ABSTRACT

Title of Dissertation: DISPATCHING AND RELOCATION OF EMERGENCY VEHICLES ON FREEWAYS: THEORIES AND APPLICATIONS

Hyoshin Park, Doctor of Philosophy, 2015

Dissertation directed by: Professor Ali Haghani
Department of Civil and Environmental Engineering

Resource allocation decisions are made to serve the current emergency without knowing which future emergency will be occurring. Different ordered combinations of emergencies result in different performance outcomes. Even though future decisions can be anticipated with scenarios, previous models assume that events over a time interval are independent. This dissertation assumes that events are interdependent, because speed reduction and rubbernecking due to an initial incident provoke secondary incidents. The misconception that secondary incidents are not common has resulted in overlooking a look-ahead concept.

This dissertation pioneers in relaxing the structural assumptions of independencies during the assignment of emergency vehicles. When an emergency is detected and a request arrives, an appropriate emergency vehicle is immediately dispatched. We provide tools for quantifying impacts based on fundamentals of incident occurrences through identification, prediction, and interpretation of secondary incidents. A proposed online dispatching model minimizes the cost of moving the

next emergency unit, while making the response as close to optimal as possible. Using the look-ahead concept, the online model flexibly re-computes the solution, basing future decisions on present requests. We introduce various online dispatching strategies with visualization of the algorithms, and provide insights on their differences in behavior and solution quality. The experimental evidence indicates that the algorithm works well in practice.

After having served a designated request, the available and/or remaining vehicles are relocated to a new base for the next emergency. System costs will be excessive if delay regarding dispatching decisions is ignored when relocating response units. This dissertation presents an integrated method with a principle of beginning with a location phase to manage initial incidents and progressing through a dispatching phase to manage the stochastic occurrence of next incidents. Previous studies used the frequency of independent incidents and ignored scenarios in which two incidents occurred within proximal regions and intervals. The proposed analytical model relaxes the structural assumptions of Poisson process (independent increments) and incorporates evolution of primary and secondary incident probabilities over time. The mathematical model overcomes several limiting assumptions of the previous models, such as no waiting-time, returning to original depot rules, and fixed depot. The temporal locations flexible with look-ahead are compared with current practice that locates units in depots based on Poisson theory. A linearization of the formulation is presented and an efficient heuristic algorithm is implemented to deal with a large-scale problem in real-time.

DISPATCHING AND RELOCATION OF EMERGENCY VEHICLES ON FREEWAYS: THEORIES AND APPLICATIONS

by

Hyoshin Park

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Ali Haghani, Chair/Advisor
Professor Paul Schonfeld
Professor Cinzia Cirillo
Professor Lei Zhang
Professor S. Raghu Raghavan

Dedication

To my grandfather who passed away during my PhD process.

Acknowledgments

I would never have been able to finish my dissertation without guidance of my committee members, colleagues, and support from my family.

First and foremost I'd like to thank my advisor, Ali Haghani for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past three years. He has always made himself available for help and advice and there has never been an occasion when I've knocked on his door and he hasn't given me time. I would also like to thank Dr. Zhang, Dr. Cirillo, and Dr. Raghavan for guiding my research for the past years and helping me to develop theories in my dissertations.

I would like to thank Dr. Schonfeld, who as a good mentor, was always willing to help and give his best suggestions. I owe many of his times that spent to encourage my research and future career. It has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank my parents for always supporting me and encouraging me with their best wishes. I would also like to thank my brother, Kyoungshin David Park, for being the best friend and pursuing lifetime research together.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation

Traffic congestion forces motorists to begin traveling much earlier for short-distance commutes, and has become a major feature of urban areas around the world [1]. Traffic incidents cause one-quarter of the congestion on US roadways, and every minute that a freeway lane is blocked creates 4-minutes extra delay [2]. When a traffic emergency is accompanied by a lane-closure, it is important for responders to arrive at the emergency scene as soon as possible. An efficient control of emergency response units (ERUs) can greatly reduce injuries and adverse impacts [3]. One way to enhance performance is applying a mobile facility concept [4], instead of a fixed facility. Once an ERU is assigned to an incident, the remaining ERUs can be relocated to better respond to future incidents.

To serve the *current* emergency request, dispatchers make a decision without knowing which request will be occurring in the *future*. One might assign or relocate a ERU near the expected location of an emergency in the next stage [5, 6]. These conventional models assume that a given number of independent events occur over a certain time interval. However, that expected location might have a request after requests at other locations. In reality, we have different *orders* in which the

emergencies take place, and dispatching action is processed at a time before next emergency occurs. It is unreasonable to assume that different orders share the same solution. Non-uniformly distributed requests on a transportation network are more likely to have different orders that lead to different outcomes of the series. Ignoring the sequence might miss out a critical location, and prepare for a completely wrong location. Shortcomings of previous studies [5, 6] become obvious when they cannot meet the standard requests required by Emergency Medical Services Act of 1973.

The property explained above motivates us to see this problem from a different perspective, that is, *online optimization with look-ahead*. Instead of unrealistic assumption that we ignore, or we know absolutely nothing about the distribution of incident sequence, it is assumed that some interdependent incidents make the system stochastic through a sequence of incidents. Speed reduction and rubbernecking due to an initial traffic incident provoke additional incidents, which are referred to as secondary incidents (specifically, secondary crashes) [7]. The emergency system evolves from one time-stage to another in such a way that chance elements are involved in progressing from one state to the next. However, a stereotype, secondary crash occurrences are not common and not easy to understand, has resulted in overlooking optimally controlling emergency response vehicles with future look-ahead. Traffic management agencies are faced with a dilemma. The consequences of misguided assumption raise a serious issue in applications when secondary incident likelihood is underestimated or overestimated. For example, counting a potential secondary incident as equal to a minor incident may result in the location of response units far away from a real secondary incident site.

This dissertation provides vehicle arrangement decisions so that emergency requests can be responded in a time-efficient manner. First of all, when an emergency is detected and a request arrives, fundamentals of incident occurrences (i.e., identification, prediction, and interpretation of secondary crashes) are made, and an appropriate vehicle is dispatched. It is an online dispatching problem because an emergency operator performs an immediate action in response to each request with partial future information. After having served a designated request, the available or/and remaining vehicles are relocated to newly updated base for the next potential emergency. A stochastic location problem is posed to build a realistic framework.

1.2 Response to Stochastic Sequence of Emergencies

The dissertation incorporates a realistic and stochastic process into the design of deployment of emergency response vehicles. The conventional optimization approach for location or allocation problem assumes that a given number of independent and identically distributed (IID) events occur over a time interval. However, the sequence is an ordered combination (permutation) of emergency requests. Suppose a set of sequences with the past request at site (2), current request at site (3), and next requests at either site 1 or site 2. Let the probability of incident at site 1 be 10% and at site 2 be 90%.

$$\sigma = \left\{ \begin{array}{ccc} (2,3) & 1 & 2 \\ (2,3) & 2 & 1 \end{array} \right\} \tag{1.1}$$

A traditional approach neglects three essential properties. First, without consideration of the order, the dispatcher would make a decision based on the anticipation of an incident at site 2. This will lead to excessive response time when an incident occurs at site 1 before site 2. Such scenario will make site 1 to be served from resources farther away than regularly assigned resources, or will not be addressed until the closest resource becomes available. Without an appropriate help, lack of tools may cause an incident to block the traffic flow and induce inefficiencies in the clearance operation.

Second, with a randomness assumption of the IID sequence, reversed times of incidents' occurrence make solutions of two different sequences the same. However, the assigned probability for each sequence is different when an initial incident provokes secondary incidents [7]. Even though primary incidents at site 2 provoke secondary incidents at site 1, reverse order (primary incidents at site 1) does not have the same mutual dependency. In reality, the probability distributions of the first and the second sequence are different. This property will cause the probability distribution of solution in Equation 1.1 to be *asymmetric*.

Lastly, probabilities associated with each transition depend on incidents earlier than the immediately preceding one. Previous studies take account of only a single step in the process. However, when primary incidents occur in a sequence of time intervals, the likelihoods of secondary incidents caused by each primary incident are accumulated. The conditional probability of a secondary incident in the future depends jointly on primary and secondary incidents that have occurred during past and present time stages. As a result, the probability of incidents evolves over time

instead of being fixed 10% at site 1 and 90% at site 2. The independent increments property of IID process (the numbers of occurrences counted in disjoint intervals are independent of each other) does not hold on freeways with secondary incidents. The cost associated with providing service to secondary incidents will exceed the original one due to capacity reductions [8]. Therefore, potential effects of secondary incidents on emergency response system have been overlooked.

1.3 Fundamentals of Incident Occurrences

To obtain property of inferences in emergency scenarios, we need to understand fundamentals of secondary crash occurrences. For instance, a Poisson process assumes that all subsequent incidents are independent of the previous incident, and all incidents have the same exponential distribution. An IID sequence rather fits into independent occurrences such as, e.g., repeated throws of loaded dice. In reality, random incidents at each location do not have the same probability distribution as other locations.

1.3.1 Identification of Secondary Crashes with Advanced Data

We cannot exaggerate the importance of precise identification since it has a direct influence on the prediction of secondary crash occurrences. Eighteen percent of traffic fatalities occur as a result of secondary crashes [9] and stuck-by secondary crashes are on the rise [2]. However, it is difficult to quantify a primary incident's impact on secondary crashes and researchers have made little progress. Previously

suggested thresholds and measurement parameters provide no universal definition of a secondary crash, regardless of discussions on the topic. In a recent survey, out of 11 practitioner responses, only 5 of them routinely measure and report secondary crashes [10]. To overcome difficulties revealed in determining the precise definition of secondary crashes, we need a robust definition.

An estimation result of traffic states significantly depends on quality of sensor data (e.g. loop detectors). However, the point sensors are prone to various errors caused by malfunctioning and communication failures. Researchers have tried to overcome the limitations of the unsatisfactory quality of point sensor data [11, 12]. Nevertheless, accurately representing the traffic conditions is a challenge [13].

In recent years, the vehicle probe industry has emerged as a viable means to monitor traffic flow. The travel time collected from vehicle probe data generally satisfies the requirements of applications for real-time travel time display [14]. This is a new opportunity to use real-time estimations of traffic congestion caused by crashes. One application [15] defines freeway segments as congested using fixed threshold (Figure 1.1). Traffic conditions can be determined by comparing the current reported speed to the reference speed (85th-percentile of the observed speeds) for each segment of road. A segment represents congestion when the actual travel speed drops below 60% of the reference speed longer than 5 min.

Unfortunately, the above static threshold method [15] cannot consider the actual representation of prevailing traffic condition when an incident occurs. In this dissertation, vehicle probe data is used to provide temporal and spatial thresholds of congestion related to primary incidents. It captures the dynamics of traffic evolution

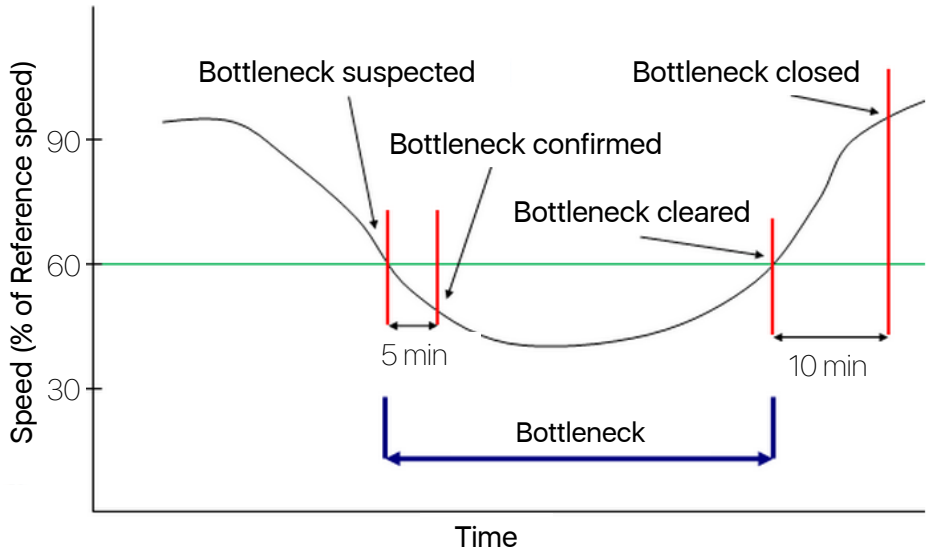


Figure 1.1: Bottleneck identification method [15]

during the primary incidents to identify secondary crashes. We further propose a clustering model [16] that considers posterior distributions to recognize congestion patterns, and propose an adjusted boxplot model to deal with obscure posterior predictive distributions in the group. Full details on the clustering model, adjusted boxplot model, real-world application, and other reference models will be presented in Chapter 4.

1.3.2 Advanced Machine Learning for Secondary Incidents

Incident duration is defined as the time between the detection and clearance of an incident. The response time contains decision-making of a responding agency and the actual travel time of the rescue personnel and equipment to the scene. The clearance time is defined as the time between the arrival of the response units and the last recovery.

It is important to understand the key cause of secondary crashes. For in-

stance, the longer an incident scene is in place, the greater the likelihood of a secondary incident [7]. Total time it takes for an incident to be cleared increases with the occurrence of secondary incidents, and travelers may experience ever-increasing congestion (Figure 1.2). We provide solutions for two main problems: an accurate prediction problem and comprehensible interpretation problem.

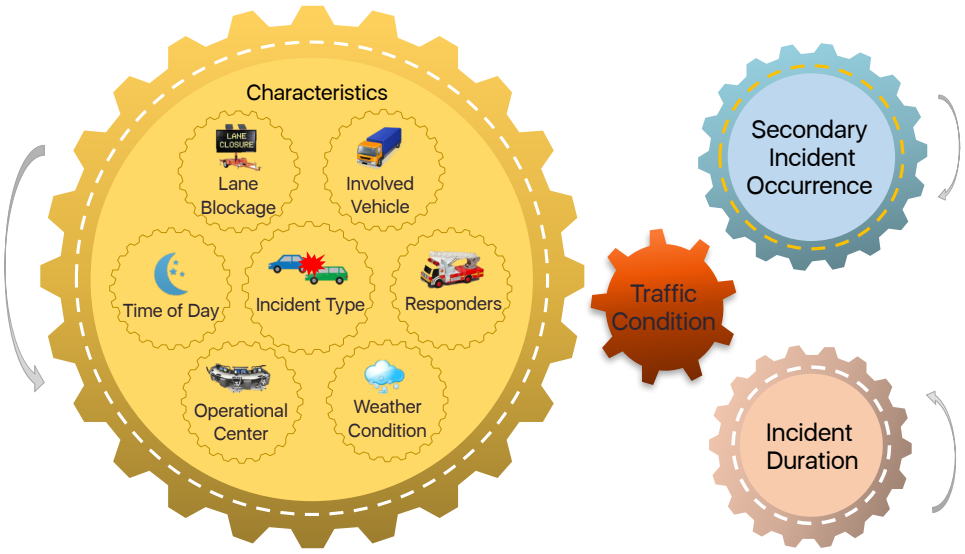


Figure 1.2: Secondary incident occurrences and contributing factors

Most previous efforts on predicting incident duration are not directly applicable to real-world due to lack of data sample, consistency in prediction, and key contributing factors. Instead, predicting incident duration still depends on the skill and field experience of the local emergency operator. Response time is quicker for severe incidents that could potentially cause greater impacts on traffic congestion [17]. We develop new models which would better perform than other existing models (back-propagation neural networks, support vector machines, and classification and regression trees).

Compared to primary incidents, secondary incidents have low sample means

and a small sample size. The wide variety of causes and impacts of non-recurring congestion make it difficult to quantify random and complex incident natures at a system level. As a result, crash prediction models have been over-fitted and have poor predictive performance [18]. We take a principled Bayesian learning approach to neural networks to predict the likelihood of secondary incidents without over-fitting.

Questions still remain about next arrival of response units. Previous studies omit a critical factor that may extend incident duration due to long travel time of a second or third response team. Even after the arrival of the first responders, the extensive travel time of the next responders can potentially influence the entire clearance operation. This dissertation considers the evolution of traffic flow by updating newly predicted incident duration according to the time point of the prediction. This helps to reach the desirable levels of prediction accuracy and will be possible by using global positioning system-based automated vehicle location on emergency vehicles. The resulting prediction value of incident duration is used to indicate the secondary incident likelihood. Full details on the Bayesian neural networks, the process of sequential prediction, and performance comparison with other reference models will be presented in this dissertation, in Chapter 5.

Challenges remain in explaining neural networks. No satisfactory interpretation of neural networks' behavior has been offered, and they have been regarded as black boxes [19]. A pedagogical rule extraction approach is introduced to improves the understanding of incident duration and secondary incidents by extracting comprehensible rules from the neural networks. The proposed algorithm branches the

tree according to the predicted values by the neural network model so it retains high accuracy while being easy to understand. The extracted decision trees provide a discovery and an explanation of previously unknown relationships present in incident nature (Figure 1.3).

For the potential mathematical utility of neural networks, multivariate and non-linear conditions should be considered because incident nature rarely occurs due to a simple cause or to a unique perturbation. We use the connection weight and stochastic gradient boosted tree to generate interpretable parameters for each explanatory variable. Unlike previous sensitivity analysis, these models determine how different values of an independent variable will impact a particular dependent variable. Full details on the pedagogical rule extraction, stochastic gradient boosted tree, and connection weight approaches will be presented in this dissertation, in Chapter 6.

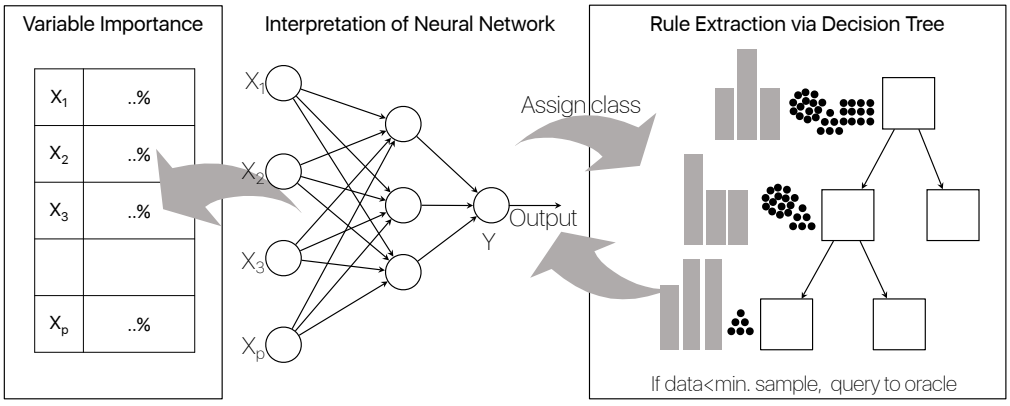


Figure 1.3: Interpretation of secondary incidents

1.3.3 Secondary Incident Delay Model (SIDM)

When we evaluate the performance of clearance of an incident, traffic delay is commonly used as a key indicator of the impacts of incidents and the benefits of emergency responses [20]. Accurate estimation of incident-induced delay helps traffic operators efficiently manage emergency response units. Highway capacity, an input to delay estimation, is an important measure in studying reliability of the transportation system [21].

A *realized* capacity reduction, after the occurrence of a secondary incident, is different from the estimation result of traditional incident delay models [22]. Applying traditional delay models may result in underestimation or overestimation of total delay. For example, the *Highway Capacity Manual* (HCM) considers the proportion of the traveled roadway that is blocked by the stopped vehicles and the number of lanes on the roadway [23]. Capacity at the secondary incident location is bounded by the maximum discharge flow rate. The freeway segment is assumed to degrade from incident-free state to primary-incident state when a primary incident occurs, and to degrade further to secondary-incident state when a secondary incident occurs [24]. As moving bottlenecks explain the capacity-drop, a backward moving shockwave from the primary incident location imposes speed reduction to traffic at a secondary incident site and discharge flow reduction in upstream location. As a result, available capacity is lower than traditional concept of capacity during clearance or recovery stages of a primary incident.

In this dissertation, a new variable represents the magnitude of capacity re-

duction over time. Secondary incident delay is described in a geometric surface area with explicit formulations considering gap/overlap between occurrence and clearance of primary and secondary incidents. To estimate this variable, we mathematically formulate a secondary incident delay model. Unfortunately, input parameters of secondary incident delay model are assumed be known, and the models can be used for after-incident evaluation [25]. We need to consider dynamic characteristics of the network [26].

In addition, a stochastic extension of secondary incident delay model is proposed to provide real-time prediction of delay. The first response unit and a secondary response unit arrival times are considered to obtain location-specific incident duration, one of input parameters for estimation of capacity reduction.

Full details on the formulation of the deterministic and stochastic secondary incident delay model, empirical analysis on capacity reduction, and comparison with capacity adjustments of Highway Capacity Manual will be presented in Chapter 7.

1.4 Proposed System for Emergency Response Unit (ERU)

Once the traffic monitoring system has detected an incident, it is necessary to efficiently manage response units to reduce negative impact. In a previous study [27], the travel time of ERUs were dependent on the traffic condition. However, in many cases, even though police units are dispatched to the scene, the left lane can be blocked until available emergency units arrive. Maryland's "clear the road" policy provides ERUs (well-equipped vehicles) for the rapid removal of vehicles from

the travel lanes rather than waiting for a private tow service. The proposed model repositions single type of ERUs to the best locations to serve future incidents. Most parts of United States and Canada enforce the "move over laws" that require motorists to move to the farthest roadside and stop, until the emergency vehicle has passed the vicinity. We consider freeway networks that have enough space on right lane/shoulder which are less likely to be influenced by severe traffic congestions. Emergency vehicles still expect delays waiting for other traveling vehicles to become aware of their presence and yield. We explore both minimum (free-flow traffic) and maximum (congested traffic) response time as an input to the model. In addition, we explore a case when each link of the network is assumed to have a fixed speed equivalent to 70% of the free-flow speed on that link.

1.5 Online ERU Dispatching

We receive a sequence of emergency calls and perform an immediate action in response to each request, without having the entire information of future. Some independent emergencies occur at unpredictable locations at unpredictable times. However, it is unrealistic to assume that we know absolutely nothing about the distribution of emergency sequence. We can have an advantage of knowing part of the future, look-ahead [28]. The online dispatching model computes a solution one-by-one in an online fashion, while minimizing the overall response time of emergency vehicles.

The flexible dispatching model uses real-time updated information to consider

reassigning an emergency response vehicle to a new emergency if the vehicle has not arrived at the previous one yet. The model minimizes the response time to the next request while making it as close to the optimal response as possible. Without knowing everything about the future, the online algorithm may turn out not to be optimal, but we focus on the quality of decision that is compared against an adversary on a worst-case input. The online model has a *look-ahead* contingent on present emergency in making future decision of which vehicle to assign. We characterize uncertainty of future emergencies conditioned on information of currently available emergencies in Chapter 8.

1.6 Stochastic ERU Location

An optimal dispatching strategy for emergency vehicles plays a crucial role in reducing the adverse effects of accidents by minimizing average response times [29]. In highway networks where traffic surveillance and incident detection are available, the key question is where to locate emergency response units. A p-median method [30] has been applied to the location-allocation problem. A single incident rate, assuming in dependency between two incidents, has been considered. However, crash risk is higher in the presence of an earlier crash [7]. Although emergency operators manage to handle a primary incident (i.e. the first incident) with this assumption, drivers suffer heavily when another incident, a secondary incident (i.e. an incident within temporal and spatial impact of a primary incident), occurs [31]. The nearest ERU might be unavailable to respond because the closest resource is

occupied by an earlier incident.

Potential delay caused by inefficient response to secondary incidents is unknown until the primary incidents′ information is given. In response to secondary incidents taking significant portion of traffic delays, emergency agency′s strategic concerns for effective response have been growing. Fortunately, scientific breakthroughs enabled us to develop thresholds as a consistent definition of secondary incidents [7] and to collect reliable data with advanced technologies [16]. This dissertation has an advantage of using reliable traffic information (i.e., INRIX) and tracking each ERUs′ performance (i.e., response, clearance) that can easily accommodate real-time operations.

The occurrence of many events has been assumed to follow statistical distributions (e.g., Poisson [6]). Another assumption of previous studies is a returning rule that limits the response units to be always dispatched from an original location. This assumption creates an unnecessary trip to the designated location.

Location-allocation solutions have been presented in two stages of decisions to address deploying response units to potential sites before an incident occurs and dispatching response units to the scene after an incident occurs [8]. In a proactive step, the emergency response units are pre-assigned to potential sites, the location problem, to promptly respond to a detected primary incident, the allocation problem. In the next step, additional emergency units, expected to suffer excessive travel time, are optimally dispatched to minimize response time to a secondary incident site. The likelihoods of secondary incidents caused by each primary incident are accumulated when primary incidents occur in a sequence of time intervals. As a

result, the probabilities of secondary incidents evolve over time. Full details on the stochastic formulation, linearization of the formulation, and performance on the real-time framework will be presented in Chapter 9.

1.7 About this Dissertation

The flow of the rest of this document is as follow. After reviewing relevant studies in Chapter 2, the dissertation′s main findings and contributions include fundamentals of incident occurrences for a reliable structure of scenario (Chapters 3, 4, 5, 6, and 7), an online dispatching with a look-ahead (Chapter 8), and a stochastic relocation (Chapter 9).

- Chapter 3 documents a stochastic process of future stages of incidents. Each sequence of incidents represents a scenario that is represented in a matrix form with an expected probability.

- Chapter 4 presents nonrecurring congestion with vehicle probe data. The clustering model considers posterior distributions to recognize congestion patterns under the impact of incidents. To deal with obscure posterior predictive distributions in the group, an adjusted boxplot model is introduced. It provides dynamic impact with temporal and spatial thresholds of congestion related to primary incidents to identify secondary crashes. Compared to static models, the proposed dynamic method has superior detection of secondary crashes.

- Chapter 5 documents the pedagogical interpretation of the secondary crash prediction. It shows sequential prediction of secondary crash likelihood from

the point of incident notification to the road clearance. It introduces a principled Bayesian learning approach to neural networks to consistently predict the likelihood of secondary crashes.

- Chapter 6 presents a comprehensive test for the developed models to determine their effectiveness. It provides decision tree approaches and a connection weight approach to improve understanding and to quantify key factors for secondary crashes.

- Chapter 7 documents capacity reduction due to incidents. Deterministic secondary incident delay is described in geometric surface area with explicit formulations considering gap/overlap between occurrence and clearance of primary and secondary incidents. A stochastic extension of delay model provides real-time prediction of delay. Empirical evidence presents significant impact of secondary incidents on capacity reduction. Without consideration of time series, the Highway Capacity Manual underestimates or overestimates capacity.

- Chapter 8 documents work on online dispatching problem where the objective is to minimize the time needed to respond to a sequence of emergency requests. The proposed dynamic model minimizes the cost of moving the next response unit while making it as close to the optimal response as possible. With updated information, the online model flexibly re-computes the solution to react in real-time. The practical online algorithm has a look-ahead setting contingent on present requests in making future decisions. We apply various

online dispatching strategies with visualization of the algorithms, and provide insights on their differences in behavior and solution quality.

- Chapter 9 presents an integrated method to solve location and routing problem of emergency response units with a stochastic approach. The principle is to begin with a location phase for managing initial incidents and to progress through a routing phase for managing the stochastic occurrence of next incidents. The proposed analytical model relaxes the structural assumptions of Poisson process (independent increments) and incorporates evolution of primary and secondary incident probabilities over time. The proposed stochastic programming model overcomes several limiting assumptions of the previous models and hedges well against a wide range of scenarios in which probabilities of a sequence of incidents are assigned. The initial non-linear stochastic model is linearized. An efficient heuristic algorithm is implemented to deal with time-consuming process of a large-scale problem in real-time. The performance model is tested for different number of available ERUs.

Chapter 10 documents conclusions and a number of possible future research directions corresponding to each problem.

Chapter 2: Literature Review

In this section, we review previous findings from related studies. We start reviewing from fundamentals of incident occurrences (i.e., including theories for identification, prediction, and interpretation of secondary crashes) to two important decisions for arrangement of emergency vehicles (i.e., dispatching and relocation).

2.1 Identification of Secondary Crashes

The identification of secondary crashes has focused on representing the temporal and spatial thresholds from the impact of primary events, and is classified in two main categories [7, 32, 33]. A static impact area is determined by maximum clearance length and time [34–37]. Compared to the static thresholds, dynamic thresholds models conclude that an incident should not be classified as secondary when it occurs far from the primary location of the event without congestion (Figure 2.1).

Different aspects of dynamic models include:

- *Simulation modeling.* It replicates rubbernecking by proportionally increasing the distances at which the vehicles are following one another. It was initiated by the study [20] to identify dynamic thresholds from the shockwave that

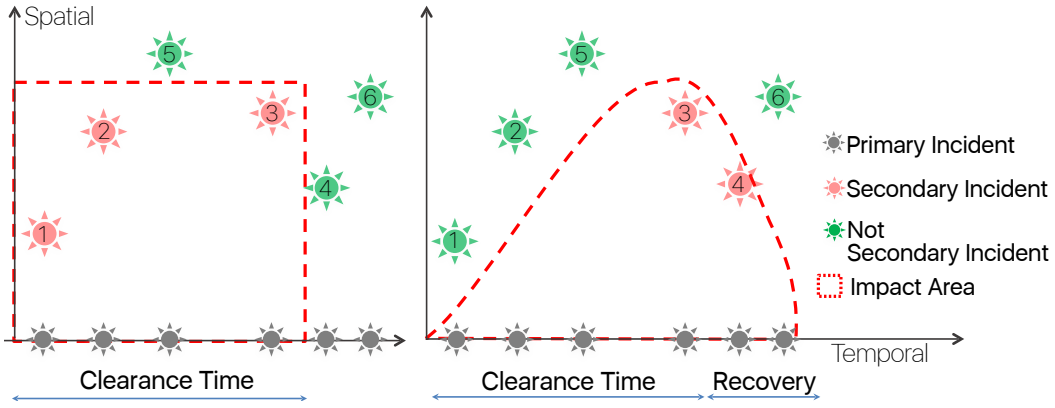


Figure 2.1: Static and dynamic models

arises as a consequence of the incident.

- *Deterministic queuing.* This uses the cumulative arrival and departure curve for deterministic estimate of traffic delays and queue lengths [38]. Deterministic queuing for real-time application might be less realistic, since it assumes exact arrival rate and capacity reduction [25].

