RC25175 (W1106-070) June 21, 2011
Mathematics

IBM Research Report

Incident Duration Prediction with Hybrid Tree-based Quantile Regression

Qing He, Yiannis Kamarianakis, Klayut Jintanakul, Laura Wynter
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Incident Duration Prediction with Hybrid Tree-based Quantile Regression

Qing He, Yiannis Kamarianakis, Klayut Jintanakul and Laura Wynter*

IBM T. J. Watson Research Center

P.O. Box 218, Yorktown Heights, NY 10598, USA

E-mail: {qhe, yiannis, kjintan, lwynter}@us.ibm.com

Abstract

Accurate prediction of incident duration is critical for efficient incident management which aims to minimize the impact of non-recurrent congestion. In this chapter, a hybrid tree-based quantile regression method is proposed for incident duration prediction and quantification of the effects of various incident and traffic characteristics that determine duration. Hybrid tree-based quantile regression incorporates the merits of both quantile regression modeling and tree-structured modeling: robustness to outliers, simple interpretation, flexibility in combining categorical covariates and capturing nonlinear associations. The predictive models presented here are based on variables associated with incident characteristics as well as the traffic conditions before and after incident occurrence. Compared to previous approaches, the hybrid tree-based quantile regression offers higher predictive accuracy.

* Corresponding author, e-mail: lwynter@us.ibm.com, Tel: +1 914 945 1708, fax: +1 914 945 3434.

1. Introduction

Incidents, including accidents, vehicle breakdowns, spilled loads, or other random events, reduce the capacity of the road and cause congestion when traffic demand exceeds the reduced capacity at the incident location. Oak Ridge National Laboratory estimates that 55% of motorist delays on freeways are incident related (Chin et al. 2004). Effective management is essential for mitigating the negative effects of incidents on congested urban freeways. Various studies have been undertaken to develop mitigation measures that minimize non-recurrent congestion due to freeway incidents. A typical example of such efforts is the development of various types of incident management systems that aim to clear traffic incidents quickly to minimize its impact on traffic flow.

In existing incident management systems, an ability to anticipate incident characteristics allows traffic managers to make better decisions on how to use management and control resources, such as Advanced Traveler Information System (ATIS) and Route Guidance Systems (RGS). Incident duration is an essential characteristic since it highly determines both the magnitude and the extent of congestion. Therefore, it is important to understand which factors can affect the incident duration. This study explores these critical factors and develops statistical models for incident duration prediction.

Incident duration can be defined as the duration between the instances of incident occurrence and of departure of the response vehicles from the accident scene (Garib et al. 1997; Nam & Mannering 2000; Smith & Smith 2001). As indicated in previous studies, an incident is composed of the following four phases: (a) incident detection and reporting time, (b) response time, (c) clearance time, and (d) recovery time. Traditionally incident duration is defined as the sum of first three phases.

Incident duration prediction models can be used as a means to improve incident management systems under non-recurrent traffic congestion. Incident management systems generally encompass three main modules, including incident detection technology, incident impact prediction, and incident-responsive traffic management and control. Incident duration prediction models are essential components in such a system, especially in the last two modules. Travelers and traffic management entities can generally realize the impact by the forecasted incident duration. In general, the impact of an incident in terms of both magnitude and extent of the congestion is significantly affected by incident duration. Virtually all existing impact prediction models developed in the literature require knowing the incident duration before producing a prediction. Since duration of an incident is usually not known until the incident is cleared, an accurate estimate is needed for accurate real-time prediction of incident impacts.

Likewise, an accurate estimate of the incident duration is also required in deriving effective response management and control strategies. Effective traffic control strategies are supposed to alleviate impacted traffic without unnecessarily interrupting the normal traffic or creating a secondary bottleneck. Rerouting factors such as diversion and merge points, diversion percentages, and diversion duration need to be derived on the basis of accurate magnitude and extent of the congestion as well as the duration of congestion (Lee et al. 2003; Srinivasan & Krishnamurthy 2003). In other words, the ability to redistribute flows over time is important for effective incident management (Oh & Jayakrishnan 2000). For instance, the projected incident duration will enable responsible traffic agencies to notify the en-route drivers of traffic congestion in a timely manner with VMS, and assess if any detour operators or control actions are needed. Drivers with better traffic information when encountering an incident can then make a proper route choice decision with less anxiety, which may consequently increase their compliance to suggestions or guidance by responsible traffic agencies (Garib et al. 1997).

The vast majority of previous related studies focused on predicting incident duration solely from incident characteristics. This work incorporates not only incident characteristics but also traffic data from both before and after the incident occurrence. Traffic data collected prior to incidents act as spatial and temporal indicators. Spatially, sequential traffic measurements indicate if the location of an incident is a bottleneck in the network. Temporally, time of day and day of week are associated with different levels of traffic variables and consequently, with different effects for incidents of the same type. Furthermore, the levels of traffic variables after an incident are associated with incident severity and hence with clearance times.

In addition, this work uses hybrid tree-based quantile regression. Other tree approaches have been used in the literature previously, including (Ozbay & Kachroo 1999; Smith & Smith 2001). A critical feature of the method used here is that it is designed to overcome the fundamental problems of previous trees such over-fitting and selection bias towards predictors with many possible splits or missing values.

This chapter is organized as follows. Section 2 is devoted to a literature review which discusses different methods for online prediction of incident duration. In section 3, we present the methodology of hybrid tree-based quantile regression, which combines conditional inference trees with quantile regression. Data description and preliminary data analysis are illustrated in section 4. Section 5 describes the calibration of the statistical models, which is followed by an evaluation of their predictive accuracy. Finally, Section 6 presents some concluding remarks.

2 Literature review

Incident duration is one of the essential characteristics of incidents that determine the magnitude and extent of the congestion. Thus, it has been extensively studied over the last few decades. Different approaches proposed in the literature can be grouped into the following categories:

- Linear regression: Garib et al. (1997) performed duration prediction using regression models, to provide real-time incident information to travelers. As the empirical distribution of incident duration is skewed (Golob et al. 1987, Giuliano 1989), linear models are based on its logarithmic transformation. Khattak et al. (1994) used a series of truncated regression models to predict incident duration, which account for the fact that incident information at a Traffic Operations Center is obtained sequentially.

