Accident Analysis and Prevention 91 (2016) 114–126

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations

Lei Lin, Qian Wang, Adel W. Sadek*

Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, USA

ARTICLE INFO

Article history:
Received 26 October 2015
Received in revised form 4 February 2016
Accepted 2 March 2016
Available online 11 March 2016

Keywords:
M5P tree
Hazard-based duration model
Accelerate failure time (AFT)
Traffic accident duration prediction
Traffic accident data heterogeneity

ABSTRACT

The duration of freeway traffic accidents duration is an important factor, which affects traffic congestion, environmental pollution, and secondary accidents. Among previous studies, the M5P algorithm has been shown to be an effective tool for predicting incident duration. M5P builds a tree-based model, like the traditional classification and regression tree (CART) method, but with multiple linear regression models as its leaves. The problem with M5P for accident duration prediction, however, is that whereas linear regression assumes that the conditional distribution of accident durations is normally distributed, the distribution for a "time-to-an-event" is almost certainly nonsymmetrical. A hazard-based duration model (HBDM) is a better choice for this kind of a "time-to-event" modeling scenario, and given this, HBDMs have been previously applied to analyze and predict traffic accidents duration. Previous research, however, has not yet applied HBDMs for accident duration prediction, in association with clustering or classification of the dataset to minimize data heterogeneity. The current paper proposes a novel approach for accident duration prediction, which improves on the original M5P tree algorithm through the construction of a M5P-HBDM model, in which the leaves of the M5P tree model are HBDMs instead of linear regression models. Such a model offers the advantage of minimizing data heterogeneity through dataset classification, and avoids the need for the incorrect assumption of normality for traffic accident durations. The proposed model was then tested on two freeway accident datasets. For each dataset, the first 500 records were used to train the following three models: (1) an M5P tree; (2) a HBDM; and (3) the proposed M5P-HBDM, and the remainder of data were used for testing. The results show that the proposed M5P-HBDM managed to identify more significant and meaningful variables than either M5P or HBDMs. Moreover, the M5P-HBDM had the lowest overall mean absolute percentage error (MAPE).

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Traffic incidents account for more than 50% of motorist delays on freeways (Farradyne, 2000; Chin et al., 2004). To reduce the societal cost of such incidents, an efficient traffic incident management system (TIM) need be developed and deployed. The TIM process can be viewed as consisting of the following five phases (Zhan et al., 2011): (1) the incident detection phase, which refers to the time interval from the occurrence of the incident to its detection; (2) the incident verification phase that covers the period from the detection to the confirmation of the incident; (3) the incident response phase spanning from the moment an incident is confirmed to the time when the first responder arrives on the scene; (4) the incident clearance phase which refers to the time interval from the arrival of the first responder to the time when the incident has been cleared from the freeway; and (5) the incident recovery phase covering the time until normal traffic conditions resume after the incident clearance phase.

A critical component of effective TIM involves the ability to predict the likely incident duration under various conditions (different local and regional traffic conditions, time of day, day of week, seasonal variations, weather conditions, work zones, etc. . .). Based on the predicted duration, authorities can allocate incident response personnel and resources more effectively, inform travelers about traffic conditions more accurately, and decide upon the appropriate response strategy.

This paper proposes a new traffic accident duration prediction model which combines a decision tree model, namely the M5P tree model, and a statistical hazard-based duration model (HBDM). The proposed model will hereafter be referred to as the M5P-HBDM. As will be discussed in more detail later, M5P-HBDM offers the advantage of minimizing data heterogeneity through dataset clas-

* Corresponding author. Fax: +1 716 645 4367.
E-mail address: asadek@buffalo.edu (A.W. Sadek).

http://dx.doi.org/10.1016/j.aap.2016.03.001
0001-4575/© 2016 Elsevier Ltd. All rights reserved.

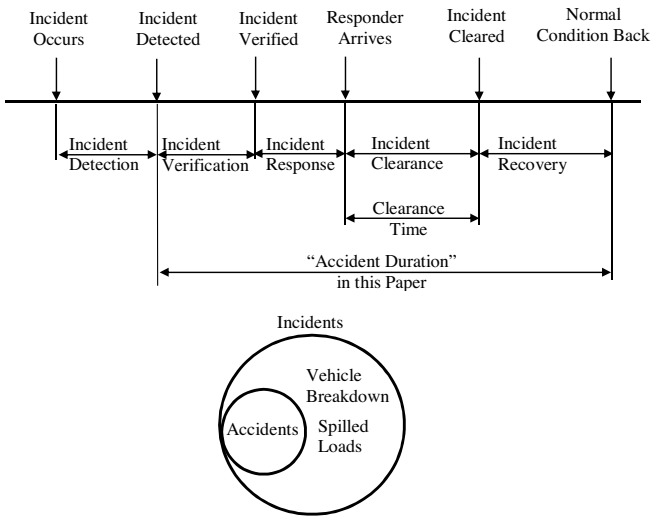

Fig. 1. Traffic incident management process and accident duration definition in this paper.

sification, while simultaneously avoiding the need for imposing restrictive assumptions regarding the distribution of traffic accident durations. The performance of the M5P-HBDM was evaluated against the performance of a stand-alone M5P tree algorithm and a stand-alone HBDM, on two freeway accident datasets.