- *Closed-circuit television.* Visual devices enable the observation of the progression of the queue formulated at the upstream. The spatial-temporal boundary for each secondary crash is defined based on maximum queue length and the duration induced by the crash [39]. It should be noted, however, that archived incident data collection are expensive and as a result may have limited queuing information.

- *Speed contour plot.* The speed threshold algorithm is widely adopted in bottleneck identification [40]. Automatic Tracking of Moving Traffic Jams (ASDA) is used to capture the propagation of wide moving jam [41]. ASDA is used for spatiotemporal evolution of traffic flow and the propagation of the traffic

disturbance upstream of the incident [42]. However, relying on loop detectors decreases the accuracy of the results. Congestion caused by crashes may not classify pronounced stop-and-go waves described as wide-moving jams [43]. Alternatively, ASDA is more appropriate for use in the context of mesoscopic traffic simulation models [44]. The end of the varying queue is marked to estimate incident progression curve [45]. A set of threshold values are used to classify a freeway segment as a congested segment [46]. Empirically obtained values may be time consuming and difficult to have a robust measurement to apply for other data. It cannot capture the skewness of the data that may appear at congested freeway sections.

In this dissertation, we apply a clustering method to contour map of probe vehicle speed to capture the dynamics of traffic evolution during primary events. It is assumed that individual component speeds may model some underlying set of hidden events with congested condition.

2.2 Prediction of Secondary Crashes

Predictions of secondary crash occurrence depend on accurate characterization of incident duration, and incident duration depends on response-unit arrival time.

2.2.1 Incident Duration Prediction Models

The complex interactions among factors affecting prediction performance make modeling challenging. For several decades, advanced data collection has made it

possible to get useful information about influential factors for incident duration. Researchers have devoted considerable efforts to this imperative issue with various methodologies outlined as following: regression model [47]; decision trees [48–51]; support vector machine [52]; log-normal distribution [53–56]; Bayesian networks and Bayesian classifier [57–59]; discrete choice models [60]; hazard-based duration models [61–66]; fuzzy logic models [67]; and, nearest neighbors [64]. They have an advantage of being easily understood with long history of application, availability of software, and deep-rooted acceptance. However, interactions between contributing factors cause estimated models' coefficients to be sensitive to omission, misclassification and time [68].

While statistical models need to specify an appropriate functional form linking the dependent and independent variables, neural networks do not require the establishment of a functional form. In recent years, artificial neural networks have provided a universal approximation of complex functions [69], especially for prediction of incident duration [70]. Compared to other parametric and non-parametric models, neural network models had a satisfactory accuracy for the incident cases and gave the best result for long duration incident cases with lowest error [71].

From a different perspective, the above models are post hoc models that can be used in planning stage. In this dissertation, sequential real-time models [72–74] are presented.

2.2.2 Secondary Crash Prediction Models

Secondary crashes have low sample mean and a small sample size compared with primary incidents. The wide variety of causes and impacts of non-recurring congestion make it difficult to quantify random and complex incident natures at a system level. As a result, crash prediction models have been over-fitted and have poor predictive performance. Because accident prediction models are non-normal and functional forms are typically nonlinear, it is shown that R^2 is not an appropriate measure [75]. It is difficult to validate secondary crash occurrence and associated delays owing to lack of field data [76]. A comparative literature review presented that neural network models perform better than logit models for classification problems [77]. In this research we take a principled Bayesian learning approach to neural networks to predict the likelihood of secondary crashes.

2.2.3 Bayesian Neural Networks

Traditional neural networks are trained to get a set of weights that minimize the error between the target values and network outputs. Back-propagation neural network (BPNN) models can fit the incident data with high precision [73]. Although BPNN has a good training result, it sometimes provides testing values with unacceptable variances (MATLAB). Starting from early works of [78,79], Bayesian framework has been used for solving complex problems: pattern recognitions [80–83]; motor vehicle collisions prediction [84]; traffic estimation and optimal counting location [85]; earthquake magnitude prediction [86]; and, bridge integrity monitoring [87].

2.3 Interpretation of Secondary Crashes

2.3.1 Pedagogical Rule Extraction

A decision tree is appealing when a good understanding of the process is essential because it has self-explained properties rooted in the structure. Previous incident duration studies have used decision trees to discover patterns in a given incident data set. Most of them are translated into if-then-else rules. However, there are a few shortcomings of decision tree algorithms [88]:

1. Decision trees typically have fewer training observations available for deciding upon the splits or leaf node class labels at lower levels of the trees.

2. Tree induction algorithms are unstable. A small addition or deletion of a few samples make the tree induction algorithm radically different. A greedy splitting selection has no backtracking in the search and is subject to all the risks of hill climbing algorithms, mainly converging to locally optimal solutions.

3. It is difficult to control the size of the trees and sometimes very large trees make comprehensibility difficult. Pruning may reduce the generalization accuracy of the tree. The user may need different size of the trees based on decision variables he or she understands.

An advantage of using a rule extraction technique is that the neural network considers the contribution of the inputs toward classification as a group, while decision tree algorithm measures the individual contribution of the inputs at a time

as the tree is grown. This dissertation introduces a pedagogical rule extraction approach, called TREPAN [89], to improve understanding of the secondary crashes by extracting comprehensible rules from the neural networks. TREPAN has been successfully applied to data mining (i.e., management science and bioinformatics), and presented better performance than traditional decision trees (i.e., C4.5, Neurorule, Nefclass, CART) [89].

2.3.2 Relative Importance of Factors

It is well-known that standardized regression coefficients have been suggested as a measure of importance of factors in regression analysis. However, when variables are correlated, the following conditions are likely to arise [90].

1. An exaggeration of the relative weight of the predictor variable most highly correlated with the dependent variable.

2. A decrease of the relative weight of other variables in the model.

3. Small differences in samples could cause large differences in regression weights.

4. A reversal of signs that could make a variable appear to have an effect the opposite of its true relationship.

Neural networks' predictive power and ability to analyze nonlinear relationships assured various researchers to study the role of variables, overcoming limitations of standardized regression coefficient. Several different algorithms that allow contribution analysis are as follows.

1. Calculation of the partial derivatives of the output according to the input variables [91].

2. Computation of weights using the connection weights [92, 93].

3. Perturbation of the input variables [94].

4. Profile method: a successive variation of one input variable while the others are kept constant at a fixed value [95].

5. Classical stepwise method: an observation of the change in the error value when an adding (forward) or an elimination (backward) step of the input variables is operated [96].

6. Improved stepwise a: the elimination of the input occurs when the network is trained and the connection weights corresponding to the input variable studied are also eliminated.

7. Improved stepwise b: involves the network being trained and fixed step by step with one input variable at its mean value.

Olden [97] compared all methods and concluded the connection weight approach is the only method that consistently identifies the correct ranked importance of all predictor variables, whereas, the other methods either only identify the first few important variables in the network or no variables at all. This dissertation uses the connection weight approach to identify critical relationships between the set of key factors and the resulting incident duration.

2.4 Capacity Adjustment for Estimation of Delay

The remaining proportion of the capacity due to an incident can be obtained in HCM [23]. Several efforts have been made to estimate incident-induced delay on freeway networks using reduced capacity. These methods include data-driven models [55, 98]; dynamic traffic assignment models [99]; and analytical link models [47, 100]. However, it is difficult to quantify an incident′s impact on drivers at different locations. As a result, secondary incidents in delay estimation have not been thoroughly analyzed.

A deterministic queuing method uses the cumulative vehicle arrivals and departures. Total delay caused by a primary and a secondary incident are calculated using following parameters: First, incident duration is defined as the time between the detection and clearance of an incident. Second, reduced capacity depends on the severity of an incident (e.g., number of lanes closed). Third, arrival rate of vehicles is calculated during an incident until the freeway capacity is restored to its normal condition.

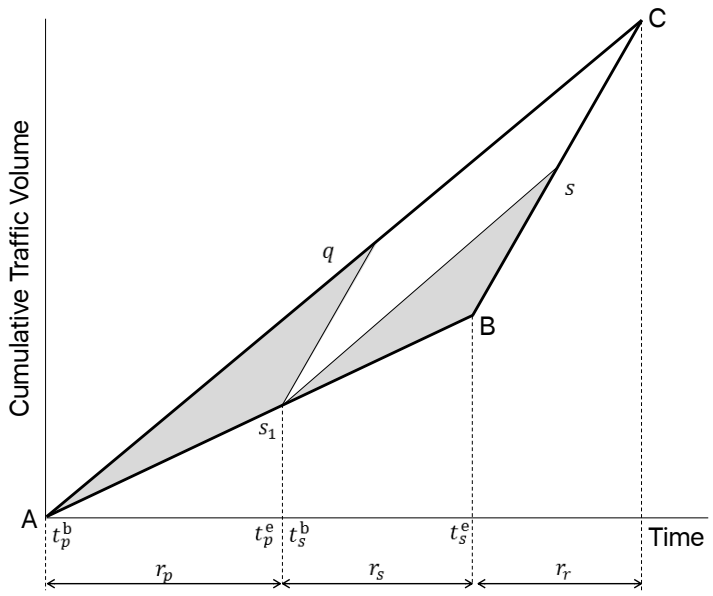


Figure 2.2: Basic deterministic queuing diagram of incident delay.

In Figure 2.2, a primary incident ends at time t_p^e, and in the meantime its secondary incident occurs at time t_s^b. A summation of the triangle areas (i.e., shaded portion) represents the total delay caused by the primary incident and the secondary incident. The durations of primary incidents are expected to be longer if secondary incidents occur as a result of additional impedance and interference [101]. Underestimation of total delay happens if two incidents are treated independently. Instead, the total delay can be calculated [38] as a significantly larger triangle area (ABC), with an extended recovery time (i.e., r_r) and is mathematically formulated.

$$delay = (r_p + r_s)^2 \frac{(s - s_1)(q - s_1)}{2(s - q)} \tag{2.1}$$

where traffic flow rate is q; primary incident duration is $r_p = t_p^e - t_p^b$; secondary incident duration is $r_s = t_s^e - t_s^b$; reduced capacity for the primary incident and the secondary incident (i.e., during the response and clearance times of the incident) is

s_1; and the normal capacity is s.

However, the example described in Figure 2.2 is a very rare case and a *gap*, or an *overlap* between occurrences of two incidents exists. In general, a secondary incident occurs before the primary incident is cleared, or before the traffic conditions are fully recovered.

In this dissertation, we introduce a delay estimation method that allows adjusting the capacity estimates to account for deviations from standard conditions. Traffic management agencies can overcome the following challenges by using our proposed method to accurately estimate the incident-induced delay.

2.5 Dispatching of Emergency Vehicles

Various aspects of different ambulance locations or dispatching rules on transportation networks have been investigated in the past. First, static covering models seek to position the least number of facilities needed to cover all points of demand within specified distance or time units [102] or with additional covering servers [103]. P-Median models involve location of facilities to minimize the total weighted distance of serving all demand [30, 104]. When taking service coverage concern, vehicle relocation is needed [105]. Dynamic relocation models pre-compute solutions in anticipation of events in the future stages [106]. Many simulation tools are used to perform a cost effectiveness analysis of emergency ambulance services [107]. Hypercube queuing model is used to incorporate theoretical queuing theory results and simulation as a tool in the dispatching police patrol cars [108] and deploying

emergency ambulances [109].

We focus more on reviewing dynamic and probabilistic dispatching models that became the motivation of this dissertation. The uncertainty in travel times of emergency vehicles was addressed [110]. Dynamic models were developed for available real-time information for more flexible time-dependent vehicle deployment [29]. Online dispatching and re-routing were considered to minimize the response time of emergency vehicles [111]. Integrated dispatching and districting policies were proposed to improve the performance of emergency medical service systems in terms of patient survival probabilities [112]. In a later study, sequential arrival of patients was considered to dispatch ambulances optimally by introducing a set of equity constraints [113]. Recently, future time-stages were considered in location-allocation of emergency response vehicles was considered [31]. The principle was to begin with a location phase for managing initial incidents and to progress through a routing phase for managing the occurrence of next incidents. The authors offered a stochastic solution to this problem by characterizing uncertainty using probability distributions.

Traditional models need input frequency of requests given in advance to get approximate answers. For example, incidents were assumed to occur on the nodes of the network in a Poisson manner with known rates [104]. However, even though fictitious play is "belief based", it is also myopic. The choice of scenarios in a probabilistic model requires data that may not be readily available about the historical request sequence that have been observed in the past, as well as faith that the future will resemble the past. It sometimes fails to capture important real-world emergency

scenarios. Therefore, a single incident rate, assuming independency between two incident sites [5, 6] cannot successfully dispatch appropriate units.

Making a sequence of decisions under unexpected events, online algorithms have been extensively studied [114]. However, a creteria of no-information about the future input is often too pessimistic. It does not involve full computation of optimal strategies. More importantly, they are not learning the"true model" considering how other decisions would actually influence on payoffs. Compared to scheduling problem, we have to make some decisions that are irrevocable along the way. We enhance the online algorithms by characterizing partial distribution conditioned on current emergency and real-time traffic information.

In this dissertation, we propose an online model that minimizes the response time to the next request while making it as close to the optimal response as possible. Without knowing everything about the future, the online algorithm may turn out not to be optimal, but we focus on the quality of decision that is compared against an adversary on a worst-case input.

In our dispatching problem, some independent emergencies occur at unpredictable locations at unpredictable times. Our practical online model has a look-ahead contingent on present emergency in making future decision of which vehicle to assign. We characterize uncertainty of future emergencies conditioned on information of currently available emergencies. The limitation of previous studies presents the urgency with which a new approach is needed. We introduce the concept of online dispatching strategies and apply the online model to a transportation network. The experimental evidence indicates the algorithm works well in practice.

2.6 Relocation of Emergency Vehicles

We focus on reviewing discrete location problems since the response units are restricted to a finite set of candidate locations. Several approaches have been proposed to solve deterministic, probabilistic, and dynamic problems of optimal facility locations.

The earlier versions of deterministic model are covering theories, such as location set-covering problems [102]. They provide coverage to all demands within a pre-determined distance range. The maximal covering location problem seeks the maximum population served within a stated service distance [115]. This model was extended to account for the chance when a demand arrives at the system that is engaged to serve other demands [103]. P-center models are equivalent to covering a given area in the plane having p identical circles where facilities are located at the centers of these circles [116].

On the other hand, a probabilistic formulation was proposed to overcome the limitations of deterministic models . P-Median models involve location of facilities on a network to minimize the total weighted distance of serving all demand [30]. One can use the maximum availability location problem [117]. An upper bound was imposed on the probability that a call on demand point does not receive immediate service [118]. To incorporate the busy probability, queuing-based models consider customers waiting for service in congested systems [108]. A spatial queuing model considers spatial and temporal demand characteristics such as the probability that a server is not available when required [119].

Location models have been applied to incident management to find optimal locations of response units. An optimal deployment of ERUs depends on incident rate at marked location and consequent delay. Optimal beat structure and truck allocation assuming that the probability of incident occurrences follows a Poisson distribution [6]. A single incident rate, assuming independencies between two incidents, has been considered [5, 55]. It assumes that all subsequent incidents are independent of previous incident, and have the exponential distribution. However, the freeway degrades from primary-incident state to secondary-incident state when a secondary incident occurs [24]. Crash risk is higher in the presence of an earlier crash [7]. Incidents frequently cause unexpected delay due to larger traffic demand than reduced capacity [8]. After a primary incident occurs, the resulting bottleneck quickly forms a queue and, the likelihood of secondary incidents and associated delay increase. Although emergency operators manage to handle a primary incident (i.e. the first incident) or an independent incident with this assumption, drivers suffer heavily when another incident, a secondary incident (i.e. an incident within temporal and spatial impact of a primary incident), occurs. However, a Poisson process does not consider dependencies in incident occurrences. Unfortunately, under traditional Poisson models, handling secondary incidents without prompt response and clearance may cause a critical issue in the efficient mitigation of incidents. Regardless of the initial response, the serving time is greatly influenced by efficiency of response-unit arrivals and consequent clearance. In our stochastic model, the probability matrix of a sequence of primary and secondary incidents varies for each request arriving in real-time.

Compared to these static models, dynamic models consider sequence of requests that are revealed incrementally over time. A mathematical model was proposed to deal with time-dependent vehicle dispatching and rerouting [29]. Solutions are computed one-by-one in an online fashion, while minimizing the response time of emergency vehicles [111]. Dynamic double standard models incorporate practical dimensions addressing the dynamic nature of the problem [120]. The real-time relocation models take service coverage concern when ERUs are dispatched [105]. Dynamic relocation models pre-compute solutions in anticipation of events in the future stages [106]. Recently, an interesting problem of determining stochastic emergency vehicle redeployment for an effective response to traffic incidents was introduced [27]. The problem under uncertainty was treated in a particularly elegant way by adjusting the scheduling plan to reposition emergency vehicles in response to service calls. In this dissertation, we estimate the number of available servers by comparing remaining time to clear the current incident and time to next incident occurrence.

Alternatively, Markov decision processes (MDPs) were used on dynamic relocation of service units in early works [121, 122]. A tree-search heuristic was applied for approaching optimal relocations to the Stockholm region in Sweden [123]. A MDP approximates response time distribution and the distribution of the number of busy ambulances to identify near-optimal compliance tables [124]. Recently, a look-ahead scheme was applied in ambulance locating models to approximate the temporally accrued rewards and discounted probabilities [125]. However, the first order Markov decision process does not capture the conditional probability of future

secondary-incidents that depends on past and present incident occurrences. To the best of our knowledge, all previous studies assume two incidents are independent without considering their spatial and temporal dependencies. In this dissertation, an analytical model is proposed to relax the restrictive assumptions of previous models and reveal mutual relationship between incidents at each site in a sequence of time stages.

System costs will be excessive if delay regarding allocation decisions is ignored when locating response units. The objective of the location-allocation problem is to accurately capture the cost of multiple-stop routes within a location model (see a comprehensive review and perspective on these models [126]). This dissertation incorporates a realistic stochastic process into the design of ERU deployment. Two decision levels are integrated for the optimal deployment of response units: a location decision of response units before an incident occurrence, and an allocation decision of vehicles after the incident occurrence. Potential delay caused by inefficient response to secondary incidents is unknown until the primary incidents / information is given. In response to secondary incidents accounting for a significant portion of traffic delays, strategic concerns to emergency responders have been growing. Fortunately, scientific breakthroughs enabled us to develop thresholds as a consistent definition of secondary incidents [7]. This dissertation uses reliable traffic information (i.e., INRIX) and tracks each ERU performance to easily accommodate real-time operations.

Another assumption of previous studies is a returning rule that limits the response units to be always dispatched from an original location. This assumption may

create an unnecessary trip to the designated location and impose hard constraints for next incidents that occur when an ERU is returning. In this study, dispatched units stay at an incident site after the clearance of the event instead of returning to their permanent or temporary place, because the plan is re-generated in the next time. The new assumption can reduce the complexity of the model without hard constraints.

Chapter 3: Stochastic Process of Incident Occurrences

In this section, we introduce a process of future stages of incidents. Each sequence of incidents represents a scenario that is represented in a matrix form with an expected probability. This section justifies learning about secondary incidents to provide a principle for stochastic incident occurrences.

3.1 Probability of Incident Occurrences

The incident occurrence includes accumulated probabilities of secondary incidents in future steps, in which the impact of primary incidents overlaps. In general, a secondary incident may occur during the clearance or recovery of a primary incident. Therefore, we look-ahead two future stages. For example, the conditional probability of a secondary incident at site 2 at the first future-stage may depend on the probability of a primary incident at site 1 during the past and site 3 during the current stage; at the second-future stage may depend on the probability of a primary incident at site 1 and site 3 during the current stage (Figure 3.1).

Let $\tau(i, r)$ be normalized probability of incidents (probability of incidents at site i over for all locations ($i \in H$) in one stage) for each stage r. The expected probability of incidents $\mathbb{E}[\tau(i, r)]$ for each site ($i = k$) and stage ($r = u$) is a sum of

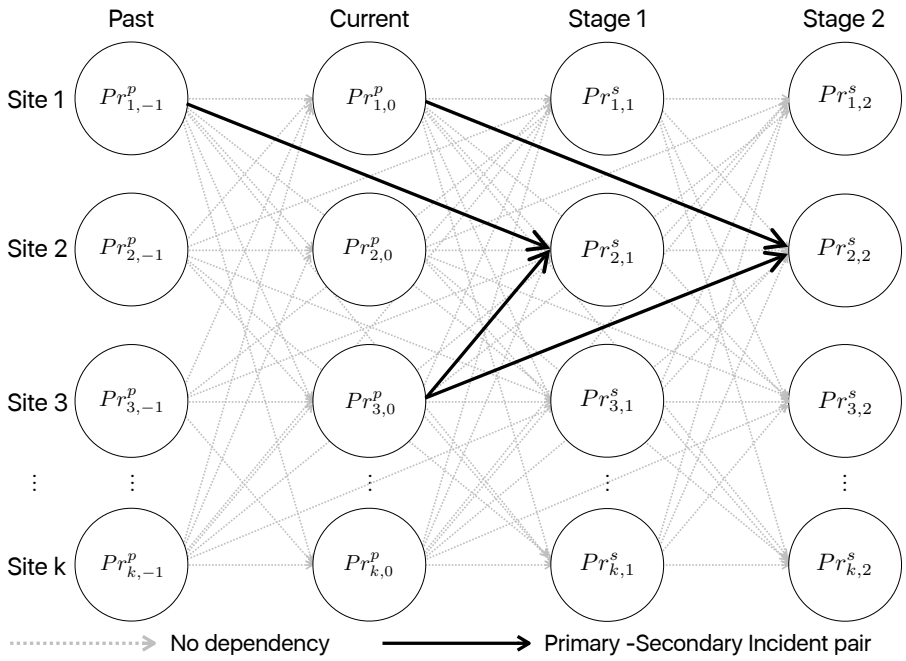


Figure 3.1: Stochastic process of incident occurrences (two future stages)

$Pr^p_{i,r}$ and $Pr^s_{k,u}$. $Pr^p_{i,r}$ denotes corresponding probability of primary and independent incidents at site i during stage r, and $Pr^s_{k,u}$ denotes corresponding probability of secondary incident occurrences at site k during stage u.

$$\mathbb{E}[\tau(i,r)] = Pr^p_{i,r} + Pr^s_{k,u} \qquad for\, i = k, r = u \tag{3.1}$$

First, we use the Poisson process [3, 6] to define $Pr^p_{i,r}$ because primary and independent incidents satisfy the IID assumption. Let parameter λ be the average number of incidents on a freeway network in a given continuous time interval T. We assume that subintervals, times between successive incidents, are exponentially distributed. An empirical analysis [5] presented inter-arrival time of incident on I-695 follow exponential distributions. They presented 8 incidents morning peak hour, one incident every 18 min, and 20 min of average incident duration. The same

freeway corridor (I-695) is used in this study. The average of subintervals is $T\lambda^{-1}$ (with the variance $T\lambda^{-2}$). The discrete random incidents are assumed to be Poisson distributed with incident rate λ_i^r indicated by $\boldsymbol{X} \sim Poisson\ (\lambda_i^r)$. Using probability mass function where the count of incidents is one, the normalized probability of incident occurring at location i for each interval r is

$$Pr_{i,r}^{p} = \lambda_i^r e^{-\lambda_i^r} (\sum_i \lambda_i^r e^{-\lambda_i^r})^{-1} \qquad \forall i, r \tag{3.2}$$

Second, the probability of secondary incidents $Pr_{k,u}^{s}$ is a function of $Pr_{i,r}^{p}$ conditioned on severity (Ω: number of blocked lanes, collision with injuries or property damage) and traffic condition at upstream (Δ: difference in speed before and after incident occurrence) of a primary incident. These are used as the main influential contributors for secondary incident occurrences [22]. Each primary incident at site i during stage r has different impact on future secondary incident occurrences. We introduce an indicator function, $I(\Omega, \Delta)_{(i,r)(k,u)}$, that equals 1 if a primary incident at site i during stage r causes a secondary incident at site k during stage u, and 0 otherwise. The primary-incident density ratio $\delta(\Omega, \Delta)_{(i,r)(k,u)}$ is defined to measure relative difference ratio and is not equal to 0 only when an interrelation between incidents exists (For example, in Figure 3.1, the bold line from $Pr_{3,0}^{p}$ to $Pr_{2,1}^{s}$ is $I(\Omega, \Delta)_{(3,0)(2,1)} = 1$). With introduced parameters and variables, we propose the probability of secondary incidents $Pr_{(k,u)}^{s}$ in an explicit form:

$$Pr_{k,u}^{s} = \sum_i \delta(\Omega, \Delta)_{(i,r-1)(k,u)} Pr_{(i,r-1)}^{p} + \sum_i \delta(\Omega, \Delta)_{(i,r-2)(k,u)} Pr_{(i,r-2)}^{p} \tag{3.3}$$

Now, we insert the $Pr^{s}_{k,u}$ from Equation (3.3) to Equation (3.1). Suppose we are interested in incidents at site 2 in the first future-stage. The expected probability of incidents is:

$$\begin{aligned}\mathbb{E}[\tau(2,1)] = Pr^{p}_{(2,1)} + \sum_{i} \delta(\Omega,\Delta)_{(i,0)(2,1)} Pr^{p}_{i,0} \\ + \sum_{i} \delta(\Omega,\Delta)_{(i,-1)(2,1)} Pr^{p}_{i,-1}\end{aligned} \tag{3.4}$$

The probability of each scenario composed of a sequence of incidents is introduced in a matrix form. Suppose there is a past incident at site 2 and a current incident at site 3. The combinatorial of future incidents (during $r+1$ at site i, $r+2$ at site j) produce $i \times j$ scenarios with probability $p(i,j)$.

$$\begin{array}{cc} r-1 & r \\ 1 & 1 \\ (2) & 2 \\ 3 & (3) \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ m & m \end{array} \quad \begin{array}{c} \\ \\ \\ r+1 \\ \\ \\ \\ \\ \end{array} \quad \begin{array}{c} r+2 \\ \begin{bmatrix} p(1,1) & p(1,2) & \cdot & \cdot & \cdot & p(1,j) \\ p(2,1) & p(2,2) & \cdot & \cdot & \cdot & p(1,j) \\ p(3,1) & p(3,2) & \cdot & \cdot & \cdot & p(1,j) \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & \cdot & & \\ p(i,1) & p(i,2) & \cdot & \cdot & \cdot & p(i,j) \end{bmatrix} \end{array} \tag{3.5}$$

The scenario space $ij(=\omega)$ is divided by two cases with probability that 1) a single incident occurs at each site: $p(\forall i \neq j)$ and 2) two incidents occur at the same site: $p(\forall i = j) = 1 - p(\forall i \neq j)$. Given the information that incidents already occurred at site 2 and site 3, the expected probability of scenarios (P_{ω}) is:

$$
\begin{aligned}
P_\omega = p(\forall i \neq j) \times & \begin{bmatrix} p(1,2) = \mathbb{E}[\tau(1,1)] \times \mathbb{E}[\tau(2,2)] \\ p(2,1) = \mathbb{E}[\tau(2,1)] \times \mathbb{E}[\tau(1,2)] \\ p(1,3) = \mathbb{E}[\tau(1,1)] \times \mathbb{E}[\tau(3,2)] \\ . \\ . \\ . \\ p(i,j) = E[\tau(i,1)] \times E[\tau(i,2)] \end{bmatrix} \\
+ p(\forall i = j) \times & \begin{bmatrix} p(1,1) = \mathbb{E}[\tau(1,1)] \times \mathbb{E}[\tau(1,2)] \\ p(2,2) = \mathbb{E}[\tau(2,1)] \times \mathbb{E}[\tau(2,2)] \\ p(3,3) = \mathbb{E}[\tau(3,1)] \times \mathbb{E}[\tau(3,2)] \\ . \\ . \\ . \\ p(i,j) = \mathbb{E}[\tau(i,1)] \times \mathbb{E}[\tau(i,2)] \end{bmatrix}
\end{aligned}
\tag{3.6}
$$

Note that the IID sequence assumes $p(1,2)$ and $p(2,1)$ are same. However, it is obvious from the equation that their expected probabilities are different ($E[\tau(1,1)] \times \mathbb{E}[\tau(2,2)] \neq \mathbb{E}[\tau(2,1)] \times \mathbb{E}[\tau(1,2)]$).

3.2 Expected Clearance Time

The server availability is an important component of the ERU deployment model. If expected available time of a busy ERU is earlier than expected occurrence time of the next incident, we can include that ERU to be one of available servers. This section extracts clearance time for each location to be used as an input parameter in emergency response problem in Chapters 8 and 9.

Clearance time has a significant influence on total delay [8]. For example, total delay, D_i, for each incident location i can be estimated using variables considered in Highway Capacity Manual 2010: traffic flow rate q_i; reduced capacity (i.e. during the response time R_i to incident site i and normal clearance time NC_i of the incident) $s_i^{'}$; and the normal capacity, s_i (i.e. during recovery). Since the total delay is a convex function of incident duration, the average delay for all vehicles affected by the incident is defined as the total delay divided by the total number of affected vehicles:

$$D_i = (R_i + NC_i)\frac{q_i - s_i^{'}}{2q_i} \tag{3.7}$$

Uncertainty of incident clearance duration is another major challenge in quantifying the impact of incidents [17]. Especially, the response delay to incidents is unknown. While existing studies considered response time to be the time between when the responding agency is notified and when the *first* response-unit arrives at the scene, arrivals of the *secondary* response units, e.g., Coordinated Highways Action Response Team (CHART), fire-board, and towing, have significant influence on clearance operation (Figure 3.2). In our optimization model, the main source of

delay is the sum of response time, response delay, and clearance time. We need a clearance time that is separated from traditional definition.