- Tree models: Ozbay & Kachroo (1998) constructed decision trees which do not require knowledge of all observable incident characteristics. A similar approach was followed in Smith & Smith (2001) where classification trees were applied to predict incident duration, defined as a categorical variable. The classification tree was shown to be well suited for forecasting the phases of incident duration with reliable and informative characteristics. Recently, Kim et al. (2008) constructed a rule-based tree model coupled with a discrete choice model, aiming at improved predictive ability. However, these models do not use any traffic data nor do they consider the tails of the distributions.

- K-nearest neighbor (KNN): Smith & Smith (2001) investigated incident duration prediction with KNN methods. Qi & Smith (2004) developed a distance metric that can be effectively used with categorical data. They argued that KNN outperformed parametric forecasting models significantly. Again, these methods did not leverage traffic data.
- Survival analysis: Incident duration can be viewed as the time period an incident can survive before being cleared. However, to implement survival analysis, selecting an appropriate probability distribution for incidence duration can be a challenging task (Jones et al. 1991; Nam & Mannering 2000; Qi & Teng 2008; Chung 2010; Chung et al. 2010).
- Artificial intelligence: Wei & Lee (2007) applied Artificial Neural Network (ANN) based models and data fusion techniques to forecast incident duration. Recently they employed Genetic Algorithms (GA) and ANNs to construct two models that forecast accident duration from the moment of accident notification to accident clearance (Lee & Wei 2010). Demiroluk & Ozbay (2011) developed three structure learning algorithms to construct Bayesian Network (BN) structures. They demonstrated that BNs were very useful in uncovering important relationships among predictors, using the concept of strength of links.

3. Methodology

The adopted methodology, proposed recently in Hothorn et al. (2006), combines unbiased recursive partitioning (URP) with piecewise constant fitting using permutation tests. The conditional distribution of statistics measuring the association between incident duration and its predictors is the basis for an unbiased selection of the predictors in the model. Multiple tests are applied to determine whether no significant association between any predictor and duration can be stated and the recursion needs to stop. The above framework aims to solve both the over-fitting and the variable selection problems of older recursive partitioning methods; (a detailed overview is provided in Murthy, 1998).

Our implementation was based on the software provided by the developers of the method (Hothorn et al., 2011). Significance levels for the test statistics were set to conventional levels (0.05) and a Bonferroni correction was applied in multiple testing procedures, in accordance with the suggestion in Hothorn et al. (2006).

Predictions from conventional tree models are compared to the ones derived from a hybrid approach that combines regression trees based on the incident characteristics with quantile regression models that use traffic variables as predictors. The latter are robust to outliers and skewed response distributions (Koenker, 2005), and are widely used in applications instead of conventional least-squares regression during the last decade. It is worth noting that our hybrid method is similar to the one adopted in GUIDE (Loh, 2008).

In Section 5 we display predictive models for the 0.5 (median regression) and the 0.9 quantiles of log-duration. Median regression models can be used as conventional incident duration predictors, while models for the 0.9 conditional quantile quantify the uncertainty associated with each prediction and can also be viewed as predictors of worst-case scenarios.

Finally, URP (prediction from only tree models) and hybrid tree-based quantile regression models are compared with the well-known older approach known as Classification and Regression Tree[†] (CART)

[†] CART is implemented in R (R Development Core Team 2009), using rpart (Therneau & Atkinson 2011).

(Breiman et al. 1984) as well as with the classic K-nearest-neighbor (KNN) methods without using traffic data as predictors, as was done in previous work found in the literature. Overall prediction accuracy is measured by mean absolute error (MAE1), median absolute error (MAE2), mean absolute percentage error (MAPE1) and median absolute percentage error (MAPE2). We also present percentages of predictions that are within a certain tolerance of their actual duration times, as suggested by Smith & Smith (2001).

4. Data Description

We examine incidents that occurred in 17 major freeways in Bay area, California, from April to June, 2010. The freeway network, shown in Figure 1, connects ten cities. Incident data were obtained from the California Highway Patrol computer aided dispatch (CHP/CAD) system (CHP 2011). Incident information was collected from two sources: the first source provided the incident type and the corresponding spatio-temporal information, while the second source provided further details on incident characteristics, such as number of vehicles involved[‡]. Original incident types were classified into three groups: collision, disabled vehicle and traffic hazard. In total, 1245 incidents with valid data were analyzed. Table 1 contains the basic summary statistics of the dataset. The empirical probability distribution of incident duration has a long tail, which is in accordance with observations from previous studies (Chung, 2007). The average incident duration is 20.61 minutes, while the median incident duration is 15.5 minutes. The set of incidents was randomly cut into a training dataset (60 percent of data) and test dataset (40 percent of data).

Figure 1 Bay area freeway network with detectors in highlighted links

---

[‡] Incidents associated with scheduled road closures or without any log were excluded from the analysis. Duplicated incidents were identified by incident reporting time and location and were excluded as well while their logs were reviewed and merged. An automatic text recognition program was developed to parse incident logs.

Traffic data were obtained from the Caltrans Performance Measurement System (PeMS). PeMS is a system designed to maintain California freeway traffic data and compute annual congestion for facilities with surveillance systems in place, typically loop detectors spaced approximately 0.5 mile apart on each freeway lane (Choe et al. 2002). There are around 850 detectors in Bay area freeways, shown as highlighted links in Figure 1. The analysis that follows uses 5-minute aggregated volume, speed, and occupancy data. Each incident was associated with traffic data spatially and temporally:

- Spatially, each incident was matched with the closest link, which satisfied the incident location descriptions. Upstream and downstream traffic detectors were also identified accordingly.
- Temporally, a modified incident detection algorithm based on the DELOS (also called Minnesota) algorithm (Chassiakos & Stephanedes 1993) was developed to trace differences in occupancy between adjacent detectors through time, and to detect an incident when these differences change significantly in a short time period. This incident detection algorithm associates incident data with upstream and downstream traffic data, locates the time stamp when the shockwave hits the nearest upstream detector, and records traffic data before and after the incident's time of occurrence.

