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).

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.

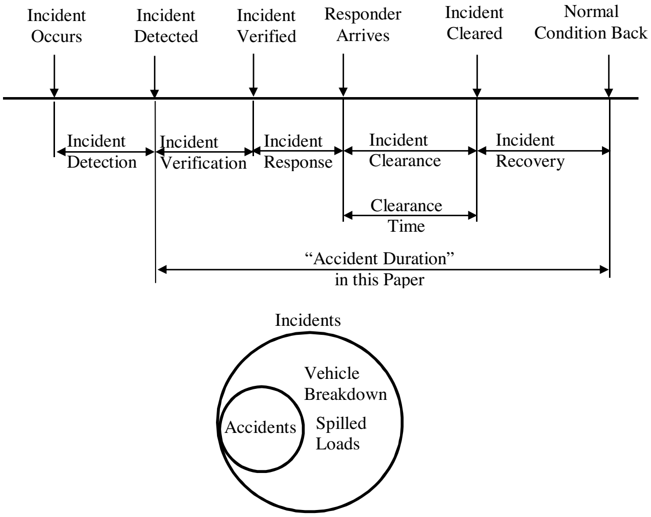


Fig. 1. Traffic incident management process and accident duration definition in this paper.

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).

Clustering and classifying the traffic accident data is another way to minimize the heterogeneity problem. One simple way to classify traffic incidents is based on the traffic incident type (Golob et al., 1987; Giuliano, 1989; Ozbay and Kachroo, 1999). In addition, some researchers recently classified traffic crash data based on factors such as visibility conditions (i.e., daylight, twilight and night conditions (Hong et al., 2014)). A few other clustering methods, including latent class clustering (Depaire et al., 2008), k-means clustering (Anderson, 2009), community detection algorithm (Lin et al., 2014), have also been applied to cluster traffic accident datasets, as a first step before accident analysis.

3. Methodology

As previously mentioned, this paper proposes a new traffic accident duration prediction model M5P-HBDM based on the decision tree model M5P tree and the statistical model HBDM. Traditional decision trees were originally proposed by Breiman et al. (1984). These trees, however, have fixed average values at their leaves that cannot model the stochastic nature of the parent-child relationship in a realistic way (Ozbay and Noyan, 2006). Considering this, Quinlan (1992) developed a new type of a tree named the M5 tree which can have multivariate linear models at its leaves; with this, more flexible predictions are allowed. In order to handle enumerated attributes and attribute with missing values, Wang and Witten (1997) proposed a modified M5 tree algorithm and called it the M5P tree algorithm. M5P tree has the advantages of being able to deal with categorical and continuous variables, and of handling variables with missing values.

The M5P tree has been applied by Zhan et al. (2011) to predict lane clearance time of freeway incidents. One problem with the M5P tree is that given that linear regression $Y=\beta X+\varepsilon$ is used to build the tree's leaves, the residuals ε have to be assumed to be normally distributed. This means that the conditional distribution of accident clearance time Y, given the explanatory variables X, has to be assumed to follow a normal distribution as well. However, the distribution for time to an event (here it is the time when the traffic returns to normal) is almost certainly nonsymmetrical (Cleves, 2008).

HBDM, on the other hand, is a statistical model used to analyze the duration of a specific event. The model allows different distributions of the duration to be assumed (e.g., Weibull distribution, log-normal distribution, log-logistic distribution and so on). The HBDM has been previously applied to analyze and predict incident duration, but on an unclassified dataset (Nam and Mannering, 2000; Chung, 2010; Alkaabi et al., 2011). To the best of the authors' knowledge, previous research did not attempt to combine a classification method with HBDMs. It would be of interest to investigate whether classifying accident dataset would, for example, yield additional insight into the relationship between accident duration and the explanatory variables, and whether the prediction performance can be improved with a combined M5P-HBDM.

The proposed M5P-HBDM retains the superior ability of the M5P tree at classifying traffic accident datasets, but replaces the linear regression models typical of the M5P algorithm with HBDMs, which in turn allows for using the probability distribution that best fits the data. The following section will introduce M5P tree and HBDM first, followed by a detailed description of the proposed the M5P-HBDM and the algorithm developed to construct the model.

3.1. M5P tree algorithm

The M5P tree algorithm mainly includes two steps (Quinlan, 1992; Wang and Witten, 1997): the tree growth step and the tree pruning step. Assume there is a collection of T^n training cases at node n ($n=0$ for the root node), and assume that each case has a fixed set of attributes, either discrete (binary or categorical) or continuous (e.g., visibility), and has a target value (i.e., the traffic accident duration). Before tree construction, all categorical attributes need to be transformed into binary variables. If a categorical attribute has c possible values, it will be replaced by $c-1$ synthetic binary attributes with one representing each possible value. Therefore, after the variable transformation, all splits in a M5P tree are binary.

In the tree growth step, the algorithm firstly calculates the standard deviation sd(T^n) of the target values of the cases in T^n. Assuming that there is a test tree that splits T^n into O outcomes (=2 for a binary split), the objective function is to find the potential test tree that maximizes the reduction in the standard deviation, calculated according to Eq. (1):

$$\Delta \mathrm{sd} = \mathrm{sd}\left(T^n\right) - \sum_{i=1}^{O} \frac{|T_i^n|}{|T^n|} \times \mathrm{sd}\left(T_i^n\right) \tag{1}$$

where T_i^n denote the subset of cases that have the ith outcome of the potential test, sd $\left(T_i^n\right)$ denote the standard deviation of the target values of cases in T_i^n, $|T_i^n|$ denote the number of cases in T_i^n, and $|T^n|$ is the number of cases in T^n. $\sum_{i=1}^{O} \frac{|T_i^n|}{|T^n|} \times \mathrm{sd}\left(T_i^n\right)$ is the weighted average standard deviation after the split.

The same process is applied *recursively* to the subsets, until the subsets at a node either contain only a small number of instances/cases, or their target values show very small variations from one another. This means that there are two termination thresholds for the algorithm: the first is $TH1$, which refers to the minimum number of cases allowed at a node, and the second is $TH2$, which is used to check whether the standard deviation of the target values at the node is less than $TH2 \times \mathrm{sd}(T^0)$. The nodes where the split terminates are marked as "leaf" nodes, whereas the other nodes are marked as interior or non-leaf nodes. After the initial tree has been grown, a multivariate linear model is constructed for each non-leaf node of the model tree by using the standard regression techniques.

In the tree pruning step, starting near the bottom of the tree, the algorithm examines each non-leaf node of the model to determine whether this node should be replaced with the linear model developed above, as a new leaf node, or whether the subtree should be kept intact. The decision is made based upon which approach (i.e., the linear model or the sub-tree) would yield the lower estimated error. The estimated error of the linear model is calculated using Eq. (2):

$$\mathrm{Error} = \frac{N+v}{N-v} \times \frac{\sum_{i=1}^{N} \mathrm{abs}\left(V_{\mathrm{act}} - V_{\mathrm{pre}}\right)}{N} \tag{2}$$

As can be seen, the estimated error is the average absolute difference between the actual target values V_{act} of the training cases and the predicted values, V_{pre}. This is given by the linear model at the current node (or the average target value for the leaf node), and adjusted by $(N+v)/(N-v)$, where N is the number of training cases going through this current node, and v is the number of the parameters in the linear model. For the estimated error of the sub-tree alternative, the error from each branch is combined into a single overall value for the node, using a linear sum in which each branch is weighted by the proportion of the training cases that go down through it (Wang and Witten, 1997).

