Needs Assessment for the Development of Data-Driven Predictive Non-Recurrent Delay Models for TRANSCOM Final Report

Submitted to:

TRANSCOM

August 2019 (V 0.1)
September 2019 (V. 0.2)
October 2019 (Final Report)

Department of Civil and Urban Engineering

School of Engineering

New York University (NYU)

Department of Civil and Environmental Engineering

Rutgers, The State University of New Jersey

Table of Contents

List of Tables

List of Figures

EXECUTIVE SUMMARY

The overall goal of this project is to review the existing predictive models and available TRANSCOM data in order to identify operations models that can best predict traffic impacts when a non-recurrent incident or event occurs. The selection process of these models is primarily driven by the needs of the TRANSCOM stakeholders as well as the available TRANSCOM data.

This project was comprised of four main tasks: The first task was a detailed literature review which relates to the development of data-driven predictive delay models for non-recurrent traffic congestion. The second task was to interview TRANSCOM stakeholders and to identify their needs for the development of a non-recurrent impact (delay) model. The third task was a detailed review of TRANSCOM data to help identify the most appropriate modeling approach, given the availability of TRANSCOM's historical as well as real-time data. The last task was to provide recommendations based on the findings of the previous tasks. The advantages and disadvantages of the recommended approaches are described using examples of their potential operations use cases. Moreover, this document provides the system requirements for an ideal predictive tool for non-recurrent traffic incidents (Section 5). This document also provides the assessment for development and implementation of recommended models (Section 6).

In this report, a detailed review of large number of past studies found in the literature is presented. The search identified any predictive operations models that can work with TRANSCOM data. Given this requirement, there was a limit on the possibilities of using off-the-shelf existing models. For example, many existing models require real-time traffic volume as one of the critical inputs; the lack of traffic volume in the TRANSCOM data limits the use of many of the existing models.

As a result of this review, the team could not identify any predictive delay model that would be compatible with the TRANSCOM data and that is currently being used by operational staff on a real-time basis. Given the feedback received from interviews with TRANSCOM stakeholders and available TRANSCOM data, this report recommends one model for each type of prediction task namely, impact duration prediction, traffic delay prediction/estimation, and queue length prediction.

For models predicting traffic impact duration, we recommend Demiroluk and Ozbay's (*1*) Bayesian network model. Their model can work with available TRANSCOM data and provide reasonable predicted results. Specifically, Demiroluk and Ozbay's (*1*) model is able to predict incident duration when there is very limited information available to traffic operators. Their model can also work with missing data and provide a predicted distribution of incident durations.

For incident delay prediction/estimation, we recommend a travel time prediction model developed by Yu (*2*) since it has the highest accuracy among all reviewed models. For the queue

length prediction, we recommend Ghosh's model (*3*) for predicting real-time queue length with reasonable accuracy using TRANSCOM's travel time data only.

Additional details of the recommended models have been provided in Section 4 of this report.

Based on the details of the recommended models, we propose and design a predictive tool for non-recurrent traffic incidents. Section 5 explains the properties of an ideal prediction tool and provides details about the designed functionalities and system requirements.

Lastly, this document includes a preliminary assessment for development and implementation of the recommended models. For each model, time requirements and development efforts for the calibration, validation and implementation are identified in Section 6. Finally, the last section provides a tentative timeline of the system development in steps.

Introduction and Study Objectives

The main goal of this document is to provide a detailed review of the previous models that were developed to predict non-recurrent traffic delay according to the task descriptions given in the scope of work. The scope of work is defined as listed below.

Task 1: Literature Review (Completed)

- Apply a comprehensive process that will focus on the review of the most recent predictive approaches that take advantage of big data from various sources.

Task 2: Interviews with TRANSCOM Stakeholders (Completed)

- Conduct a minimum of four face-to-face interviews that can be supplemented by several one-on-one phone interviews.
- MTA B&T, PANYNJ, NYSTA, NYSDOT, NYCDOT, MTA NYCT, NJ Transit, NJDOT and NJ Turnpike.

Task 3: Detailed Review of TRANSCOM Data (Completed)

- Determine geographical scope (ICM 495 corridor)
- TRANSCOM's historical traffic and event data will be obtained with the goal of identifying the most appropriate modeling approach(es).

Task 4: Final Recommendations and Final Report (Completed)

- Provide a final recommendation in the form of a final report that clearly documents findings of above tasks

Task 5: Project Management

- Meetings, quarterly and final reports, and other project management tasks

As mentioned above, in addition to the literature review, a detailed review of the TRANSCOM data for the ICM 495 corridor shown in Figure 1 below was conducted with the ultimate objective of determining the minimum data and ideal dataset requirements and providing recommendations.

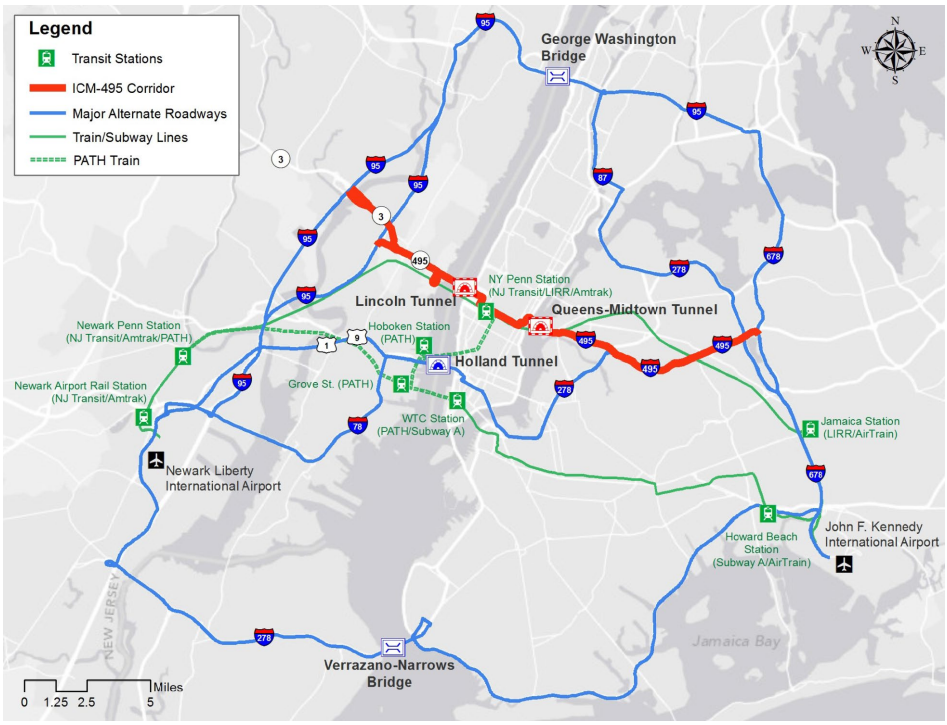


Figure 1. ICM-495 Corridor and Alternate Roadways from NJ Turnpike to JFK. (Source: https://www.nymtc.org/portals/0/pdf/presentations/MMN-ITS_ICM_Presentation_MM.pdf)

It is essential to emphasize further that our final recommendations will be made with a clear understanding that developed models will be used by operators at a Traffic Management Center to manage an incident in the best way possible. This has a few critical implications including the need for the models to:

1) work with existing real-time data;
2) be computationally efficient in order to produce almost instantaneous predictions;
3) generate easy to understand and disseminatable predictions;
4) adaptive to real-world changes as the incident removal operations progress. In the rest of this document, we provide a detailed review of relevant impact duration-delay estimation models.

Overview

A non-recurrent traffic incident comprises of four distinct intervals: detection, response, clearance, and recovery. This definition (*4*) is consistent with the incident timeline, which starts when an incident occurs, identifies key interim activities, notes when clearance of the roadway occurs and ends with traffic returning to normal conditions. Figure 2 shows the timeline of the elements of a typical traffic incident management operation.

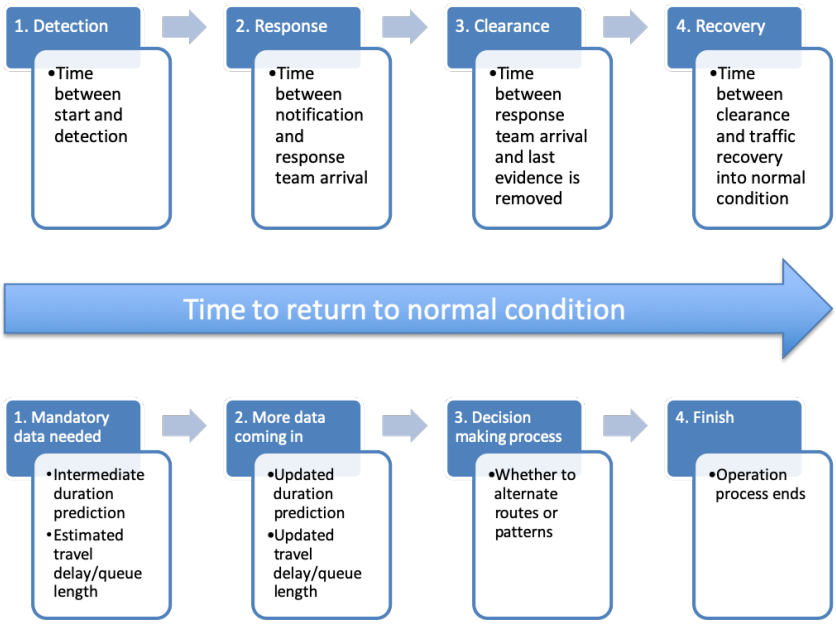


Figure 2. Timeline of the elements of a traffic incident.

At the operations level, when an incident occurs, operators first need to know how long this incident will last. This is the point where traffic impact duration prediction models are needed to provide an estimated impact duration. The availability of duration information will allow operators to assess the potential impacts of the incident. Next, operators need to quantify the traffic impact of the incident in order to make operations decisions. This is the point where traffic delay estimation models can help to provide traffic impact information, including traffic delay, the increase of travel time, and queue length. Figure 3 shows a general operations flowchart of a non-recurrent traffic incident.

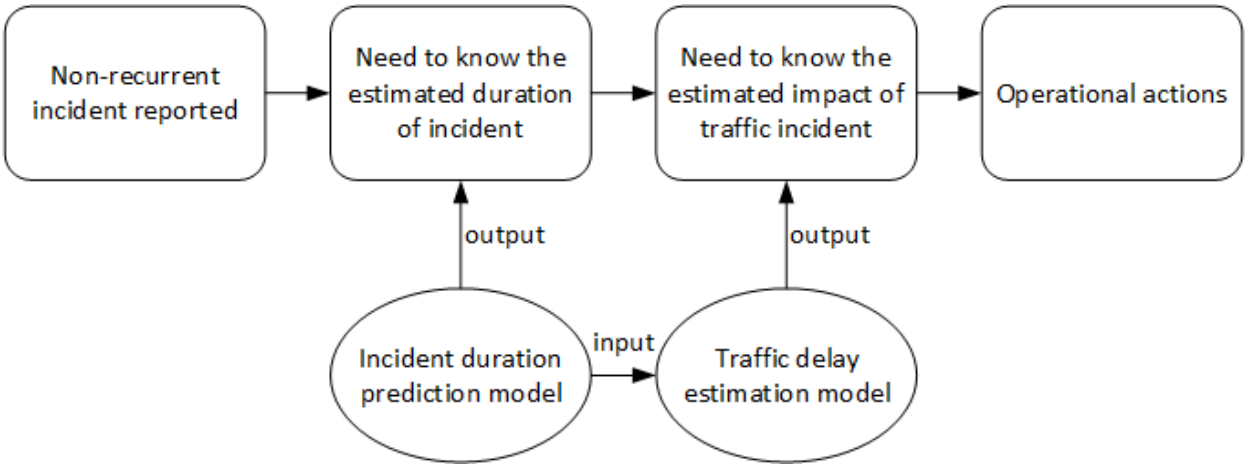


Figure 3 General operations process for a non-recurrent traffic incident event.

In the literature review, impact duration prediction and traffic delay estimation models are considered separately. This document has seven main sections. The first section is the literature review of traffic impact duration prediction models. This section categorizes models under five types of modeling approaches (regression, classification tree, Bayesian network, artificial neural network, hazard-based model). The second section is the literature review of traffic delay estimation/prediction models. This section divides delay estimation/prediction models into two categories, analytical and data-driven models. For each model under both categories, we provide a brief description of the modeling methodology, data needs, and detailed information about model evaluation in terms of their performance and advantages/disadvantages of using them. At the end of each subsection, a brief summary of models will also be provided. The summary includes an individual model performance comparison, compatibility of model data requirements compared with the data in the TRANSCOM database as well as highlights of each model.

The third section is a detailed data analysis towards the selected estimation models. This section analyzed and described the available TRANSCOM data by providing data fields, data quality check and potential usage for selected estimation models.

The forth section of this document provides a final recommendation of candidate models and the comparison between TRANSCOM data and data needs of these recommended models. In the fifth section, this document proposes an "**ideal data-driven predictive non-recurrent duration / delay estimation framework**" based on the outcomes of literature review study. The sixth section provides a preliminary assessment of development and implementation of recommended models. In the last section, this document proposes a timeline for the development of designed systems.

1.Traffic impact duration prediction

Impact duration prediction is one of the most critical steps of the overall incident management process. An accurate and reliable prediction of the impact duration can be the main difference between an effective incident management operation versus an unacceptable one. When a traffic incident occurs, a fast and accurate prediction will affect the effectiveness of the overall decision-making process of incident management operators. Thus, computationally efficient models that can work in real-time is a vital requirement.
To predict impact duration, there is a wide range of approaches that are proposed in the literature. This document divides these approaches into five main categories:

1. Regression-based models
2. Classification Tree Method (CTM) based models
3. Artificial neural networks
4. Machine-learning (Bayesian networks, SVM) based models
5. Hazard-based duration models

Table 1 below is a summary of duration prediction models reviewed in this task. It also provides each model's data compatibility with respect to TRANSCOM data. During the literature search task, data needs of reviewed models and their compatibility with data received from TRANSCOM were considered. We use three levels to represent models' data compatibility with TRANSCOM data:

1. **Low compatibility**: TRANSCOM data covers less than 40% of the data needs of a model, or TRANSCOM data does not have some fundamental inputs required by a model for its real-time operations use. For instance, many analytical models that attempt to estimate delay require real-time traffic volume.
2. **Medium compatibility**: TRANSCOM data does not currently cover some of the data requirements for real-time use of a model but it is expected that some of the missing data will be obtained at a later stage. For example, incident operation data such as the number of police vehicles involved cannot be reported to the operators instantaneously but can be updated/provided as the incident moves on.
3. **High compatibility**: TRANSCOM data covers most of the data needs of a model.

Table 1 Summary of traffic impact duration prediction models

	Model	TRANSCOM data compatibility	Highlights –
Regression models	Khattak et al., 1995 (1.1.1)	Low	Statistical regression, sequential/real-time model, operations, unreliable
	Garib et al., 1997 (1.1.2)	Medium	Statistical regression, one-time model, not operations, unreliable
	Peeta et al., 2000 (1.1.3)	Low	Statistical regression, one-time model, not operations, unreliable

	Khattak et al., 2016 (1.1.4)	High	Statistical regression, sequential/real-time, operations, too simplistic and too many categorical variables
	Yu and Xia et al., 2012 (1.1.5)	Low	Statistical regression, one-time model, not operations, unreliable
	Weng et al., 2015 (1.1.6)	Low	Statistical regression, one-time model, probabilistic, not operations, reliable
Classification Tree Methods	Ozbay and Kachroo, 1999 (1.2.1)	Low	Classification tree, sequential model, real-time prediction
	Smith et al., 2002 (1.2.2)	Low	Classification tree, sequential model, bad performance, not operations, real-time, unreliable
	Knibbe et al., 2006 (1.2.3)	Low	Classification tree, sequential model, simple, not operations, real-time, unreliable
	He et al., 2013 (1.2.4)	Medium	Classification tree, sequential model can extend to real-time model, operations, interpretable, reliable
	Zhan et al., 2011 (1.2.5)	Low	Classification tree, can extend to real-time model, deal with missing values, operations, unreliable
Artificial neural network	Wei and Lee, 2007 (1.3.1)	High	Artificial neural network, provide immediate and updated duration, operations, one-time and real-time capable, reliable
	Park et al., 2016 (1.3.2)	Medium	Artificial neural network, probabilistic, interpretable, one-time model, reliable
Bayesian networks	Ozbay and Noyan, 2006 (1.4.1)	Medium	Bayesian network, interpretable, capture stochasticity, sequential model, operations, reliable
	Boyles et al., 2007 (1.4.2)	High	Bayesian network, interpretable, capture stochasticity, sequential model, operations, unreliable
	Ji et al., 2008 (1.4.3)	Low	Bayesian network, deal with missing data, sequential model, not operations, reliable
	Shen and Huang, 2011 (1.4.4)	Low	Bayesian network, interpretable, capture stochasticity, sequential model, not operations, reliable
	Demiroluk and Ozbay, 2014 (1.4.5)	Medium	Bayesian network, interpretable, adaptive learning, sequential model, real-time prediction, operations
Hazard-based model	Qi and Teng, 2008 (1.5.1)	High	Hazard-based model, three-stage model, provide immediate and updated duration, operations, reliable

SVM	Yu et al., 2016(1.6.1)	Low	Support vector machine, interpretability, one-time model, not operations, reliable

1.1 Regression model-based impact duration prediction models

Traffic researchers applied several well-known statistical methods to predict the traffic impact duration. Regression is one of the most popular statistical approaches used for this goal. There are a few studies that applied regression models for the duration prediction problem in the literature.

1.1.1 A simple time-sequential procedure for predicting freeway impact duration. Khattak et al. (1995)

In one of the earliest academic studies, Khattak et al., 1995 (*5*) developed a truncated regression model and applied it using a time-sequential methodology. They predicted impact duration as the TMC receives the incident information based on a dataset of 109 large-scale incidents. In this study, it is assumed that the relationship between impact durations, y, and independent variables $x_1, x_2, \dots, x_k$ is of the form:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \beta' x_i + \epsilon_i$$

Where i refers to the i the observation; the set of n observations can be denoted as:

$$Y = X\beta + \epsilon$$

Where:

$Y =$ Vector of n dependent variable observations on impact duration
$X =$ Matrix of k independent variables and n observations
$\beta =$ Vector of k parameters
$\epsilon =$ The error term with expected value zero and variance σ^2

This study then applied several truncation points $\tau \in (10,15,20,25,30\ min)$ to compare model performance under different truncation points.

$$y_i = \beta' x_i + \epsilon_i \quad > \quad \tau_0 \ are\ included\ in\ the\ data\ observed, and$$
$$y_i = \beta' x_i + \epsilon_i \quad \leq \quad \tau_0 \ are\ exclued$$

In this way, the truncated regression model can receive input data in time-sequential order.

Data needs

Traffic data: Traffic flow conditions for the time of day and day of the week
Incident data: Incident type, vehicle type, number of vehicles involved, injuries and fatalities, state property damage.
Operations data: Response times, number of rescue vehicles, whether a heavy wrecker was needed if sanding/salting was done because of a spill/ice on the pavement, whether other agencies such as medical services and owners of the vehicles involved provided assistance, whether incident information is disseminated to motorists or not.
Time data: Time when incident is detected, time when incident is cleared, month of the year.
Location data: Freeway where the incident occurred, distance from the city center.
Weather data: Rainy or dry

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Can deal with time-sequential data and provide real-time impact duration prediction. • Considers ten distinct stages of impact duration based on the available information.	• Developed using a minimal data set, and it is questionable that it will work under various real-world conditions.	• Not tested under real-world conditions due to the lack of actual field data.

1.1.2 Estimating magnitude and duration of incident delays. Garib et al. (1997)

Garib et al., 1997 (*6*) introduced a statistical model for predicting impact duration using a linear regression model. They estimated the model using 205 incidents and claimed that their model reported adjusted R-square value as 81%. The input variables include the number of lanes affected, number of vehicles involved, truck involvement, time of day, police response time, and weather condition. The estimated model is shown below:

$$Log(Duration) = 0.87 + 0.27X_1X_2 + 0.2X_5 - 0.17X_6 + 0.68X_7 - 0.24X_8$$

Where:
$Duration$ = impact duration in minutes
X_1 = number of lanes affected by the incident
X_2 = number of vehicles involved in the incident
X_5 = dummy variable representing truck involvement in the incident
X_6 = dummy variable representing the time of day
X_7 = natural logarithm of the police response time in minutes
X_8 = dummy variable representing weather condition.

Data needs

Traffic data: None.
Incident data: Incident type, number of lanes affected, vehicle type, vehicle color.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Direction of an incident, lanes affected, upstream/downstream to the nearest exit.
Operations data: Time of police arrival, number of tow trucks.
Weather data: Rainy or dry.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**

• Simple and easy to use for operation purpose.	• Cannot deal with time-sequential data. • Cannot provide real-time impact duration prediction. • Developed using a minimal data set (205 incidents), and it is questionable that it will work under various real-world conditions.	• Best adjusted R^2: 81%

1.1.3 Providing real-time traffic advisory and route guidance to manage Borman incidents online using the Hoosier helper program. Peeta et al. (2000)

Peeta (*7*) estimated a linear regression model to estimate the clearance time of one incident using 835 crashes and 1176 debris (debris on the roadway). A simple linear regression model with four categories of explanatory variables was estimated: incident severity (including number of vehicles, trucks), incident lateral location variables (including locations on-ramp, median, left lane), environmental condition variables (such as night, temperature, vision) and current traffic condition variables (such as rush hour). Their linear regression model was estimated for both crashes and highway debris. The statistical performance was reported as $R^2 = 0.234$ for crashes and $R^2 = 0.362$ for debris. The model for predicting the duration of crashes is shown below:

$$Duration(Crashes) = 12.774 * ONE + 7.349 * NVEH + 2.930 * TRUCK + 18.055 * RAMP + 4.496 * MEDIAN + 9.095 * LL + 15.846 * CL + 9.780 * RL + 16.596 * NIGHT - 0.065 * TEMP - 0.136 * VIS + 32.842 * RAINH + 13.571 * RAINL + 6.527 * SNOW - 1.150 * RUSH$$

Where:

$Duration\ (Crashes) =$ Predicted impact duration that is caused by crashes
$NVEH =$ number of vehicles involved in the incident
$MEDIAN =$ if the incident occurred on the median
$LL =$ if the incident occurred on the left lane
$CL =$ if the incident occurred on the center lane
$RL =$ if the incident occurred on the right lane
$RAMP =$ if the incident occurred on the freeway ramp
$RAINH =$high intensity rain
$RAINL =$ low-intensity rain
$SNOW =$ if snowing during the incident clearance process
$NIGHT$ = if the incident clearance process occurs at night
$TRUCK =$ if a truck is involved in the accident

Data needs

Traffic data: Traffic volume, average traffic speed
Incident data: Incident type, vehicle type, number of vehicles involved, percentage of trucks at the time of incident.
Operations data: number of emergency crew at the time of incident, type of equipment used, number of equipment used, whether incident information is disseminated to motorists or not.
Time data: Time when incident is detected, time when incident is cleared.
Location data: None.
Weather data: Rain or snow
Light conditions: Night or daytime.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose.	• Cannot deal with time-sequential data. • Cannot provide real-time impact duration prediction. • Developed using a minimal and biased data set (only two types of incidents), and it is questionable that it will work under various real-world conditions.	• Best R^2: 0.234

1.1.4 Modeling traffic impact duration using quantile regression. Khattak et al. (2016)

Khattak (*8*) developed dynamic impact duration models and provided better prediction results than his previous models due to the capability of integrating additional information into the dynamic models. Their approach was based on ordinary least squares (OLS) regression models and able to predict primary and secondary impact durations.
They claimed that dynamic impact duration models predict impact duration more accurately since different time stages will support successively more information as incident progress. They tested both OLS and truncated regression models (their previous study) and claimed that truncated regression models under-predicted impact durations, especially when longer duration incidents were involved. The OLS regression model is shown below:

$$Y_{Duration} = \beta_0 + \beta_1(TOD) + \beta_2(WEATHER) + \beta_3(LOCATION) + \beta_4(AADT) + \beta_5(DETECTION) + \beta_6(VEHICLES) + \beta_7(TYPE) + \beta_8(LANECLOSURE) + \beta_9(EMS) + \beta_{10}(RTSHOULDER) + \beta_{11}(RAMP) + \beta_{12}(LFSHOULDER) + \epsilon$$

Where:
β = estimated parameters

$\epsilon =$ error term
$Y_{Duration} =$ impact duration (minutes)
$TOD =$ time of day incident occurred
$WEATHER =$ is the bad weather or not
$LOCATION =$ incident location
$AADT =$ average annual daily traffic
$DETECTION =$ incident detection source
$VEHICLES =$ number of vehicles involved in the incident
$TYPE =$ incident type
$LANECLOSURE =$ whether traffic lane was closed or not
$EMS =$ emergency medical service was present or not
$RTSHOULDER =$ the right shoulder affected by the incident
$LFSHOULDER =$ the left shoulder affected by the incident
$RAMP =$ ramp affected by the incident

The input information will be updated as traffic operation center (TOC) is involved, and new prediction based on new input data will be provided.

Data needs

Traffic data: AADT, detection source
Incident data: Incident type, number of vehicles involved.
Operations data: Whether response agencies are involved or not.
Time data: Time when incident is detected, time when incident is cleared, the peak time of day.
Location data: Location of the incident, number of lanes closed, whether left/right shoulder is affected, whether a ramp is affected.
Weather data: Whether severe weather or not.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Can deal with time-sequential data. • Can provide real-time impact duration prediction. • Can predict both primary and secondary impact durations. • Developed using a large dataset (59804 incidents).	• The model has low accuracy.	• Best MAPE: 37%.

1.1.5 A methodology for freeway impact duration prediction using computerized historical database. Yu and Xia. (2012)

Yu and Xia (*9*) proposed a linear model with stepwise regression. Their model could generate a preliminary prediction of impact duration when limited information about the incident is known. They then compared their proposed model with a more traditional linear model and claimed a more precise and dynamic prediction result by their model.
In their study, they provided a simple linear model shown below:

$$duration = 54.4 \times \exp(0.63 weather + 0.147 vehicle\ numbers + 0.263\ lane\ blockage)$$

Where
$duration =$ predicted the impact duration
$weather =$ whether rain or not
$vehicle\ numbers =$ number of vehicles involved in the incident
$lane\ blockage =$ number of lanes blocked due to the incident

Before using their data to estimate a linear model, their distribution estimation results indicated that their traffic incident followed a log-normal distribution and vehicle assistance data followed a logistic distribution. They then estimated the accumulative probability of log-normal and logistic distribution using historical data. To overcome the lack of available data, they introduced a stepwise procedure. They aggregated the cumulative probability distributions of different variables and used them to infer missing input data. They claimed that when more incident data becomes available, their model will provide more accurate predictions since the fitness of the estimated distribution would be improved with additional new data.

Data needs
Traffic data: None
Incident data: Incident type, number of vehicles involved, vehicle type, severity and fatality, property damage
Operations data: Response time for the relief station, travel time for the relief station, process time for an incident
Time data: Time when incident is detected, time when incident is cleared, day of the week, time of day
Location data: Number of lanes closed
Weather data: rainy or dry

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Can deal with time-sequential data. • Can provide real-time impact duration prediction.	• Developed using a minimal and biased data set (only 503 incidents), and it is questionable that it will	• Best prediction error for duration less than 60 minutes: 77.8%.

• Can deal with missing data. • Can provide better prediction with updated incoming data.	work under various real-world conditions.	

1.1.6 Cluster-based lognormal distribution model for accident duration. Weng et al. (2015)

Weng and his colleagues, (*10*) developed a cluster-based log-normal distribution model to predict accident duration. They first used a decision tree approach to split the entire dataset into three clusters, which are then treated as additional variables in modeling accident duration.

In their study, they modeled impact duration as a random variable which follows a log-normal distribution. Their decision tree method adopted F-test as the splitting criterion, and a detailed variable selection procedure was provided. The lognormal distribution model is as below based on 2512 incidents data:

$$\ln y = 2.49 - 0.16x_1 - 0.03x_3 + 0.07x_4 + 0.13x_5 - 0.19x_6 + 0.17x_7 + 4.6 \times 10^{-5}x_8 + 0.10x_{12} + 0.98 \times Cluster1 + 1.22 \times Cluster2 + 1.60 \times Cluster3 + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

Where

$$\sigma^2 = 0.25x_1 + 0.66 \times Cluster1 + 0.33 \times Cluster2 + 0.29 \times Cluster3$$

$$Cluster1 = \begin{cases} 1 & if\ x_3 \leq 2, \\ 0 & otherwise \end{cases}$$

$$Cluster2 = \begin{cases} 1 & if\ x_3 > 2\ and\ x_2 \leq 2, \\ 0 & therwise \end{cases}$$

$$Cluster3 = \begin{cases} 1 & if\ x_3 > 2\ and\ x_2 > 2, \\ 0 & otherwise \end{cases}$$

Data needs

Traffic data: Traffic volume, traffic speed.
Incident data: Severity and fatality, property damage, number of vehicles involved.
Operations data: Number of notifications sent from operation center, number of responders on the scene.
Time data: Time when incident is detected, time when incident is cleared, day of the week, time of day.
Location data: Number of lanes closed.
Weather data: Rain, wind and visibility.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose.	• Need to predetermine the form of prediction distribution.	• Best MAPE: 34.1%.