Table 1 Summary statistics

Incident data	
Number of incidents	1245
Median incident duration (min)	15.5
Average incident duration (min)	20.61
Proportion of incidents in "Collision"	0.52
Proportion of incidents in "Disabled"	0.26
Proportion of incidents in "Hazard"	0.22
Proportion of incidents with injuries	0.08
Average number of vehicle involved	1.30
Traffic data	
Average historical speed across all incident sites (mph)	52.84
Average historical volume across all incident sites (veh/hr/ln)	1361
Average historical occupancy across all incident sites	0.139
Average speed before incident (mph)	48.75
Average speed after incident (mph)	29.27
Average volume before incident (veh/hr/ln)	1303
Average volume after incident (veh/hr/ln)	1300
Average occupancy before incident	0.147
Average occupancy after incident	0.316

Table 1 depicts summary statistics of traffic data before and after the incident. Prior knowledge suggests that incidents will cause congestion on an upstream detector whereas traffic conditions will become less congested at downstream stations (Payne & Tignor, 1978). In our study, it is found that speed and occupancy are affected dramatically by incidents, while volume remains relatively stable. On average, speed drops 40% after an incident, while occupancy increases by 115%. The impact of incidents on traffic data is illustrated in Figure 2. Speeds, volumes and occupancies at the first upstream detector before and after an incident's time of occurrence are normalized and plotted in the same graph. Points on the 45 degree line correspond to data that are not affected by an incident. One notes that speeds tend to decrease while occupancies tend to increase after the incident, in accordance with prior expectations. On the other hand, volumes may increase or decrease, depending on the levels of traffic congestion before and after the incident.

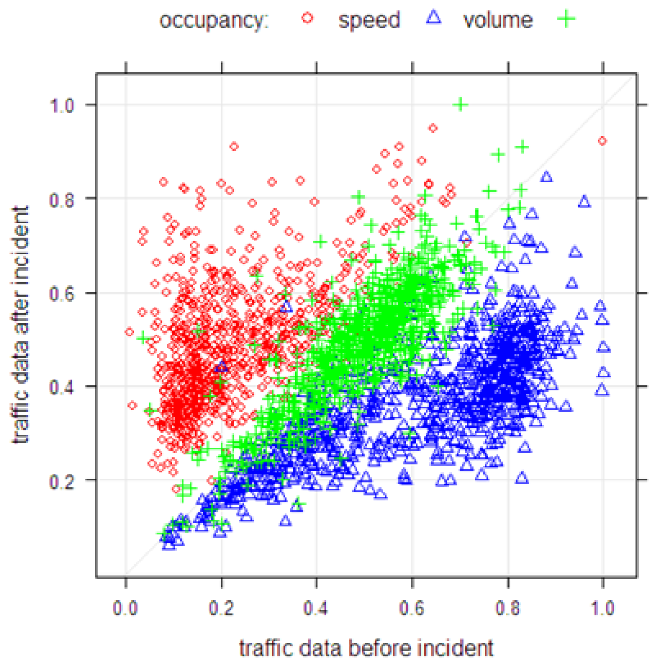

Figure 2 Scatter-plots of normalized traffic data (speed, volume and occupancy) at the first upstream detector before and after incident occurrence.

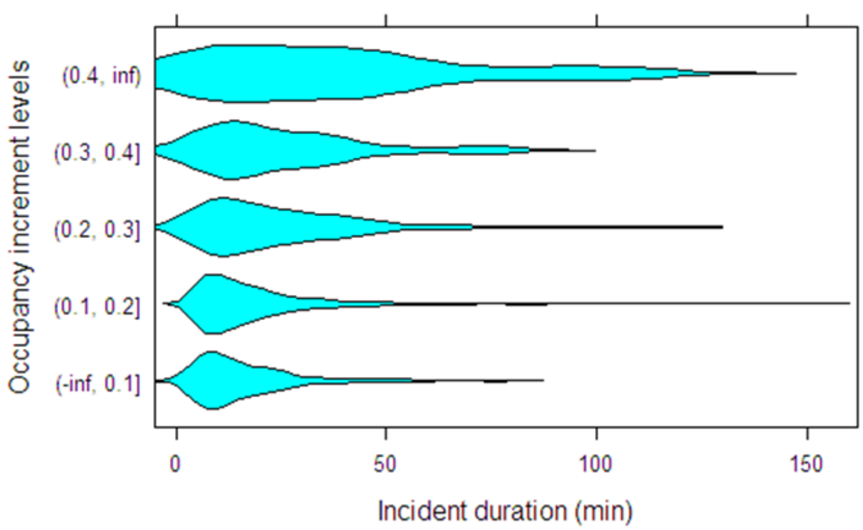

Figure 3 Empirical distributions of incident duration for different occupancy increment levels after incident detection at the first upstream detector.

The candidate predictor variables are displayed in Table 2. All incident-related variables are categorical except for the number of vehicles involved in the incident. Figure 3 depicts a heteroscedastic relationship between incident duration and occupancy range. The latter is measured by occupancy differences, i.e., $Occ(s,t) - Occ(s,t\text{-}1)$, where section s indicates the first upstream detector, and t is the time when the incident-induced impact is observed, with $t\text{-}1$ being the preceding time period to t. Each violin-type plot represents the empirical probability density of incident duration at different ranges of occupancy. Clearly, the variability of incident duration increases as occupancy increment increases. Low occupancy ranges are associated with short incident durations, while high occupancy increments may be related to both short and long incident durations. This suggests that traffic data may provide significant predictive power for incident duration. An increasing relationship between incident duration and the number of vehicles involved in an incident can be observed in Figure 4.

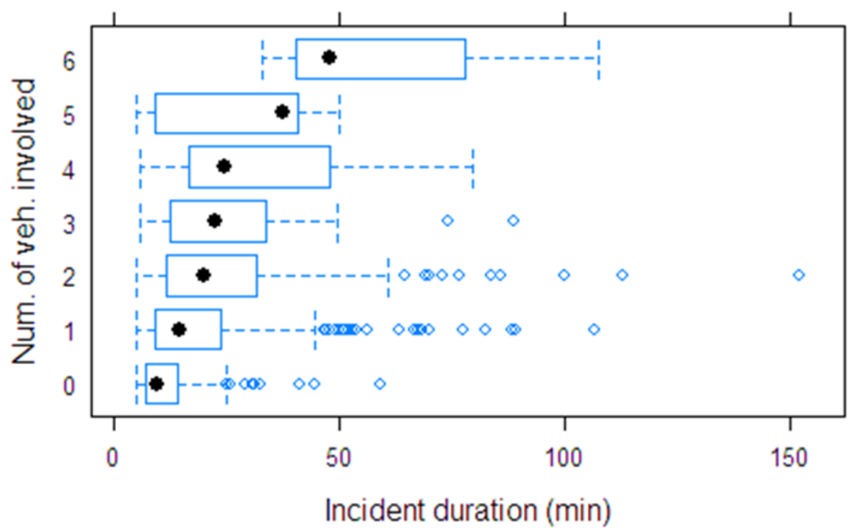

Figure 4 Box-plots of incident duration for different numbers of vehicles involved.