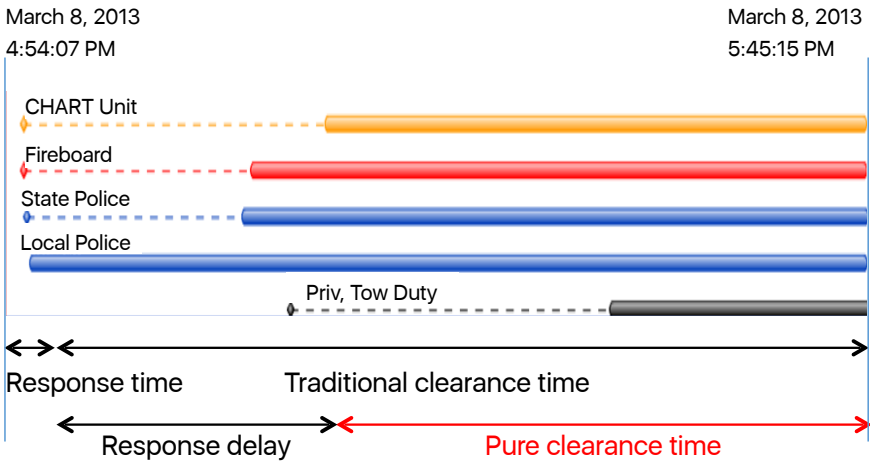


Figure 3.2: The concept of pure clearance time [22]

Potentially delayed clearance can be modeled by integrating delay-type with normal clearance time. A test [22] reveals that time to clear the incident is significantly longer when combinations of response units are delayed. Instead of the original delay graph, a new figure presents the concept of pure clearance time.

We define β_i^η as an indicator of response delay (categorized for each type η: 1 = no delay, 2 = CHART delay, 3 = other response delay, 4 = CHART and other response delay, 5 = not responded by CHART), to extract pure clearance time C_i (when $\eta = 1$) from traditional normal clearance time NC_i at each location i,

$$C_i = NC_i\beta_i^\eta \tag{3.8}$$

In our optimization problem, the clearance time without delay is used as an input to minimize the total delay. For example, when we have the delay type (β_1^1=0.68) at location 1, the value of clearance time purely depends on the characteristic of

incidents (C_1) which is 68% of normal clearance time (NC_1). In this way, we have less chance of overestimating clearance times. Our main goal is getting required ERUs to the incident site as quickly as possible to reduce total incident-induced delay. More details are provided in [22].

Chapter 4: Detection of Delay and Secondary Crashes

We propose the properties of incident impact. A framework is introduced to estimate the feasible area for secondary crashes in real-world cases.

4.1 Problem and Assumptions

The relationship between primary and secondary crashes is revealed by real-world degradations of traffic conditions. Previously suggested thresholds and measurement parameters provide no universal definition. The definition of secondary crashes has still not been finalized. We provide a methodology that would apply to any incident, at any time and location, having available speed data collected from any type of speed sensors. This dissertation answers the following questions:

1. How can we estimate impacts of incidents under varying traffic conditions? Congestion can be defined as a localized section of highway that experiences speed reduction due to inherent delays resulting from recurring or nonrecurring events. This dissertation develops a systematic methodology to classify congested and non-congested conditions.

2. How can we define a highway segment around a boundary of congested and

uncongested condition? Once traffic is congested, crash severity is greatly reduced when all lanes present similar flow conditions [53]. On the contrary, rear-end crashes may occur at the tail of the queue due to large differences in speed [127]. This dissertation introduces an advanced statistical tool to accurately judge if a highway segment is classified as a congested segment.

3. Which incident cases should be regarded as primary incidents? How should we define secondary incidents without noticeable congestion under low traffic demand? The effect of incidents (e.g. vehicle on fire, weather conditions, road maintenance, disabled vehicle) as potential primary incidents is considered minor. This dissertation investigates blocked lane cases only. Moreover, if abrupt speed reduction does not exist, the proposed algorithms for identifying incident impact are not required. Without congested conditions, we assume secondary crashes only occur within incident clearance and upstream within one mile due to the relationship between secondary incidents and incident duration.

4.2 Methodology

Accurately identifying secondary crashes is a challenge. Misclassified incidents lead to biased modeling and unreliable decisions on emergency systems.

4.2.1 Secondary Crash Feasibility Area

Analytical congestion models present a situation in which road users cannot drive at their desired speed. Instead of subjective boundaries, proximity limits can be determined from the mean occupancy rates of each road segment [20]. Defining the boundaries of the impact area from an incident can approximate real traffic conditions with high density (Figure 4.1).

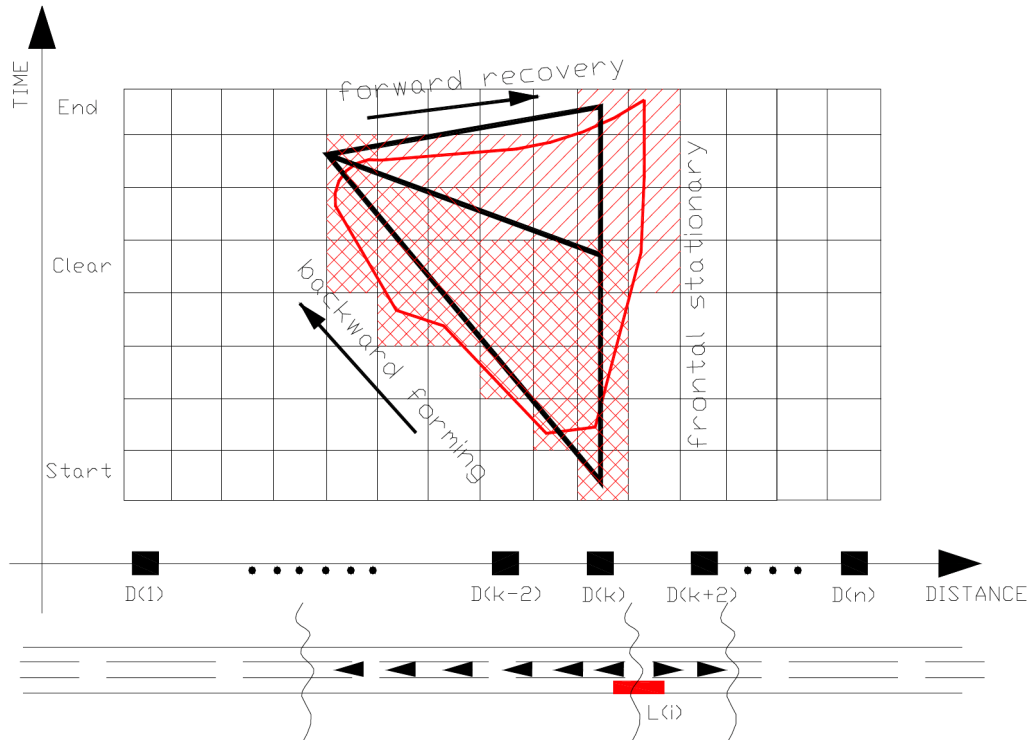


Figure 4.1: Incident impact defined by high density area [20]

The characteristics of traffic conditions within the high-density area represent the boundary between the congested and recovery conditions. For example, a shockwave is characterized by the sudden change in the vehicle speeds downstream of a disturbance. However, the shockwave formation method does not consider the nonlinearities existing in the queue formation. Another analytical procedure, a de-

terministic queuing model, needs a threshold for capacity reduction not suitable for real-time application.

This dissertation adopts the idea of detecting congestion in Figure 4.1. The travel speeds of probe vehicles are represented on traffic message channels (TMC) to account for a feasible area with speed variations near the incident location. Given the speed profiles, the analysis of the secondary crashes determines the piece-wise time and space extent of the feasible region. The hypothetical correlation between the secondary crashes and the primary incidents can be examined using the time-space evolution of disturbance boundaries while considering the effects of isolated incidents. The impact of an incident is schematically described in a speed contour map on a specific day on freeway segments (Figure 4.2).

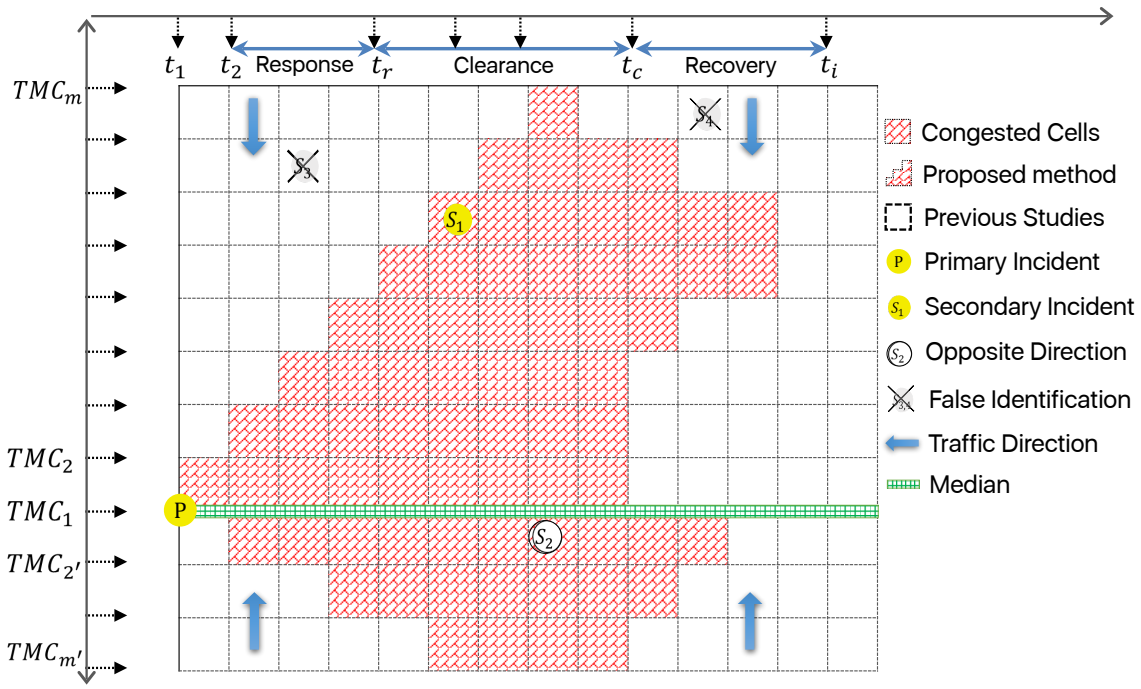


Figure 4.2: Systematic spatial-temporal freeway sections impacted by an incident

We use the following procedure to check each coordinate of congestion boundaries on the contour plot.

- *Step 1.* Build a speed contour plot using speed measurement: Each cell represents speed measurements, $\hat{V_s}(t_n)$, for a section s at a time interval, .

- *Step 2.* Decide whether each speed cell is in congested or non-congested condition and build a binary speed contour plot. A red dot shaded cell describes congestion speeds affected by incident impact, compared to the distribution of historical speed measurement, $V_s^i(t_n)$, for i^{th} day of the week ($i = 1, 2, , 7$ analogous to the day of the week from Monday to Sunday). The proposed methodology will be discussed in the next section.

- *Step 3.* Draw boundaries over time. Detect whether each incident is related to primary incidents. After the occurrence of a primary incident P at time t_1, the traffic grows in upstream direction. A formation of queues and shockwaves can be observed until the incident is partially cleared at t_c or completely cleared at t_r and the dissipation of the queue to normal traffic conditions can be shown when the traffic condition is fully recovered at t_n. By summing the adjacent dot-shaded areas, the total queue lengths for each interval are estimated. The maximum queue, right before the partial clearance of incident with reopened lane at time t_{so}, can be calculated as $TMC_m - TMC_1$.

The essential idea of estimating the feasibility area is to check each coordinate of congestion boundaries on the contour plot. A crash S_1 that occurs within congestion area paring with primary incident, P can be defined as a secondary crash.

On the country, crash S_3 and S_4 that occur in the outside of the region of feasibility cannot be related with primary incident and regarded as an irrelevant incident. The crash, S_2 that occurs in the opposite direction within the congestion area during total incident duration $(t_f - t_i)$ is assumed to be a potential secondary crash due to rubbernecking.

4.2.2 A Gaussian Mixture Model

Contrary to the static approach [15], this dissertation presents variable criteria to decide whether each road segment is under a congested or non-congested condition. Speed data are available as early as 1-min time intervals from INRIX traffic data, and using shorter interval may enable microscopic estimation of impact area. Many bottleneck detections are based on 5-min aggregated data for stable performance of algorithms [15, 40]. In this dissertation, it is assumed that traffic conditions of each segment have unique patterns for each time interval of 5 min.

To estimate congested regions affected by an incident, a scientific approach is introduced. This approach provides a visible interpretation of each speed state and informs the frequency and magnitude. As a side benefit, the interval from the model tells us whether any particular $\hat{V}_s(t_n)$ is drawn from the distribution of $V_s^i(t_n)$ or not. Each incident case has a coefficient of variation (COV) indicating the variation of speed on the segment during peak hours in a day. Since the mean value of speed may fall in any range across different segments, the coefficient of variation in speed would be a better indicator compared to variance itself. If representative speeds on

a segment do not change significantly, the travel speed does not change during the peak hour. The lower the speed cell is, the more potential the daily COV has to be under congestion.

Our job is to make sense of this data, even though no one has provided us with correct labels. First, we must make sense of clustering. Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups (Figure 4.3). Each data point contains corresponding speed measurements and COV that have potential to be under congested or non-congested conditions.

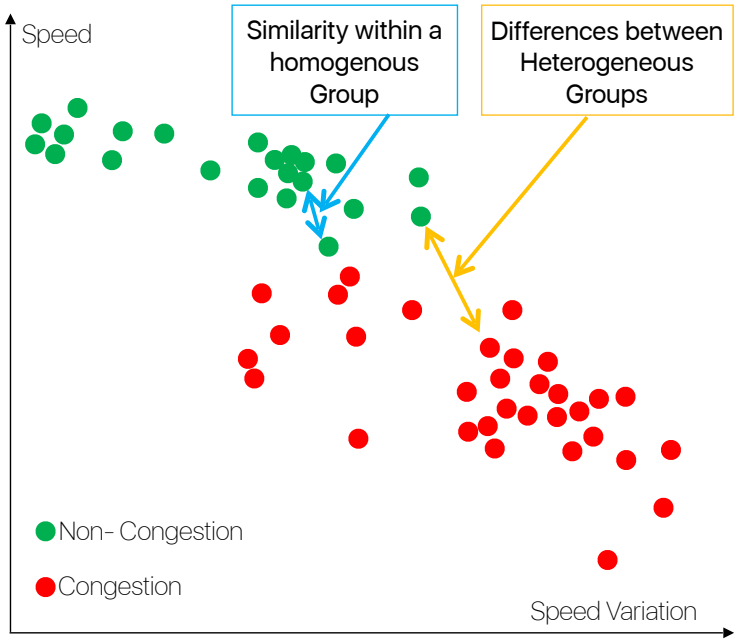


Figure 4.3: Congestion versus non-congestion

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [128]. This model has been successfully used [16] to uncover temporal relations by classifying the consecutive time windows into similar error patterns. The application of GMM

to represent distribution in traffic conditions was motivated by the intuitive notion that the individual speeds can be grouped by hidden events describing congested or non-congested conditions. Due to unlabeled features to train the GMM, classes are hidden based on the level of services of an observation.

This dissertation assumes that the speed measurements and coefficient of variations are related to features corresponding to traffic conditions. The level of services reflects general configurations useful for characterizing congestion identity. In turn, the spectral shape of the class can be represented by density of the congested or non-congested condition, and variation of the average spectral shape can be represented by the covariance matrix. Since data are distributed in different areas of the space, we must decide how much weight to give to each group of the networks. The Gaussian density distribution, given N input vectors, can be written in the form

$$g(x|\mu,\Lambda) = \frac{1}{2\pi^{\frac{D}{2}}}\frac{1}{|\Lambda|^{\frac{1}{2}}}exp\Big\{-\frac{1}{2}(x-\mu)'\Lambda^{-1}(x-\mu)\Big\} \tag{4.1}$$

where x is a D-dimensional continuous-valued data vector, μ is a mean vector, Λ is a covariance matrix, and $|\Lambda|$ denotes the determinant of Λ.

A linear superposition of two Gaussians better characterizes the data set. Such superposition, formed by taking linear combinations of more basic distributions such as Gaussians, can be formulated as probabilistic models known as mixture distributions. A Gaussian mixture distribution can be written as a weighted sum of M component Gaussian densities as given by:

$$P(x|\lambda) = \sum_{i=1}^{K}\omega_i g(x|\mu_i,\Lambda_i) \tag{4.2}$$

Each Gaussian density $g(x|\mu_i,\Lambda_i)$ is called a component of the mixture and

has its own mean and covariance. The parameter $\omega_i (i = 1, 2, ..., M)$ is called a mixing coefficient. If we integrate both sides of equation 4.3 with respect to x and note that both $P(x)$ and the individual Gaussian components are normalized, and we obtain

$$\sum_{i=1}^{K} \omega_i = 1 \tag{4.3}$$

Also, the requirements $P(x|\lambda) \geq 0$ and $g(x|\mu_i, \Lambda_i) \geq 0$ implies $\omega \geq 0$ for all i, and we obtain

$$0 \leq w_i \leq 1 \tag{4.4}$$

Given training vectors and a GMM configuration, we need to estimate the parameters of the GMM, λ, to best match the distribution of the training feature vectors. The most popular and well established method is the maximum likelihood (ML) estimation. The aim of the ML estimation is to find model parameters that maximize the likelihood of the GMM given the training data. Assuming independence between a sequence of T training vectors $X = \{x_1, ..., x_T\}$, the GMM likelihood can be written as:

$$P(X|\lambda) = \prod_{t=1}^{T} p(x_T|\lambda) \tag{4.5}$$

ML parameter estimates can be obtained iteratively using Maximum a Posteriori estimation. The first step is where estimates of the sufficient statistics of the training data are computed for each mixture in the prior model. The second step is for adaptation; these "new" sufficient statistic estimates are then combined with the "old" sufficient statistics from the prior mixture parameters using a data-dependent mixing coefficient. The data-dependent mixing coefficient is designed so that mix-

tures with high counts of new data rely more on the new sufficient statistics for final parameter estimation, and mixtures with low counts of new data rely more on the old sufficient statistics for final parameter estimation. Given a prior model and training vectors from the desired class, we first determine the probabilistic alignment of the training vectors into the prior mixture components. Then a posteriori probability for component i is given by

$$P(X|\lambda) = \frac{w_i g(x|\mu_i, \Lambda_i)}{\sum_{i=1}^{K} w_i g(x|\mu_i, \Lambda_i)} \tag{4.6}$$

We then compute the sufficient statistics for the weight, mean and variance parameters.

4.2.3 An Adjusted Boxplot Model

To test if the congested group has points not under the incident impact, this dissertation applies an adjusted boxplot method. This approach defines a case as an outlier if a given speed cell is outside the data interval.

The boxplot is one of the most frequently used graphic tools for visualizing the distribution of continuous data [129]. It can be constructed by putting a line at the height of the sample median Q_2, drawing a box from the first quartile Q_1 to the third quartile Q_3. The length of this box equals the inter-quartile range, $IQR = Q_3 - Q_1$, as a robust measure of the scale. All points outside the interval in Equation 4.7 can be classified as potential incident cases.

$$[Q_1 - 1.5IQR; Q_3 + 1.5IQR] \tag{4.7}$$

However, observations outside the fence are not necessary real incident cases that behave differently from the majority of the data. At thick-tailed symmetric distributions, many regular observations will exceed the outlier cutoff values defined in Equation 4.7, whereas data from thin-tailed distributions will hardly exceed the fence [130]. We use the *medcouple* (MC) to measure the skewness of a univariate sample from a continuous distribution F,

$$MC = med\, h(x_i, x_j) \qquad (x_i \leq Q_2 \leq x_j) \tag{4.8}$$

for all $x_i \neq x_j$, kernel function h is defined as

$$h(x_i, x_j) = \frac{(x_j - Q_2) - (Q_2 - x_i)}{x_j - x_i} \tag{4.9}$$

The medcouple always lies between -1 and 1. A distribution skewed to the right has a positive medcouple, whereas a distribution skewed to the left has a negative medcouple. As shown in [131], we use the exponential model in the definition of our adjusted boxplot to define the boundaries of the interval.

$$[Q_1 - h_l(MC)IQR; Q_3 + h_u(MC)IQR] \tag{4.10}$$

Additionally, we require that $h_l(0) = h_u(0) = 1.5$ to obtain the standard boxplot at symmetric distributions. Note that by using different functions h_l and h_u in Equation 4.11, we allow the fence to be asymmetric around the box so that adjustment for skewness is possible. To decrease potential outliers of the model, lower values of a and b are preferred. For a simple application, we consider fence given by $a = -3.5$ and $b = 3.5$ as suggested by [131].

$$h_l(MC) = 1.5e^{aMC}, h_u(MC) = 1.5e^{bMC} \tag{4.11}$$

We can define speed at section i at time $t_n : S_i(t_n)$, and consider if $S_i(t_n) \leq Q_1 - h_l(MC)IQR$: under crash impact area; $S_i(t_n) > Q_1 - h_l(MC)IQR$: free-flow area. A continuous region affected by crashes can be described and used for identifying secondary crashes.

$$\begin{bmatrix} S_i(t_n) \leq Q_1 - h_l(MC)IQR & \text{under crash impact area;} \\ S_i(t_n) > Q_1 - h_l(MC)IQR & \text{under free-flow area} \end{bmatrix} \tag{4.12}$$

4.3 Numerical Examples

4.3.1 Description of Incident and Traffic Data

The model was developed with travel speeds from a 51-mile section of the I-695 corridor, beginning in the MD-150/Eastern Blvd/Exit 38, and ending at the MD-151/North Point Blvd/Exit 40. Because of local commute patterns, the highest demand and congestion appear during peak hours. This corridor was selected because of the density of traffic message channel (TMC) sections, the availability of continuous probe vehicle travel speeds at five-minute intervals, and the frequency of non-recurrent congestion. It is also a major route to M&T Bank Stadium where the Baltimore Ravens draw tens of thousands of attendees to home games during the National Football League season. The archived incident and probe vehicle database are provided by Center for Advanced Transportation Technology Laboratory at the University of Maryland. Based on incident location, traffic data from TMC codes are used to present the traffic state of each segment

Data from the Vehicle Probe Project comes primarily from the vehicles operating as anonymous probes. The pooling capacity of the probe vehicles detectors defining the time slices accuracy is considered as 5 min. Meaningful travel time information for each TMC segment is achieved after data processing methods of aggregation, filtering and smoothing.

Table 4.1 shows the list of TMC segments covered in the I-695 corridor, including the beginning and endpoint as well as the length of each TMC segment.

Table 4.1: List of TMC Segments on I-695

TMC	Start Lat	Start Long	End Lat	End Long	Length(mi)
110P04555	39.2063	-76.5913	39.2066	-76.6119	1.11
110P04520	39.2968	-76.7426	39.313	-76.7445	1.12
110-04523	39.382	-76.7376	39.378	-76.744	0.45
110-04520	39.3122	-76.7447	39.3118	-76.7447	0.03
110-04519	39.2994	-76.7432	39.2894	-76.7414	0.7
110+04527	39.3922	-76.7071	39.3959	-76.6877	1.08
110N04535	39.4017	-76.5629	39.4019	-76.5683	0.3
110P04512	39.2363	-76.6677	39.2391	-76.6685	0.2
110+04542	39.3346	-76.4904	39.3344	-76.4902	0.02
110P04514	39.246	-76.6749	39.2561	-76.6914	1.15
110P04532	39.4131	-76.604	39.4136	-76.5958	0.44
110P04549	39.2336	-76.5043	39.232	-76.5071	0.18
110-04558	39.2044	-76.6392	39.2025	-76.6347	0.28
⋮	⋮	⋮	⋮	⋮	⋮
110+04519	39.2818	-76.7308	39.2841	-76.7351	0.28
110+04560	39.2067	-76.6424	39.2099	-76.6482	0.38
110+04531	39.4151	-76.6252	39.4129	-76.616	0.53
110+04541	39.3446	-76.4949	39.3379	-76.4937	0.47
110P04523	39.365	-76.7473	39.378	-76.7438	0.94
110N04537	39.3867	-76.5265	39.3899	-76.5339	0.45
110+04546	39.2835	-76.4897	39.2802	-76.4783	0.65
110P04518	39.2768	-76.7253	39.2818	-76.7308	0.47
110+04528	39.398	-76.684	39.4066	-76.6691	1.01
110N04549	39.2319	-76.5071	39.2336	-76.5042	0.19
110-04526	39.3954	-76.6944	39.3928	-76.7065	0.68
110+04529	39.4146	-76.6602	39.4207	-76.6449	0.96
110-04536	39.3899	-76.5339	39.3946	-76.5447	0.67

The incident data along this I-695 corridor are investigated. In total, 30,284 incidents (e.g., disabled vehicle, weather event, road maintenance, collision incidents, vehicle on fire, debris) from May 2011 to September 2013 are collected. Additionally, 1,738 collisions (e.g., fatality, personal injury, and property damage) and lane-

blockage incidents are regarded as candidates for primary-secondary crash pairs.

4.3.2 Modeling Results

The GMM is used to divide the population into subgroups. The TMC segment 110-04523 is used for illustration. Note that a TMC segment in a specific time with speed information has not been labeled as being in a congested or non-congested condition. We classify each event based on an estimate of the proportion of the population that lies in each group. In Figure 4.4, the proportion of congested condition in the population is estimated to be 0.619. The posterior distribution in the window shows that the proportion of events belonging in the non-congested group is certainly between 0.45 and 0.75. There is a high probability that the proportion is between 0.55 and 0.65. Comparatively, the proportion of the congested group is around 0.381.

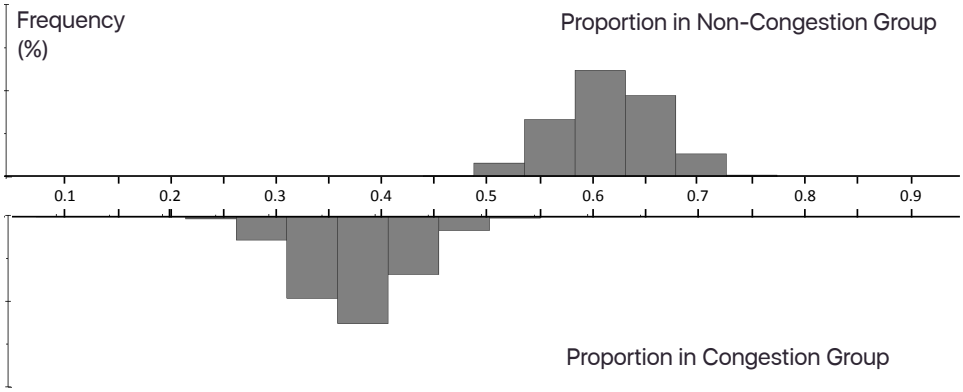


Figure 4.4: Posterior distribution of a population proportion

If label switching, an abrupt shift in the trace plot between groups, occurs during iteration, posterior distribution may not provide a meaningful estimate in a mixture modeling analysis. The graphs in Figure 4.5 show that label switching

didn't occur during 58,000 iterations of the MCMC algorithm.

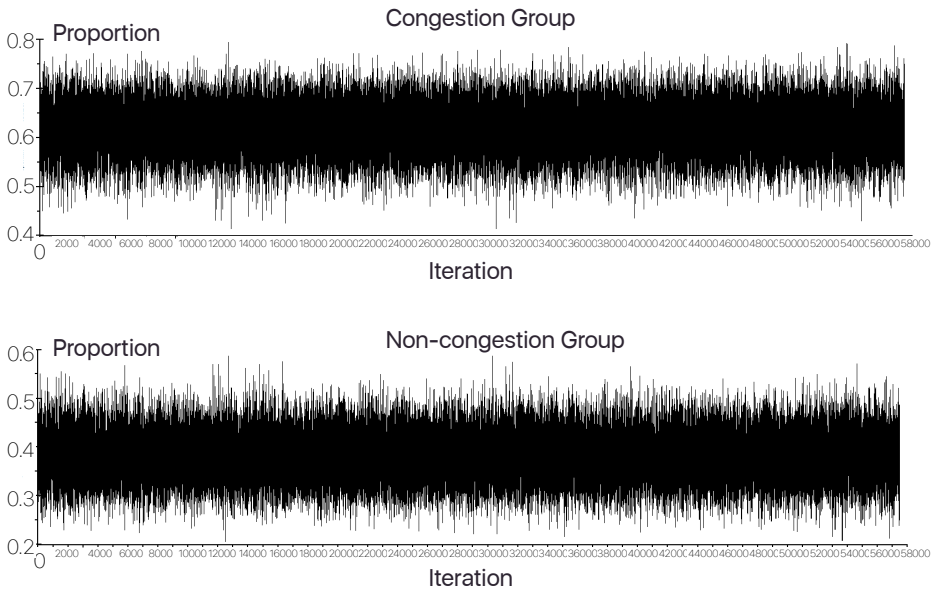


Figure 4.5: Label switching test

A total of 126 speed cases on Fridays for 29 months are analyzed. In Table 4.2, it is clear that the first six cases are placed in congested group with a probability of one, while the seventh case lies in non-congested group with a probability of one. However, case 111 has a lower confidence level. We need to make those confusing data clear. It can be argued that this phenomenon is not a major problem in exploratory data analysis. On the contrary, the observations outside of the fence give an additional graphical indication of the shape of the distribution. Unfortunately, classical methods do not distinguish between 'potential' outliers and 'real' outliers.

We pick congested group for our distribution analysis and present how a proposed adjusted boxplot defines speed data compared to other methodologies. From the normal quantile plot in Figure 4.6, it can be seen that the distribution of speed is right skewed due to recurring congestion at peak hours, with relatively high MC

Table 4.2: Posterior Predictive Distributions (TMC110-04523, 4:30PM)

Case	Date	Speed	Variance	Non-congestion	Congestion
1	5/6/2011	21.4	125.8	0	1
2	5/13/2011	17.6	127.5	0	1
3	5/20/2011	10	149.2	0	1
4	5/27/2011	25.4	102.1	0	1
5	6/3/2011	12.8	134.3	0	1
6	6/10/2011	35	129.4	0	1
7	6/17/2011	62	13.9	1	0
⋮	⋮	⋮	⋮	⋮	⋮
111	6/7/2013	45	167.2	0.62	0.38
⋮	⋮	⋮	⋮	⋮	⋮
125	9/20/2013	11.4	263.4	0	1
126	9/27/2013	18.2	175.6	0	1

value, which equals 0.12. If a distribution describes less congestion, the data will be more symmetric. The smaller value of MC will slightly affect the fence therefore the adjusted boxplots are very similar to the standard boxplot.