3.2. Hazard-based duration model

Suppose the duration of a specific traffic accident is represented by a continuous random variable D with a cumulative probability distribution function, $F(d)$. $F(d)$ represents the probability that duration D is less than a time value d, and is called the failure function in HBDM. It is defined as shown in Eq. (3):

$$F(d) = \int_0^d f(u)\,du = P(D < d), 0 < d < \infty \tag{3}$$

The corresponding probability density function is thus given as:

$$f(d) = \frac{\delta F(d)}{\delta d} = \lim_{\Delta d \to 0} \frac{P(d \le D < d + \Delta d)}{\Delta d}, \tag{4}$$

where $f(d)$ describes the instantaneous failure rate in the infinitesimally small interval $[d, d+\Delta d]$. Also given $F(d)$, the survival function, $S(d)$, is defined as in Eq. (5):

$$S(d) = 1 - F(d) = P(D \ge d),$$

where $S(d)$ denotes the probability that the duration D is longer than time value d.

At last, with the probability density function $f(d)$ and the survival function $S(d)$ known, the hazard function $h(d)$ is defined in Eq. (6) as follows:

$$h(d) = \frac{f(d)}{S(d)} = \lim_{\Delta d \to 0} \frac{P(d \le D \le d + \Delta d | D \ge d)}{\Delta d}, \tag{6}$$

where $h(d)$ can be interpreted as the instantaneous failure rate at time d, given that the duration has lasted at least d minutes.

The accelerated failure time model (AFT) is a main approach to investigate the effects of explanatory variables on accident durations using HBDMs (Alkaabi et al., 2011; Chung, 2010). AFT assumes a distribution for

$$\tau = \exp\left(-x_i\beta\right) \times d_i \tag{7}$$

where τ may have a specified distribution like the Weibull distribution, the Log-normal distribution, or the Log-logistic distribution, d_i is the duration of case i, x_i is its value vector of explanatory variables, and β is the vector of estimated coefficients. After taking the logarithm for both sides, the AFT model can be framed as a linear model as shown in Eq. (8):

$$\ln(d_i) = x_i\beta + \ln(\tau) \tag{8}$$

where $\ln(d_i)$ is the natural logarithm of the survival time. With the parameters in β and τ estimated, for a new observation, the mean or median of the failure time distribution can be calculated and used as the prediction for the accident duration (Cleves, 2008).

3.3. M5P-HBDM model

This section will describe the process of building the proposed M5P-HBDM and how it is designed to take advantage of the strengths of each of the M5P and HBDM methods, described above; Appendix A shows the pseudo-codes of the M5P-HBDM algorithm, and compares it with the original M5P algorithm described in Wang and Witten (1997). As can be seen from Appendix A, the building process of the M5P-HBDM model is very similar to that for the M5P model in that the two main steps of tree growth and tree pruning are preserved. Nevertheless, there are a few differences between the original M5P tree and the proposed M5P-HBDM algorithms.

First, in the split step for tree growth, when the stop criteria are met and the node is marked as a leave node, the original M5P tree algorithm uses the average of the target values for that leave node. In the HBDM-M5P algorithm, on the other hand, the algorithm proceeds to build a HBDM model using the training cases at that leave node. If the prediction performance of the HBDM model is better than the constant average value, we use the HBDM model as the model of the leave node.

Second, in the pruning step where a model needs to be built for each interior/non-leaf node, the original M5P tree algorithm (Wang and Witten, 1997) builds a linear regression model for the current node, using only the variables that are referenced by the subtree. The algorithm then greedily drops the variables, if doing so decreases the prediction errors calculated using Eq. (2). This means that the linear regression models in the original M5P algorithm do not consider problems such as whether the variables are significant, or whether the signs of the variables are meaningful. For the M5P-HBDM algorithm, a HBDM model is built for a node using all the variables except those that have been taken by the higher-level nodes in the path from the root to the current node. The prediction performance of a HBDM model, along with the p-values of the variables and the signs of the variables, are all checked to make sure that the variables included are significant and that the signs of their coefficients agree with intuition.

Third, in the proposed M5P-HBDM, the model of the node can consist only of the constant value calculated by taking the average or the median of the target values (which will thus constitute the predicted value of the traffic accident duration). It can also be a HBDM, where the predictions of the target values would be the mean or median value of the AFT with a selected distribution shown in Eq. (8). This is different from the prediction calculation using the constant average value or the linear regression models in the original M5P tree algorithm, as will be explained in more detail later.

4. Modeling datasets

4.1. Virginia traffic accident dataset

The Virginia dataset included traffic accident records reported in 2005 and 2006 on a segment of interstate highway I-64 in Norfolk, Virginia. The accidents were monitored and recorded by Virginia Department of Transportation (VDOT's) Archived Data Management System (ADMS). For this study, 602 accident records were selected; for each record, 17 variables are used to describe the accident. These variables are summarized in Table 1.

As can be seen, there are: (a) three temporal variables in the dataset (season, weekday and hour of the day); (b) one environmental variable (weather conditions); (c) four geographic or spatial variables (direction, location code, lane number at main road, and road structure); and (d) nine accident outcome variables (detection source, accident type, moving to shoulder, fire, roll over, number of vehicles involved, blocked lanes, injured number and duration).

Among the traffic accident relevant variables, the "location code", which takes on values from "1" to "9", refers to the nearest traffic detector code (there are nine detectors in this segment of I-64) to the accident location. "Detection source" is included to investigate whether the accident reporting way has any impact on accident duration. "Accident type" is included, since the type of the accident naturally affects the manner followed to remove the accident, and the equipment used, which in turn may affect accident duration (Chung, 2010). Finally, the variable "Moving to shoulder" is included, because it is generally assumed that moving vehicles to the shoulder after an accident contributes to shorter recovery time and thus shorter accident duration.

4.2. Buffalo–Niagara traffic accident dataset

This dataset included 616 traffic accidents observed on I-190 from 01/01/2008 to 10/31/2012. Incidents and traffic flow infor-

Table 1
Traffic accident variables in I–64 dataset.

Variables	Values
Season	Spring (March–May); Summer (June–August); Autumn (September–November); Winter (December–February)
Weekday	Yes (Monday 2 AM–Friday 9 PM, except holidays); No
Hour of the day	Morning (7 AM–9 AM); early afternoon (10 AM–12 Noon); afternoon (1 PM–3 PM); evening rush (4 PM–6 PM); evening (7 PM–9 PM); night (10 PM–6 AM)
Weather conditions	Clear; Rain; Snow
Direction	East Bound; West Bound
Location code	1; 2; 3; 4; 5; 6; 7; 8; 9 (the codes mean different detectors)
Lane number at main road	2; 3; 4
Road structure	Ramp; Highway
Detection source	CCTV; FIRT; Phone Call; SSP; TMS Camera; VSP CAD; VSP Radio; Other
Accident type	Car; Wrong Way; Truck/Tractor trailer; Motorcycle; car to facility; Others
Moving to shoulder	Yes; No
Fire	Yes; No
Roll over	Yes; No
Number of vehicles involved	1; 2; greater than 2
Blocked lanes	0; 1; 2; 3; 4
Injured number	0, 1, . . .
Duration	0, 1, . . .

