TransInfo
Tier I University Transportation Center

FINAL REPORT

Novel Machine Learning Methods for Accident Data Analysis

Date of report: January 2019

Lei Lin, Ph.D.
Graduate Research Assistant, University at Buffalo
Research Scientist, Goergen Institute for Data Science at the University of Rochester

Qian Wang, Ph.D.
Teaching Assistant Professor, University at Buffalo

Adel W. Sadek, Ph.D.
Professor, University at Buffalo
Director, Transportation Informatics University Transportation Center
Associate Director, Institute for Sustainable Transportation & Logistics

Prepared by:
Department of Civil, Structural & Environmental Engineering, Univ. at Buffalo

Prepared for:
Transportation Informatics Tier I University Transportation Center
204 Ketter Hall
University at Buffalo
Buffalo, NY 14260

1. Report No.

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
Novel Machine Learning Methods for Accident Data Analysis

5. Report Date

January 2018

6. Performing Organization Code

7. Author(s)
Lei Lin & Adel W. Sadek

8. Performing Organization Report No.

9. Performing Organization Name and Address
Department of Civil, Structural and Environmental Engineering
University at Buffalo
204 Ketter Hall
Buffalo, NY 14260

10. Work Unit No. (TRAIS

11. Contract or Grant No.
DTRT13-G-UTC48

12. Sponsoring Agency Name and Address
US Department of Transportation
Office of the
UTC Program, RDT-30
1200 New Jersey Ave., SE
Washington, DC 20590

13. Type of Report and Period Covered

Final: June 2014 – January 2018

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The field of traffic accident analysis has long been dominated by traditional statistical analysis. With the recent advances in data collection, storage and archival methods, the size of accident datasets has grown significantly. This in turn has motivated research on applying data mining and Machine Learning algorithms, which are specifically designed to handle datasets with large dimensions, to traffic accident analysis. This project explores three specific applications of Data Mining and Machine Learning algorithms to traffic accident analysis. The first application explores the potential for using a modularity-optimizing community detection algorithm and association rules learning algorithm, to identify important accident characteristics. The second application proposes a novel Frequent Pattern tree (FP tree) based variable selection method, and then develops models for the real-time prediction of traffic accident risk. Finally, the third application proposes a novel approach to developing accident duration prediction models. The approach improves on the original M5P tree algorithm through the construction of a M5P-Hazard-Based Duration Model (HBDM).

17. Key Words
Data mining; Complex Network Analysis; Frequent Pattern tree (FP tree); Fuzzy C-means clustering (FCM); Bayesian network; Random forest; M5P Tree; Hazard-based Duration Model.

18. Distribution Statement

No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)

Unclassified

20. Security Classif. (of this page)

Unclassified

21. No. of Pages

70 pages

22. Price

Insert your own project cover page here

Acknowledgements

The authors would like to acknowledge the Niagara International Transportation Technology Coalition for providing data utilized in this research. They also would like to acknowledge the Journal of the Transportation Research Board, the Transportation Research – Part C journal, and the Accident Analysis and Prevention journal for publishing the material compiled in this report.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The field of traffic accident analysis has long been dominated by traditional statistical analysis. With the recent advances in data collection, storage and archival methods, the size of accident datasets has grown significantly. This in turn has motivated research on applying data mining and Machine Learning algorithms, which are specifically designed to handle datasets with large dimensions, to traffic accident analysis. This project explores three specific applications of Data Mining and Machine Learning algorithms to traffic accident analysis, as briefly described below.

The first application explores the potential for using a *modularity-optimizing community detection* algorithm and *association rules learning* algorithm, to identify important accident characteristics. As a case study, the algorithms are applied to an accident dataset compiled for Interstate 190 in the Buffalo-Niagara metropolitan area. Specifically, the *community detection algorithm* is used first to cluster the data in order to reduce the inherent heterogeneity, and then the *association rule learning* algorithm is applied to each cluster to discern meaningful patterns within each, particularly related to high accident frequency locations (hotspots) and incident clearance time. To demonstrate the benefits of clustering, the *association rule* algorithm is also applied to the whole dataset (before clustering) and the results are compared to those discovered from the clusters. The study results indicate that: (1) the *community detection algorithm* was quite effective in identifying clusters with discernible characteristics; (2) clustering helped in unveiling relationships and accident causative factors that remained hidden when the analysis was performed on the whole dataset; and (3) the association rule learning algorithm yielded useful insight into accident hotspots and incident clearance time along I-190.

The second application focuses on the development of models for the real-time prediction of traffic accident risk. The data required for the development of such models are usually complex, noisy, and even misleading. This raises the question of how to select the most important explanatory variables to achieve an acceptable level of accuracy for real-time traffic accident risk prediction. To address this, the project proposes a novel Frequent Pattern tree (FP tree) based variable selection method. The method works by first identifying all the frequent patterns in the traffic accident dataset. Next, for each frequent pattern, we introduce a new metric, herein referred to as the Relative Object Purity Ratio (ROPR). The ROPR is then used to calculate the importance score of each explanatory variable which in turn can be used for ranking and selecting the variables that contribute most to explaining the accident patterns. To demonstrate the advantages of the proposed variable selection method, the study develops two traffic accident risk prediction models, based on accident data collected on interstate highway I-64 in Virginia, namely a k-nearest neighbor model and a Bayesian network. Prior to model development, two variable selection methods are utilized: (1) the FP tree based method proposed herein; and (2) the random forest method, a widely used variable selection method, which is used as the base case for comparison. The results show that the FP tree based accident risk prediction models perform better than the random forest based models, regardless of the type of prediction models (i.e. k-nearest neighbor or Bayesian network), the settings of their parameters, and the types of datasets used for model training and testing. The best model found is a FP tree based Bayesian network model that can predict 61.11% of accidents while having a false alarm rate of 38.16%. These results compare very favorably with other accident prediction models reported in the literature.

The third application develops models for predicting incident duration, based on the M5P algorithm. 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*. This project 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 that model, 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 is 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 are 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).

Key Words: Data mining; Complex Network Analysis; Frequent Pattern tree (FP tree); Fuzzy C-means clustering (FCM); Bayesian network; Random forest; M5P Tree; Hazard-based Duration Model.

INTRODUCTION

Given the enormous societal cost of traffic accidents, the transportation community has consistently been interested in accident analysis methods to reveal patterns, identify causative factors, and suggest countermeasures. The field of traffic accident analysis, however, has for long been dominated by traditional statistical analysis methods which over the years have yielded invaluable insight and helped guide policy. With the recent advances in data collection, storage and archival methods, the size of accident datasets has grown significantly. This in turn has motivated research into data mining and Machine Learning algorithms, which are specifically designed to handle datasets with large dimensions, for traffic accident analysis.

This project explores three specific applications of Data Mining and Machine Learning algorithms to traffic accident analysis. The first application explores the potential for using a *modularity-optimizing community detection* algorithm and *association rules learning algorithm*, to identify important accident characteristics. The second application proposes a novel *Frequent Pattern tree (FP tree)* based variable selection method, and then develops models for the real-time prediction of traffic accident risk. Finally, the third application proposes a novel approach to developing accident duration prediction models. The approach improves on the original M5P tree algorithm through the construction of a *M5P-Hazard-Based Duration Model (HBDM).*

Besides the Introduction and the Conclusions section, this report is divided into three major sections, each dedicated to discussing one of the three applications studied in this project, namely: (1) the application of Data Mining and Complex Network Algorithms for Traffic Accident Analysis; (2) the use of a novel variable selection method based on Frequent Pattern Tree for real-time traffic accident risk prediction; and (3) the development of a combined M5P Tree and Hazard-based Duration Model for predicting urban freeway traffic accident durations. It should be noted that the current report represents a compilation of the material previously published by the authors in the following papers, Lin et al. (2014); Lin et al. (2015) and Lin et al. (2016).

DATA MINING AND COMPLEX NETWORK ALGORITHMS FOR TRAFFIC ACCIDENT ANALYSIS

This section, which is based on Lin et al. (2014), is organized as follows. First, background information on clustering, complex networks analysis techniques, and on the methods used to extract the relationship between crash involvement and risk factors, is provided. The study's methodology is then described including a description of: (1) how the *modularity optimization algorithm for community detection* was applied to cluster the data; (2) the *association rule* data mining method; and (3) the characteristics of the dataset used. The clustering results are then presented, followed by a description of the discovered association rules for: (1) identifying hotspots and their characteristics; and (2) understanding the factors affecting incident clearance time. A discussion of the difference between the association rules derived from the whole data set and those derived from each cluster is also included.

BACKGROUND

Clustering and Data Heterogeneity

Several researchers have recently pointed out that heterogeneity inherent in traffic accident data often prevents the further exploration of these data (Savolainen et al., 2011; Depaire et al., 2008). To deal with the issue, random effects and random parameters models have been proposed for traffic accident data analysis (Karlaftis et al., 1998; Miaou et al., 2003). 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 & Mannering, 2010). Anastasopoulos and Mannering (2009), for example, demonstrated that random parameters model can account for the heterogeneity arising from a number of factors in accident records and other unobserved factors in their accident frequency study. However, random effects model and random parameters model may not be easily transferable, and are often difficult to estimate (Lord & Mannering, 2010). Clustering the traffic accident data is another way to minimize the heterogeneity problem. For example, Valent et al. (2002) found that “Sundays” and “holidays” arise as significant risk factors when the analysis was performed for clustered data. Moreover, Mohamed et al. (2013) identified “bad visibility due to bad weather” as a factor that can increase the risk of fatal crashes in Montreal Canada, based on an analysis performed on a clustered dataset.

In traffic accidents studies, the two most widely used clustering techniques are: (1) the latent class clustering (LCC); and (2) the K-means clustering method. On one hand, LCC has the advantages of being able to provide statistical criteria for deciding the number of clusters, and to calculate the probabilities for the new data points to belonging to a given cluster (Depaire et al., 2008; de Oña et al., 2013). On the other hand, LCC heavily relies on the assumption of local independence among traffic accident variables to reduce parametric complexity and computing time, and was found to sometimes reach the local rather than global maximum. As for K-means clustering, Anderson (2009) applied the method to classify accident hotspots into relatively homogenous types based on their environmental characteristics. In addition, Mohamed et al. (2013) reported that for the Montreal accident dataset the K-means clustering method appeared to do a better job compared to LCC which tended to classify 90% of the accidents into the first two clusters.

Modularity Optimization Community Detection Method

Recently, complex network analysis methods have been intensively used to understand the features of complex systems such as biological, social, technological and information networks. In the analysis, communities, also called clusters or modules, denote groups of system components that probably share common properties and/or play similar roles in graphs (Fortunato, 2010). For example, for a Facebook social network, communities represent people who share common interests, and therefore exploiting the affiliations of users to these communities provides an effective way to provide them with targeted recommendations and advertisements (Ferrara, 2012). For these methods to work, however, the problem needs to be formulated in the form of a network graph.

The modularity optimization method is one of the most popular methods used for community detection in graph and network analysis (Fortunato, 2010). Its premise is that the network is divided the best when the modularity (i.e., the degree to which a system's components may be divided) is maximized. Due to the generality of the method, the concept of modularity optimization can be applied to traffic accident clustering, by representing each accident record as one node in the network (analogous to a person in a social network).

Discerning Relationships between Crash Involvement and Risk/Causative Factors

For traffic accidents analysis, many statistical, non-parametric and data mining methods have been previously used, with or without clustering, to identify hotspots and to extract relationships between crash involvement and risk factors. As for hotspots, various approaches have been used to define and detect hotspots (Anderson, 2009). Some studies defined hotspots (or black spots) as geographical locations with highly concentrated traffic accidents (Geurts, 2003: Xie & Yan, 2008), while some others detected hotspots based on quantitative measures such as the number of accidents divided by the traffic flow rate per period of time (Gregoriades & Mouskos, 2013). Among those studies, Kernel Density Estimation has gained more and more popularity (KDE) (Anderson, 2009; Xie & Yan, 2008; Okabe et al., 2009; Bil et al., 2013) especially in conjunction with Geographic Information Systems (GIS). In terms of statistical methods previously used to model other aspects of traffic accidents, examples include: (1) hazard-based duration models which have been applied to identify accident characteristics that affect clearance time (Alkaabi et al., 2011; Ghosh, 2010); (2) ordered probit models for estimating the likelihood of injury severity (Lee & Abdel-Aty, 2005); and (3) Bayesian networks for detecting the factors explaining rural highway accident severity (used in conjunction with LCC clustering in de Oña et al., 2013). Among data mining methods proposed for accident analysis, on the other hand, is the *association rule* method used for example in Geurts et al., 2003 and Xi et al., 2004.

METHODOLOGY

Suppose the accident dataset contains N records, each of which contains information about a set of variables $A = \{c_1, c_2, \dots c_m, a_1, a_2, \dots a_n\}$. We divide those variables intro two groups: (1) the c_l variables, $1 \leq l \leq m$, which represents the causative factors behind the accident such as time of day, weather conditions, road geometric features (e.g. number of lanes), etc.; and (2) the accident attributes, a_k, $1 \leq k \leq n$, which represents the specific characteristics of a crash such as associated injuries, location, incident clearance time, etc.

Clustering Analysis

This study used the *community detection* algorithm, for the first time, to cluster the data and reduce heterogeneity. The first step was to represent the data in the form of the network by treating each accident record as one vertex in the network (similar to a friend in a Facebook network). Then, the problem becomes to find out how these vertices are connected in the network. Because in this study the objective is to find out how causative factors contribute to the outcome (i.e. the accident characteristics), the grouping is based on the causative factors (i.e. the c_l variables).

According to the algorithm, two vertices (i.e. two accidents) i and j, $1 \leq i, j \leq N, i \neq j$ will be connected if the following condition is satisfied:

$$\sum_{1 \leq l \leq m} \boldsymbol{e_l} \geq \boldsymbol{e_{th}}, \qquad \text{Equation 1}$$

Where $e_l = 1$, if the values of the factor c_l of i and j are the same, otherwise $e_l = 0$, and e_{th} is the similarity threshold defined by the user (i.e. this counts how many attributes are similar). If the two vertices i and j are connected, an undirected edge is drawn between them, and the weight of the edge can be calculated as:

$$\boldsymbol{W_{ij} = \frac{\sum_{1 \le l \le m} e_l}{m},} \qquad \text{Equation 2}$$

Following the network formation, the *community detection* algorithm is applied to divide it into clusters so that each vertex belongs to only one cluster. The most popular quality function of a partition is the modularity of Newman and Girvan *(22)*, which can be calculated as following:

$$\boldsymbol{Q = \frac{1}{2T} \sum_{i,j} [W_{ij} - \frac{f_i f_j}{2T}] \delta(o_i, o_j),} \qquad \text{Equation 3}$$

Where W_{ij} represents the weight of the edge between vertex i and j; $f_i = \sum_j W_{ij}$ is the summation of the weights for the edges attached to vertex i; o_i is the index of community or cluster vertex i is assigned to in a given iteration, and $\delta(o_i, o_j) = 1$, if $o_i = o_j$, otherwise $\delta(o_i, o_j) = 0$; and $T = \frac{1}{2} \sum_{i,j} W_{ij}$. As defined above, the *modularity* basically reflects the concentration of vertices within communities compared with random distribution of edges between all vertices regardless of communities. A positive modularity means that the weights of the edges within the communities exceed the weights expected on the basis of chance, and this is the main motivation behind maximizing modularity. However, because it is too difficult to enumerate and test all the ways to partition a graph, algorithms such as the one proposed by Blondel et al., 2008 for the fast unfolding of the communities are needed. Blondel et al.'s algorithm was the one utilized in this study (Blondel et al., 2008; Arenas et al., 2007).

As compared to traditional clustering techniques such as LCC and K-means clustering, the *community identification* algorithm offers several advantages. First, the network transformation and the modularity optimization method are intuitive and easy to implement. Second, when building the network, because the causative factors are compared one by one and because there is no distance measure involved, as is the case with other techniques such as *K-means*, there is no need to normalize the data (which often introduces imprecision). Third, unlike the LCC method, the modularity optimization algorithm does not rely on the assumption of the independence among variables to decrease the complexity of computation; instead, it is extremely fast since the number of possible communities decreases drastically after a few iterations (Blondel et al., 2008). Fourth, the method provides a modularity based quality function, which can be used to measure the effect of clustering. Finally, the method, even for large dimensional datasets, requires the specification/calibration of only one parameter, the threshold e_{th}, as opposed to classical statistical analysis methods where the number of parameters may exponentially increase as the number of variables increases (Chen & Jovanis, 2000).