• Can provide predicted probabilistic distribution of impact duration.	• Cannot deal with time-sequential data. • Cannot provide real-time prediction.	

1.1.7 Summary of Regression-Based Duration Models

For regression-based models, one common property is that the implementation of models can be straightforward. However, most of the regression models require additional information that is not currently available in the TRANSCOM data set provided to the research team. Moreover, most of the regression models are estimated using a limited number of incidents, which reduces their reliability and thus makes them unsuitable for real-world operations. In this study, the emphasis is on operations use, reliability, and ability to dealing with sequential data. Therefore, the model from Khattak et al., 2016 appears to be a better-suited model among all the regression-based models reviewed in this section.

Table 2 Summary of regression-based duration prediction models

Regression-based duration prediction models

Model	**Performance**	**TRANSCOM Data Compatibility**	**Highlights**
Khattak et al, 1995	Not test	Low	Sequential/real-time model, operations, unreliable
Garib et al, 1997	Best Adj. R^2: 81%	Medium	One-time model, not operations, unreliable
Peeta et al, 2000	Best R^2:0.234	Low	One-time model, not operations, unreliable
Khattak et al, 2016	Best MAPE: 37%	High	Sequential/real-time, operations, reliable
Yu and Xia, 2012	Best prediction error for duration less than 60 minutes: 77.8%.	Low	One-time model, not operations, unreliable
Weng et al, 2015	Best MAPE: 34.1%	Low	One-time model, probabilistic, not operations, reliable

1.2 Classification Tree Method (CTM) based impact duration prediction methods

There are several studies that employed classification tree-based methods for the impact duration prediction.

1.2.1 Incident Management in Intelligent Transportation Systems. Ozbay and Kachroo. (1999)

Ozbay and Kachroo (*11*) were among the first researchers to recognize that the non-homogenous nature of the impact duration data interferes with the ability to use traditional linear regression for model estimation. They reported that the impact duration values did not follow either a lognormal or log-logistic distribution. They then employed the classification tree to estimate the impact duration.

Data needs

Traffic data: None.
Incident data: Incident type, whether heavy vehicles are involved or not, severe injuries and fatalities, property damage or not.
Operations data: Whether heavy wrecker is used or not, whether assistance from response agencies is needed or not.
Time data: Time of day, day of week.
Location data: Total number of lanes, number of closed lanes, whether shoulders exist or not.
Weather data: Extreme weather or not.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Can provide real-time predictions. • Assumed log-normal distribution instead of general Gaussian distribution. • Requires low computation efforts.	• Developed using a minimal data set.	• Best correct classification rate: 60%.

1.2.2 Forecasting the clearance time of freeway accidents. Smith et al. (2002)

Smith et al. (*12*) investigated three forecasting models that can predict the clearance time of a freeway accident, namely, a stochastic model, nonparametric regression model, and a classification tree model. However, the results in this paper indicate that the classification models are not promising, they also do not show a meaningful performance improvement from the nonparametric regression models. Figure 4 shows the classification tree model diagram presented in this paper.

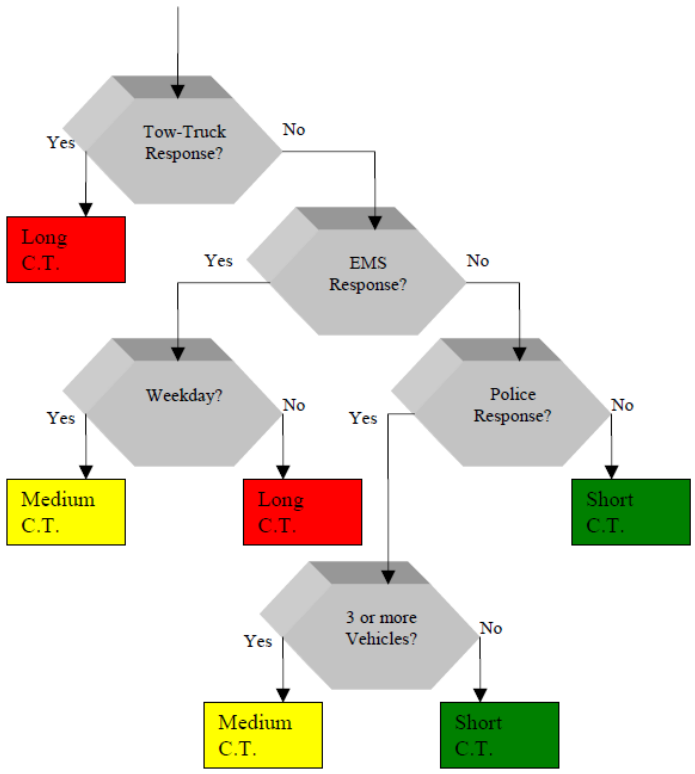


Figure 4 Classification tree model. Breiman et al. (*13*).

Data needs

Traffic data: None.
Incident data: Number of vehicles, whether trucks are involved or not, whether buses are involved or not.
Operations data: Whether agencies response or not (EMS, police, FIRT, hazardous material agency, VDOT), whether tow trucks involved or not.
Time data: Time of day, day of week.
Location data: None.
Weather data: Severe weather or not.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Can provide real-time predictions. • Can deal with time-sequential data.	• Model performance is not satisfactory with only 58% correct classification rate.	• Best correct classification rate: 58%.

1.2.3 Automated estimation of impact duration on Dutch highways. Knibbe et al. (2006)

Knibbe et al. (*14*) proposed a classification tree method for real-time impact duration estimation. In this approach, sequences of decision trees are constructed and used for determining the expected duration interval of an incident. Table 3 shows the decision tree's main parameters in this paper. This approach can also be used for real-time impact duration estimation.

Table 3 Main parameters for incident classification. Knibbe et al. (*14*).

Main Parameters Used For Incident Classification			
Incident	**Accident**	Passenger car	No casualties Casualties
		Truck	No casualties
	Stopped vehicle	Passenger car	Casualties Malfunction Fire
		Truck	Malfunction Fire
	Load	-	-

Data needs

Traffic data: None.
Incident data: Incident type, vehicle type, number of vehicles involved, property damage.
Operations data: Whether response agencies are involved or not (Police, ambulance, road manager, fire department), whether tow trucks are involved or not, whether repair service is required or not, whether a police investigation is required or not, Type of towing required, whether traffic control is required or not.
Time data: Time of day, day or week.
Location data: None.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Can provide real-time predictions. • Can deal with time-sequential data.	• Model performance is not satisfactory.	• Best correct classification rate: 29%.

1.2.4 Impact duration prediction with hybrid tree-based quantile regression. He et al. (2013)

He et al. (*15*) used a hybrid tree-based quantile regression method, which incorporates the merits of both quantile regression modeling and tree-structured modeling. The implementation in this paper was based on the software provided by the developers of this method (Hothorn et al. 2011). Significance levels for the test statistics were set to conventional levels (0.05) as suggested in Hothorn et al. (2006). There were two unbiased recursive partitioning (URP) trees with different sets of predictors. The first one, called URP tree1, shown in Figure 5, was created using all candidate variables. The second one (URP) was obtained using all but traffic variables and is depicted in Figure 6. Specifically, URP tree2 is a subset of URP tree1 that did not contain traffic data variables. The decision path of the tree model was followed by answering a yes or no question at each node.

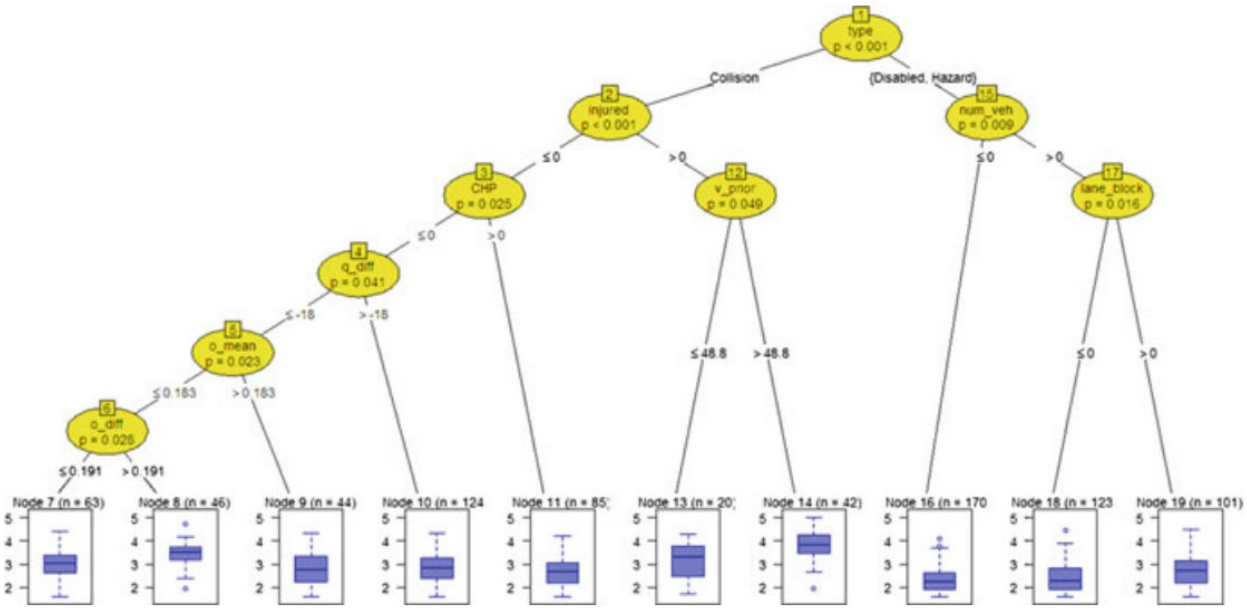


Figure 5 URP tree1 (with traffic data). He et al. (*15*)

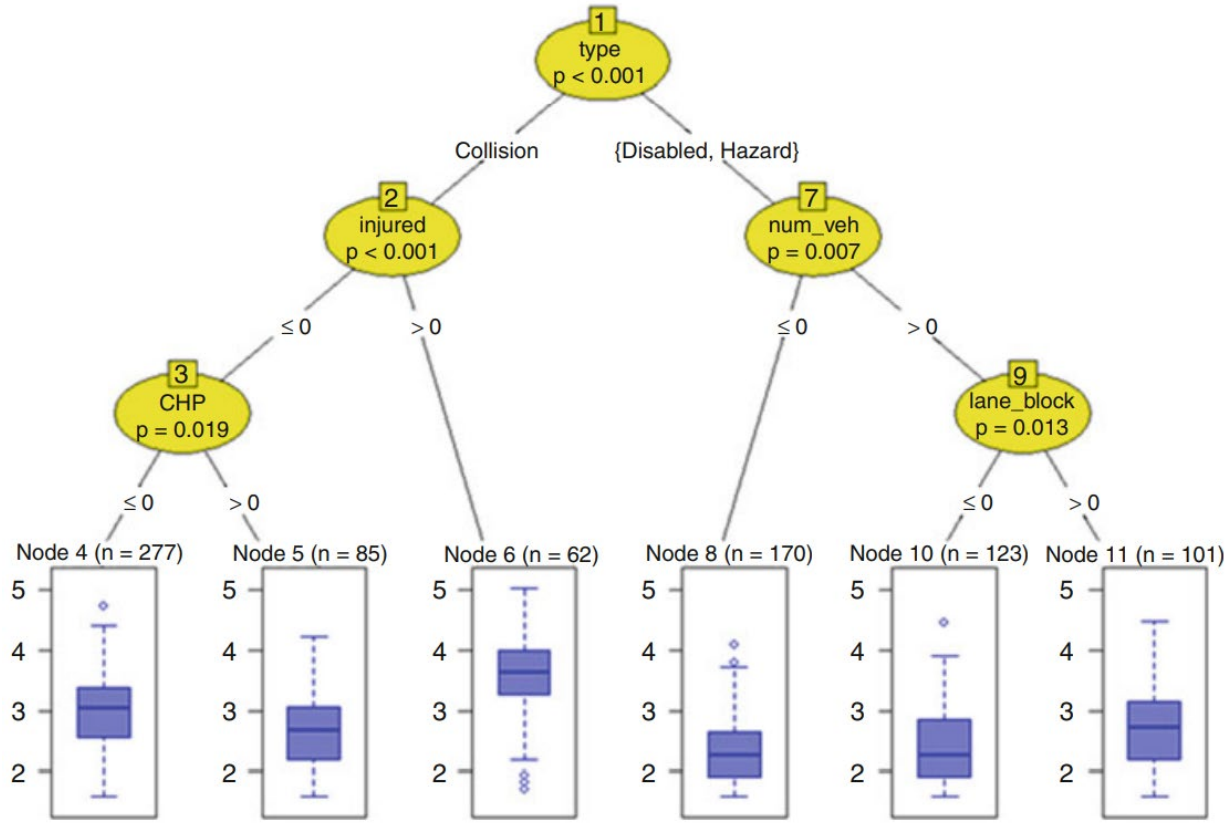


Figure 6 URP tree2 (without traffic data). He et al. (*15*)

Data needs

Traffic data: Average speed, average traffic volume, average occupancy
Incident data: Incident type, number of vehicles involved, severities and fatalities, property damage
Operations data: None
Time data: Time of day, day of week
Location data: Whether a ramp exists, whether happened on highway
Weather data: Rain or snow

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Can provide real-time predictions. • Can deal with time-sequential data.	• Model performance is not satisfactory.	• Best MAPE: 49.1%.

1.2.5 Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. Zhan et al. (2011)

Zhan et al (*16*) proposed an M5P tree algorithm for lane clearance time prediction. This algorithm can work with categorical and continuous variables as well as variables with missing values.

Figure 7 shows the three significant steps for M5 tree development: 1) tree construction; 2) tree pruning; and 3) tree smoothing. The M5 tree construction process attempts to maximize a measure called the standard deviation reduction (SDR). The SDR is defined as

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i)$$

Where T is the set of cases, T_i is the ith subset of cases that result from the tree splitting based on a set of variables (attributes), $sd(T)$ is the standard deviation of T, and $sd(T_i)$ is the standard deviation of T_i as a measure of error.

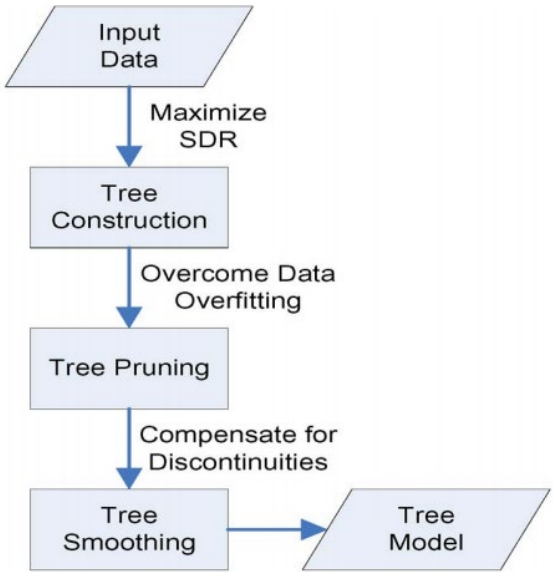


Figure 7 M5 tree flowchart. Zhan et al. (*16*).

The developed M5P regression tree model is shown in Figure 8. The regression sub-models [see (LM1)–(LM5) in Figure 8] are listed as follows:

$$\begin{aligned} \text{LM1}: \tau(Y,\lambda) = {} & 2.912 + 1.117 \times \text{NumRRAssists} - 0.09 \times \text{TMCResponse} \\ & + 0.091 \times \text{TMCVerification} + 0.892 \times \text{Injury} - 0.999 \times \text{ShoulderAvailable} \\ & + 2.093 \times \text{hasFullBlockage} + 0.542 \times \text{Weekend} + 0.908 \times \text{Tractor} \\ & + 1.602 \times \text{Truck} - 0.496 \times \text{DisabledVehicle} - 0.372 \times \text{CCTV} \\ & + 0.023 \times \text{DMSCount} \end{aligned}$$

$$\begin{aligned} \text{LM2}: \tau(Y,\lambda) = {} & 5.219 + 1.997 \times \text{NumRRAssists} - 0.154 \times \text{TMCResponse} \\ & + 0.887 \times \text{TMCV erif ication} + 4.875 \times \text{SIRV} + 12.104 \times \text{BUS} \\ & + 3.613 \times \text{Tractor} \end{aligned}$$

$$\begin{aligned}
\text{LM3}: \tau(Y,\lambda) &= 7.142 - 4.971 \times \text{ShoulderAvailable} + 1.694 \times \text{NumRRAssists} \\
&\quad - 0.155 \times \text{TMCResponse} + 2.752 \times \text{Weekend} + 0.080 \times \text{DMSCount} \\
&\quad + 7.017 \times \text{BUS} + 7.025 \times \text{Emergency} + 1.825 \times \text{Illumination} \\
&\quad + 2.080 \times \text{Rollover} + 0.393 \times \text{VehicleCount} + 2.826 \times \text{HasFullBlockage} \\
&\quad + 1.629 \times \text{Tractor} \\
\text{LM4}: \tau(Y,\lambda) &= -330.463 + 2328.506 \times \text{TotalActivities} + 2058.012 \times \text{Injury} \\
&\quad - 1649.351 \times \text{NumRRDispatches} + 4103.359 \times \text{SIRV} + 1743.637 \times \text{I595E} \\
&\quad + 851.413 \times \text{Weekend} - 68.838 \times \text{TMCResponse} \\
&\quad + 60.161 \times \text{TMCVerification} \\
\text{LM5}: \tau(Y,\lambda) &= 5.581 + 2.095 \times \text{NumRRAssists} - 2.466 \times \text{ShoulderAvailable} \\
&\quad - 3.436 \times \text{Midday} + 1.735 \times \text{Rollover} - 3.422 \times \text{PM} - 2.285 \times \text{AM} \\
&\quad + 1.989 \times \text{Tractor} - 0.087 \times \text{TMCResponse} + 4.554 \times \text{Truck} \\
&\quad + 0.581 \times \text{TotalLanes} + 2.276 \times \text{Fire} + 0.915 \times \text{Injury}
\end{aligned}$$

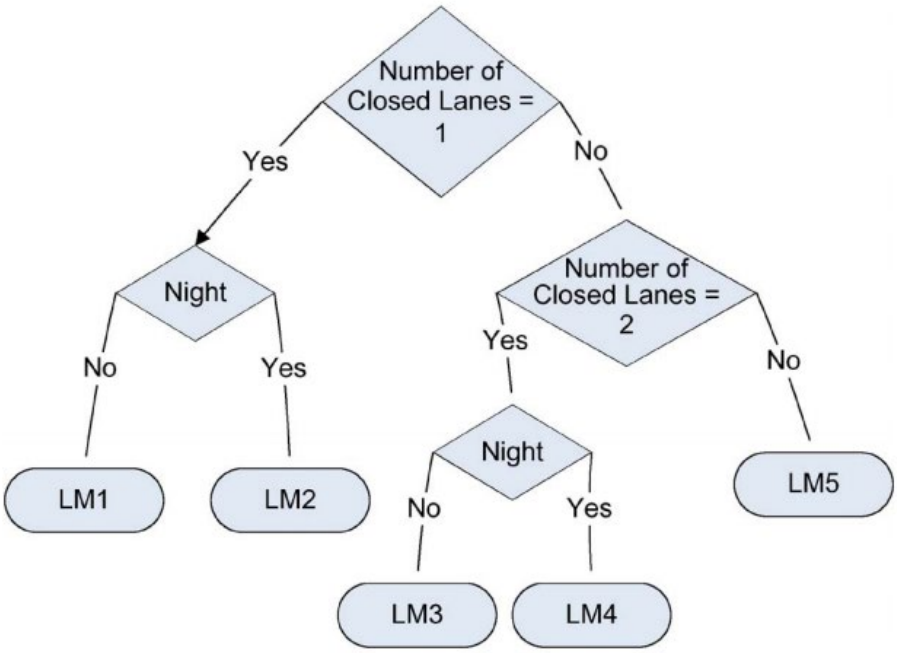


Figure 8 M5P regression tree for lane clearance time prediction. Zhan et al. (*16*).

Data needs

Traffic data: None

Incident data: Incident type, vehicle type, number of vehicles involved, injury and fatalities.

Operations data: Response times from the operation center, whether response agencies involved or not (such as Road ranger, Highway patrol), whether detected by CCTV or not, whether dynamic message sign (DMS) activated or not, number of on-site assists by the road ranger, number of on-site assists performed by agencies

Time data: Time of day, day of week, time when incident is detected, time when incident is cleared

Location data: Total number of lanes, number of lanes closed, whether shoulders exist or not and whether shoulders blocked or not.
Weather data: Rainy or dry, severe weather or not.
Visibility data: Clear or foggy.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use for operation purpose. • Provides real-time predictions. • Deals with time-sequential data. • Deals with missing data.	• Model performance is not satisfactory.	• Best MAPE: 42.7%.

1.2.6 Summary of Classification Tree Method (CTM) models

Most of classification tree methods require incident attributes such as severity, number of vehicles involved, and incident operation condition. This type of data is not provided in TRANSCOM database. Moreover, most CTM models do not have good classification rates.

Table 4 Summary of Classification Tree (CTM) based Impact duration Methods
Classification Tree (CTM) based Impact duration Methods

Model	**Performance**	**TRANSCOM Data Compatibility**	**Highlights**
Ozbay and Kachroo, 1999	Correct rate: 60%	Low	Sequential model, real-time prediction
Smith et al, 2002	Correct rate: 58%	Low	Sequential model, bad performance, not operations, real-time, unreliable
Knibbe et al, 2006	Correct rate: 29%	Low	Sequential model, simple, not operations, real-time, unreliable
He et al, 2013	MAPE: 49.1%	Medium	Sequential model can extend to real-time model, operations, interpretable, reliable
Zhan et al, 2011	MAPE: 42.7%	Low	Can extend to real-time model, deal with missing values, operations, unreliable

1.3 Artificial neural network-based impact duration methods

1.3.1 Sequential forecast of impact duration using Artificial Neural Network. Wei and Lee. (2007)

Wei and Lee (*17*) used an Artificial Neural Network (ANN) as well as data fusion technique to build a multi-period forecast model for predicting the impact duration. They proposed two types of impact duration models (Model A and B) to perform forecasts in impact duration. When an incident is noticed for the first time, they used Model A to perform a preliminary forecast of the impact duration. After the incident, Model B takes over from Model A to perform forecasts with updated data. Model A and B together provide a sequential forecast for the impact duration. Their study only considers the car accident data for modeling building and evaluation. The model structure is as follows:

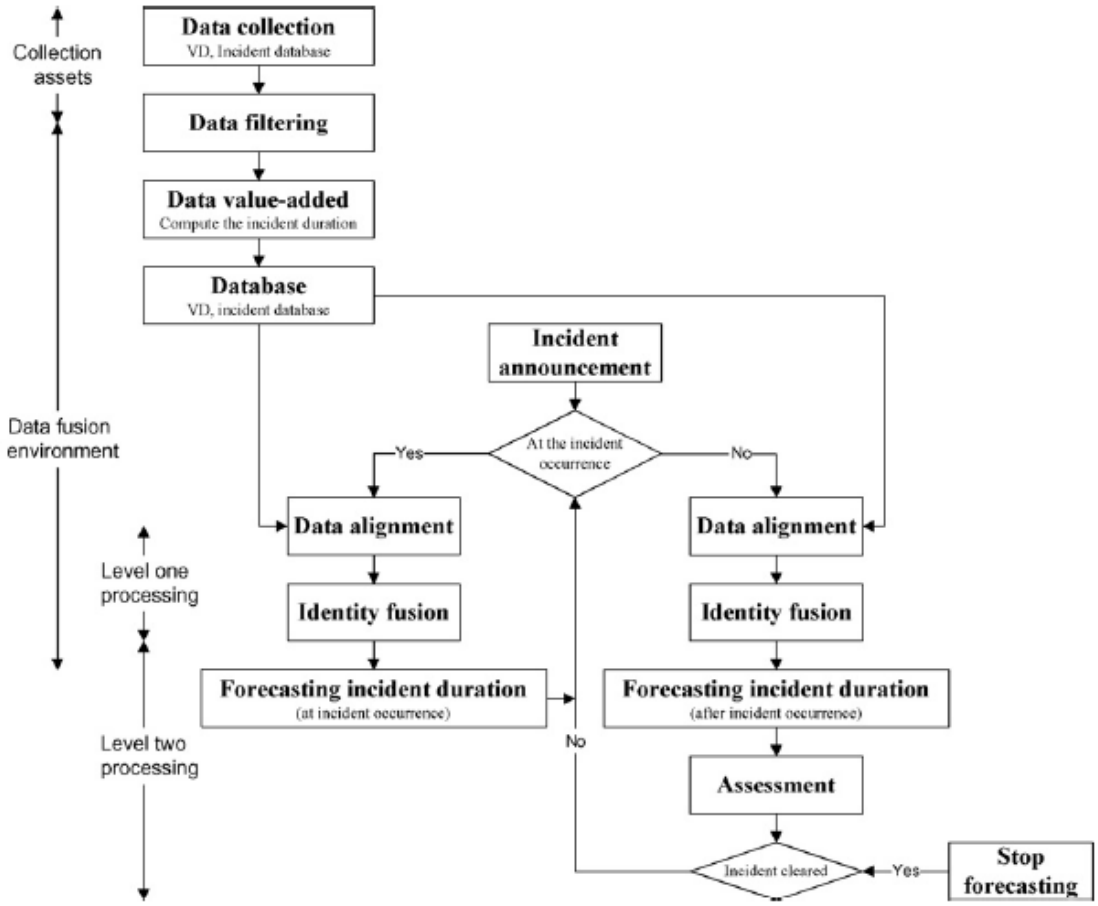


Figure 9. Impact duration forecast flowchart. Wei and Lee (*17*).

Data needs

Traffic data: Traffic volume, traffic speed.
Incident data: Incident type, vehicle type, number of vehicles involved.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Location of the incident, whether an interchange exists between the incident and the detector, whether a toll plaza or service area exists between the incident and the detector, the distance between the incident and detector locations.
Weather data: None.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides real-time prediction (both immediate and updated prediction).	• The model is trained with a minimal and biased dataset (only one incident type and 39 incidents).	• Best MAPE: 29%.

• Deals with time-sequential data.	• Requires heavy computation effort and time-consuming.	

1.3.2 Interpretation of Bayesian neural networks for predicting the duration of detected incidents. Park et al. (2016)

Park et al (*18*) introduced a Bayesian neural network model to predict the impact duration. They applied Monte Carlo algorithm to update BNN parameters and adopted a pedagogical rule extraction algorithm (TREPAN) to extract decision trees to explain potential relationships present in incident nature. In other words, they combined a Bayesian neural network model with decision tree technique to provide both predictive and explanatory impact duration results. The methodology is as follows:

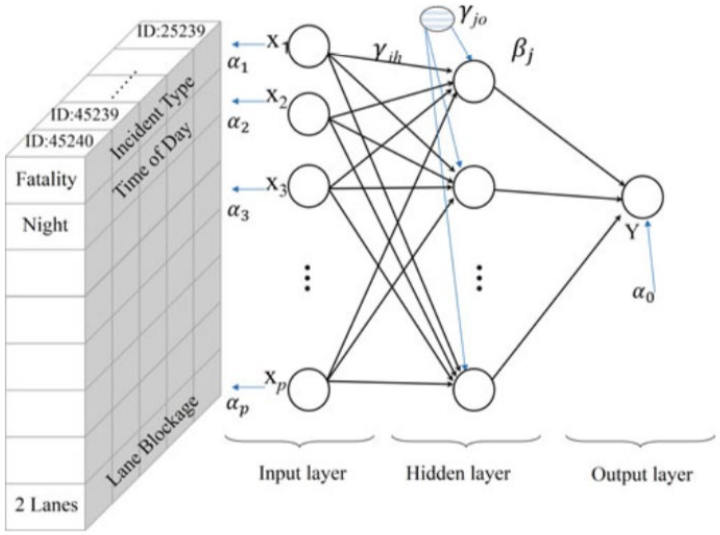


Figure 10 Structure of the Bayesian neural network. Park et al. (*18*)

During the implementation of the Bayesian neural network, they applied hybrid Monte Carlo (HMC) to sample the posterior distribution to get predictive results. They then applied TREPAN to extract rules from Bayesian neural network models and form a decision tree to interpret the predicted impact duration as follows:

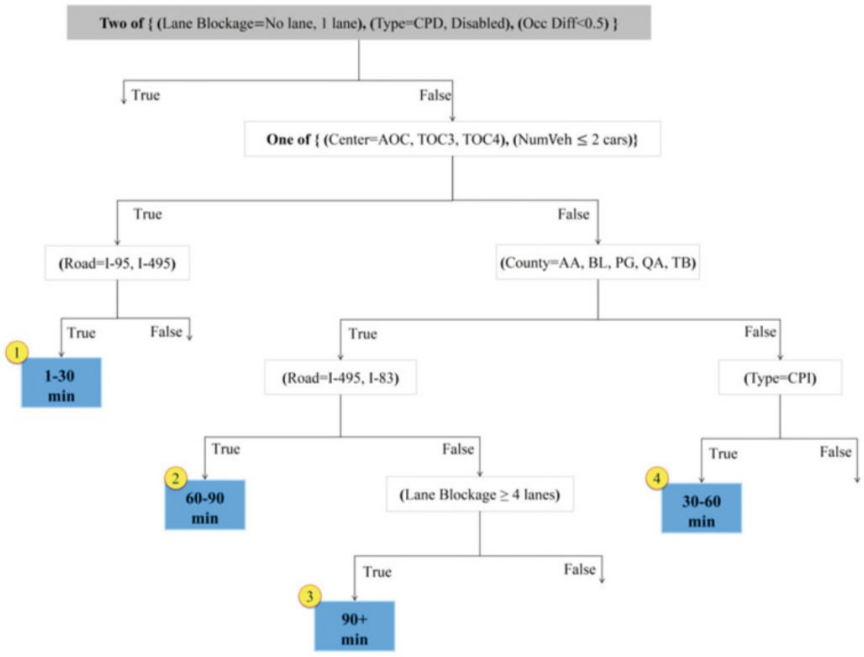


Figure 11 Extracted decision tree from Bayesian neural network. Park et al. (*18*).