Before proceeding, a clarification of a few terms is in order. In this paper, we assume that accidents are a *subset* of incidents as shown in Fig. 1. Incidents, on the other hand, include events such as vehicle breakdowns, spilled loads or other random events, besides accidents (He et al., 2013). The focus of this paper is on accidents and not incidents. We further assume that the duration of an accident refers to the time interval from the moment an accident is detected to the time when normal traffic conditions return as also shown in Fig. 1.

The organization of the paper is as follows. The paper begins with a review of previous research on incident duration prediction models and approaches to deal with heterogeneity in traffic accident data. Next, the basic methodologies of M5P tree and HBDM are introduced, and the proposed algorithm to build the M5P-HBDM is described. The two traffic accident datasets used in this research are then presented, and three different incident duration models are constructed for each dataset: (1) a stand-alone M5P Tree model; (2) a stand-alone HBDM; and (3) the proposed M5P-HBDM. The performances of the three models, in terms of prediction accuracy and the significant variables identified, are then compared. Finally, the study's conclusions are summarized and suggestions for future are provided.

2. Literature review

2.1. Traffic accident duration analysis

Given the enormous societal cost of traffic accidents, the transportation research community has always been interested in models and methodologies for predicting the likelihood of traffic accidents, the factors behind their occurrences, and their likely durations. In terms of accident duration analysis, the methods proposed in the literature can be grouped into the following categories: (1) statistical methods; and (2) artificial intelligence (AI)-based methods.

For statistical methods, previous research has examined the candidate probability distributions that fit traffic accident durations. Golob et al. (1987) analyzed truck-involved incident durations in California, and reported that the durations of the incidents, categorized by the type of collisions, followed a log-normal distribution. On the other hand, Ozbay and Kachroo (1999) identity a normal distribution of incident durations for homogeneous incidents grouped by incident type and severity. In terms of statistical methods, regression models have been applied in the past to predict traffic accident durations and identify the contributing factors. For example, Giuliano (1989) assigned incidents into multiple categories and, for each category, estimated a model for predicting incident durations using linear regression techniques. Garib et al. (1997) also developed a polynomial regression model to predict incident durations. Their results showed that, in terms of adjusted R-square, 81% of the variability in incident durations, in a natural logarithm format, can be predicted as a function of six independent variables such as the number of lanes affected, the number of vehicles involved, whether a truck was involved or not, the time of day, the police response time, and weather conditions. Naturally, standard regression models have the advantage of being easily understood and interpreted (Khattak et al., 1995). Besides regression, Nam and Mannering (2000) built hazard-based duration models to evaluate incident durations, based on a two-year dataset from the state of Washington. They mentioned that, compared to regression approaches, hazard-based duration models have the advantage of allowing the explicit study of duration effects (i.e., the relationship between how long an incident has lasted and the likelihood of it ending soon). Recently, Alkaabi et al. (2011) and Chung (2010) also developed hazard-based duration models to predict traffic accident durations, and to analyze the factors affecting such durations.

For AI-based methods, a few previous studies employed decision trees to predict incident durations (He et al., 2013; Ozbay and Kachroo, 1999; Smith and Smith, 2001). The main advantage of decision trees is that they require no assumption regarding the probability distribution of the incident duration data (Alkaabi et al., 2011). On the negative side, however, Ozbay and Noyan (2006) pointed out that the decision trees can sometimes become unstable and insensitive to the stochastic nature of the data. Many other AI techniques have also been applied to accident duration prediction. Examples include Bayesian networks (BN) (Ozbay and Noyan, 2006), artificial neural networks (ANN) (Wei and Lee, 2007), genetic algorithms (GA) (Lee and Wei, 2010) and support vector machines (Valenti et al., 2010). Recently, Lin et al. (2014) proposed a complex network algorithm, which combines the modularity-optimizing community detection algorithm and the association rules learning algorithm, to unveil the factors that affect incident clearance time.

2.2. Data heterogeneity

The heterogeneity inherent in traffic accident data often prevents their further exploration (Savolainen et al., 2011). In the presence of data heterogeneity, the patterns/distributions observed at the population level may be surprisingly different from the underlying patterns at the individual level (Vaupel and Yashin, 1985). In other words, the aggregated behavior of a heterogeneous population, composed of two or more homogeneous but differently behaving subpopulations, will differ from the behavior of any single individual (Lerman, 2013).

To deal with the issue, random effects and random parameters models have been proposed for traffic accident data analysis (Karlaftis and Tarko, 1998; Miaou et al., 2003; Anastasopoulos and Mannering, 2009). Such models capture the unobserved heterogeneity by using random error terms, and allow each estimated parameter of the model to vary across each individual observation in the dataset (Lord and Mannering, 2010). This can prevent the problems of inconsistent coefficient estimates and inferences based on inappropriate standard errors (Nam and Mannering, 2000).

Download English Version:

https://daneshyari.com/en/article/571986

Download Persian Version:

https://daneshyari.com/article/571986

Daneshyari.com