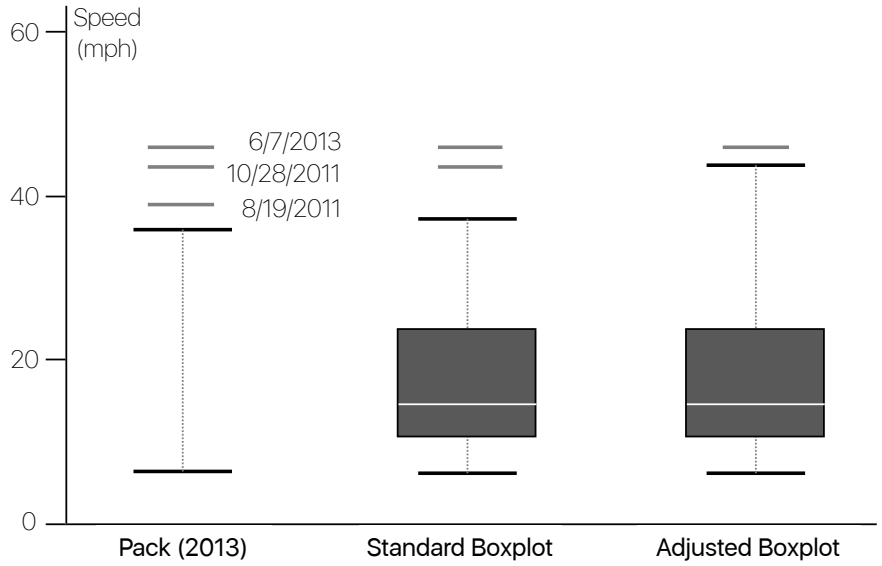


Figure 4.6: A comparison of Pack [15], standard, and the adjusted boxplot approaches

Figure 4.6 describes how a proposed adjusted boxplot defines speed data compared to other methodologies. Pack's theory [15] defines three cases to be uncongested because their values are larger than 60 percent of reference speed. It considers a constant threshold for all cases without the observation of data. Consequently, this static method incorrectly defines a congested case as a non-congested case. In the standard boxplot, it can be seen that underlying distribution is skewed to the right. The median does not lie in the middle of the box and the lower bound is much smaller than the upper bound. Here, two observations exceed the upper bound. Clearly, it would be incorrect to classify them as not congested segments. The proposed adjusted boxplot yields a more accurate representation of the data. The upper bound has become much larger and now reflects better the skewness of the underlying distribution. As a result, the proposed adjusted boxplot causes fewer observations outside of the boundary. Potential secondary crashes occurring at the tail of the queue or at the head of the queue due to large differences in speed can be successfully captured.

Considering the potential influence of outliers in our model, a contour map can be described using the information from each cell being grouped to congested or non-congested cluster. To facilitate the illustration of secondary crash phenomena, 5-min intervals of speed contour plots from onset of incident to recovery are investigated. As shown in Figure 4.7, a speed contour map for I-695 corridor (Thursday, September 26, 2013) is developed. The length of the queue for each time interval is interpreted from traffic speed contour plots by adding the length of each road segment associated with the bottleneck. Red cells represent temporal-spatial area under congestion.

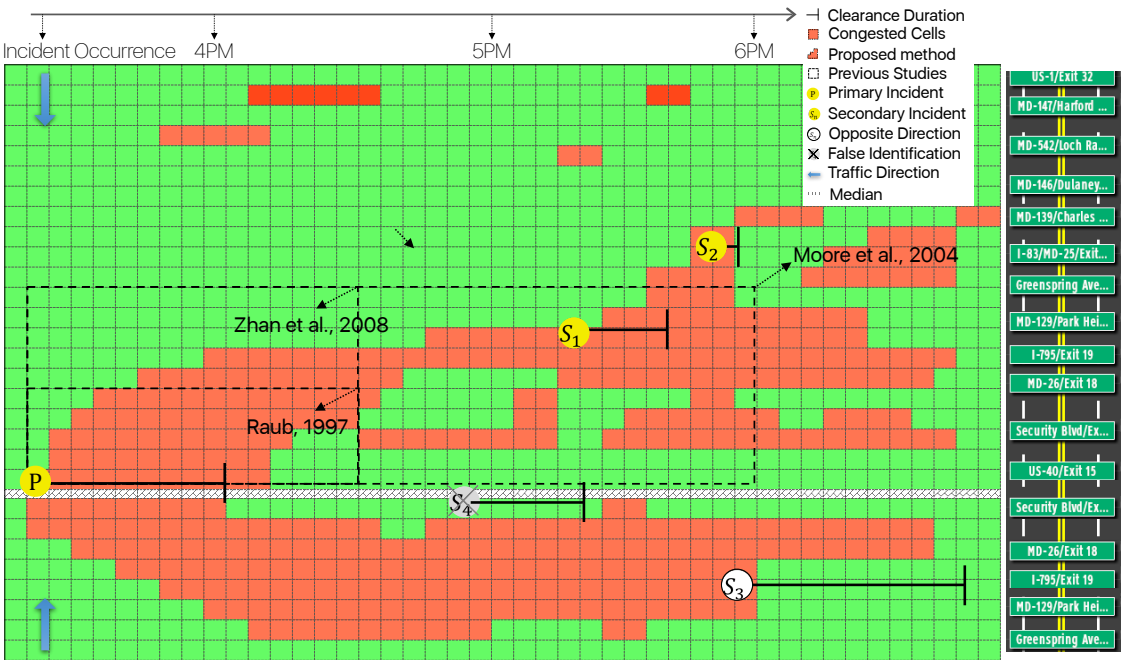


Figure 4.7: Detection of secondary crashes

There are four incidents on the map. First, an incident, labeled as 1, which occurred at Exit 13 at 3:14 p.m., is defined as a primary incident. Incident 1, which involved a two-vehicle collision with injuries, caused a three-lane closure and a 5-operation-unit deployment to clear the scene. Vehicles upstream of the incident are in a slow-moving queue because following vehicles suffer from congestion with traffic conditions rapidly deteriorating from normal driving speed to stop-and-go traffic. The speed reduction from the primary incident may have had an impact on the possibility of secondary crash, labeled as 2, which occurred at the Exit 20 at 5:19 p.m. This may have made the period of congestion even longer and caused an additional secondary crash, labeled as 3, which occurred at the Exit 23 at 5:45 p.m. While the above secondary crashes occurred in the same direction as the primary incident, there is an additional secondary crash, labeled as 4, which occurred in the

opposite direction at the Exit 7 at 5:52 p.m. However, incident 3, outside of the influence of primary incident, is not identified as a secondary crash. The dissipation pattern in opposite direction of primary incident follows dissipation type 1 [42], and incident 3 is classified false secondary crash. When drivers pass the incident, they may speed up to normal driving speed or even free-flow speed. When the speeds of vehicles return to normal after an incident, the queue has dissipated. Traffic flow conditions will return to normal.

An incident′s characteristics indicate different rubbernecking phenomena that perpetuates in the impact area with different intensities depending on the cross and longitudinal location with respect to the incident. Especially, in this larger scale event, multiple secondary crashes have a higher likelihood of occurrence, and their clearance takes longer.

Compared to static threshold methods in previous studies, the probe-based filtering method has superiority. Including the methods proposed by [34–37], only the proposed method can capture incidents 2 as a secondary crash. Moreover, incident 3 in the opposite direction can be identified in the method proposed by this dissertation and [35].

Overall, it was difficult to configure the formation and dissipation of the influence areas upstream of an accident. Table 4.3 shows the total number and percentage of secondary crashes detected by each method respectively. We believe our proposed method describes a real influential area with a good level of certainty, to the extent permitted by the quality of vehicle probe data.

The number of secondary crashes detected by proposed method was 317 out

Table 4.3: Performance Comparison of Secondary Crash Detection (May 2011 to September 2013)

TMC	True	False detected incidents	
	Secondary crashes	Secondary	Primary
Proposed method	317	-	-
Raud [34]	280	0	37
Hirunyanitiwattana and Mattingly [36]	314	20	23
Zhan et al. [37]	348	42	11
Moore et al. [35]	379	68	6

of 1,738 incidents. Comparing the results of each of the methods to the results of the proposed method, the number of false detected secondary and primary incidents are calculated and presented.

An important outcome of the analysis is that all the tested methods tend to wrongly characterize independent incidents as secondary crashes (false detected secondary crashes), or miss others that were secondary crashes (false detected primary incidents). False detected secondary crashes are common for methods that used predefined static thresholds. The main reason for these errors is that static methods do not account for traffic conditions upstream of incidents. Another reason is that the spatial and temporal size of the influence area suggested by these methods is often quite larger than the real influence areas boundaries. As a result, incidents that have no relation with a prior incident can possibly be detected as secondary.

4.4 Conclusions

We contribute to the literature on estimation of incident impacts and the identification of incident detection by using probe vehicle techniques, which generally

satisfy the applications for real-time travel time display. The integration of traffic and incident database enable us to look into critical factors for incident impacts and capture the dynamics of traffic evolution during the primary incident. Compared to static methods in previous studies, the dynamic filtering method has a better result in identifying secondary crashes. The proposed model can be applied to real transportation cases once we build a universal acceptance of a definition and corresponding set of parameters of secondary crashes.

The proposed methodology can be applied to any freeway segments in which speed information is available. Since vehicle probe technology is increasingly becoming more attractive for real-time system state estimation, and it is a common practice for data-providers to report data on TMC codes, we hope more accurate sources of traffic data are available. Including more incidents in a larger network will improve the accuracy of the results. Accurate and understandable information provided by the tool may help emergency operators make better decisions and maximize the effectiveness of incident management.

Chapter 5: Prediction of Secondary Crash Occurrence

In this chapter we introduce models to predict the likelihood of secondary crashes, given the primary incident types and road conditions. The results of prediction models provide incident management agencies with useful information.

5.1 Methodology

In this dissertation, a principled Bayesian learning approach to neural networks (Figure 5.1) is used to predict secondary crashes more accurately and robustly than current neural networks models. The main difference between Bayesian neural networks (BNN) and BPNN is the variable structure of the BNN and fixed structure of the BPNN [17]. The multilayer perceptron (MLP) type of neural networks is used.

Let $(x_1, y_1), , (x_n, y_n)$ be a set of incident data. Link function $f_B(x_i, \theta)$ can be obtained by parameters α_p the parameter for the weights between the input layer, the bias, and the output layer with normal prior distribution; β_j the parameter for the weights between hidden layer and output layer; γ_{jh}, the parameter for the weights between the input layer, the bias, and the hidden layer; $\mathbf{P}$, the input dimension; and $\mathbf{M}$, the maximum number of hidden neurons specified by the user. x_{ik} is the

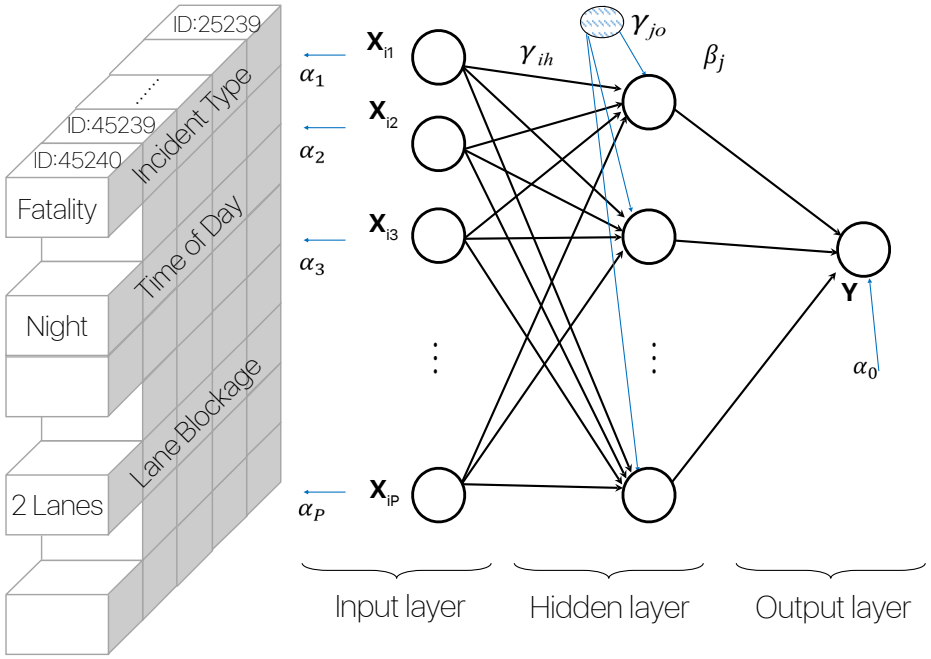


Figure 5.1: Structure of Bayesian neural network

k^{th} element of the i^{th} input. The link function is as follows:

$$f_B(x_i, \theta) = \alpha_o + \sum_{i=1}^{p} \alpha_k \cdot x_{ik} + \sum_{i=1}^{m} (\beta_j \cdot tanh \left(\gamma_{jo} + \sum_{i=1}^{p} \gamma_{jk} \cdot x_{ik} \right) \tag{5.1}$$

Probabilistic learning models can be defined as a conditional distribution $P(y|x)$ for an output y, given the input vector x, and a standard deviation σ_i. They are as follows:

$$P(y|x) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} exp \left(\frac{-(y_i - \hat{y}_i)^2}{2\sigma_i^2} \right) \tag{5.2}$$

The objective in the Bayesian approach is to find the predictive distribution for the target values in a new test case (x_{n+1}, y_{n+1}), given the input for that case and the targets and the inputs for the training cases. This distribution is obtained by integrating the predictions of the model with respect to the posterior distribution

of the network parameters and is shown below.

$$\begin{aligned} &P(y_{n+1}|x_{n+1},(x_1,y_1),...,(x_n,y_n)) \\ &= \int P(y_{n+1}|x_{n+1},\theta)P\left(\theta|(x_1,y_1),...,(x_n,y_n)\right)d\theta \end{aligned} \tag{5.3}$$

$P(\theta|(x_1,y_1),...,(x_n,y_n))$ is the posterior distribution of θ given observed incident data (x_n,y_n). The posterior distribution for these parameters is proportional to the product of the prior distribution and the likelihood function, and it varies during training in response to how well a particular set of weights model the data. The predicted clearance time value $(\hat{y}_i)$is given by:

$$\hat{y}_i = \int f_B(x_i,\theta)\cdot P(\theta|(x_1,y_1),...,(x_n,y_n))d\theta \tag{5.4}$$

The posterior distributions θ in case of multilayer perceptrons are complex, and above integrals are difficult to evaluate. The Markov chain Monte Carlo (MCMC) methods have been used to simulate the distribution of states of a system with combinatorial inference problems. However, the Hybrid Monte Carlo (HMC) techniques (called Hamiltonian Monte Carlo), which integrate molecular dynamics approaches and MCMC, perform better than traditional MCMC algorithms in high-dimensional, continuous, correlated spaces [132, Chapter 5]. The HMC method is used in this dissertation to approximate the integral by sampling the posterior distribution of the models. Hamiltons equations, which come from classical mechanics and assume that one can calculate the instantaneous position and momentum of a particle, are applied for the HCM. The Hamiltonian function operates on a *d*-dimensional position vector and a *d*-dimensional momentum vector so that the full state space has

$2d$-dimensions,

$$H(q,p) = U(q) + K(p) \tag{5.5}$$

where $U(q)$ is called the potential energy. Potential energy is defined to be negative of the log probability density of the distribution for q that we wish to sample, plus any constant that is convenient. $K(p)$, the kinetic energy, is defined as

$$K(p) = \frac{p^T M^{-1} p}{2} \tag{5.6}$$

Here, M is a symmetric, positive-definite "mass matrix" a typically diagonal scalar multiple of the identity matrix. This is an elaborate Metropolis Hastings Monte Carlo method that makes efficient use of gradient information to reduce random walk behavior. The Metropolis Hastings defines the Markov chain where the new sample $W(n+1)$ is generated from the old sample $W(n)$ by first generating a candidate state from a proposed distribution and then deciding whether or not to accept the candidate state. The HMC combines the Metropolis Hastings algorithm with sampling techniques based on a dynamic simulation, allowing us to incorporate gradient information from the distribution of interest. The gradient indicates the direction one should go to find states with high probability, and it can be calculated relatively easily for neural networks using error back-propagations. Details on the algorithm can be found in [132, Chapter 5].

5.2 Empirical Analysis: Key Factors

This section justifies key factors that can be used for incident duration prediction based on an exploratory analysis by [17]. Since only 10.5%, of incident

data includes pavement condition as a proxy variable for weather, this dissertation gathers the *actual* weather information for incident duration. The Clarus Initiative System provides weather information collected from a large network of stationary roadside weather detectors. In addition, archived traffic data before and after the incident events is collected from Center for Advanced Transportation Technology Laboratory (CATT Lab). These three different sources are incorporated through matching latitudes and longitudes of each record. In total, data pertaining to 13,987 incidents from year 2010 to 2011 are collected; the average incident duration under different categories is computed at 24.39 minutes, and the relationship between potential contributing factors is investigated.

Both the complex interactions among factors and the high dispersion in the data make predictions challenging. We intended to find an effective way to identify variables that may have affected the operational duration. Previous approaches assumed response times an independent variable, while the incident duration was a dependent variable. However, it is apparent from the following preliminary analysis that each category of factors has a different contribution to the incident duration. In this study we investigate the role of factors for response and clearance duration.

1. *Lane blockage*. This variable represents the number of lanes closed. The categories considered are: no blockage, shoulder lane blockage, one lane blockage, two lanes blockage, three lanes blockage and more than three lanes of blockage. The response times to incidents involving lane blockage are quicker than these causing only shoulder-lane rubbernecking impact (Figure 5.2). The exact fac-

tors contributing to such performance discrepancies are to be identified, but the resource limitations or personnel constraints may naturally cause response units to prioritize incidents that potentially have a greater impact. An opposite pattern exists with respect to clearance times: multi-lane blocked incident categories take much more time, an average of 51 minutes, to be cleared.

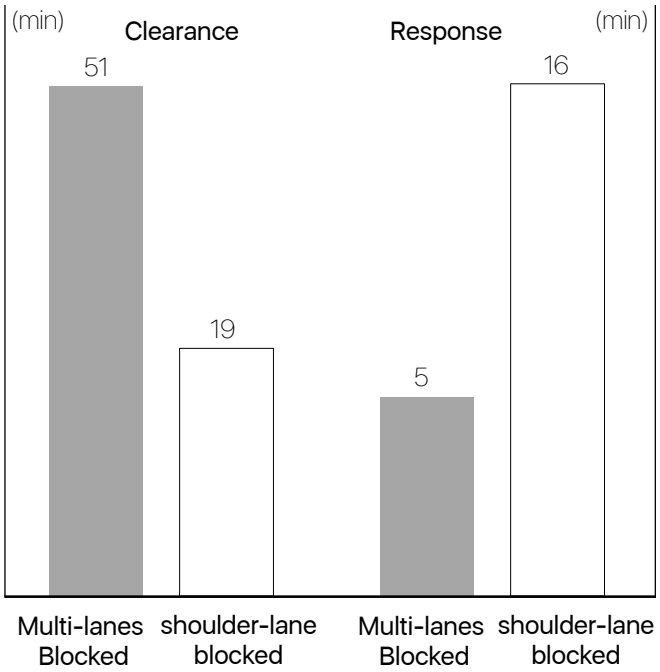


Figure 5.2: The average duration for lane blockage

2. *Incident type.* This variable includes collision with fatality (CF), collision with personal injury (CPI), collision with property damage (CPD), vehicle fire, disabled vehicles, and others (Figure 5.3). CF generally takes much longer to clear because of legal concerns, including the need for thorough incident investigation and documentation and the need for medical examiner investigation [2]. Also, the response times for incidents involving CF are quicker than those with only disabled vehicles. In the interest of safety, response drivers may not take

all allowable risks as they travel to less urgent incidents with lights and sirens, one of the most dangerous parts of their job [133]. Actually, 50 percent of all police, emergency medical services personnel and firefighter fatalities in 2002 occurred as a result of transportation incidents [134]. Another potential reason is that disabled vehicles have few effects on drivers′ safety or transportation and consequently gain little priority to be disposed [135].

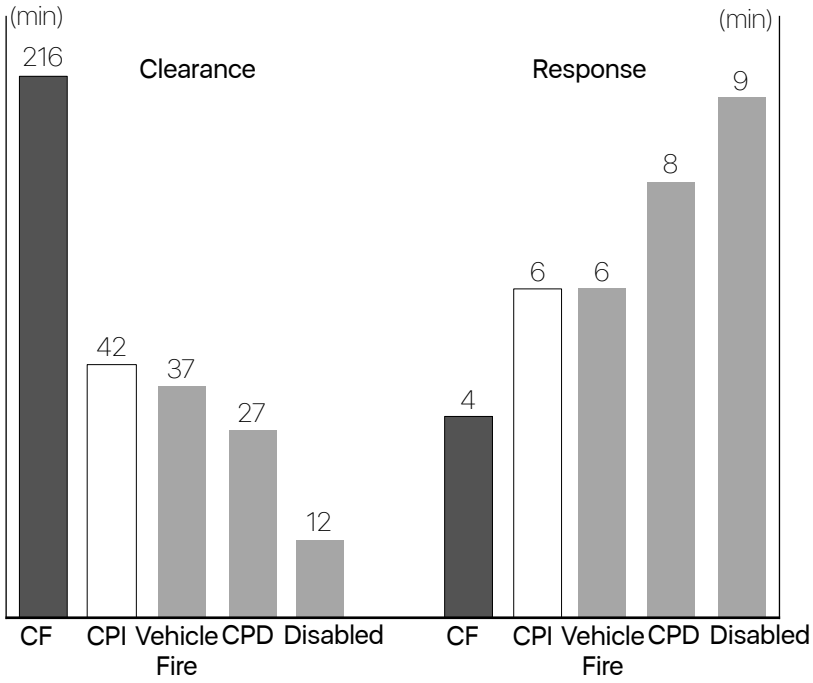


Figure 5.3: The average duration for incident type

3. *Incident location*. Figure 5.4 presents that incidents have their own patterns along spatial distribution on 11 major primary highway segments divided by 10 counties (abbreviations from Maryland State Archives, 1990). Note that even though incidents occur within the same area, the timeline for clearance and response are different. This is partly because response time includes travel time, which increases as the location of an incident site is farther from the

nearest emergency center.

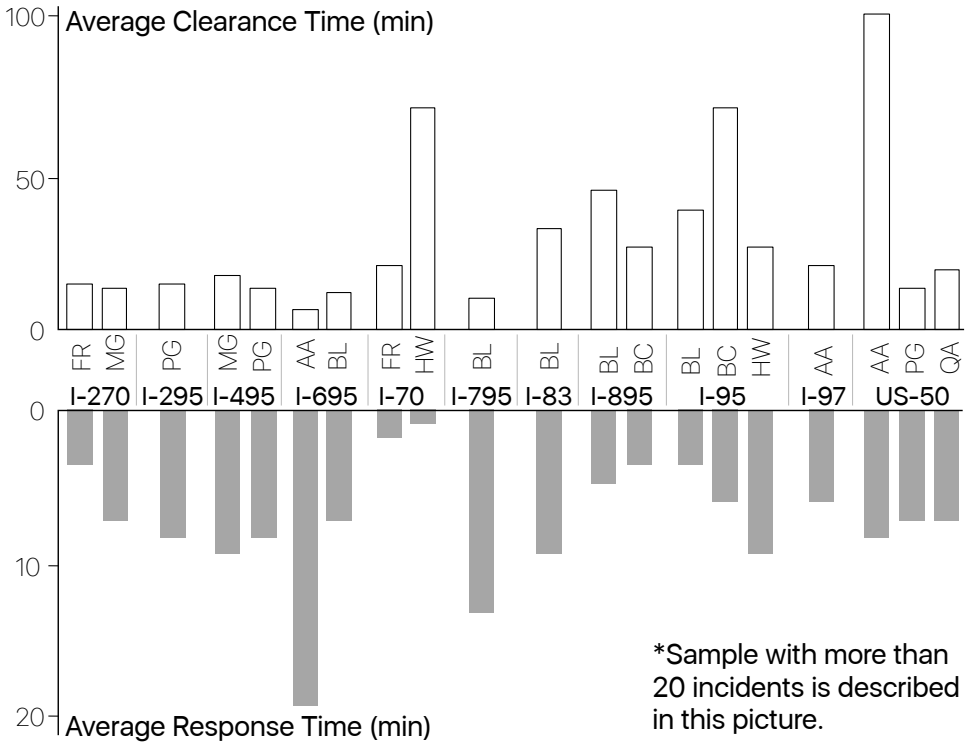


Figure 5.4: The spatial distribution of incidents

4. *Operation center*. Due to the available resources, the response efficiency of operations centers varies (Figure 5.5). Statewide operation centers (SOCs) distributed throughout almost all parts of a state generally outperform all other centers, and Traffic operation center (TOC)7 tends to take the longest response. On the contrary, SOCs' longest clearance time is partly because it is responsible for managing the most severe incidents.

5. *Traffic data*. As incidents will cause traffic congestion on an upstream detector [136], higher occupancy increase can be brought into relation with longer clearance duration. Five minutes of aggregated occupancy before and after incident′s time of occurrence are recorded at the first upstream loop detector,

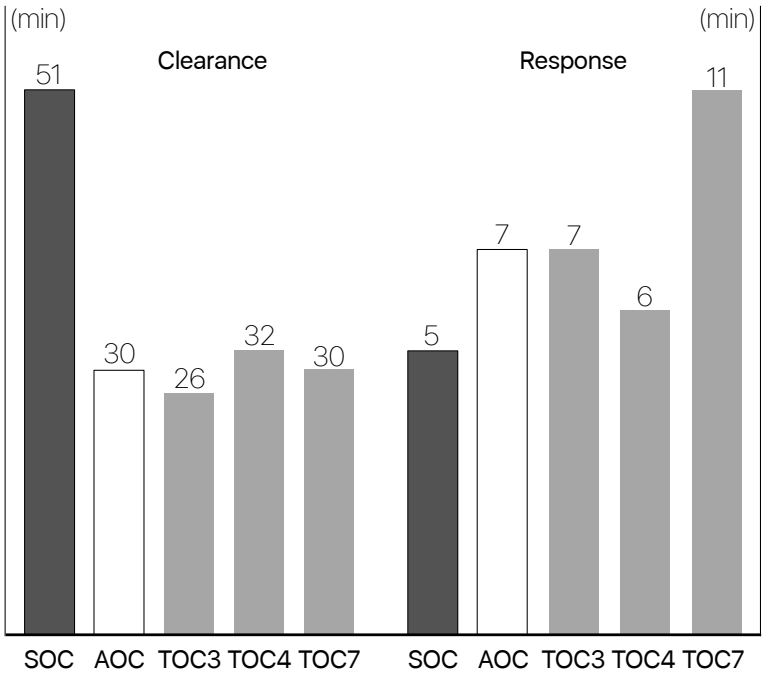


Figure 5.5: Performances by operational centers

and their increase rates are calculated. The variation of clearance duration increases as the rate of occupancy difference increases, where average value is 0.08, or 8%. Five minutes of aggregated probe vehicle data are used. Based on incident location, travel speed on the first upstream Traffic Message Channel (TMC) and second upstream TMC.

6. *Time of day.* Incidents occurring at night hours took an average of 26 minutes, seven minutes of clearance duration longer than incidents occurring in the day hours, which had an average of 19 minutes of clearance duration. This is due to fewer response teams being available, thus contributing to longer times to clear incidents.

7. *Vehicle involvement.* As the numbers of vehicles or heavy vehicles (truck-trailer, single unit truck, pickup/van) involved in the incidents increases, spe-

cial equipment for clearance operation are required, increasing duration.

8. *Detection source.* Clearance is more timely and effective if Coordinated Highways Action Response Team or Maryland Transportation Authority Police first identified the incident. By contrast, if incident alarm comes from the driver or passengers passing the site, it would take more time to clear.

5.3 Model Results

5.3.1 One-time Prediction of Clearance Time

In this section, 13,987 incidents from year 2010 to 2011 were used to compare the proposed BNN model with other advanced computing models: Back-propagation neural network BPNN, CART, and Support vector machine (SVM) [17].

Currently, the most frequently used performance metric for traffic incident management center is 30, 60, and 90 min clearance times based on severity [137]. A minor incident typically lasts no more than 30 min and does not require lane closures or extensive traffic control. Statewide incident clearance performance goals are 90 min for collision with fatalities. For the presentation of reliable system and potential application of the estimated model, classification rather than regression tool is preferred by traffic incident operators. The clearance duration is categorized into four groups: 1∼30 min, 30∼60 min, 30∼90 min, and over 90 min.

Both BPNN and BNN structures employed *Hyperbolic tangent transfer* function for the hidden units with 1 hidden layer, *Softmax transfer* function for the output

units, and were run 10 independent times to get their average performance. The optimal network models have 16 hidden units for BPNN, and 11 hidden units for BNN. CART defines impurity function as maximum homogeneity of child nodes. *Gini* splitting rule is used to maximize the change of impurity measure. Splitting is stopped when the number of observation of incidents is no more than 2 at a particular node. *Gaussian Radial Basis* kernels are used for SVM. We set the ϵ in loss function=0.005, tolerance of termination criterion=0.01, shrinking heuristics=1, and the parameter cost=1000. For a thorough discussion, readers are referred to CART [138] and the SVM [139].

Three key measures of effectiveness are applied to evaluate the models: (1) Mean Absolute Error (MAE) is used to measure the prediction accuracy, (2) Mean Squared Prediction Error (MSPE) is employed for determining the variance of the difference between predicted and observed results, (3) The percentage of underestimated cases is analyzed as a tool for an operational view point. The predicted traffic impact and following response strategy will also be underestimated if the incident duration is underestimated (if $\hat{y}_i - y_i < 0, U_i = 1$, otherwise $\hat{y}_i - y_i \geq 0, U_i = 0$). In Equations 5.7 through Equations 5.9, $\hat{y}_i$ and y_i are the predicted and observed values, respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5.7}$$

$$MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{5.8}$$

$$Underestimation = \frac{1}{n}\sum_{i=1}^{n} U_i \tag{5.9}$$

The 2010 incident data is randomly divided into 80% for training and 20% for cross validation. After training phase, the network is tested based on 20% of randomly selected 2011 incident data, not included in the training set. This is to test temporal transferability of the developed model, which have not been treated much in previous studies.

Table 5.1 summarizes the training and testing performance of the four models. CART does not perform very well compared to the other three models, but produces slightly lower percentage of underestimated prediction compared to SVM. Moreover, the output of CART presents that it may also suffer from over-fitting problem when we compare MAPE values for training and testing. This seems to support that the neural network models and SVM can better approximate nonlinear functions. BPNN and SVM perform approximately the same for testing MAPE values. BNN shows the best performance compared to all these three models. This research produced results which corroborate the discussion in [140].