Data needs

Traffic data: Travel time before and after the incident occurrence.
Incident data: Incident type, number of vehicles involved, vehicle type, severities, and fatalities.
Operations data: Type of operations center agencies, whether the incident clearance is operated by highway response team or police department, type of response equipment involved.
Time data: Time of day, time when incident is detected, time when incident is cleared.
Location data: Number of lanes closed.
Weather data: Snow or rain.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides interpretable results. • Provides probabilistic distribution of predicted duration.	• Require heavy computation effort (Monte-Carlo simulation) and time-consuming.	• Best MAPE: 18%.

1.3.3 Summary of Artificial Neural Network models

For artificial neural network models, one significant property is that they can achieve high accuracy by training historical dataset. Both of the two selected models can provide immediate duration prediction when an incident is detected/reported. Furthermore, Wei and Lee's model can update the model itself by occupying the second ANN model and read in upcoming data to provide an accurate incident prediction. One drawback of ANN models is that it usually requires heavy computation, which may lead to a low prediction frequency and cannot be used for real-time operations.

Table 5 Summary of artificial neural network models

Artificial neural network models			
Model	**Performance**	**TRANSCOM Compatibility**	**Highlights**
Wei and Lee, 2007	Best MAPE: 29%	High	Provide immediate and updated duration, operations, one-time and real-time capable, reliable
Park et al, 2016	Best MAPE: 18%	Medium	Probabilistic, interpretable, one-time and real-time capable, reliable

1.4 Bayesian Network-based impact duration prediction methods

1.4.1 Estimation of incident clearance times using Bayesian Networks approach. Ozbay and Noyan. (2006)

Ozbay and Noyan (*19*) were the first researchers to use Bayesian Networks (BNs) to model the incident clearance durations. Considering the stochastic variation and presence of incomplete information of incident data, BNs is a powerful modeling and analysis tool to create dynamic impact duration estimation trees because of its three main advantages, which are bi-directional induction, incorporation of missing variables and probabilistic inference. BNs consist of two components, one is a directed acyclic graph, and the other is the probability distribution over a set of random variables. By learning over the space of possible graph structures and model parameters with the relationships suggested by the data, a Bayesian scoring algorithm is used to find the BN that maximizes the scoring criterion. The BN with the highest score is shown in Figure 12. Figure 13 shows the conditional probability distributions for some of the variables in this BN. Through rigorous validation of the estimated trees using real-world data set collected in Northern Virginia, the prediction methodology is shown to be fully capable of representing the stochastic nature of incidents.

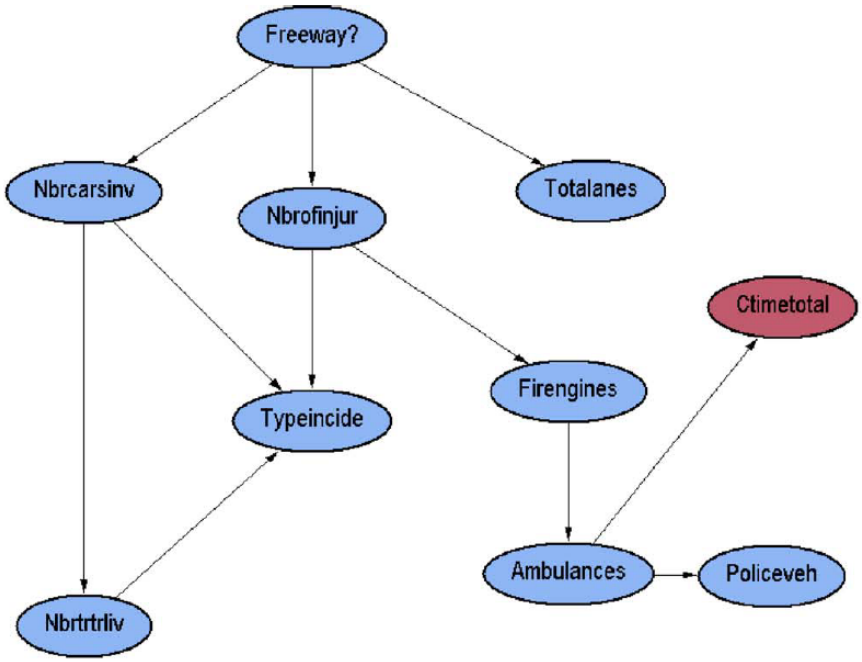


Figure 12 BN structure: nodes and arcs. Ozbay and Noyan (*19*).

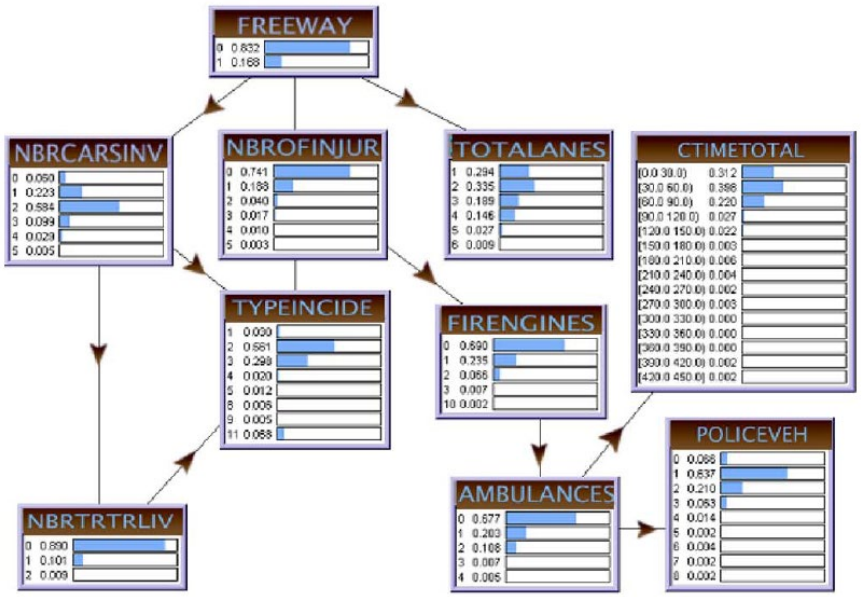


Figure 13 Posterior conditional probability distributions of the nodes. Ozbay and Noyan (*19*)

Data needs

Traffic data: None.
Incident data: Incident type, injuries and fatalities, number of vehicles involved, vehicle type.
Operations data: Number of response agencies involved.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Type of roadway, number of lanes.

Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides real-time predictions. • Deals with time-sequential data. • Deals with missing data. • Captures the stochastic nature of incidents. • Provides interpretable results and easy for operations use.	• Needs to be improved by testing various prior distributions on the decision variable.	• Accuracy rate: 80%.

1.4.2 A naïve Bayesian classifier for impact duration prediction. Boyles et al. (2007)

Boyles et al. (*20*) developed a probabilistic model based on a naïve Bayesian classifier (NBC) for prediction of impact duration. The proposed model can readily accommodate incomplete information or information received at different points in time, both of which are characteristics of the incident management process. Similar to other classifiers, NBC can calculate the probability of our objective belonging to a discrete set of categories, conditioning on the observed attributes. The final result of the objective is typically assigned to the category with the highest probability. In the context of impact duration prediction, the observed attributes correspond to observable incident characteristics, such as number of injured persons, number of blocked lanes, location of the incident, weather conditions, and so on. The NBC classifies incidents into one of three categories: those lasting less than half an hour, between half an hour and an hour, and longer than an hour. The proposed model includes sixty-two attributes. When applied to the validation set, the results of NBC classifier are compared with a linear regression model. The validation results showed that NBC can provide a more straightforward, more flexible, and more useful approach than the regression model without scarifying prediction accuracy.

Data needs

Traffic data: None
Incident data: Incident type, injuries and fatalities, number of vehicles involved, vehicle type, property damage
Operations data: None
Time data: Time when incident is detected, time when incident is cleared
Location data: Type of roadway, number of lanes affected
Weather data: None

Model Highlights

Advantages	Disadvantages	Model performance
• Provides real-time predictions. • Provides robust prediction to outliers than regression models. • Captures the stochastic nature of incidents. • Provides interpretable results and easy for operations use.	• Model performance is not satisfactory with the correct classification rate of 50%. • Cannot provide a distribution of impact duration.	• Correct classification rate: 50%.

1.4.3 Traffic impact duration prediction based on the Bayesian decision tree method. Ji et al. (2008)

Ji et al. (*21*) presented a prediction model based on a Bayesian decision tree model to estimate traffic impact duration. This model is defined as a Bayesian decision tree model because Bayesian nodes are inserted into the generic decision tree model, as shown in Figure 14. Each Bayesian node contains a value which is either "*0*" or "*f*". If the characteristic object information is complete, the value of Bayesian node is "*0*" and there are no calculations. However, if the object characteristic is missing, the value of Bayesian node will be set to "*f*" and need to be calculated later. The proposed model is capable of dealing with "dirty" traffic incident data, which may contain incomplete or inconsistent information. The theoretical accuracy of this model is higher than traditional classification tree models.

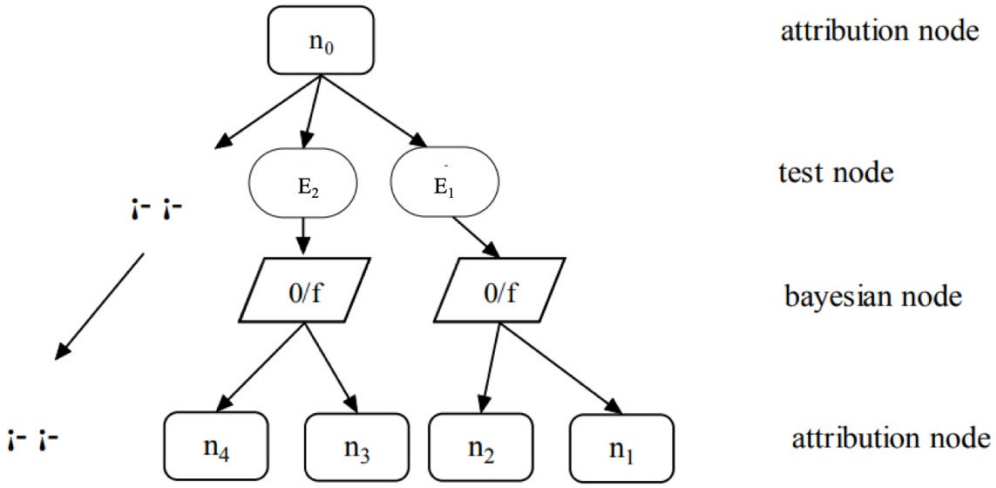


Figure 14 Illustration of Bayesian decision tree model. Ji et al. (*21*)

Data needs

Traffic data: None.

Incident data: Incident type, vehicle type, property damage.
Operations data: Whether roadway is closed or not, whether response agencies involved or not (police, road manager, tow truck).
Time data: Time when incident is detected, time when incident is cleared, day of week.
Location data: None.
Weather data: None.

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides real-time predictions. • Provides robust prediction to outliers than CTM models. • Captures the stochastic nature of incidents. • Deals with missing data. • Provides interpretable results and easy for operations use.	• Cannot deal with time-sequential data. • Cannot provide a distribution of impact duration.	• Best correct classification rate: 74%.

1.4.4 Data mining method for impact duration prediction. Shen and Huang. (2011)

Shen and Huang (*22*) developed a Bayesian Network (BN) model for predicting impact duration based on the time sequence of incident management stages after an incident is verified by the Fort Lauderdale Traffic Management Center (TMC) in Florida and response vehicle arrived at the incident location. A BN represents the cause-effect relationships and conditional dependencies between variables of interest by a directed acyclic graph and local conditional probability distribution for each node that defines the joint probability distribution. Since BN cannot handle continuous variables, it is necessary to discretize the continuous impact duration variables into nominal variables first. Through structural learning and probability inference, the structure of a BN can be determined. This network is used as the basis to compute probabilities of interest-based on the Bayes' theorem. The constructed model structure of the final graph is shown in Figure 15. The advantage of this model is that the probability results are straightforward and the prediction accuracy is acceptable. In addition, given the apparent varying nature of impact duration data, this model is robust with respect to outliers by classifying incidents into broader categories according to some field applications.

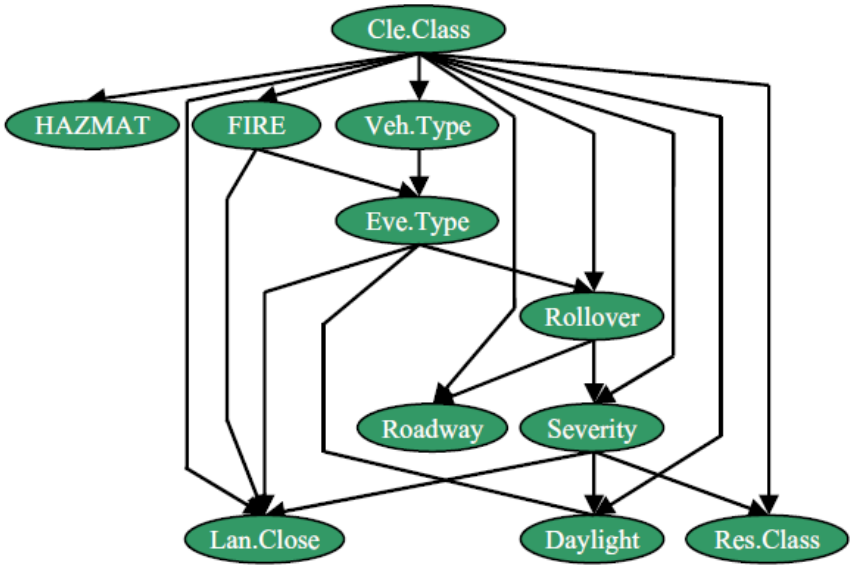


Figure 15 Structure of the learned Bayesian Network. Shen and Huang (*22*)

Data needs

Traffic data: None.
Incident data: Incident type, vehicle type, number of vehicles involved, severity, injuries and fatalities, property damage.
Operations data: Whether response agencies involved or not, whether first notify Traffic management center or not.
Time data: Time when incident is detected, time when incident is cleared, time of day, day of week.
Location data: Roadway type, total number of lanes, number of lanes closed, pavement conditions.
Weather data: Rainy or dry, daylight on or not

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides real-time predictions. • Provides robust prediction to outliers than CTM models. • Captures the stochastic nature of incidents. • Provides interpretable results and easy for operations use.	• Cannot deal with continuous variables.	• Best correct classification rate: 72.6%.

1.4.5 Adaptive learning in Bayesian networks for impact duration prediction. Demiroluk and Ozbay. (2014)

Demiroluk and Ozbay (*1*) developed a new adaptive model based on Bayesian networks for impact duration prediction. They adopted three types of Bayesian network structures, including Naïve Bayes model, tree-augmented naïve Bayes model (TAN) and K2 model to discover the best Bayesian network for impact duration prediction. In the validation section, they used BIC (Bayesian Information Criterion) scores to assess the overall fitness of models and facilitate the comparison of these three models. They then proposed an adaptive learning algorithm for real-time prediction of impact durations. Their model showed an increase in prediction accuracy with the use of the adaptive learning algorithm and provided reasonable (best correct classification rate as 93.3%) prediction results. Figure 16 shows the mechanism of adaptive learning as part of the best Bayesian network model identified in the previous step.

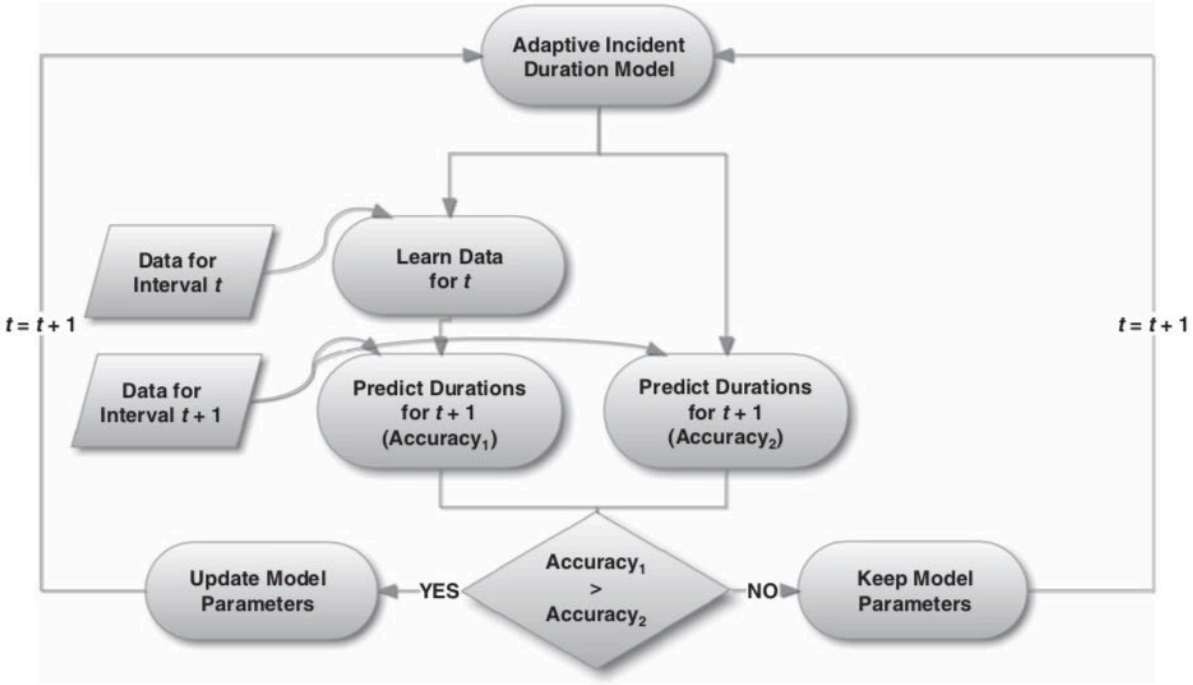


Figure 16 Adaptive learning mechanism in the context of Bayesian network model. Demiroluk and Ozbay (*1*)

Data needs

Traffic data: None.
Incident data: Incident type, vehicle type, number of vehicles involved, severity, injuries and fatalities, property damage.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared, time of day, day of week, month of year.
Location data: Roadway type, pavement conditions, distance from the closest exit.

Weather data: clear or not, rain or not, snow or not, fog or not
Light data: daylight or not, dawn or not, dusk or not, dark or not.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides real-time predictions. • Deals with time-sequential data. • Provides robust prediction to outliers than CTM models. • Captures the stochastic nature of incidents. • Provides interpretable results and easy for operations use. • Deals with missing data.	• The prediction accuracy is relatively low with limited real-time data.	• Best correct classification rate: 63.1% (without adaptive learning), 93.3% (with adaptive learning).

1.4.5 Summary of Bayesian Network models

Bayesian Network-based impact duration prediction models require more data than the data that is currently available in the TRANSCOM database. One significant feature of Bayesian network models is interpretability. An operator can obtain the significance of each variable in the prediction process. Demiroluk and Ozbay's model can automatically adapt itself to future conditions by learning the patterns of new incidents and their respective conditions. Their model is not only able to work with variables with missing values, but also provide a distribution of predicted impact duration. Their model provides relatively low accuracy due to the limited on-line real-time data without the use of adaptive learning structure. When more operations data becomes available, this model can, however, provide reasonably accurate predictions with the use of adaptive learning structure.

Table 6 Summary of Bayesian Network-based impact duration prediction models

Bayesian Network-based impact duration prediction models

Model	**Performance**	**TRANSCOM Compatibility**	**Highlights**
Ozbay and Noyan, 2006	Best correct classification rate: 80%.	Medium	Interpretable, capture stochasticity, sequential model, operations, reliable
Boyles et al, 2007	Best correct classification rate: 50%.	High	Interpretable, capture stochasticity, sequential model, operations, unreliable
Ji et al, 2008	Best correct classification rate: 74%.	Low	Deal with missing data, sequential model, not operations, reliable
Shen and Huang, 2011	Best correct classification rate: 72.6%.	Low	Interpretable, capture stochasticity, sequential model, not operations, reliable
Demiroluk and Ozbay, 2014	Best correct classification rate: 63.1%.	Medium	Interpretable, adaptive learning, real-time prediction, operations

1.5 Hazard-based impact duration prediction models

Various hazard-based models have been employed to predict impact duration. The log-logistic distribution was first used to describe freeway impact duration in the model of Jones et al., 1999. They adopted hazard-based regression to identify influencing factors of highway impact duration. However, their model used the same model and parameters during the whole process of highway incidents. Later on, Nam and Mannering, 2000 improved their model by introducing multiple stages (detection, response, and clearance) of traffic incidents and provided different models at different stages, respectively. They adopted hazard-based regression to identify influencing factors of traffic incidents at different stages. In this section, we will introduce another hazard-based model which employs hazard-based regression at multiple stages and provides real-time predictions for traffic incidents.

1.5.1 An information-based time-sequential approach to online impact duration prediction. Qi and Teng. (2008)

Qi and Teng (*23*) proposed a time-sequential procedure which can provide an online prediction of impact duration. The procedure contains multiple stages during the incident management process. For each stage, they applied a hazard-based duration regression model with different variables representing the available information. They used the remaining impact duration as the definition of impact duration and concluded that the accuracy of the prediction of impact duration increases as more information becomes available and then is incorporated into their models.

Their procedure for on-line incident prediction contains three stages; different hazard-based regression models are applied at each stage. Figure 17 shows the incident management process with different stages.

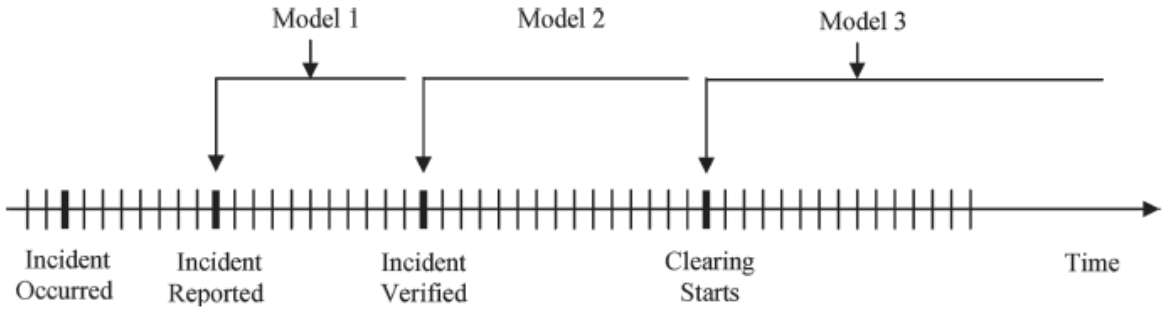


Figure 17 Time-sequential procedure for the prediction of remaining impact duration. Qi and Teng (*23*)

Stage 1: Started by an incident being just reported, an operator in a traffic management center may want to know how long it may take to clear an incident from a road. The input of Model 1 includes where the incident occurred, the weather, and the time when the incident happened.

Stage 2: Started with the verification of the incident. Model 2 receives extra information with the type of the incident and the types of vehicles involved in the incident.

Stage 3: Begins at the onset of the clearance of the incident, the operator may want to update the prediction of the time needed to clear the incident based on additional information such as

which agency and what facilities are involved in the clearance activities, such information are also inputs for Model 3.

The hazard function is written as:

$$h(t) = \lambda p(\lambda t)^{p-1}/[1 + (\lambda t)^{p}]$$

This function suggests that, if $p \leq 1$, the likelihood that an incident will end soon monotonically decreases with the length of impact duration. Otherwise (i.e., if $p \geq 1$), this likelihood will first increase from 0 to a maximum at a critical point $t = (p-1)^{\frac{1}{p}}/\lambda$ and then decrease.
The effect of external covariates, x_i, on impact duration can be incorporated by writing

$$\lambda = \exp(-\beta X)$$

The parameters of the probability distribution p and λ, and the coefficients of the duration model β, can be estimated by maximum likelihood estimation using the likelihood function:

$$\ln L = \sum_{i=1}^{n} h_0[t_i \exp(-\beta X_i)]\exp(-\beta X_i) + \sum_{i=1}^{n} S_0[t_i \exp(-\beta X_i)]$$

Where $S_0(t) = 1/(1 + (t)^p)$ and $h_0(t) = p(t)^{p-1}/[1 + (t)^p]$

Data needs

Traffic data: None
Incident data: Incident type, vehicle type, number of vehicles involved, severity, injuries and fatalities, property damage.
Operations data: Type of response agencies involved (Police, NYCDOT), whether tow truck is involved or not.
Time data: Time when incident is detected, time when incident is cleared, time of day, day of week.
Location data: Roadway type, number of lanes closed.
Weather data: Snow or clear, rain or dry

Model Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides real-time predictions. • Deals with time-sequential data. • Provides interpretable results and easy for operations use. • Provides better prediction with updated incoming data.	• The form of predicted distribution needs to be pre-determined.	• Better accuracy as more data coming into the model.

1.6 Support Vector Machine (SVM) based impact duration prediction models

1.6.1 A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. Yu et al. (2016)

Yu et al. (*24*) applied a comparative study of the performance of Artificial Neural Network (ANN) and support vector machine (SVM) for the prediction of traffic impact durations. SVM is a type of learning algorithms based on statistical learning theory, which can be adjusted to map the input-output relationship for the non-linear system.
An SVM estimator (f) on regression can be expressed as:

$$f(x) = w\,\phi(x) + b$$

Where ϕ denotes a nonlinear transfer function that maps the input vectors into a high-dimensional feature space in which the sample data are linearly separable.
With the induced loss function, the SVM estimator can be converted to an optimization problem:

$$R(a_i, a_i^*) = \sum_{i=1}^{n} (a_i - a_i^*) K(x, x_i) + b$$

Where $K(x, x_i)$ is the kernel function which maps the nonlinear regressors into linear regressors by adopting Lagrange multipliers. The structure of the SVM is shown in Figure 18.

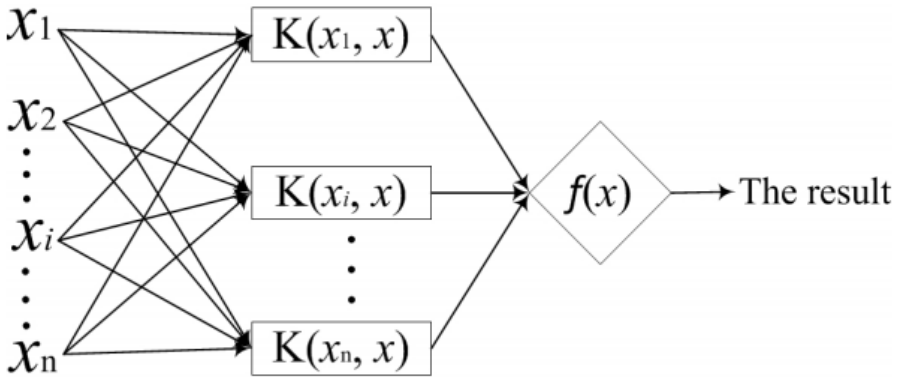


Figure 18 Structure of SVM. Yu et al. (*24*)

This study applied a K-means clustering method to select significant variables from incident dataset and input such variables into the SVM model in Figure 19.