Association Rule Learning

The concepts of association rules learning were firstly introduced by Agrawal et al., 1993. Given a traffic accident related variable set $A = \{c_1, c_2, \dots c_m, a_1, a_2, \dots a_n\}$, it can be transformed to a set of binary attributes called items $I = \{I_{1c}, I_{2c}, \dots I_{Lc}, I_{1a}, I_{2a}, \dots I_{Ka}\}$, where I_{lc}, $1 \leq l \leq L$ are the binary attributes associated with the causative factors, and I_{ka}, $1 \leq k \leq K$ are the binary variables related to accident attributes (i.e. the outcome). For example, the factor "Season" can be represented by four binary attributes, i.e., "spring", "summer", "autumn", and "winter". Each of the N accident records, referred to here as transactions T, has a unique transaction ID and is a subset of I. An association rule is an implication of the form, $X \Rightarrow Y$, where X and Y are sets of items in I, $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$.The sets of items X and Y are called the body and head of the rules, respectively.

At a very high level, generating the association rules involves two basic steps. The first is to generate the frequent item sets in the data. X is called a frequent item set when its support, which refers to the frequency at which X appeared in the N transactions, is equal to or greater than the minimum support defined by user.

$$\frac{\boldsymbol{supp}\{\boldsymbol{X}\}}{\boldsymbol{N}} \geq \boldsymbol{\sigma}, \qquad \text{Equation 4}$$

Where $supp\{X\}$ is the number of transactions in N that contains item set X, and σ is the minimum support.

Now suppose item sets X and $X \cup Y$ are frequent item sets, the second step is to calculate the confidence of $X \Rightarrow Y$. This is based on the ratio of the number of transactions that contains $X \cup Y$ to transactions that only contains X. If the confidence is equal to or higher than the user-defined minimum confidence, $X \Rightarrow Y$ is an association rule.

$$conf(X \Rightarrow Y) = \frac{supp\{X \bigcup Y\}}{supp\{X\}} \geq \varepsilon\,, \qquad \text{Equation 5}$$

Where ε is the minimum confidence. Methods are then available to distinguish between the trivial and non-trivial rules (Geurts et al., 2003).

DATA PROCESSING

The dataset used in this study included 999 traffic accidents observed at I-190 from 01/01/2008 to 10/31/2012. I-190 runs 28.34 miles (45.61 km) from its intersection with I-90 near Buffalo, NY up north to Lewiston, NY via Niagara Falls. I-190 plays a critical role in the Buffalo-Niagara transportation network, especially in terms of connecting Western New York to Southern Ontario, Canada. Incidents and traffic flow are monitored 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 1 lists both the causative factors and accident attributes variables that were available in NITTEC's incident logs, and were thought to be useful for analysis. After initial screening of the data, a total of 15 variables were selected (nine causative factors and six accident attributes) as shown in Table 1. The variables that were excluded did not exhibit enough variation over the dataset compiled (i.e., more than 95% of the records had the same value for the variable).

Table 1. Traffic Accident Variables in the I-190 Data

Variables	Values	Included
Causative Factors		
Season	Spring (March, April, May); Summer (June, July, August); Autumn (September, October, November); Winter (December, January, February)	Yes
Weekday	Yes (Monday 2 AM-Friday 9 PM, except holidays); no	Yes
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)	Yes
Wind Speed	0 mph (miles per hour); 10 mph; 20 mph; 30 mph	Yes
Weather Conditions	clear; rain; snow	Yes
Direction	North; South	Yes
Lane Number on Main Road	1; 2; 3	Yes
Lane Number on Ramp	0 (away from exit); 1; 2;	Yes
Ramp Type	on ramp; off ramp; highway to highway on ramp; highway to highway off ramp;	Yes
Vehicle Type	Car; Truck/Tractor Trailer; Motorcycle	No
Accident Attributes:		
Location – Exit Number	Exit 1; …; Exit 25; Highway	Yes
Location relative to Road Configuration	Before the exit; at the exit; beyond the exit; highway; ramp; bridge; before the bridge; after the bridge	Yes
Number of Vehicles Involved	1; 2; more than 2	Yes
Clearance Time	0-15minutes; 16-30 minutes; 31-45 minutes; 46-60 minutes; 61-75 minutes; 76-90 minutes; more than 90 minutes	Yes
Blocked Lane Index	left lane at main road; middle lane at main road; right lane at main road; all lanes at main road; left lane at ramp; right lane at ramp; all lanes at ramp;	Yes
Blocked Lane Number	one lane at main road; two lanes at main road; three lanes at main road; one lane at ramp; two lanes at ramp	Yes
Injury	Yes; No	No
Roll Over	Yes; No	No
Congestion	Yes; No	No

RESULTS

Community Detection

The only parameter that needed to be calibrated was the similarity threshold e_{th}, and given that the number of causative variables used for the comparison was 9 ($m = 9$), the range for that parameter was from 1 to 9. Furthermore, because e_{th} determines the similarity criterion between two accident records, at least more than half of the variables should have the same values. This further narrowed the range to between 5 and 8 (it also does not make sense to require all 9 parameters to be similar). Given this, we experimented with four possible values for e_{th}: 5, 6, 7, and 8. This process was conducted with the help of the open visualization software Gephi (Bastian et al., 2014) and the resulting network characteristics are shown in Table 2.

Table 2. Network Clusters with Respect to the Similarity Threshold

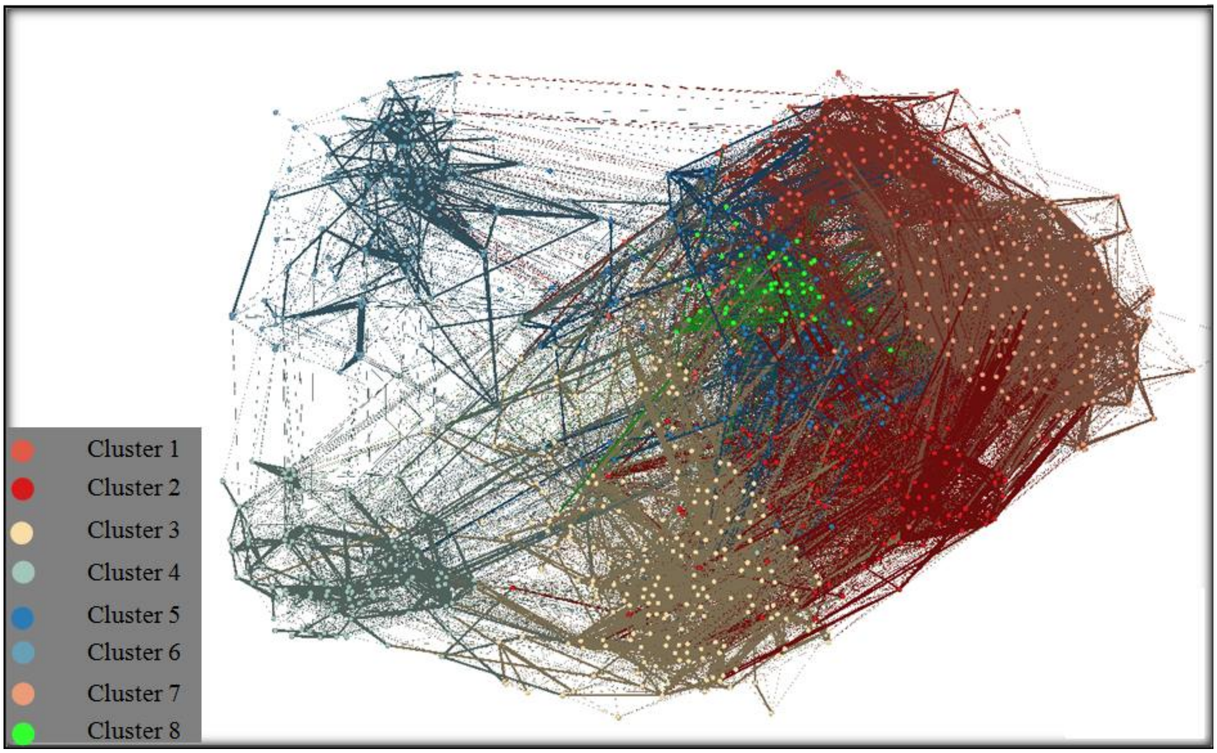
Resulting Network Characteristics	$e_{th} = 5$	$e_{th} = 6$	$e_{th} = 7$	$e_{th} = 8$
Number of vertices	999	999	997	930
Number of edges	180,480	83,945	27,552	5,705
Number of clusters founded	3	5	8	33
Maximum modularity	0.213	0.296	0.47	0.647

Causative Factors and Their Probabilities in Each Cluster (%)

Variable: Value (Environmental Feature)	**Cluster**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Season: Winter	14	50	21	34	16	45	29	94
Weekday: Yes	**99**	**95**	0	66	54	73	73	70
Weekday: No	1	5	**100**	34	46	27	27	30
Weather Conditions: Clear	80	70	84	65	85	44	73	0
Weather Conditions: Snow	1	16	5	24	0	31	14	**100**
Direction: South	**99**	0	60	55	**100**	**88**	0	**98**
Direction: North	0	**100**	40	45	0	12	**100**	0
Lane Number at Main Road: 3	**99**	**98**	**99**	61	0	2	0	74
Lane Number on Main Road: 2	0	0	0	37	**100**	**98**	**99**	26
Lane Number on Ramp: 1	99	81	90	0	**100**	15	72	**100**
Lane Number on Ramp: 2	1	19	10	0	0	**85**	28	0
Lane Number on Ramp: 0	0	0	0	**100**	0	0	0	0

As can be seen from Table 2, with the increase in the value of the similarity threshold e_{th}, the number of edges in the network decreases (since it becomes harder to find similar vertices to connect), and the number of clusters as well as the associated maximum modularity of the network increase. Since modularity represents the concentration of nodes within communities in comparison to the random distribution of edges among nodes regardless of communities, lower e_{th} makes the network more randomly connected. Therefore, it is better to choose larger e_{th}. However, when $e_{th} = 8$, although the maximum modularity is 0.647, the number of clusters is as high as 33. Besides, because connection requirements are more demanding, only 930 out of the 999 vertices are connected in that network (the remaining accidents were not found to be similar to any other accident which defies the purpose behind clustering). Given this, 7 was selected as the value for e_{th}, resulting in a total of 8 clusters. Figure 1 shows the resulting traffic accident network and the clustering results.

Figure 1. Resulting traffic accidents network and community detection ($e_{th} = 7$).



To identify the attributes of each cluster (in terms of describing a given accident type or condition), we followed the method used by Depaire et al. (2008), where the distributions of the variables in each cluster are analyzed to identify the dominant or skewed features (the cluster could then be named based on these features. For example, if 100% of traffic accidents in one cluster happen at non-weekdays, while the other clusters have low probabilities for that feature, we can refer to that cluster as the non-weekday accidents cluster). Table 2 shows the probabilities for each feature within the 8 clusters, where the dominant or skewed feature probabilities are underlined and highlighted.

The probabilities in Table 2 can clearly be used to characterize each cluster. For example, the first three clusters are all most likely to occur on the highway sections with three lanes at main road (with the occurring probabilities of 99%, 98% and 99%, respectively). Moreover, Cluster 1 and 2 can be claimed as weekday accidents in the southbound and northbound directions of I-190, respectively, while Cluster 3 includes non-weekday accidents only. All the Cluster 4 accidents (100%) occurred on highway sections away from exits, where the lane number on the ramp is 0. Clusters 5, 6 and 7 all involve accidents on roads with only two lanes. However, Cluster 5 seems to have involved accidents close to a ramp with one lane, whereas for Cluster 6, the ramp had two lanes. Moreover, Clusters 5 and 6 seem to involve accidents in the southbound direction, whereas accident s in Cluster 7 occurred in the northbound direction. Finally, Cluster 8 appears to involve accidents happening during snowy conditions (100%). Based on the results, the eight clusters can be described as shown in Table 3.

Table 3. Traffic Accident Types

Cluster	Traffic accident types	Size (%)
1	Traffic accidents on southbound sections with three lanes at main road on weekdays	17
2	Traffic accidents on northbound sections with three lanes at main road on weekdays	10
3	Traffic accidents on sections with three lanes at main road on non-weekdays	11
4	Traffic accidents on sections away from exits	13
5	Traffic accidents on southbound sections with two lanes at main road and 1 lane at ramp	9
6	Traffic accidents on southbound sections with two lanes at main road and two lanes at ramp	13
7	Traffic accidents on northbound sections with two lanes at main road	22
8	Traffic accidents on southbound sections with one lane at ramp in snowy days	5

Association Rule Analysis to Identify Hotspots

In this study, for the association rule analysis, a “hotspot” is defined as the place where the ratio of the number of accidents at that particular spot, to the number of accidents on the whole transportation system under consideration is greater than the minimum support σ, under the conditions defined by the body of an association rule. In order to identify accident hotspots and the characteristics of accidents that occur there, the association rule analysis algorithm was then run using the 9 causative factors as candidate variable for the body of each rule, and using the “Location-Exit Number” accident attribute as the head of each rule. The minimum support parameter was set to 0.05, and the minimum confidence to 0.50. The results are shown in Table 4 which lists the rules that had the highest confidence for a given location, along with a few other rules that provide some insight for the study. As can be seen, the analysis was performed twice: first, on the whole dataset without clustering, and then on each cluster. The dominant or skewed features for each cluster, as determined from the previous analysis, are shown in bold. Finally, the confidence level values shown in parentheses are those that result when the value of one causative factor is perturbed (e.g. for rule #5 in cluster 2, the confidence drops from 1.00 to 0.38, when the environmental condition changes from rain to clear).

Table 4. Rules on Hotspots from the Whole Dataset and the Clusters

Datasets	ID	Body	Head	Confidence
Whole Dataset	1	[direction: north]+[lane number at main road: 2]+[ramp type: off ramp]	[Exit 9: Peace Bridge]	0.67
	2	[lane number at main road: 2]+[lane number at ramp:1]+[ramp type: highway to highway off ramp]	[Exit 11: route 198]	1
	3	[lane number at main road: 2]+[lane number at ramp: 2]+[ramp type: highway to highway off ramp]	[Exit 16: I-290]	0.60
Cluster1	4	**[Weekdays: yes]**+[weather condition: clear]+**[direction: south]**+**[lane number at main road: 3]**+[lane number at ramp: 1]+[ramp type: highway to highway off ramp]	[Exit 7 Skyway]	1
Cluster2	5	**[weekdays: yes]**+[hour: 4 PM-6 PM]+[weather condition: rain (clear)]+**[direction: north]**+**[lane number at main road: 3]**+[lane number at ramp: 1]	[Exit 8: Niagara Street]	1 (0.38)
	6	([season: Winter]+)**[weekdays: yes]**+ **[direction: north]**+**[lane number at main road: 3]**+[lane number at ramp: 2]+[ramp type: off ramp]	[Exit 6: Elm/Oak Street]	0.90 (1)

Cluster3	7	**[weekdays: no]**+[direction: north]+**[lane number at main road: 3]**+[lane number at ramp: 2]+[ramp type: off ramp]	[Exit 6: Elm/Oak Street]	0.89
Cluster4	8	[season: winter]+[weekdays: yes]+[lane number at main road: 2]+**[lane number at ramp: 0]**	Milepost 10-12	0.54
Cluster5	9	**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 1]**+[ramp type: highway to highway off ramp]	[Exit 11: Route 198]	1
	10	([season: winter]+)[hour: 7 AM-9 AM]+**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 1]**+[ramp type: off ramp]	[Exit 17: South Grand Island Bridge]	0.54(0.90)
Cluster6	11	[weekdays: yes]+[hour: 4 PM-6 PM]+**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 2]**+[ramp type: highway to highway off ramp]	[Exit 16: I-290]	0.63
	12	[weekdays: yes]+[hour: 7 AM-9 AM]+[direction: north]+**[lane number at main road: 2]**+**[lane number at ramp: 2]**+[ramp type: highway to highway off ramp]	[Exit 16: I-290]	1
Cluster7	13	[weekdays: yes]+[hour: 4 PM-6 PM]+**[direction: north]**+**[lane number at main road: 2]**+[lane number at ramp: 2]+[ramp type: off ramp]	[Exit 9: Peace Bridge]	1
	14	[hour: 4 PM-6 PM]+[weather condition: clear]+**[direction: north]**+**[lane number at main road: 2]**+[lane number at ramp: 2]+[ramp type: off ramp]	[Exit 9: Peace Bridge]+[road structure: beyond the exit]	0.52
	15	**[direction: north]**+**[lane number at main road: 2]**+[lane number at ramp: 1]+[ramp type: highway to highway off ramp]	[Exit 11: Route 198]	1
Cluster8	16	[weekdays: yes]+**[weather condition: snow]**+**[direction: south]**+[lane number at main road: 3]+**[lane number at ramp: 1]**+[ramp type: highway to highway off ramp]	[Exit 7: Skyway]	1
	17	**[weather condition: snow]**+**[direction: south]**+[lane number at main road: 3]+**[lane number at ramp: 1]**+[ramp type: highway to highway off ramp]	[Exit 7: Skyway]+[road structure: before the exit]	0.6
	18	[weekdays: yes]+[hour: 10 PM-6 AM]+([wind speed: 10])+**[weather condition: snow]**+**[direction: south]**+[lane number at main road: 2]+**[lane number at ramp: 1]**+[ramp type: off ramp]	[Exit 9: Peace Bridge]	0.5 (0.75)

From the analysis on the whole dataset, three association rules with the highest confidence, for the corresponding three hotspots (Exits 9, 11 and 16), are selected. One common feature in body parts of the three rules is there are two lanes at main road, and two out of the three rules contain highway to highway off ramp feature, which appear to be problematic areas with a high accident frequency (this is quite intuitive because of the limitation of capacity and the excessive weaving that takes place there). As can be seen, the analysis on the non-clustered dataset yielded limited insight about the hotspots.