Table 2
Traffic accident variables in I-190 dataset.

Variables	Values
Season	Spring (March–May); Summer (June–August); Autumn (September–November); Winter (December–February)
Weekday	Yes (Monday 2 AM–Friday 9 PM, except holidays); No
Hour of the day	Morning (7 AM–9 AM); early afternoon (10 AM–12 Noon); afternoon (1 PM–3 PM); evening rush (4 PM–6 PM); evening (7 PM–9 PM); night (10 PM–6 AM)
Visibility	0–10
Wind speed	0 mph (miles per hour), . . .
Weather conditions	Clear; Rain; Snow
Direction	North Bound; South Bound
Location code	1; 2; . . .; 24; 25; 26 (the codes represent different exits at I-190)
Lane number at main road	2; 3; ≥3
Lane number at ramp	0 (away from exit); 1; 2
Ramp type	On ramp; off ramp; highway to highway on ramp; highway to highway off ramp
Ramp layout	On ramp, off ramp; off ramp, on ramp; only off ramp; only on ramp
Road structure	Before the exit; at the exit; beyond the exit; highway; ramp; bridge; before the bridge; after the bridge
Accident type	Car; Wrong Way; Truck/Tractor trailer; Motorcycle; car to facility; Others
Blocked lane	N/A at main road; Left lane at main road; middle lane at main road; right lane at main road; left two at main road; right two at main road; left and right lanes at main road; all lanes at main road; N/A at ramp; left lane at ramp; right lane at ramp; all lanes at ramp
Blocked lanes number at main road	0; 1; 2; 3
Blocked lanes number at ramp	0; 1; 2
Injured	Yes; No
Roll over	Yes; No
Congestion	Yes; No
Fire	Yes; No
Number of vehicles involved	1; 2; greater than 2
Duration	0, 1, . . .

mation are monitored and recorded by the Niagara International Transportation Technology Coalition (NITTEC), which serves as the region's Traffic Operations Center (TOC). Incident details are recorded every day through detailed incident log forms, which formed the basis for compiling the dataset used in this study. Table 2 summarizes the variables included in the Buffalo–Niagara dataset.

In this dataset, there are 23 variables in total for each accident record. The three temporal variables are the same as those in the I-64 dataset: season, weekday and hour of the day. There are: (a) three environmental variables: visibility, wind speed and weather conditions; (b) seven geographic or spatial variables: direction, location code, lane number on main road, lane number on ramp, ramp type, ramp layout and road structure; and (c) ten accident outcome variables: accident type, block lane index, blocked lanes number at main road, blocked lanes number at ramp, injured, roll over, congestion, fire, number of vehicles involved and clearance time.

The "Location code" variable in this dataset can range from "1" to "26", and refers, in this case, to the ID of the nearest exit from the accident location. For example, "1" means the accident is closest to Exit 1 on I-190. "Ramp type" can be one of the following: (1) a "highway to highway on ramp"; or (2) "highway to highway off ramp", since I-190 is connected to other two highways "I-290" and "I-90". If the ramp is from the other highway to I-190, we classify the ramp as "highway to highway on ramp". "Ramp layout" is the layout of the ramps at the exit. The relative location order of "on-ramps" and "off-ramps" may impact the accident duration. "Blocked lane" records the blocked lane at the main road or the ramp, as a result of the traffic accident.

Comparing the two datasets, we can see that the records have different emphasis on traffic accidents characteristics. The I-64 accident dataset records detailed information about moving the vehicles to the shoulder and the detection source. In contrast, the I-190 accident dataset includes information such as on which lane the accident occurred, whether the accident happened on the mainline or on the ramp, among other attributes.

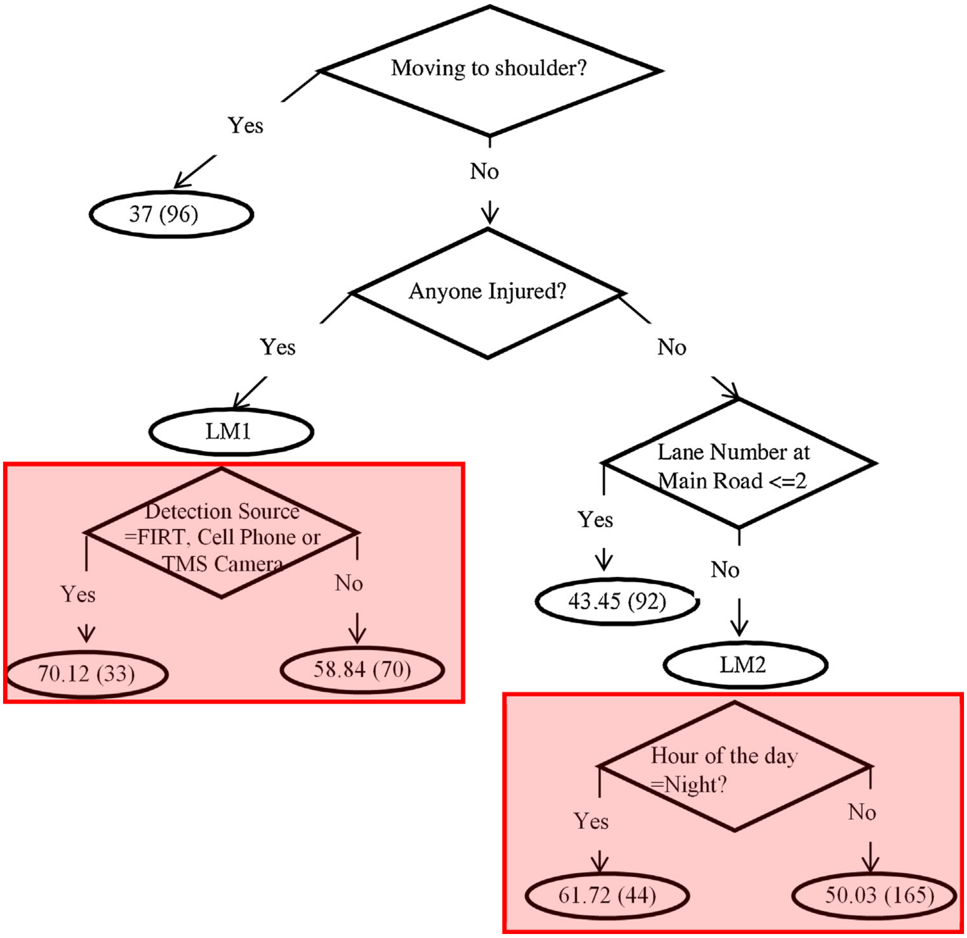


Fig. 2. M5P tree model for I-64 training dataset. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

5. Model development

As mentioned before, the I-64 dataset included 602 traffic accident records and the I-190 dataset included 616 traffic accident records. For each dataset, the first 500 records were used for model training, and the remainder data points for testing. For each dataset, three different models are developed: (1) a stand-alone M5P tree; (2) a stand-alone HBDM; and (3) the proposed combined M5P-HBDM.

5.1. M5P tree

In this study, a Matlab package called M5PrimeLab (Jekabsons, 2010) was used for the M5P tree model development. To build the tree, the modeler needs first to decide upon the values of the two thresholds, namely: (1) the minimum number of training records at one node *TH*1; and (2) the ratio of the standard deviation *TH*2, mentioned in Section 3.1.