Table 5.1: Model Performance

	Training (2010)				Testing (2011)			
	BNN	BPNN	CART	SVM	BNN	BPNN	CART	SVM
MAE	0.18	0.22	0.23	0.21	0.22	0.25	0.29	0.26
MAPE	0.25	0.29	0.35	0.3	0.29	0.38	0.56	0.38
Underestimation	10.20%	11.60%	11.95%	12.30%	12.06%	13.64%	14.16%	14.46%

From the result in Figure 5.6, it is apparent that BNN can consistently achieve lower average deviation in absolute value of the predicted class from the true class

compared to other procedures for incidents with duration of 1∼30 minutes, 30∼60 minutes, 60∼90 minutes, and larger than 90 minutes. BPNN, SVM, and CART models performed as well as BNN for the incident durations of 1∼30 minutes, but they show relatively lower prediction accuracy for incidents with durations over 30 minutes. It seems possible because BNN can superiorly approximate nonlinear function in spite of the fact that the dataset has a relatively smaller number of incidents over 30 minutes. BNN is the only model that can predict duration with MAE value within 0.6 for over 90 minutes.

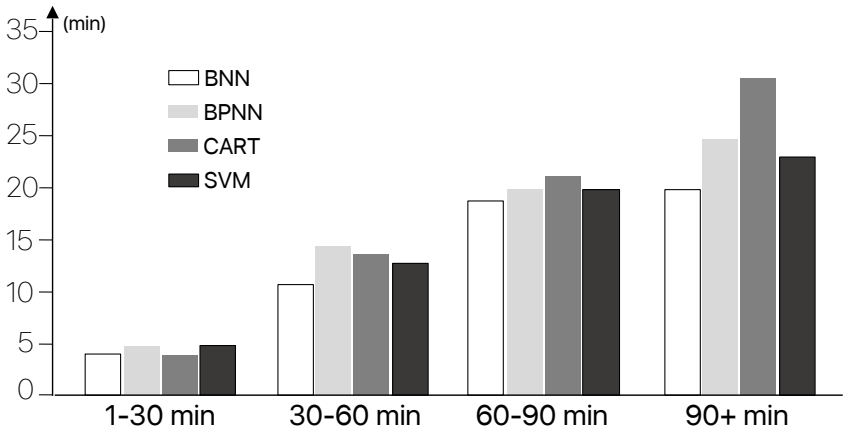


Figure 5.6: Mean absolute error (MAE) for different classifications

5.3.2 Sequential Prediction of Clearance Time

Previous studies considered response time to be the time between when the responding agency is notified and when the first response-unit arrived at the scene. However, as Figure 5.7 shows, if the first response unit (e.g. local police) is insufficient to clear the incident, clearance duration is extended until a second or greater response-unit (e.g. CHART units) arrives.

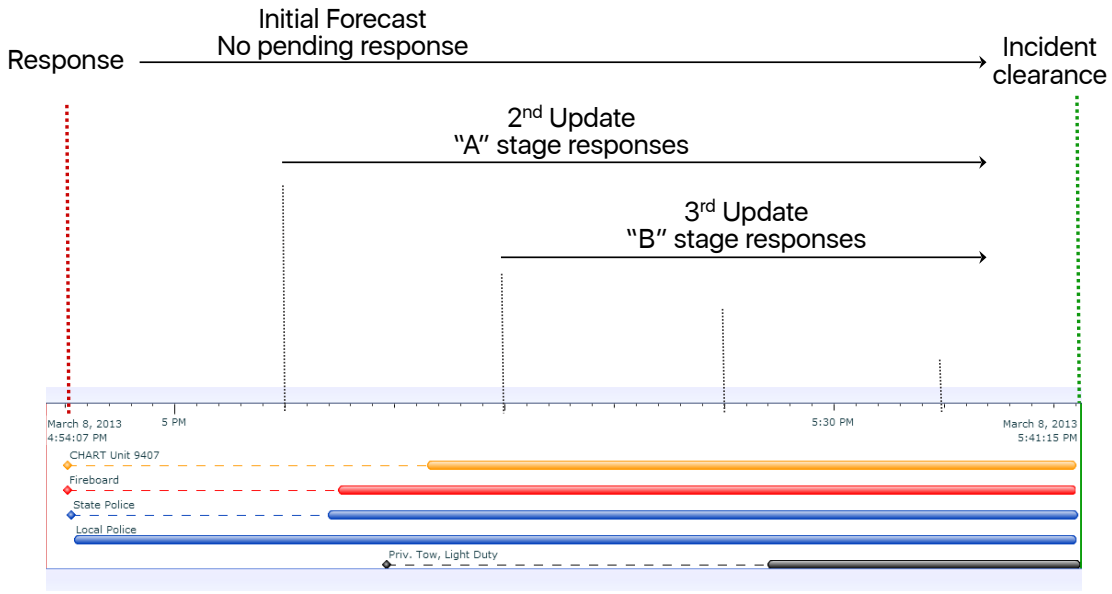


Figure 5.7: Sequential forecasting framework

In sequential prediction, each stage of prediction evaluates the response-units present, and notified. In Figure 5.7, at the initial prediction, CHART and Fireboard have been notified, but have not arrived. At the second update, CHART and Fireboard have yet not arrived, and the updated prediction is equal to the initial prediction. At the 3^{rd} update, CHART and Fireboard have arrived, but a third response unit has been notified and is in transit. The prediction of clearance time is updated accordingly.

The analysis of variance (ANOVA) test (Table 5.2) shows that there is a statistical difference in clearance time when combinations of response units are delayed ($F(8, 1146) = 67.458$, $p = 0.000$). A turkey post-hoc test reveals that the time to clear the incident is statistically significantly longer when

1. Arrival of the first CHART unit is delayed ($29.6 \pm 1.9min, p = .000$);

2. Fireboard or towing delayed ($41.4 \pm 1.9min, p = .000$);

3. First, second CHART delayed ($44.4 \pm 2.3min, p = .000$);

4. First CHART, Fireboard, and towing delayed ($74.8 \pm 2.0min, p = .000$);

5. Second CHART, Fireboard, and towing delayed ($50.5 \pm 2.0min, p = .000$);

6. First, second CHART, fireboard, towing delayed ($116.4 \pm 1.8min, p = .000$);

7. Clearance is delayed without CHART unit ($42.8 \pm 3.5min, p = .000$);

8. Compared to no responding delay ($15.9min$)

Table 5.2: One-Way ANOVA (Post Hoc Tests)

Delay caused by	Difference	Std.error	Sig.	Lower bd	Upper bd
1^{st} CHART unit	13.7	2.3	0	20.8	6.6
2^{nd} CHART unit	14.4	7	0.495	36.1	7.2
Fireboard, towing	25.5	2.9	0	34.2	16.9
1^{st}, 2^{nd} CHART unit	28.6	4.5	0	42.7	145
2^{nd} CHART, Fireboard, towing	58.9	10.6	0	91.9	26.1
1^{st} CHART, Fireboard, towing	34.6	3.1	0	44.4	24.9
1^{st}, 2^{nd} CHART, fireboard, towing	100.6	5.6	0	117.9	83.2
Without CHART unit	27	4	0	39.4	14.5

Key contributing factors for sequential prediction are as follows (the numbers in parentheses present the code):

- Number of lanes blocked (BL): 0, 1, 2, ... ;
- Time of day (TOD): peak (1), day non-peak (2), night non-peak (3);
- Traffic operation center (TOC): TOC 4 (1), AOC (2), SOC (3);
- Number of involved vehicles (NUM): 0, 1, 2, ... ;

- Truck Involvement (TK): no (1), one truck (2), more than one truck (3), truck overturn (4);
- Location (AREA): Exit 1-5 (1); Exit 6-10 (2); Exit 11-13 (3); Exit 14-18 (4); Exit 19-26 (5); Exit 27-31 (6); Exit 32-40 (7);
- Incident (TYPE): collision (1), injury (2), fatality (3);
- Response delay type (DY): 0 - 8, as described in Table 3 Above;
- Require firefighter (FIRE): yes (1), no (2);
- Severity (SEV): just off ramp closed (1), normal (2), guardrail damaged (3)

Two one-time prediction models differ in their use of response delay: basic prediction model and the proposed prediction model. The basic prediction model doesn′t use response delay for the prediction (MAE = 15.2 min). The proposed prediction model uses a delay-adjusted incident duration, the time that second and next response units spend in transit (MAE = 14.3 min). In previous research, error lower than 15 min is difficult to predict [74], but the proposed model makes the prediction more useful.

Sequential models update predictions periodically, e.g. every 10 min. At each reevaluation point, we also obtain observed value. The quality of predictions should improve as new information becomes available (e.g. response-unit arrival after travel time and damage to freeway infrastructure from traffic management center communications).

In Figure 5.8, we present the MAPE diagram through time. More outliers, which may cause more prediction errors, are observed in duration lower than 5 min. It origins from the lack of incident information at the beginning, but the model will have updated information as time goes to the end of incident clearance. The model has better performance after 10 min.

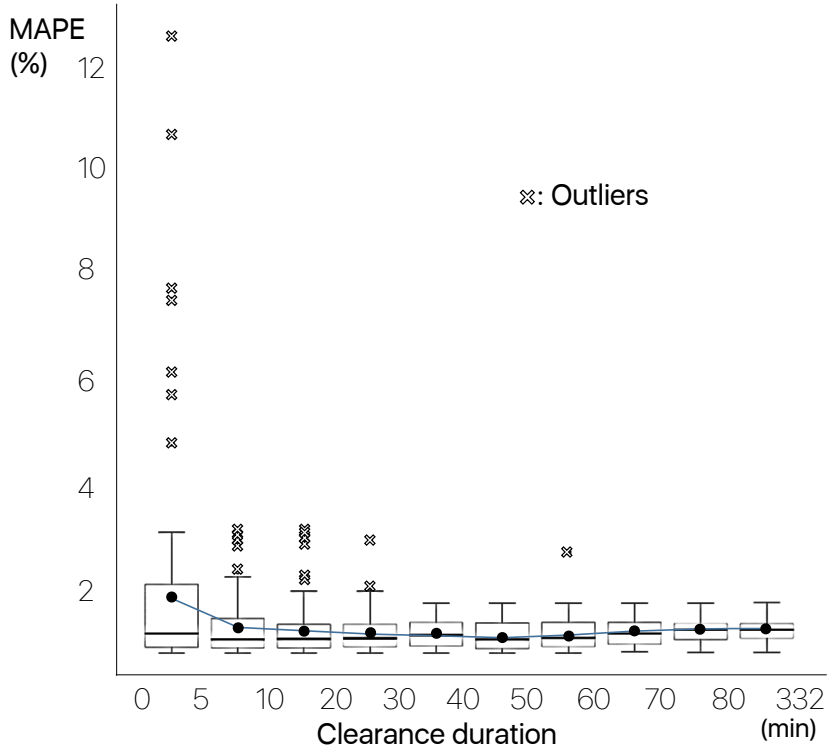


Figure 5.8: MAPE performance of models

We turn our interest to prediction result of secondary crash likelihoods in the next section.

5.3.3 Sequential Prediction of Secondary Incident Likelihood

Secondary incident occurrence is predicted using BNN, and compared with a Binary Logit model. One-time and sequential-prediction models are investigated using the nine variables, except for response delay type, in the clearance time prediction. In addition, the following four variables are added. Traffic condition variables were found to influence significantly the probability of having a secondary incident [141].

In contrast to previous study [141], this dissertation uses sequentially predicted clearance duration to predict the probability of having a secondary incident. The parenthesis presents the code.

- Traffic condition of the first upstream (FI): congested (1), not congested (2);
- Traffic condition of the second upstream (SE): congested (1), not congested (2);
- If the incident caused the traffic congestion (CTC): no (1), yes (2);
- Predicted clearance duration (CL): 0-5 (1), 5-10 (2), 10-20 (3), 20-30 (4), 30-40 (5), 40-50 (6), 50-60 (7), 60-70 (8), 70-80 (9), 80+ (10)

The MATLAB and modified NETLAB toolbox were used to implement BNN. The HMC return 100 samples to form the posterior probability. For each run, the average computing time for was 49 sec (Core 2 Duo 3.00 GHz CPU and 8 GB memory). For optimal setting of the models, numbers are chosen after obtaining results from many tests that involved trying potential combinations of parameters. BNN

employs *Hyperbolic tangent transfer* function for the hidden units with 1 hidden layer, 11 hidden units, and *Softmax transfer* function for the output units. It is well known that each running multiple neural networks may produce different results. Thus neural network models were run 10 independent times to get average performance. To evaluate temporal transferability of the models, data is randomly divided into two parts: 70% of the data set for training and 30% for testing set.

Table 5.3 presents that performance was improved when we considered traffic condition factor in our prediction. When transition of upstream traffic condition to congestion occurs within primary incident duration, it increases the chance of secondary incident occurrences. For more realistic situations, a predicted duration value is required, which is not significantly different from observed incident duration. Table 5.3 shows BNN outperforms the Logit model. This can be explained by the fact that Bayesian methods update network parameters using the Hybrid Monte Carlo algorithm, and improve the generalization ability of neural networks without compromising their nonlinear approximation ability.

Table 5.3: Comparison of One-Time Prediction Models with Different Conditions

Models (MAE)	With traffic condition		Without traffic condition
	Predicted duration	Observed duration	Predicted duration
BNN	15.60%	14.90%	25.60%
Logit	20.80%	-	-

In Figure 5.9, we sequentially tested the prediction performance using trained BNN. The proportion of false predicted primary incidents continuously decreases as new information (e.g. traffic condition upstream) updates, until the clearance

stage becomes more than 60 min. However, after 60-min clearance duration, both errors increase resulting in prediction performance as low as the clearance duration less than 5 min. The increase in error stems from relatively smaller sample size of secondary incidents after 60 min. The proportion of false predicted secondary incidents increases as traffic conditions become more congested. Since it takes time for a secondary incident to occur after primary incident [24], we will have better predictions after information is updated. Without updating required information, accuracy will not improve by time. Secondary incidents are more likely to occur when clearance duration is between 10 min and 20 min, or more than 75 min.

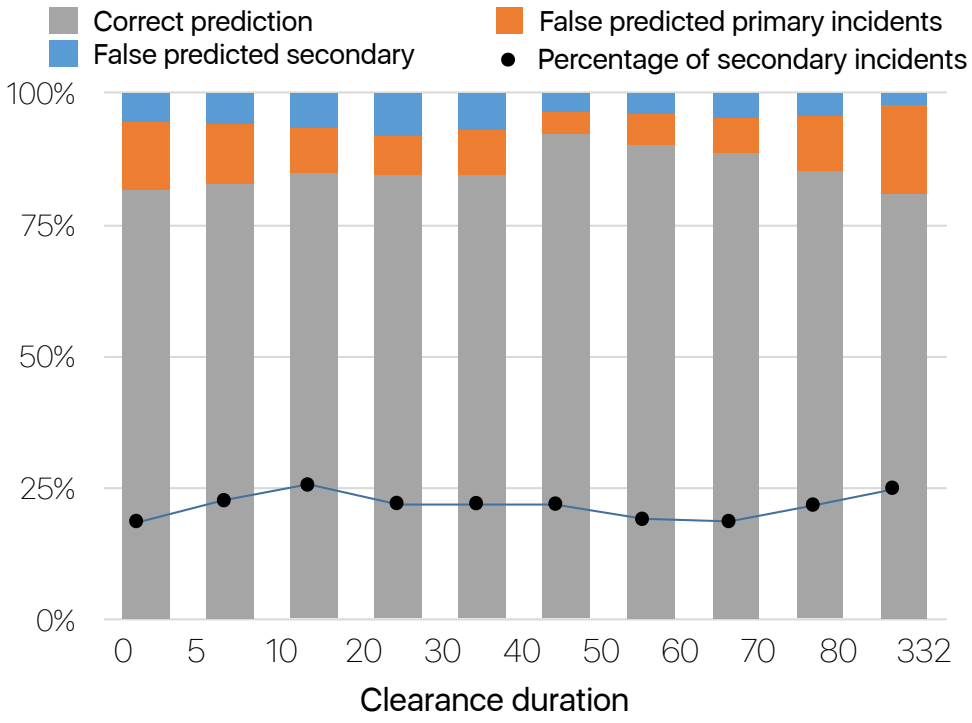


Figure 5.9: MAPE performance of models with different stages of clearance duration

5.4 Applications

A secondary crash is the product of factors relating to human, environmental, and vehicle. This dissertation aims to explore an importance of factors according to their weight value influencing secondary crash risk. Figure 5.10, known as *Haddon Matrix*, provides a framework for targeting different stages and influential factors of a crash. The phases are: pre-crash phase, the crash phase, and the post-crash phase. In the pre-crash phase, it is necessary to select all countermeasures that prevent secondary crashes from occurring. Interventions can reduce the chance of crash occurrences. In the crash phase, countermeasures prevent injury from occurring or reduce its severity. In the post-crash phase, all activities attempt to reduce the adverse outcome of the primary crash.

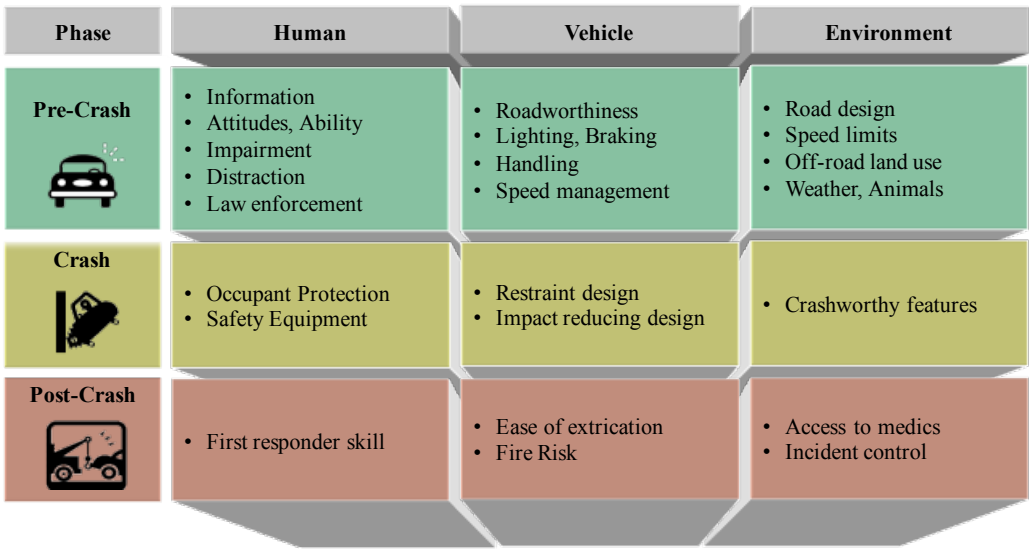


Figure 5.10: The contributing factors for crashes

This research plays an important role in the real-time incident management system. Collected traffic, incident, and weather information, typically from different

agencies, can be combined into one source (Figure 5.11). We should be careful of catastrophic forgetting, when new learning disrupts information previously learned by the network [142], to apply the proposed model to sequential forecast. Moreover, with a small sample size, caution must be applied, as the findings might not be transferable to other type of data. After data transition, this piece of information can be used by the traffic management center operators to take actions by prioritizing monitoring and freeway patrol service coverage. More quickly dispatched emergency responders and the right personnel and equipment dispatched to the scene can strengthen efficient response and manage effective incident scene clearance. In addition, this tool can enhance real-time information for the travelers through ways such as changeable or variable/dynamic message signs, highway advisory radio, and the internet.

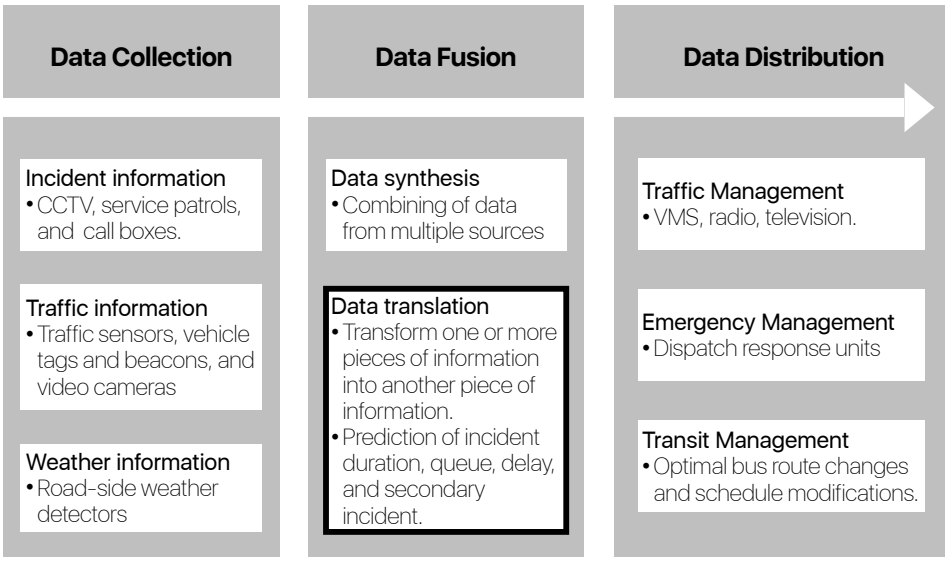


Figure 5.11: Advanced traveler information system by incident management

5.5 Conclusions

Modern data collection technologies enable us to look into critical factors of incident durations and establish an incident management plan. In this dissertation, artificial intelligence based on the Bayesian inference is used to design real-life pattern recognition problems of the likelihood of secondary crashes accurately and efficiently. BNN have shown promise to provide superior prediction performance compared to other tools. Accurate information provided by the tool may help emergency operators make better decisions, and it may maximize the effectiveness of incident management.

Chapter 6: Interpretation of Secondary Crash Occurrence

Even though a trained model has learned interesting and possibly universal approximation properties, these relationships are encoded incomprehensibly as weight vectors and cannot easily support the generation of scientific theories. In this research we introduce pedagogical rule extraction, stochastic gradient boosted tree, and connection weight approaches to interpret the prediction models.

6.1 Pedagogical Rule Extraction

The main difference between TREPAN and the Classification and Regression Tree (CART) is that CART builds a tree from the original data [138] while TREPAN branches the tree according to the predicted values by the neural network model. Therefore, the decision tree retains good prediction performance of the actual neural networks. Additional data from oracle (described below) provides higher predictive accuracy as the nodes lack sufficient data with the increase in tree size. The key aspects of the TREPAN are described in detail below.

Table 6.1: TREPAN Algorithm [89]

TREPAN	
Input:	Oracle(), training set S, feature set F, min sample, stopping criteria
1.	for each example x ∈ S
2.	class label for x := Oracle(x)
3.	initialize the root of the tree, R, as a leaf node
4.	construct a model M of the distribution of instances covered by node R
5.	query instances R := DrawSample({}, min sample $\|S\|$, M)
6.	use S and query instances R to determine class label for R
7.	initialize Queue with tuple (R, S, query instances R)
8.	**while** Queue not empty and global stopping criteria not satisfied
9.	remove node N , S_N , query instances N , constraints N) from Queue
10.	T := ConstructTest(F , $S_N \cup$ query instances N)
11.	make N an internal node with test T
12.	**for each** outcome, t, of test T
13.	make C, a new child node of N
14.	constraints C := constraints N ∪ {T = t}
15.	S_C := members of S_N with outcome t on test T
16.	construct a model M of the distribution of instances covered by node C
17.	query instances C := DrawSample (constraints C, min sample $\|S_C\|$, M
18.	use S_C and query instances C to determine class label for C
19.	**if** local stopping criteria not satisfied **then**
20.	put (C, S_C , query instances C , constraints C) in Queue
	Return: tree with root R

1. *Oracle and queries:* The primary goal of the TREPAN algorithm is to mimic the behavior of the trained neural networks. Instead of using the original training observations, TREPAN re-labels training data according to the classifications made by the network. The re-labeled data set is then used to initiate the tree-growing process. Training data become enriched with additional training instances, which are then also labeled by the neural network itself. The network is thus used as an oracle to answer class membership queries about

artificially generated data points. Each node split or leaf node class decision is based upon at least S_m data points. In other words, if a node has only m training available data points and $m < S_m$, then $S_m - m$ data points are additionally generated and labeled by the network.

2. *Drawing Query Instances:* Given a model and a set of constraints for each feature, TREPAN generates a value for the feature by sampling the distribution that is defined by the model conditioned on the constraints. The empirical distribution for a discrete-valued feature is represented by a parameter. Each possible value of the feature indicates the frequency of that value in the training set.

3. *Expansion:* Unlike most decision tree algorithms, TREPAN grows trees using a best-first expansion. Each node is assigned a priority defined to be the proportion of examples misclassified by the node. The algorithm maintains a queue of leaf nodes ordered by priority, and it successively expands the node, at the head of the queue into a fork with two children. Nodes with higher priorities are processed first because they offer the greatest chance of increasing the information gain: $G(n)$ in Equation 6.1. $R(n)$ is the number of original samples reaching the node divided by total number of original training samples, and $F(n)$ is the number of correctly classified samples in the node divided by the number of all samples in the node.

$$G(n) = R(n)(1 - F(n)) \tag{6.1}$$

4. *Splitting* TREPAN uses an *M-of-N* expression for splitting test. An *M-of-N* expression is a Boolean expression specified by an integer threshold, M, and a set of N Boolean literals. At least two of $\{C1, C2, C3\}$ are logically equivalent to $\{C1$ and $C2\}$ or $\{C1$ and $C3\}$ or $\{C2$ and $C3\}$. These *M-of-N* splits are constructed by the heuristic search procedure that uses a beam-search method with a beam width of two at each point; the best two splits are retained for further examination, and a best-first method for selecting the order in which nodes of the tree are expanded is used. To avoid over-fitting, a χ^2 test is used to determine whether the proposed change to the *M-of-N* test results in a significantly different partitioning of the instances than the partition induced by the test before the proposed change. Since each feature presents an opportunity to spuriously reject the null hypothesis, a *Bonferroni* correction is used to adjust the significant test downward for the individual tests.

5. *Stopping* TREPAN uses both global and local criteria to determine when to stop growing the tree. A global stopping criterion provides users control over the comprehensibility of the trees by limiting the size of the tree TREPAN returns. A local stopping criterion provides how many instances are needed to get a sufficiently tight confidence interval. If a proportion of instances have already reached the node of interest, then TREPAN makes it a leaf.

6. *Pruning* After the stopping criteria are met, TREPAN employs pruning to detect sub-trees that predict the same class at all of their leaves, and to collapse

each such sub-tree into a single leaf. TREPAN estimates the proportion of examples that fall into the most common class at a given node. Then, it calculates a confidence interval around this estimated proportion. The modifications made to a tree by this process do not change the predictive behavior of the tree at all. Then TREPAN returns the final tree.

6.2 Relative Importance of Factors

For the potential mathematical utility of neural networks, the connection weight method which generates interpretable parameters for each explanatory variable is used. This method involves partitioning the hidden-output connection weights of each hidden neuron into components associated with each input neuron [97]. Unlike previous sensibility analysis [90], multivariate and non-linear conditions are considered in this method because incident nature is rarely due to a simple cause or to a unique perturbation. The direction of the input-hidden-output of raw weights and the absolute values are considered in the present work to rank the factors as shown in Equation 6.6.

$$RI_i = \mid \sum_{h=1}^{M} \gamma_{ih}\beta_h \mid \tag{6.2}$$

where γ_{ih} denotes the value of the input hidden layer connection weight and β_h denotes the value of the hidden-output layer connection weight. The contribution of each input to the output is calculated as the product of the inputhidden weight and the hidden-output weight. The relative importance is the sum of products across all hidden weights. The interested readers are referred to [97] for further information.

6.3 Stochastic Gradient Boosted Decision Trees

Random forests, as an ensemble learning, generate a classification tree forest. Two well-known methods are bagging and boosting. In bagging, successive trees do not depend on earlier trees but are built independently using bootstrap sample of the data set. By contrast, in boosting methods, models are constructed sequentially and one tries to reduce the bias of the combined model. The motivation is to combine several weak models to produce a powerful ensemble. This dissertation uses stochastic gradient boosting decision trees (GBDT), which combine gradient boosting with bagging [143]. At each iteration, the base classifier is trained on a fraction subsample of the available training data.

Let $\{(x_1, y_1), ..., (x_i, y_i)\}_1^n$ be a set of incident data, consisting of output y_i (i.e. secondary crash occurrences) and input x_i (i.e. primary incident characteristics). Given historical training sample, our goal is find a function $F(x)$ that minimizes the expected value of loss function $\Psi(y, F(x))$. Gradient tree boosting considers weak leaners, $h_m(x)$ for the function

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{6.3}$$

We can build the additive model in a forward stage-wise fashion

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{6.4}$$

At each stage the decision tree $h_m(x)$ is chosen to minimize the loss function

given the current model $F_{m-1}(x)$ and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + argmin \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x)) \tag{6.5}$$

At each iteration m, a tree partitions the x-space into L-disjoint regions and predicts a separate constant value in each one.

$$\gamma_{im} = argmin \sum_{x_i \in R_{im}} \Psi(y_i, F_{m-1}(x_i) + \gamma) \tag{6.6}$$

Gradient Boosting attempts to solve this minimization problem numerically via steepest descent. The steepest descent direction is the negative gradient of the loss function evaluated at the current model F_{m-1} which can be calculated for any differentiable loss function. A shrinkage parameter υ is used to control the learning rate of the procedure. The stochastic gradient boosting incorporates randomness as an integral part of the procedure. A subsample of the training data is drawn at random from the full training data set. This randomly selected subsample is then used, instead of the full sample, to fit the base learner and compute the model update for the current iteration.

GBDT contains interpretable additive predictors. The partial effect of predictor is used to estimate the importance of each variable. To measure the importance of each variable after training, the values of the feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences. GBDT are built in Python 3.3 (scikit-learn toolkit).

6.4 Extracted Decision Trees

6.4.1 Settings

TREPAN accurately represents the network from which the rules are extracted, becoming a useful tool for eliciting comprehensible representation of neural networks. The main difference between TREPAN and the CART is that CART builds a tree from the original data while TREPAN branches the tree according to the predicted values by the neural network model. Therefore, our decision tree retains good prediction performance of the actual neural networks.

