Transportation Research Part C 57 (2015) 30–41

Contents lists available at ScienceDirect

Transportation Research Part C

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

# Development of a maximum likelihood regression tree-based model for predicting subway incident delay

Jinxian Weng [a,b,*], Yang Zheng [b], Xiaobo Qu [c], Xuedong Yan [b]

[a] *College of Transport and Communications, Shanghai Maritime University, Shanghai 201306, China*
[b] *MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China*
[c] *Griffith School of Engineering, Griffith University, Gold Coast, 4222 QLD, Australia*

A R T I C L E   I N F O

*Article history:*
Received 10 September 2014
Received in revised form 11 February 2015
Accepted 1 June 2015
Available online 17 June 2015

*Keywords:*
Subway incidents
Delay
Maximum likelihood regression tree
Accelerated failure time

A B S T R A C T

This study aims to develop a maximum likelihood regression tree-based model to predict subway incident delays, which are major negative impacts caused by subway incidents from the commuter's perspective. Using the Hong Kong subway incident data from 2005 and 2009, a tree comprising 10 terminal nodes is selected to predict subway incident delays in a case study. An accelerated failure time (AFT) analysis is conducted separately for each terminal node. The goodness-of-fit results show that our developed model outperforms the traditional AFT models with fixed and random effects because it can overcome the heterogeneity problem and over-fitting effects. The developed model is beneficial for subway engineers looking to propose effective strategies for reducing subway incident delays, especially in super-large-sized cities with huge public travel demand.



## 1. Introduction

Subway transportation is an efficient and green transportation mode that can produce low carbon emissions. In general, it plays a vital role in the urban public transportation service. However, breakdowns of subway services caused by system component failures (e.g., power failures), which are defined as subway incidents, can cause great disruption to the commuters though the occurrence likelihood is low. From the commuter's perspective, the subway incident delay is the major consequence resulting from a subway incident, measured as the difference between the scheduled and actual subway train departure times. Note that the delay is not the only factor, but is a major factor affecting commuters' trips. For example, the cancellation of trains may sometimes influence the frequency of trains. However, this is a rare event because subway companies typically have a few spare or reserve trains in case there is an emergency. In fact, the subway delay is able to capture the exposure of train cancellations. For example, the subway delay that results from a train cancellation is generally larger than that when there is no train cancellation.

It is thus necessary for public transportation authorities to implement effective management strategies for clearing subway incidents as quickly as possible. The quick clearance of a subway incident usually requires an efficient allocation of resources so as to dispatch a crew in a timely manner. In an attempt to achieve this objective, there is a critical need to develop a model to comprehensively explore the influencing factors and predict the delays caused by subway incidents (Chung, 2010).

* Correspondence to: Dr. Jinxian Weng, Professor, Shanghai Maritime University, China.
*E-mail addresses:* jxweng@bjtu.edu.cn, wjx206@gmail.com (J. Weng).

http://dx.doi.org/10.1016/j.trc.2015.06.003
0968-090X/

So far, many studies have been conducted on the analysis of the causes of subway incidents. For example, fire may be a disaster in an underground subway system. Based on the archived subway fire incident data, Cheng et al. (2001) investigated the major causes of fire accidents and then provided some suggestions to reduce the occurrence likelihood of subway fire incidents. Some researchers have also examined the major causes of terrorist attacks (Staten, 1997; Okumura et al., 2005; Mumpower et al., 2013) and drivers' mental state (Mishara, 1999). In general, subway incidents might be induced by failures of multiple factors including the power system, subway vehicles, ventilation and smoke exhaust systems, the water supply and drainage system, and the communication and signal system (He et al., 2005).

Providing information to the public about delays caused by subway incidents could alert commuters to the necessity of rescheduling their trips. Therefore, it is of the utmost importance to accurately predict subway incident delays. To date, various parametric models, such as the accelerated failure time (AFT) models (Chung et al., 2010; Tavassoli et al., 2013), have been developed to estimate freeway incident delay. However, parametric models have predetermined functional forms and assumptions that are not usually met in the case of subway incidents, especially given the large number of influencing factors. In order to eliminate the drawbacks of parametric models, many researchers (e.g., Wei and Lee, 2007; Ma and Kockelman, 2008) have proposed nonparametric models for estimating subway/freeway incident delay. However, it should be noted that these nonparametric models make it difficult to in examine the marginal effect of influencing factors on the incident delay, which is useful for subway staff in establishing priorities for reducing subway incident delays.