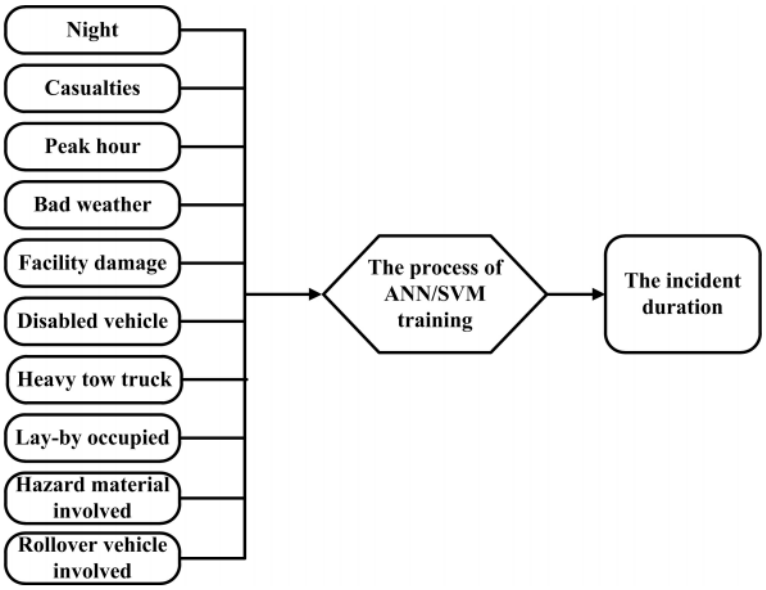


Figure 19 Structure of SVM for predicting the impact duration. Yu et al. (*24*)

Data needs

Traffic data: None.
Incident data: Incident type, vehicle type, severity, injuries and fatalities, property damage.
Operation data: None.
Time data: Time when incident is detected, time when incident is cleared, time of day, day of week.
Location data: None.
Weather data: Whether severe weather or not.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides interpretable results and easy to use operations.	• Cannot deal with time-sequential data. • Cannot provide real-time predictions. • Only tested with freeway accident data, it is, therefore, questionable for other real-world applications.	• Best MAPE: 19%.

1.6.2 Summary of hazard-based and SVM-based impact duration prediction models

For hazard-based impact duration prediction models, Qi and Teng et al. provided a three-stage model, which covered the time when an incident is reported, the time when the incident is confirmed/verified, and the time when the incident is cleared. This model can provide immediate duration prediction when an incident is first reported with limited available data. The model is also able to provide updated predicted duration when more information is coming in and provide better accuracy. This model is suitable for operations use and provides reliable results.

For supported vector machine (SVM), the selected model can provide reliable results. However, the model requires additional data that is not currently available in the TRANSCOM database. Moreover, the model was only tested and trained using freeway accident data, which may not be able to be compatible with other types of non-recurrent incidents in our study region.

Table 7 Summary of hazard and support vector machine (SVM) based models

Hazard-based model			
Model	**Performance**	**TRANSCOM data compatibility**	**Highlights**
Qi and Teng, 2008	Better accuracy with more data becoming available	High	Three-stage model, provide immediate and updated duration, operations, reliable
Support vector machine (SVM)			
Yu et al, 2016	Best MAPE: 19%	Low	Interpretability, one-time model, not operations, reliable

1.7 Estimation of incident recovery time

Incident recovery time refers to the time difference between the clearance of the incident and the time when the traffic flow conditions return to normal. The estimation of incident recovery time plays an important role at the operational level. When an incident is detected, operators need to know when the affected traffic flow will return to normal conditions. Usually, they regard the recovery time as an essential measure in their decision-making process. However, it is not easy to determine the incident recovery time using only travel time data since it is challenging to be sure that incident is the only reason for increased link travel times. Moreover, most impact duration prediction studies have so far ignored the problem of the prediction of the incident recovery time (*25*) and (*26*). In this study, we include an empirical method (*25*), which allows estimating incident recovery time based only on travel time data. The model is able to provide a reasonable estimation of incident recovery time by comparing the background travel time profile to the current travel time under incident conditions. Moreover, the model can capture the incident recovery time without any model assumptions and calibrations.

1.7.1 Empirical methods for estimating traffic incident recovery time. Zeng and Songchitruksa. (2010)

Zeng and Songchitruksa (*25*) adopted travel time for incidents and non-incidents to develop the model of estimating incident recovery time. Their proposed method uses percentile statistics to establish the background conditions that represent travelers' anticipation under incident-free conditions and then employs the concept of the difference in the travel time and information from the incident database to estimate traffic recovery time. Their proposed method involved four main steps:

1. Determine a background travel time profile.
2. Obtain a current travel time profile under incident conditions.
3. Estimate incident recovery time from the difference-in-travel-time profile.
4. Determine the reliability of the estimates.

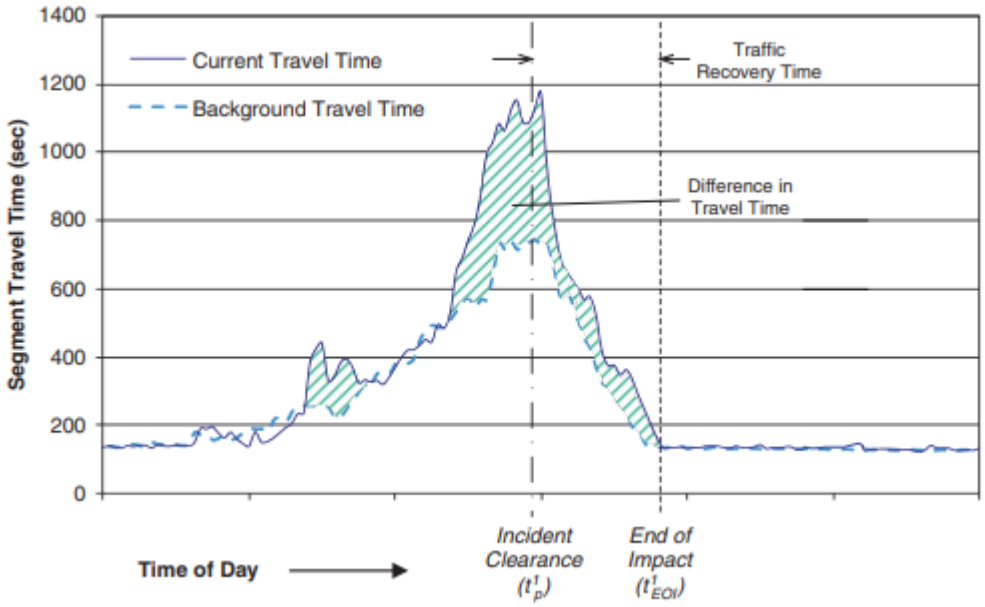


Figure 20. Travel time profiles for estimating traffic recovery time. (*25*)

They used a so-called "median-based profile approach" to determine the background travel time profile. The background profile should be constructed from the data that share common traffic patterns (e.g., same peak periods, the same day of the week, or weekdays versus weekends). For each of these profiles, the background travel time value (BTT) at the jth interval will be:

$$BTT_j = median\left(tt_{1j}, tt_{2j}, \dots, tt_{nj}\right) \; n \geq 3$$

Where n is the number of days considered in constructing the median profile.
They obtain the current travel time profile from both recurrent and incident-induced congestions. By superimposing the incident-affected travel time profile on the background profile, the difference-in-travel-time profile can be determined and used as a basis for estimating the traffic recovery time. The traffic recovery time can be expressed as:

$$TR^k = \begin{cases} t_{EOI}^k - t_p^k; TT_j - BTT_j > a \\ 0; \ TT_j - BTT_j \leq a \end{cases}$$

Where
$TR^k =$ traffic recovery time for incident k
$TT_j^k =$ travel time under the impact of incident k at time interval j
$t_p^k, t_{EOI}^k =$ incident clearance time (removal time) of incident k and time at the end of impact (EOI) respectively
$a =$ tolerance value that specifies the maximum difference between current travel time and background travel time before the end of the traffic recovery process can be specified.

Data needs

Traffic data: Travel time data (5 minutes aggregation).
Incident data: Incident type, vehicle type, number of vehicles involved, severity, injuries and fatalities, property damage.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared, time of day, day of week.
Location data: Number of lanes closed.
Weather data: Whether severe weather or not.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides incident recovery time estimation.	• May not represent actual background traffic conditions accurately (using median travel time profile only).	• Within +/- 10 minutes of median recovery time.

1.8 Data needs from reviewed models and their compatibility with TRANSCOM data

Below we provide a summary of data needs based on all the impact duration models reviewed versus available data from TRANSCOM. It is important to note that every model does not need all the data shown in Table 8. The team will make its final predictive model selection recommendation for the short-run based on the currently available data. Moreover, if a model is deemed promising but not recommended due to the immediate unavailability of data from TRANSCOM then it will be identified as a candidate model that can be tested in the mid-term contingent upon the availability of required data in the near future. For example, TRANSCOM communicated with the Team their plans for acquiring more incident response data in real-time such as number and types of response vehicles on-site and when and if this data becomes

available in the future, there will be an opportunity to test other suggested but not selected model(s) that will be included in the recommendations section.

Table 8 TRANSCOM data compatibility on reviewed duration prediction models

		TRANSCOM
Incident attributes	Incident type	●
	Impact duration	●
	Injury/fatality/property damage	Not currently available
Traffic attributes	Real-time traffic volume	Not currently available
	Traffic speed (before, during and after traffic incidents)	●
Time information	Response time	●
	Time first/last witnessed	●
	Time of police/tow truck arrival	
	Time of clearance	●
	Time of day	●
	Day of week	●
	Month of year	●
Geometry	Number of lanes affected	●
	Incident direction	●
	Which lane	●
	Left/right shoulder	
	Ramp/exit/corridor	●
Operation	Number of notifications sent	
	Workload of crew	
	Number of agencies involved	●
	Provision of traffic information to motorists	
Vehicle involvement	Number of vehicles involved	
	Number of trucks involved	
	Number of rescue vehicles/equipment used	
Weather	Rain/snow/sunny	●
Visibility	Dark/bright	

2. Traffic delay estimation/prediction

Traffic delay estimation/prediction is the second part of the predictive non-recurrent delay modeling methodology described in the beginning sections of the document, which mainly includes analytical and data-driven approaches. For analytical models, there are both deterministic and stochastic approaches. For data-driven models, methods including statistical regression, machine-learning techniques are reviewed in detail.

Impact duration is one of the critical inputs for traffic delay estimation/prediction models. This section first starts with the review of traffic delay estimation/prediction models for non-recurrent incidents as well as short-term work zone-related delay models for completeness purposes. The main reason for including work-zone literature is because a short-term work zone with mainly local impact is a particular type of non-recurrent incident where some of the theoretical model developments can also be applied to all non-recurrent incidents in general.

2.1 Analytical models for the estimation/prediction of traffic delay

2.1.1 Incident management integration tool: dynamically predicting impact durations, secondary incident occurrence, and incident delays. Khattak et al. (2012)

Khattak et al. (*27*) proposed a deterministic delay model that can deal with dynamic incident delay prediction. The main inputs to the delay prediction model are:
1. incident severity which is directly related to incident reduced capacity
2. impact duration, which affects the length of time it takes to clear the incident
3. arrival rate (traffic demand) and road geometry information such as the number of lanes.
Moreover, the predictive outputs of the model include total traffic delay and maximum queue length.
The calculation of queue length at a given time and the remaining total delays on a specified freeway segment are illustrated in Figure 21. Traffic arrives at the incident location according to curve $A_c(t)$. The departure curve $D_c(t)$ shows the departure from the incident bottleneck. The departure flow rate is initially μ^*, the reduced capacity of the bottleneck and then after the incident blockage is cleared at the time T_c, the capacity is restored at μ. The variables t_{n-1}, t_n represent the $(n-1)$th and nth time intervals from the incident start time – the time interval is set at 10min, representing the minimum period when a traffic arrival rate remains steady. The traffic arrival curve consists of a number of small time-dependent arrival rates. The current queue length for a given time t_i can be expressed as:

$$q(t_i) = q(t_{n-1}) + (t_i - t_{n-1})(\lambda_n - \mu^*) \quad for\ t_{n-1}, t_i < T_c$$
$$q(t_i) = q(t_{n-1}) + (t_i - t_{n-1})(\lambda_n - \mu) \quad for\ t_{n-1}, t_i < T_c$$

As long as all of the queue lengths for $t_i, t_n, \dots, t_e$ are calculated, the remaining total delay for a given time t_i is the shaded area between t_i and T_e, which is the summation of small trapeziums between arrival and departure curves right after t_i. The areas of the first three trapeziums can be written as:

$$A_1 = \frac{1}{2}\big(q(t_n) + q(t_i)\big) \times (t_n - t_i)$$

$$A_2 = \frac{1}{2}\left(q(t_{n+1}) + q(t_n)\right) \times (t_{n+1} - t_n)$$

$$A_3 = \frac{1}{2}\left(q(t_{n+2}) + q(t_{n+1})\right) \times (t_{n+2} - t_{n+1})$$

The remaining total delay at t_i is the sum of A_k, where $k = 1,2, \ldots$ represents the trapeziums.

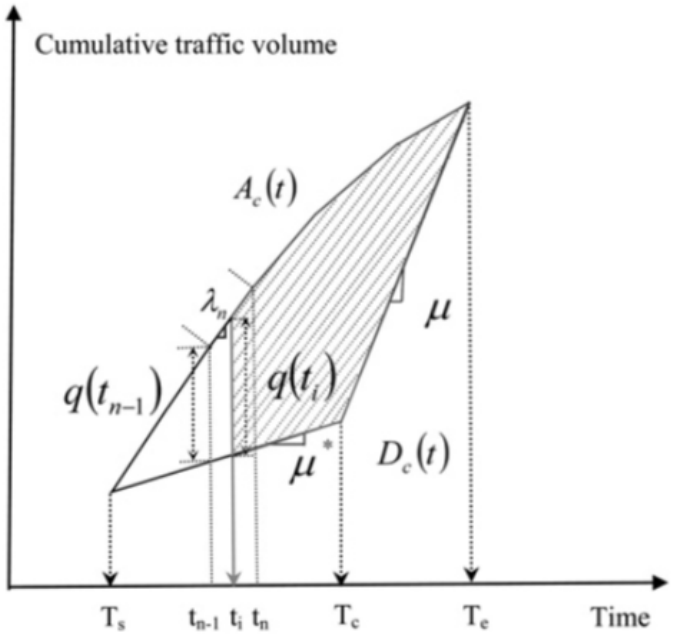


Figure 21 General deterministic queuing diagram of incident delay. Khattak et al (*27*)

Data needs

Traffic data: Traffic volume, roadway capacity.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week.
Location data: Total number of lanes, number of lanes closed.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimation.	• Deterministic model may overestimate or underestimate traffic delay. • Cannot provide travel time prediction.	• Not provided.

2.1.2 Estimation of incident delay and its uncertainty on freeway networks. Li et al. (2006)

Li et al (*28*) stated that traditional deterministic traffic delay estimation methods could not account for the stochastic attributes of dynamic traffic networks. They introduced a stochastic traffic delay model, which can calculate the variance and expected a total delay in dynamic networks. Their model was developed from the deterministic delay model and calculate the mean traffic delay in the same way as the deterministic model does. They incorporated the coefficient of variation of impact duration into the variance of delay and captured the total delay with its stochasticity.

The variance of delay function:

$$Var[d(t,r,s_1)] = \left\{\frac{\left[(q-\bar{s_1})^2+\sigma_{s_1}^2\right](1+x^2)}{3} - \frac{(q-\bar{s_1})^2}{4q^2}\right\}\bar{r}^2$$

The expected total delay function:

$$E[TD(t,r,s_1)] = \frac{\left[\left(\bar{s_1}^2+\sigma_{s_1}^2\right)-(s+q)\bar{s_1}+sq\right](1+x^2)\bar{r}^2}{2(s-q)}$$

Where

$s =$ freeway capacity, which is also the departure rate after the incident

$s_1 =$ reduced freeway capacity during the incident

$q =$ traffic flow rate

r = impact duration

$t_c =$ congestion clearance time

$x = \frac{\sigma_r}{\bar{r}}$ is the coefficient of variation of impact duration

Data needs

Traffic data: Traffic volume, roadway capacity.

Incident data: Impact duration.

Operations data: None.

Time data: Time of day, day of week, time when incident is cleared.

Location data: Length of affected roadway.

Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimation. • The stochastic model provides a distribution of traffic delay.	• Cannot provide real-time travel time prediction. • Low compatibility with TRANSCOM data.	• Not provided.

2.1.3 Proposed model for predicting motorist delays at two-lane highway work zones. Cassidy and Han. (1993)

One particularly important factor that directly affects delay and queue length is the length of the work zone. Cassidy and Han (*29*) estimate delay and queue length as a function of work zone length.
The model splits the work zone delay into two delay components, one is queueing delay, and the other is a travel time delay. The definition of queueing delay is that once cycle length and effective green and red times are computed, queueing delays and queue lengths can be determined using queueing theory. The definition of travel time delay is the difference between the actual average travel times through the work zone and the average travel times without the work zone.

Data needs

Traffic data: Traffic volume, roadway capacity, saturation headway, start-up lost time, ending lost time, travel time, traffic speed.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week.
Location data: Length of affected roadway.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimations.	• Deterministic model, may overestimate or underestimate traffic delay. • Cannot provide real-time travel time prediction. • Low compatibility with TRANSCOM data.	• Not provided.

2.1.4 Traffic characteristics and estimation of traffic delays and user costs at Indiana freeway work zones. Jiang. (1999)

Jiang (*30*) estimated work zone delays under several different categories: vehicle deceleration before entering work zones, moving delays experienced by vehicles passing through work zones at lower speeds, acceleration delays experienced by vehicles accelerating after existing work zones, and queuing delays caused by the ratio of vehicle arrival to discharge rates.
Note: this model applied M/M/1 queueing theory to calculate the length of the queue when the traffic flow rate is below the work zone capacity. Vehicles may arrive at a Poisson distribution and exponentially distributed through the work zone.

Logistics of the model:

$$DELAY_I = F_{aI}\,[d_d + d_z + d_a + (1 - t_I)d_w] + D_I$$

Where,

F_{aI} = hourly volume of arrival vehicles at hour i

d_d = delay due to vehicle deceleration before entering the work zone

d_z = delay due to reduced speed through the work zone

d_a = delay due for resuming freeway speed after exiting the work zone

d_w = delay due to vehicle queues during uncongested traffic

D_I = delay due to vehicle queues during congested traffic

$$QUEUE\ LENGTH = Q_0 + \sum_{i=1}^{m} F_{ai} - mF_d$$

Where,

Q_0 = original vehicle queue

F_{ai}= hourly volume of arrival vehicles at hour i

F_d= vehicle queue discharge rate

Data needs

Traffic data: Traffic volume, roadway capacity, travel time, traffic speed, vehicle acceleration rate.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week.
Location data: Length of affected roadway.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimations.	• Deterministic model, may overestimate or underestimate traffic delay. • Built for work zone delay only, additional efforts may need for adapting other incident types.	• Not provided.

	• Cannot provide real-time travel time prediction. • Low compatibility with TRANSCOM data.	

2.1.5 Optimal work zone lengths for four-lane highways. Chien and Schonfel. (2001)

The main objective of Chien and Schonfel's (*31*) study is to optimize the work zone length when the traffic flow rate is lower than the work zone capacity. When the traffic flow rate is lower than the work zone capacity, they also proposed a model to estimate the queue delay.
Logistics of the model:

$$t_q = \frac{1}{2}\left(1 + \frac{Q - c_w}{(c_0 - Q)}\right)(Q - c_w)(z_3 + z_4 L)^2$$

Where,
c_w = the work zone capacity
c_0 = roadway capacity in normal conditions
Q = approaching traffic flow
z_3 = the work zone setup time
z_4 = the additional time required per work zone kilometer
L = the work zone length

Data needs

Traffic data: Traffic volume, roadway capacity, travel time, traffic speed, vehicle saturation headway.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week, fixed setup time of work zone.
Location data: Length of work zone.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimations.	• Deterministic model, may overestimate or underestimate traffic delay. • Built for one-lane closure work zone delay only, additional efforts may need for adapting other incident types.	• Not provided.

	• Cannot provide real-time travel time prediction. • Low compatibility with TRANSCOM data.	

2.1.6 Freeway work zone traffic delay and cost optimization model. Jiang and Adeli. (2003)

Jiang and Adeli (*32*) proposed a deterministic queuing model for both short term and long term work zones based on average hourly traffic flow. The model splits the total delay into two parts: upstream queue delay time (t_q) and the moving delay time (t_m).
Logistics of the model:

$$t_d = t_q + t_m = \sum_{t=t_i}^{t_i+D-1} \left(\frac{T_t + T_{t+\Delta t}}{2}\Delta t\right) + \sum_{t=t_i}^{t_i+D-1} \Delta t_m = \sum_{t=t_i}^{t_i+D-1} \left(\frac{T_t + T_{t+\Delta t}}{2}\Delta t + \Delta t_m\right)$$

$$T_{t+\Delta t} = \max\{T_t - s, 0\}$$

$$\begin{cases} s = c_w - \alpha_s f_{\Delta t} & Long\ term\ work\ zone \\ s = c_0 - \alpha_s f_{\Delta t} & Short\ term\ work\ zone \end{cases}$$

Where,
t_d= total queueing delay
t_i = the starting time at the work zone in hours ranging from 1 to 24
D = the time period required to complete the maintenance for the work zone
Δt = the given time period
T = the cumulative number of vehicles
Δt_m = the moving delay time
c_w = work zone capacity
c_0 = freeway capacity without work zone
α_s = seasonal demand factor

Data needs

Traffic data: Traffic volume, roadway capacity, travel time, traffic speed, vehicle saturation headway.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week, fixed setup time of work zone, seasonal demand factor.
Location data: Length of work zone.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use.	• Deterministic model, may overestimate or	• Not provided.

• Provides real-time delay estimations. • Provides both short-term and long-term delay estimations.	underestimate traffic delay. • Built for work zone delay only, additional efforts may need for adapting other incident types. • Cannot provide real-time travel time prediction. • Low compatibility with TRANSCOM data. • Can only provide hourly delay prediction.	

2.1.7 Methodology for computing delay and user costs in work zones. Chitturi et al. (2008)

Chitturi et al. (*33*) proposed a step-by-step methodology to estimate capacity, queue length, and delay at work zones. By applying with the lane width factor, heavy vehicle factor, and PCE values from HCM, they estimated the adjusted capacity of the work zone.
Logistics of the model:

$$d_{total} = d_q + d_{spd} = \sum_{i=0}^{t-1}\left(\frac{n_i + n_{i+1}}{2}\right) + \sum_{i} V_i * (\frac{L}{U_0} - \frac{L}{U_{lim}})$$

$$n_{i+1} = n_i + V_{i+1} - C_{adj} * N_{op}$$

Where,
d_{total}= total delay with work zone
d_q= delay due to queueing
d_{spd}= delay due to the slower speed
n_i=number of vehicles in the queue at hour i
L= length of the work zone
V_i= demand flow rate in hour i
U_0= operating speed
U_{lim}= posted speed limit inside the work zone
C_{adj}= adjusted work zone capacity
N_{op} = number of lanes opened at the work zone

Data needs

Traffic data: Traffic volume, roadway capacity, travel time, traffic speed, speed limit, vehicle saturation headway.
Incident data: Impact duration.

Operations data: None.
Time data: Time of day, day of week.
Location data: Length of work zone.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimation. • Provides both short-term and long-term delay estimation.	• Deterministic model, may overestimate or underestimate traffic delay. • Built for work zone delay only, additional efforts may need for adapting other incident types. • Cannot provide real-time travel time predictions. • Low compatibility with TRANSCOM data. • Can only provide hourly delay predictions.	• Not provided.

2.1.8 Methodology to analyze queue length and delay in work zones. Ramezani and Benehokal. (2011)

Ramezani and Benehokal (*34*) proposed that there may be more than one bottlenecks in a single workspace and/or the transition area (within the single work zone). When the traffic flow rate exceeds the transition area and work zone capacity, there will be active bottlenecks not only in the workspace but also in the transition area. When the traffic flow rate is less than capacity, there will be only one bottleneck throughout the work zone.
The model calculated the queueing delay by setting up multiple volume conditions among demand, transition capacity, and workspace capacity. When estimating the queue length, the model induced shockwave theory to calculate shockwave speed and arriving volume minute-by-minute.

Data needs
Traffic data: Traffic volume, roadway capacity, travel time, traffic speed, speed limit, vehicle saturation headway.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week.

Location data: Length of work zone.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time delay estimations. • Provides both short-term and long-term delay estimations. • Provides 1-, 3-, 5-min delay estimations.	• Deterministic model, may overestimate or underestimate traffic delay. • Built for work zone delay only, additional efforts may need for adapting other incident types. • Cannot provide real-time travel time predictions. • Low compatibility with TRANSCOM data.	• Not provided.

2.1.9 Theoretical approach to predicting traffic queues at short-term work zones on high-volume roadways in urban areas. Ullman and Dudek. (2003)

The model is designed for estimating queue length of the short-term work zone on urban highways. Ullman and Dudek (*35*) had a concern that the current models have an overestimation of queue length due to the assumption of far apart on and off-ramp. Instead, drivers may choose alternative routes to avoid work zone area if they can live in urban highways since the distance between on and off-ramp is usually short. Therefore, they applied macroscopic fluid-flow theory to estimate the queue length of work zones on urban highways.
Logistics of the model:

$$q_{side(1)} = KiA = K' \frac{\Delta_{\overline{p_1}}}{TE} \Delta x_i$$

Where,
$q_{side(1)}$ = the flow permeating out the sides of the pipe through each segment (VPH)
Δp1 = the average traffic stream pressure differential between the roadway and the rest of the corridor within Δx1
A = area through which flow is occurring
K = coefficient of permeability
TE = total energy of the traffic stream, and
i = energy gradient across the permeable medium

Data needs

Traffic data: Traffic volume, roadway capacity, travel time, traffic speed, speed limit, vehicle saturation headway.
Incident data: Impact duration.
Operations data: None.
Time data: Time of day, day of week.
Location data: Length of work zone.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Stochastic model. • Provides real-time delay estimations. • Provides both short-term and long-term delay estimations.	• Built for work zone delay only, additional efforts may need for adapting other incident types. • Cannot provide real-time travel time predictions. • Low compatibility with TRANSCOM data.	• Not provided.

2.1.10 Summary of analytical models for traffic delay estimation/prediction

This section summarized available analytical methods for traffic delay estimation/prediction. Both deterministic and stochastic models for work zone and other general non-recurrent incidents were introduced. For deterministic models, one significant advantage is that they can provide average traffic delay and queue lengths fast. However, all deterministic models suffer from the problem of overestimation. Stochastic models can predict expected total delay and variance of total delay. However, due to low compatibility of their data needs with TRANSCOM's database, they may not be well suited to be used as part of TRANSCOM's operators. Table 9 shows a summary of all analytical models of delay estimation/prediction that are reviewed so far.

Table 9 Summary of analytical models of delay estimation/prediction

Analytical models for delay estimation

Model	TRANSCOM Compatibility	Highlights
Khattak et al, 2012	Low	Deterministic model, average delay, not operations, overestimated
Li et al, 2006	Low	Stochastic model, the variance of delay, not operations, reliable
Cassidy and Han, 1993	Medium	Work zone related, deterministic model, can only model one lane closure event, not operations, overestimated
Jiang, 1999	Medium	Work zone related, deterministic model, average delay, not operations, overestimated
Chien and Schonfel, 2001	Low	Work zone related, deterministic model, average delay, not operations, overestimated
Jiang and Adeli,2003	Low	Work zone related, deterministic model, good for both long and short term work zone, not operations, overestimated
Chitturi et al, 2008	Medium	Work zone related, can use sensor data, good for both long and short term work zone impact assessment, reliable
Ramezani and Benehokal, 2011	Low	Work zone related, deterministic model, similar to Chitturi but was modeled with 5-min aggregated traffic data (volume)
Ullman and Dudek, 2003	Low	Work zone related, stochastic model, shock-wave analysis, not operations, reliable

2.1.11 Traffic incident management decision support tools for planning purposes

In this section, we introduce several decision support tools that are used for work zone analysis and mainly for planning purposes. These decision tools are developed using deterministic models mentioned above.

QuickZone.

QuickZone (*36*) is the most used software packages for estimation of queue lengths and delays in work zones. It is a work zone delay estimation program developed in Microsoft Excel. The primary functions of QuickZone include quantification of corridor delay resulting from capacity decreases in work zones, identification of delay impacts of alternative project phasing plans, supporting tradeoff analyses between construction costs and delay costs, examination of impacts of construction staging, by location along mainline, time of day (peak vs. off-peak) or season, and assessment of travel demand measures and other delay mitigation strategies. QuickZone can provide estimation/prediction of traffic delay and queue lengths.

RILCA.

One drawback of QuickZone is that users need to input highly detailed traffic data to analyze the impacts of long-term lane closures. This drawback can be addressed by RILCA (Rutgers Interactive Lane Closure Application) (*37*), an interactive computer tool to plan lane closures for work zones. It is a tool that was developed with the ArcView geographic information system (GIS) software package as the main development environment. One of its features is that RILCA gives users the flexibility to export the corresponding traffic volume from RILCA data to QuickZone, if a detailed long-term lane closure analysis is required. It reduces the effort of inputting detailed data into QuickZone for long-term work zone analysis.

Work Zone Coordination tool

Work Zone Coordination tool (*38*) is an online tool that can evaluate the feasibility and effectiveness of coordinating short- and long-term work zones and to measure the benefits. It integrates all scheduled and active construction projects, identifies conflicts between work zone projects. It provides the estimation of traffic delay and queue length using deterministic queuing model.

There are other decision support tools embedded with deterministic models, such as LCAP (Lane Closure Analysis Program) (*39*) adopted by Maryland State Highway Administration. There are also other work zone related studies (*40*) (*26*) that provide models to quantify the traffic impacts of work zones or estimate the reduced capacity caused by work zones. For example, Bian and Ozbay (*41*) proposed an artificial neural network model to estimate the uncertainty of work zone capacity and provide its predicted distribution.