Although the value of *TH*1 can be set as low as 2, it is generally not desirable for a non-lead node to have too few records, in order to allow for building good linear regression models after the tree growth step. In this study, we experimented with *TH*1 values ranging from 5% to 10% of the total number of training cases (i.e., values between 25 and 50). After some experimentation, *TH*1 was set to 30, and *TH*2 was set to 0.95. Fig. 2 shows the resulting M5P tree model for the I-64 dataset.

As can be seen in Fig. 2, for some leaf nodes, there are a constant value and a number in the parenthesis. The constant value is the average of the accident durations (in minutes) for the cases in that node, and the number in parenthesis is the number of those cases. There are also two linear models in two leaf nodes, LM1 and LM2. In the tree pruning step, these two models replaced the original sub trees (enclosed by the red rectangles in Fig. 2).

The details of LM1 and LM2 are listed below.

LM1: duration = 62.46 min (103 cases);

LM2: duration = 52.49 min (209 cases).

As can be seen, the two linear regression models developed here are basically two constants. As discussed in Section 3.3, after building a linear regression model for an interior node, the M5P algorithm uses a greedy search to remove variables that do not improve the predictions for the cases going through that node. In our case, the algorithm ended up removing all variables, and the linear models ended up with just the constant. The number in the parentheses refers to the number of training cases at that leave.

Insight into the factors affecting accident duration can be gained from studying the developed tree. First from the splitting rule at the root node, it can be seen that if the vehicles involved were moved to the shoulder once the accident happened, the average accident duration was only 37 min. On the other hand, if the vehicles were not moved to shoulder, the duration was significantly longer. Specifically, with the vehicles not moved to the shoulder and with someone injured, the accident duration was estimated to be as long as 62.46 min (according to the LM1 model). With no injury, involved vehicles not moved to the shoulder, and when the number of lanes on the freeway equal to 2, the accident duration was estimated to be equal to 43.45 min, which is shorter than the cases when the accidents happened on freeways with more than 2 lanes (for that case, the estimated duration was 52.49 min as given

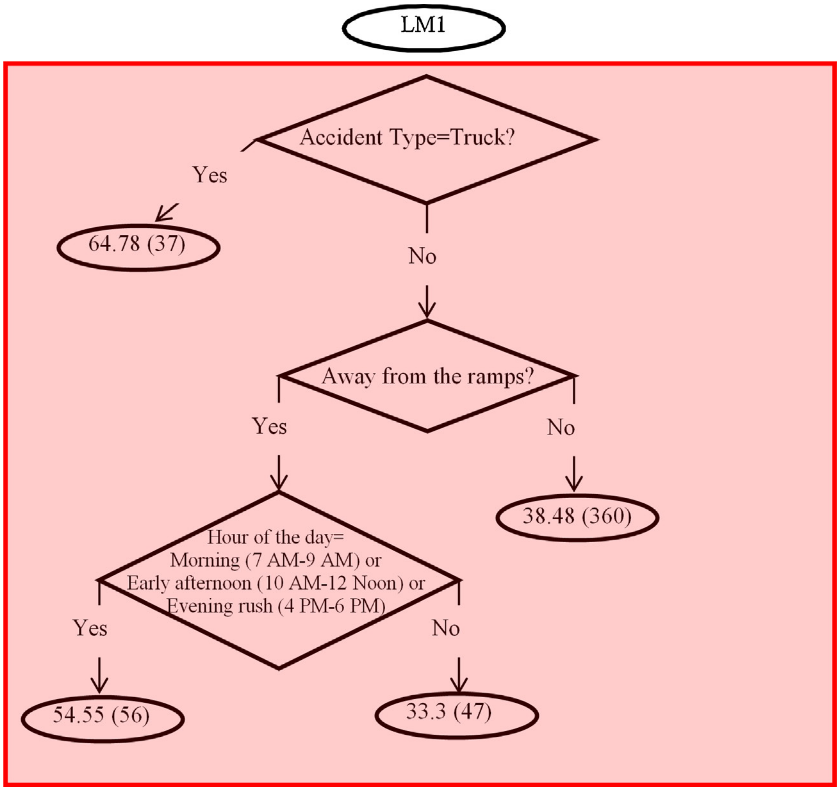


Fig. 3. M5P tree model for I-190 training dataset.

by LM2). This is probably because there is lighter traffic on freeways with lower number of lanes.

Similarly, an M5P tree was developed for the Buffalo–Niagara I-190 accident dataset. After experimentation as before, *TH*1 was set to 35, and *TH*2 to 0.75. Fig. 3 shows the M5P tree model that resulted.

We can see that this is an extreme situation for the algorithm, when the whole grown M5P tree is replaced by one linear regression model LM1 in the tree pruning step (shown below).

LM1: duration = 37.95 + 6.92*Hour of the day = Morning (7 AM–9 AM) or Early afternoon (10 AM–12 Noon) or Evening rush (4 PM–6 PM)? (500 cases).

The developed LM1 shows that the estimated duration of an accident is at least 37.95 min, and that there is only one independent variable, which is the "hour of the day". If the hour is one of the following time intervals, the morning (7 AM–9 AM) period, or early afternoon (10 AM–12 Noon) or evening rush (4 PM–6 PM)" hour, the duration will be increased by 6.92 min.

In conclusion, it can be seen that while the tree pruning step of the original M5P is designed to allow for the use of the linear regression model when it can bring the lower estimated error calculated in Eq. (2), that step has resulted, for both the Virginia and Buffalo datasets utilized in this study, in models with very weak explanatory power (i.e., few independent variables).

5.2. Hazard-based duration model

Before applying HBDM models, there are two issues that need to be addressed. First, a probability distribution form needs to be specified for τ in Eq. (8) (Section 3.2). Secondly, the significant explanatory variables x_i need to be determined. In this paper, we followed the four-step procedure outlined, aided by STATA software, to develop the HBDM (Collett, 2003; Alkaabi et al., 2011).

1. Fit models using exponential, Weibull, Log-normal, Log-logistic and Generalized Gamma models with no explanatory variables. Record the log likelihood for each model.
2. For each model, add the explanatory variables from the candidate variable list, one by one, test the new model, and select the one which increased the log likelihood the most.
3. For each model, repeat step 2 by adding one additional variable from the remainder of the candidate variables. Stop when no variable can increase the log likelihood.
4. For each model, calculate the value of the Akaike information criterion (AIC), which can be calculated as shown in Eq. (9) below (Alkaabi et al., 2011; Cleves, 2008):

$$\mathrm{AIC} = -2\ln L + 2(k + c) \tag{9}$$

where L is the likelihood, k is the number of model covariates, and c is the number of model-specific distributional parameters. Finally select the model with the lowest value of AIC as the HBDM model.

The AIC values of the HBDMs developed for the I-64 and I-190 datasets are listed in Table 3. As can be seen, for both the I-64 and the I-190 datasets, the HBDM model with the log-normal distribution had the lowest AIC, and hence this was the model employed to analyze the accident duration in this paper. It is to be noted that this is consistent with other studies reported in the literature (Golob et al., 1987; Chung, 2010).