When the analysis was performed on the clusters, several more rules and causative factors are revealed. Specifically, 15 association rules are revealed, along with eight hotspots. For the hotspots, only one is located away from exits, and the rest are all close to exits. Furthermore, these seven exits identified are spatially correlated with one another, and fall very neatly in two definite *geographic* clusters; the first is [Exit 6, Exit 7, Exit 8, Exit 9, and Exit 11] – note that there is no Exit 10 on I -190; and the second is [Exit 16 and Exit 17]. Through comparing and analyzing the rules describing the same hotspot, a few additional insights can be gained as below:

Firstly, for Exit 6, when comparing Rules 6 and Rule 7, it becomes clear that the problem is consistently in the north direction no matter if it is a weekday or a non-weekday. Secondly, for Exit 7, when comparing Rules 4 and Rule 16, we can see that Exit 7 is always a hotspot with (confidence level = 1) regardless of the weather condition (both clear and snow). Rule 17 shows that the segment before Exit 7 is a hotspot in south direction when it snows. Thirdly, for Exit 9, Rule 13 provides more specific conditions than Rule 1. According to the rule, Exit 9 is a hotspot with confidence level 1 in the north direction for the peak hour 4 PM-6 PM on weekdays. Rule 14 shows that if it is the peak hour 4 PM-6 PM with clear weather, the segment beyond Exit 9 in north direction is also a hotspot. And Rule 18 shows that, in the south direction, Exit 9 may also be a hotspot when it is 10 PM-6 AM on weekdays with snow. Fourthly, for Exit 11, by checking Rule 9 and Rule 15, Exit 11 is always a hotspot with confidence 1 in both the north and southbound direction. This is consistent with the conclusion of Rule 2 on the whole dataset. Finally, for cluster 4 describing traffic accidents on highways away from exits, only one hotspot is found with a relatively low confidence 0.54, although it contains 13% of the total records. This seems to indicate that accidents along I-190 tend to happen close to exits more often.

Besides insight regarding hotspots, the associative rules shed additional light on the conditions under which accidents happen at those locations. This additional insight is gained by considering the role of the variables in the "body" parts of the rules. A few examples are described below.

Firstly, the variables "weekdays" and "hour of the day" appear to affect whether a location becomes a hotspot. Nine out of the 15 association rules generated from the clusters contain "[weekdays: yes]" in the body parts, and five of the nine rules contain "[hour: 7 AM-9 AM]" or [hour: 4 PM-6 PM]." This reveals the effect of weekday peak hours on traffic accidents. Another convincing example comes from Rule 11 and Rule 12. Exit 16-I-290 is a hotspot when it is 7 AM-9 AM in the morning towards north direction, and Exit 16 is also a hotspot when it is 4 PM-6 PM in the afternoon towards south direction.

Secondly, the feature “[season: winter]” can increase the confidence in claiming a location as a hotspot. For example, Rule 6 in Cluster 2 shows that if it is in winter, the confidence for Exit 6 to be a hotspot on weekdays will increase from 0.90 to 1. Similarly, Rule 10 in Cluster 5 shows that if it is 7 AM-9 AM on someday in winter, the confidence in claiming Exit 17 as a hotspot witness a large increase from 0.54 to 0.90. Besides that, the variable “wind speed” and “weather condition” are found to affect the confidence for some locations. Rule 18 shows that Exit 9-Peace Bridge has a higher risk 0.75 than the previous 0.50 if the wind speed is 10 miles per hour. Rule 5 shows that with the other features in the body part being the same, the “[weather condition: rain],” rather than “[weather condition: clear],” tend to make Exit 8 a hotspot with confidence level 1.

Association Rule Analysis to Identify Factors Affecting Incident Clearance Time

The *association rule* analysis was then repeated, this time using the accident attribute “incident clearance time” as the “head of the rules” to gain some insight into the factors affecting incident clearance time. For clearance time analysis, the minimum support is set as 0.05, and the minimum confidence is lowered to 0.30 (experiments showed this set of rules to have lower confidence levels compared to the hotspot analysis). The results are shown in Table 5.

Table 5. Rules on Clearance Time from the Whole Dataset and the Clusters

Datasets	**ID**	**Body**	**Head**	**Confidence**
Whole Dataset	1	[weekdays: yes]+[hour: 4 PM-6 PM]	[Clearance time: 31-45 minutes]	0.32
	2	[season: winter]+[lane number at main road: 3]	[Clearance time: 16-30 minutes]	0.34
Cluster1	3	**[weekdays: yes]**+[hour: 4 PM-6 PM]+[wind speed: 10]+**[direction: south]**+**[lane number at main road: 3]**	[Clearance time: 31-45 minutes]	0.35
Cluster2	4	**[weekdays: yes]**+[hour: 4 PM-6 PM]+[weather condition: clear]+**[direction: north]**+**[lane number at main road: 3]**+[lane number at ramp: 1] +[ramp type: off ramp]+[road structure: at the exit]	[Clearance time: 31-45 minutes]	0.58
	5	**[weekdays: yes]**+[weather condition: clear]+[Exit 8: Niagara Street]+**[direction: north]**+**[lane number at main road: 3]**	[Clearance time: 31-45 minutes]	0.55
	6	[season: winter]+**[weekdays: yes]**+[weather condition: clear]+**[direction: north]**+**[lane number at main road: 3]**+[lane number at ramp: 1]	[Clearance time: 16-30 minutes]	0.30
Cluster3	7	[season: autumn]+**[weekdays: no]**+[direction: north]+**[lane number at main road: 3]**+[lane number at ramp: 1]+[ramp type: off ramp]	[Clearance time: 46-60 minutes]	0.60
	8	**[weekdays: no]**+[Exit 8: Niagara Street]+ **[lane number at main road: 3]**+[lane number at ramp: 1]+[ramp type: off ramp]	[Clearance time: 46-60 minutes]	0.33
Cluster4	9	[season: autumn]+[weekdays: yes]+[lane number at main road: 3]+**[lane number at ramp: 0]**	[Clearance time: 46-60 minutes]	0.50

	10	[season: winter]+[direction: south]+[lane number at main road: 3]+**[lane number at ramp: 0]**	[Clearance time: 16-30 minutes]	0.47
	11	[weekdays: no]+[direction: south]+[lane number at main road: 3]+**[lane number at ramp: 0]**	[Clearance time: 31-45minutes]	0.37
	12	[weekdays: yes]+[direction: south]+[lane number at main road: 3]+**[lane number at ramp: 0]**	[Clearance time: 16-30 minutes]	0.41
	13	[weekdays: yes]+[direction: north]+[lane number at main road: 3]+**[lane number at ramp: 0]**	[Clearance time: 31-45minutes]	0.31
Cluster5	14	[weekdays: no]+**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 1]**	[Clearance time: 16-30 minutes]	0.31
	15	[weekdays: yes]+**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 1]**	[Clearance time: 31-45minutes]	0.32
	16	[weekdays: yes]+ [Exit 9: Peace Bridge]+**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 1]**+[ramp type: off ramp]	[Clearance time: 31-45minutes]	0.60
Cluster6	17	[Exit 16: I-290]+**[direction: south]**+**[lane number at main road: 2]**+**[lane number at ramp: 2]**+[ramp type: highway to highway off ramp]+[road structure: at the exit]	[Clearance time: 31-45minutes]	0.35
	18	[hour: 7 AM-9 AM]+**[lane number at main road: 2]**+**[lane number at ramp: 2]**+[ramp type: highway to highway off ramp]	[Clearance time: 46-60 minutes]	0.33
Cluster7	19	[weekdays: yes]+[hour: 1 PM-3 PM]+**[direction: north]**+**[lane number at main road: 2]**	[Clearance time: 0-15minutes]	0.52
	20	[weekdays: yes]+[Exit 9: Peace Bridge]+**[direction: north]**+**[lane number at main road: 2]**	[Clearance time: 16-30 minutes]	0.31
	21	[weekdays: yes]+[hour: 4 PM-6 PM]+**[direction: north]**+**[lane number at main road: 2]**	[Clearance time: 31-45minutes]	0.31
	22	[Exit 11: Route 198]+**[direction: north]**+**[lane number at main road: 2]**	[Clearance time: 31-45minutes]	0.34
	23	[season: winter]+([weather condition: snow])+**[direction: north]**+**[lane number at main road: 2]**	[Clearance time: 31-45 minutes]	0.34 (0.46)
Cluster8	24	[season: winter]+**[weather condition: snow]**+**[direction: south]**+[lane number at main road: 3]+**[lane number at ramp: 1]**	[Clearance time: 16-30 minutes]	0.52

As shown in Table 1, clearance time is divided into seven intervals, each 15 minutes long. When the analysis was performed for the whole dataset, two rules are shown: Rule 1 is associated with peak-hour 4 PM-6 PM on weekdays, and the clearance time of accidents is shown to be 31-45 minutes (with a confidence level of 0.32); Rule 2 is for winter, if accidents happen at sections with three lanes main road, the clearance time tend to be between 16-30 minutes (with confidence level of 0.32). As before, when the associate rule analysis is performed on the whole dataset, limited insight is gained.

For the clusters, 22 rules are selected; four have a clearance time of 46-60 minutes, 12 have 31-45 minute clearance times, 5 have 16-30 minutes, while the remainder has 0-16 minutes clearance times. Some of the main observations are summarized below.

Firstly, with respect to the "Weekday" variable, its impact on the incident clearance time appears to be mixed. For example, Rule 8 shows that on non-weekdays, accidents at Exit 8 have clearance time between 46 and 60 minutes with confidence 0.33. Also, according to Rule 11 and 12, on the southbound sections with 3 lanes on the main road, accidents on non-weekdays tend to have a longer clearance time than accidents on weekdays. On the other hand, when comparing Rule 14 and 15, we can see that with other factors being the same, accidents on non-weekdays are more likely to have clearance time of 16-30 minutes, while those on weekdays tend to have longer clearance time of 31-45 minutes. This indicates that there are other factors besides whether the accident is on a weekday or not that affects clearance time, but perhaps the dataset was not rich enough to reveal such factors.

Secondly, the variable "Hour of the Day" may have an impact on the clearance time of traffic accidents. Rules 3, 4 and 21, which correspond to a clearance time 31-45 minutes, all have the same feature "the peak hours 4 PM-6 PM" in their body parts; Rule 18 shows that at peak hours 7 AM-9 AM, accidents on sections with two lanes at main road and two lanes at highway to highway off ramp have a probability of 0.33 to experience 46-60 minutes. And Rule 19 which shows on weekdays at 1 PM-3 PM (i.e. off-peak) the clearance time of accidents on sections towards north with two lanes at main road tends to be short, 0-15 minutes with confidence equal to 0.52.

Thirdly, the feature "snow" appears to increase the likelihood of longer clearance time. According to Rule 23, in the winter for sections towards north with two lanes at main road, the confidence in the clearance time being 36-45 minutes (i.e. on the long side) is 0.34. During snowy condition, the confidence increases to 0.46.

Finally, the "direction" of the road may also affect the clearance time (because it could potentially impact the time needed to get to the incident scene). By comparing Rule 12 and Rule 13, we can see that for sections with 3 lanes on the main road on weekdays, accidents in the north direction has clearance time of 31-45 minutes with confidence 0.31, while accidents in the south direction has a probability of 0.41 to have clearance time of 16-30 minutes. Another similar example is for hotspot at Exit 9. Based on Rule 16 and Rule 20, on weekdays, the clearance time for accidents at Exit 9 in the southbound direction may be 31-45 minutes with a confidence level of 0.60, which is longer than 16-30 minutes at the same exit in the north direction (confidence of 0.31).

CONCLUSIONS

In this study, the *modularity-optimizing community detection* algorithm was used first to cluster accident data recorded for I-190 in the Buffalo-Niagara area. Following this, the *association rules learning* algorithm was used to gain some insight into accident hotspots and incident clearance times. To demonstrate the benefits of clustering, the *association rule* algorithm was applied to both the whole dataset (before clustering) and then to the clusters and the results were compared. The main findings are summarized as below:

1) The community detection algorithm appears to do an excellent job in clustering the data into well-defined clusters;
2) Clustering the data first before running the association rule learning algorithm appears to be a necessary step that can significantly improve the quality of the insight to be gained from the rules extracted. Specifically, when the association rule algorithm was run on the whole dataset in this study, the insight gained was very limited compared to that gained from running the analysis on the clusters.

3) The association rule learning algorithm has the potential to reveal interesting insight about the characteristics of accidents, where they tend to occur, and the factors that affect incident clearance time.

For future research, the authors plan to test the community detection and association rule learning algorithms on larger and richer data sets, and to explore additional relationships between causative factors and accident attributes. They also plan to apply some of the previously used statistical traffic accident techniques, in particular hazard-based duration models, to the analysis of the accident clearance time and to compare the results to those from the data mining techniques utilized herein.

A NOVEL VARIABLE SELECTION METHOD BASED ON FREQUENT PATTERN TREE FOR REAL-TIME TRAFFIC ACCIDENT RISK PREDICTION

Traffic accidents cause a great deal of loss of lives and property. According to the accidents report of the United States Census Bureau, there were 10.8 million accidents and 35,900 persons killed in 2009 (US census bureau, 2013). To address this, many studies have been conducted to predict accident frequencies and analyze the characteristics of traffic accidents, including studies on hazardous location/hot spot identification (Lin et al., 2014), accident injury-severities analysis (Milton et al., 2008), and accident duration analysis (Zhan et al., 2011).

With the development of intelligent transportation systems technologies, there currently exists a wealth of real-time traffic data collected from fixed-locations sensors, automatic vehicle identification systems and other sensing technologies. These data sources can be fused and analyzed to develop real-time management strategies and applications for the purpose of improving efficiency, safety, resiliency and reliability of transportation systems. Particularly in the area of transportation safety, researchers have started to develop real-time traffic accident risk prediction models that take advantage of complex and rapidly and continuously flowing data for predicting traffic accidents.

New issues are emerging accompanying the new opportunities offered by real-time traffic data. One issue is that related to explanatory variable selection, a topic that has received increased attention in real-time traffic accident risk prediction. The wealth of real-time traffic data offer more explanatory variables that may contribute to explaining traffic accident risk and patterns. However, as has been widely recognized, “more is not always better”, particularly for accident prediction. Inclusion of a large number of explanatory variables may cause model overfitting (Sawalha and Sayed, 2006). In addition, it can cause application related issues such as long prediction running time and unreliable prediction results, particularly when a model is applied to new locations and larger data instances (Fernández et al., 2014).