Instead of using the original training observations, TREPAN re-labels training data according to the classifications made by the network. The re-labeled data set is then used to initiate the tree-growing process. Training data become enriched with additional training instances, which are then also labeled by the neural network itself. The network is thus used as an oracle to answer class membership queries about artificially generated data points. Additional data from oracle provides higher predictive accuracy as the nodes lack sufficient data with the increase in tree size.

The process of expanding a node in TREPAN uses a best-first expansion so that as it adds each node it tries to maximize the gain in fidelity of the tree to the network that it is trying to model; a splitting test is selected for the node; and a child is created for each outcome of the test. Each child is either made a leaf of the tree or put into the queue for future expansion. Readers are referred to [89] for a more detailed description of the algorithm.

6.4.2 Results

The parameters are set as follows: at least 200 instances (training examples plus queries) are considered before selecting each split; significance level for comparing m-of-n tests are set to 0.05; maximum tree size is set to 35 internal nodes, which is the size of a complete binary tree of depth six. The extracted tree showed high fidelity (90.9%) to the network from which they were generated, resulting in 25 total nodes and 13 leaves. Lets assume that we have three main contributors for secondary incidents: location, time of day, and type. Then, we can build a decision tree using If-Then-Else, commonly used in CART (Figure 6.1).

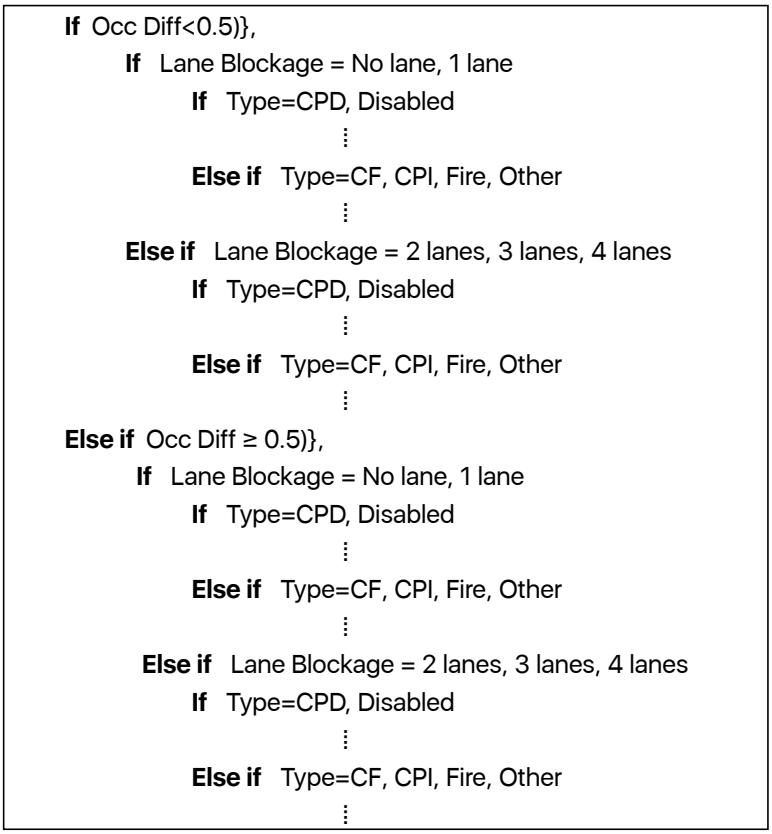


Figure 6.1: Extracted if-then-else rules for second split from decision tree

The label "True" indicates that the relevant entailment holds; the label "False" indicates that relevant entailment fails to hold. Incidents having different natures and characteristics are associated with different contributing factors. Contributing factors vary, occurring in different combinations per each incident. Extracted decision trees are simpler than complex CART for expressing rules. Figure 6.2 illustrates "If-Then-Else" statements can be transformed to "M-of-N" rule corresponding to the second node in the left.

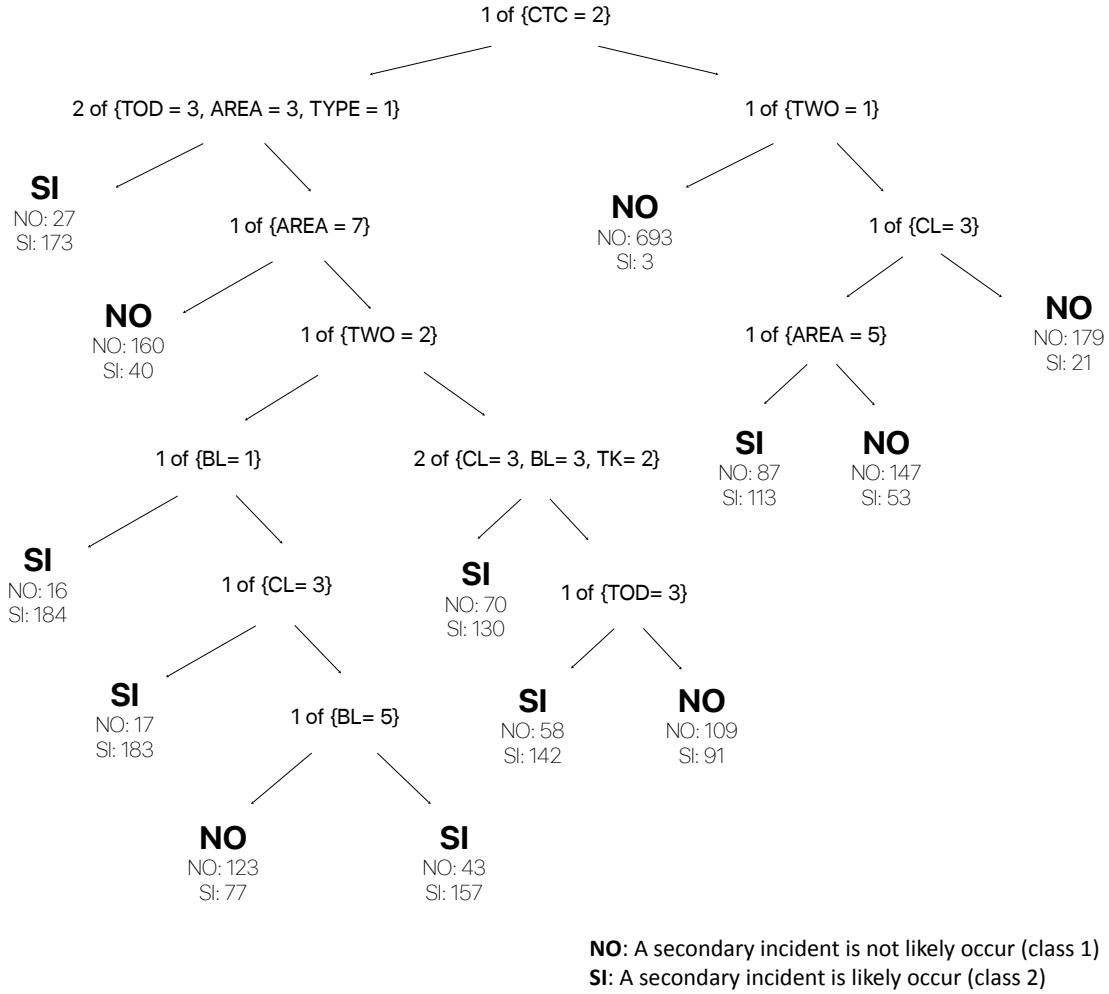


Figure 6.2: Extracted decision tree from prediction

Extracted decision trees are presented in the M-of-N rule, which has three

Boolean features, location, time of day, and type. Two of {(Location (area 3) = Exit 11, or 12, or 13), (Time of Day=peak hour), and (Type (1) = Collision with property damage)} is logically equivalent to {(Location = Exit 11, or 12, or 13) and (Time of Day=peak hour)} or {(Location = Exit 11, or 12, or 13) and (Type=Collision with property damage)} or {(Time of Day=peak hour) and (Type=Collision with property damage)}. If this condition is satisfied, we reach the leaf node which is classified to class 2 (secondary incident). The occurrence of secondary incidents (SI) is predicted to be 70.5% (173 among a total of 200 incidents).

In addition to this simple structure, Figure 6.3 provides a full version of the decision tree. Each node is assigned a priority, defined to be the proportion of examples misclassified by the node [89]. To decide how to partition the part of the instance space by the internal node, the M-of-N search uses information gain as its heuristic evaluation function. The result is the greatest information gain for each node (e.g., for example, the second node has information gain: $6.407149e^{-2}$ and priority: 0.046611). The numbers assigned to #class 2 (141/111) represent real/false examples reaching that node.

It is clear that "If-Then-Else" statements have more decision points; as a result, the M-of-N expressions better facilitate comprehensibility of the tree. In this way, TREPAN reduces the tree depth compared to the "If-Then Rules" statements used in the CART. Thus, TREPAN rules are straightforward to code in any incident management software. Traffic operators can easily understand TREPAN outputs by following the branches related to the conditions of variables. Moreover, this tool can also generate predictions when only partial information is available, since each node

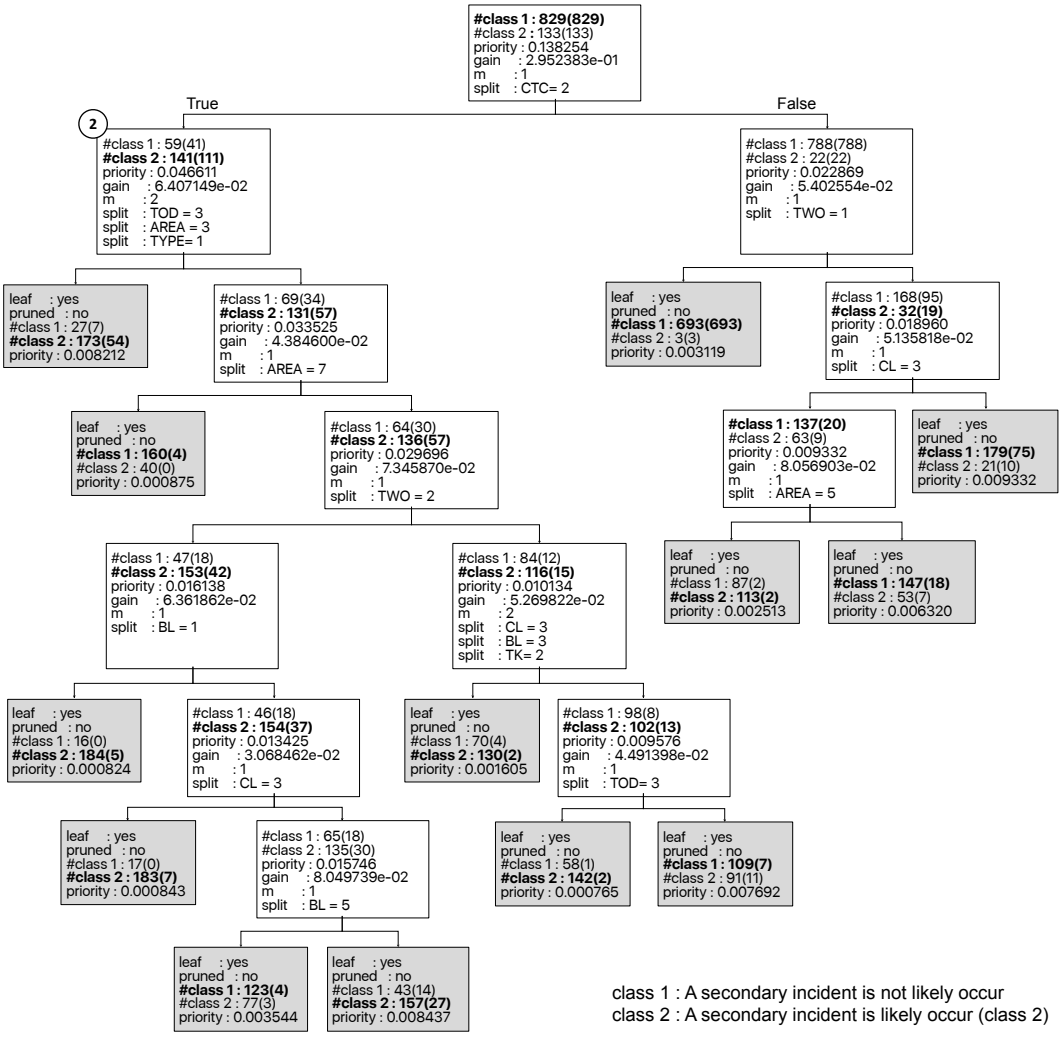


Figure 6.3: Full decision tree from prediction of secondary incident occurrences

can generate the maximum likelihood estimation of how long the incident may last. This information may contribute to the accurate selection of appropriate emergency response units.

The decision rules are cast in a form that appear to be particularly suitable for the representation of an incident that requires quick and concise action. Since each incident is different, the sequence of individual responder actions depends upon a variety of factors, such as who arrives first on scene, the severity of the incident,

and the surrounding traffic conditions, among others.

Table 6.2 presents that the increase in error stems from relatively smaller sample size of secondary incidents after 60 min. BNN outperform GDBT except for the first two clearance stages (i.e. 10 min), after the primary incident occurrence. BNN tend to underestimate when upstream of incident scene has no congestion caused by negative impact of the primary incident.

Table 6.2: Performance of Models for Each Update (GDBT and BNN)

Clearance	GDBT			BNN		
	True	False		True	False	
		primary	secondary		primary	secondary
0-5 min	82.10%	13.10%	4.80%	81.60%	13.00%	5.40%
5-10 min	83.20%	11.20%	5.60%	82.90%	11.10%	6.00%
10-20 min	84.70%	8.50%	6.80%	84.90%	8.50%	6.60%
20-30 min	84.10%	7.20%	8.70%	84.50%	7.30%	8.20%
30-40 min	84.20%	8.20%	7.60%	84.50%	8.30%	7.30%
40-50 min	89.80%	4.20%	6.00%	92.10%	4.30%	3.60%
50-60 min	90.10%	5.80%	4.10%	90.30%	5.80%	3.90%
60-70 min	88.40%	6.70%	4.90%	88.50%	6.70%	4.80%
70-80 min	84.90%	10.40%	4.70%	85.10%	10.40%	4.50%
80 min +	80.50%	16.60%	2.90%	81.00%	16.70%	2.40%

The decision tree is a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by logic. By contrast, GDBT treats the decision tree model as a black box. It is hard to interpret and it does not take advantage of the tree structure itself. Use of small shrinkage parameter GDBT could lead to a huge tree model, which is very undesirable as it leads to high computational cost of applications.

6.5 Relative Importance

Table 6.3 describes the connection weight matrices of 11×11 (input hidden) and 11×1 (hidden output) extracted from trained Bayesian neural networks with best performance. The relative contribution to clearance time depends on the magnitude and direction of the connection weights. Input variables with larger connection weights represent greater intensities of signal transfer, and therefore are more important in the prediction of incident duration compared to variables with smaller weights. This result also shows that negative value of input variable "center" represents TOCs typically associated with shorter duration than SOCs, while other factors are positively associated with incident duration. For example, incident associated with higher occupancy increase with more number of involved vehicles and blocked lanes, and collision with fatalities or injuries occurring at night time result in longer duration. These findings further support the idea of preliminary analysis.

As shown in Figure 6.4, contributions of each input variable to the output are divided by the sum of contributions and expressed as a percentage to ease the interpretation of relative importance. The incident type, lane blockage, and occupancy are the strongest indicators of clearance duration compared to the other factors. It also corresponds to the discrepancy patterns from the preliminary analysis. It is important to use this method in accordance with an emergency operator's opinion regarding the ranking of importance of inputs and their mode of action on the output. The result of connection weight approach provides an insight into the critical factors that affect decision support in the context of emergency response manage-

Table 6.3: The Connection Weight Productions (Clearance Time) [17]

Hidden	1	2	3	4	5	6	7	8	9	10	11
Weather	0.32	-0.94	0.13	1.07	1.82	0.08	0.09	-0.4	-0.71	-0.05	-0.87
Type	-1.33	1.86	0.23	0.06	1.42	2.35	-0.01	0.86	-0.21	-1.07	-0.17
Occupancy	0.58	-0.25	-0.97	1.07	-0.41	0.02	0.26	0.04	0.28	0.55	0.59
Center	0.27	-0.36	0.71	0.75	0.16	0.82	-0.12	0.11	-1.05	1.01	-0.49
Road	0.69	-1.07	-0.36	-2.91	-1.39	0.11	0.44	-1.55	-2.42	0.47	-0.77
County	-0.42	-0.05	0.23	1.21	0.58	0.33	-0.58	-0.02	-0.44	-0.43	-0.99
NumVeh	1.03	0.23	-1.35	-1.63	0.3	-0.86	-1.61	0.45	0.23	1.66	0.37
Time	0.81	0.6	0.09	0.34	0.32	-0.86	-0.35	-0.25	-0.39	0.75	-0.85
HeavyVeh	0.34	-0.35	-0.38	0.09	0.55	0.16	-0.07	-0.06	0.49	0.51	0.04
Blockage	0.39	0.15	-0.48	2.5	0.2	-0.21	1.01	0.24	-0.85	-0.65	-1.13
Detection	0.34	0.33	-0.48	0.79	0.98	-0.72	0.22	-0.79	0.71	1.41	-0.07
×											
Hidden	1	2	3	4	5	6	7	8	9	10	11
Output	1.2	0.72	-0.7	0.35	0.67	0.88	0.31	-0.29	-0.62	0.07	0.15
=											
Hidden	1	2	3	4	5	6	7	8	9	10	11	Sum
Weather	0.38	-0.18	0.68	0.37	-0.27	0.02	0.08	-0.01	-0.18	0.04	0.09	1.01
Type	-1.59	1.35	-0.16	0.02	0.95	2.08	0	-0.25	0.13	-0.07	-0.02	2.43
Occupancy	0.38	-0.68	-0.09	0.37	1.22	0.07	0.03	0.12	0.44	0	-0.13	1.72
Center	0.33	-0.26	-0.49	0.26	0.11	0.73	-0.04	-0.03	0.65	0.07	-0.07	1.25
Road	0.82	-0.77	0.25	-1.02	-0.93	0.1	0.14	0.45	1.51	0.03	-0.12	0.46
County	-0.5	-0.04	-0.16	0.42	0.39	0.29	-0.18	0.01	0.27	-0.03	-0.15	0.33
NumVeh	1.23	0.17	0.94	-0.57	0.2	-0.76	-0.5	-0.13	-0.14	0.11	0.06	0.61
Time	0.97	0.43	-0.07	0.12	0.21	-0.76	-0.11	0.07	0.24	0.05	-0.13	1.04
HeavyVeh	0.4	-0.25	0.26	0.03	0.37	0.14	-0.02	0.02	-0.31	0.04	0.01	0.69
Blockage	0.47	0.11	0.33	0.87	0.13	-0.18	0.32	-0.07	0.53	-0.05	-0.17	2.3
Detection	0.41	0.24	0.33	0.28	0.66	-0.64	0.07	0.23	-0.44	0.1	-0.01	1.22

ment, identifies and highlights potential areas for improvement, and allocates more resources for response to severe incident type such as collision with fatality rather than just disabled vehicles.

Pedagogical interpretation is one of the most powerful interpretation tools. A comprehensive summary of GDBT's dependence on the joint values of the input variables is presented in Figure 6.5.

Regardless of clearance stages, the main effects that explain the secondary incident occurrences are from the decision on whether the primary incident mainly

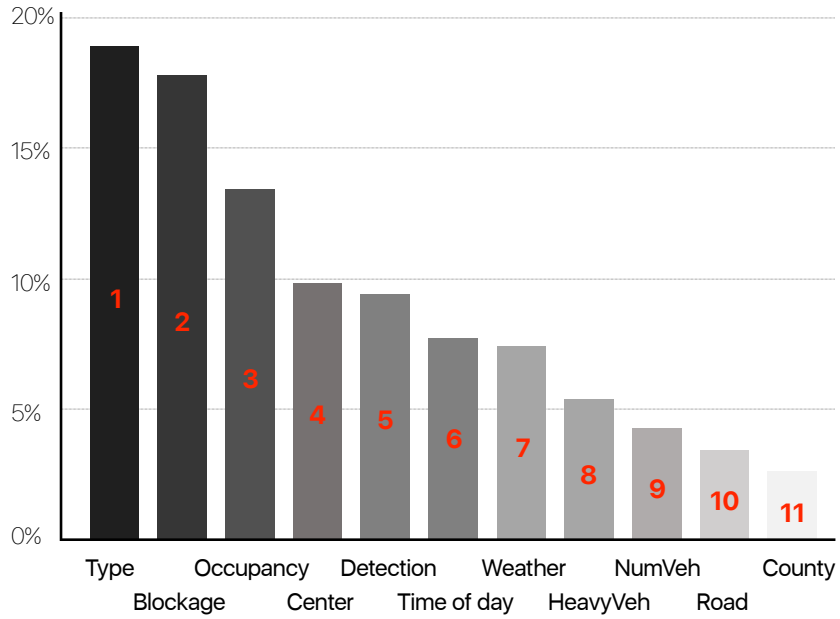


Figure 6.4: Relative importance for incidnet duration

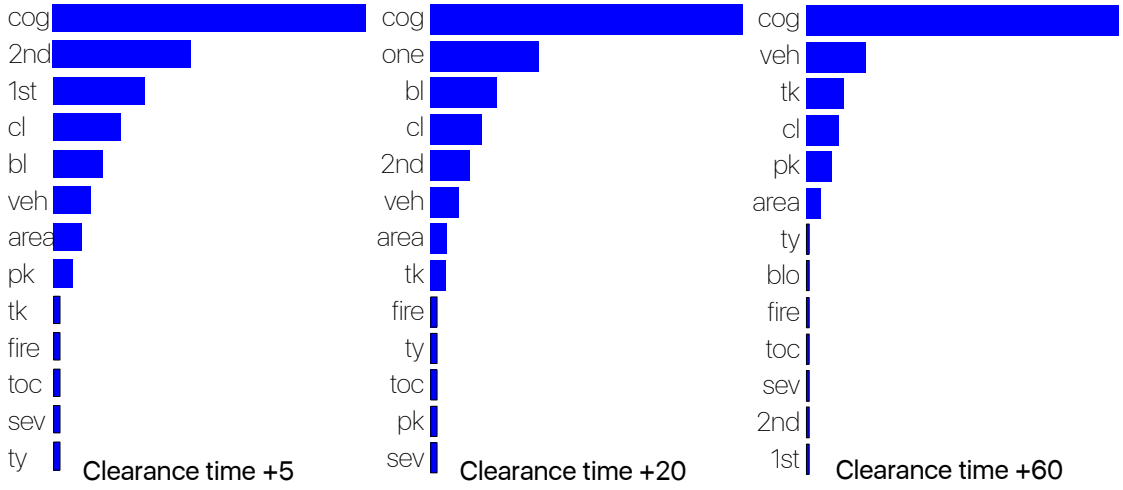


Figure 6.5: Relative importance for secondary incident likelihood

caused the congestion on the road. However, the relative contribution of this predictor variable, the main cause of congestion, becomes less significant within the group of shorter clearance stages (i.e. more than 5 min). Instead, the relative contribution of the predictor variable, the traffic condition of the first or second upstream from the incident location, becomes more significant.

6.6 Conclusions

Good performance of prediction models will be worthless without reasoning behind the learning system. It will be valuable to investigate a device simple to understand and interpret. The integration of the Bayesian neural network with an algorithm to extract knowledge from the trained networks takes an advantage of both worlds to an incident management coordinator attempting to make predictions of a detected incident and understand it. In contrast with shortcomings of traditional decision trees, TREPAN embeds not only higher predictive accuracy with data re-labeling from developed BNN and additional data using an oracle, but also provides improved comprehensibility with simpler M-of-N rule expression. Furthermore, using connection weights from Bayesian neural networks, relative importance is identified, which provides an insight into the critical factors that affect decision support in the context of emergency response management. It also highlights potential areas for improvement and allocates more resources. The extraction of decision trees from trained Bayesian neural networks is an important addition to the Advanced Traveler Information Systems (ATIS) toolkit of knowledge extraction technique.

Chapter 7: Stochastic Capacity Adjustment Considering Secondary Incidents

7.1 Deterministic SIDM

Grounded on a related study [22], a secondary incident delay model (SIDM) is formulated to estimate reduced discharge flow by considering both primary and secondary incidents. When durations of two interrelated incidents overlap, total delay is underestimated, or when there is a gap, it leads to an overestimation of total delay. Without consideration of time series of incident occurrences, total delay is calculated in the traditional way (two smaller triangles in Figure 2.2). A secondary incident occurs during the clearance or recovery stage of a primary incident. r_{ps} is introduced as the gap/overlap between the beginning of the secondary incident and the end of the clearance stage of the primary incident. A consolidated area $A'B'C'$ (Figure 7.1 (c)) is developed from two types of isolated individual areas (Figure 7.1 (a, b)).

For a primary incident, the time t_p of congestion clearance (including recovery time) is expressed as a function of queue formed during primary incident duration:

$$t_p = r_p \frac{s - s_1}{s - q} \tag{7.1}$$

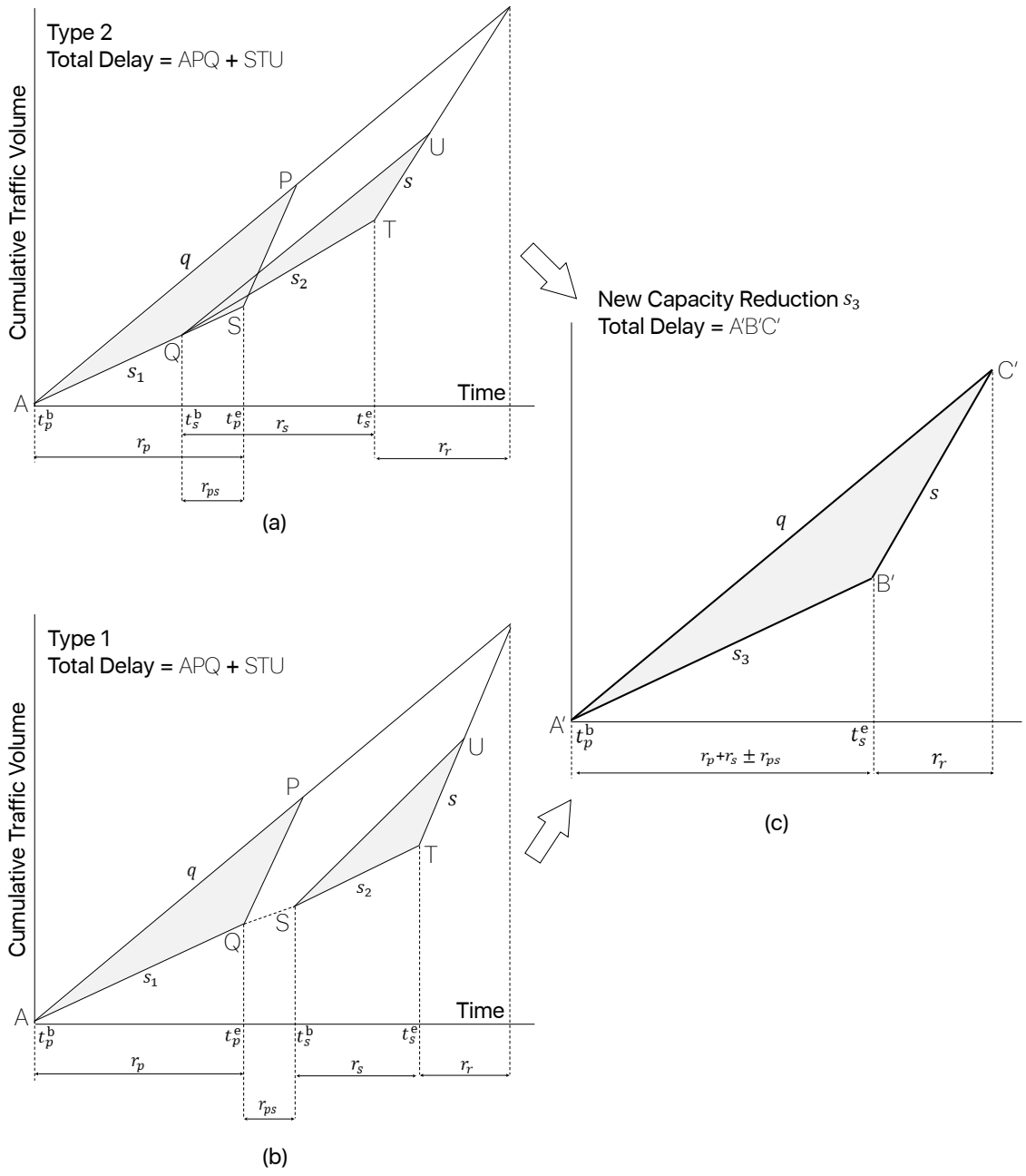


Figure 7.1: The proposed incident delay model considering secondary incidents that occurred in (a) the clearance stage of primary incidents (b) the recovery stage of primary incidents, and (c) new discharge flow s_3.

When a secondary incident occurs during the clearance stage of a primary incident (*Type 2*), recovery time is extended because the queue has not dissipated.

The congestion clearance t_p for the secondary incident is expressed as follows:

$$t_s(Type\ 2) = \frac{r_p(s - s_1) + r_s(s - s_2)}{s - q} \tag{7.2}$$

On the contrary, when a secondary incident occurs during the recovery stage of a primary incident (*Type 1*), the dissipated queue is deducted from previous Equation 7.2:

$$t_s(Type\ 1) = \frac{r_p(s - s_1) - r_{ps}(s - q) + r_s(s - s_2)}{s - q} \tag{7.3}$$

The total delay caused by the remaining queue remaining from a primary incident and the queue formed because of a secondary incident is shown in the gray area in the Figure 7.1(a)-(c). The queue upstream of the secondary incident will dissipate to free flow after the secondary incident is fully cleared. For a simple calculation of total delay, gray areas in the figure are transferred to the triangular area ABC.

We estimate the total delay caused by a primary incident and a secondary incident. Now discharge flow rate (s_3) is calculated as a function of s_1, s_2, s, q, r_p, r_s, r_{ps}. Assuming the constant arrival rate of vehicle q for all incidents, the total delay (*Type 2*) can be calculated as follows:

$$\begin{aligned} delay &= (r_p + r_s \pm r_{ps})^2 \frac{(s - s_3)(q - s_3)}{2(s - q)} \\ &= (r_p)^2 \frac{(s - s_1)(q - s_1)}{2(s - q)} + r_s(q - s_2)\frac{r_p(s - s_1) + r_s(s - s_2)}{2(s - q)} \end{aligned} \tag{7.4}$$

This dissertation defines the impact of a new discharge flow s_3 as a function of a primary incident discharge flow s_1 and a secondary incident discharge flow s_2.