Table 2 Candidate independent variables

Information type	Independent variables	Notation
Weather characteristics	Rainy	rain
	Snowy	snow
Temporal characteristics	Time of day (AM, PM, Mid, Off-peak)	t_am, t_pm, t_mid, t_off
	Day of week (Weekday or not)	weekday
Incident characteristics	Incident type (collision, disabled or hazard)	type
	Num of vehicles involved	num_veh
	Lanes blocked (binary)	lane_block
	Truck involved (binary)	truck
	Person injured (binary)	injured
	CHP officer assigned (binary)	CHP
Geometric characteristics	Freeway (CA-17, CA-237, CA-24, CA-242, CA-4, CA-84, CA-85, CA-87, CA-92, I-238, I-280, I-580, I-680, I-80, I-880, I-980, US-101)	freeway1~freeway17
	City (Castro Valley, Contra Costa, Dublin, Hayward, Marin, Oakland, Redwood City, San Francisco, San Jose, Solano)	city1~city10

	Interstate highway	interstate
	Ramp exists near incident location (upstream/downstream on-ramp/off-ramp; binary)	uponramp,upofframp, downonramp, downofframp
	Upstream off-ramp and a downstream on-ramp exist near incident location (binary)	junction
	Upstream on-ramp and/or downstream off-ramp exist near incident location (binary)	junctionbwt
	number of lanes (2 or 3, 4, 5+)	ln23, ln4, ln5
Traffic characteristics	Historical mean of traffic data (speed, volume and occupancy) at the time of incident	v_mean, q_mean, o_mean
	Traffic data at the first upstream detector before incident detection	v_prior, q_prior, o_prior
	Traffic data at the first upstream detector after incident detection	v_inc, q_inc, o_inc
	Traffic data after incident occurrence divided by measurements collected before incident occurrence	v_ratio, q_ratio, o_ratio
	Traffic data increments after incident occurrence	v_diff, q_diff, o_diff

Note: v_ratio = v_inc/v_prior; v_diff = v_inc – v_prior.

5. Model Estimation and Validation

Two URP trees were built based on different sets of predictors. The first one, called URP tree1, shown in Figure 5, was created using all candidate variables in Table 2. The second one (URP) was obtained using all but traffic variables and is depicted in Figure 6. The decision path of the tree model is followed by answering a yes or no question at each node. Eventually, at each terminal node, a prediction is made based on the mean of incident duration of the data in that category.

Specifically, URP tree2 is a subset of URP tree1 that does not contain traffic data variables. According to the p-values in each node in both URP tree1 and URP tree2, the most significant predictor variables are incident characteristics (type, injured, num_veh and lane_block). As shown below, URP tree1 turns out to yield improved prediction accuracy than URP tree2, demonstrating that the incorporation of traffic data provides increased predictive power to the model.

In both URP trees the first node separates incidents according to type. This finding is in accordance with earlier studies which suggest that the empirical distribution of incident duration depends significantly on incident type (Kim et al. 2008). In the case of traffic collisions, the second node divides incidents according to the presence or not of an injury. In the URP tree with traffic data, URP tree1, if both collision and injury occur, v_prior, the level of speed prior to an incident, divides the dataset further. Hence, collisions with injuries and high prior speeds (>48.8 mph), cause the longest incident durations in URP tree. High speeds prior to the incident are usually associated with off-peak periods; severe off-peak incidents are expected to last longer due to fewer available response units.

Node 3 indicates that CHP officer involvement reduces incident duration while node 4 indicates that a large reduction in traffic volume is associated with elevated incident duration. Node 5 uses historical occupancy to split incidents. Large values correspond to peak-periods, which tend to have short incident duration. Information from node 6 is consistent with the observations from Figure 3: a larger occupancy increment is associated with larger incident duration. Incidents with disabled vehicles (num_veh=1) have longer expected duration than traffic hazard (num_veh=0). Node 17 shows that information on blocked lanes is significant for duration prediction with disabled vehicles.

