Accident Analysis and Prevention 88 (2016) 9–19

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Predicting crash risk and identifying crash precursors on Korean expressways using loop detector data

Ho-Chan Kwak[a,*], Seungyoung Kho[b]

[a] Green Transport and Logistics Institute, Korea Railroad Research Institute, Uiwang, Republic of Korea
[b] Department of Civil and Environmental Engineering, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 6 April 2015
Received in revised form 4 November 2015
Accepted 4 December 2015

Keywords:
Crash risk prediction
Segment type
Traffic flow state
Genetic programming
Conditional logistic regression analysis
Loop detector

ABSTRACT

In order to improve traffic safety on expressways, it is important to develop proactive safety management strategies with consideration for segment types and traffic flow states because crash mechanisms have some differences by each condition. The primary objective of this study is to develop real-time crash risk prediction models for different segment types and traffic flow states on expressways. The mainline of expressways is divided into basic segment and ramp vicinity, and the traffic flow states are classified into uncongested and congested conditions. Also, Korean expressways have irregular intervals between loop detector stations. Therefore, we investigated on the effect and application of the detector stations at irregular intervals for the crash risk prediction on expressways. The most significant traffic variables were selected by conditional logistic regression analysis which could control confounding factors. Based on the selected traffic variables, separate models to predict crash risk were developed using genetic programming technique. The model estimation results showed that the traffic flow characteristics leading to crashes are differed by segment type and traffic flow state. Especially, the variables related to the intervals between detector stations had a significant influence on crash risk prediction under the uncongested condition. Finally, compared with the single model for all crashes and the logistic models used in previous studies, the proposed models showed higher prediction performance. The results of this study can be applied to develop more effective proactive safety management strategies for different segment types and traffic flow states on expressways with loop detector stations at irregular intervals.

1. Introduction

With the tremendous growth of Intelligent Transportation System (ITS) during the past decade, it is now possible to collect traffic parameters such as traffic volume, speed, and occupancy from a variety of detectors in real time. Several studies have attempted to demonstrate the potential application of loop detector data in order to reveal the relationship between crash occurrence and traffic parameters (Hughes and Council, 1999; Oh et al., 2001, 2005; Lee et al., 2002, 2003). Also, proactive safety management strategies utilizing Advanced Traffic Management Systems (ATMS) such as variable speed limits (VSL) and ramp metering have showed effects on improving traffic safety (Abdel-Aty et al., 2006, 2007; Lee et al., 2006a,b; Lee and Abdel-Aty, 2008). Within these studies, real-time crash risk prediction models were estimated to predict crash occurrence likelihood based on real-time traffic data.

The crash precursors mean traffic flow characteristics leading to crashes, which are identified by comparing the crash cases and non-crash cases. The crash precursors also will be different depending on traffic flow characteristics by surrounding environment on expressways. Especially, these traffic flow characteristics are divided by segment type and traffic flow state in Highway Capacity Manual (HCM). Therefore the real-time crash risk prediction models based on the crash precursors are required to develop for different segment types and traffic flow states on expressways. However, the majority of studies have developed a single model (e.g., aggregated model) for overall crashes (Abdel-Aty et al., 2004, 2008; Abdel-Aty and Pande, 2005; Ahmed and Abdel-Aty, 2012, 2013; Yu and Abdel-Aty, 2013). A major drawback of the aggregated modeling technique is that it cannot consider different effects of traffic flow characteristics on crash risk by roadway geometry and traffic flow states. Therefore, separated crash risk prediction models were developed by segment type and traffic flow state on expressways in this study. Also Korean expressways have irregular intervals between loop detector stations. Therefore, we investigated on the effect and application of the detector stations at irregular intervals for the crash risk prediction on expressways.

* Corresponding author.
E-mail addresses: kwak01@krri.re.kr, kwak.hochan@gmail.com (H.-C. Kwak).

http://dx.doi.org/10.1016/j.aap.2015.12.004
0001-4575/

Data used in this study were collected from the Gyeongbu expressway in Korea. The real-time traffic data (volume, speed and occupancy) was obtained from loop detectors and matched with historical crash data. The real-time crash risk prediction models were developed by segment type and traffic flow state. The mainline of expressways is divided into two segments based on ramp presence, basic segment and ramp vicinity, and the traffic flow states are classified into uncongested and congested conditions. Genetic programming technique was used in this study to develop the real-time crash risk prediction models. The genetic programming is a relatively new modeling technique that was proposed to solve classification and regression problems. Compared with traditional statistical regression methods and machine learning algorithms, two major advantages of genetic programming were proposed (Xu et al., 2013b). First, genetic programming can find a solution to a problem without any pre-specified functional forms. Second, in contrast with different machine learning algorithms, genetic programming can remove the "black box" effect and make the model understandable. Due to genetic programming models lack the ability to select significant variables that play a role to increase the reliability of prediction models, conditional logistic regression analysis was first estimated to select the most significant traffic variables contributing to crash occurrence. The conditional logistic regression analysis has an advantage which can control confounding factors by matching. Based on the chosen explanatory variables by conditional logistic regression analysis, the genetic programming models have been estimated and the relationship between traffic variables and crash risk has been investigated for each condition. Finally, the prediction performance of the proposed models have been compared to single model based on Receiver Operating Characteristics (ROC) curves and areas under the ROC curve (AUC).