2.2 Data-driven methods for estimating/predicting impacts of non-recurrent traffic events

Different from the analytical/statistical methods mentioned in the previous sections, this section will provide data-driven approaches that provide traffic impacts of non-recurrent traffic events by learning from the speed profiles with and without non-recurrent traffic events. Given the availability of specific data from TRANSCOM, this type of models can be the most practical ones for TRANSCOM's operations needs.

2.2.1 Estimating magnitude and duration of incident delays. Garib et al. (1997)

Garib et al. (*6*) proposed a multivariate regression model that is based on predictors such as the number of lanes affected, the number of vehicles involved and the impact duration. They proposed two models that best predicted the incident delay based on their available data:
Model 1:

$$Delay = -4.26 + 9.71X_1X_2 + 0.5X_1X_3 + 0.003X_2X_4 + 0.0006X_3^2$$

Model 2:

$$Delay = -0.288 + 3.8X_1X_2 + 0.51X_1X_3 + 0.06X_3 + 0.356X_2^3$$

Where,

$Delay =$ cumulative incident delay
$X_1 =$ number of lanes affected by the incident
$X_2 =$ number of vehicles involved in the incident
$X_3 =$ impact duration (the difference between the incident detection time and the incident clearance time)
$X_4 =$ traffic demand upstream of the incident in the last 15 minutes before the incident starting time

Data needs

Traffic data: Traffic volume, reduced roadway capacity.
Incident data: Impact duration, number of vehicles involved, incident type, truck involvement.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Occurrence within bottleneck, number of segments upstream of the incident, number of lanes affected.
Weather data: Rain or dry.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Requires low computation effort.	• Cannot provide real-time travel time predictions.	• Prediction accuracy: 74%.

2.2.2 Modelling the impact of traffic incidents on travel time reliability. Hojati et al. (2016)

Hojati et al. (*42*) proposed a method to quantify the impacts of traffic incidents on travel time on freeways. They adopted historical data to establish recurrent speed profiles and identified non-recurrent congestion based on their negative impacts on speeds. The locations and times of incidents are used to identify incidents among non-recurrent congestion events.
They firstly defined the recurrent speed profile as a benchmark to quantify the impact of traffic incidents. Therefore, the extra travel time due to traffic incidents is calculated for each time interval as the difference between the recurrent speed profile and the daily speed profile.
The procedure can be described as follow:
1. Apply Quantum-Frequency Algorithm to identify recurrent speed profile (RSP_g) and daily speed profile (DSP_g). The difference between RSP_g moreover, DSP_g highlighted the impact of non-recurrent congestion events.
2. Non-recurrent events duration prediction.
3. Estimation of total travel time due to the non-recurrent event. The equation of getting estimated travel time over affected links is:

$$DTT_{t_k}^{i} = \sum_{m=1}^{j} dtt_i(g_m, \overline{tt_m}) = \sum_{m=1}^{j} \frac{l_{g_m}}{DSP_{g_m,d,\overline{tt_m}}}$$

Where,

$DTT_{t_k}^{i}$ = total travel time due to an event i on a set of affected links, in the time interval t_k, in hours (h)
$dtt_i(g_m, \overline{ti_m})$ = travel time due to an event i on link g, mth affected link, in the time interval $\overline{ti_m}$, in hours (h)
$DSP_{g_j,d,\overline{ti_m}}$ = speed on link g, mth affected link, day d, in the time interval $\overline{ti_m}$
l_{g_m} = link length of link g, mth affected link, (km)
t_k = time interval of an event $t_k \in \{t_s, \dots t_e\}$
m = set of affected links of an event $m \in \{g_1, .. g_j\}$
$\overline{ti_m}$ = time interval of jth affected link based on departing the event at the time interval t_k, $t_k - [[\sum_{m=g_2}^{g_j} dtt(g_{m-1}, \overline{ti_{g_{m-1}}})] \times \alpha]$, α is an aggregated factor, the default value is 12

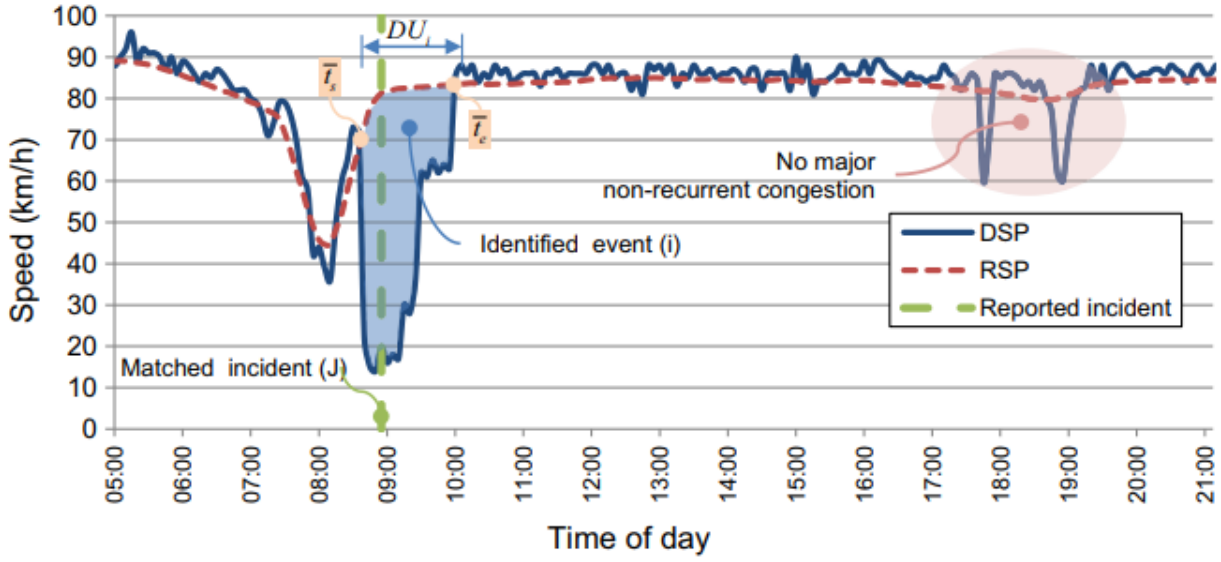


Figure 22 Schematic event identification in a typical day. Hojati et al. (*42*)

Data needs

Traffic data: Traffic volume, traffic speed.
Incident data: Impact duration, incident type.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Link segment location, incident direction, segment length.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use. • Provides travel time estimations.	• Cannot make real-time predictions, the travel time can only be predicted after the clearance of incident.	• Not given.

	• Cannot provide a range of predicted travel time. • Cannot provide queue length estimations.	

2.2.3 A framework for travel time variability analysis using urban traffic incident data. Javid et al. (2018)

This study (*43*) developed a framework to estimate travel time variability caused by traffic incidents using integrated traffic, road geometry, incident, and weather data. They adopted a two-year data in the California highway system to develop robust regression models. Their models estimate highway clearance time, which shares the same definition of recovery time defined in Section 1.7.1. Their models also estimate speed changes in percentages in both upstream and downstream links of the incident bottleneck. Based on their proposed speed change models, they estimated travel time variability due to non-recurrent incidents. Such travel time variability can be regarded as one measurement to quantify the impact of non-recurrent incidents.

Their methodology for speed change model is relatively easy, they employed a method called Iteratively Reweighted Least Squares (IRLS) to implement robust regressions.

$$\hat{\beta}_{IRLS} = \arg\min_{\beta} \sum_{i=1}^{n} \omega_i r_i^2(\beta)$$

$$\omega_i = \frac{\rho(\frac{r_i}{\sigma})}{r_i^2}$$

$$\rho(u) = \begin{cases} 1 - \left\{1 - \left(\frac{u}{4.685}\right)^2\right\}^3 & if\ |u| \le 4.685 \\ 1 \quad if\ |u| > 4.685 \end{cases}$$

Where,

ω_i= weight for observation i

σ = standard deviation of the residuals

ρ= loss function

The equations above will be iteratively implemented in a step-wise algorithm and stop until the maximum changes in weights is less than 95%.

The regression function for predicting speed changes and highway clearance time is listed as below:

$$y_i = \beta x_i + \sum \gamma_i + \sum \gamma_i \gamma_j$$
$$i, j = 1, \dots, n$$

The descriptions of variables in the model are shown in Table 10.

Table 10 Descriptions of variables in model.

Variable in the model	Description

y_1	Percent reduction in speed
y_2	Highway clearance time (min)
x	Incident clearance time (min)
γ_1	Whether the incident occurred over the weekend or not (1 or 0)
γ_2	Whether the incident occurred during peak hours or not (1 or 0)
γ_3	Whether all lanes are engaged or not (1 or 0)
γ_4	The highway has 12 ft width and 18 ft shoulders or not (1 or 0)
γ_{ij}	Interaction variable of γ_i and γ_j

Data needs

Traffic data: Traffic speed.
Incident data: Impact duration, incident type.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared, time of day, day of week.
Location data: Link segment location, incident direction, segment length, segment width, shoulder width.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use.	• Cannot provide queue length estimations.	• Low performance R^2: 0.33

2.2.4 Estimating freeway route travel time distributions with consideration to time-of-day, inclement weather, and traffic incidents. Caceres et al. (2016)

This paper (*44*) developed a probabilistic model for estimating route travel time variability with consideration of factors like time-of-day, inclement weather, and traffic incidents. They applied Monte Carlo simulation to estimate the total travel time from origin to destination by creating condition probability function for each link travel time.
Their diagram for modeling link and route travel time is shown below.

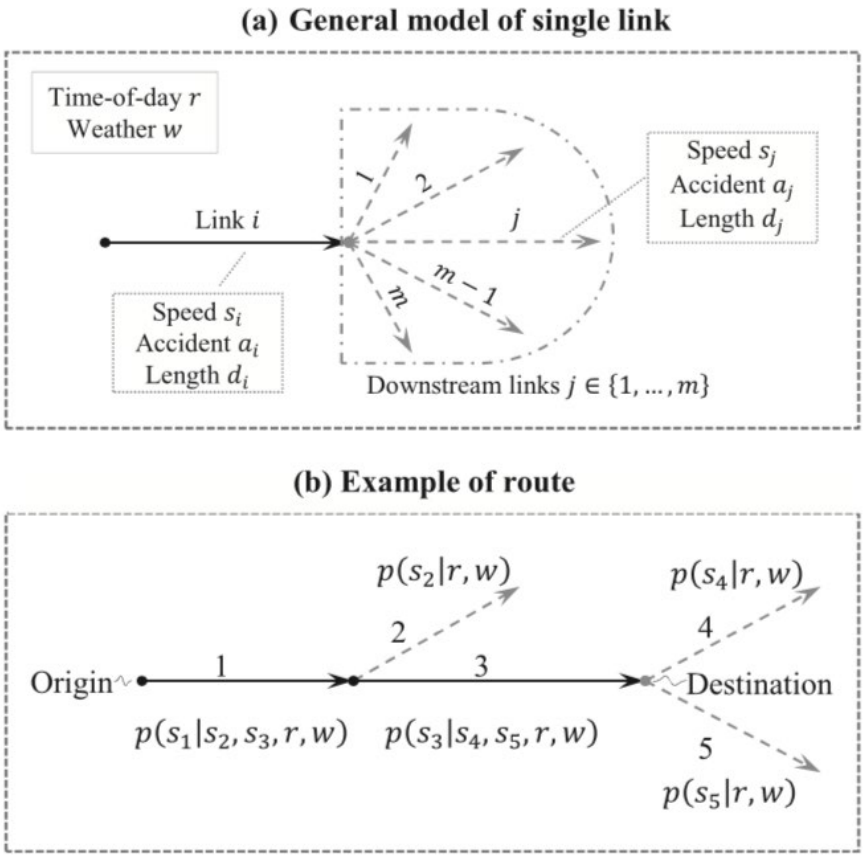


Figure 23 Diagram for modeling link and route travel time. (a) the general model of a single link. (b). example of the route (*44*).

One significant shortcoming of this model is that all variables are defined in a discretized way. Let T_i represent travel time in link i, where $T_i: \Omega \rightarrow \mathbb{R}^+$.

$$S_i: \Omega \rightarrow \{5, 15, 25, 35, 45, 55, 65\}$$
$$R: \Omega \rightarrow \{\text{Peak}, \text{Off} - \text{Peak}, \text{Weekend}\}$$
$$W: \Omega \rightarrow \{\text{Clear}, \text{Moderate}, \text{Rain}, \text{Snow}\}$$
$$A_i: \Omega \rightarrow \{\text{None}, \text{Accident}\}$$

Where,
S_i = speed in link i
R = time-of-day
W = weather condition
A_i = incident present in link i

They use probability mass function (pmf) to obtain the probability distribution for total travel time of a route.

$$p(s_i|s_1, \dots, s_m, r, w) = \sum_{a_i} p_1(s_i|a_i, s_1, \dots, s_m, r, w)\, p_2(a_i|s_1, \dots, s_m, r, w)$$

Estimating p_1 and p_2 directly from the data requires having observations for all levels of s_i for each combination of $(a_i, s_1, \dots, s_m, r, w)$ and observations for all levels of a_i for each combination of $(s_1, \dots, s_m, r, w)$.
The dependency of all the pmfs needed to find the probability distribution of the speed of link i is shown as below.

Conditional *pmf* for the speed of link i	Simplification		To be obtained from data
$p(s_i\|s_1, \dots, s_m, r, w)$	$p(s_i\|a_i, s_1, \dots, s_m, r, w)$	$p(s_i\|a_i, r, w)$	$p(s_i\|a_i, r)$ $p(s_i\|a_i, w)$ $p(s_i\|a_i)$
		$p(s_j\|a_i, s_i, r, w)$	$p(s_j\|a_i, s_i, r)$ $p(s_j\|a_i, s_i, w)$ $p(s_j\|a_i, s_i)$
		$p(s_j\|a_i, r, w)$	$p(s_j\|a_i, r)$ $p(s_j\|a_i, w)$ $p(s_j\|a_i)$
	$p(a_i\|s_1, \dots, s_m, r, w)$	$p(a_i\|r, w)$	$p(a_i\|r)$ $p(a_i\|w)$ $p(a_i)$
		$p(s_j\|a_i, r, w)$	$p(s_j\|a_i, r)$ $p(s_j\|a_i, w)$ $p(s_j\|a_i)$
		$p(s_j\|r, w)$	$p(s_j\|r)$ $p(s_j\|w)$ $p(s_j)$
4th level	3rd level	2nd level	1st level

Note that links $j \in \{1, \dots, m\}$ are downstream links of link i.

Figure 24 Variable indication of pmf derivations (*44*).

When constructing each pmf level by level (from level 1 to level 4), a recursive probability tree is built to derive the conditional pmf for the speed of link i with the consideration of multiple combinations of time-of-day, weather, and incidents.

Data needs

Traffic data: Traffic speed.
Incident data: Impact duration, incident type.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Link segment location, incident direction, segment length.
Weather data: Clear, moderate, rain or snow.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use.	• Cannot provide queue length estimations.	• Best KS difference: 0.175, p-value: 0.573.

• Provides probabilistic distribution of travel time.	• Only considered the occurrence of the incident, without taking other incident attributes.	

2.2.5 Predicting the spatial impact of planned special events. Martino et al. (2019)

This study (*45*) proposed a model to quantify planned special events (PSE) such as concerts, soccer games, and so on. They employed a K-Nearest Neighbor (KNN) classifier and Dynamic Time Warp (DTW) to predict the spatial impact of PSE. By training traffic data of event and non-event days for each road, using DTW, this model identified all road segments around a venue that show a different traffic behavior on event days than non-event days.
Their approach of identifying road segments that are potentially affected by PSE is by comparing events on different time-spans. They introduced the definition of Relative Timespan of Interest and cited an approach to identify any non-recurring influencing factor perturbing the flow on a specific date. The queue length can, therefore, be provided by adding up the total numbers of affected road segments.
They applied a binary classification approach to identify the road segments affected by a PSE. In order to search for correlations between Non-Recurring Traffic and the presence of an event, they employed a binary classifier to discriminate road segments of the dataset into positive (Non-Recurring Traffic) and negative (no abnormal traffic behavior) classes.

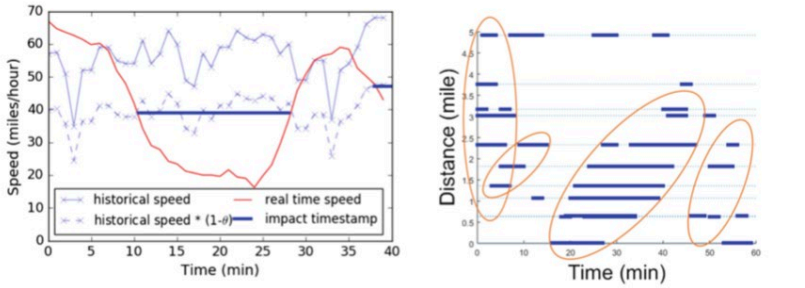
Data needs
Traffic data: Traffic speed.
Incident data: Impact duration, information about planned special events.
Operations data: None.
Time data: Planned special events schedules and time.
Location data: A description of the road network.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use. • Provides queue length predictions.	• Cannot provide travel time predictions. • Model-based on planned special events, additional efforts may need for adapting to other types of incidents.	• Best F-measure: 0.97.

2.2.6 Traffic accident detection with spatiotemporal impact measurement. Yue et al. (2018)

This study (*46*) adopted Impact Interval Grouping (IIG) to capture the spatiotemporal impact of traffic accidents to upstream locations. IIG compares real-time traffic speed with historical data and generates impact intervals to determine the presence of accidents. They then take a

multivariate time series classification approach to extract features to quantify the impact of traffic accidents.
This study used a 2-norm multi-dimensional dynamic time warping (WD-DTW) as the baseline model.
This study compared real-time speed with historical speed to quantify the incident impact. The historical speed is calculated using the average speed at the same location and same time. The unusual speed drop is modeled discretely by extracting impact intervals. Through such discretization, they converted the complex time series into a concise formulation which is easier to model as shown below.



(a) Generate Impact Intervals (b) Impact Interval Groups

Figure 25 Impact intervals and impact interval groups of an incident (*46*).

The definition of impact interval is a tuple (t_s, t_e), $t_s \le t \le t_e$, $\frac{|x(t)-\bar{x}(t)|}{\bar{x}(t)} \ge \theta$. Here $x(t)$ denotes the real-time speed at time t, and $\bar{x}(t)$ denotes the historical average speed of the same sensor, at time t. θ is a tuning parameter determining how strict the impact is measured. The IIG procedure includes three steps, 1) Discretization, 2) Smoothing, 3) Grouping. With the implementation of IIG, three features will be extracted and calculated to capture the impact of traffic incidents.
Dropping severity λ: the drops in traffic speed. Given a multivariate time series, $X = \{x_1, x_2, \dots x_k\}$, the historical speed is denoted as $\bar{X} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k\}$. The dropping severity is measured as:

$$\lambda_{max} = \max_{i,k}(1 - x_k(i)/\bar{x}_k(i))$$
$$\lambda_{avg} = avg_{i,k}(1 - x_k(i)/\bar{x}_k(i))$$

Lasting severity τ: after an accident happens, the drop in speed will last for a certain time. This term is defined similarly as impact duration. Impact interval is used to measure lasting severity because the discretization provides an easy extraction of temporal patterns. A list of impact intervals I_k is generated. $|x_k|$ denotes the length of time series x_k. The lasting severity is measured as:

$$\tau_{max} = \max_{i,k}(I_k(i)[1] - I_k(i)[0])/|x_k|$$
$$\tau_{avg} = avg_k(\max_i(I_k(i)[1] - I_k(i)[0]))/|x_k|$$

Distant severity σ: an accident will affect a certain distance in the upstream traffic. This term can be used as the queue length of an incident. The distant severity is measured based on the existence of impact intervals. d_k denotes the distance of the sensor s_k.

$$\sigma_{cons} = d_k / d_K$$
$$k = \arg\max_k \{I_1 \; to \; I_k \neq \emptyset\}$$
$$\sigma_{disc} = d_k / d_K$$
$$k = \arg\max_k \{I_k \neq \emptyset\}$$

Data needs

Traffic data: Traffic speed.
Incident data: Impact duration, incident type.
Operations data: None.
Time data: Time when incident is detected, time when incident is cleared.
Location data: Link segment location, incident direction, segment length.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy to use. • Detects traffic accidents.	• Cannot provide travel time predictions. • Cannot provide queue length estimations.	• Not given.

2.2.7 Utilizing real-world transportation data for accurate traffic prediction. Pan et al. (2012)

Pan et al. (*47*) adopted an enhanced ARIMA (auto-regressive integrated moving average) to predict traffic. They proposed a method to predict traffic by incorporating historical and real-time data into time-series mining technique. The first method used H-ARIMA approach, which utilizes both historical traffic patterns and current traffic speed for traffic prediction under normal conditions and the presence of traffic incidents.
The method is a hybrid forecasting model named Historical ARIMA (H-ARIMA) that selects in real-time between ARIMA or HAM (Historical Average Model) based on their accuracy.

ARIMA: this model is a generalization of the autoregressive moving average model with an initial differencing step applied to remove the non-stationary of the data. The model is formulated as:

$$Y_{t+1} = \sum_{i=1}^{p} \alpha_i Y_{t-i+1} + \sum_{i=1}^{q} \beta_i \epsilon_{t-i+1} + \epsilon_{t+1}$$

Where $\{Y_t\}$ refers to time-series data (e.g., the sequence of speed readings). In the autoregressive component of this model ($\sum_{i=1}^{p} \alpha_i Y_{t-i+1}$), a linear weighted combination of previous data is calculated, where p refers to the order of this model and α_i refers to the weight of $(t-i+1)$-th reading. In the second part ($\sum_{i=1}^{q} \beta_i \epsilon_{t-i+1}$), the sum of weighted noise from the moving average model is calculated, where ϵ denotes the noise, q refers to its order and β_i represents the weight of $(t-i+1)$-th noise.

Historical Average Model (HAM): they introduced HAM that uses the average of previous speed readings for the same time and location to forecast the future data. The HAM is formulated as:

$$v(t_{d,w}+h)=\frac{1}{|V(d,w)|}\sum_{s\in V(d,w)}v(s)$$

Where $V(d,w)$ refers to the subset of past observations that happened at the same time d on the same day w. Specifically, d captures the daily effects (i.e., the traffic observations at the same time of the day are correlated), while w captures the weekly effects (i.e., the traffic observations at the same day of the week are correlated). h refers to the prediction horizon (the time step in the future).

They proposed a decision-tree model that selects between ARIMA and HAM whichever reports a lower prediction error to forecast the speed at individual time stamps. In this model, the decision parameter and threshold are denoted as λ and ϕ. The detailed approach is shown in Figure 26.

Algorithm 1 Get $\lambda(\{v(j)\}$, d, w)

Output: λ

Let S = $\{V(\{\text{v(j)}\},d,w)\}$
Let Err_{ARIMA} = 0; $Err_{\text{HAM}} = 0$
Initialize ARIMA model with training dataset $\{v(j)\}$
v_{HAM}= Average(V{d,w});
for all $v_i \in S$ **do**
    v_{ARIMA} = ARIMA(i);
    Err_{ARIMA} = Err_{ARIMA}+ RMSE(v_i, v_{ARIMA});
    Err_{HAM} = Err_{HAM}+ RMSE(v_i, v_{HAM});
end for
λ = Err_{ARIMA} / (Err_{ARIMA}+Err_{HAM})
Return λ.

Figure 26 Algorithm of hybrid ARIMA and HAM (*47*).

Data needs

Traffic data: Traffic speed or travel time.
Incident data: Incident type.
Operations data: None.
Time data: Time of day, day of week, time when incident is detected.
Location data: Incident direction, incident location, number of affected lanes.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**

• Simple and easy for operations use. • Provides real-time travel time predictions. • Can be extended to predict travel time with the presence of incidents.	• Cannot provide queue length estimations. • Cannot provide a range of predicted travel time.	• Best MAPE: 80% (Incident condition).

2.2.8 Analysis and prediction of the queue length for non-recurring road incidents. Ghosh et al. (2017)

Ghosh et al. (*3*) combined incident records with traffic speed data from the expressways of Singapore to compute the queue length. They proposed a hybrid classification-regression model to predict the queue length of the incidents in real-time. Their model contains multiple stages. The first stage of the model is binary classifier. The second stage is activated if the queue length of an incident is predicted to be higher than a predetermined threshold value. The model will perform a regression analysis to predict the queue length of incidents for fine-tuning. The third stage of the model will evaluate the performance of different classification and regression methods based on accuracy.
Their definition of queue length is the spread of upstream congestion links from the incident location. For the prediction of the queue length, they employed three methods, classification and regression tree (CART), support vector machine (SVM) and Treebagger. The procedure of their queue length prediction model is shown below.

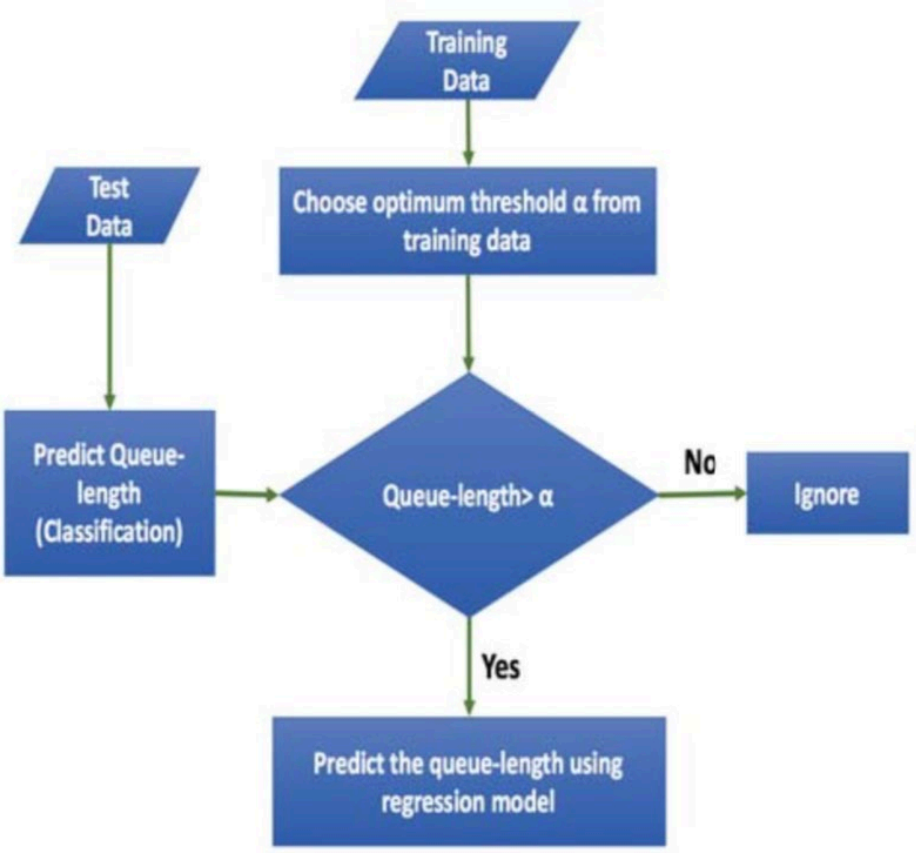


Figure 27 Flowchart of queue length prediction model (*3*).

Data needs

Traffic data: Traffic speed.
Incident data: Incident type, impact duration.
Operations data: None.
Time data: Time of the incident.
Location data: Incident direction, segment length, condition of the shoulder, total number of lanes, number of lanes affected, type of affected lanes (1st, 2nd or 3rd, from extreme right).
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time queue length predictions.	• Cannot provide travel time predictions.	• Best MAPE: 25% (Incident condition).

2.2.9 Real-time travel time prediction using particle filtering with a non-explicit state-transition model. Chen and Rakha (2014)

This paper (*48*) presents a methodology for a short to medium-term travel time prediction which is based on the real-time and historical traffic data collected. They proposed a new algorithm based on particle filter algorithm which selects particles from a historical database and propagates particles using historical data sequences as opposed to using a state-transition model. This particle method does not require an underlying physical model in order to model the state transition function but rather only depends on historical travel time trends. They apply a partial resampling method to address the degeneracy problem by replacing invalid or low weighted particles with historical data that provide similar data sequences to real-time traffic measurements.

For test cases and evaluation, they applied INRIX probe data to learn historical and real-time travel time trends. Their model shows an increased performance when compared to KNN and Kalman filters. The prediction horizon of their model is as far as 60 minutes (10 minutes time intervals).