Therefore, this study aims to develop a maximum likelihood regression tree-based model to predict the delays caused by subway incidents. Each terminal node of the regression tree is assigned a single AFT model to describe the distribution of the subway incident delay. The contribution of this study is twofold. First, it makes an initial attempt to build a model that will compensate for the weak points of parametric AFT models. Unlike the AFT models, our developed model is able to account for the heterogeneity effect as well as avoid the over-fitting problem. Second, the developed model could help subway staff to quickly implement the most effective strategies for reducing subway incident delays, especially in the super-large-sized cities with huge public travel demand.

## 2. Literature review

To date, various parametric models have been proposed for the analysis of traffic incident durations. Among these parametric models, the AFT models have been the most widely used in previous studies (Haque and Washington, 2015). For example, Giuliano (1989) developed a lognormal distributed AFT model to determine the traffic incident duration. Chung (2010) found that the log-logistic AFT model fitted the freeway incident duration best. Other parametric models, including time sequential models (Khattak et al., 1995), Bayesian models (Ma and Kockelman, 2008; Kim and Chang, 2012), linear regression models (Valenti et al., 2010), and the mechanism-based approach (Hou et al., 2013) have also been developed for the analysis of traffic incidents. However, it should be pointed out that parametric models have predetermined functional forms and assumptions that are sometimes violated in reality.

Thus, many nonparametric models, such as fuzzy logic models (Kim and Choi, 2001), artificial neural networks (Wang et al., 2005; Wei and Lee, 2007), tree-style classification models (Smith and Smith, 2001; Ma and Kockelman, 2008), and the text analysis approach (Pereira et al., 2013), have been proposed in an attempt to eliminate the drawbacks of parametric models. Ferdous et al. (2011) discussed the event tree and fault tree methods. Fuzzy set theory and evidence theory were used to describe the uncertainties of the inputs associated with the event occurrence likelihoods so as to mitigate the impact of the assumption that all events are independent. Because of their simplicity and high prediction accuracy, many tree-based models have been developed for analyzing accident injury severity. Nevertheless, it is difficult to interpret the marginal effect of influencing factors using nonparametric models, whereas it is a key step for the transportation authority toward establishing priorities for shortening subway incident delays.

Some other researchers have employed least square tree methods (Quinlan, 1992) to analyze accident injury severity. Kuhnert et al. (2000) utilized multivariate adaptive regression splines (MARS), classification and regression trees (CART), and logistic regression to analyze motor vehicle injury data. Yan et al. (2010) adopted the hierarchical tree-based regression (HTBR) method to investigate train–vehicle crashes at highway–rail grade crossings. Weng et al. (2013) presented a tree-based logistic regression approach to assess work zone casualty risk. Their results showed that the proposed approach outperformed the decision tree approach and the logistic regression approach.

However, one shortcoming of the least square tree approach is that it cannot be applied to analyze a dependent variable with a large variance. To overcome this shortcoming, Torgo (2000) built a least absolute deviation regression tree as an extension of the least square regression tree. It was found that the built tree can alleviate the effect of large variance to some extent. Moreover, it performed well at handling data that followed a skewed distribution. Nevertheless, both least absolute deviation regression trees and least square regression trees have poor stability. Hence, some researchers (e.g., Su, 2002; Mohamed et al., 2013) have proposed a maximum likelihood regression tree and found that it has a rigorous mathematical justification and better tree stability than the least square regression tree.

In summary, the existing literature clearly shows that both parametric AFT approaches and tree-based methods are available for traffic safety analysis. However, these approaches cannot be applied to predict subway incident delay for the following reasons. First, the parametric AFT approaches may yield biased and implausible results even though they can be used to interpret the marginal effects of influencing factors on subway incident delay. This is because AFT approaches cannot explain