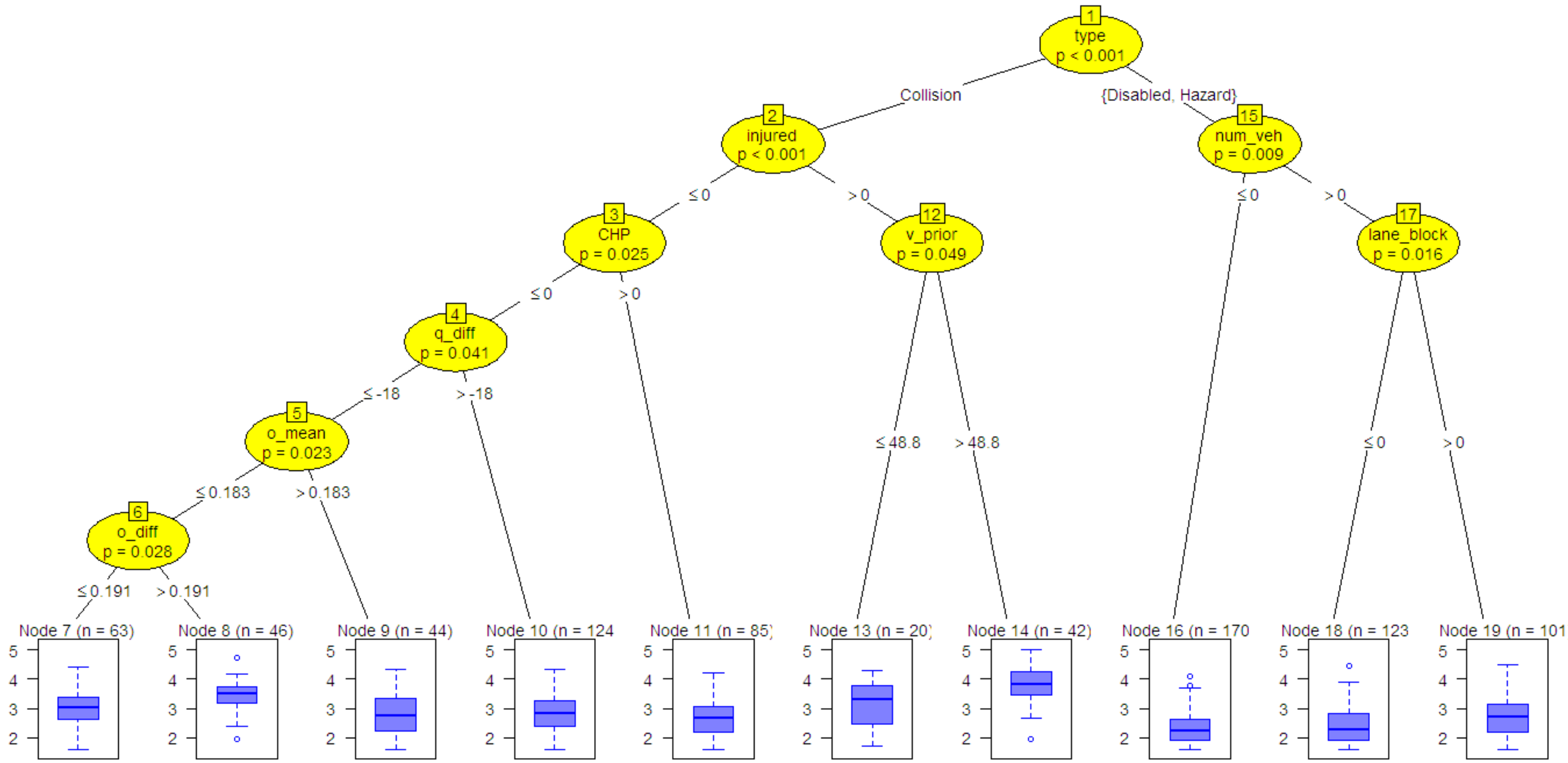

Figure 5 URP tree1 (with traffic data): Unbiased recursive partition tree using all candidate predictors in training data. For each inner node, the Bonferroni-adjusted p-values are given. A box-plot of the log of incident duration is displayed in each terminal node.

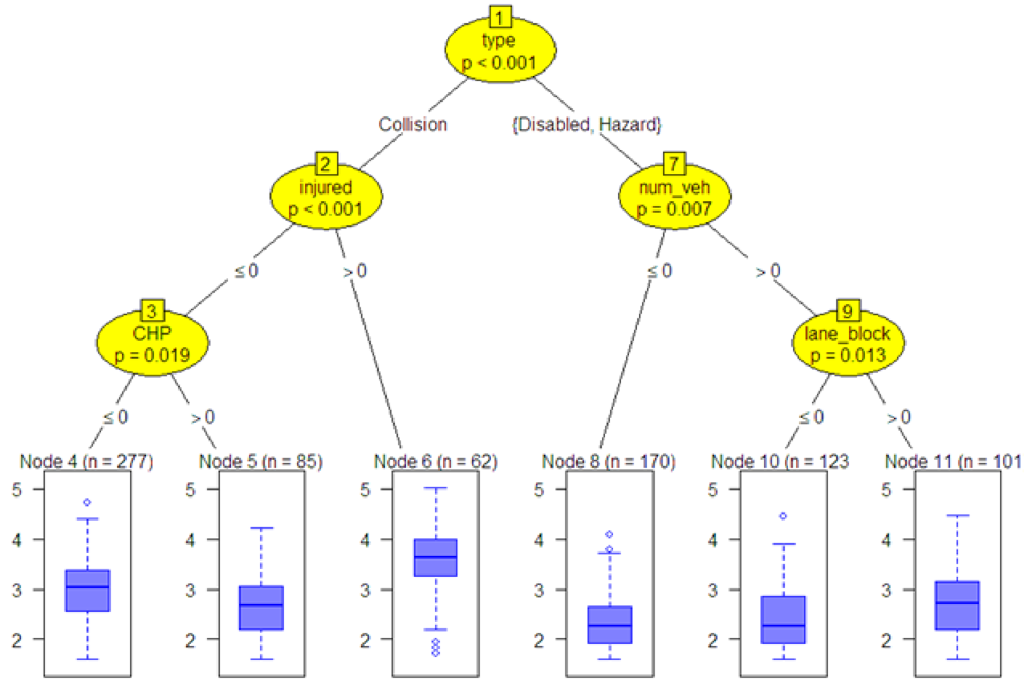

Figure 6 URP tree2 (without traffic data): Unbiased recursive partition tree using only categorical variables in training data. For each inner node, the Bonferroni-adjusted p-values are shown. A box-plot of the log of incident duration is displayed in each terminal node.

To gain better prediction accuracy, quantile regression models are built for each terminal node in URP tree2. Unlike least-squared regression trees, which concentrate on modeling the relationship between the response and the covariates at the centre of the response distribution, quantile regression can provide insight into the nature of that relationship at the centre as well as the tails of the response distribution. Table 3 shows the coefficients of the six estimated regression models for the 0.5 (median) and 0.9 quantiles of the logarithm of incident duration. Besides traffic characteristics, geometric characteristics appear in most of the estimated regression models, such as ramp, city and freeway junction information. This implies that incident duration varies significantly for different geometry factors, as well as different jurisdictions. For example, incidents that happen in freeway junctions are related to increased clearance times, while the presence of an upstream off-ramp may decrease incident duration.

By replacing the mean in the final nodes of tree2 by quantile regression models, the forecasting accuracy (measured by median absolute percentage error) on each terminal node was improved on average by 15%, as can be observed in Figure 7. To better visualize the difference between 0.5 and 0.9 quantile estimates, the corresponding predictions are plotted in Figure 8; the average ratio of 0.9 and 0.5 quantile estimates is 2.29.