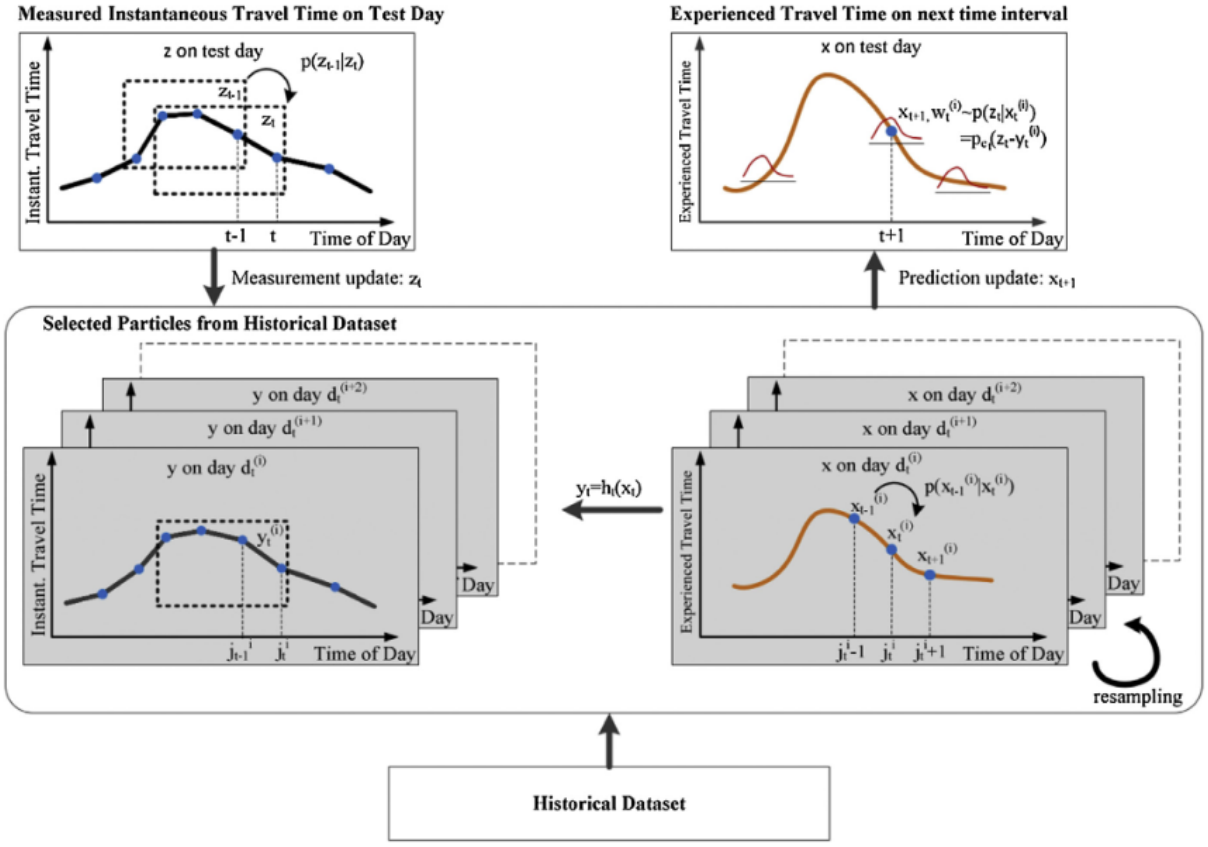


Figure 28 Demonstration of the proposed particle filter approach (*48*).

For their methodology, they used a graphical representation in Figure 28 to show their proposed approach: non-explicit state-transition particle filter (NSPF). The input data are the

measured instantaneous travel times for each time interval, the update of measurement data from z_{t-1} to z_t is conducted by shifting the data sequence window one-time step forward. Each particle can be recognized as a data sequence of instantaneous travel times and a data sequence of experienced travel times on the same historical day. The time update of the particle filter from x_{t-1}^i to x_t^i is accomplished by shifting one step ahead along the data sequence of experienced travel time. For each particle, the corresponding traffic pattern y_t^i can be derived according to the relationship with x_t^i represented by $y_t = h_t(x_t)$. At the same time, the associated weight w_t^i can be calculated as the likelihood $p(z_t|x_t^i)$, which can be accomplished by comparing the dissimilarity between real-time and historical traffic pattern as $p_{e_t}(z_t - y_t^i)$. The likelihood function is normal distribution $N(0,1)$. The distribution of experienced travel time on the next time interval $t+1$ can be predicted as $\{x_{t+1}^i, w_t^i\}_{i=1}^N$. For multi-step prediction with prediction horizon $t+p$, the predicted travel time is $\{x_{t+p}^i, w_t^i\}_{i=1}^N$. The proposed algorithm is shown in Figure 29.

$[\{x_t^{(i)}\}_{i=1}^N] = \text{NSPF}[\{x_{t-1}^{(i)}\}_{i=1}^N, z_t, \Omega]$

–Initialize particles x_0 : $\{x_0^{(i)} | x_0^{(i)} = \Omega_{\exp}(d_0^{(i)}, j_0^{(i)}), i \in [1:N]\}$

$d_0^{(i)}$ = randomly select a day from $[1, 2, \ldots, D]$, $j_0^{(i)}$ = randomly select a time index at day $d_0^{(i)}, i \in [1:N]$

Step 1: Time update

Propagate the particles by drawing $x_t^{(i)} \sim p(x_t^{(i)} | x_{t-1}^{(i)})$

$d_t^{(i)} = d_{t-1}^{(i)}, j_t^i = j_{t-1}^i + 1, i \in [1:N]$

Identify valid particles with respect to prediction horizon p

$\Psi_t = \{i | j_t^{(i)} \leqslant H_{d_t^{(i)}} - p, i \in [1:N]\}$

Step 2: Measurement update

$w_t^{(i)} \propto p(z_t | x_t^{(i)}) = p_{e_t}(z_t - y_t^{(i)}), i \in \Psi_t$

Select Nth number of particles with least weight values

For $j = 1$: Nth

$x_t^{(j)} = x_t^{(i)}, w_t^{(j)} = w_t^{(i)}$, when $i = \arg\max_{i \in \Psi_t} w_t^{(i)}, \Psi_t = \Psi_t - \{i\}$

End For

Step 3: Resampling

For $j = N\text{th} + 1 : N$

Calculate the probability of selecting each historical day λ_t^n

$\lambda_t^n = p_{e_t}(z_t - \Omega_{inst}(n, \sigma_t^n))$, when $\sigma_t^n = \arg\min_{k \in [1, H_n - p]} (z_t - \Omega_{inst}(n, k)), n \in [1:D]$

$d_t^{(j)}$ = randomly select a day from $[1, 2, \ldots, D]$ according to the probability $[\lambda_t^1, \lambda_t^2, \ldots, \lambda_t^D]$

$x_t^{(j)} = \Omega_{\exp}(d_t^{(j)}, \sigma_t^{d_t^{(j)}}), w_t^{(j)} = \lambda_t^{d_t^{(j)}}$

End For

Step 4: Prediction

Draw $x_{t+p}^{(i)} \sim p(x_{t+p}^{(i)} | x_t^{(i)}), i \in [1:N]$

$\overline{x_{t+p}} = \sum_{i=1}^N w_t^{(i)} \cdot x_{t+p}^{(i)} / \sum_{i=1}^N w_t^{(i)}$

Figure 29 Multi-step travel time prediction by NSPF (*48*).

Data needs

Traffic data: Travel time.
Incident data: None.
Operations data: None.
Time data: Time of day, day of week.
Location data: Segment location, segment length.
Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Simple and easy for operations use. • Provides real-time travel time predictions. • Provides a range of predicted travel time. • Fast computation time.	• Cannot provide queue length predictions. • No test cases for non-recurrent incidents, additional efforts may be needed for non-recurrent incident cases.	• Best MAPE: 7.32%.

2.2.10 Deep learning: a generic approach for extreme condition traffic forecasting. Rose Yu, et al. (2017)

This study (*2*) provides a deep learning method to predict traffic speed under non-recurrent congestion conditions. They propose two methods for peak-hour speed prediction and non-recurrent congestion prediction. They apply a deep long short term memory (LSTM) for peak hour traffic speed prediction. They further improve the model to predict traffic speed under non-recurrent congestion conditions through a mixture deep LSTM model. Their model is tested using traffic dataset in Los Angeles.

They mainly adopt LSTM model for peak-hour prediction and post-accident prediction. LSTM is a special type of Recurrent neural network (RNN). RNN is a feature map that contains at least one feedback loop. Denote the input vector at timestamp t as x_t, the hidden layer vector as h_t, the weight matrices as W_h and U_h, and the bias term as b_h. The output sequence o_t is a function over the current hidden state. RNN iteratively computes the hidden layer and outputs using the following recursive procedure:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h)$$

And

$$o_t = \sigma(W_o h_t + b_o)$$

Where W_o and b_o represent the weight and bias for the output respectively.

LSTM is a special type of RNN since it replaces the summation unit in RNN with memory cell state which contains gates to protect and control the cell state. In this way, LSTM avoids vanishing gradient issues and is able to model the long-term sequence problem.

Data needs

Traffic data: Traffic speed.
Incident data: Incident type.
Operations data: None.
Time data: Time of day, day of week.
Location data: Segment location, segment length, incident direction.

Weather data: None.

Highlights		
Advantages	**Disadvantages**	**Model performance**
• Provides accurate travel time prediction. • Designed specifically to predict traffic speed with/without non-recurrent incidents.	• Cannot provide queue length predictions. • Cannot provide a range of predicted travel time. • Requires heavy computation efforts.	• Best MAPE: 0.97%.

2.2.11 Summary of data-driven models for traffic delay estimation

In summary, with the investigation of available data-driven models for travel time and queue length prediction, we found that there are not any available models which can provide both travel time and queue length prediction with the presence of non-recurrent incidents. A future effort could determine if a model exists that would combine prediction models for impact duration, traffic delay, and queue length to satisfy the main objective of this project.

Table 11 Summary of data-driven models of delay estimation/prediction.

Data-driven models for delay estimation/prediction		
Model	**TRANSCOM data compatibility**	**Highlights**
Garib et al, 1997	Low	Statistical regression, operations, not reliable
Hojati et al, 2016	High	Able to provide travel time increase, cannot be used for prediction purpose, operations, reliable
Javid et al, 2018	High	Travel time prediction, statistical regression, operations, not reliable
Caceres et al, 2016	High	Travel time prediction, able to provide travel time distribution, can provide link and route travel time. Modeled with discretized data interval, operations.
Martino et al, 2019	High	Queue length prediction, machine learning model, designed for planned special events, reliable
Yue et al, 2018	High	Queue length identification, machine learning model, designed for incident detection

		purpose, not able to predict queue length, reliable
Pan et al, 2012	High	Travel time prediction, time-series method, real-time, operations, reliable
Ghosh et al, 2017	High	Queue length prediction, machine learning model, real-time, operations, reliable
Chen and Rakha, 2014	High	Travel time prediction, provide a range of predicted travel time, operations, reliable
Yu et al, 2017	High	Travel time prediction, deep learning model, high accuracy, reliable

2.3 Data needs from reviewed models and their compatibility with TRANSCOM data

Below we provide a summary of data needs based on all the incident delay estimation/prediction models reviewed versus available data from TRANSCOM. It is important to note that every model does not need all the data shown in Table 12. The team will make its final predictive model selection recommendation for the short-run based on the currently available data in addition model's predictive capabilities and accuracy. Moreover, if a model is deemed promising but not recommended due to the immediate unavailability of data from TRANSCOM then it will be identified as a candidate model that can be tested in the mid-term contingent upon the availability of required data in the near future.

Table 12 Data compatibility with TRANSCOM for traffic delay estimation/prediction

		TRANSCOM
Incident attributes	Incident type	●
	Impact duration	●
Traffic attributes	Real-time traffic volume	Not currently available.
	Traffic speed before, during and after traffic incidents	●
	Startup/end lost time	●
	Acceleration rate	●
Geometry	Number of lanes affected	●
	Incident direction	●
	Length of incident	
	Roadway capacity	●
Vehicle attributes	Number of vehicles involved	
	Number of trucks involved	
Weather	Rain/snow/sunny	●
Visibility	Dark/bright	

3. Data analysis towards estimating selected operations models

TRANSCOM provided the research team with three types of data from 2015 to 2018 in the ICM-495 corridor (Figure 1), which included highway events, highway trip, and HPMS volume data. There are 7 types of data files in highway events data, 2 types of files in highway trip data and 1 type of file in HPMS volume data. Table 13 shows the description of collected datasets from TRANSCOM.

Table 13. Description of data obtained from TRANSCOM.

Type	Details	Years	Export Type
Highway Events	Incidents	2015, 2016, 2017, 2018	CSV
	Construction	2015, 2016, 2017, 2018	CSV
	Special Event	2015, 2016, 2017, 2018	CSV
	Facility - Event Type Mapping		CSV
	Incident Type - Event Category Mapping		CSV
	Event - link ID mapping	2015, 2016, 2017, 2018	CSV
	Event Actions	2015, 2016, 2017, 2018	CSV
Highway Trip Data	Link travel time every 2 minutes by day of week for following		
	1. Monthly	2015, 2016, 2017, 2018	CSV
	2. Quarterly	2015, 2016, 2017, 2018	CSV
	3. Yearly	2015, 2016, 2017, 2018	CSV
	Link Definition including no. of lanes details		CSV
	Link shapefile		ESRI Shapefile
	Holiday Calendar	2015, 2016, 2017, 2018	CSV
HPMS Volume	AADT by link IDs	2017	CSV
	Hourly distribution factor		CSV

3.1 Highway events

Highway events dataset includes seven types of data files: Highway Events-Incidents, Highway Events-Construction, Highway Events-Special Events, Facility-Event Type Mapping, Incident Type-Event Category Mapping, Event-Link ID Mapping, Event Actions.

Highway Events-Incidents, Construction, Special Events

These files include 21,277 individual records of non-recurrent traffic events (5,265 incidents, 14,778 construction activities, and 1,234 special events) that occurred from 2015 to 2018, as shown in Figure 30. A total of 67 columns are included in each of these data files, including details of non-recurrent traffic events such as event type, start/end date time, direction of the event and so on. The purpose of this study is to provide traffic impact duration and delay/queue length predictions. After a detailed literature search and review, we selected the attributes that we thought would be useful for operation duration and impact model calibration and validation. Table 14 shows a description of these 20 selected attributes. The description of the entire 67 attributes is provided in Table 38 in Appendix.

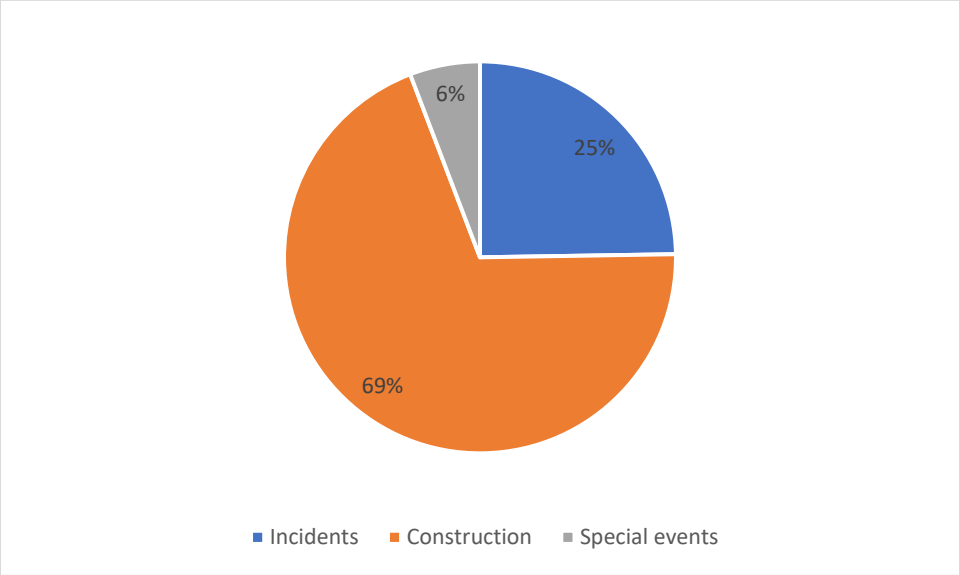


Figure 30. Type and percentage of events that occurred from 2015 to 2018.

Table 14. Description of selected data fields from highway events data files.

Sr. No.	Field Name	Description	Format	Missing Rate	Example
1	Id	Unique identifier of event	String	0.02%	ORI171242207
4	eventstatus	Status of an event List of values 0 - New 1 - Updated 2 - Closed 255 - Scheduled	Integer	0.02%	Closed

5	StartDateTime	Start date - time of event	Datetime	0.02%	12/25/17 1:39:00 AM
6	EndDateTime	End date - time of event	Datetime	0.03%	1/1/18 1:54:21 AM
7	LastUpdate	Last updated date-time of an event	Datetime	0.02%	12/25/17 1:39:40 AM
9	Organization_ShortName	Reporting organization Name	String	0.02%	MTA Bridges & Tunnels
10	eventType	Contains event type of current event	String	0.02%	Disabled vehicle
11	LanesTotalCount	Total lanes of roadway	Integer	99.29 %	4
12	LanesAffectedCount	Lanes affected by this event	Integer	95.15 %	2
13	LanesDetail	Contains lane affected detail. Example, all lanes at least one lane closed for repairs	String	45.29 %	right lane
14	LanesStatus	Contains lane status Example, open close traffic disruption	String	45.59 %	blocked
15	Facility	Event location facility name	String	0.02%	I-495
16	Direction	Contains Event direction	String	6.38%	westbound
17	City	City name based on Event	String	12.48 %	New York
18	County	County name based on Event	String	0.64%	Queens
19	State	State abbreviation of Event Example, NJ – New Jersey PA – Pennsylvania	String	0.02%	NY
27	PointDatum	Any reference point/co-ordinates from which measurement may be taken. Here the default Point Datum is NAD83(North American 1983 Datum)	Float	0.02%	NAD83
28	PointLAT	Latitude of an event	Float	0.02%	40.73690033
29	PointLON	Longitude of an event	Float	0.02%	-73.9312973
56	xcm_WeatherCondition	Weather condition during event	String	99.98 %	sunny

The data type and the percentage of missing values are obtained after conducting a complete process of raw data analysis and quality check of the provided files. In Table 14, it can be seen that most of the selected attributes report missing values less than 10%. However, attributes such as “LanesTotalCount”, “LanesAffectedCount” have more than 90% missing values. Furthermore, “LanesDetail” and “LanesStatus” are found to have 45% missing values.

Facility-Event Type Mapping

This file provides all highway facility names in the ICM-495 corridor that had non-recurrent traffic incidents from 2015 to 2018. Each event type is referenced with a single facility ID and facility name. Table 15 shows the field names with descriptions, data types and percentages of missing values. There are no reported missing values in this file.

Table 15. Facilities mapped into various event types.

Sr. No.	Field Name	Description	Format	Missing Rate	Example
1	EventTypeID	An ID referencing an IncidentType object, which describes the type of incident (e.g. accident, delay).	Integer	0%	1
2	EventType	A string representing the type(s) of event (e.g. "Accident, Delays").	String	0%	Construction
3	FacilityID	Event location facility ID	Integer	0%	109
4	Facility	Event location facility name	String	0%	I-495

Incident Type-Event Category Mapping

This file categorizes 141 event types into 5 distinct categories: congestion, construction, incident, special event and weather. For example, traffic events such as overhead sign repair, bridge rehabilitation and barrier repairs are categorized as “construction activities”. Traffic events such as disabled truck, accident and overturned tractor and trailer are categorized as “incidents”. Events like concert, parade and hockey game are categorized as “special events”. Events like downed tree and flooding are categorized as “severe weather events”. Table 16 shows the field names of this file with description, data type and percentage of missing values. There are no missing values reported in this file.

Table 16. Traffic event type mapped into 5 categories.

Sr. No.	Field Name	Description	Format	Missing Rate	Example
1	EventTypeID	An ID referencing an IncidentType object, which describes the type of incident (e.g. accident, delay).	Integer	0%	1

2	EventType	A string representing the type(s) of event (e.g. "Accident, Delays").	String	0%	paving operations
3	CategoryName	Event Category Example : Congestion Construction	String	0%	Construction

Event-Link ID Mapping

This file mapped each individual traffic event using a link ID of the location where the event has actually occurred. Using the link ID and provided coordinates, we are thus able to locate a specific traffic event in the ICM-495 corridor. With the help of the provided shapefile, we can find the upstream and downstream links of the target link where the event takes place. Table 17 shows the field names of this file with description, data type and percentage of missing values. There are no missing values found in this file.

Table 17. Individual traffic event mapped using a unique link ID.

Sr. No.	Field Name	Description	Format	Missing Rate	Example
1	TripMasterId	Unique ID of Trip	Integer	0%	3440
2	EventId	Unique identifier of event	Integer	0%	ORC14662711
3	LinkId	Native Link ID	Integer	0%	119959579

Event Actions

This file includes the operations actions for each individual traffic event from 2015 to 2018 by corresponding organization. Actions such as event opened, event updated and event closed are described in each traffic event record. The field name “TypeId” refers to a list of action types provided by TRANSCOM, which is shown in Table 39 in the Appendix. Table 18 shows the field names of this file with description, data type and percentage of missing values. There are no missing values reported in this file.

Table 18. Event actions for each individual traffic event.

Sr. No.	Field Name	Description	Format	Missing Rate	Example
1	EventID	An ID referencing the event to which event this action corresponds to. This should be used in conjunction with the "eventClass" field to	String	0%	ORI171242207

		determine the correct event.			
2	EventClass	Suggest a class of an event. Refer to the "Event Class" excel sheet tab.	Integer	0%	1
3	OccureAtTime	Represents date and time when this action was created	Datetime	0%	1/1/18 1:52
4	Description	Represents description of the action	String	0%	Closed via client action
5	TypeId	An integer representing the type of action.	Integer	0%	32
6	OrganizationId	An ID referencing the Organization that created this action. Refer to the "Org Name" excel sheet tab.	Integer	0%	1104

3.2 Highway trips

Highway trip data includes two types of data files: Link travel time and Link shapefile.

Link travel time

This file provides travel time data from 2015 to 2018 for each individual link in the ICM-495 corridor mapped with a unique link ID. The travel time data is provided every two-minute interval and aggregated by month, quarter and year. TRANSCOM also provides this real-time travel time in seconds. In this project, we will mainly use real-time link travel time data to calibrate and validate our recommended models.

This file maps link ID with link travel time. This link ID is also referenced in file “Event-Link ID Mapping” and Link shapefile. Therefore, we are able to obtain the link travel time before, during and after a specific traffic event. The link travel time for downstream and upstream of the link where the traffic event occurs are obtained after conducting a whole process of data analysis.

Table 19. Link travel time mapped using a unique link ID.

Details	Field Name	Description	Format	Missing Rate
Link travel time every 2 minutes by day of week	linkid	Unique Identifier for each link in ICM-495	Integer	0%
	hhmm	Hour : Minute	Datetime	0%

for following a. Monthly b. Quarterly c. Yearly d. Real-time (not averaged)	day0avgtt	Average TravelTime of all sunday of month/quarter/year in seconds	Integer	0%
	day1avgtt	Average TravelTime of all monday of month/quarter/year in seconds	Integer	0%
	day2avgtt	Average TravelTime of all tuesday of month/quarter/year in seconds	Integer	0%
	day3avgtt	Average TravelTime of all wednesday of month/quarter/year in seconds	Integer	0%
	day4avgtt	Average TravelTime of all thursday of month/quarter/year in seconds	Integer	0%
	day5avgtt	Average TravelTime of all friday of month/quarter/year in seconds	Integer	0%
	day6avgtt	Average TravelTime of all saturday of month/quarter/year in seconds	Integer	0%
	Real-time_tt	Real-time travel time of all 2-minute intervals in seconds	Integer	0%

Link shapefile

This file provides an ESRI shapefile that uses NAD83 coordinate system. This file provides a map of all links in the ICM-495 corridor in both directions. The attribute table of this shapefile includes details such as number of lanes, length of links and direction of the link. In the files "Highway Events-Incidents/Construction/Special Events", there are a lot of missing values (99.29%) in the field total number of lanes. This information is provided in the link shapefile. Table 20 shows detailed information that is provided in the link shapefile.

Table 20. Attribute table in the link shapefile.

Field Name	Description
link_id	Refers to link ID that maps with TMC
functionclass	The function class of specific link
highway_nm	The highway name where specific link belongs
length	The length of link in meters
postedspeedinmph	The posted speed of link in mph
direction	The facing direction of the link, i.e. North, East, South, West, North East, North West, South West and South East
state	The state where link belongs
county	The county where link belongs
ramp	Ramps are connectors that provide access between roads that do not cross at grade.

route_type	The route type indicates that the road's name is actually a route number and in many countries is displayed in a shield symbol (i.e. Interstate and State routes in the U.S.).
tollway	This attribute identifies a link for which a fee must be paid to use the road.
roundabout	A roundabout is a contiguous loop with consistent one-way traffic throughout the circle that controls the traffic flow from converging roads.
poing_desc	The location description
roadway	The roadway name
bridge	Bridge is a structure that allows a road, railroad, or walkway to pass over another road, railroad, water feature, or valley.
tunnel	Tunnel is a covered passageway through or under an obstruction.
phys_lanes	Physical Number of Lanes indicates the total number of all lanes on a link across all travel directions.

3.3 HPMS volume data

HPMS volume data includes AADT by link IDs and corresponding hourly distribution factors. We can convert AADT for each link to hourly traffic volume by each link using an hourly distribution factor. Table 21 shows a description of HPMS volume data.

Table 21. Description of HPMS volume data.

Details	Field Name	Description
AADT by link IDs	link_id	Unique ID of an link
	aadt	Annual Average Daily Traffic
	aadt_singl	Annual Average Daily Traffic for single-unit trucks and buses
	aadt_combi	Annual Average Daily Traffic for Combination Trucks
Hourly distribution factor	VPROFILE	Profilename
	hour	value 0 to 23
	pct_hrly	Hourly distributed percentage

4. Comparison of reviewed models and recommendations

After a detailed review of the open literature, it was not possible to identify non-recurrent delay prediction models that are currently used by operations staff at a TMC as part of their real-time incident management operations. It is important to note that there might be operational models that are embedded in the proprietary software used by some TMCs, but such implementations are not generally published in the open literature. Thus, it is not possible to identify these as part of a literature review such as the one conducted in this project. One alternative way to get this type of additional information is to conduct a nation-wide interview of all the major TMCs. However, this can be very time-consuming and expensive effort which is definitely beyond the scope of this limited study. Finally, any predictive model should be able to work with TRANSCOM data for it to be appropriate for deployment by TRANSCOM and this requirement further limits the possibility of using off-the-shelf existing predictive models. For example, many existing models require real-time traffic volume as one of the critical inputs; the lack of traffic volume data in TRANSCOM data limits the use of many existing models.

In this section, based on our comprehensive review of the literature presented in the previous sections, we recommend one model for each type of prediction task namely, impact duration prediction, traffic delay prediction/estimation, and queue length prediction. We first summarize the feedback obtained from the interviewed stakeholders. Based on this feedback, we propose several performance measures to compare and evaluate reviewed prediction models. This section is concluded by comparing the data needs of recommended models with TRANSCOM data.

Based on the scope of work in this project, we identified that both travel time prediction and impact duration prediction can and should be done at the operations level. After interviewing TRANSCOM stakeholders, we compiled their responses and created a checklist of model requirements as shown in Table 22.

4.1 Ideal model vs. existing models

An ideal model should contain two components: the impact duration module, and travel time and queue length prediction module. The ideal model should be able to provide a prediction of these two components at the same time. Moreover, the ideal model should satisfy all of the requirements raised by the scope of work and stakeholders in our interviews, which are shown in Table 22.

For the impact duration prediction module, an ideal model should satisfy only 4 points shown in the checklist, namely: (1), (4), (6), (8) and (10).

For traffic delay and queue length prediction module, an ideal model should satisfy 8 points shown in the checklist: (1), (2), (3), (4), (5), (7), (8), (9) and (10).

Table 22 Checklist based on the scope of work (SOW) and interview feedbacks

Number	Requirements from SOW and interviews	Ideal	Possible
①	Both travel time prediction and impact duration prediction should be done at an operations level.	√	√
②	Provide travel time prediction at least for the impacted zone, then expand to a corridor, and/or alternative corridors with further effort.	√	√
③	Provide travel time prediction and parameters by vehicle type and by lane.	√	×
④	Should work with the current TRANSCOM dataset.	√	√
⑤	Provide travel time prediction with consideration of roadway closures.	√	√
⑥	Provide duration prediction for incidents that last more than 30 minutes. The model should also be able to predict incidents within 30 minutes, especially at peak hour/high demand routes.	√	√
⑦	Provide a range of predicted travel times instead of a single value. This predicted travel time should be updated every 5 minutes. It is better to also provide the distribution of predicted travel times with corresponding confidence levels.	√	√
⑧	The accuracy of the predicted travel time/impact duration within +/- 10% error. Stakeholders agree to sacrifice accuracy to get a longer prediction time window.	√	√
⑨	Provide real-time prediction of the queue length.	√	√
⑩	Can disseminate different levels of prediction information to different levels of agencies, stakeholder, decision-makers, and partners.	√	√

However, after a careful investigation of the literature, we find that a single model cannot address all of the above points included in the checklist at the same time. However, by combining several candidate approaches, one can manage to cover 9 out of 10 points in the checklist. Besides, there are no operations models reported in the literature that can predict travel time by vehicle type and by individual lane to the best of our knowledge. TRANSCOM data, as of its current version, does not have lane-based or vehicle-type-based travel time. Therefore, it is not possible to meet the third requirement with existing data sources and prediction models available in the literature.