In terms of usage, as a preprocessing step before building any prediction models, variable selection can help researchers identify and extract meaningful information (patterns, structure, underlying relationships, etc.) from the data. Only a small representative subset of the original feature space of the data may be needed to interpret the results (Fernández et al., 2014).
Real-time traffic accident risk prediction models can be broadly classified into two categories, namely statistical models and data mining/machine learning models. Statistical models, such as matched case-control logistic regression models (Abdel-Aty et al., 2004), binary logit models (Xu et al., 2013) and aggregate log-linear models (Lee et al., 2003), have been tested and used in the previous studies. Typical examples of the data mining/ machine learning modeling approach include k nearest neighbor models (Lv et al., 2009), neural networks (Abdel-Aty et al., 2008), Bayesian network models (Hossain and Muromachi, 2012) and support vector machines (Yu and Abdel-Aty, 2013). Those methods have been gaining more and more popularity in recent years.

As previously mentioned, the variable selection problem has attracted attention in previous real-time traffic accident risk prediction research. For statistical models, Sawalha and Sayed (2006) found that using less but statistically significant explanatory variables can avoid over-fitting and improve the reliability of a model. They suggested combining the t-statistics test and the likelihood ratio based scaled deviance test, for selecting significant explanatory variables. Different procedures were suggested for Poisson regression and negative binomial regression respectively due to the additional complexity introduced to the scaled deviance test for negative binomial regression models. As for the data mining models, classification and regression tree (CART) has been used to perform variable selection (Yu and Abdel-Aty, 2013; Pande and Abdel-Aty, 2006). Another ensemble learning method for classification and regression, called random forest, has also been widely used to rank explanatory variables (Abdel-Aty et al., 2008; Ahmed and Abdel-Aty, 2012). Recently, a hybrid model random multinomial logit (RMNL), formed by combining the random forest and logit models, was applied to calculate traffic accidents variable importance (Hossain and Muromachi, 2012).

Different from previous research, this study proposes a novel frequent pattern tree (FP tree) based variable selection method for real-time traffic accident risk prediction, using the data collected on interstate highway I-64 in Virginia as the case study. A new algorithm was built to rank explanatory variables based on the "calculated variable importance score". To verify the model performance, the study then develops two traffic accident risk prediction models, namely a k-nearest neighbor model and a Bayesian network model. Prior to the model development, two variable selection methods are utilized: (1) the FP tree based method proposed by the present research; and (2) the baseline random forest tree based method. The results show that the models trained with the FP tree selected explanatory variables always outperformed the others. To the best of the authors' knowledge, this study is the first attempt toward applying the FP tree based models to traffic accident related research.

This section is organized as below. First, an introduction to the Frequent Pattern (FP) tree model and its variable importance score calculation algorithm is provided. Second, we describe the traffic accident datasets used for model training and testing. Third, we describe and compare the FP tree and the random forest based variable selection methods, in terms of their variable importance ranking results. Fourth, based on the variables selected by the FP tree and the random forest methods respectively, two traffic accident risk predictions models are discussed and compared in terms of their prediction performance, namely the k-NN model and the Bayesian network model. The section ends with a summary of the main conclusions of the work and suggestions for future research.

MODEL METHODOLOGY

This section discusses the FP-tree algorithm used in this study for explanatory variable selection. The algorithm consists of two steps: variable discretization and variable importance score calculation. For the former step, the fuzzy c-means clustering method is used to convert a continuous variable to a series of discrete categorical variables; for the latter, we propose the "Relative Object Purity Ratio (ROPR)" as an importance score for each explanatory variable. This section will also introduce the random forest method that is used as the bench-marking variable selection method. Finally, the two methods used for accident risk prediction, namely the k-NN model and Bayesian network, are briefly introduced.

Frequent-pattern tree (FP-tree)

The Frequent pattern tree (FP-tree) algorithm was proposed by Han et al. (2004). It yields a compact representation of all relevant frequency information in a dataset. A brief introduction of the FP-tree algorithm follows. Suppose $I = \{i_1, i_2, i_3, \ldots, i_m\}$ be a set of items. Let TN be a set of transactions or records in a database DB, and each transaction $Tran$ is a set of items, $Tran \subseteq I$. A pattern X also contains a set of items, $X \subseteq I$. X is called a frequent pattern when its support, referring to the frequency at which X appears in the TN transactions, is equal to or greater than the minimum support threshold, σ .

$$\frac{supp\{X\}}{TN} \geq \sigma \qquad \text{Equation 6}$$

where, σ is a threshold value defined by user.

A FP-tree includes a root labeled as "null". It also includes a set of item-prefix sub-trees as the children of the root. There are two important fields for each node in the item-prefix sub-trees: *item name* and *count*. *Item name* tells which item this node represents, and *count* records the number of transactions represented by the portion of the path reaching this node.

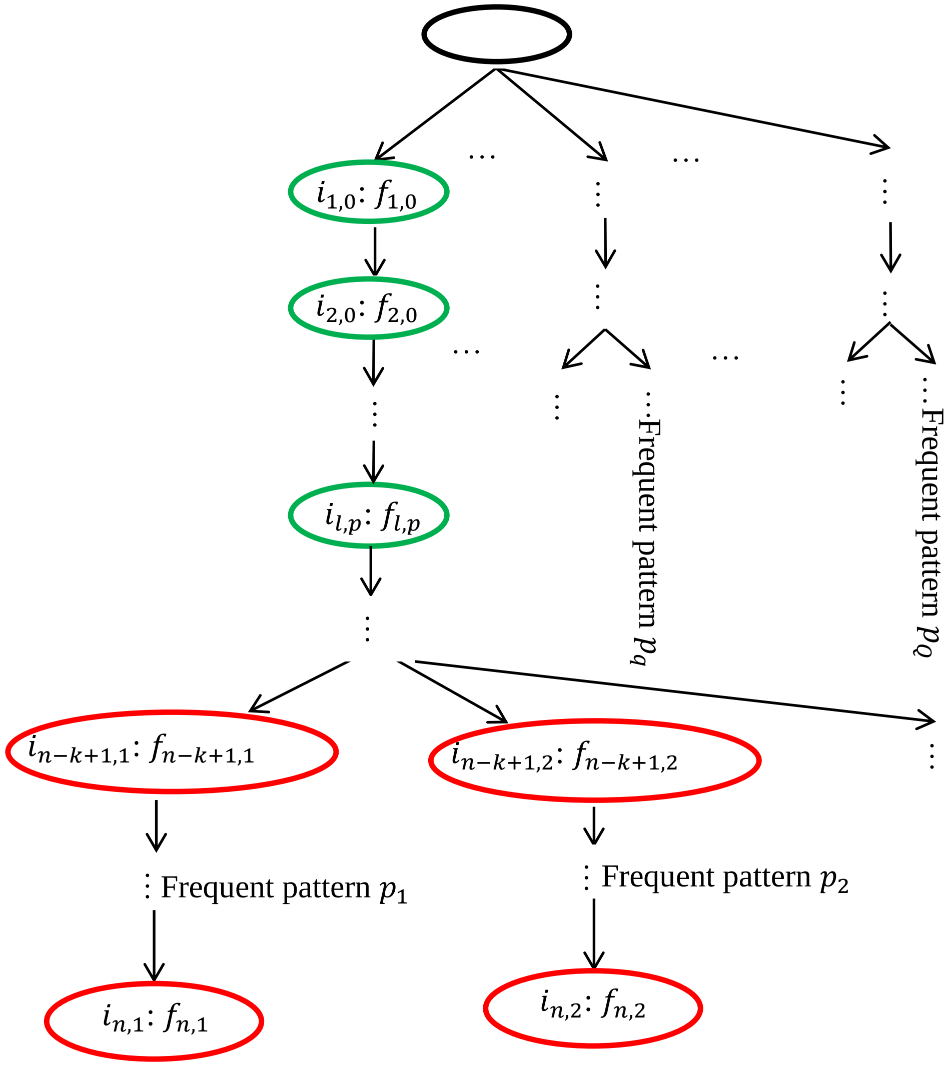


Figure 2. Frequent pattern (FP) tree.

Figure 2 shows an example of a FP tree. Suppose there are TN transactions $[x_1, x_2, \ldots, x_{TN}]$ in database DB, each transaction contains the values of n explanatory variables V_e, $1 \leq e \leq n$, and one response variable V_r which, in our case, denotes whether an accident occurs or not. The FP tree is then built on the TN transactions with n explanatory variables, among which the continuous variables are first transformed to discrete variables by using the Fuzzy C-means clustering method (FCM) as will be discussed in a later section. For more details about how the FP tree is constructed, the reader is referred to Lin et al. (2015).

After the FP tree is constructed and the shared and exclusive nodes identified, the next step is to assign credits or scores to the discrete items in the exclusive nodes, given that these exclusive nodes differentiate the frequent patterns from one another. In this study, we propose a novel variable importance score calculation method based on the Relative Object Purity Ratio, as we describe later in this report.

Figure 3 summarizes the different steps of the variable selection method. In that Figure, we distinguish between the novel aspects of the proposed method (highlighted in bold and italic), and those which we borrow from the previous work reported in the literature. In our subsequent discussion, we focus on those novel aspects but we still briefly describe the other steps as well for the convenience of the reader.

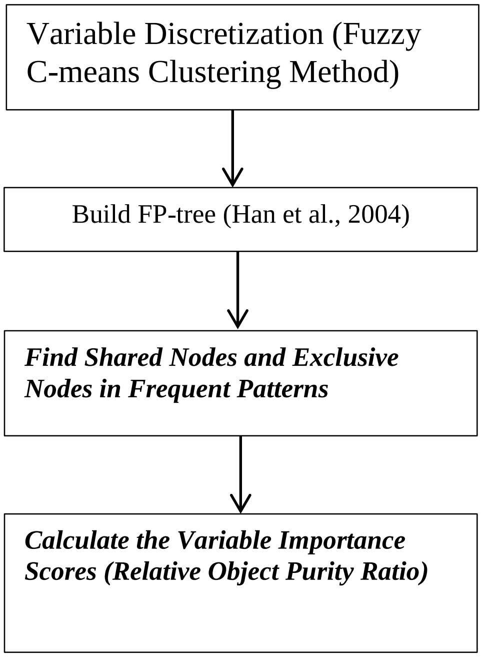

Variable Discretization (Fuzzy C-means Clustering Method)
Build FP-tree (Han et al., 2004)
Find Shared Nodes and Exclusive Nodes in Frequent Patterns
Calculate the Variable Importance Scores (Relative Object Purity Ratio)

Figure 3. Flow chart of variable selection method based on FP tree

Variable importance calculation:

A novel FP tree based variable importance score calculation method is proposed to rank and select the significant explanatory variables for accident risk prediction. The method proceeds as follows.

1. For each frequent pattern p_q, calculate its *object purity ratio* r_q (OPR). OPR refers to the proportion of records falling into this frequent pattern, where their response variable V_r takes the object value o (in this study the object value o is set as 1 which indicates an accident occurrence). r_q can thus be calculated as follows:

$$r_q = \frac{num_q(V_r = o)}{f_{n,q}} \qquad \text{Equation 7}$$

where,

$num_q(V_r = o)$ is the number of records in frequent pattern q which have the response variables V_r as o;

$f_{n,q}$ is the number of records allocated to frequent pattern q.

One issue associated with OPR is that its value is in reference to the proportion of records taking the object value in the whole dataset DB, which can thus lead to inconsistent variable ranking. In this context, it is the difference between the OPR value of a pattern and the average behavior of the entire data that actually distinguishes a pattern. Therefore, we propose the *relative object purity ratio* r_{rq} (ROPR) in this study, where, in its modified version, ROPR represents the absolute difference between the OPR and the proportion of records taking the object value in the whole dataset DB.

$$r_{rq} = abs(\frac{num_q(V_r = o)}{f_{n,p}} - \frac{num_{DB}(V_r = o)}{TN}) \qquad \text{Equation 8}$$

where,

$num_{DB}(V_r = o)$ is the number of records with the response variables V_r as the object value o.

2. Given an observed record located in this frequent pattern, one intuitive thought is that the higher the ROPR is, the purer the frequent pattern is and the more likely the object response value will take place (i.e., in our case, that an accident will occur) or will not happen. Again, we assume that only the discrete items that are in the exclusive nodes play a role in differentiating one frequent pattern from the others. Therefore, the importance score of an item is determined as follows: for each transaction $Tran$ in DB, find its corresponding frequent pattern p_q and exclusive nodes E_q; for each item in $Tran$, if it exists in E_q, add the ROPR to the item's importance score IS_i, otherwise, keep IS_i unchanged.

$$IS_i = \sum_{1\le q\le Q}\sum_{1\le Tran\le TN} r_{rq} * d_{Tran} * d_q * d_e \text{ , } 1 \le i \le m \qquad \text{Equation 9}$$

where,

$d_{Tran} = 1$ if item i is in transaction $Tran$; otherwise $d_{Tran} = 0$;

$d_q = 1$ if p_q is the frequent pattern of the corresponding transaction $Tran$; otherwise $d_q = 0$;

$d_e = 1$ if item i is in the exclusive node set E_q; otherwise $d_e = 0$.

3. After the importance score of each item is calculated, the remaining step is to calculate the importance score of a variable (IS_v).

$$IS_v = \sum_{1 \le i \le m} IS_i * d_v, \; 1 \le v \le n \qquad \text{Equation 10}$$

where,
$d_v = 1$ if item i is one discrete value of variable v; otherwise $d_v = 0$.
At last, the explanatory variables can be ranked based on the variable importance scores.

Variable discretization for FP tree

The FP-tree algorithm requires each transaction in the database to be a set of discrete items. However, in traffic accident risk prediction database, continuous variables such as traffic speed and traffic volume are quite common. In this study, the Fuzzy C-means clustering method (FCM) is used to transform the continuous variables to the discrete variables. FCM is an extension of the k-means methods in which each data point can be a member of multiple clusters with a membership value (soft assignment) (Jain, 2010). For details about how FCM was applied in this study, the reader is referred to (Lin et al. 2015) and to (Hung and Yang, 2001).

Random forest

Random forest is an ensemble learning method for classification and regression. It is widely used to rank the importance of variables in a natural way. Again, suppose there are *TN* records or transactions $[x_1, x_2, \dots, x_{TN}]$ in database DB, each record includes one response variable V_r and a set of explanatory variables $V = [V_1, \dots, V_n]$, a classification and regression tree (CART) $\hat{f}$ for predicting V_r can be built (Breiman et al., 1984) . The prediction error of $\hat{f}$ based on a validation subset of DB is then defined as

$$R\left(\hat{f}, \overline{DB}\right) = \frac{1}{|\,\overline{DB}\,|} \sum_{i \in \overline{DB}} I\left(\hat{f}\left(V_i\right) = V_{ir}\right), \qquad \text{Equation 11}$$

where,

$$I(e) = \begin{cases} 1, & if\ e\ is\ true \\ 0, & if\ e\ is\ false \end{cases};$$

$\overline{DB}$ is the validation data subset;
V_{ir} is the observed value of the response variable of the i^{th} record.

However, CART is known to be unstable as a small perturbation of the training sample may change the prediction results. To overcome this, Breiman introduced the random forest algorithm (Breiman, 2001): the trees are built over n_{tree} bootstrap samples $\overline{DB}^1, \dots, \overline{DB}^{n_{tree}}$ of the training data DB; for each tree, different from the CART algorithm, a subset of variables n_{var} is randomly chosen for the splitting rule at each node; each tree is then fully grown until each node is pure. The trees are not pruned. The resulting learning rule is the aggregation of all the tree-based estimators denoted by $\hat{f}_1, \dots, \hat{f}_{n_{tree}}$ (Gregorutti et al., 2013). The class with the maximum number of votes among the n_{tree} trees in the forest is the predicted class of an observation.