2. Background

Real-time crash risk prediction models were estimated with the purpose of identifying the crash precursors and the results were applied in proactive safety management strategies. With the advanced traffic surveillance system (loop detectors, remote traffic microwave sensors, automatic vehicle identification systems), traffic flow characteristics prior to crash occurrence could be identified and matched with the crashes. Abdel-Aty et al. (2004) developed crash likelihood prediction model for freeway using matched case-control logistic regression based on loop detector data. The average occupancy at the upstream and the coefficient of variation in speed at downstream was found to affect crash occurrence most significantly. Abdel-Aty and Pande (2005) proposed probabilistic neural network to classify traffic speed pattern for crash and non-crash conditions from historical crash and loop detector data collected from the Interstate-4 corridor in the Orlando metropolitan area.

In order to increase prediction performance of the crash risk models, the separate models have been developed by the crash characteristics. Abdel-Aty et al. (2005) identified that multivehicle crashes under high- and low-speed traffic conditions on freeways were found to differ in severity and in their mechanism. The speed regime was divided into two speed conditions by the distribution of average speeds obtained immediately before the crash from the loop detector and the two models by speed condition were estimated using matched case-control logistic regression. Pande and Abdel-Aty (2006a,b) identified the traffic parameters leading to rear-end collisions and lane-change related collision from loop detector data on freeways. Classification tree was used in variable selection procedure, and then the classification models (crash versus non-crash) based on selected variables were developed by neural network. Xu et al. (2012) divided freeway traffic flow into different states, and evaluated the safety performance associated with each state. K-means clustering analysis was conducted to classify traffic flow into five different states, and conditional logistic regression models using case-controlled data were developed to identify the relationship between crash risks and traffic states. Hassan and Abdel-Aty (2013) investigated whether real-time traffic flow data, collected from loop detectors and radar sensors on freeways, can be used to predict crashes occurring at reduced visibility conditions. The result indicated that traffic variables leading to visibility related crashes are slightly different from those variables leading to clear visibility crashes. Xu et al. (2013a) developed a model that predicts the crash likelihood at different levels of severity with a particular focus on severe crashes. Crash severity was divided into three levels: fatal/incapacitating injury crashes (KA), non-incapacitating/possible injury crashes (BC), and property-damage-only crashes (PDO). The sequential logit model was used to link the likelihood of crash occurrences at different severity levels to various traffic flow characteristics identified from detector data.

Also the studies applying more advanced vehicle detection systems and modeling techniques have been performed to improve the prediction performance. Abdel-Aty et al. (2008) attempts to address the issues of transferability through analysis of crash data and loop detector data collected from Dutch freeways. In addition to these transferability issues, the application of a new data mining technique, random forests, was investigated for identifying significant variables associated with the crash. Ahmed and Abdel-Aty (2012) examined the identification of freeway locations with high crash potential using real-time speed data collected from automatic vehicle identification (AVI). Travel time and space mean speed data by AVI systems and crash data on the freeway network in Orlando were collected. Utilizing a random forest technique for significant variable selection and matched case-control method to account for the confounding effects, the log odds of crash occurrence were calculated. The results showed that the length of the AVI segment was found to be a crucial factor that affects the usefulness of the AVI traffic data and the likelihood of a crash is statistically related to speed data obtained from AVI segments within an average length of 1.5 mile. Hossain and Muromachi (2012, 2013a) employed random multinomial logit model to identify the most important predictors as well as the most suitable detector locations to build models for the basic freeway segment and the ramp areas of urban expressways. And Bayesian belief net (BBN) was applied to build the real-time crash prediction model. Hossain and Muromachi (2013b) identified crash influential factors and traffic patterns for different types of crashes on urban expressways. However, these studies were focused on crashes and traffic patterns on urban expressways rather than those on conventional regional expressways in this study. Compared to the conventional expressways, crashes and traffic patterns on urban expressways is different since urban expressways have a relatively lower speed limit and more frequent ramps. Xu et al. (2013b) evaluated the application of the genetic programming model for real-time crash prediction on freeways. Traffic, weather, and crash data were obtained from I-880N freeway in California, United States, and the random forest technique was conducted to select the variables that affect crash risk. The genetic programming models were found to increase the crash prediction accuracy compared with binary logit models. Ahmed and Abdel-Aty (2013) proposed a framework for real-time crash risk assessment on a freeway in Colorado fusing traffic data from two different detection systems (AVI and Remote Traffic Microwave Sensors (RTMS)), real-time weather data and roadway geometry data. Stochastic Gradient Boosting (SGB) is used to calibrate the model. Yu and Abdel-Aty (2013) introduced support vector machine (SVM) to evaluate real-time crash risk. Classification and regression tree (CART) model has been developed to select the most important traffic variables. The SVM models with