We then compared reviewed models using the checklist in Table 22, and four other performance measures explained in the following section. After a detailed comparison, we recommend the most appropriate models for the prediction of impact duration, travel and

queue length in the presence of a non-recurrent event, given the availability of both historical and real-time TRANSCOM data.

4.2 Model comparison

Based on the scope of work of this project and the feedback from the stakeholder meetings, we developed four performance measures to further evaluate reviewed models.

4.2.1 Operations versus planning

One of the significant needs identified from user feedbacks is the requirement for the recommended model(s) to be for "operations" use only. An operations model should be able to provide predictions based on the limited information that becomes available during real-time incident management operations. The recommended model should be able to work in real-time and provide on-line predictions. For impact duration prediction, it should be able to use time-sequential data and provide updated predicted duration with new information becoming available as the incident management operations progress. However, many studies in the literature propose "one-time" models which can only be used mostly for planning purposes. These models can only work with historical data and provide impact duration prediction with complete data that can only be available after the full clearance of an incident. In summary, the first performance function in terms of recommending a model in this study is that it should be a model specifically suited for real-time "operations".

4.2.2 Prediction of a single value versus a range of values

Most stakeholders interviewed in the first task of this study mentioned a need for a range of predicted travel times rather than a single value. Non-recurrent congestion can cause significant interruptions to regular traffic patterns. Travel times in the presence of such non-recurrent congestion can thus fluctuate due to the stochasticity and possible modeling errors on a case by case basis. The probabilistic distribution of predictions can capture such randomness and uncertainties in a way a single point estimate cannot. This travel time prediction approach can also help agencies in the decision-making process by providing them with a range of values including, minimum, maximum, and average travel times. In fact, during the agency interviews, the team found out that many agencies prefer to disseminate different levels of predicted information (an upper and lower bound or an average expected travel time) within their agency, to their stakeholders and travelers depending upon such factors as their confidence in the model predictions and severity of the non-recurrent event. Thus, the second performance measure is that the recommended model should be able "to generate a range of predicted values rather than a single value to give the agencies flexibility in interpreting and disseminating results."

4.2.3 Analytical versus data-driven

Based on the findings of the extensive literature review, it is apparent that most of the non-recurrent delay models are analytical models. These analytical models include queuing-based

delay models or shock-wave based models. One common shortcoming of these models is that they cannot provide predictions with missing data/parameters. Moreover, analytical models cannot generally work in real-time due to extensive data input and model output analysis requirements that are not suitable for real-time operations. More importantly, all of these models require actual volumes and reduced capacities in order to predict delays. However, TRANSCOM currently does not acquire real-time volume data from most of its agencies and lack of real-time volume (demand) data makes all of these analytical models infeasible for operations use at this time. On the other hand, data-driven models can learn traffic patterns such as speed-profiles without knowing the details of non-recurrent traffic events as well as current traffic demand. The third performance measure is that the recommended predictive model should be "data-driven and should be able to be trained using currently available TRANSCOM data only."

4.2.4 Compatibility with TRANSCOM data

During the process of managing incidents in real-time, it is common that operators need to make decisions based on limited information. For impact duration prediction models, many of the reviewed models require data that is not currently available from TRANSCOM in real-time although some of it may become available after the incident is cleared. For example, most Classification Tree Method (CTM) models require operations data as key inputs, which is not provided in the TRANSCOM dataset. Therefore, we require our recommended model(s) to work with limited data, especially in real-time. Thus, the fourth performance measure is that the recommended model should be highly compatible with TRANSCOM data and be able to predict delay with some data missing.

4.2.5 Summary of model comparison

Impact duration prediction

For impact duration predictions, we compared models by checking if they meet the performance measures of (3.2.1) operations vs. planning, (3.2.2) prediction of a single value versus a range of values, (3.2.3) analytical versus data-driven and (3.2.4) compatibility with TRANSCOM data.

It is apparent from Table 23 that most regression models and classification tree methods suffer from low compatibility with TRANSCOM data. For Bayesian network models, although Ozbay and Noyan (*19*) and Demiroluk and Ozbay (*1*) are found to have medium compatibility with TRANSCOM data. However, due to the flexibility of Bayesian networks (BN), their model can provide reasonable impact duration predictions with limited data information and missing values. When more data becomes available, their model can update itself and provide updated and reliable predictions. Wei and Lee (*17*) achieved an accurate impact duration prediction through Artificial Neural Networks. Their model is highly compatible with the TRANSCOM dataset and can deal with sequential data. However, their model is computationally expensive and that may lead to a low prediction frequency and cannot be used for real-time operations. Khattak's model (*8*) reported a reasonable MAPE as 37%, and the model can deal with sequential data and is highly compatible with the TRANSCOM dataset. Since it is a regression

model, it requires low computational effort and can provide a prediction for less than 5 minutes. However, its actual accuracy for complex networks and traffic conditions is not tested using extensive field data. Qi and Teng (*23*) proposed a hazard-based method and reported to provide better accuracy as more data becomes available in a time sequence. Their model is highly compatible with the TRANSCOM dataset and requires low computational effort.

Table 23 Comparison results of impact duration prediction models.

	Model	Operations use?	Sequential/one-time model	TRANSCOM compatibility	Checklist –
Regression models	Khattak et al., 1995 (1.1.1)	Yes	Sequential	Low	①⑥
	Garib et al., 1997 (1.1.2)	No	One-time	Medium	④
	Peeta et al., 2000 (1.1.3)	No	One-time	Low	④
	Khattak et al., 2016 (1.1.4)	Yes	Sequential	High	①④⑥
	Yu and Xia, 2012 (1.1.5)	No	One-time	Low	⑥
	Weng et al., 2015 (1.1.6)	No	One-time	Low	⑥
Classification Tree Methods	Ozbay and Kachroo, 1999 (1.2.1)	Yes	Sequential	Low	①⑥
	Smith et al., 2002 (1.2.2)	No	Sequential	Low	⑥
	Knibbe et al., 2006 (1.2.3)	No	Sequential	Low	⑥

	He et al., 2013(1.2.4)	Yes	Sequential	Medium	①④⑥
	Zhan et al., 2011 (1.2.5)	Yes	Sequential	Low	①⑥
Artificial neural network	Wei and Lee, 2007 (1.3.1)	Yes	Sequential	High	①④⑥
	Park et al., 2016 (1.3.2)	Yes	Sequential	Medium	①④⑥
Bayesian networks	Ozbay and Noyan, 2006 (1.4.1)	Yes	Sequential	Medium	①④⑥
	Boyles et al., 2007 (1.4.2)	Yes	Sequential	High	①④⑥
	Ji et al., 2008 (1.4.3)	No	Sequential	Low	⑥
	Shen and Huang, 2011 (1.4.4)	No	Sequential	Low	⑥
	Demiroluk and Ozbay, 2014 (1.4.5)	Yes	Sequential	Medium	①④⑥⑧⑩
Hazard-based model	Qi and Teng, 2008 (1.5.1)	Yes	Sequential	High	①④⑥⑧
SVM	Yu et al., 2016 (1.6.1)	No	One-time	Low	⑥

Traffic delay estimation

For traffic delay estimation models, we compared models by checking if they meet the performance measures of (3.2.1) operations vs. planning, (3.2.2) prediction of a single value versus a range of values, (3.2.3) analytical versus data-driven and (3.2.4) compatibility with TRANSCOM data.

Different from analytical delay prediction models, data-driven models can learn the features from real-time data and generate reliable predictions as long as the models are well trained/customized. Hojati's model (*42*) can quantify the increase in travel time due to the occurrence of non-recurrent incidents by learning standard travel speed profiles and comparing them with incident-based travel speed profiles. However, Hojati's model cannot predict travel times in the presence of incidents. The hybrid ARIMA model proposed in (*47*) is able to predict travel times with particular timestamps after the occurrence of an incident. However, it cannot generate a range of predicted travel times. For queue length estimation, we identified three methods presented in (*45*), (*46*), and (*3*). All of them employed machine-learning techniques and can identify affected road segments due to the presence of an incident. However, certain important drawbacks exist in Martino's (*45*), and Yue's (*46*) model. Martino's model focused only on special events such as sports events and concerts. Substantial additional effort will be required if one wants to extend this model to other types of non-recurrent incidents. Yue's model is not directly used for prediction purposes. It is, however, possible to borrow the idea of defining the impacted roadway segments and adapt it to TRANSCOM's database to estimate new predictive models. Ghosh's (*3*) model is used directly for predicting queue length of non-recurrent incidents, and this model can be re-estimated with TRANSCOM data for operations use. Yu's model (*2*) can provide real-time travel time predictions after the occurrence of non-recurrent traffic incidents. Their model achieves the highest accuracy among all reviewed models for traffic delay estimation/prediction. One added advantage of this model is that it was trained and tested with 5-minute link travel time data which makes it promising in terms of compatibility of its findings given the similarity of TRANSCOM travel time database.

Table 24 Comparison of traffic delay estimation/prediction models.

Model	Operations use?	Analytical or data-driven	A range or single value	TRANSCOM Compatibility	Checklist
Khattak et al., 2012 (2.1.1)	Yes	Analytical	Single value	Low	①②⑤⑨
Li et al., 2006 (2.1.2)	Yes	Analytical	Range	Low	①②⑤⑦⑨
Cassidy and Han, 1993 (2.1.3)	Yes	Analytical	Single value	Medium	①②⑤⑨
Jiang, 1999 (2.1.4)	Yes	Analytical	Single value	Medium	①②⑤⑨
Chien and Schonfel, 2001 (2.1.5)	Yes	Analytical	Single value	Low	①②⑤⑨

Jiang and Adeli, 2003 (2.1.6)	Yes	Analytical	Single value	Low	①②⑤⑨
Chitturi et al., 2008 (2.1.7)	Yes	Analytical	Single value	Medium	①②⑤⑨
Ramezani and Benehokal, 2011 (2.1.8)	Yes	Analytical	Single value	Low	①②⑤⑨
Ullman and Dudek, 2003 (2.1.9)	No	Analytical	Single value	Low	①②⑤⑨
Garib et al., 1997 (2.2.1)	Yes	Data-driven	Single value	Low	①②⑤
Hojati et al., 2016 (2.2.2)	Yes	Data-driven	Single value	High	①②④⑤
Javid et al., 2018 (2.2.3)	Yes	Data-driven	Single value	High	①②④⑤
Caceres et al., 2016 (2.2.4)	Yes	Data-driven	Range	High	①②④⑤⑦⑩
Martino et al., 2019 (2.2.5)	No	Data-driven	Single value	High	④⑨
Yue et al., 2017 (2.2.6)	No	Data-driven	Single value	High	④⑨
Pan et al., 2012 (2.2.7)	Yes	Data-driven	Single value	High	①②④⑤
Ghosh et al., 2017 (2.2.8)	Yes	Data-driven	Single value	High	①④⑨
Chen and Rakha, 2014 (2.2.9)	Yes	Data-driven	Range	High	①②④⑤⑦⑧⑩

Yu et al., 2018 (2.2.10)	Yes	Data-driven	Single value	High	①②④⑤⑧⑩

4.3 Final model recommendation

During the literature search process, we aimed to find a model that can both provide real-time impact duration prediction and traffic delay prediction/estimation as well as queue length prediction. However, as a result of the detailed review of the existing literature, we found that no single model is able to predict impact duration, traffic delay, and queue length at the same time. Moreover, we could not find models that are currently used by operations staff for real-time operations. Therefore, we classified and then reviewed models for impact duration prediction, traffic delay prediction/estimation, and queue length prediction seperately.

After comparing models in detail in the light of the feedback obtained from interviews, we recommend three separate models. Specifically, we recommend one approach for impact duration prediction, one approach for incident delay estimation/prediction, and one for queue length prediction. Table 25 shows a summary of our recommended models.

The approach recommended for impact duration prediction is the Bayesian network approach proposed by Demiroluk and Ozbay (*1*) since it is the most appropriate model for use in real-time operations. Although the current TRANSCOM database does not contain all the needed data to calibrate the parameters of this model, to the best of our knowledge, several TRANSCOM agencies such as NJ Turnpike Authority and NY Thruway Authority collect the missing information. Thus, this model can be estimated in a limited fashion to test its accuracy and usefulness. This model can also deal with incident data becoming sequentially available during the incident management operation, have reasonable accuracy, and very low computational cost. Moreover, this model covers most of the requirements identified in the interview checklist shown in Table 22. For traffic delay estimation/prediction, we recommend Yu's model (*2*) due to its capability of online prediction and high prediction accuracy. Moreover, this model has automatic calibration which is convenient for re-calibration. Finally, for the queue length prediction, we recommend Ghosh's model (*3*) for predicting real-time queue length with reasonable accuracy using TRANSCOM's travel time data only.

Table 25 Summary of recommended models

Model	Operations use?	TRANSCOM Compatibility	%Checklist Satisfaction	Highlights
Demiroluk and Ozbay, 2014 (1.4.5)	Yes	Medium	100%	Bayesian network, interpretable, adaptive learning, real-time prediction, operations

Yu et al., 2018 (2.2.10)	Yes	High	71%	Time-series (RNN), travel time prediction, deep learning model, high accuracy, reliable
Ghosh et al., 2017 (2.2.8)	Yes	High	100%	Queue length prediction, machine learning model, real-time, operations, reliable

4.4 Comparison of TRANSCOM data and data needs of recommended models

In this section, we will compare TRANSCOM current data availability with the data needs of our recommended models.

Demiroluk and Ozbay's model (2014) (1)

Demiroluk and Ozbay's model (*1*) was developed with incident data obtained from transportation agencies in New Jersey. Incident and operations data such as the number of response agencies involved and the number of vehicles involved were used in the development of this model. No such information is currently available in the TRANSCOM dataset, as shown in Table 26. Therefore, we will not be able to calibrate this duration model using all of the variables employed in the original model. However, due to the flexibility of Bayesian networks (BN), we can remove unavailable variables in the TRANSCOM data and calibrate BN model using the available information. As shown in Table 26, both weather and pavement data in the TRANSCOM database have a lot of missing values (more than 99%). Therefore, we can calibrate the BN model using attributes/variables that have a small number of missing values, such as time data ("Month", "DayofWeek", "TimeofDay"), incident data ("CrshType") and location data ("Location", "Distance"). It is important to note that the shaded area in Table 26 represents the additional variables needed to be collected by TRANSCOM in the next stepto improve the capability of this model. With the collection of these variables, we can replicate the settings of Demiroluk and Ozbay's model (*1*) and provide reliable predictions. Their model can also provide duration prediction with limited data information. Therefore, at an early stage of a traffic incident, this model can work and provide a short-term prediction with missing data. The model becomes more accurate as it gets more data from the response team at the scene of an incident. This model can work with missing data and provide a predicted distribution of impact durations. Therefore, this model will provide more reliable prediction if and when more detailed incident information becomes available.

It is also important to note that this model can produce the prediction of incident clearance and incident recovery times. The incident clearance time can be determined directly by the start and end time of a reported incident. However, there is no direct way to determine the incident recovery time through the reported start and end time of a traffic incident. Instead, the recovery time needs to be estimated by comparing the travel time under incident conditions

and normal conditions. In this study, we reviewed and proposed a way of estimating incident recovery time using the approach described in Section 1.7.

Table 26 Detailed data needs from model and its compatibility with TRANSCOM data

	Variables	Description	TRANSCOM data	File name	Field name	Missing rate
Weather data	Weather	Weather conditions	●	Highway Events	xcm_WeatherCondition	99.98 %
Time data	Month	Month of year	●	Highway Events	StartDateTime / EndDateTime	0.02%
	DayofWeek	Day of week	●	Highway Events	StartDateTime / EndDateTime	0.02%
	TimeofDay	Time of day	●	Highway Events	StartDateTime / EndDateTime	0.02%
Incident data	NumFat	Number of fatalities				
	NumInj	Number of injuries				
	CrshType	Type of crash	●	Highway Events	eventType	0.02%
	VehNo	Number of vehicles involved				
	Roadwaydamage	Presence of roadway damage				
	NumTrkInv	Number of trucks involved				
Location data	Pavement	Pavement conditions	●	Highway Events	xcm_PavementCondition	100%

	Location	Link where incident is located	●	Event-Link ID mapping	LinkId	0%
	Distance	Distance from the closest exit	●	Highway Events	PointLAT/PointLON	0.02%
Light data	Light	Lighting conditions				

The shaded green area represents the data that is not currently collected or collected rarely by TRANSCOM and needs to be collected in the future to improve model estimation and prediction. "●" represents the data currently available in the TRANSCOM database.

Yu's model (2017) (2)

Yu et al (*2*) proposed two neural network models for travel time prediction. The format of the travel time dataset they used is similar to the TRANSCOM dataset's format. Their model requires time data ("Month", "DayofWeek", "TimeofDay"), incident data ("IncidentType"), location data ("Direction", "Location") and travel time data ("Travel Time") as input variables. Specifically, the model requires travel time data with 5-minute aggregation, TRANSCOM can provide travel time with aggregation as small as 2-minutes. As shown in Table 27, TRANSCOM provides available data for all these required variables with a small percentage (less than 1%) of missing values. Therefore, we will be able to calibrate and validate the model using the provided TRANSCOM data.

Training and calibrating this model requires relatively substantial computational resources. We recommend implementing this model if high-performance computing resources that exist at NYU are available for initial training, calibration, and testing. Please note that this is only required for initial calibration and validation and these models when successfully calibrated can be operationalized in a standard PC for day to day usage.

Table 27 Detailed data needs of Yu's model and its compatibility with TRANSCOM data

	Variables	Description	TRANSCOM data	File name	Field name	Missing rate
Time data	Month	Month of year	●	Link Travel Time	hhmm	0%
	DayofWeek	Day of week	●	Link Travel Time	hhmm	0%

	TimeofDay	Time of day	●	Link Travel Time	hhmm	0%
Incident data	IncidentType	Type of incident	●	Highway Events	eventType	0.02%
Location data	Direction	Incident direction	●	Highway Events	Direction	0.02%
	Location	Incident location	●	Event-Link ID mapping	LinkId	0%
Travel time data	Travel time	Link travel time (5 minutes aggregation)	●	Link Travel Time	Real-time_tt	0%

"●" represents the data currently available in the TRANSCOM database.

Ghosh's model (2017) (3)

Ghosh et al. (*3*) proposed a cascaded classification-regression model to predict the queue lengths. They adopted travel time data having a similar format to the format of TRANSCOM dataset. Their model requires time data ("Month", "DayofWeek", "TimeofDay"), incident data ("IncidentType"), location data ("Direction", "SegementLength", "Shoulder", "Total_Lanes", "Num_Lanes", "Type_Lanes") and travel time data ("Travel Time"). Table 28 shows a relatively high percentage of missing values (45.29%) in the data for variables "Shoulder", "Num_Lanes" and "Type_Lanes". Therefore, there is a potential need to remove data records with missing values and calibrate the model using the rest of the available data. Therefore, through proper data processing, we can calibrate and validate their model using available TRANSCOM data.

Table 28 Detailed data needs of Ghosh model and its compatibility with TRANSCOM data

	Variables	Description	TRANSCOM data	File name	Field name	Missing rate
Time data	Month	Month of year	●	Link Travel Time	hhmm	0%
	DayofWeek	Day of week	●	Link Travel Time	hhmm	0%
	TimeofDay	Time of day	●	Link Travel Time	hhmm	0%
Incident data	IncidentType	Type of incident	●	Highway Events	eventType	0.02%
Location data	Direction	Incident direction	●	Highway Events	Direction	0.02%

	SegmentLength	Length of segment	●	Link Shapefile	length	0%
	Shoulder	Whether the shoulder is involved	●	Highway Events	LanesDetail	45.29%
	Total_Lanes	Total number of lanes	●	Link Shapefile	phys_lanes	0%
	Num_Lanes	Number of affected lanes	●	Highway Events	LanesDetail	45.29%
	Type_Lanes	Type of affected lanes	●	Highway Events	LanesDetail	45.29%
Travel time data	Travel time	Link travel time (5 minutes aggregation)	●	Link Travel Time	Real-time_tt	0%

The shaded green area represents the data that is not currently collected or collected rarely for by TRANSCOM and needs to be collected in the future to improve model estimation and prediction. "●" represents the data currently available in the TRANSCOM database.

5. System requirements for an ideal predictive tool

Based on the assessment of the needs previously described in this report, it can be claimed that TRANSCOM and their member agencies must collaborate and adopt a computerized map-based/table-based tool or a data-feed service that is part of an implementation framework that can employ TRANSCOM's real-time data feed to provide operations personnel with predictive duration/delay/ queueing information in the presence of non-recurrent delays. This section lays out the desired functional requirements of this predictive framework that can be implemented as an operations software.

The main goal of this section is thus to clearly describe the functionalities of an "**ideal data-driven predictive non-recurrent duration/delay estimation framework**" based on the outcomes of the previous literature study. The framework presented in this section needs to be integrated into a map-based/table-based software tool or a data-feed service that incorporates all the functionalities that are deemed essential for the operations of non-recurrent traffic events.

Figure 31 illustrates the framework of an ideal computerized tool for helping operate non-recurrent traffic incidents/events. As seen in the figure, all current and historical traffic incident information and travel time data from TRANSCOM and other agencies are fed to an extended

traffic database for calibration and validation purposes. This extended database should be hosted at a server within the agency and automatically extracted to an online database via periodic XML feeds. The information within this online database can then be reached, queried, and used via a map-based/table-based software interface or a data-feed service. This map-based/table-based software interface or data-feed service can be built into TRANSCOM's software platform and accessed by only authorized users.

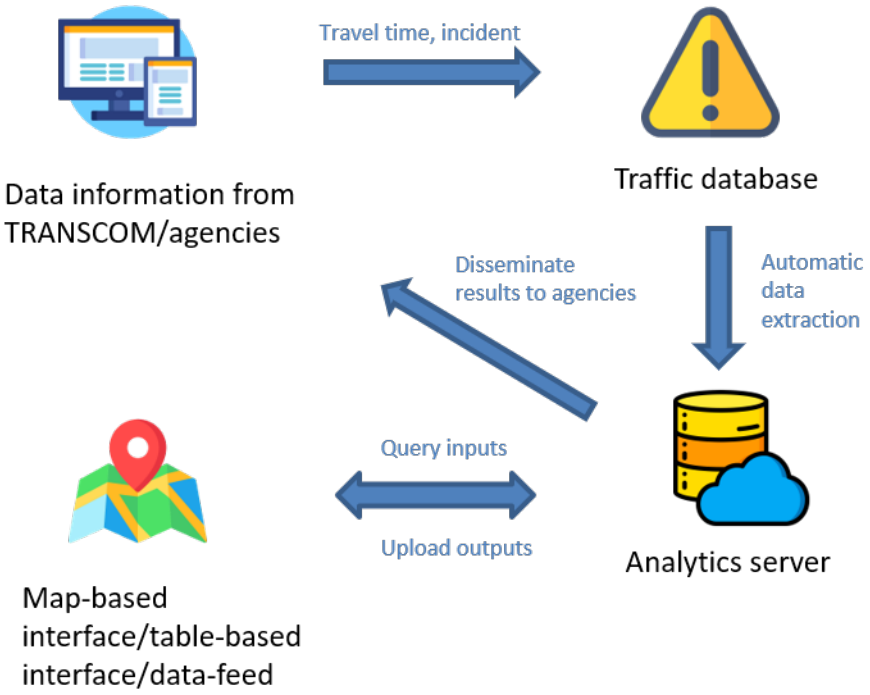


Figure 31. Framework of ideal predictive non-recurrent estimation delay tool.

The following subsections briefly outline the desired functionalities of the proposed ideal predictive non-recurrent estimation delay tool.

5.1. Maintain a database of non-recurrent events

This tool should be able to store and display detailed information about all historical and current non-recurrent traffic events (accident, construction event, special event). Information should include the event ID, description, event type, number of lanes affected, type of lanes affected, event start and end time, event location (coordinates, link ID), and the division of operations information.

This tool should provide a map-based/table-based interactive user-interface to allow the ease of entering input information and visualizing outputs for users. For example, on the front-end, when a traffic incident is reported, a traffic operator should be able to enter necessary information via a user-interface and visualize the location of this incident on the map. On the

back-end, the online traffic database should be able to query the inputs from operators and match the link ID of the incidents. The map-based user-interface should also display outputs such as estimated delay, duration and queue length on the map. This functionality provides operators with a good representation of predicted results and a better understanding of the impact caused by current traffic incidents.

This tool should also be able to store and update a certain period of recent real-time travel time information. For example, when a traffic incident is reported using the tool, the traffic database should automatically acquire a certain period of recent travel time data needed for prediction functions.

The traffic database should also automatically calculate and then disseminate the model predictions such as impact duration, delay and queue length to relevant stakeholders. Moreover, the database should also be designed in a way that it can easily be integrated with TRANSCOM's OpenReach database as well as its current user interface.

5.2. Traffic impact duration prediction

It is vital to provide a real-time prediction of traffic impact duration for a non-recurrent traffic incident. As outputs of this functionality, this tool must provide online and reliable predictions for the clearance and recovery times of a non-recurrent traffic event. When an incident is reported, this tool must provide a short-term prediction of incident duration with limited information such as incident type, start time and location. As the incident clearance operation progresses, additional information acquired by the operator will be entered and an updated predicted impact duration will be calculated and made available to traffic operators.

It should be noted that the successful implementation of the duration prediction functionality significantly depends on the available incident information. As mentioned in previous sections, the minimum required parameters for traffic impact duration prediction include weather conditions, time of the incident, incident location (coordinates, link ID) and incident type. These parameters are used to predict the duration of traffic incidents at a very early stage. In order to provide updated and more reliable predictions of traffic incident duration, this functionality requires additional data such as the existence of property damage, injuries and fatalities, existence and number of disabled vehicle, whether or not road repair/work is involved, the number of vehicles involved, number of lanes closed, whether or not a police department is involved and, whether or not a tow truck or fire truck is involved.

Therefore, this functionality requires the acquisition of real-time incident scene data that needs to be uploaded continuously to obtain updated prediction results. As per the interviews, the ideal tool requires this functionality to provide updated predictions every 5-minutes. Moreover, this functionality should provide an immediate update of its prediction when there are major changes in terms of the real-time incident information.

In summary, the outputs of this functionality include the initial predicted incident clearance time, updated predicted incident clearance time and recovery time. It is important to note that incident clearance time can be calculated directly through the reported start and end time. However, there is no direct way to calculate incident recovery time from the reported historical data because unlike incident clearance times, recovery times are not recorded. One alternative way to determine the recovery time is to compare travel time observed during the incident and normal traffic conditions (see Section 1.7) and try to indirectly identify incident recovery times (time to normal flow). Table 29 shows a summary of the system requirements of this functionality.

Table 29. Summary of system requirements for traffic impact duration prediction.

Model inputs	**Minimum:** weather conditions, day and time of the incident, incident location, incident type
	Ideal: existence of property damage, injuries and fatalities, existence and number of disabled vehicles involved, existence of road repair/ work zone, the number of vehicles involved, number of lanes closed, whether or not police department is involved, whether or not tow / fire truck is involved
Model outputs	Incident clearance time
	Incident recovery time*
Time from incident detection to the first prediction	5 minutes (immediate if major changes), work with limited information
Time interval to next updated prediction	5 minutes, update as soon as new and significantly different information becomes available

* Incident recovery time needs to be indirectly estimated through travel time data. Thus, it will be an approximation of the actual recovery time since ground truth data that can be used to validate the exact recovery time does not exist. A heuristic method to estimate recovery time from observed travel times and recorded clearance times need to be developed as part of this functionality. However, the details of this approach will require additional work that is beyond the scope of this work. It is also important to note that physical queueing models cannot be used due to the absence of volume data required by all the queuing models.

5.3 Traffic delay estimation/prediction

It is important to provide real-time estimation/prediction of the delay that would be caused by the non-recurrent traffic events. The outputs of this functionality include incident induced travel time prediction and queue length prediction.

It should be noted that the successful implementation of this functionality depends greatly on the available real-time link travel times as well as details of the incident. As mentioned in the previous subsection, this functionality requires recent real-time travel time data as the main input. For example, a successful prediction of one-hour after the occurrence of a traffic incident may require the past one-week travel time data. Such data should be automatically extracted from the historical traffic information database and updated in the online database.
The required parameters for travel time prediction include 5-minute link travel time data, incident type, day and time of the incident, incident direction and incident location (coordinates, link ID). These parameters are used to estimate travel time distribution after the detection of a traffic incident. For queue length estimation/prediction, the required parameters include day and time of the incident, incident type, incident direction, length of the segment where the incident occurred, whether or not shoulder is involved, number of lanes affected, type of lanes affected and 5-minute link travel times. It is important to note that this tool will only focus on predicting queues on highways only not on urban streets due to their inherent complexities.

In summary, the outputs of this functionality include predicted link travel time after the detection of a traffic incident, the average estimated traffic delay and queue length estimation for individual links. It should be noted that real-time prediction of travel time and queue length might require a processing time which may be up to 5 minutes. Users will obtain predictions once they input incident related data into the tool through the use of a map-based interface mentioned previously in this section.
Table 30 shows a summary of the system requirements for this functionality.

Table 30. Summary of system requirements for traffic delay estimation/prediction.

