Transportation Research Part C 54 (2015) 74–85

Contents lists available at ScienceDirect

Transportation Research Part C

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

Competing risk mixture model and text analysis for sequential incident duration prediction

Ruimin Li [a,d,*], Francisco C. Pereira [a,b,1], Moshe E. Ben-Akiva [c,1]

[a] Intelligent Transportation Systems Laboratory, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 0219, USA
[b] Singapore-MIT Alliance for Research and Technology, 1 CREATE Way, #09-02 CREATE Tower, Singapore 138602, Singapore
[c] Edmund K. Turner Professor of Department of Civil and Environmental Engineering, MIT. Room 1-181, 77 Massachusetts Avenue, Cambridge, MA 02139, USA
[d] Institute of Transportation Engineering, Department of Civil Engineering, Tsinghua University, Beijing 100084, China

A R T I C L E   I N F O

Article history:
Received 12 September 2014
Received in revised form 7 March 2015
Accepted 7 March 2015
Available online 24 March 2015

Keywords:
Traffic incident management
Incident duration prediction
Competing risk mixture model
Topic modeling

A B S T R A C T

Predicting the duration of traffic incidents sequentially during the incident clearance period is helpful in deploying efficient measures and minimizing traffic congestion related to such incidents. This study proposes a competing risk mixture hazard-based model to analyze the effect of various factors on traffic incident duration and predict the duration sequentially. First, topic modeling, a text analysis technique, is used to process the textual features of the traffic incident to extract time-dependent topics. Given four specific clearance methods and the uncertainty of these methods when used during traffic incidents, the proposed mixture model uses the multinomial logistic model and parametric hazard-based model to assess the influence of covariates on the probability of clearance methods and on the duration of the incident. Subsequently, the performance of estimated mixture model in sequentially predicting the incident duration is compared with that of the non-mixture model. The prediction results show that the presented mixture model outperforms the non-mixture model.

1. Introduction

Traffic incidents are main contributors to traffic congestion, particularly non-recurrent congestion, not only on expressways but also on urban arterial networks (Owens et al., 2010). Over the past two decades, various traffic incident management systems have been deployed in many cities and expressway traffic management centers worldwide to reduce traffic incidents and alleviate the related congestion. Traffic flow management, as well as providing accurate and timely information for travelers during traffic incident clearance periods, are two key aspects of efficient traffic incident response (Al-Deek and Kanafani, 1993; Lou et al., 2011). The response strategy during the clearance period depends largely on the duration of an incident. Thus, accurate prediction of traffic incident duration has attracted research attention.

Using limited data, many studies (Chung, 2010; Zhan et al., 2011) on traffic incident duration prediction focus on “one time” prediction, which means that at the time when an incident is reported and the traffic control center receives the initial information about the incident, the incident duration is predicted but not updated subsequently. In fact, the residual incident

* Corresponding author at: Institute of Transportation Engineering, Department of Civil Engineering, Tsinghua University, Beijing 100084, China. Tel.: +86 10 62770985.

E-mail addresses: lrmin@tsinghua.edu.cn (R. Li), camara@smart.mit.edu (F.C. Pereira), mba@mit.edu (M.E. Ben-Akiva).

[1] Tel.: +1 617 253 5324; fax: +1 617 253 0082.

http://dx.doi.org/10.1016/j.trc.2015.03.009

duration at one point is mainly dependent on the following: (1) the length of time of the incident, (2) the measurements conducted, and (3) the information about the traffic incident at that time. These factors mean that the expected residual incident duration should be updated as the traffic control center receives new information, which helps in efficient traffic incident management.

Traffic incident duration is usually defined as the time between incident notification and incident clearance, mainly including three stages: preparation, travel, and clearance, although which are not included in every incident (Sun and Chilukuri, 2010). Preparation delay and travel time are combined to be considered as response time (Hou et al., 2012). During the past two decades, several methods have been used to model traffic incident duration or clearance time prediction, which is typically the longest stage of traffic incident duration.

Smith and Smith (2001) investigated the application of three models (i.e., stochastic, nonparametric regression, and classification tree) on traffic incident clearance time prediction, but these models inefficiently predicted the clearance time of future incidents.