A quadratic equation is derived from the Equation 7.1 as follows:

$$\begin{aligned}&(r_p + r_s - r_{ps})^2(s - s_3)(q - s_3)\\&-(r_p)^2(s - s_1)(q - s_1) - r_s(q - s_2)\left\{r_p(s - s_1) + r_s(s - s_2)\right\} = 0\end{aligned} \tag{7.5}$$

After dividing the quadratic equation by $(r_p + r_s - r_{ps})^2$ and the method of completing the square can be applied.

$$\begin{aligned}&s_3^2 - s_3(s + q) + sq\\&-\frac{(r_p)^2(s - s_1)(q - s_1)}{(r_p + r_s - r_{ps})^2} - \frac{(r_s)(q - s_2)\left\{r_p(s - s_1) + r_s(s - s_2)\right\}}{(r_p + r_s - r_{ps})^2} = 0\end{aligned} \tag{7.6}$$

Isolating s_3 gives solutions of the quadratic equation:

$$s_3 = -\frac{B}{2A} - \frac{\sqrt{B^2 - 4AC}}{2A}\left\{\begin{array}{l}A = 1\\B = s + q\\C = 4\Bigg\{sq - \dfrac{(r_p)^2(s - s_1)(q - s_1)}{(r_p + r_s - r_{ps})^2}\\\quad + \dfrac{(r_s)(q - s_2)\left\{r_p(s - s_1) - r_{ps}(s - p) + r_s(s - s_2)\right\}}{(r_p + r_s - r_{ps})^2}\Bigg\}\end{array}\right\} \tag{7.7}$$

Cases are excluded when a secondary incident occurred and cleared before the clearance of the primary incident.

7.2 Stochastic SIDM

The secondary incident delay model in Section 7.1 assumes that all the parameters are known with certainty. For example, traffic demand (q), incident duration (r), capacity (s) and, reduced capacity (s_1, s_2, s_3) are assumed to be known. However, in the real-time operations, this information could be obtained through incident responders or data collection in real time, which result in different estimations or

realizations of these parameters. To address the case when r, s_1, s_2, and s_3 are not known with certainty, a stochastic extension of SIDM is proposed: Stochastic Secondary Incident Delay Model (SSIDM). Variables that have relatively smaller variability and easier prediction (q and s) are assumed to be constant.

Incident duration and capacity reduction are assumed to be random variables with their probability density functions (see stochastic form of delay model [144]). SSIDM is expected to estimate greater total delay because it takes the uncertainty of incident duration and reduced capacity into consideration in estimating the delay. First, relax primary incident duration $r_p \sim f(r_p)$ and integrate the deterministic version Equation 7.3 with probability density function (PDF), assuming that other variables are constant. Let $\bar{r}_p$ be mean and σ_{r_p} be standard deviation of primary incident duration. The expected total incident-induced delay (TD) is:

$$\begin{aligned} &E[TD(t, r_p \mid r_s, s_3)] \\ &= \int_0^q f(r_p) \frac{(s - s_3)(q - s_3)(r_p + r_s \pm r_{ps})^2}{2(s - q)} dr_p \\ &= \frac{(s - s_3)(q - s_3)\left\{\bar{r}_p^2 + \sigma_{r_p}^2 + 2\bar{r}_p(r_s \pm r_{ps}) + (r_s \pm r_{ps})^2\right\}}{2(s - q)} \end{aligned} \tag{7.8}$$

Second, relax secondary incident duration $r_s \sim f(r_s)$ and integrate the deterministic version Equation 7.5 with PDF. The expected TD is:

$$\begin{aligned} &E[TD(t, r_p, r_s \mid s_3)] \\ &= \left\{\bar{r}_p^2 + \sigma_{r_p}^2 + \bar{r}_s^2 + \sigma_{r_s}^2 \pm 2r_{ps}(\bar{r}_p + \bar{r}_s) + 2\bar{r}_p\bar{r}_s + r_{ps}^2\right\} \times \frac{(s - s_3)(q - s_3)}{2(s - q)} \end{aligned} \tag{7.9}$$

Third, relax reduced discharge flow $s_3 \sim f(s_3)$ and integrate the deterministic version Equation 7.6 with probability density function, assuming that other variables

are constant. The expected TD is:

$$E[TD(t, r_p, r_s, s_3)]$$
$$= \left\{ \bar{r}_p^2 + \sigma_{r_p}^2 + \bar{r}_s^2 + \sigma_{r_s}^2 \pm 2r_{ps}(\bar{r}_p + \bar{r}_s) + 2\bar{r}_p\bar{r}_s + r_{ps}^2 \right\} \times \frac{\left\{ \bar{s}_3^2 + \sigma_{s_3}^2 + \bar{s}_3(s-q) + sq \right\}}{2(s-q)} \tag{7.10}$$

Finally, the coefficient variations of reduced discharge flow is $x = \frac{\sigma_{s_3}}{\bar{s}_3}$ and $\bar{s}_3$ is estimated. Part of Equation 7.7 can be transformed into Equation 7.11

$$\bar{s}_3 = -\frac{B'}{2A'} - \frac{\sqrt{B'^2 - 4A'C'}}{2A'}$$
$$\left\{ \begin{array}{l} A' = 1 + x^2 \\ \\ B' = s + q \\ \\ C' = 4\left\{ sq - \dfrac{(\bar{r}_p^2 + \bar{\sigma}_{r_p}^2)(s - \bar{s}_1)(q - \bar{s}_1)}{\bar{r}_p^2 + \sigma_{r_p}^2 + \bar{r}_s^2 + \sigma_{r_s}^2 + \bar{r}_p\bar{r}_s - r_{ps}(\bar{r}_p + \bar{r}_s) + r_{ps}^2} \right. \\ \left. + \dfrac{(\bar{r}_s)(q - \bar{s}_2)\left\{ \bar{r}_p(s - \bar{s}_1) - r_{ps}(s - p) + \bar{r}_s(s - \bar{s}_2) \right\}}{\bar{r}_p^2 + \sigma_{r_p}^2 + \bar{r}_s^2 + \sigma_{r_s}^2 + \bar{r}_p\bar{r}_s - r_{ps}(\bar{r}_p + \bar{r}_s) + r_{ps}^2} \right\} \end{array} \right\} \tag{7.11}$$

To make the sum of different distributions possible, it is assumed that new variable $\bar{s}_3$ does not follow any specific distribution. Interested readers may find methodologies for approximation of sum of differences from [145].

7.3 Location-Dependent Incident Duration

An explicit function is introduced based on response efficiency and incident type to get incident duration parameters unique at each location. Response efficiency is an important explanatory variable that can be described as coordination of the first and second responses. If the first response unit (e.g., local police) is insufficient to clear the incident, clearance duration is extended until a second or greater response-unit (e.g., coordinated highways action response team, CHART) arrives.

In general, arrivals of the response units (i.e., $\overline{x}$ is defined as average response time to incident locations, i.e., exits) depend on where the response units are previously assigned. In addition, arrivals of the secondary response units depend more on the possibility of the server being busy for responding to prior incidents.

Clearance times of all incident locations are averaged to $\overline{y}$, as least amount of duration without influence of incident severity. The extra time between when the second response units have been notified, but have not arrived, is defined as z_i for each incident location i. The response delay is multiplied by constant coefficient variable ω to represent magnitude of secondary response delay for each incident location i (see [16] for calculating different response times). The incident duration is significantly influenced by the clearance operation at each location i and can be calculated as follows:

$$r_i = \overline{x} + \overline{y}(\omega \cdot z_i) \tag{7.12}$$

The differences in incident duration means and variations between group (location i) will be analyzed. The observed variance in the explanatory variable, $\omega \cdot z_i$, is partitioned into components attributable to different source of variation. Our hypothesis is that the time to clear the incident will be significantly longer when the arrival of CHART or Fireboard units are delayed (i.e., $1 < \omega \cdot z_i$). On the contrary, the time to clear the incident will be minimized and lower than the average value ($\overline{y}$), when the arrival of emergency units is quicker (i.e., $0 < \omega \cdot z_i \leq 1$). Total delay caused by primary incidents and secondary incidents is calculated by using parameters ($\overline{x}$, $\overline{y}$) for the freeway network and coefficients ($\omega \cdot z_i$) unique to each

location.

7.4 Impact of Secondary Incidents

After emergency response units clear the primary incident, closed lanes are re-opened, and traffic conditions recover to free-flow. The secondary incident, occurring within the vehicle queue caused by the primary incident, gives rise to additional shoulder and/or main lane closure. If a queue, caused by a primary incident at the upstream of the secondary-incident site, has not been fully dissipated when the secondary-incident recovery starts, the traffic recovery will be disturbed. Because of the disturbance of departure at the secondary incident site, impact on the road can be expressed as a drop of discharge flow rate. The reduction in flow arrival at the downstream location is not observed in this research.

In general, a primary incident associated with a secondary incident, or multiple secondary incidents, causes more speed reduction than an independent incident when their severities are same. Also, the definition of secondary incidents includes severe collisions (e.g., resulting in injuries) that impose more speed reduction, clearing activities, and rubbernecking than minor incidents (e.g., disabled vehicles). In this dissertation, delay and incident duration are averaged for each incident site assuming that some severe incidents more likely occur at specific locations.

Figure 7.2 is the stretch of the site. For example, zone 2 consists of a free-way segment located from "0.04 miles west of Thornton road" to "1.32 miles east of Greenspring Ave" on I-695 in Maryland, USA. The westbound direction of the

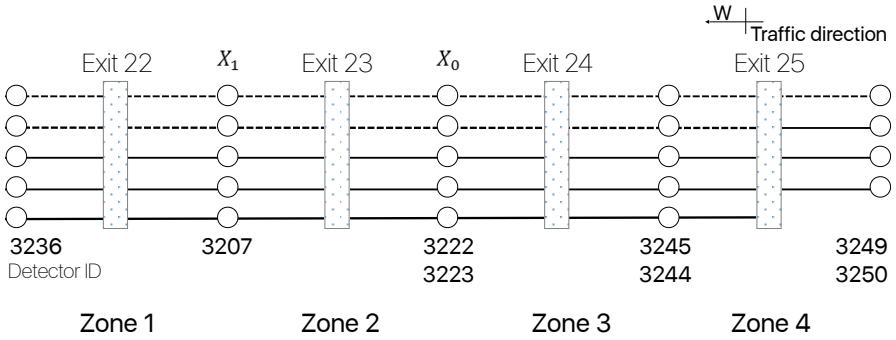


Figure 7.2: Westbound I-695 corridor (Exit 22-25).

freeway has three main lanes with a typical two-lane exit ramp at location X_0 and X_1. The exit ramp is enough apart from the nearest upstream on-ramp such that the weaving effect is disregarded inside of each zone. Without on-ramp, the common bottleneck from recurring congestion does not occur on this segment. Instead, incidents occurring within 0.5 miles upstream of study site (i.e., near I-83 Exit 23) are considered as a main cause of congestion. This research only includes secondary incidents occurring in the same zone where primary incidents occurred. To minimize the impact of distance between two incidents, secondary incidents caused by primary incident at different zone are excluded. Each lane has inductive loop detectors to measure vehicle counts, speed, and occupancy. Loop detectors located at X_0 station have zone ID: 3207 and X_1 station have zone ID: 3223.

The Center for Advanced Transportation Technology Laboratory at the University of Maryland provided the station speed as the volume-weighted average of the detector speeds and extrapolated station volume if only few of the constituent detectors return data. Identified missing data are replaced with the average value between upstream and downstream detector at the same time interval.

Figure 7.3 shows traffic condition changes of upstream X_1 station during the

day on weekdays from January 2 to January 20, 2012. Lane-by-lane (totaling five lanes) speed data is averaged to zone speed. During normal conditions, speeds are in free-flow state, from 55 to 65 mph. When an incident occurs, drivers at the front of the queue move slowly and accelerate away from the incident site. The upstream segment of the incident site experiences congestion due to the queue formation. The speed observed in this segment depends on how far the incident is from the position in the queue.

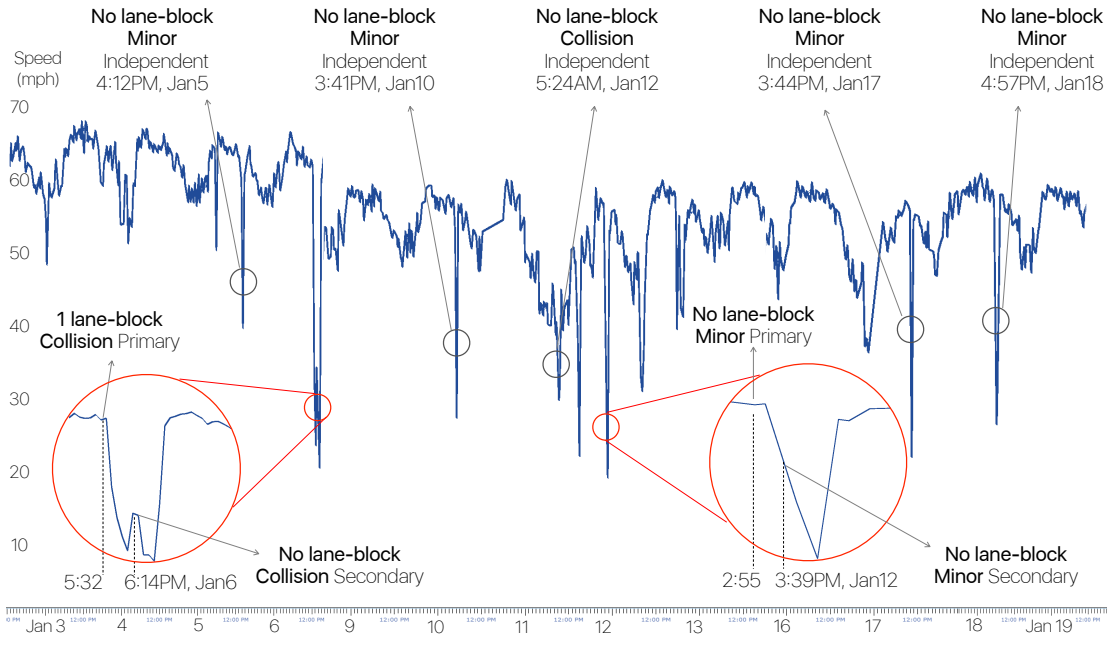


Figure 7.3: Speed reduction due to different types of incidents

Two types of secondary incidents are presented over two weeks of time horizon (only weekdays): *Type 1* (e.g., January 6, 2012) is a secondary incident that occurred in the recovery stage at t = 6:14 PM, 11 minutes after its primary incident was cleared. Discharge flows recovered partially at t = 18:00, but the curve reveals another speed drop due to clearing the secondary incident. The speed drop (during recovery time) depends on how much time has elapsed since the primary incident

was cleared and the queue was discharged. *Type 2* (e.g., January 12, 2012) is a secondary incident occurred in the clearance stage at t = 3:39 PM, 34 minutes after its primary incident occurred. These are queued speeds caused by discharge flow drop.

These two types of secondary incidents are mapped in Figure 7.4 by presenting the basic diagram of traffic for flow-density (occupancy) relationship curve. An "inverted-V" shape is a plausible representation of their relationship to identify the amount of capacity drop. As provided in the previous studies, occupancies of 17% or less denote free flow traffic conditions, where flow = demand; and occupancies greater than 17% roughly denote queues. There appears to be strong evidence that the traffic operations on a freeway can move from one normal branch of the curve (e.g, 4-5PM, January 20) to the incident condition (e.g, 4-5PM, January 6) without going all the way around the capacity point, when secondary incident occurs.

The discharge flow ranges from normal traffic conditions to non-recurring congestion at the same time on same weekdays, e.g., Mondays, at different dates during the month (e.g., 4-5 PM, January 12 and January 19, 2012). There are significant discharge flow drops for both types of incidents (i.e., Type 1 from January 20, 2012; Type 2 from January 19, 2012). Compared to Type 2, larger reduction of discharge flow is observed in Type 1, from the case when a collision primary incident and a minor secondary incident occurred. Capacity reduction caused by a minor primary incident with a minor secondary incident is higher than capacity reduction caused by independent incidents. Without lane blockage, a realized capacity reduction due to secondary incidents can be larger than primary incidents.

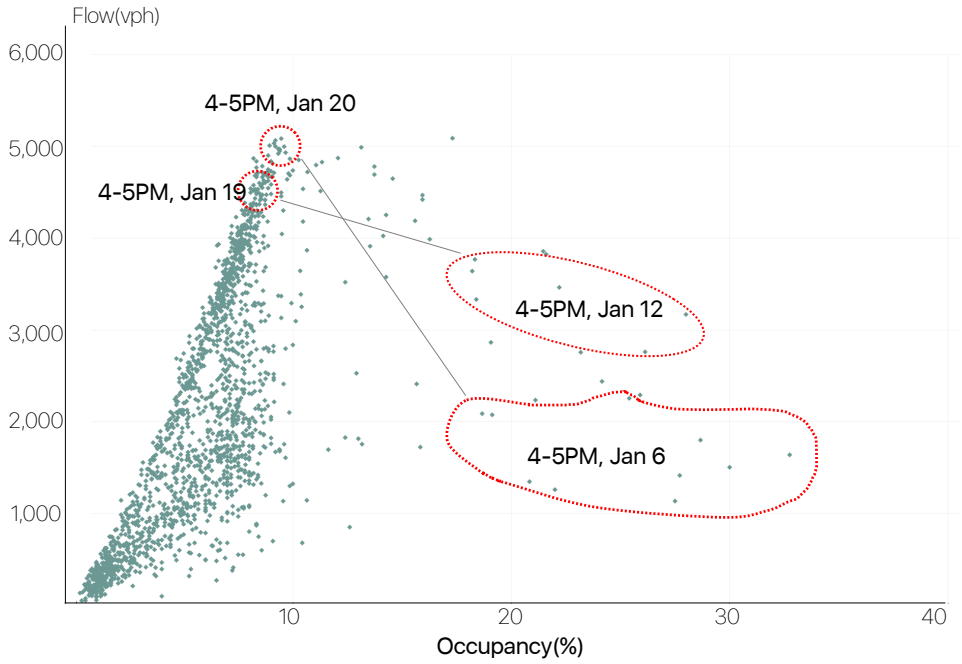


Figure 7.4: Flow-occupancy curve considering congestion caused by secondary incidents.

The above analysis shows a need for estimation of different parameters for capacity reductions from secondary incidents. Different occurrence time of secondary incidents is also expected to have impact on capacity reduction.

7.5 Case Study

This section consists of preparing parameters that can be used in the estimation of total delay. The rate of discharge flow drop is empirically analyzed under impact of independent incidents and secondary incidents. The following example illustrates the application of the model to a freeway segment.

7.5.1 Data Description

Incidents occur on freeway sections of Baltimore Beltway (I-695) extending around Baltimore, Maryland, USA. It is a 51-mile-long segment, with 40 exits and intersects with other major roads (e.g. I-97, I-70, I-83, etc.). The relationship between primary incidents and secondary incidents is based on secondary incidents identified in a previous study [16]. A total of 206 primary-secondary incident pairs were identified from January 2012 to August 2013. In this research, a location at upstream and downstream of I-83 Exit 23 (see Figure 7.2) is considered to analyze the impacts of independent, primary, and secondary incidents. Any incident that lasts shorter than 1 minute, remains longer than 2 hours, or has no valid traffic data is regarded as an outlier, and is not considered.

Table 7.1 shows results of the duration parameters for primary incidents and secondary incidents introduced in Equation 7.12. Response time is quicker than secondary response delay for primary incidents and secondary incidents. In general, the time to clear primary incidents is longer than secondary incidents. However, the contribution of response delay and its coefficient is higher for secondary incidents, and total incident duration is extended to 15.9 minutes on I-695 at Exit 23.

Table 7.1: Incident Duration of Primary and Secondary Incidents

Time	Primary incidents	Secondary incidents
$\overline{x}$ (response time)	3.2 min	3.6 min
$\overline{y}$ (clearance time)	27.7 min	9.0 min
z_i (response delay)	6.6 min	7.2 min
ω (delay coefficient)	0.15	0.19
r_i (incident duration)	30.6 min	15.9 min

What comes first is the empirical evidence linking independent incidents to reduction in capacity. When an incident occurs, the sustained flow that can be observed as the capacity of an active bottleneck can be differentiated from high flows that can occur in a roadway. The capacity of the freeway section is 1,800 vphpl and the traffic flow rate is observed on the field. After the incident is removed, traffic condition is restored to normal flow and the traffic dissipates at a rate of capacity. Total delay for all vehicles influenced by the incidents will be estimated by using the proposed model. The HCM provides a general method to categorize the remaining capacity of a road segment under incident conditions. The number of opened lanes in the freeway section and incident severity are qualitative representations of roadway operating conditions. Furthermore, larger capacity drop is presented with empirical data related with secondary incidents.

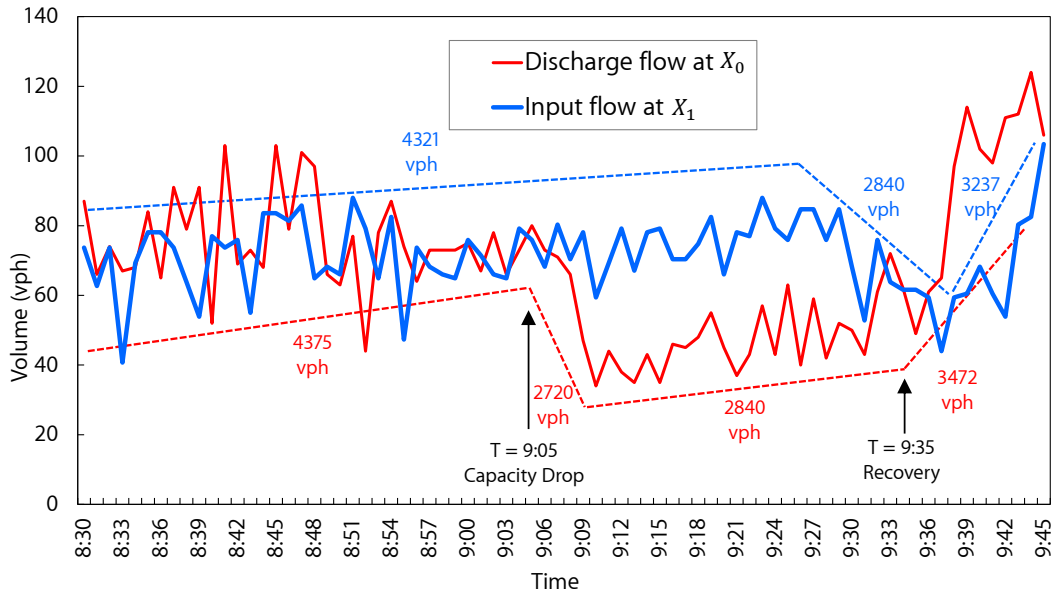
7.5.2 Independent Incident Impact

Before studying the model, the impact of incidents is empirically analyzed. Figure 7.5(a) shows a high input flow of 4,321 vph persisted at X_1 prior to t = 9:24 (Dec. 8, 2012). Emergency response units blocked one shoulder lane and three main lanes until incident was partially cleared at t = 9:35. Only one opened lane remained to vehicles to pass by during this period (from t = 9:05 to t = 9:35). The dotted trend lines highlight a reduction in discharge flow. Input flow at X_1 diminished soon thereafter, constrained by the arrival of the queue from the lane blockage downstream. When the queue′s front passed over X_1, the flow began to increase. By about t = 9:38, flow at X_1 rose to a rate of 3237 vph.

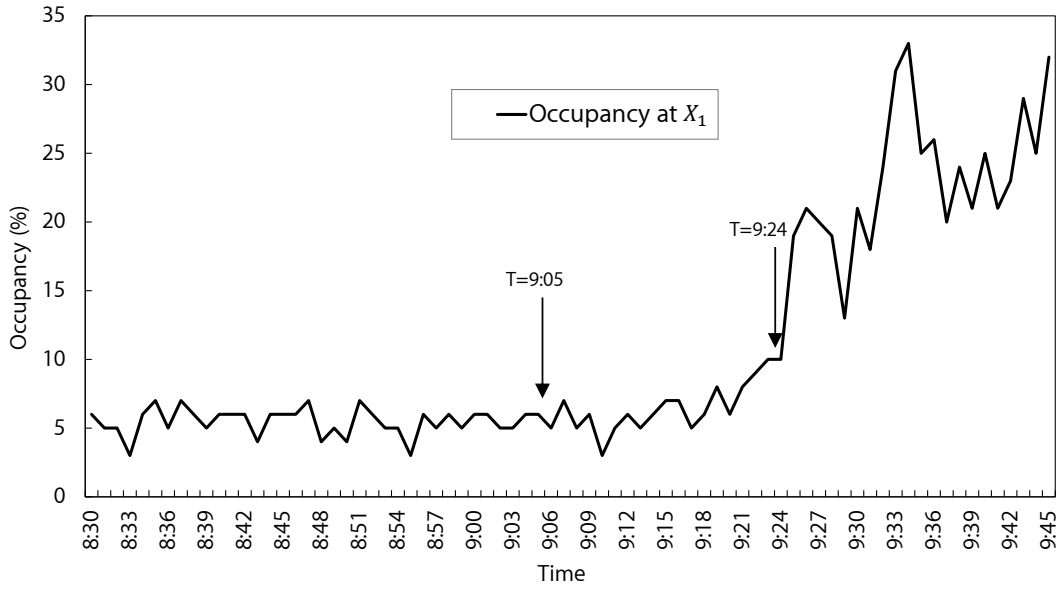
The lower curve in Figure 7.5(a) was constructed from the counts at location X_0 located downstream of the lane blockage. The figure shows that the high rate of flow 4375 vph (averaged for 35-minutes period, 8:30-9:05) prevailed until the capacity drops at t = 9:05. Discharge flow at this time dropped to an average of 2720 vph, a 38% reduction from preceding average rate and the lowest rate observed during the peak. However, after 4-minute period, discharge flow starts to recover to an average 2840 vph from t = 9:09 to t = 9:35, after emergency units finished their job. Visual comparison of the two curves shows that in the recovery, discharge flow was slightly higher than the input flow.

Figure 7.5(b) presents a time series of the occupancies that accompanies capacity drop. Occupancy was obtained by measuring the percentage of time during which the detector is "occupied" by vehicles across all freeway lanes. The figure

shows that occupancy rose steadily, 19 minutes after capacity drop, beginning at about t = 9:24 when the queue from the downstream lane closure moved upstream.



(a) Discharge flow drop



(b) Occupancy increase

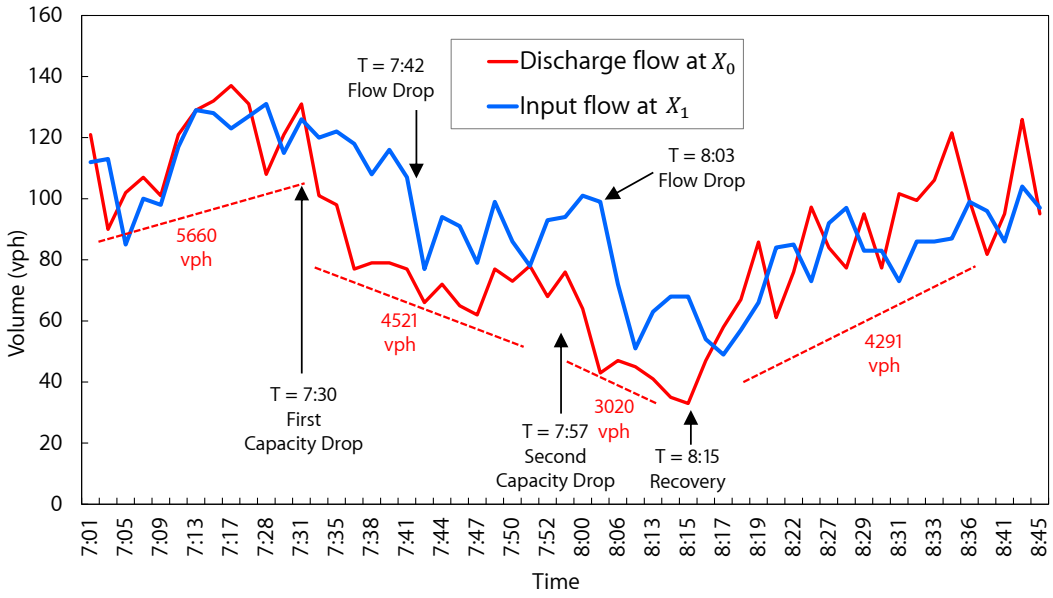
Figure 7.5: Impact of an independent incident

7.5.3 Secondary Incident Impact

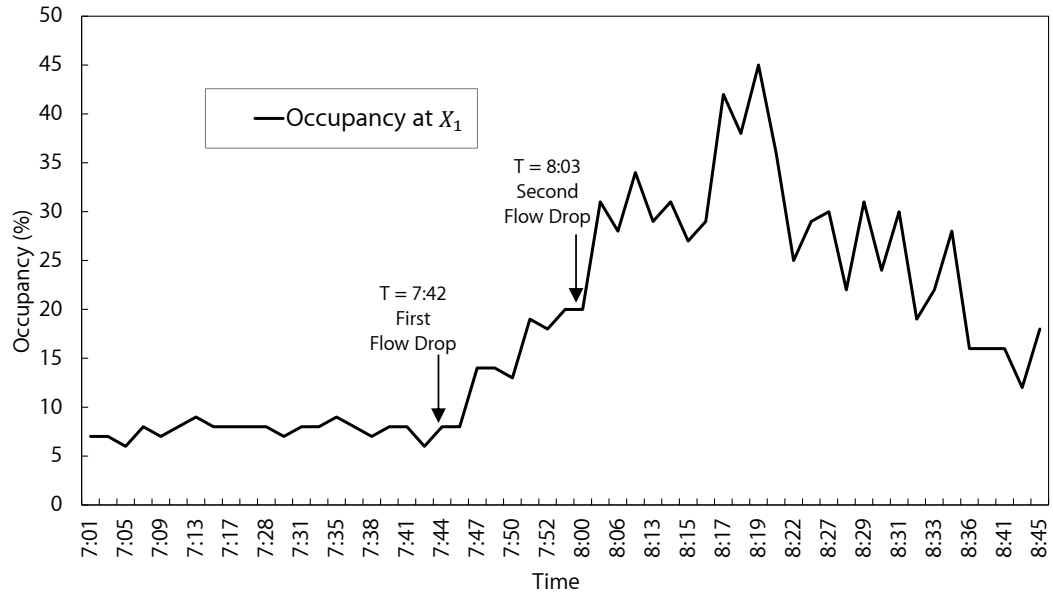
Figure 7.6(a) shows two times of capacity drop due to impact of incidents. Vehicles from morning rush hours present discharge flow at X_0, an average of 5660 vph. To clear a *primary incident* upstream location at t = 7:30, emergency response units periodically blocked traffic on the shoulder and one main lane. This resulted in a drop in capacity to an average of 4521 vph, a 20% reduction from preceding average rate (5660 vph). The vertical displacement between two curves (input flow at X_1 and discharge flow at X_0) are the excess vehicle accumulations (queuing) on the intervening freeway segment. After queue build up downstream, there is noticeable increase in occupancy (see Figure 7.6(b)). A capacity drop ultimately occurred when the segment′s density reached a certain point. The figure shows the occupancy rise starting at t = 7:42, corresponds to the onset of queuing.