For the log-normal regression AFT model, τ is distributed as log-normal with parameters (β_0, σ). The log-normal AFT function can thus be expressed as in Eq. (10) below (Cleves, 2008):

$$\ln(d_i) = \beta_0 + x_i\beta + \mu \tag{10}$$

where μ follows a normal distribution with mean 0 and standard deviation σ.

For the I-64 dataset, Table 4 shows the estimated coefficients of the explanatory variables, the standard error, the *P*-value, and percentage change (%) for the log-normal AFT model. The percent-

Table 3
AIC values of HBDMs for I-64 and I-190 training datasets.

Model	I-64 dataset				I-190 dataset			
	$-2\ln L$	k	c	AIC	$-2\ln L$	k	c	AIC
Exponential	1169.42	9	1	1179.42	1223.04	2	1	1226.04
Weibull	952.92	9	2	963.92	1105.78	9	2	1116.78
Log-normal	949.08	6	2	957.08	1107.72	3	2	1112.72
Log-logistic	954.62	8	2	964.62	1186.34	9	2	1197.34
Generalized gamma	957.24	5	3	965.24	1185.3	9	3	1197.3

Table 4
Log-normal AFT models on I-64 training dataset.

Variable	Coefficient	Standard error	*P* value	Percentage change (%)
Night	0.19	0.07	0.016	21
Move to shoulder?	−0.36	0.07	0.000	−30
Road structure	0.26	0.10	0.017	30
Injured number	0.22	0.04	0.000	25
Detection = 7 (VSP Radio)	−0.16	0.08	0.025	−15
Roll over	0.51	0.25	0.041	67
β_0	3.41	0.11	0	
σ	0.62	0.02		

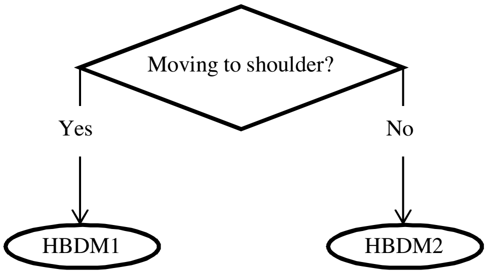


Fig. 4. M5P-HBDM model for I-64 training dataset.

age change represents the change in the duration of the incident resulting from a one unit change in the value of the variable under consideration. According to Eq. (7), that percentage change (%) can be calculated by taking the exponent of the estimated coefficient of the significant independent variable. For example, the coefficient of variable “night” in Table 4 has a positive value 0.19, its exponential value is exp(0.19) = 1.21. This means that the incident duration would be expected to be 21% longer when the accident occurs at night. This underscores the importance of ensuring there is enough staff at night for incident clearance work. On the other hand, if the coefficient of a variable has a negative value, the contribution of that variable is a *decrease* in the incident duration.

As shown in Table 4, the variables that are likely to result in an increase in the traffic accident duration include the following variables: (1) “Night”, 21% longer when the accident occurs at night defined as between 10 PM and 6 AM; (2) “road structure”, 30% longer when the accident occurs near the ramp; (3) the number of injured people, 25% longer every time one more person gets injured; and (4) whether the accident involved a roll over, 67% longer than those not involving one. In contrast, moving an accident to the shoulder can decrease the accident duration by 30%. Also, according to this particular dataset, the detection source (if it were to be from source 7—the VSP (Virginia State Police) Radio), can lead to a 15% decrease in incident duration. This may be because the response in this case by the police may be immediate, compared to other sources of detection.

Similarly, Table 5 lists the coefficients of the significant independent variables, along with the corresponding standard error, P-value, and percentage change (%), for the log-normal AFT model of the I-190 training dataset (i.e., the Buffalo–Niagara dataset).

As can be seen, the only variable with negative percentage change (%) is the variable “Afternoon” (1 PM–3 PM), which shows that if the accident were to happen during this time interval, the duration would be 15% shorter. Also similar to the results for the I-64 training dataset, the rolling over of the involved vehicles can lead to a dramatic increase in the accident duration (in this case of about 129%). Finally, the duration of the accident is increased by 23% with the additional involvement of a single vehicle in the accident.

5.3. M5P-HBDM model

Now with the stand-alone M5P and HBDM models developed for the two datasets, the study proceeded to construct the new M5P-HBDM proposed herein, following the procedure described in Section 3.3. Fig. 4 shows the M5P-HBDM built for the I-64 or the Virginia training dataset.

As can be seen from Fig. 4, the M5P-HBDM model has only one splitting rule, namely “moving to shoulder?”. The AIC test shows the log-normal distribution is still the best assumption for the accelerated failure time functions of HBDM1 and HBDM2. Table 6 shows the relevant parameters for the two models.

As can be seen, for the log-normal AFT model HBDM1, based on the 96 cases in which the involved vehicles are moved to the shoulder, no significant variables are found; only the constant β_0 and the sigma in the log-normal distribution are estimated.

On the other hand, for the HBDM2 based on the 404 cases when the involved vehicles are not moved to the shoulder, one can make a few additional observations, beyond the insight made possible from the stand-alone HBDM. First, the “blocked lane number” variable shows that one more lane being blocked can increase the accident duration by 6%. Second, the detection source “detection = 5” (TMS camera) demonstrates that the accidents detected by camera have a longer duration (this was also shown in the M5P tree model before pruning in Fig. 2). Finally, one more observation is that if the vehicle in the traffic accident is on fire, the duration is likely to increase by 12%.

Similarly, the M5P-HBDM of I-190 dataset is shown in Fig. 5. Note that the 44 min for the branch with the 37 cases is the *median* value which we found to be more accurate than the mean, based on the estimated error in the building process of M5P-HBDM.

For both HBDM1 and HBDM2, the AIC test still shows that the log-normal distribution appears to be the best assumption for the

Table 5
Log-normal AFT models on I-190 training dataset.

Variable	Coefficient	Standard error	*P* value	Percentage change (%)
Afternoon (1 PM–3 PM)	−0.16	0.10	0.007	−15
Roll over?	0.83	0.26	0.001	129
Vehicle number	0.21	0.10	0.050	23
β_0	3.06	0.20	0	
σ	0.75	0.02		

Table 6
Log-normal AFT models in M5P-HBDM of I-64 training dataset.

Branches	Variable	Coefficient	Standard error	*P* value	Percentage change (%)
HBDM1 (96 cases)	β_0	3.36	0.08	0	
	σ	0.74	0.05		
HBDM2 (404 cases)	Night	0.14	0.07	0.06	15
	Blocked lane number	0.06	0.04	0.007	6
	Road structure	0.27	0.10	0.005	31
	Injured number	0.18	0.05	0.000	20
	Detection = 5 (TMS Camera)?	0.06	0.07	0.007	6
	Detection = 7 (VSP Radio)?	−0.13	0.09	0.008	−12
	Roll over?	0.54	0.27	0.05	72
	Fire or not?	0.11	0.09	0.02	12
	β_0	3.31	0.12	0	
	σ	0.60	0.02		

Table 7
Log-normal AFT models in M5P-HBDM of I-190 training dataset.