The Gini criterion is used to select the split with the lowest impurity at each node. As a useful byproduct of random forests, the Gini variable importance measure can be calculated once the forest is formed: at each split, the decrease in the Gini node impurity is recorded for variable V_i in $[V_1, \dots, V_n]$, and the average of all the decreases in the Gini impurity in the forest where V_i

forms the split is its Gini variable importance. At last, the variables can be ranked according to the Gini variable importance measure (Archer and Kimes, 2008). Besides this, Breiman also proposed other measures like the permutation importance, the z-score and so on (Breiman, 2001).

k nearest neighbor (k-NN)

k-NN is a classification method that decides the class of an object by finding its k-nearest neighbors (i.e. most similar) based on its explanatory variables in the training dataset. The Euclidean distance is typically used to assess similarity (Lin et al., 2013). When k nearest neighbors are found, the following equation (12) $R\left(\hat{f},\overline{DB}\right)=\frac{1}{|\overline{DB}|}\sum_{i\in DB} I\left(\hat{f}\left(V_i\right)=V_{ir}\right)$,

Equation 11can be used to determine the class of the object (Murphy, 2012):

$$p(y=c\mid X,D,k)=\frac{1}{k}\sum_{i\in N_k(X,D)} I(y_i=c) \qquad \text{Equation 12}$$

where,

$N_k(X,D)$ are the k nearest neighboring points to object X in point set D;

$I(e)=\begin{cases}1, \, if\ e\, is\, true\\ 0, if\ e\, is\ false\end{cases}$;

y_i is the response variable of neighboring point i;

y is the response variable of object X;

c is the one of the possible classes.

Bayesian network

By the chain rule of probability, a joint distribution can be represented as follows:

$$p\left(x_{1:V}\right)=p\left(x_1\right)p\left(x_2|x_1\right)p\left(x_3|x_2,x_1\right)p\left(x_4|x_1,x_2,x_3\right)\ldots p(x_V\mid x_{1:V-1}) \qquad \text{Equation 13}$$

where,

V is the number of variables;

$1{:}V$ denotes the set $\{1,2,\ldots,V\}$.

Suppose all the variables have K discrete states, we can create $p(x_1)$ as a table of $O(K)$ numbers, representing a discrete distribution (there are actually only K-1 free parameters because of the sum-to-one constraint, but we write $O(K)$ for simplicity). Similarly, we can create $p(x_2|x_1)$ as a table of $O(K^2)$ numbers, and $p(x_3|x_2,x_1)$ as a table with $O(K^3)$ numbers, and so on. These tables are called conditional probability tables (CPTs). As can be seen, the conditional distributions p$(x_t|X_{1:t-1})$ become harder to estimate as t gets larger (Murphy, 2012).

A Bayesian network is an efficient tool to overcome this problem. Specifically, a Bayesian network is a directed graphical model representing a joint distribution by making conditional independence (CI) assumptions. The nodes in the graph represent random variables, and the edges represent the CI assumptions. More details can be found in Lin et al. (2015).

MODELING DATASET

The dataset used in this study includes the traffic accident records collected on a segment on interstate highway I-64 in Norfolk, Virginia in 2005, as marked in Figure 4.

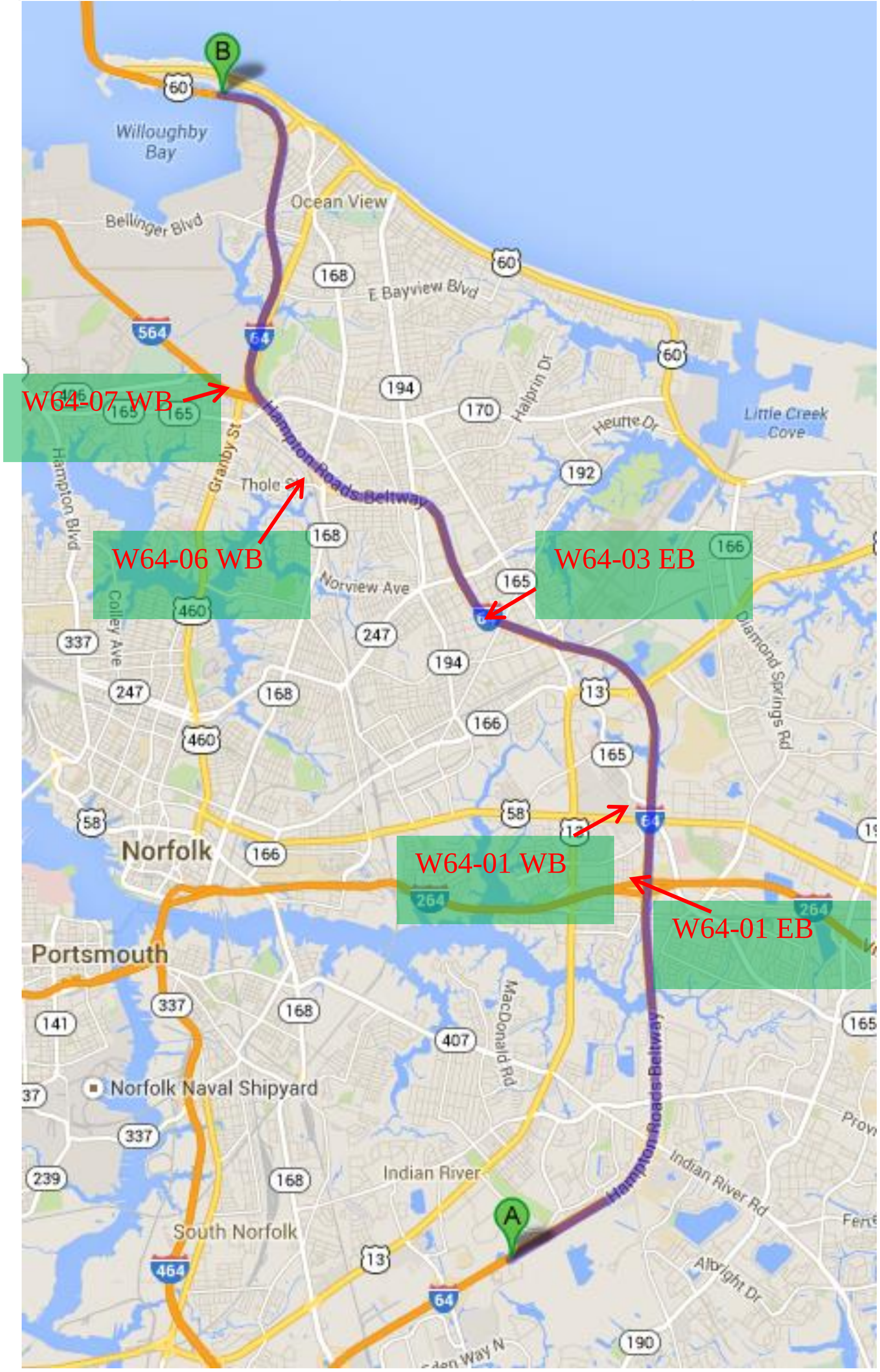


Figure 4. Part of I-64 in Norfolk, Virginia.

The accidents were stored in the Virginia Department of Transportation (VDOT's) Archived Data Management System (ADMS). Besides that, this dataset also contains weather, visibility, traffic volume, speed, and occupancy information, with one minute resolution.

However, this dataset by itself cannot be directly applied to predict real-time traffic risk directly. As a classification problem, the pre-crash condition and normal traffic condition have to be defined first (Hossain and Muromachi, 2012). Some studies defined the pre-crash condition as a time period starting right before an accident and extending up to 5 or 10 minutes (Oh et al., 2005; Zheng et al., 2010), while some studies defined it as a 5 minute time period starting from a close time point such as 4 or 5 minutes before the accident (Abdel-Aty et al., 2008; Hossain and Muromachi, 2012).

In this study, as shown in Figure 5. we used two temporal settings to define the pre-crash condition: the first one is a 10-minute time period starting from 5 minute before the accident, and the other is in a 5-minute time period starting from 5 minute before the accident. The normal condition is defined as the same time period as the pre-crash condition, but taking place on the same day of the other weeks from two weeks earlier to two weeks later than the day of the week with an accident. It needs to note that a normal condition data point is excluded if there is an accident happening within one hour before or after the designated time (Hossain and Muromachi, 2012).

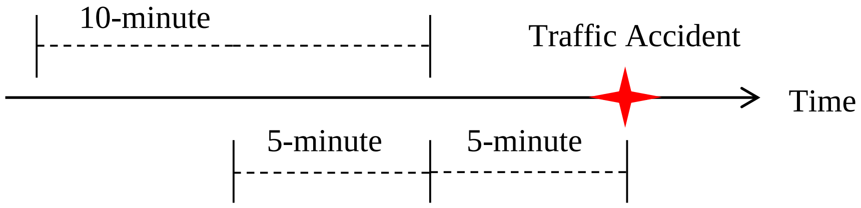


Figure 5. Temporal settings of pre-crash and normal traffic conditions

After the pre-crash condition and normal traffic condition are defined, the relevant data can be extracted given the number and locations of traffic detectors in place. Most of the previous studies considered more than one detector during the extraction process, such as one upstream detector and one downstream detector (Abdel-Aty et al., 2008), and two upstream detectors, two downstream detectors and one detector covering the accident location (Hossain and Muromachi, 2012). Due to the problem of missing data, we were forced to rely on only one detector, that is to say, the one reporting an accident. There are five such detectors, labeled W64-01 EB, W64-01 WB, W64-03 EB, W64-06 WB and W64-07 WB, their approximate locations are marked in Figure 4.

At last, two datasets were obtained, which differ from each other in terms of the time period used to define the pre-crash and normal traffic condition (the first DB has a time period of 10-minute long, and the second one is 5-minute long). Eight explanatory variables were contained in the data, including: the mean of the weather condition ($Mean_{wea}$) as defined below, the mean of visibility ($Mean_{vis}$), the mean and standard deviation of the traffic volume ($Mean_{vol}$ and Std_{vol}, unit: vehicle per hour), the mean and standard deviation of the traffic speed (Mean_{spe} and Std_{spe}, unit: mph), and the mean and standard deviation of the occupancy (Mean_{ocu} and Std_{ocu}). The accident response variable is defined as a binary variable with value 1 for the pre-crash situation and 0 for normal traffic. It is worth noting that the weather variable was a categorical variable originally with 26 possible different weather types.

We used the numbers 0 to 25 to represent these different weather types that range from fine weather like “clear” to extreme inclement weather like “thunderstorm”. Although typically, the weather condition will not change significantly within a 5- or 10- minute period, we nevertheless, take the mean value of the weather over that period. The resulting variable, therefore, may theoretically assume a non-integer value and can be assumed as a continuous (and not discrete) variable. The same applied for “visibility”, which is also a continuous variable ranging from 0 to 10 miles.

After processing, the 5-minute accident dataset included 170 pre-crash records and 555 normal traffic records, and the 10- minute accident dataset included 174 pre-crash records and 569 normal traffic records. Note that the 5-minute accident dataset has fewer records because of the higher probability of data missing for 5 minute period than the 10 minute period. For each dataset, 80% of the pre-crash records and normal traffic records were randomly chosen as the training dataset while the remaining 20% were taken as the test dataset.

MODEL DEVELOPMENT AND RESULTS

Variable importance calculation

Two *training* datasets are generated through the random sampling with the 80% rate, including a 5-minute training dataset with 136 pre-crash records and 444 normal traffic records and a 10-minute training dataset with 139 pre-crash records and 455 normal traffic records. For each training dataset, FCM was first applied to transfer a continuous variable to a discrete cluster variable.

Table 6. Clustering results for 5-minute and 10-minute accident training datasets

datasets	Variable	Cluster 1 low	Cluster 2 medium	Cluster 3 high
5-minute training dataset	Mean_{wea}	[0, 5]	[6, 16]	[17, 25]
	Mean_{vis}	[0.13, 4.25]	[5, 8]	[8.8, 10]
	Mean_{vol}	[60, 564]	[576, 1164]	[1176, 1908]
	Mean_{ocu}	[1, 8.2]	[8.4, 27.6]	[31.2, 66.4]
	Mean_{spe}	[0, 33.6]	[34, 59.8]	[60, 93]
	Std_{vol}	[0, 112.24]	[115.41, 245.19]	[247.38, 642.58]
	Std_{ocu}	[0, 3.96]	[4, 15.66]	[26.62, 27.07]
	Std_{spe}	[0, 4.15]	[4.21, 11.73]	[12.19, 33.16]
10-minute training dataset	Mean_{wea}	[0, 5]	[6, 16]	[17, 25]
	Mean_{vis}	[0.25, 4.8]	[5, 8]	[8.5, 10]
	Mean_{vol}	[60, 560]	[564, 1152]	[1170, 1890]
	Mean_{ocu}	[1, 7.9]	[8, 28]	[29.7, 66.4]
	Mean_{spe}	[0, 31.6]	[34.37, 59.8]	[59.85, 94.5]
	Std_{vol}	[0, 124.73]	[124.9, 245.68]	[248.51, 699.74]
	Std_{ocu}	[0, 4.17]	[4.36, 16.06]	[17.79, 31.10]

Std$_{spe}$	[0, 4.63]	[4.65, 13.48]	[13.62, 31.78]

The clustering results are shown in Table 6. Three clusters were generated for each continuous variable, representing: low, medium and high value ranges. The two numbers in each bracket denotes the lower bound and upper bound of a cluster. Through this process, the original eight continuous explanatory variables were transferred into 24 discrete variables (called items in the following analysis). The support of each item or the size of each cluster were obtained and sorted in a descending order as shown in Table 7:

Table 7. Supports of items in 5-minute and 10-minute accident training datasets

Index	5-minute training dataset		10-minute training dataset	
	Item	Support	Item	Support
1	Std$_{ocu}$ low	542	Std$_{ocu}$ low	554
2	Mean$_{wea}$ low	435	Mean$_{wea}$ low	436
3	Mean$_{vis}$ high	401	Mean$_{vis}$ high	405
4	Mean$_{ocu}$ low	385	Mean$_{ocu}$ low	383
5	Std$_{spe}$ low	372	Std$_{spe}$ low	373
6	Mean$_{spe}$ medium	337	Mean$_{spe}$ medium	332
7	Std$_{vol}$ low	321	Std$_{vol}$ low	327
8	Mean$_{vol}$ medium	261	Mean$_{vol}$ medium	261
9	Mean$_{spe}$ high	216	Mean$_{spe}$ high	231
10	Mean$_{vol}$ low	170	Mean$_{vol}$ low	188
11	Mean$_{ocu}$ medium	169	Std$_{spe}$ medium	188
12	Std$_{vol}$ medium	169	Std$_{vol}$ medium	184
13	Std$_{spe}$ medium	169	Mean$_{ocu}$ medium	178
14	Mean$_{vol}$ high	149	Mean$_{vol}$ high	145
15	Mean$_{vis}$ medium	125	Mean$_{vis}$ medium	132
16	Mean$_{wea}$ medium	98	Mean$_{wea}$ medium	109
17	Std$_{vol}$ high	90	Std$_{vol}$ high	83
18	Mean$_{vis}$ low	54	Mean$_{vis}$ low	57
19	Mean$_{wea}$ high	47	Mean$_{wea}$ high	49
20	Std$_{spe}$ high	39	Std$_{ocu}$ medium	34
21	Std$_{ocu}$ medium	35	Mean$_{ocu}$ high	33
22	Mean$_{spe}$ low	27	Std$_{spe}$ high	33
23	Mean$_{ocu}$ high	26	Mean$_{spe}$ low	31
24	Std$_{ocu}$ high	3	Std$_{ocu}$ high	6

When screening frequent items, we set the threshold value σ in equation (1) to 0 so that all the items shown in Table 7 are considered. The rationale behind this is to prevent any information loss in the variable importance score calculation. Since the items have already been sorted in a support-based descending order, Table 7also provides the F-List to build the FP Tree. The reader is referred to Lin et al. (2015) for an example FP tree built from the training dataset.

With the FP Tree constructed, the variables’ importance scores are calculated using equation (3), (4) and (5). The results are shown in Table 8.

Table 8. Variable importance calculations results based on FP Tree and random forest methods

Variables	5-minute training dataset		10-minute training dataset	
	FP tree	Random Forest	FP tree	Random Forest
Mean_{vol}	46 (1)*	27.31 (3)	48.6 (1)	27.51 (4)
Std_{vol}	43.2 (2)	26.98 (4)	42.8 (2)	29.15 (2)
Mean_{spe}	16.6 (7)	28.56 (2)	20.8 (7)	29.02 (3)
Std_{spe}	35.6 (4)	29 (1)	29 (6)	30.11 (1)
Mean_{ocu}	21.6 (6)	25.99 (5)	35.8 (4)	24.89 (6)
Std_{ocu}	15.2 (8)	22.19 (6)	15 (8)	26.41 (5)
Mean_{wea}	40.2 (3)	8.79 (8)	37.6 (3)	9.88 (8)
Mean_{vis}	33.8 (5)	13.77 (7)	30.8 (5)	13.39 (7)

Notes: * The first number is the variable importance score, and the number in the following parentheses is the ranking of variable (“1” means the most important, and “8” means the least important).