Table 3 Coefficients of regression models estimated at each terminal node of URPtree2

0.5 quantile				**0.9 quantile**			
Regressor	**Value**	**Std. Error**	**t value**	**Regressor**	**Value**	**Std. Error**	**t value**
Node 4				**Node 4**			
constant	2.7325	0.1266	21.5818	constant	3.3950	0.1967	17.2611
city2	-0.2881	0.1692	-1.7025	interstate	-0.1924	0.1353	-1.4215
q_diff	-0.0007	0.0003	-2.3113	junctionbwt	-0.2772	0.1076	-2.5768
o_diff	1.8049	0.6023	2.9968	o_prior	1.3523	0.6291	2.1495
Node 5				o_diff	1.8274	0.7205	2.5364
constant	2.6672	0.0910	29.2972	Node 5			
Node 6				lane_block	0.3862	0.1603	2.4096
v_diff	0.0175	0.0144	1.2180	city2	-0.9949	0.4525	-2.1985
v_prior	0.0566	0.0074	7.6151	city6	-0.7597	0.3037	-2.5012
o_prior	6.6888	1.3426	4.9818	city7	-0.3904	0.2175	-1.7954
city8	-0.8139	0.3257	-2.4987	city9	-0.6949	0.1800	-3.8601
city2	0.5440	0.2358	2.3071	q_mean	0.0006	0.0003	1.8679
Node 8				v_prior	0.0441	0.0108	4.0738
constant	1.7954	0.1688	10.6352	o_prior	4.3970	1.2343	3.5624
junction	0.6095	0.1287	4.7376	v_diff	0.0631	0.0182	3.4759
uponramp	0.2327	0.1087	2.1402	o_diff	8.0788	2.2776	3.5470
upofframp	-0.4717	0.1057	-4.4644	Node 6			
v_prior	0.0060	0.0028	2.1330	interstate	0.4334	0.3388	1.2792
Node 10				o_diff	3.1075	1.5401	2.0177
constant	1.7715	0.2757	6.4263	v_prior	0.0426	0.0071	5.9916
t_mid	0.4259	0.1787	2.3832	o_prior	7.9935	1.7221	4.6418
t_off	0.6355	0.2892	2.1977	city8	-0.8302	0.3984	-2.0837
t_pm	0.2677	0.1373	1.9501	Node 8			
junction	0.2560	0.1296	1.9751	constant	2.8679	0.1357	21.1301
CHP	-0.3997	0.1244	-3.2133	junction	0.8737	0.2749	3.1787
downonramp	-0.2272	0.1599	-1.4208	junctionbwt	0.5531	0.2895	1.9109
city6	0.3466	0.2346	1.4772	upofframp	-1.1451	0.2178	-5.2583
q_mean	0.0002	0.0002	1.5193	downonramp	0.3267	0.1980	1.6497
o_diff	1.7876	1.0319	1.7324	Node 10			
Node 11				constant	3.7102	0.3994	9.2896
constant	1.6724	0.6992	2.3918	junction	1.2756	0.4000	3.1891
weekday	-0.5759	0.2481	-2.3209	upofframp	-1.1815	0.3823	-3.0906
interstate	-0.6921	0.2970	-2.3303	downonramp	-0.8234	0.3498	-2.3537
truck	0.4832	0.2702	1.7882	downofframp	-0.8556	0.4051	-2.1121
uponramp	-0.2953	0.1660	-1.7790	city6	0.3029	0.1863	1.6260
freeway12	0.3397	0.2504	1.3564	v_diff	0.0110	0.0057	1.9181
v_prior	0.0198	0.0087	2.2799	q_diff	-0.0007	0.0004	-2.0189
o_prior	4.8327	1.7361	2.7836	o_diff	2.4608	1.4465	1.7012

o_diff	2.5521	1.3578	1.8796	Node 11			
				constant	3.6738	0.1235	29.7529
				uponramp	-0.4549	0.1742	-2.6114

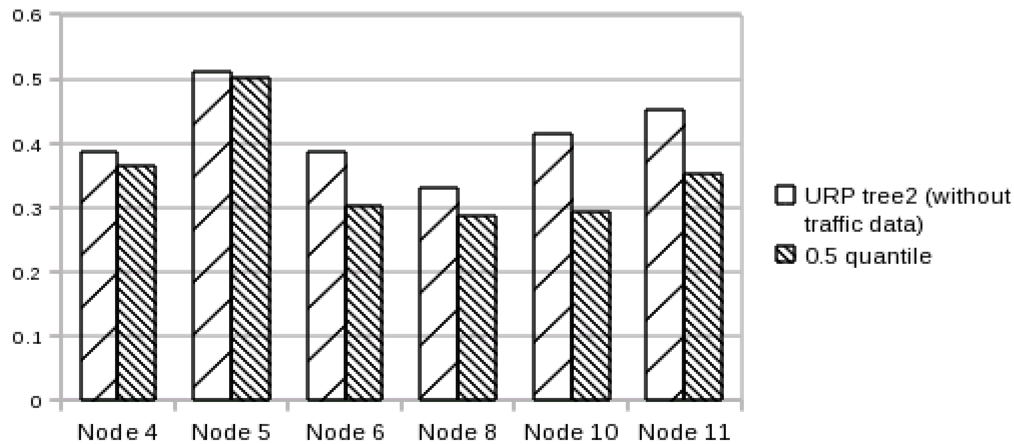

Figure 7 Comparisons of median absolute percentage error for URPtree2 and 0.5 quantile regression in each terminal node.

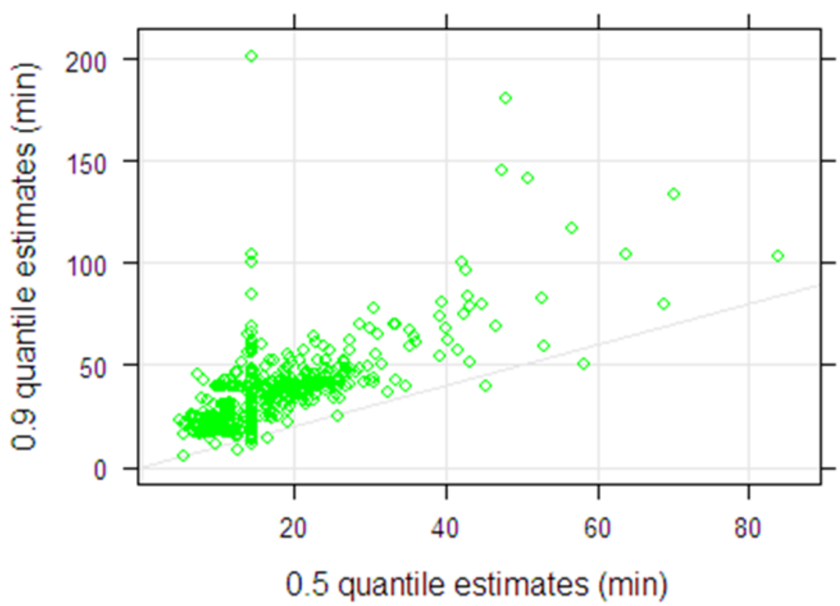

Figure 8 Comparisons of 0.5 and 0.9 quantile estimates

Finally, the proposed hybrid tree-based quantile regression model is compared with the well-known Classification and Regression Tree[4] (CART) (Breiman et al. 1984) and K-nearest-neighbor (KNN) methods that do not use traffic data, as was reported in earlier studies in the literature. Hence, we consider