Branches	Variable	Coefficient	Standard error	*P* value	Percentage change (%)
HBDM1 (103 cases)	Evening rush (4 PM–6 PM)	0.44	0.22	0.05	55
	β_0	3.38	0.10	0	
	σ	0.87	0.06		
HBDM2 (360 cases)	Morning (7 AM–9 AM)	0.06	0.11	0.02	6
	Afternoon (1 PM–3 PM)	−0.21	0.11	0.05	−19
	Vehicle number	0.32	0.14	0.007	38
	Location = exit 16	0.34	0.14	0.019	40
	Main road lane number = 2	−0.90	0.67	0.02	−59
	Main road lane number = 3	−0.96	0.67	0.01	−62
	β_0	3.72	0.73	0	
	σ	0.67	0.02		

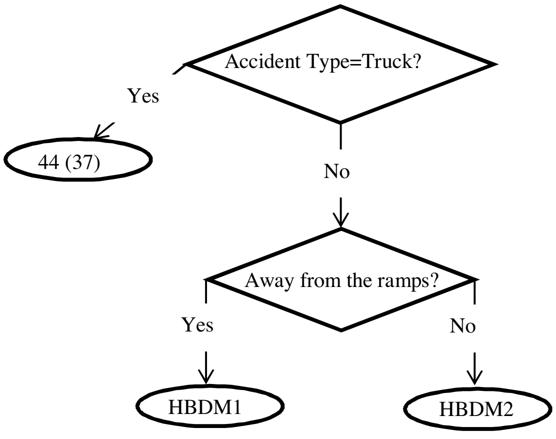


Fig. 5. M5P-HBDM model for I-190 training dataset.

AFT functions. The relevant parameters of HBDM1 and HBDM2 are shown in Table 7.

From Table 7, we can see that for HBDM1 based on the cases when the accidents happen away from the ramps, the accident duration is increased by 55% if the accident were to occur during the evening rush period (4 PM–6 PM).

The main interesting observations from the HBDM2 based on the 360 cases include the following. First, from the values of the percentage change caused by the variables, “Main Road Lane Number = 2” and “Main Road Lane Number = 3”, we can see that, for accidents not involving trucks that occur close to the ramps, roads with three lanes can help reduce the accident duration by 62%, which is a little larger than the 59% reduction, under the scenario when the main road close to the ramps has two lanes. Although the main road with three lanes signify that there are more traffic going through the ramp than the road with two lanes, it is also possible that a wider main road can provide more space for the accident clearance work and therefore the duration could be shorter. Secondly, the data shows that accidents happening at Exit 16 (I-190/I-290 Interchange) have significantly longer durations than those occurring elsewhere (40% longer). This observation makes perfect sense, given the extremely high volumes at the I-190 and I-290 interchange in Buffalo. In fact, our previous research also showed that Exit 16 is one of the accident hotspots on I-190 (Lin et al., 2014), as well as one where significant traffic and weaving maneuvers take place all the time.

6. Model comparison

In order to demonstrate the advantages of the proposed M5P-HBDM, this section will compare the significant independent variables identified by the three models and for each dataset, as well as the models' prediction accuracy.

Table 8
Significant variables in M5P, HBDM and M5P-HBDM of I-64 training dataset.

I-64 training dataset	M5P	HBDM	M5P-HBDM
Lane number at main road ≤2?	X (−)		
Move to shoulder?	X (−)	X (−)	R
Injured number	X (+)	X (+)	X (+)
Road structure (0 for highway, 1 for ramp)		X (+)	X (+)
Hour of the day = night?		X (+)	X (+)
Roll over?		X (+)	X (+)
Detection source = Virginia State Police Radio		X (−)	X (−)
Detection source = Camera?			X (+)
Blocked lane number at main road			X (+)
Fire or not?			X (+)

6.1. Significant independent variables comparison

Table 8 lists all the significant variables identified by each of the M5P, HBDM and M5P-HBDM models, for the I-64 training dataset. The significant variables for the corresponding model are marked with "X", and the sign in the parenthesis indicates the impact of that variable in terms of increasing (a plus sign) or decreasing (a minus sign) the accident duration. The symbol "R" indicates that the variable resulted in a splitting rule for the model.

As can be seen from Table 8, the M5P model helped identify only three significant independent variables affecting accident duration. HBDM, on the other hand, identified six significant variables, whereas eight significant variables and one splitting rule were identified by the M5P-HBDM model. Two significant variables "moving to shoulder?" and "injured number" were identified by all the three models.

Similarly, the significant independent variables identified by the M5P, HBDM and M5P-HBDM models for the I-190 training dataset are summarized in Table 9. As can be seen, the number of significant variables identified by M5P-HBDM far exceeds those identified by either the stand-alone M5P or the stand-alone HBDM. Specifically M5P-HBDM identified eight significant variables and two splitting rules, compared to only three significant variables identified by either HBDM or M5P.

Finally, when considering the modeling results for both the I-64 and I-190 training datasets, it can be clearly seen that the use of the proposed combined M5P-HBDM helps identify the largest number of significant independent variables. Specifically, the tree growth step (the splitting rule at the nodes of M5P tree which attempts to find the variable that can bring the maximum reduction in the standard deviation of the target value or accident duration in this case), helps alleviate the data heterogeneity problem, and thus helps reveal more previously unobserved factors that impact traffic accident duration. This agrees with our previous research (Lin et al., 2014).

6.2. Accident duration prediction comparison

The prediction accuracy of the three models was compared, using a test set not previously utilized in model development. As mentioned before, this test set consisted of 102 records from the I-64 data set, and 116 records from the I-190 set. For prediction performance evaluation, the Mean Absolute Percentage Error (MAPE), a widely used measure to assess the accuracy of models developed, was utilized. MAPE can be calculated as follows:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - P_i}{A_i}\right| \tag{11}$$

where A_i is the ith actual value, P_i is the ith predicted value.

To calculate the predictions, for the M5P tree model, each testing record will be directed toward the corresponding leave, and the linear functions, or the mean target values at that leave, will be used to estimate the accident duration. For HBDMs, the mean and the median values of the survival time (accident duration) for the log-normal AFT models are calculated and used for prediction (the study calculated both the median and the mean values to see which approach yielded better predictive accuracy). This is shown below, in Eqs. (12) and (13) after the relevant parameters and variable coefficients in the log-normal AFT models are estimated.

$$\mathrm{Median}\,(d_i) = \exp\left(\beta_0 + x_i\beta\right) \tag{12}$$

$$\mathrm{Mean}\,(d_i) = \exp\left(\beta_0 + x_i\beta + \sigma^2/2\right) \tag{13}$$

For M5P-HBDMs, similar to the M5P tree, the testing record is first directed toward the corresponding leave. If there is no log-normal AFT model at the leave, as was mentioned earlier in Section 5.3, we use the median value of the cases at that node as the prediction. Otherwise, if there is a log-normal AFT model at the corresponding leave, the predictions can be derived by using the same method as in HBDMs based on Eq. (12) or (13).

Table 10 shows the MAPEs of the M5P tree model, HBDM model and the M5P-HBDM model for the two testing datasets. In Table 10, the column labeled "HBDM (median)" lists the HBDM's MAPE resulting from using the median values of the survival times using Eq. (12), whereas the column entitled "HBDM (mean)" lists the model's MAPE resulting from using the mean values calculated using Eq. (13). The same is true for the columns entitled M5P-HBDM (median) and M5P-HBDM (mean) in connection with the M5P-HBDM.