We also calculated the variables importance scores based on random forest method (see Table 8), using the package “randomForest” within the statistics software R (Liaw and Wiener, 2002). For more details about calculating the importance scores and the random forest method, the reader is referred to Lin et al., 2015 and Efron and Tibshirani, 1997.

With this, for the 5-minute training dataset, the sample size was set as 366, and for the 10-minute training dataset, the sample size was set as 375. The package “randomForest” produced the mean decrease of the Gini index for each variable as an output. As mentioned before, the mean decrease of the Gini index, measures the contribution of a variable to the homogeneity of the nodes and leaves in the random forest (Metagenomics Statistics, 2014). The higher the mean decrease of the associated Gini index is, the more important the variable is.

Through the comparison of the variable importance scores generated from the FP tree and the Random forest, we can see that the two models produce different variable importance rankings. The FP tree models tended to rank traffic volume related variables, such as Mean_{vol} and Std_{vol} as the top two most important variables while resulting in much lower scores for speed related statistics, particularly for Mean_{spe}. In contrast, traffic speed related statistics variables were deemed slightly more important by the random forest. Nevertheless, the volume related variables were judged important by both of the methods (among the top four). As for the weather related variables, Mean_{wea} was ranked as the third most important variable based on the FP method, while it was scored as the least important by the random forest method.

k-NN

This study tested the performance of k-NN for the 5-minute and 10-minute testing datasets. k was set as 2 and 3 separately, and each time k-NN was run for three scenarios: (1) using all the variables; (2) using all variables except for Meanspe and Stdocu which were ranked as the least important by the FP tree method; and (3) using all variables except for Meanwea and Meanvis that were ranked as the least important by random forest. The voting criterion of k-NN in this study is that once one of k nearest neighbors has the response variable equal to 1 (indicating the occurrence of an accident), the predicted response of the observation is set as 1. The results can be seen in Figure 6.

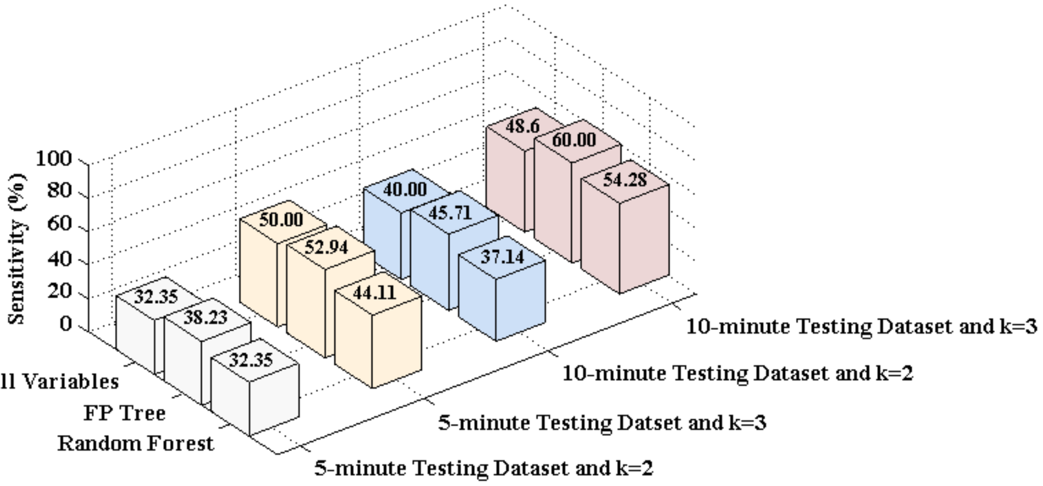


Figure 6a. Comparison of Sensitivity

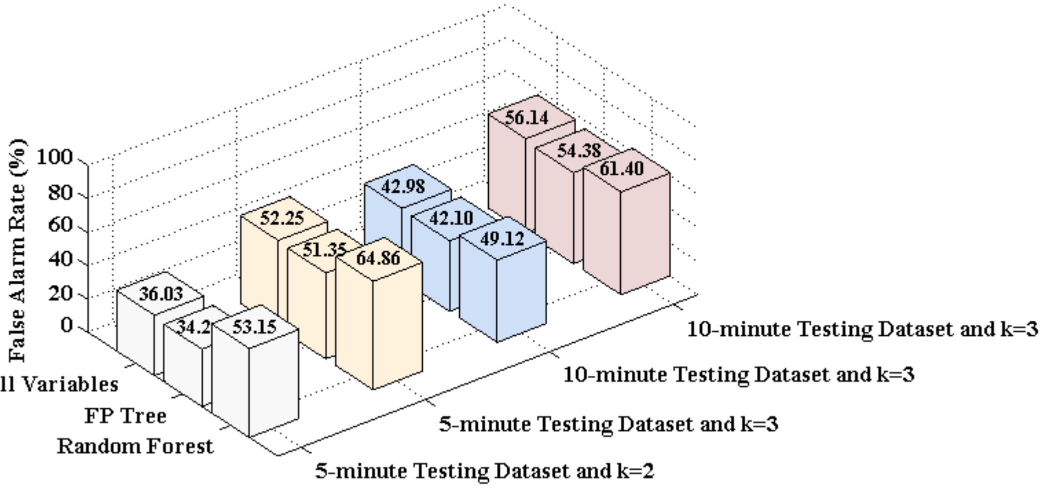


Figure 6b. Comparison of False Alarm Rate

Figure 6. Performance of k-NN for different variable selection

Note that there are two prediction performance measures used as shown in Figure 6a. and Figure 6b. These are: (1) the sensitivity, which measures the proportion of actual accidents that were accurately predicted as such; and (2) the false alarm rate that refers to the proportion of normal situations that were wrongly predicted as accidents. A good traffic accident risk prediction model should yield a high sensitivity and a low false alarm rate.

The major findings are summarized below according to Fig.6. First of all, although k-NN doesn’t perform well in general, using the FP tree to pre-select the explanatory variables significantly improved the prediction accuracy. In comparison to the “all variables case”, the FP tree based k-NN model consistently produced higher prediction sensitivity values and lower false alarm rates, regardless of the testing dataset used. In contrast, there was generally no benefit from the random forest based variable selection, with the only exception of the case of k=3 with the 10-minute testing dataset where the variable selection with the random forest method generated a higher prediction sensitivity than the case using all the variables. This indicates the advantage of FP tree in sorting out important affecting factors and improving model prediction performance. Second, regardless of the type of the testing datasets used, the comparison between the k-NN model with k=3 and the one with k=2 shows that adding one nearest neighbor will significantly increase the prediction sensitivity; however, as can be seen, this will also increase the false alarm rate. Lastly, the k-NN models work better for the 10-minute testing dataset than for the 5-minute testing dataset in terms of prediction sensitivity. However, the false alarm rates tend to be higher for the 10-minute testing dataset as well. This indicates that the pre-crash time period may also affect model performance.

Bayesian network

Bayesian network models were also built to predict accident risk for comparison. As a crucial step to perform Bayesian network modeling, the continuous variables need to be discretized. How to transform a continuous variable to discrete category variables vastly depends on the objectives set by researchers (Hossain and Muromachi, 2012). Among the discretization techniques available in the literature, we selected the normalized equal distances (NED) method, using the software Bayesialab due to its promising performance (Bayesia, 2013). The values of each variable are first normalized based on the mean and standard deviation of the variable (Han et al., 2006). Then, the normalized values are split to the user-defined number of equal width discrete intervals (Kotsiantis and Kanellopoulos, 2006). For this research, we set the number of equal width discrete intervals as 3 and 4, respectively.

We considered one of the most plausible Bayesian network structures, which just let the response variable be the child node of the possible explanatory variables (Hossain and Muromachi, 2012). Three scenarios, as before, were tested under the structure: (1) using all the variables; (2) using all except of $Mean_{spe}$ and Std_{ocu} that are ranked as the least important by FP tree; and (3) using all except $Mean_{wea}$ and $Mean_{vis}$ that are ranked as the least important by random forest. The software Netica was used to learn the Bayesian network parameters (Netica tutorial, 2014). For more details, the reader is referred to Lin et al., 2015.

The performance of Bayesian networks with different NED numbers (in parentheses), and for the 5-minute and 10-minute testing datasets are shown in Figure 7.

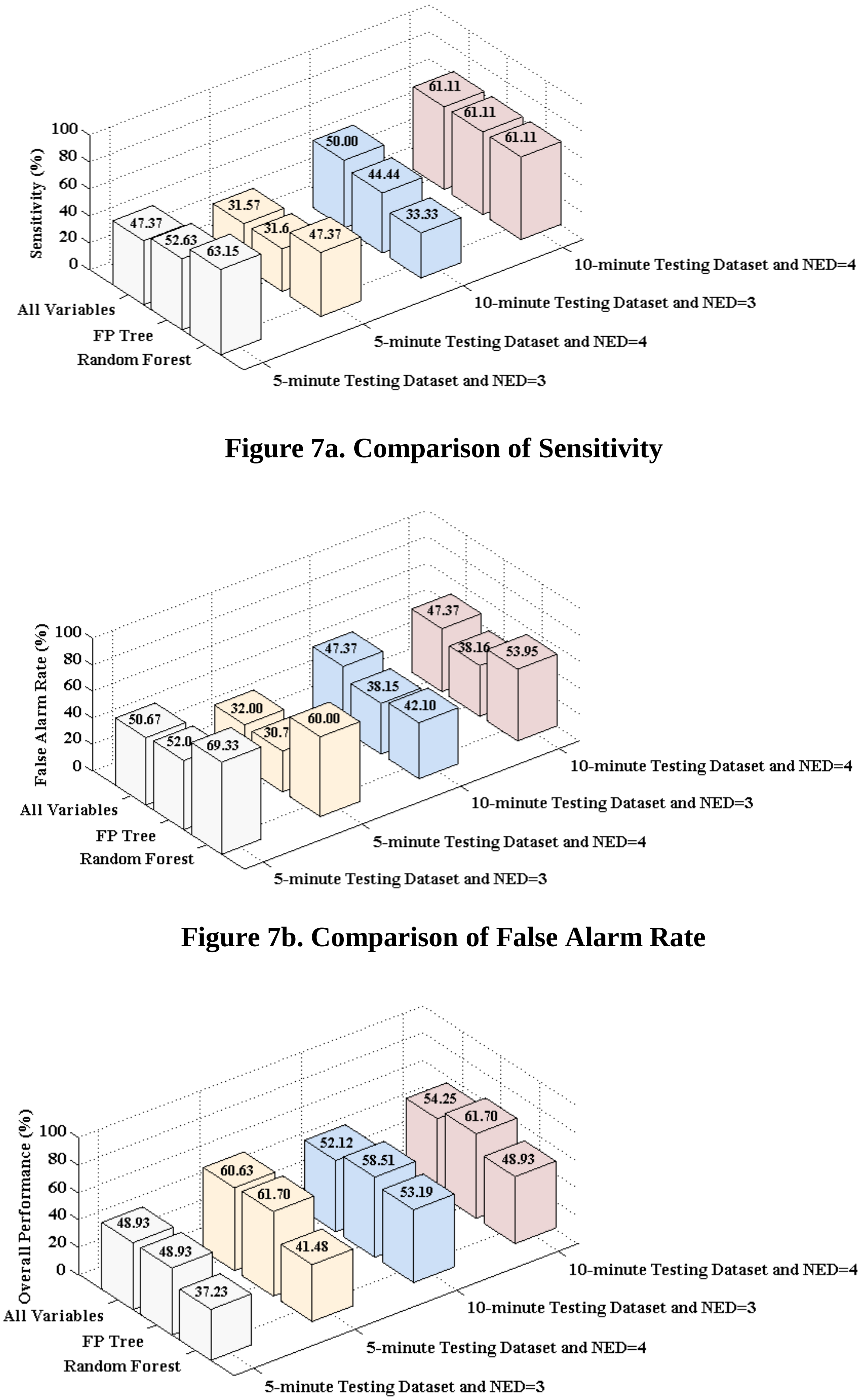


Figure 7a. Comparison of Sensitivity

Figure 7b. Comparison of False Alarm Rate

Figure 7c. Comparison of Overall Performance

Figure 7. Bayesian network Performance with different variable selection strategies.

Several observations can be discerned from Figure 7. First, based on Figure 7a., which compares the sensitivity values, and Figure 7b., which compares the false alarm rate, the best Bayesian network model results in a sensitivity value as high as 61.11% and a false alarm rate as low as 38.16%, when trained based on the 10-minute dataset with the NED number equal to 4. These results compare very favorably to those obtained by previous studies reported in the literature as shown in Table 9. This is especially true given that the current study, because of missing data, had to rely on data collected from only a single detector (the one reporting the traffic accident), whereas most of the previous studies extracted the relevant variables from both upstream and downstream detectors relative to the crash location.

As can be seen, for the previous studies the sensitivity values are usually around 60%, and the false alarm rate ranges between 20% and 50%. The best result from previous studies is that reported by Hossain and Muromachi (2012) with a sensitivity of 66% and a false alarm rate of 20%. In that study, however, for each record in the database, information were extracted from two upstream detectors, two downstream detectors and the one nearest to the traffic accident; we did not have the luxury of such data in the current study.

Table 9. Comparison with the previous studies

Authors	More than One Detector	Variable Selection Method	Traffic Accident Prediction Method	Sensitivity	False Alarm Rate
Abdel-Aty et al., (2004)	Yes	N/A	Logistic Regression	69%	N/A
Pande and Abdel-Aty (2006)	Yes	Classification Tree	Neural Network	57.14%	28.83%
Abdel-Aty et al., (2008)	Yes	Random Forecast	Neural Network	61%	21%
Hossain and Muromachi (2012)	Yes	Random multinomial logit	Bayesian Network	66%	20%
Ahmed and Abdel-Aty (2012)	Yes	Random Forecast	Matched Case-Control Method	68%	46%

Secondly, the results, shown on Figure 7a. and Figure 7b., show that the number of NED could affect the performance of the Bayesian network. For the 5-minute dataset, the sensitivity and false alarm rate both decreased when the number of NED was set to 4 instead of 3. On the other hand, for 10-minute dataset, the sensitivity improved, but the false alarm rate remained almost the same when NED number is changed from 3 to 4, except for the situation using the variables based on random forest, for which the false alarm rate also increased.

Third, for the majority of cases, the Bayesian network models using variables selected by FP tree perform better than the ones using the random forest selected variables. For example, for the 10-minute dataset, when NED number is 4, although the sensitivity values of the two types of models are somewhat similar (around 61%), the false alarm rate of the random forest based Bayesian network model is much higher than its FP tree based counterpart. For other cases, however (e.g., the models based on the 5-minute training dataset), it is hard to decisively conclude that the models based on FP tree performed better than those based on random forest because the sensitivity and false alarm rate of the former are *both* lower than those of the latter. Because of this, we introduced a third criterion shown in Fig. 8c., called the overall performance to measures the ratio of correct predictions (no matter whether it is accident or a non-accident) in the whole testing dataset. Based on the overall performance criterion, we can easily see that the models based on variables selected by FP tree significantly outperform those based on all the variables or based on random forest for the 10-minute testing dataset. For the 5-minute testing dataset, the models based on variable selected by FP tree have the same as or a little higher overall performance than the models using all variables. The models based on random forest variable selection in this case had the lowest overall performance.

CONCLUSIONS AND FUTURE WORK

In this part of the study, we proposed a novel variable selection algorithm based on FP tree for real-time traffic accident risk prediction. The importance score of each explanatory variable in the dataset is calculated and ranked through the calculation of the ROPR of the corresponding frequent patterns. This variable selection algorithm was tested on the Virginia traffic accident dataset collected in 2005 in comparison to the widely used random forest variable selection. Based on the variables selected by the two methods, two traffic accident risk prediction models, the k-NN and Bayesian network models, were developed and tested for three situations: using all variables, using the important variables selected by FP tree, and using the important variables selected by random forest. The major findings are summarized as below:

1. Generally, the accident risk prediction results are quite acceptable when using the Bayesian network model with NED number equal to 4 and based on a 10-minute dataset. This is especially true for the case using variables selected by FP tree, where the sensitivity was as high as 61.11% and the false alarm rate was as low as 38.16%. Considering that only data from one detector were available in this study, these results are very promising.