Model inputs	**Travel time prediction:** 5-minute link travel time data, incident type, day and time of the incident, incident direction and incident location
	Queue length estimation: day and time of the incident, incident type, incident direction, length of the segment where incident happens, whether or not shoulder is involved, number of lanes affected, type of lanes affected and 5-minute link travel time data
Model outputs	Post-incident travel time
	Average traffic delay
	Link-based queue length
Time from incident detection to the first prediction	**Travel time prediction**: up to 5 minutes. **Queue length estimation**: 5 minutes, work with limited information

Time interval to next updated prediction	5-minutes, update as soon as new and significantly different information becomes available

6. A preliminary assessment of development, calibration and implementation efforts required for recommended models

In Section 4.4, we compared TRANSCOM data with the data needs of our recommended models. With the comparison between TRANSCOM data and recommended data needs, we identified the efforts for data preparation, cleaning, and mining in order to calibrate each recommended model in this section. Moreover, we also identified the time and efforts for model development and training for each recommended model.
At the very end, users will have a clear understanding of the potential time and efforts for the development, calibration, and implementation of these recommended models.

Demiroluk and Ozbay's model (2014) (*1*)

In this subsection, we will describe the data preparation effort in order to calibrate and implement Demiroluk and Ozbay's model. Specifically, we include the efforts of collecting additional data compared to available TRANSCOM data, data processing, and model training. At the end of this subsection, we will also provide a rough estimated time of applying data preparation and calibration respectively.

Data preparation and calibration

The data preparation effort will start with additional data collection. As shown in Table 31, we carefully compared TRANSCOM data with the data needs of this recommended model and determined what additional variables are needed for calibrating this duration prediction model.

Table 31. Data collection required for additional variables needed by this model.

	Variables	Description
Incident data	NumFat	Number of fatalities
	NumInj	Number of injuries
	VehNo	Number of vehicles inolved
	Roadwaydamage	Presence of roadway damage
	NumTrkInv	Number of trucks involved
Light data	Light	Lighting conditions

Table 32 shows data processing and preparation effort required to calibrate Demiroluk and Ozbay's model. As shown in Table 32, the minimum required variables for calibration include weather conditions, day and time of the incident, incident location (coordinates, link ID) and the type of incident. The ideal variables required by the calibration include the existence of property damage, injuries, and fatalities, the number of vehicles involved, number of trucks involved, pavement conditions, distance from the closet exit, light conditions.

Table 32 also shows minimum amount of data required for calibration and implementation. Specifically, calibrating this model requires traffic incident data for a period of at least six months or more. To implement this model in real-world cases, traffic operators will need to input the required variables for the specific incident.

Table 32. Data preparation effort required for model calibration and implementation.

Minimum variables required for calibration	Weather conditions (snow, rain), time and day of incident, incident location (coordinates, link ID), incident type
Ideal variables required for calibration	Existence of property damage, injuries and fatalities, number of vehicles involved, number of trucks involved, pavement conditions, distance from the closet exit, light conditions
Minimum amount of data needed for calibration	6-months of traffic incident data
Minimum amount of data needed for real-world implementation	Details of the current traffic incident in real-time

As mentioned in the data analysis section (Section 3), a large amount of noise exists in the raw dataset. Therefore, some potential efforts of data processing are listed:

- Remove the data with missing values from more than 60% of all the data
- Match event data with event action data via event ID
- Match link travel time with traffic incidents by link ID and coordinates
- Convert "type of affected lanes" to "number of lanes affected"
- Calculate incident clearance time based on reported start and end time of the incident
- Calculate incident recovery time based on the difference of mean and variance of normal traffic speed and reduced traffic speed
- Other data cleaning tasks on an as needed basis

As a preliminary estimate, data processing can be as long as 6 months depending on the specifics of the databases that need to be processed and combined. The data processing takes time when received volume of data is large and the time of downloading and acquiring data may be long. Moreover, the research team may have to create a database to maintain all received data and update it when there are changes in data format. For example, the research team may receive and process the dataset with variables that are minimum required for calibration. When more data information mentioned in ideal variables required become

available, the research team may need to update the database and process the updated dataset. Furthermore, the research team has not worked on how to filter the erroneous data which may require additional work and literature search.

This model will require a minimum time of six months for model training, calibration, validation, and computer implementation. However, one advantage of this model is its self-learning capability, which avoids the need for re-calibration. As per the interview results, the desired accuracy of calibration between ground truth data and trained prediction results is +/- 10% error and this model will re-calibrate automatically to maintain this level of accuracy. Table 33 shows a summary of estimated time for data processing, model calibration, training, and computer implementation.

Table 33. Estimated time for data processing and calibration efforts for Demiroluk and Ozbay's model.

Estimated minimum time for data processing	6 months
Estimated minimum time for model training, calibration, validation, and computer implementation[1]	6 months
Need to be re-calibrated?	Automatic calibration
Training accuracy need to achieve	+/- 10% error

Yu's model (2017) *(2)*

In this subsection, we will describe the effort of data preparation in order to calibrate and implement Yu's model (2). As mentioned above, the data needs of this model are highly compatible with TRANSCOM's available data. Therefore, there is no need to collect additional variables for further calibration. Instead, we mainly describe the efforts of data processing and model training. At the end of this subsection, we will also provide a rough estimated time of applying data preparation and calibration, respectively.

Data preparation and calibration

Table 34 shows the required variables in order to calibrate Yu's model (2), including 5-minute link travel time data, day and time of the incident, incident type, direction of the incident, and incident location. Moreover, Table 34 also mentions the minimum data required for calibration and implementation. Specifically, calibrating this model requires at least one year of historical traffic incident data. For a computer implementation of this model, we will need the actual coding of this specific calibrated, trained and validated model to make it operational in the sense of obtaining 1-week travel time data from the previous week along with the details of the actual traffic incident from the Analytics Server shown in Figure 31.

[1] Computer implementation refers to the actual coding of the specific calibrated, trained, and validated predictive model to make it operational in the sense of obtaining real-time data from the Analytics Server shown in Figure 31 and passing its predictive output to the desired user interface in the desired format.

Table 34. Data preparation efforts required for model calibration and implementation.

Minimum variables required for calibration	5-minute travel time data, day and time of the incident, incident type, direction, incident location
Minimum data amount for model calibration	1-year of travel time data and traffic incident data
Minimum data amount for real-time model implementation	1-week of travel time data from the previous week, details of current incident

As mentioned in the data analysis section (Section 3), there is a large amount of noise in the raw dataset. Therefore, efforts required for data processing and preparation are listed below:

- Remove the data with missing values from more than 60% of all the data
- Match event data with event action data via event ID
- Match link travel time with traffic incidents by link ID and coordinates
- Convert "type of affected lanes" to "number of lanes affected"
- Other data cleaning tasks on an as-needed basis
- Need high-performance computing resources for model training

As a preliminary estimate, data processing can be as long as 3 to 6 months depending on the specifics of the databases that need to be processed and combined. It is important to note that as model calibration requires data from different sources to be in the same format, a number of scripts that will automate the process have to be developed. However, for future re-calibration of the same model, the time it takes to process the new data will be significantly less than the original data processing effort since the developed scripts can be re-used as long as the format of the new datasets is not significantly different from the original one. In other words, once the initial data processing and preparation task is completed, the development team can re-use the same scripts to process new data for re-calibration purposes. However, it is important to note that if the data format changes, it will take some time to modify the scripts in order to process the new data.

For model training, calibration, validation, and implementation efforts, this model require data covering a period of 6-9 months. However, this model is capable of automatic re-calibration, which avoids additional manual re-calibration. In other words, as long as new incoming data is ready for re-calibration, this model is able to re-calibrate itself and find the optimal training accuracy (+/- 5%). Table 35 shows the estimated required time and effort for data processing, model calibration, and training[2] and computer implementation.

Table 35. Estimated time for data processing and calibration efforts Yu's model.

Estimated minimum time for data processing	**3 to 6 months**

[2] It is important to note that all the estimation of required effort in terms of time assumes a high level of familiarity with the TRANSCOM data as well as the specific aspects of the model to be calibrated. It will take considerably longer time if there is a need for learning specific aspects of each model and data needed to calibrate and operationalize them.

Estimated time for model training, calibration, validation and implementation	6-9months
Need to be re-calibrated?	Automatic calibration
Training accuracy need to achieve	+/- 5% error

Ghosh's model (2017) (*3*)

In this subsection, we describe the data processing and preparation effort required to calibrate and implement Ghosh's model (*3*). Similar to Yu's model (2), the data needs of this model are highly compatible with TRANSCOM's available data. Therefore, we mainly describe the efforts of data processing and model training. At the end of this subsection, we provide a rough estimate of effort in terms of time required for data preparation and calibration respectively.

Data preparation and calibration

Table 36 shows the required variables in order to calibrate Ghosh's model (*3*), which include 5-minute travel time data, day and time of the incident, incident type, direction, incident location, length of segment, whether or not shoulder is involved, total number of lanes, number of affected lanes, and type of affected lanes.

Moreover, Table 36 shows the minimum amount of data required for calibration and implementation. Specifically, calibrating this model requires at least 6-month of traffic incident data. For the real-time implementation of this model, one-week-long travel time from the past week along with the details of the current incident are required.

Table 36. Data preparation efforts required for model calibration and implementation.

Variables required for calibration	5-minute link travel time, day and time of the incident time, incident type, direction, incident location, length of segment, whether or not shoulder is involved, total number of lanes, number of affected lanes, type of affected lanes
Minimum data amount for calibration	6-month of travel time data and traffic incident data
Minimum data amount for implementation	1-week of recent travel time data, 1 real-time incident with details

As mentioned in the data analysis section (Section 3), there is a large amount of noise in the raw dataset provided to the research team by TRANSCOM. Therefore, substantial data processing effort listed below is required:

- Remove the data with missing values from more than 60% of all the data
- Match event data with event action data via event ID
- Match link travel time with traffic incidents by link ID and coordinates

- Match event data with shapefile via link ID
- Convert “type of affected lanes” to “number of lanes affected”
- Other data cleaning tasks on an as-needed basis

As a preliminary estimate, data processing can be as long as 3 to 6 months depending on the specifics of the databases that need to be processed and combined. However, it is important to note that as model calibration requires the same data formats, the time it takes to process the data for re-calibration efforts will be decreased. In other words, once the first data processing is finished, the research team can adopt the same computation scripts to process any further incoming data for re-calibration purposes.

For model calibration, training, validation and implementation efforts, this model will require a period of 6-9 months for model training, calibration., and computer implementation Moreover, this model needs to be re-calibrated every year in order to keep up-to-date roadway conditions. This model is required to have a +/- 5% error between ground truth data and trained prediction results. Table 37 shows the estimated time and efforts for data processing, model calibration, and training.

Table 37. Estimated time for data processing and calibration efforts Ghosh’s model.

Estimated minimum time for data processing	3 to 6 months
Estimated time for model training, calibration, validation, and implementation	6-9 months
Need to be re-calibrated?	Yes, every 1 year.
Training accuracy need to achieve	+/- 5% error

7. Timeline of the system development

Based on the literature review, detailed data analysis and designed system requirements for the ideal predictive incident delay estimation tool, we suggest a tentative timeline for the model development effort:
1. **Immediate action** (if approved): start focusing on the development of a delay prediction model based on the models recommended in this study, availability of data, and needs identified from surveys.
2. **Longer-term action (**2-3 years) When more data becomes available, consider the development of a duration prediction model based on the model recommended in this report.

Furthermore, based on the development efforts, we propose a step-wise model development approach[3]:

- **Step1:** Develop link-based delay / queue prediction models, validate their usefulness under real-world conditions, and integrate them in a software environment where operators can start experimenting with them. (Year 1)

[3] It is important to note that this step-wise approach will be revised based on the availability of new data as well as the success of each step in terms of the adoption of each model by member agencies.

- **Step 2:** Develop corridor-based delay / queue prediction models and validate their usefulness under real-world conditions and integrate them in a software environment where operators can start experimenting with them. (Year 2)
- **Step 3:** Develop alternative route-based delay / queue prediction models and validate their usefulness under real-world conditions and integrate them in a software environment where operators can start experimenting with them. (Year 3)
- **Step 4:** Consider development of duration prediction models and integrate them in a software environment where operators can start experimenting with them. (Year 3)

Appendix

Table 38. Field description of Highway Events data.

Sr. No.	Field Name	Description	Format	Missing Rate	Example
1	Id	Unique identifier of event	String	0.02%	ORI171242207
2	AssociatedEventID	Associated schedule/plan Event ID.	String	82.11%	ORI171242207
3	EventClass	Suggest a class of an event. List of values are: 0 – incident 1 – construction 2 – special event	Integer	0.02%	0
4	eventstatus	Status of an event List of values 0 - New 1 - Updated 2 - Closed 255 - Scheduled	Integer	0.02%	Closed
5	StartDateTime	Start date - time of event	Datetime	0.02%	12/25/17 1:39:00 AM
6	EndDateTime	End date - time of event	Datetime	0.03%	1/1/18 1:54:21 AM
7	LastUpdate	Last updated date-time of an event	Datetime	0.02%	12/25/17 1:39:40 AM

8	SummaryDescription	Description of an event	String	0.02%	MTA Bridges & Tunnels: Truck restrictions on I-495 westbound near Greenpoint Avenue (New York) Trucks over 12 feet restricted from using the Queens Midtown Tunnel. All vehicles over 12 feet must use alternate route.
9	Organization_ShortName	Reporting organization Name	String	0.02%	MTA Bridges & Tunnels
10	eventType	Contains event type of current event	String	0.02%	Disabled vehicle
11	LanesTotalCount	Total lanes of roadway	Integer	99.29%	4
12	LanesAffectedCount	Lanes affected by this event	Integer	95.15%	2
13	LanesDetail	Contains lane affected detail. Example, all lanes at least one lane closed for repairs	String	45.29%	right lane
14	LanesStatus	Contains lane status Example, open close traffic disruption	String	45.59%	blocked
15	Facility	Event location facility name	String	0.02%	I-495
16	Direction	Contains Event direction	String	6.38%	westbound
17	City	City name based on Event	String	12.48%	New York
18	County	County name based on Event	String	0.64%	Queens
19	State	State abbreviation of Event Example, NJ – New Jersey PA – Pennsylvania	String	0.02%	NY
20	PrimaryCity	Primary city name of Event	String	13.12%	New York

21	SecondaryCity	Secondary city name of Event	String	76.52%	Weehawken Twp
22	CityArticle	Article used with City name Example, at around between	String	38.98%	near
23	PrimaryMarker	Primary mile marker	Float	64.19%	1.2
24	SecondaryMarker	Secondary mile marker	Float	78.70%	0.5
25	MarkerArticle	Article used with mile marker Example, at around between	String	100.00%	at
26	MarkerUnits	Unit of measurement specified in mile marker	String	64.19%	mi
27	PointDatum	Any reference point/co-ordinates from which measurement may be taken. Here the default Point Datum is NAD83(North American 1983 Datum)	Float	0.02%	NAD83
28	PointLAT	Latitude of an event	Float	0.02%	40.73690033
29	PointLON	Longitude of an event	Float	0.02%	-73.9312973
30	PrimaryLoc	Primary Location of an event Example, Mile Post: 8.5 Exit: US 1 NORTH - MORRISVILLE {# 5A} (Beginning of I - 295)	String	1.41%	Greenpoint Avenue
31	SecondaryLoc	Secondary Location of an event	String	59.16%	New Jersey Side - Center Tube

32	LocArticle	Article used with Location of an event Example, at near	String	38.49%	near
33	Comments	Comments about an event	String	32.45%	until further notice
34	EventTypeDesc	Description of event type. Example, Highway	String	0.02%	Highway
35	EventImpactType	Impact of an event Example, Major Minor	String	89.20%	Minor
36	xcm_ShortDesc	Description of an event	String	0.02%	MTA Bridges & Tunnels: Truck restrictions on I-495 westbound near Greenpoint Avenue (New York) Trucks over 12 feet restricted from using the Queens Midtown Tunnel. All vehicles over 12 feet must use alternate route.
37	xcm_SortCategory	Contains the combination of sort order, sort weightage and event type id as per defined in DFE system Example, A040.200.196 Here, “A0” is prefix, “40” is sort order, “200” is sort weightage and “196” is an event type id	String	0.02%	A020.400.255
38	xcm_SortOrder	Sort order of an event type as per defined in DFE system	Integer	0.02%	20

39	xcm_PresentationHint	Image file which contains the icon, used for representing event on Operations Map	String	0.02%	incident.png
40	OR_TrackingID	Open Reach Id of an event	String	0.02%	ORI-171242207
41	xcm_Source	Source name of an event Example, TRANSCOM-OpenReach	String	0.02%	TRANSCOM OpenReach
42	xcm_Local	Flag value with a value of either 0 or 1. 0 means that an event is a public event 1 means that an event is a local event	Integer	0.02%	0
43	xcm_Transit	Flag value with a value of either 0 or 1. 0 means that an event is a highway event 1 means that an event is a transit event	Integer	0.02%	0
44	xcm_FacilityShortName	Facility's Short name where event occurred. Example, I-295	String	0.63%	I-495
45	xcm_CountyTo	Affected County due to event	String	27.03%	Hudson
46	OR_ToPointLat	Affected "To" Point Latitude	Float	64.84%	40.765298
47	OR_ToPointLon	Affected "To" Point Longitude	Float	64.84%	-74.014992
48	xcm_EarliestScheduleStart	Earliest Schedule Start date-time of an event	Datetime	100.00%	12/25/17 1:39:00 AM
49	xcm_LatestScheduleEnd	Latest Schedule End date-time of an event	Datetime	100.00%	1/1/18 1:54:21 AM

50	xcm_EventID	Combination of creation time and reporting org id Example: 2014081522113401104 Here, “20140815221134” is date in format of yyyyMMddHHmmss “01104” is reporting org id	String	0.02%	2.01712E+18
51	xcm_ReportingOrgName	Reporting Organization name	String	0.02%	MTA Bridges & Tunnels
52	xcm_UpdateCount	Number of times events got updated	Integer	0.02%	29
53	xcm_RaEventType	Event Type reference to Event Archive System Example, roadway vehicle fire accident	String	0.02%	truck restrictions
54	xcm_IncExpEndDttm	Event’s expected end date-time	Datetime	0.10%	12/25/17 1:39:00 AM
55	xcm_CountyFrom	Affected “From” county name	String	0.58%	Queens
56	xcm_WeatherCondition	Weather condition during event	String	99.98%	sunny
57	xcm_PavementCondition	Pavement condition during event	String	100.00%	N/A
58	xcm_OtherInformationTwo	Additional Other Information about event	String	100.00%	N/A
59	xcm_LaneDetails	Contains lane affected detail	String	89.74%	service road
60	xcm_Impact	Impact of event. Example, MAJOR MINOR	String	89.22%	Minor

61	xcm_RespondingOrgName	Organization who responded to event	String	100.00%	MTA Bridges & Tunnels
62	xcm_IncidentOccured	Date-time when event occurred	Datetime	100.00%	12/25/17 1:39:00 AM
63	xcm_IncidentReported	Date-time when event was reported	Datetime	100.00%	12/25/17 1:39:00 AM
64	xcm_IncidentVerified	Date-time when event was verified	Datetime	100.00%	12/25/17 1:39:00 AM
65	xcm_ResponseIdentifiedAndDispatched	Date-time when response was identified and dispatched to event location	Datetime	100.00%	12/25/17 1:39:00 AM
66	xcm_AllLanesOpenToTraffic	Date-time when all lanes were open to traffic	Datetime	100.00%	12/25/17 1:39:00 AM
67	eventDuration	Duration of the event	Datetime	0.02%	7 - 00:15

Table 39. Action type with type id.

Id	TypeName
0	Verification
1	Notification
2	VMS
3	HAR
4	Diversion Route
5	IMRT
6	Crew
7	State Police
8	Other
9	Fatality
10	Construction
11	HAZMAT
12	Bridge Plates

13	Buses Ordered
14	On-board Announcement
15	Station Announcement
16	Station Displays
17	Alternates
30	Event Created
31	Event Updated
32	Event Closed
33	Event Reopened
34	Event Copied To
35	Event Copied From
36	Event Spawned
37	Event Terminated
38	Event Archived
39	Event Modified
40	Event Prepopulated
50	Initial Event Snapshot
51	Event Update Snapshot
52	Event Closed Snapshot
53	Event Created Snapshot

References

1. Demiroluk, S., and K. Ozbay. Adaptive learning in bayesian networks for incident duration prediction. *Transportation Research Record,* Vol. 2460, No. 1, 2014, pp. 77-85.
2. Yu, R., Y. Li, C. Shahabi, et al. Deep learning: A generic approach for extreme condition traffic forecasting.In *Proceedings of the 2017 SIAM International Conference on Data Mining*, SIAM, 2017. pp. 777-785.
3. Ghosh, B., J. Dauwels, and U. Fastenrath. Analysis and prediction of the queue length for non-recurring road incidents.In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2017. pp. 1-8.
4. Haule, H. J., T. Sando, R. Lentz, et al. Evaluating the impact and clearance duration of freeway incidents. *International Journal of Transportation Science and Technology,* Vol. 8, No. 1, 2019, pp. 13-24.
5. Khattak, A. J., J. L. Schofer, and M.-H. Wang. A simple time sequential procedure for predicting freeway incident duration. *Journal of Intelligent Transportation Systems,* Vol. 2, No. 2, 1995, pp. 113-138.
6. Garib, A., A. Radwan, and H. Al-Deek. Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering,* Vol. 123, No. 6, 1997, pp. 459-466.
7. Peeta, S., J. L. Ramos, and S. Gedela. Providing real-time traffic advisory and route guidance to manage Borman incidents on-line using the hoosier helper program. 2000.
8. Khattak, A. J., J. Liu, B. Wali, et al. Modeling traffic incident duration using quantile regression. *Transportation Research Record,* Vol. 2554, No. 1, 2016, pp. 139-148.
9. Yu, B., and Z. Xia. A methodology for freeway incident duration prediction using computerized historical database.In *Twelfth COTA International Conference of Transportation ProfessionalsAmerican Society of Civil EngineersTransportation Research Board*, 2012.
10. Weng, J., W. Qiao, X. Qu, et al. Cluster-based lognormal distribution model for accident duration. *Transportmetrica A: Transport Science,* Vol. 11, No. 4, 2015, pp. 345-363.
11. Ozbay, K., and P. Kachroo. Incident management in intelligent transportation systems. 1999.
12. Smith, K., and B. L. Smith. Forecasting the clearance time of freeway accidents. 2002.
13. Breiman, L. *Classification and regression trees*. Routledge, 2017.
14. Knibbe, W. J. J., T. P. Alkim, J. F. Otten, et al. Automated estimation of incident duration on Dutch highways.In *2006 IEEE Intelligent Transportation Systems Conference*, IEEE, 2006. pp. 870-874.
15. He, Q., Y. Kamarianakis, K. Jintanakul, et al. Incident duration prediction with hybrid tree-based quantile regression.In *Advances in dynamic network modeling in complex transportation systems*, Springer, 2013. pp. 287-305.
16. Zhan, C., A. Gan, and M. Hadi. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. *IEEE Transactions on Intelligent Transportation Systems,* Vol. 12, No. 4, 2011, pp. 1549-1557.
17. Wei, C.-H., and Y. Lee. Sequential forecast of incident duration using Artificial Neural Network models. *Accident Analysis & Prevention,* Vol. 39, No. 5, 2007, pp. 944-954.
18. Park, H., A. Haghani, and X. Zhang. Interpretation of Bayesian neural networks for predicting the duration of detected incidents. *Journal of Intelligent Transportation Systems,* Vol. 20, No. 4, 2016, pp. 385-400.
19. Ozbay, K., and N. Noyan. Estimation of incident clearance times using Bayesian Networks approach. *Accident Analysis & Prevention,* Vol. 38, No. 3, 2006, pp. 542-555.

20. Boyles, S., D. Fajardo, and S. T. Waller. A naive Bayesian classifier for incident duration prediction.In *86th Annual Meeting of the Transportation Research Board, Washington, DC*, Citeseer, 2007.
21. Yang, B.-b., X. Zhang, and L. Sun. Traffic incident duration prediction based on the Bayesian decision tree method.In *The First International Symposium on Transportation and Development–Innovative Best Practices (TDIBP 2008) American Society of Civil EngineersChina Academy of Transportation Sciences*, 2008.
22. Shen, L., and M. Huang. Data mining method for incident duration prediction.In *International Conference on Applied Informatics and Communication*, Springer, 2011. pp. 484-492.
23. Qi, Y., and H. Teng. An information-based time sequential approach to online incident duration prediction. *Journal of Intelligent Transportation Systems,* Vol. 12, No. 1, 2008, pp. 1-12.
24. Yu, B., Y. Wang, J. Yao, et al. A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. *Neural Network World,* Vol. 26, No. 3, 2016, p. 271.
25. Zeng, X., and P. Songchitruksa. Empirical method for estimating traffic incident recovery time. *Transportation Research Record,* Vol. 2178, No. 1, 2010, pp. 119-127.
26. List, G. F. Quantifying non-recurring delay on New York City's arterial highways. 2008.
27. Khattak, A., X. Wang, and H. Zhang. Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays. *IET Intelligent Transport Systems,* Vol. 6, No. 2, 2012, pp. 204-214.
28. Li, J., C.-J. Lan, and X. Gu. Estimation of incident delay and its uncertainty on freeway networks. *Transportation Research Record,* Vol. 1959, No. 1, 2006, pp. 37-45.
29. Cassidy, M. J., and L. D. Han. Proposed model for predicting motorist delays at two-lane highway work zones. *Journal of Transportation Engineering,* Vol. 119, No. 1, 1993, pp. 27-42.
30. Yi, J. Traffic characteristics and estimation of traffic delays and user costs at Indiana freeway work zones.In, Indiana. Dept. of Transportation, 1999.
31. Chien, S., and P. Schonfeld. Optimal work zone lengths for four-lane highways. *Journal of Transportation Engineering,* Vol. 127, No. 2, 2001, pp. 124-131.
32. Jiang, X., and H. Adeli. Freeway work zone traffic delay and cost optimization model. *Journal of Transportation Engineering,* Vol. 129, No. 3, 2003, pp. 230-241.
33. Chitturi, M. V., R. F. Benekohal, and A.-Z. Kaja-Mohideen. Methodology for computing delay and user costs in work zones. *Transportation Research Record,* Vol. 2055, No. 1, 2008, pp. 31-38.
34. Ramezani, H., R. F. Benekohal, and K. A. Avrenli. Methodology to analyze queue length and delay in work zones.In, 2011.
35. Ullman, G. L., and C. L. Dudek. Theoretical approach to predicting traffic queues at short-term work zones on high-volume roadways in urban areas. *Transportation Research Record,* Vol. 1824, No. 1, 2003, pp. 29-36.
36. Curtis, D. QuickZone [software that estimates traveller delay due to road work zones]. *Public Roads,* Vol. 65, No. 1, 2001.
37. Bartin, B., K. Ozbay, and S. Mudigonda. Interactive lane closure and traffic information tool based on a geographic information system. *Transportation Research Record,* Vol. 2272, No. 1, 2012, pp. 44-55.
38. Bartin, B., K. Ozbay, M. D. Maggio, et al. Work zone coordination software tool. *Transportation Research Record,* Vol. 2617, No. 1, 2017, pp. 60-70.
39. Chang, G.-L., and N. Zou. An Integrated Work-Zone Computer System for Capacity Estimation, Cost/Benefit Analysis, and Design of Control.In, 2009.
40. Ozbay, K., and B. Bartin. Development of uniform standards for allowable lane closure: final report, September 2008.In, New Jersey. Dept. of Transportation, 2008.

41. Bian, Z., and K. Ozbay. Estimating uncertainty of work zone capacity using neural network models. *Transportation Research Record,* Vol. 2673, No. 2, 2019, pp. 49-59.
42. Hojati, A. T., L. Ferreira, S. Washington, et al. Reprint of: modelling the impact of traffic incidents on travel time reliability. *Transportation Research Part C: Emerging Technologies,* Vol. 70, 2016, pp. 86-97.
43. Javid, R. J., and R. J. Javid. A framework for travel time variability analysis using urban traffic incident data. *IATSS research,* Vol. 42, No. 1, 2018, pp. 30-38.
44. Caceres, H., H. Hwang, and Q. He. Estimating freeway route travel time distributions with consideration to time-of-day, inclement weather, and traffic incidents. *Journal of Advanced Transportation,* Vol. 50, No. 6, 2016, pp. 967-987.
45. Di Martino, S., S. Kwoczek, and S. Rossi. Predicting the Spatial Impact of Planned Special Events.In *International Symposium on Web and Wireless Geographical Information Systems*, Springer, 2019. pp. 102-117.
46. Yue, M., L. Fan, and C. Shahabi. Traffic Accident Detection with Spatiotemporal Impact Measurement.In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2018. pp. 471-482.
47. Pan, B., U. Demiryurek, and C. Shahabi. Utilizing real-world transportation data for accurate traffic prediction.In *2012 IEEE 12th International Conference on Data Mining*, IEEE, 2012. pp. 595-604.
48. Chen, H., and H. A. Rakha. Real-time travel time prediction using particle filtering with a non-explicit state-transition model. *Transportation Research Part C: Emerging Technologies,* Vol. 43, 2014, pp. 112-126.