Various decision tree models have been applied in traffic incident duration predictions. A rule-based supplemental model was combined with an ordered probit model to predict incident duration with different time ranges based on the incident information obtained in 2001 and 2002 (Lin et al., 2004). Similar methods were applied based on the data from 2003 to 2006 (Kim et al., 2008). The results indicated that the combined model can improve prediction accuracy. Recently, a rule-based tree model, a multinomial logit model, and a naïve Bayesian classifier (Kim and Chang, 2012) were combined to develop a hybrid model for predicting freeway traffic incident duration in Maryland. With the ability to readily accommodate incomplete information, various Bayesian classifiers (Boyles et al., 2007; JiYang et al., 2008; Li and Cheng, 2011) were also used to model traffic incident duration prediction, and the results showed that under certain conditions, the Bayesian classifier-based models can provide a useful approach to traffic incident prediction. Zhan et al. (2011) applied the M5P tree algorithm to predict the lane clearance time of freeway incidents. As shown in the test results, the developed model can generally provide better prediction results than the traditional regression and decision tree models.

Survival analysis models have been used to predict traffic incident duration and other fields in transportation (Haque and Washington, 2015; Hasan et al., 2013). Chung (2010) developed a log-logistic accelerated failure time (AFT) model based on a two-year accident duration dataset and tested the temporal transferability; the results indicated that the developed model can provide reasonable prediction through the mean absolute percentage error (MAPE) measurement. Similarly, based on a different dataset from an expressway in China, the Weibull (Kang and Fang, 2011) and log-logistics AFT model (Wang et al., 2013a) were used to model the traffic duration and predict the duration of future incidents, with a reasonable prediction accuracy, as shown through the MAPE. The log-logistics AFT model was also used in the incident duration prediction for in-vehicle navigation systems with similar prediction accuracy based on the MAPE (Hu et al., 2011). Based on the incident dataset from Beijing, China, Li (2014) developed a series of AFT models that analyze and predict different incident duration stages, considering the unobserved heterogeneity, time-varying covariates, and the relationship between consecutive traffic incident duration stages. The results show that the developed model can generally achieve reasonable prediction results, except for very short or very long extreme values. Subsequently, Li et al. (2014) applied a mixture competing risks model that predicts the incident duration for the same dataset as this paper (incidents in Singapore). However, the developed model is *static* in that it does not consider the sequential nature of the information, as happens now by including the textual stream. This apparent small difference implies that, (1) the time sequential model proposed in this paper is more practical and provides the possibility for implementation in the real world, (2) the information extraction may include more factors which will influence the traffic duration.

All the aforementioned traffic incident duration prediction models were developed based on the assumption that when a model predicts the duration of an incident, all the possible information has already been acquired. However, during the initial stage of a traffic incident, limited information (e.g., incident location, time, weather, and traffic conditions) may be known about the incident; after the traffic response team arrives at the incident site, more information can be acquired. With more and more information about incident and traffic condition available, more dynamic analysis models of incident evolution have been developed (Imprialou et al., 2014; Sun and Chilukuri, 2010; Yang et al., 2014). Thus the traffic incident duration prediction model also should accommodate new information as the information arrives in its own time sequence.

First, using the data on 109 large-scale incidents that occurred between 1989 and 1990 on six major Chicago freeways, Khattak et al. (1995) developed a time-sequential methodology to predict the incident duration as traffic operation centers receive the incident information. The results show that this approach can offer a reasonable balance between the values of early and accurate incident duration prediction.

Subsequently, Wei and Lee (2007) applied artificial neural network (ANN)-based models and data fusion techniques to develop a time-sequential traffic incident duration prediction procedure, which mainly included two different adaptive ANN-based models. This study showed that the developed models can provide a reasonable prediction based on the performance of the MAPE, which was mostly under 40%. Lee and Wei (2010) employed ANNs and genetic algorithms to construct two models to provide a sequential prediction of accident duration from accident notification to clearance, with mostly less than 29% of the MAPE at each time point.

Qi and Teng (2008) applied a hazard-based model to develop a time-sequential procedure that included different hazard-based duration regression models with different variables for each stage, according to the specific information available. This study found that the incident duration prediction accuracy increased as more information became available and was incorporated into the developed models.