2. In terms of the time resolution to be used in compiling the datasets, no decisive conclusions can be made regarding whether a 5-minute or a 10-minute resolution would yield better performance. For Bayesian network, the overall performances are improved by using the 10-minute dataset except the cases with NED number set as 4, using all variables and FP tree based variables.

3. The most important finding of this part of the study is that the accident risk prediction models based on FP tree variable selection outperform the models based on all variables and the ones based on random forest, regardless of the settings of the prediction models such as the selection of k for k-NN, the NED number selected for Bayesian network, and the pre-crash time period used in the datasets. Being insensitive to the selection of the models' parameters is a good quality that the FP tree variable selection algorithm appears to possess.

4. For the applications of the novel variable selection method and traffic accident risk prediction model, given that this is a classification problem in essence, both traffic accident records and normal traffic conditions extracted from the same segment of the road are needed to train and test the models. However, this study shows that records from different segments of the road can be put together in order to generate a bigger dataset. For example, the dataset in this study include the corresponding records reported by five detectors from different road segments of I-64.

As a novel algorithm, there are still a lot of details to be finalized in the future. For example, we may test the impact of clustering number in FCM on the FP tree variable importance calculation (in this study, we just set it as 3), and we may also try other variable discretization methods. Besides that, there are some other variable reduction/selection algorithms, such as stratified random forest (Ye, et al., 2013), and random projection (Fan, et al., 2013) that deserve to be explored. We will also test other accident risk prediction methods such as support vector machine (SVM) as our future work.

A COMBINED M5P TREE AND HAZARD-BASED DURATION MODEL FOR PREDICTING URBAN FREEWAY TRAFFIC ACCIDENT DURATIONS

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 5 phases (Zhan et al., 2011): (1) incident detection, which refers to the time interval from the occurrence of the incident to its detection; (2) incident verification that covers the period from the detection to the confirmation of the incident; (3) incident response spanning from the moment an incident is confirmed to the time when the first responder arrives on the scene; (4) incident clearance 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) incident recovery covering the time until normal traffic conditions resume.

A critical component of effective TIM involves the ability to predict the likely incident duration under various conditions. 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.

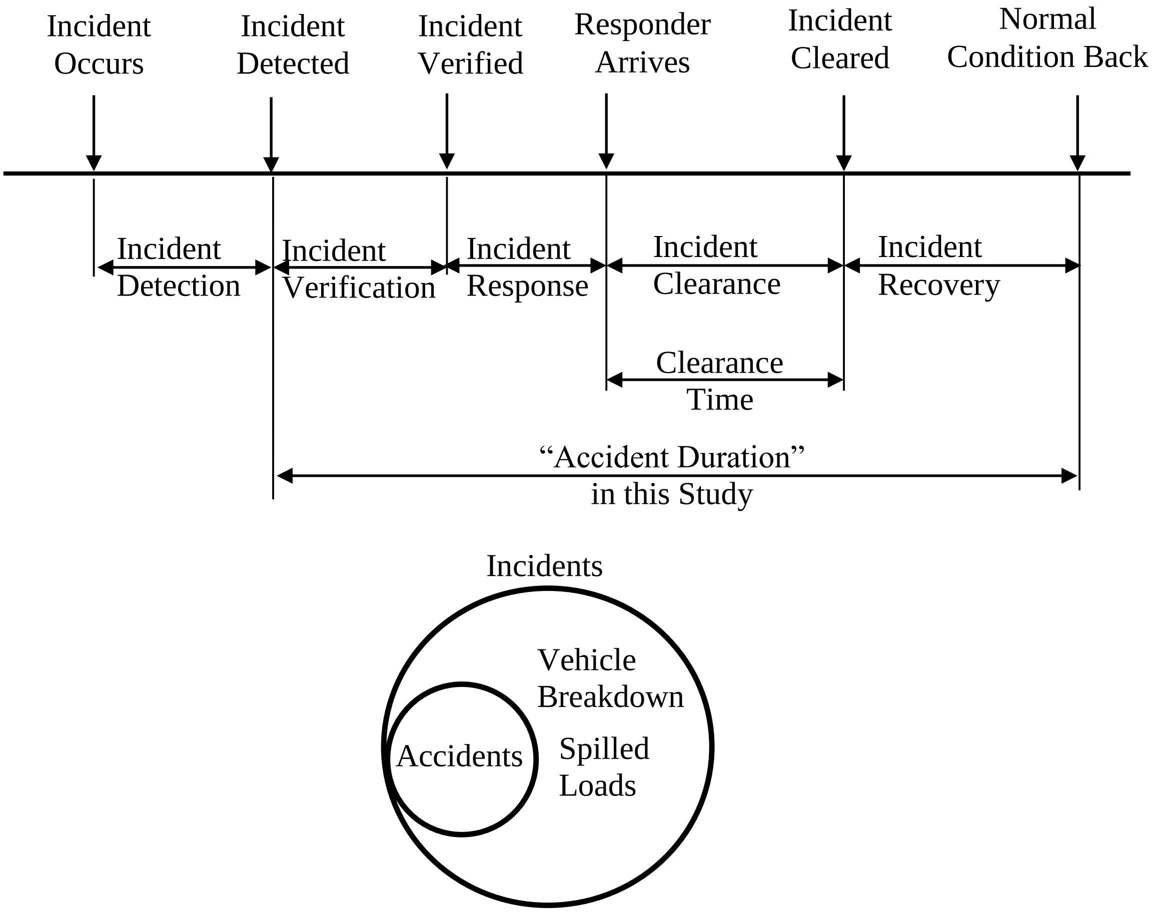


Figure 8. Traffic incident management process and accident duration definition

This study 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 classification, 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.

The organization of this third major section of the report is as follows. The section 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.

PREVIOUS RESEARCH ON INCIDENT DURATION PREDICTION

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, decision trees were used in previous research to predict incident durations (He et al., 2013; Ozbay et al., 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 utilized. 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.

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 (Nam and Mannering, 2000).

Clustering and classifying the traffic accident data is another way to minimize the heterogeneity problem. One way to classify traffic incidents is based on 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, as a first step before accident analysis.

METHODOLOGY

As previously mentioned, this study 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 et al., 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.

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 Equation 14

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

Equation 14

Where T_i^n denote the subset of cases that have the *ith* outcome of the potential test, $sd(T_i^n)$ 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. $\sum_{i=1}^{O} \frac{|T_i^n|}{|T^n|} \times sd(T_i^n)$ 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 $TH2 * sd(T^0)$. 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 Equation 15:

$$\boldsymbol{Error = \frac{N+v}{N-v} * \frac{\sum_{i=1}^{N} abs(V_{act} - V_{pre})}{N}}$$

Equation 15

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).

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 Equation 16:

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

Equation 16

The corresponding probability density function is thus given as:

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

Equation 17

where $f(d)$ describes the instantaneous failure rate in the infinitesimally small interval $[\mathrm{d}, \mathrm{d} + \Delta\mathrm{d}]$. Also given *F(d),* the survival function, $S(d)$, is defined as in Equation 18

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

Equation 18

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 Equation 19 as follows:

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

Equation 19

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

$$\tau = exp(-x_i\beta) * d_i$$

Equation 20

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 Equation 21:

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

Equation 21

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 et al., 2008).

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 equation (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 **Equation 20**. 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.

MODELING DATASETS

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 10.

Table 10. Traffic accident variables in I-64 dataset

Variables	Values
Season	Spring (March, April, May); Summer (June, July, August); Autumn (September, October, November); Winter (December, January, 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, …

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.

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 information 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 11summarizes the variables included in the Buffalo-Niagara dataset.

Table 11. Traffic accident variables in I-190 dataset

Variables	Values
Season	Spring (March, April, May); Summer (June, July, August); Autumn (September, October, November); Winter (December, January, 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, …

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.

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.

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 $TH1$; and (2) the ratio of the standard deviation $TH2$, previously mentioned.

Although the value of $TH1$ 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 $TH1$ values ranging from 5% to 10% of the total number of training cases (i.e., values between 25 and 50). After some experimentation, $TH1$ was set to 30, and $TH2$ was set to 0.95. Figure 1Figure 9 shows the resulting M5P tree model for the I-64 dataset.

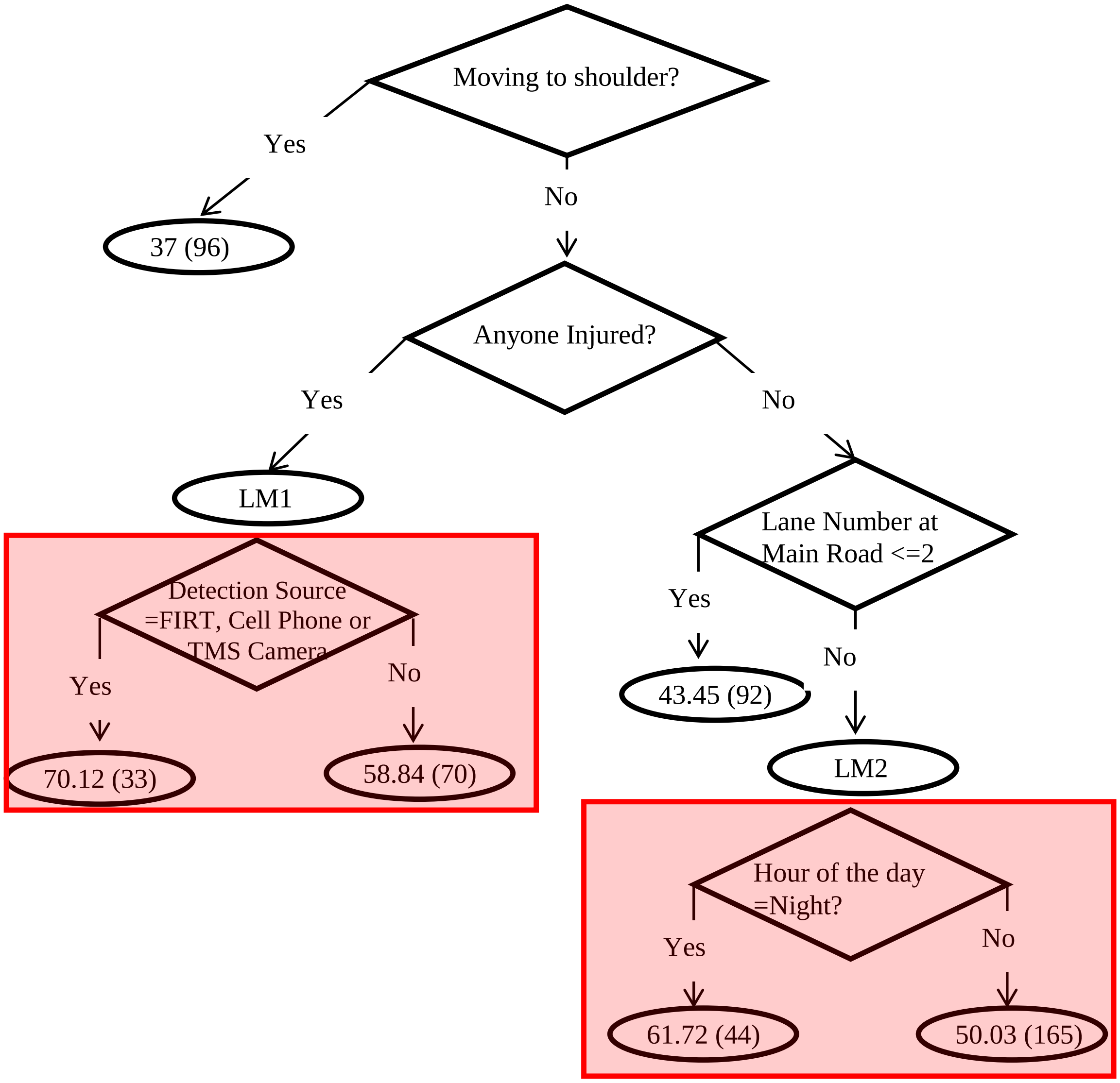


Figure 9. M5P tree model for I-64 training dataset.

As can be seen in Figure 9., 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 Figure 9.). The details of LM1 and LM2 are listed below.

LM1: Duration=62.46 minutes (103 cases);

LM2: Duration=52.49 minutes (209 cases);

As can be seen, the two linear regression models developed here are basically two constants. As discussed before, 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 leaf.

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 minutes. 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 minutes (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 minutes, 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 minutes as given 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, $TH1$ was set to 35, and $TH2$ to 0.75. Figure 10 shows the M5P tree model that resulted.

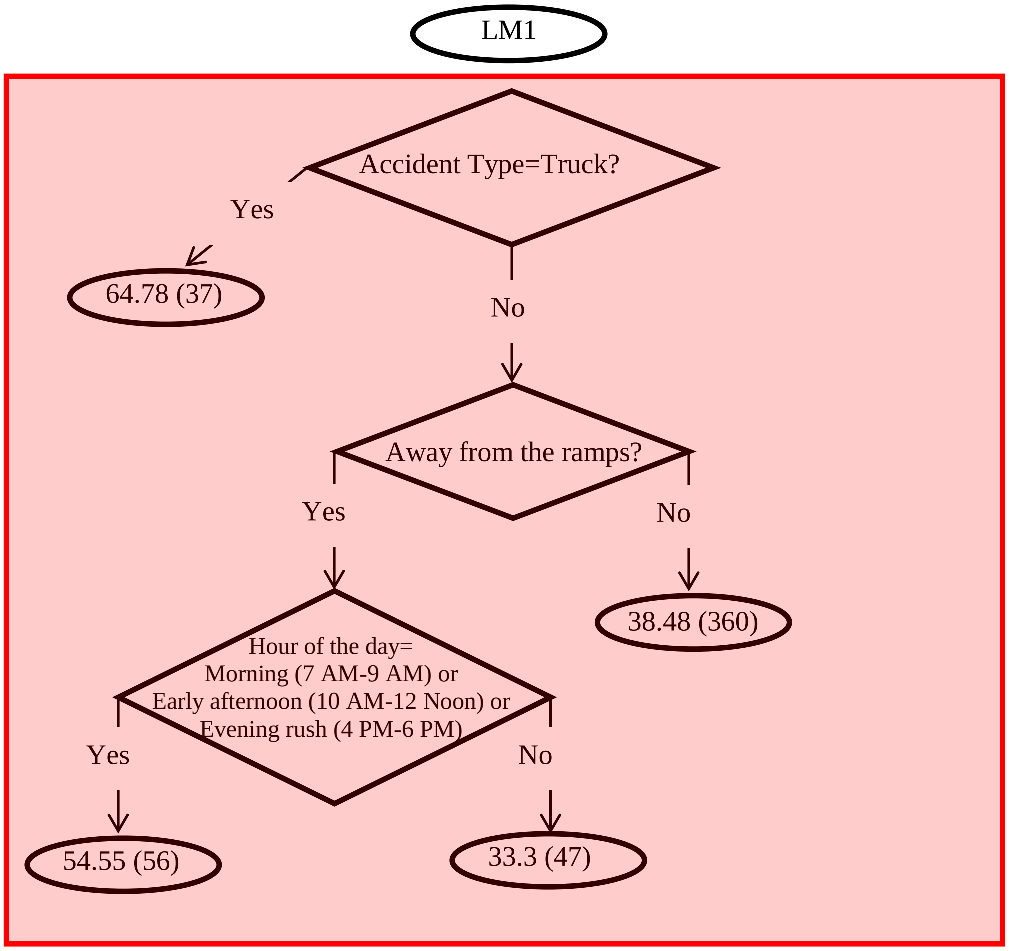


Figure 10. M5P tree model for I-190 training dataset.

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 minutes, 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 minutes.

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, 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).

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 **Equation 20**. Secondly, the significant explanatory variables x_i need to be determined. In this study, 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 below (Alkaabi et al., 2011; Cleves et al., 2008):

$$AIC = -2lnL + 2(k + c)$$

Equation 22

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 12 . 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 study. It is to be noted that this is consistent with other studies reported in the literature (Golob et al., 1987; Chung, 2010).

Table 12. AIC values of HBDMs for I-64 and I-190 training datasets

Model	I-64 dataset				I-190 dataset			
	-2lnL	k	c	AIC	-2lnL	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

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 Equation 23 below (Cleves et al., 2008):

$$ln(d_i) = \beta_0 + x_i\beta + \mu$$

Equation 23

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

For the I-64 dataset, Table 13 shows the estimated coefficients of the explanatory variables, the standard error, the P-value, and percentage change (%) for the log-normal AFT model. The percentage change represents the change in the duration of the incident resulting from a one unit change in the value of the variable under consideration.