[4] CART is implemented in R (R Development Core Team 2009), using rpart (Therneau & Atkinson 2011).

the CART and the KNN approaches as performed here to be benchmarks for this research. The tree model from CART is depicted in Figure 9. Again, the first nodes are incident type and the presence of an injury. KNN selects *k* past incidents that are closest to the current one, and takes the mean or median of incidents in the neighborhood. A similar KNN approach implemented for incident prediction was reported in previous studies (Qi & Smith 2004; Smith & Smith 2001). The predictors in URPtree2 (Figure 6) were used as the set of descriptors for each incident in KNN. The distance metric of Qi & Smith (2004) was adopted for measuring similarity between current and past incidents.

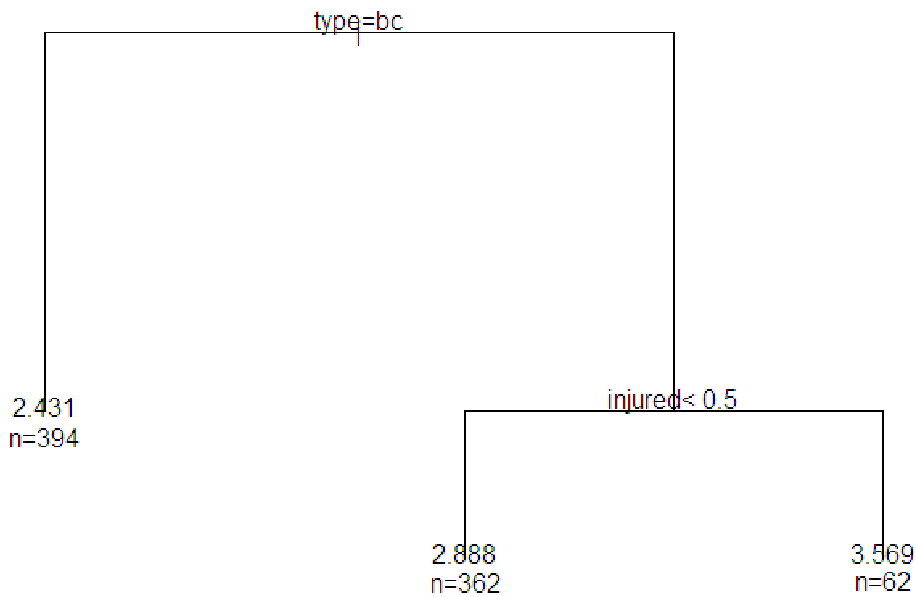

Figure 9 Regression tree from CART for the training data. The split is beneath each intermediate node. Types a,b and c represent collision, disabled vehicle and hazard, respectively. The number beneath each terminal node is the predicted logarithm of incident duration.

Table 4 reports measures of predictive accuracy for all examined methods. Overall prediction accuracy was measured by mean absolute error (denoted as MAE1), in minutes, median absolute error (denoted as MAE2), in minutes, mean absolute percentage error (denoted as MAPE1) and median absolute percentage error (denoted as MAPE2). As can be observed from the table, the URP tree approaches, and specifically the hybrid tree-based quantile regression, reduced error across the board as compared to the KNN and CART approaches used in the literature.

Table 4 Evaluation of predictive error with different methods

	KNN	CART	URP tree2 (without traffic data)	URP tree1 (with traffic data)	hybrid tree-based quantile reg.
MAE1 (min)	9.77	9.62	9.39	9.15	8.54
MAPE1 (%)	59.2%	57.1%	55.1%	53.2%	49.1%
MAE2 (min)	6.2	6.02	5.81	5.78	4.99

MAPE2 (%)	42.2%	42.2%	40.4%	39.8%	34.5%

An alternative measure of effectiveness is related to a certain tolerance of the prediction error. As suggested by Smith & Smith (2001), it is useful to know the percentage of predictions that are within a certain tolerance of their actual duration times. Table 5 reports accuracy in terms of tolerance levels. Five tolerance values were used: 5, 10, 15, 30, and 60 min.

Over 50% of incidents have been predicted with less than 5 min prediction error with hybrid tree-based quantile regression, while other published methods reached at most 44%. For the ranges of prediction error under 5min, 10min, and 15min, there were clear advantages to using the URP approaches proposed here. Note that for thresholds of 30min and 60min, the benefits of the proposed approach decrease. That is not surprising for this data set in that the average and median incident durations were 20 and 15 minutes, respectively. Hence relatively few incidents fall into the range of 30minutes or more, and the benefits of predicting better those durations is therefore not as visible. Nonetheless, the most important tolerance levels for the incidents in the dataset used in this study, namely the 5min, 10min, and 15min thresholds, all demonstrated significant improvements via the use of the URP techniques developed here.

Table 5 Comparisons of percentage of test samples in different prediction tolerances

	KNN	CART	URP tree2 (without traffic data)	URP tree1 (with traffic data)	hybrid tree-based quantile reg.
prediction error <= 5min	42.8%	42.6%	43.4%	43.96%	50.1%
prediction error <= 10min	69.1%	70.1%	71.2%	72.1%	72.3%
prediction error <= 15min	82.5%	82.4%	84.8%	83%	84.6%
prediction error <= 30min	94.2%	94.6%	94.7%	94.5%	94.3%
prediction error <= 60min	98.8%	99.2%	99%	99.2%	99.1%

6. Concluding Remarks

In this chapter, the use of unbiased recursive partitioning (URP) on both incident characteristic data as well as traffic data is proposed for incident duration prediction. In particular, a hybrid tree-based quantile regression method was developed; hybrid tree-based quantile regression modeling incorporates the merits of both quantile regression modeling and tree-structured modeling. Its merits include simple interpretation and ease of handling categorical covariates, robustness, and flexibility for nonlinearity. Given a URP tree, the hybrid method works by obtaining quantile regression models for each terminal node. With both 0.5 and 0.9 quantile estimates, traffic operators may understand not only the actual prediction but also the worst case results, and visualize the prediction range easily. Compared with the classic classification and regression tree (CART) approach, as well as a K-nearest neighbor approach, the URP trees and hybrid tree-based quantile regression proposed here appear to offer higher prediction accuracy.

