miranda-Moreno, 2008; Xu et al., 2014).

When developing the crash risk models, the problem of limited real-time traffic data may occur. For example, the traffic surveillance equipment has just been placed on a freeway, or the local transportation agencies are not expected to store all the historic real-time traffic data. To develop a crash risk model for a freeway on which only limited data are available, one possible method is based on Bayesian inference approach (Lord and Miranda-Moreno, 2008; Khondakar et al., 2010; Xu et al., 2014; Yu and Abdel-Aty, 2013b). The central idea of Bayesian inference approach is to combine the information from observed data with the prior information from additional knowledge that does not depend on the observed data. The Bayesian inference approach achieves this by adding a prior distribution of the parameters in the model. Most of the existing crash risk models based on the Bayesian inference approach used the non-informative priors, as the dataset in these studies were

* Corresponding author at: Jiangsu Key Laboratory of Urban ITS, Southeast University, Si Pai Lou #2, Nanjing 210096, China.
E-mail addresses: xuchengcheng@seu.edu.cn (C. Xu), wangwei@seu.edu.cn (W. Wang), pan_liu@hotmail.com (P. Liu), lizhibin-2002@163.com (Z. Li).

http://dx.doi.org/10.1016/j.aap.2015.09.016
0001-4575/

enough to establish an accurate relationship between crash risks and real-time traffic variables. However, when the sample size of the dataset is small, the informative priors with plausible knowledge can increase the accuracy of model estimation and predictive performance (Lord and Miranda-Moreno, 2008; Khondakar et al., 2010; Hadayeghi et al., 2006; Xu et al., 2014).

Recently, several studies have been conducted to developed crash frequency model with limited data using the Bayesian inference method with informative priors (Lord and Miranda-Moreno, 2008; Khondakar et al., 2010; Hadayeghi et al., 2006). For example, Lord and Miranda-Moreno found that the Bayesian inference approach can greatly minimize the risk of inaccurate model estimation caused by limited data when an appropriate informative prior distribution is used (Lord and Miranda-Moreno, 2008). The study conducted by Hadayeghi et al. (2006) suggested that, when limited data are available, the accident prediction models based on the informative priors can produce better prediction accuracy than the model that was directly developed using the limited data. Most of these studies focused on developing the aggregate crash frequency models when only limited data are available. However, relatively fewer studies have considered how to develop real-time crash risk models with limited data. Besides, most of the previous studies developed the informative priors from a single model for another freeway (Hadayeghi et al., 2006; Xu et al., 2014). The disadvantage of this method is that the inaccurate informative priors from one single freeway may affect the accuracy of the model estimation.

This study fills the gap in developing the informative priors for the real-time crash risk models by using the Bayesian meta-analysis. The meta-analysis is a quantitative method that combines the results from numerous studies about real-time crash risk assessment and produces a more accurate estimate for the effects of traffic flow variables. The primary objective of this study is to develop a real-time crash risk model with limited data in China by using the Bayesian meta-analysis and the Bayesian inference method. More specifically, this study first conducted a systematic review of previous literature about real-time crash assessment using the Bayesian meta-analysis. Three different Bayesian meta-analyses, including the fixed effect meta-analysis, the random effect meta-analysis and the meta-regression, were used to formulate the informative priors. Then, the Bayesian inference method was used to develop the crash risk models based on the informative priors obtained from the three Bayesian meta-analyses. The effects of three Bayesian meta-analyses on the accuracy of model estimation were examined. Finally, the Bayesian predictive densities analysis for identifying the outliers in the limited data has been introduced and its effects on the accuracy of model estimation were examined.

2. Data sources

2.1. Data from systematic review of literature

A literature search for relevant studies published from 2001 to 2015 was conducted using several databases, including the transportation research international documentation (TRID) database, the Sciencedirect database, the Springer database, the Taylor and Francis database, and the Wiley database. In general, "real-time crash risk", "crash likelihood", "traffic flow", "traffic data", and "loop detector data" were used as the search terms.

The initial literature review indicated that 114 published papers used the high-resolution traffic flow data to conduct real-time crash risk assessment. These papers were selected for further examination in detail, of which 36 were included in the meta-analyses. The following criteria were used when selecting papers for the meta-analyses:

(1) The selected studies should focus on the general crashes instead of specific types of crashes. The studies focusing on specific crash types were excluded.
(2) The selected studies should report the effects of traffic variables as logarithms of odds ratios or odds ratios. The crash risks defined in the selected studies are consistent. Numerous studies using artificial intelligent (AI) or machine learning technique were excluded, because most of the AI models work as black-boxes. They cannot provide the estimated effects of traffic variables as logarithms of odds ratios or odds ratios.
(3) The traffic flow variables in the included studies should be aggregated during 5-min time interval. Because this study aimed to develop a real-time crash risk model based on the traffic data with 5-min aggregation interval on the freeway in China. Besides, most of the previous studies about real-time crash risk assessment aggregated traffic flow data at 5-min level. And they should also be observed in the time interval between 5 and 10 min prior to the crash occurrence time.
(4) The traffic flow variables in the included studies should be at the same location with respect to crash site (i.e., upstream or downstream).

To conduct meta-analyses, the following type of information was extracted from each selected study:

(1) Publication year. Since the real-time crash risk assessment is an emerging field in traffic safety analysis, the selected papers were published from 2001 to 2015.
(2) The freeway and country where the data were collected. Most of the studies were conducted in the United States. Only a few studies were conducted in other countries, including Korea, Japan, Belgium, and China.
(3) Considered traffic flow characteristics. Different traffic variables for measuring traffic conditions in previous studies were extracted, including the means and standard deviations of traffic flow measurements, the difference in traffic flow measurements between adjacent lanes, as well as the difference in traffic flow measurements between upstream and downstream stations.
(4) The estimator of the effects of traffic variables. The estimated coefficient and the standard error of the coefficient for each traffic flow variable were extracted from each study. Note that some studies only reported the odds ratio and the standard error of the odds ratio. In this case, the coefficient was calculated as the logarithm of the odds ratio; and the standard error of the coefficient is calculated using the following equation (Ntzoufras, 2009):

$$\hat{\sigma} = log \frac{U/L}{2 \times 1.96} \tag{1}$$

where $\hat{\sigma}$ represents the standard error of the estimated coefficient of the traffic flow variable; U and L, respectively, represents the upper and the lower limits of the 95% confidence interval of the odds ratio.

Table 1 gives the studies included in the meta-analysis. The main reasons for excluding studies were as follows: (1) the studies did not meet the above selection criteria; (2) the studies cannot provide the sufficient information, i.e., the above four type information; (3) the studies were review, i.e., a secondary source based on existing literature.