Table 13. 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		

Similarly, Table 14 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).

Table 14. 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		

As can be seen, the only variable with negative percentage change (%) is the variable "Afternoon" (1 PM-3PM), which shows that if the accident were to happen during this time interval, the duration would be 15% shorter, most probably because of lighter traffic during that time period. 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%).

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 previously described. Figure 11 shows the M5P-HBDM built for the I-64 or the Virginia training dataset.

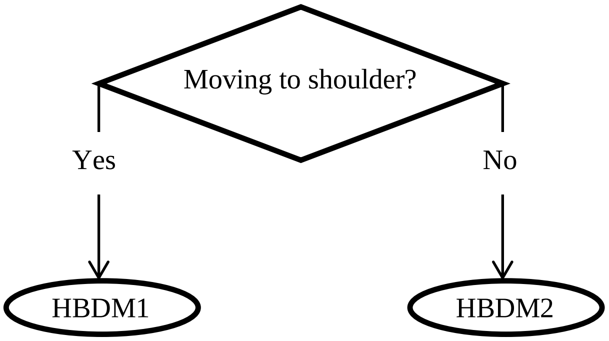


Figure 11. M5P-HBDM model for I-64 training dataset.

The M5P-HBDM model ended up having 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 15 shows the relevant parameters for the two models.

Table 15. 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		

As can be seen, for the log-normal AFT model HBDM1, no significant variables are found; only the constant β_0 and the sigma in the log-normal distribution are estimated. For the HBDM2, 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). 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 Figure 12. M5P-HBDM model for I-190 training dataset.

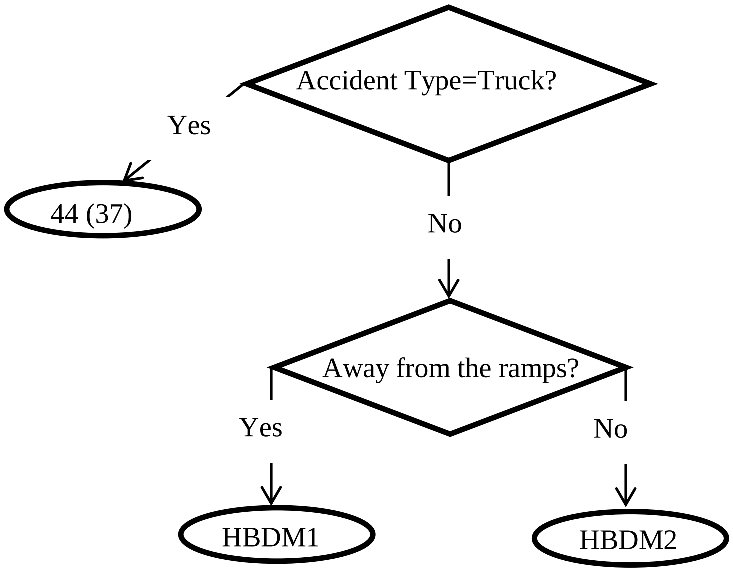


Figure 12. M5P-HBDM model for I-190 training dataset.

For both HBDM1 and HBDM2, the AIC test still shows that the log-normal distribution appears to be the best assumption for the AFT functions. The relevant parameters of HBDM1 and HBDM2 are shown in Table 16.

Table 16. 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		

From Table 16, 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). This makes sense since it is definitely harder to clear an incident during heavy traffic. Regarding the HBDM2 based on the 360 cases, the results show, for example, 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.

MODEL COMPARISON

Significant Independent Variables Comparison

Table 17. 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 (+)

Table 17 lists all the significant variables identified by each of the M5P, HBDM and M5P-HBDM models, for the I-64 training dataset (the sign in the parenthesis indicates the impact of that variable in terms of increasing or decreasing accident duration. The symbol “R” indicates that the variable resulted in a splitting rule for the model, as a part of the M5P algorithm.

As can be seen from Table 17, 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 18. 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.

Table 18. 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

Accident Duration Prediction Comparison

The prediction accuracy of the three models was compared, using a test set not previously utilized in model development. 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:

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

Equation 24

where A_i is the i^{th} actual value, P_i is the i^{th} predicted value.

To calculate the predictions, for the M5P tree model, each testing record will be directed toward the corresponding leaf, and the linear functions, or the mean target values at that leaf, are 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). 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 leaf, we use the median value of the cases at that node as the prediction.

Table 19 shows the MAPEs of the M5P tree model, HBDM model and the M5P-HBDM model for the two testing datasets. The column labelled “HBDM (median)” lists the HBDM’s MAPE resulting from using the median values of the survival times, whereas the column entitled “HBDM (mean)” lists the model’s MAPE resulting from using the mean values. The same is true for the columns entitled M5P-HBDM (median) and M5P-HBDM (mean) in connection with the M5P-HBDM.

Table 19. MAPEs of M5P tree, HBDM model and M5P-HBDM model

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%

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.

CONCLUSIONS AND FUTURE WORK

This study has proposed a novel approach for accident duration prediction, which constructs a M5P-HBDM model in which the leaves of the M5P tree model are HBDMs instead of linear regression models. Two traffic accident duration datasets were then used to construct and evaluate the performance of three modeling approaches, 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.
2. Because M5P-HBDM can build an AFT model as its leaf, 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.

REFERENCES

Abdel-Aty, M., Pande, A., Das, A., & Knibbe, W. J., 2008. Assessing safety on Dutch freeways with data from infrastructure-based intelligent transportation systems. *Transportation Research Record: Journal of the Transportation Research Board* 2083(1), 153-161.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F. M., & Hsia L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. *Transportation Research Record: Journal of the Transportation Research Board* 1897(1), 88-95.

Agrawal, R., T. Imieliński, and A. Swami, 1993. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, Vol. 22, No. 2, pp. 207-216.

Ahmed, M. M., & Abdel-Aty, M., 2012. The viability of using automatic vehicle identification data for real-time crash prediction. *IEEE Transactions on Intelligent Transportation Systems*, 13(2), 459-468.

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. *Transportation Research Record: Journal of the Transportation Research Board*, (2229), 46-54.

Anastasopoulos, P. C., & Mannering, F. L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention*, 41(1), 153-159.

Anderson, T. K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis & Prevention*, 41(3), 359-364.

Archer, K. J., & Kimes, R. V., 2008. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* 52(4), 2249-2260.

Arenas, A., J. Duch, A. Fernández, and S. Gómez, 2007. Size reduction of complex networks preserving modularity. *New Journal of Physics*, Vol. 9, No. 6, pp. 176-190.

Bastian M., S. Heymann, and M. Jacomy, 2014. *Gephi: an open source software for exploring and manipulating networks*. www.aaai.org/ocs/index.php/ICWSM/09/paper/viewFile/154Forum/1009. Accessed Feb. 18, 2014.

Bayesia, S. A. S., 2013. *BayesiaLab 5.1. The technology of Bayesian networks at your service.*

Bíl, M., R. Andrášik, and Z. Janoška, 2013. Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. *Accident Analysis & Prevention*, Vol. 55, pp. 265-273.

Blondel, V. D., J. L. Guillaume, R. Lambiotte, and E. Lefebvre, 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, Vo. 10, P10008.

Breiman, L., 2001. Random forests. *Machine learning* 45(1), 5-32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A., 1984. *Classification and regression trees*. CRC press.

Chen, W. H., and P. P. Jovanis, 2000. Method for identifying factors contributing to driver-injury severity in traffic crashes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1717*,, pp. 1-9.

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. *United States Department of Energy*.

Chung, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis & Prevention*, 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.

de Oña, J., G. López, R. Mujalli, and F. J. Calvo, 2013. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention* Vol. 51, pp. 1-10.

Depaire, B., Wets, G., & Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention*, 40(4), 1257-1266.

Efron, B., & Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438), 548-560.

Fan, J., Han, F., & Liu, H., 2013. Challenges of Big Data Analysis. *arXiv preprint*. arXiv:1308.1479.

Farradyne, P. B., 2000. *Traffic incident management handbook*. Prepared for Federal Highway Administration, Office of Travel Management.

Fernández, A., Gómez, Á., Lecumberry, F., Pardo, Á., & Ramírez, I., 2015. Pattern Recognition in Latin America in the "Big Data" Era. Pattern Recognition. *Pattern Recognition*, Vol. 48(4), p. 1185-1196,

Ferrara, E. A large-scale community structure analysis in Facebook. *EPJ Data Science*, Vol. 1, No. 1, 2012, pp. 1-30.

Fortunato, S. Community detection in graphs. *Physics Reports,* Vol. 486, No. 3, 2010, pp. 75-174.

Garib, A., Radwan, A. E., & Al-Deek, H., 1997. Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering*, 123(6), 459-466.

Geurts, K., G. Wets, T. Brijs, and K. Vanhoof, 2003. Profiling of high-frequency accident locations by use of association rules. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* pp. 123-130.

Ghosh, I, 2012. Examination of the factors influencing the clearance time of freeway incidents. *Journal of Transportation Systems Engineering and Information Technology*, Vol. 12, No. 3, pp. 75-89.

Giuliano, G., 1989. Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research Part A: General*, 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. *Accident Analysis & Prevention*, 19(5), 375-395.

Gregoriades, A., and K. C. Mouskos, 2013. Black spots identification through a Bayesian Networks quantification of accident risk index. *Transportation Research Part C: Emerging Technologies*, Vol. 28, pp. 28-43.

Gregorutti, B., Michel, B., & Saint-Pierre, P., 2013. Correlation and variable importance in random forests. *arXiv preprint* arXiv:1310.5726.

Han, J., Kamber, M., & Pei, J., 2006. *Data mining: concepts and techniques*. Morgan kaufmann.

Han, J., Pei, J., Yin, Y., & Mao, R., 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery* 8(1), 53-87.

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* (pp. 287-305). Springer New York.

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. In *Transportation Research Board 93rd Annual Meeting* (No. 14-2418).

Hossain, M., & Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. *Accident Analysis & Prevention* 45, 373-381.

Hung, M. C., & Yang, D. L., 2001. An efficient fuzzy c-means clustering algorithm. In *Proceedings of the IEEE International Conference on Data Mining*, 225-232.

Jain, A. K., 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31(8), 651-666.

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. *Accident Analysis & Prevention*, 30(4), 425-433.

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

Kotsiantis, S., & Kanellopoulos, D., 2006. Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering* 32(1), 47-58.

Lee, C., and M. Abdel-Aty. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. *Accident Analysis & Prevention*, Vol. 37, No. 4, 2005, pp. 775-786.

Lee, C., Hellinga, B., & Saccomanno, F., 2003. Real-time crash prediction model for application to crash prevention in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board* 1840(1), 67-77.

Lee, Y., & Wei, C. H., 2010. A computerized feature selection method using genetic algorithms to forecast freeway accident duration times. Computer ‐ Aided Civil and Infrastructure Engineering, 25(2), 132-148.

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

Liaw, A., & Wiener, M., 2002. Classification and Regression by Random Forest. *R News* 2(3), 18-22.

Lin, L., Wang, Q., & Sadek, A. W., 2013. Short-Term Forecasting of Traffic Volume: Evaluating Models Based on Multiple Data Sets and Data Diagnosis Measures. *Transportation Research Record: Journal of the Transportation Research Board* 2392(1), 40-47.

Lin, L., Wang, Q., & Sadek, A.W., 2014. Data Mining and Complex Network Algorithms for Traffic Accident Analysis. *Transportation Research Record: Journal of the Transportation Research Board* 2460(1), 128-136.

Lin, L., Wang, Q. and Sadek, A.W., 2016. A Combined M5P Tree and Hazard-based Duration Model for Predicting Urban Freeway Traffic Accident Durations. *Accident Analysis and Prevention*, Volume 91, Pages 114–126.

Lin, L., Wang, Q. and Sadek, A.W., 2015. A Novel Variable Selection Method based on Frequent Pattern Tree for Real-time Traffic Accident Risk Prediction. *Transportation Research – Part C*, Vol. 55, pp. 444–459.

Lord, D., & Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291-305.

Lv, Y., Tang, S., & Zhao, H., 2009. Real-Time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method. In *IEEE International Conference on Measuring Technology and Mechatronics Automation*, 3, 547-550.

Metagenomics Statistics, 2014. <http://dinsdalelab.sdsu.edu/metag.stats/index.html>.

Miaou, S. P., Song, J. J., & Mallick, B. K., 2003. Roadway traffic crash mapping: a space-time modeling approach. *Journal of Transportation and Statistics*, 6, 33-58.

Milton, J. C., Shankar, V. N., & Mannering, F. L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis & Prevention* 40(1), 260-266.

Mohamed, M. G., N. Saunier, L. F. Miranda-Moreno, and S. V. Ukkusuri. A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. *Safety Science*, Vol. 54, 2013, pp. 27-37.

Murphy, K. P., 2012. *Machine learning: a probabilistic perspective*. MIT Press.

Nam, D., & Mannering, F., 2000. An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A: Policy and Practice*, 34(2), 85-102.

Netica tutorial, 2014. <https://norsys.com/netica.html>.

Newman, M. E., and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, Vol. 69, No. 2, 2004, 15 pages.

Oh, C., Oh, J. S., & Ritchie, S. G., 2005. Real-time hazardous traffic condition warning system: framework and evaluation. *IEEE Transactions on Intelligent Transportation Systems*, 6(3), 265-272.

Okabe, A., T. Satoh, and K. Sugihara, 2009. A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, Vol. 23, No. 1, pp. 7-32.

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. *Accident Analysis & Prevention,* 38(3), 542-555.

Pande, A., & Abdel-Aty, M. 2006. Assessment of freeway traffic parameters leading to lane-change related collisions. *Accident Analysis & Prevention* 38(5), 936-948.

Quinlan, J. R., 1992. Learning with continuous classes. In *5th Australian joint conference on artificial intelligence,* Vol. 92, pp. 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. *Accident Analysis & Prevention,* 43(5), 1666-1676.

Sawalha, Z., & Sayed, T., 2006. Traffic accident modeling: some statistical issues. Canadian *Journal of Civil Engineering* 33(9), 1115-1124.

Smith, K., & Smith, B., 2001. *Forecasting the clearance time of freeway accidents*. Center for Transportation Studies, University of Virginia.

Valent, F., F. Schiava, C. Savonitto, T. Gallo, S. Brusaferro, and F. Barbone. Risk factors for fatal road traffic accidents in Udine, Italy, 2002. *Accident Analysis & Prevention,* Vol. 34, No. 1, pp. 71-84.

Valenti, G., Lelli, M., & Cucina, D., 2010. A comparative study of models for the incident duration prediction. *European Transport Research Review,* 2(2), 103-111.

Vaupel, J.W. and Yashin, A.I., 1985. Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician,* 39(3), pp.176-185.

Xi, J., Z. Gao, S. Niu, T. Ding, and G. Ning, 2013. A Hybrid Algorithm of Traffic Accident Data Mining on Cause Analysis. *Mathematical Problems in Engineering,* Vol. 2013, 8 pages.

Xie, Z., and J. Yan. Kernel density estimation of traffic accidents in a network space, 2008. Computers, *Environment and Urban Systems,* Vol. 32, No. 5, pp 396-406.

Xu, C., Tarko, A. P., Wang, W., & Liu, P., 2013. Predicting crash likelihood and severity on freeways with real-time loop detector data. *Accident Analysis & Prevention* 57, 30-39.

US census bureau, 2013. http://www.census.gov/compendia/statab/2012/tables/12s1103.pdf.

Wang, Y., & Witten, I. H., 1997. Inducing model trees for continuous classes. In *Proceedings of the Ninth European Conference on Machine Learning* (pp. 128-137).

Wei, C. H., & Lee, Y., 2007. Sequential forecast of incident duration using Artificial Neural Network models. *Accident Analysis & Prevention,* 39(5), 944-954.

Ye, Y., Wu, Q., Zhexue Huang, J., Ng, M. K., & Li, X., 2013. Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition* 46(3), 769-787.

Yu, R., & Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. *Accident Analysis & Prevention* 51, 252-259.

Zhan, C., Gan, A., & Hadi, M., 2011. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), 1549-1557.

Zheng, Z., Ahn, S., & Monsere, C. M., 2010. Impact of traffic oscillations on freeway crash occurrences. *Accident Analysis & Prevention* 42(2), 626-636.