The overall findings of this chapter can be summarized as follows:

- Incident characteristics (type, injuries, blocked lanes, number of vehicle involved etc) are the most significant predictors of incident duration.
- Traffic data can provide additional information that improves forecasting accuracy. Incidents with high prior speeds (occurring for instance during the night or during off-peak hours) generally last longer than those in daytime due to the lack of sufficient response units for incident clearance operations. Incidents with large occupancy increment tend to have longer duration than those with small occupancy changes.
- Incident location matters. Different geometry factors and jurisdiction may result in different incident duration.

In summary, it is essential to forecast the spatial-temporal incident impact based on both incident duration prediction and traffic conditions. Spatial-temporal incident impact aims to capture how congestion propagates over space and time. Future work in this area should leverage not only the model structure developed here, but in an online decision support system would incorporate real-time traffic predictions as predictor variables, in addition to the incident characteristics as they become available. Together, such a system can provide traffic operators with important components of an optimized control strategy for non-recurrent congestion.

References

Breiman, L. et al., 1984. *Classification and Regression Trees*, New York: Chapman & Hall.

Chassiakos, A.P. & Stephanedes, Y.J., 1993. Smoothing Algorithms for Incident Detection. *Transportation Research Record: Journal of the Tranportation Research Board*, 1394, pp.8-16.

Chin, S.M. et al., 2004. *Temporary Loss of Highway Capacity and Impacts on Performance: Phase 2*, Oak Ridge, Tennessee: Oak Ridge National Laboratory.

Cho, H.J. & Hong, S.M., 2008. Median regression tree for analysis of censored survival data. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 383, pp.715-726.

Choe, T., Skabardonis, A. & Varaiya, P., 2002. Freeway Performance Measurement System: Operational Analysis Tool. *Transportation Research Record: Journal of the Transportation Research Board*, 1811, pp.67-75.

CHP, 2011. CHP Traffic Incident Information Page. Available at: http://cad.chp.ca.gov/ [Accessed April 27, 2011].

Chung, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems. *Accident Analysis and Prevention*, 42, pp.282-289.

Chung, Y., Walubita, L.F. & Choi, K., 2010. Modeling Accident Duration and Its Mitigation Strategies on South Korean Freeway Systems. *Transportation Research Record: Journal of the Transportation Research Board*, 2178, pp.49-57.

Demiroluk, S. & Ozbay, K., 2011. Structure Learning for the Estimation of Non-Parametric Incident Duration Prediction Models. In *Proceedings 90th Annual Meeting of TRB (CD-ROM)*. Washington D.C.

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

Giuliano, G., 1989. Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research*, 23A, pp.387-396.

Golob, T.F., Recker, W.W. & Leonard, I.D., 1987. An analysis of truck involved freeway accidents. *Accident Analysis and Prevention*, 19, pp.375-395.

Hothorn, T. et al., 2011. CRAN - Package party. *party: A Laboratory for Recursive Partytioning*. Available at: http://cran.r-project.org/web/packages/party/ [Accessed May 3, 2011].

Hothorn, T., Hornik, K. & Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics,* 15, pp. 651-674.

Jones, B., Jassen, L. & Mannering, F.L., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis and Prevention*, 23, pp.239-255.

Khattak, A.J., Schofer, J.L. & Wang, M.-H., 1994. A simple time sequential procedure for predicting freeway incident duration. *IVHS Journal*, 1, pp.1-26.

Kim, W., Natarajan, S. & Chang, G., 2008. Empirical Analysis and Modeling of Freeway Incident Duration. In *Proceedings of the 11th International IEEE Conference on Intelligent Transportation Systems*. Beijing, China.

Koenker, R., 2005. Quantile Regression.Cambridge University Press.

Lee, D.-H., Jeng, S. & Ng, M., 2003. Defining the incident impact area for traffic diversion: knowledge discovery via a data mining approach. In *Proceedings 82th Annual Meeting of TRB (CD-ROM)*. Washington D.C.

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), pp.132-148.

Lewis, C.D., 1982. *Industrial and Business Forecasting Methods*, London: Butterworth- Heinemann.

Loh, W.-Y., 2008. Classification and Regression Tree Methods. *Encyclopedia of Statistics in Quality and Reliability,* F. Ruggeri, R. Kenett and F.W. Faltin (Eds.) Wiley, pp. 315-323.

Murthy, S.K., 1998. Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey, *Data Mining and Knowledge Discovery*, 2, pp. 345-389.

Nam, D. & Mannering, F.L., 2000. An exploratory hazard-based analysis of highway incident duration. *Transportation Research Part A*, 34, pp.85-102.

Oh, J. & Jayakrishnan, R., 2000. Temporal Control of Variable Message Signs Toward Achieving Dynamic System Optimum. In *Proceedings 79th Annual Meeting of TRB (CD-ROM)*.

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

Payne, H.J. & Tignor, S.C., 1978. Freeway Incident-Detection AlgorithmsBased on Decision Trees with States. *Transportation Research Record: Journal of the Tranportation Research Board*, 682, pp.30-37.

Qi, Y. & Smith, B.L., 2004. Identifying Nearest Neighbors in a Large-Scale Incident Data Archive. *Transportation Research Record: Journal of the Transportation Research Board*, 1879, pp.89-98.

Qi, Y. & Teng, H., 2008. An Information-Based Time Sequential Approach to Online Incident Duration Prediction. *Journal of Intelligent Transportation Systems*, 12(1), pp.1-12.

R Development Core Team, 2009. *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing.

Smith, K.W. & Smith, B.L., 2001. *Forecasting the Clearance Time of Freeway Accidents*, Charlottesville, VA: Center for Transportation Studies, University of Virginia.

Srinivasan, K. & Krishnamurthy, A., 2003. Roles of spatial and temporal factors in variable message sign effectiveness under nonrecurrent congestion. *Transportation Research Record: Journal of the Tranportation Research Board*, 1854, pp.124-134.

Therneau, T.M. & Atkinson, B., 2011. CRAN - Package rpart. *rpart: Recursive Partitioning*. Available at: http://cran.r-project.org/web/packages/rpart/index.html [Accessed May 4, 2011].

Wei, C.H. & Lee, Y., 2007. Sequential forecast of incidentduration using artificial neural network models. *Accident Analysis and Prevention*, 39(5), pp.944-54.