Firstly, as can be seen, our experiments in this study seem to indicate that the use of the median values of the survival time results in better prediction performance compared to the mean values for both HBDMs and M5P-HBDMs. Secondly, for the I-64 testing dataset, the lowest MAPE was 36.20% given by the M5P-HBDM (median), followed by the HBDM (median) with an MAPE of 38.32%. The MAPE of the M5P model is the highest (i.e., 48.69%). For I-190 testing dataset, the M5P-HBDM (median) still had the best prediction performance with an MAPE value equal to 31.87%, followed by M5P-HBDM (mean), HBDM (median), HBDM (mean) and then M5P. It thus seems that, regardless of the testing dataset, the M5P-HBDM model based on the median value of AFT model appears to perform the best.

Finally, considering that the M5P tree model previously developed by Zhan et al. (2011) for accident duration had an MAPE of 42.70%, as reported in the literature, and that HBDM traffic accident duration prediction developed by Chung (2010) had an MAPE of 47%, our results appear to be superior to those of previous studies.

7. Conclusions and future work

This study has proposed 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

Table 9
Significant variables in M5P, HBDM and M5P-HBDM of I-190 training dataset.

I-190 training dataset	M5P	HBDM	M5P-HBDM
Hour of the day = morning (7 AM–9 AM)	X (+)		X (+)
Hour of the day = early afternoon (10 AM–12 Noon)	X (+)		
Hour of the day = evening rush (4 PM–6 PM)	X (+)		X (+)
Hour of the day = afternoon (1 PM–3 PM)		X (−)	X (−)
Vehicle number		X (+)	X (+)
Roll over?		X (+)	X (+)
Location = exit 16			X (+)
Lane number at main road = 2			X (−)
Lane number at main road = 3			X (−)
Accident type = truck?			R
Away from the ramps?			R

Table 10
MAPEs of M5P tree, HBDM model and M5P-HBDM model for I-64 and I-190 testing datasets.

Datasets	M5P	HBDM (median) (%)	M5P-HBDM (median) (%)	HBDM (mean) (%)	M5P-HBDM (mean) (%)
I-64	48.69	38.32	36.20	41.21	39.10
I-190	38.45	33.61	31.87	35.21	33.15

models. Two traffic accident duration datasets, namely the I-64, Virginia and I-190, Buffalo data sets, were then used to construct and evaluate the performance of three modeling approaches, namely a stand-alone M5P tree, a stand-alone HBDM, and the proposed M5P-HBDM model. Among the main conclusions of the study with respect to the proposed new algorithm are:

1. Thanks to the tree growth step of the M5P algorithm, the proposed M5P-HBDM is able to reduce data heterogeneity through the splitting rules at the nodes. With this, the new algorithm is able to identify more factors as significantly affecting incident duration, compared to either M5P or HDBM alone. This is also consistent with our previous research which shows the advantages of reducing data heterogeneity through dataset grouping and clustering.
2. Because M5P-HBDM can build an AFT model as its leave, and since the AFT model does not need to assume that the conditional distribution of traffic accident durations, given the independent variables, follows the normal distribution (as was the case with the linear regression model in M5P model), the analyst is free to experiment with other distributions such as the Weibull distribution, the log-normal distribution, the log-logistics. In this study, we found that the log-normal AFT model appeared to be the best choice, based on the AIC values.
3. The comparison of the prediction performances of the three models shows that, for both testing data sets, the M5P-HBDM based on the median value of the survival time for the log-normal AFT model always had the lowest overall MAPE.

For future research, one possible idea to investigate, involves combining the M5P tree algorithm with a *random parameter* HBDM. This may further improve accident duration prediction, by allowing the coefficients of the variables in the model to vary across each individual observation in the dataset. Another possible idea is to test the transferability of M5P-HBDM by building a unique model for two or more datasets.

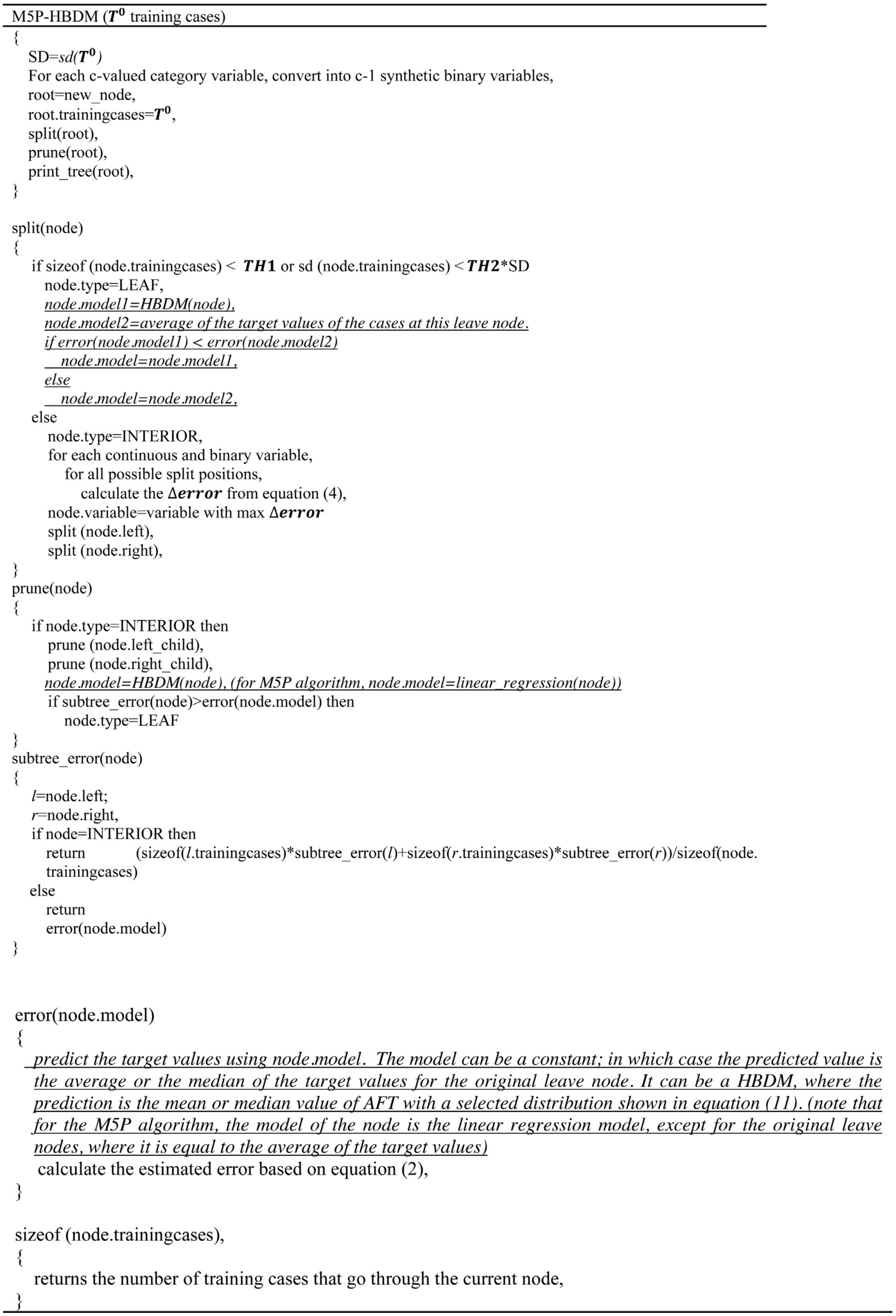
Acknowledgements

This research was partially funded by the Transportation Informatics (TransInfo) Tier I University Transportation Center (UTC) headquartered at the University at Buffalo. TransInfo is one of the UTCs established by the U.S. Department of Transportation's Office of the Assistant Secretary for Research and Technology (OST-R). The authors would like to thanks OST-R for its support.

Appendix A. Pseudo-code of M5P-HBDM algorithm and comparison with M5P tree

Pseudo-code of M5P-HBDM algorithm and comparison with M5P tree

The different parts of between M5P tree and M5P-HBDM algorithms are marked as bold, italic and underlined in the following table.



```
M5P-HBDM ($T^0$ training cases)
{
  SD=sd($T^0$)
  For each c-valued category variable, convert into c-1 synthetic binary variables,
  root=new_node,
  root.trainingcases=$T^0$,
  split(root),
  prune(root),
  print_tree(root),
}

split(node)
{
   if sizeof (node.trainingcases) < $TH1$ or sd (node.trainingcases) < $TH2$*SD
     node.type=LEAF,
     node.model1=HBDM(node),
     node.model2=average of the target values of the cases at this leave node.
     if error(node.model1) < error(node.model2)
       node.model=node.model1,
     else
       node.model=node.model2,
   else
      node.type=INTERIOR,
      for each continuous and binary variable,
         for all possible split positions,
            calculate the $\Delta error$ from equation (4),
      node.variable=variable with max $\Delta error$
      split (node.left),
      split (node.right),
}
prune(node)
{
   if node.type=INTERIOR then
      prune (node.left_child),
      prune (node.right_child),
     node.model=HBDM(node), (for M5P algorithm, node.model=linear_regression(node))
      if subtree_error(node)>error(node.model) then
         node.type=LEAF
}
subtree_error(node)
{
   l=node.left;
   r=node.right,
   if node=INTERIOR then
     return        (sizeof(l.trainingcases)*subtree_error(l)+sizeof(r.trainingcases)*subtree_error(r))/sizeof(node.
     trainingcases)
   else
     return
     error(node.model)
}


error(node.model)
{
  predict the target values using node.model. The model can be a constant; in which case the predicted value is
  the average or the median of the target values for the original leave node. It can be a HBDM, where the
  prediction is the mean or median value of AFT with a selected distribution shown in equation (11). (note that
  for the M5P algorithm, the model of the node is the linear regression model, except for the original leave
  nodes, where it is equal to the average of the target values)
   calculate the estimated error based on equation (2),
}

sizeof (node.trainingcases),
{
  returns the number of training cases that go through the current node,
}
```

References

Alkaabi, A., Dissanayake, D., Bird, R., 2011. Analyzing clearance time of urban traffic accidents in Abu Dhabi, United Arab Emirates, with hazard-based duration modeling method. Transp. Res. Rec. 2229, 46–54.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accid. Anal. Prev. 41 (3), 359–364.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Chin, S.M., Franzese, O., Greene, D.L., Hwang, H.L., Gibson, R.C., 2004. Temporary Losses of Highway Capacity and Impacts on Performance: Phase 2. Department of Energy, United States.

Chung, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems. Accid. Anal. Prev. 42 (1), 282–289.

Cleves, M., 2008. An Introduction to Survival Analysis Using Stata. Stata Press.

Collett, D., 2003. Modelling Survical Data in Medical Research, vol. 57. CRC Press.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40 (4), 1257–1266.

Farradyne, P.B., 2000. Traffic Incident Management Handbook. Prepared for Federal Highway Administration, Office of Travel Management.

Garib, A., Radwan, A.E., Al-Deek, H., 1997. Estimating magnitude and duration of incident delays. J. Transp. Eng. 123 (6), 459–466.

Giuliano, G., 1989. Incident characteristics, frequency, and duration on a high volume urban freeway. Transp. Res. Part A Gen. 23 (5), 387–396.

Golob, T.F., Recker, W.W., Leonard, J.D., 1987. An analysis of the severity and incident duration of truck-involved freeway accidents. Accid. Anal. Prev. 19 (5), 375–395.

He, Q., Kamarianakis, Y., Jintanakul, K., Wynter, L., 2013. Incident duration prediction with hybrid tree-based quantile regression. In: Advances in Dynamic Network Modeling in Complex Transportation Systems. Springer, New York, pp. 287–305.

Hong, S., Kim, J., Oh, C., Ulfarsson, G.F., 2014. The effect of road environment factors on freeway traffic crash frequency during daylight, twilight, and night conditions. Transportation Research Board 93rd Annual Meeting (No. 14-2418).

Jekabsons, G., 2010. M5PrimeLab: M5′ regression tree and model tree toolbox for Matlab/Octave. Available at http://www.cs.rtu.lv/jekabsons/.

Karlaftis, M.G., Tarko, A.P., 1998. Heterogeneity considerations in accident modeling. Accid. Anal. Prev. 30 (4), 425–433.

Khattak, A., Schofer, J., Wang, M.-H., 1995. A simple time sequential procedure for predicting freeway incident duration. IVHS J. 2 (2), 113–138.

Lee, Y., Wei, C.H., 2010. A computerized feature selection method using genetic algorithms to forecast freeway accident duration times. Comput. Aided Civil Infrastruct. Eng. 25 (2), 132–148.

Lerman, K., 2013. The Curse of Heterogeneity in Big Data. http://wp.sigmod.org/?p=960.

Lin, L., Wang, Q., Sadek, A., 2014. Data mining and complex network algorithms for traffic accident analysis. Transp. Res. Rec. 2460, 128–136.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A Policy Pract. 44 (5), 291–305.

Miaou, S.P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space-time modeling approach. J. Transp. Stat. 6, 33–58.

Nam, D., Mannering, F., 2000. An exploratory hazard-based analysis of highway incident duration. Transp. Res. Part A Policy Pract. 34 (2), 85–102.

Ozbay, K., Kachroo, P., 1999. Incident Management in Intelligent Transportation Systems. Artech House, Bonston.

Ozbay, K., Noyan, N., 2006. Estimation of incident clearance times using Bayesian Networks approach. Accid. Anal. Prev. 38 (3), 542–555.

Quinlan, J.R., 1992. Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence, vol. 92, 343–348.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43 (5), 1666–1676.

Smith, K., Smith, B., 2001. Forecasting the Clearance Time of Freeway Accidents. Center for Transportation Studies, University of Virginia.

Valenti, G., Lelli, M., Cucina, D., 2010. A comparative study of models for the incident duration prediction. Eur. Transp. Res. Rev. 2 (2), 103–111.

Vaupel, J.W., Yashin, A.I., 1985. Heterogeneity's ruses: some surprising effects of selection on population dynamics. Am. Stat. 39 (3), 176–185.

Wang, Y., Witten, I.H., 1997. Inducing model trees for continuous classes. Proceedings of the Ninth European Conference on Machine Learning, 128–137.

Wei, C.H., Lee, Y., 2007. Sequential forecast of incident duration using Artificial Neural Network models. Accid. Anal. Prev. 39 (5), 944–954.

Zhan, C., Gan, A., Hadi, M., 2011. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. IEEE Trans. Intell. Transp. Syst. 12 (4), 1549–1557.