Before road condition is recovered to normal condition, a *secondary incident* at the upstream location occurred at t = 7:57 and dropped discharge flow to an unprecedented level of 3,020 vph. This is an approximately 47% drop from the maximum capacity observed at peak hour of same day (5,660 vph). This drop was caused by not only lane closures but distracted drivers (i.e., "gawking" or "rubbernecking effects"). Existing vehicles could not cut through the queue because of the denser exit queue. Since drivers already were in queue from primary incidents, rubbernecking effects to secondary incidents are more obvious causing slower speed. As a result, discharge flow rate diminishes. These rubbernecking effects result in the lowest capacity observed at this site. At t = 8:03, 8 minutes after the secondary

incident occurred, the occupancy rose up to 45% as a result of exiting queue that had been formed since the primary incident occurred. Beginning at t = 8:15, after incident clearance, the discharge flow partially recovered to 4291 vph.



(a) Discharge flow drop



(b) Occupancy increase

Figure 7.6: Impact of a secondary incident

7.5.4 Results

The proposed delay models (i.e., SIDM and SSIDM) were tested for 35 day samples on freeway. For comparison, experiments were performed for different types of incident occurrences depending on when secondary incidents occurred (Figure 7.7). Each clearance period of secondary incidents spanned the time from initial formation of the traffic queue to the extension depending on the overlaps (-): Type 1 and gaps (+): Type 2. Type 1 incidents have gaps range from 1 min to 57 min and Type 2 incidents have overlaps range from 1 min to 71 min. Type 2 incidents count on the un-recovered queue generated by a primary incident, and incidents with overlapping period within 10 minutes result in more capacity reduction for secondary incidents.

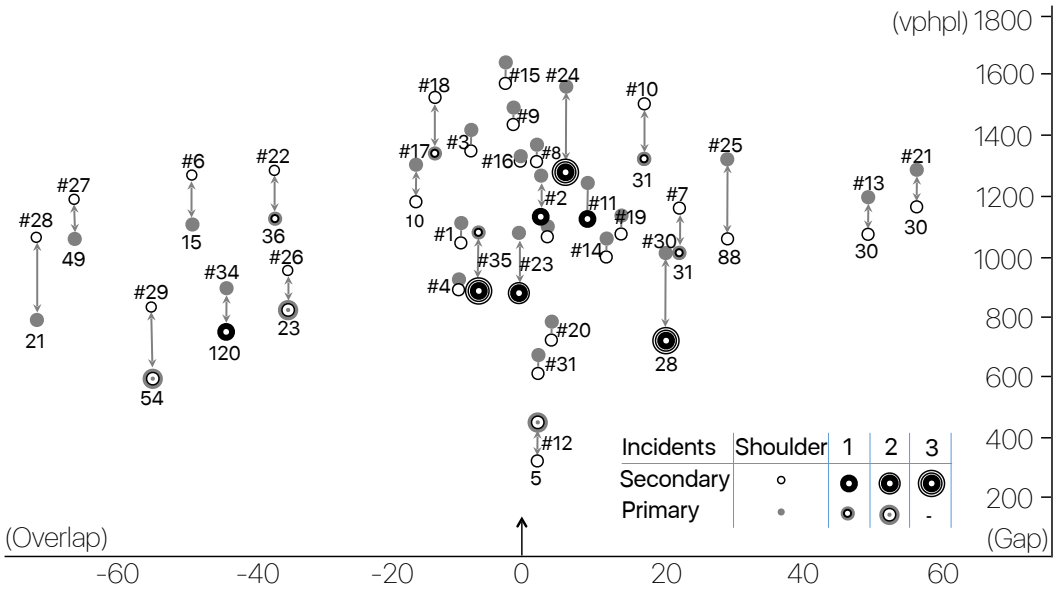


Figure 7.7: Capacity difference between primary and secondary incidents.

Note that secondary incidents starting and ending before clearance of primary

incident have less capacity reduction than primary incident one with the same condition. The queue generated by a primary incident is partially cleared and therefore this type of incident generates less capacity reduction as the gap increases.

The outcomes are summarized in Table 7.2 with observation of reduced capacity due to primary and secondary incidents in the fourth and fifth column. The capacity of the freeway section (s) was measured as the maximum of the observed flow (q) at the downstream detector one week before and after the incident. The HCM capacity adjustments for incidents are given in next two columns with residual capacity depending on remaining number of lanes before and after the incidents. Compared with the field observations, the results from the HCM show an underestimation for 3 lane-closure cases and an overestimation for 1,2, and shoulder lane-closure cases.

Depending on the time when a secondary incident occurs, there is a capacity drop due to additional queue formation, or capacity recovery due to queue dissipation. In general, a smaller consolidated-capacity (s_3) is observed when secondary incidents occur much earlier than clearance of primary incidents. For type 2 incidents, as a secondary incident occurs earlier, more time passes without recovery from impact of the primary incident, resulting in rise of capacity reduction. This larger capacity reduction causes higher estimation of total delay with the same lane-closure condition. This is due to existing impact of primary incidents on the road. By definition, a secondary incident always occurs under the impacts of a primary incident. Proposed new capacity values can be used to clearly describe incident-induced delay in geometric surface area and estimate delay with simple calculation

when secondary incident occurs.

Table 7.2: Deterministic Estimation for Secondary Incident Delay

ID #	Date	r_{ps}	Observed capacity (vphpl)		HCM capacity (vphpl)		s_3 (vphpl)	Total delay(SIDM) (vehicle hours)
			Primary	Secondary	Primary	Secondary		
1	12/6/11	-9	1,100	1,056	1,566	1,566	1,063	3,538,156
2	9/18/12	3	1,376	1,297	1,566	1,170	1,359	5,048,040
3	11/27/13	-7	1,366	1,214	1,566	1,566	1,229	2,709,395
4	12/2/13	1	1,311	1,300	1,566	1,170	1,315	6,386,449
5	12/23/13	-1	1,141	1,123	1,566	1,566	1,151	35,226,933
6	1/30/14	-48	1,143	1,274	1,566	1,566	600	338,780
7	4/22/14	20	1,033	1,175	1,170	1,566	1,257	10,010,233
8	10/2/14	2	1,301	1,236	1,566	1,566	1,271	36,006,881
9	7/18/11	0	1,501	1,425	1,170	1,566	1,499	72,881,696
10	6/19/12	15	1,520	1,340	1,170	1,566	1,504	55,751,235
11	1/2/13	14	1,140	1,030	1,566	1,170	1,251	5,708,578
12	1/15/13	1	450	210	720	1,566	481	452,366
13	12/26/13	50	1,112	1,102	1,566	1,566	1,343	39,030,951
14	1/21/14	1	600	530	1,566	1,566	625	163,464
15	4/17/14	-3	1,650	1,590	1,566	1,170	1,597	15,151,421
16	7/10/14	-2	1,130	1,206	1,566	1,566	1,125	1,646,314
17	10/31/14	-10	1,325	1,101	1,566	1,566	1,178	387,458
18	11/11/14	-7	1,390	1,425	1,170	1,566	1,333	2,807,023
19	11/24/14	12	1,065	1,010	1,566	1,170	1,059	1,922,397
20	12/20/11	4	773	752	1,170	1,170	979	860,028
21	6/12/12	57	1,276	1,206	1,566	1,566	1,338	43,291,288
22	11/27/12	-36	1,145	1,294	1,170	1,566	1,105	5,901,420
23	4/5/13	0	1,091	892	1,566	720	938	5,944,458
24	5/27/13	3	1,218	1,077	1,566	360	1,110	3,622,903
25	8/9/13	27	1,316	1,079	1,170	1,566	1,260	109,163,098
26	12/11/13	-34	877	961	720	1,566	789	5,744,970
27	2/4/14	-66	1,127	1,189	1,566	1,566	872	3,155,116
28	7/14/14	-71	852	1,077	1,566	1,566	472	179,981
29	3/7/13	-54	618	843	720	1,566	556	1,811,853
30	3/10/14	19	1,024	766	1,566	360	974	4,004,676
31	1/6/12	5	811	623	1,170	1,170	792	1,085,903
32	9/18/14	14	986	988	1,170	1,566	1,057	956,470
33	11/4/14	9	1,210	1,124	1,566	1,566	1,246	2,798,324
34	11/6/14	-43	906	786	1,566	1,170	824	26,005,543
35	11/19/14	-9	1,001	903	1,170	360	918	1,107,447

Table 7.3 presents use of the SSIDM in real time. For example, in the zone 1, primary incident is predicted to remain 20 min with standard deviation of 15 min and secondary incident is predicted to last 45 min (standard deviation of 30 min). From observed capacity data, the remaining capacity with primary incident (1,162 vphpl) and secondary incident (1,095 vphpl) can be estimated. Furthermore, assuming that the standard deviation of consolidated-capacity is the average capacity-reduction of primary and secondary incident (242vphpl), stochastic total delay can be estimated. The total delay under stochastic delay model (SSIDM) is expected 4.5 times larger than those under the deterministic total delay model (SIDM). This is due to the effect of the variability of the reduced capacity and incident duration. However, when the expected capacity (1162 vphpl) is significantly different from that of observed capacity (902 vphpl), the SSIDM may underestimate total delay.

Such limitation can be improved with an adjustment of traffic update from detectors. In the planning stage of delay estimation, the proposed model can be applied. Once new data are available after 10 minutes, the quality of the mean and standard deviation of new capacity value will increase so that more accurate delay estimation is possible.

In incident management, local ramp metering can be a solution to minimize the impact of incidents. When the estimation of capacity reduction is high, the input flow can be reduced by controlling ramp flow to the road. The queue formations at the incident cite can be alleviated during clearance duration.

Table 7.3: Stochastic Estimation for Secondary Incident Delay

Zone	ID #	r_{ps}	Stochastic capacity (vphpl)						Total delay (SSIDM) (vehicle hours)	SSIDM ÷ SIDM
			$\bar{s}_1$	σ_{s_1}	$\bar{s}_2$	σ_{s_2}	$\bar{s}_3$	σ_{s_3}		
1	1	-9	1162	217	1095	266	1216	242	15,965,604	4.5
1	2	3	1162	217	1061	331	1198	274	37,783,081	7.5
1	3	-7	1162	217	1095	266	1255	242	48,700,172	18.0
1	4	1	1162	217	1061	331	1200	274	32,468,988	5.1
1	5	-1	1162	217	1095	266	1224	242	24,489,515	0.7
1	6	-48	1162	217	1095	266	1299	242	7,664,952	22.6
1	7	20	745	226	1095	266	1075	246	43,942,526	4.4
1	8	2	1162	217	1095	266	1227	242	43,925,776	1.2
2	9	0	1160	269	1095	266	1213	267	23,553,785	0.3
2	10	15	1160	269	1095	266	1137	267	77,346,939	1.4
2	11	14	1162	217	1061	331	1162	274	24,629,510	4.3
2	12	1	745	226	1095	266	938	246	3,967,898	8.8
2	13	50	1162	217	1095	266	1357	242	70,107,166	1.8
2	14	1	1162	217	1095	266	1233	242	7,831,953	47.9
2	15	-3	1162	217	1061	331	1223	274	85,883,476	5.7
2	16	-2	1162	217	1095	266	1231	242	8,461,791	5.1
2	17	-10	1162	217	1095	266	1251	242	6,164,932	15.9
2	18	-7	1160	269	1095	266	1259	267	23,523,462	8.4
2	19	12	1162	217	1095	331	1259	274	9,996,890	5.2
3	20	4	1160	269	1061	331	1205	300	27,041,868	31.4
3	21	57	1162	217	1095	266	1227	242	116,847,241	2.7
3	22	-36	1162	269	1095	266	1280	267	18,228,982	3.1
3	23	0	1162	217	892	239	1095	228	24,590,924	4.1
3	24	3	1162	217	915	127	1112	172	35,133,215	9.7
3	25	27	1160	269	1095	266	1140	267	176,834,810	1.6
3	26	-34	745	217	1095	266	1052	242	10,246,643	1.8
3	27	-66	1162	226	1095	266	1330	246	8,762,996	2.8
3	28	-71	1162	217	1095	266	1237	242	4,949,998	27.5
4	29	-54	745	226	1095	266	1036	246	5,705,029	3.1
4	30	19	1162	217	915	127	1139	172	35,399,075	8.8
4	31	5	1160	269	1061	331	1190	300	21,340,512	19.7
4	32	14	1160	269	1095	266	1227	267	28,854,504	30.2
4	33	9	1162	217	1095	266	1240	242	53,643,428	19.2
4	34	-43	1162	217	1061	331	1143	274	7,847,492	0.3
4	35	-9	1160	269	915	127	1118	198	18,158,085	16.4

7.6 Conclusions

We proposed a model to estimate overall capacity reduction when secondary incidents occur. Overlaps and gaps between occurrence and clearance time of primary incidents and secondary incidents generate different magnitude of capacity reduction. A traditional deterministic queuing model was revised to include the impact of secondary incidents with observation of queue formation and dissipation. There is empirical evidence that occupancy (a dimensionless measure of density extracted by loop detectors) correlates with additional capacity drop due to secondary incidents. The first response unit and a secondary response unit arrival times are considered to obtain location-specific incident duration, one of the input parameters. More accurate estimation of total delay can be performed by applying the proposed reduced-capacity value.

Chapter 8: Online ERU Dispatching Problem

Previous models have focuses on solving optimal location problem with an assumption that the closest vehicles are dispatched to the request. In reality, non-uniformly distributed requests on a transportation network are more likely to have different orders that lead to different cost of the series. Under uncertainty, this approach may not capture inherently the dynamic nature of emergency response systems, especially when incidents occur at unpredictable locations at unpredictable times. We approach this challenge from an operational perspective, *online optimization*. Unlike popular nearest-origin assignment strategy that searches for greedy decisions, we consider both past and future requests. With updated information, the proposed dynamic model flexibly re-computes the solution to react in real-time. Our practical online algorithm has a look-ahead setting contingent on present requests in making future decisions.

8.1 Online Algorithm

In this section, a brief example illustrates the general concept of online optimization. We introduce a dynamic model under real-time framework, and enhance the performance.

8.1.1 Problem Statement

The k-server problem was first posed [146] as a special case of the online metrical task systems. Let $G = (V, E, d)$ be a complete graph, where V is a set of vertices on which incident may occur and $E = \{x, y | x, y \in V\}$ is a set of edges. To serve a request at y, a corresponding algorithm moves a server to y when the requested point is not served. When the algorithm moves a server from a location x to y, there incurs a cost ($\mu : E \rightarrow G$) equal to travel time between x and y in G. Our objective is to find the minimum cost function that is non-negative satisfying reflexivity and the triangle inequality.

$$\begin{aligned} &\mu(x, y) > 0 (x \neq y), \quad \mu(x, y) = 0 (x = y) \\ &\mu(x, z) \leq \mu(x, y) + \mu(y, z) \end{aligned} \tag{8.1}$$

Figure 8.1 illustrates how the online algorithm works on the k-server problem. Emergency vehicles (k-mobile servers) residing in some vertices of the graph move from a point to a different point in a network (metric service).

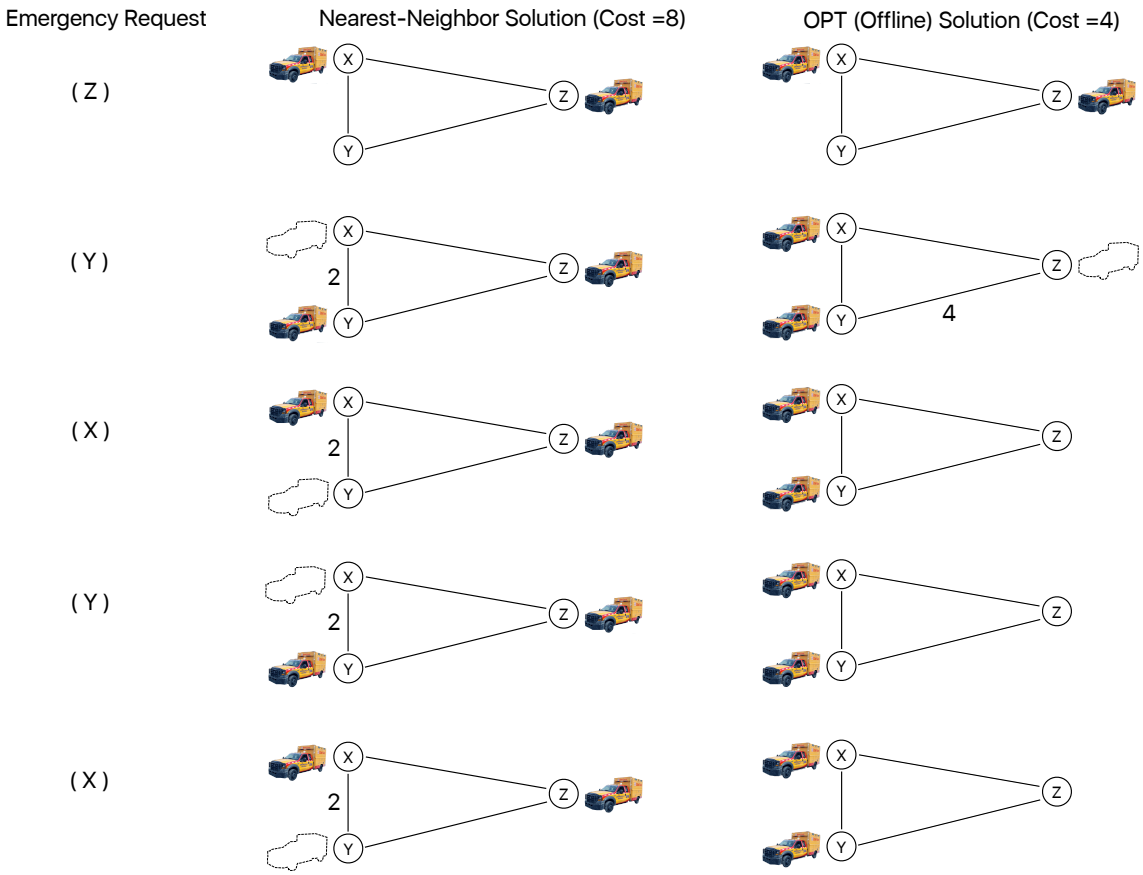


Figure 8.1: The k-server problem.

The algorithm in Figure 8.1 receives a sequence of emergency requests, in which each incident is on a point in the metric space. Consider a *2*-server problem on three points x, y, and z. A total of n $(=5)$ incidents are predicted during a fixed time-period. Emergency requests arrive for the point z followed by a long sequence of requests for the points x and y, alternating between them ($\sigma = r_z, r_x, r_y, r_x, r_y$). An online algorithm has to decide first which of the two servers should be moved to z. The initial location is x and z, therefore we do not need a cost for the first request.

A potential algorithm is the *nearest-neighbor* (GREEDY) algorithm that has been popular in most of previous vehicle dispatching strategies. It has an immediate

benefit of minimizing the cost of moving a vehicle to the emergency request. First, GREEDY assigns one of its two vehicles at z. Then all future requests will be served by the same vehicle moving back and forth between x and y. In other words, even when there is only one candidate emergency request on the network, GREEDY fails to serve this request (e.g., when an online server is far away).

On the other hand, an optimal offline algorithm (OPT) will know that such a choice is not optimal in the long run. The request is satisfied by OPT that moves the server from z to x or y after the first request is served. Then it is easy to demonstrate GREEDY does poorly (Cost=8) compared to OPT (Cost=4).

8.1.2 Model Framework

The following assumptions are made in the development of the model:

1. This dissertation deals with one type of emergency vehicle (e.g., CHART truck). We mainly focus on major emergencies with well-equipped vehicles that provide a rapid removal of incident-involved vehicles from the travel lanes. Major incidents are less likely to occur concurrently over a short time period. Instead, for minor incidents, different types and less number of emergency vehicles can provide service for gas, tire change, and hot shot. Based on incident data used in this dissertation, it takes an average of 19.8 min to clear an emergency and 1.9 min to clear a minor incident.

2. We focus on freeway emergencies in a metrical task system [147] with a symmetry. Each link of the network is assumed to have a fixed speed equivalent to 70% of the free-flow speed on that link. We consider freeway networks that have enough space on right lane/shoulder which are less likely to be influenced by severe traffic congestions.

Figure 8.2 shows the framework of the real-time dispatching system. In this dissertation, we provide an assistant for making assignment decision in daily emergency response operations. Unlike traditional approaches, vehicles do not have to return to their permanent or temporary stations, because the plan is re-generated in the next time. Emergency vehicles wait at their last stop until they receive the next order from the dispatcher.

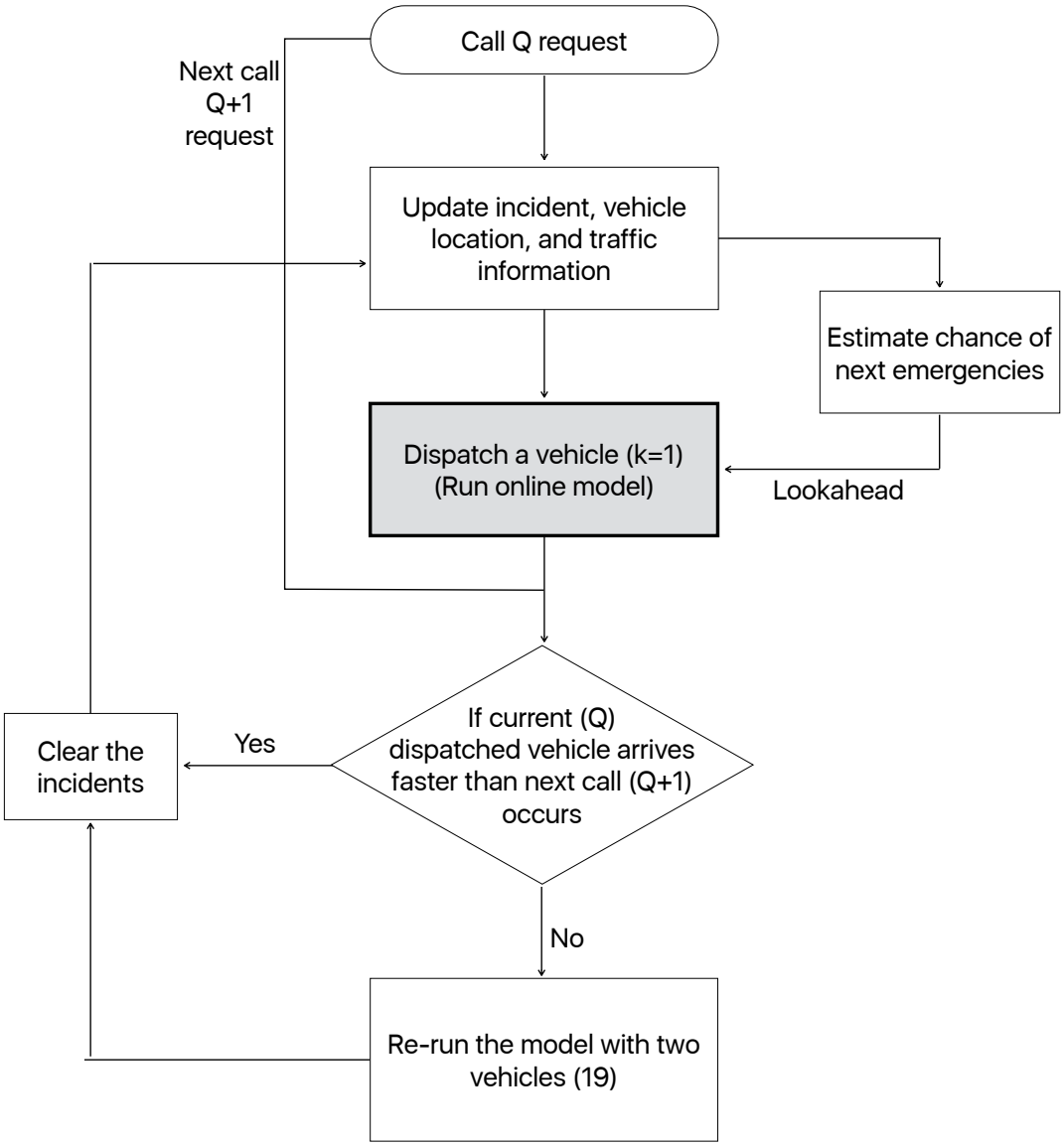


Figure 8.2: Real-time emergency dispatching framework.

In this dissertation, we incorporated the reassignment strategy [111] in the online framework with updated real-time information. If next emergency occurs before previous emergency vehicle arrives at the destination, we re-run the model with shifted sequences and choose a better solution. Estimated probability of secondary incident is used to look into the future. Without consideration of sequence, unnecessary relocation may cause more delay in responding to an emergency. Therefore, relocating vehicles is not required. Interested reader can consider that frequent relocation of vehicles may cause drivers to be confused and to make mistakes. Without a siren and an active emergency call, emergency vehicles are more likely influenced by severe traffic congestions in peak hours.

As Mirchandani and Odoni [104] assumed, an incident is serviced by the optimal available unit. If all units are unavailable, or severe emergency needs extra personnel, a unit from outside the system (e.g., depot backup) can be dispatched.

8.1.3 Work Function Algorithm with Look-ahead

Poor performance of GREEDY stems from an approach that is too conservative when it fails to capture a particular region of the network with a lot of requests (non-uniform). It would be profitable to move more vehicles nearer to this region. We try to remedy the forgetfulness of GREEDY by taking the past request points into account while determining the next assignment. Looking at the sequence until now, we determine the configuration that the algorithm would be in after servicing the sequence.

To serve the emergency request r_t with the lowest cost, the *work function* $w_t(s)$ switches possible locations of an emergency vehicle starting from s_0, serving in turn, $r_1, r_2, \ldots, r_t$, and ending up in s. Let $u(s_t, s)$ be the response time of an emergency vehicle from s_t to s at time $t + 1$. The objective is to calculate the minimum of the sum of $w_t(s)$ and $u(s_t, s)$. The Work Function Algorithm (WFA) computes the solution incrementally by generalizing dynamic programming.

$$s_{t+1} = argmin_s(w_t(s) + u(s_t, s)) \tag{8.2}$$

To evaluate the proposed models, we use competitive analysis. An online algorithm (ALG) can only approximate the performance of the corresponding OPT, which knows the whole sequence of requests in advance and deals with request at minimum total cost. Such desirable approximation property of ALG is formally described by the notion of competitiveness.

We implement WFA that has been regarded as one of the most competitive

algorithm $(2k-1)$ for the k-server problem, compared to GREEDY that is (2^k-1) competitive [147]. Many researchers have widely used WFA in practice. For further details, the reader is referred to [146, 147].

Even though considerable research has focused on studying competitive analysis, the global strategy is sometimes unrealistic in real-life. We modify WFA for optimal solution achievable by local strategies. Diffuse adversary [148] is applied to our problem for an adversary to choose the input distribution $\boldsymbol{D}$. It removes the assumption that we know nothing about the incident distribution. Instead, we have member of a given class $\boldsymbol{D} \in \Delta$ of probability distributions:

$$V(\Delta) = \min_{A} \max_{D \in \Delta} \left\{ \frac{E_D\big[\mathrm{WFA}(\sigma)\big]}{E_D\big[\mathrm{OPT}(\sigma)\big]} \right\} \tag{8.3}$$

where the expectations are taken over all incident sequences weighted with the respective probability according to the probability distribution $\boldsymbol{D}$. Instead of choosing a worst possible input, the adversary picks a distribution $\boldsymbol{D}$ so that the expected performance of the algorithm and the online optimum algorithm are as far apart as possible.

The particular class of distributions is denoted Δ_α and Δ_β $(\alpha, \beta \in D)$. The class of Δ_α contains the conditional probability $Pr_{D,\pi,\omega}(x = \tau|\sigma)$ of a secondary incident at location τ occurring after incident sequence σ, incident severity π, and environmental and traffic information ω. The class contains conditional probability of incidents $Pr_{D,\pi,\omega}(x \neq \tau|\sigma)$ at locations other than τ. Under two visions $V(\Delta_\alpha)$ and $V(\Delta_\beta)$, we choose one with better solution.

8.2 Application Design

8.2.1 Data Description

We apply the proposed online algorithms on I-695/MD-695 on which a high frequency of emergencies is present over 51 miles. An average of eight incidents occurred during the morning peak-hours (5:30-9:00AM) on weekdays from October 2012 to September 2013. Secondary incidents, within temporal and spatial impact of primary incidents, occurred once every two days. After detection of an incident, we use updated real-time traffic information in predicting the likelihood of secondary incidents [7]. Based on incident locations, we matched the travel speed of probe vehicles information, which was provided by Center for Advanced Transportation Technology Laboratory at the University of Maryland. All algorithms in this study (written in C++ programming language) compute the solution of each request in 10 sec and react in real-time.

We introduce other strategies that can be applied to online dispatching on a transportation network, is evaluated the performance.

8.2.2 Greedy Strategy

It is worthwhile to note that the proposed online dispatching strategy has different behaviors compared with the following two heuristics.

As demonstrated in Figure 8.1, greedy is a well-known heuristic. It finds the nearest emergency vehicle to each request, and moves it while ignoring history. How-

ever, in the long term, GREEDY may not benefit from strategically placed emergency vehicles close to later requests.

8.2.3 BALANCE Strategy

Balance algorithm (BALANCE) [146] considered both distance and history for the k-server problem. For each server, BALANCE maintains the total distance it has moved since the start of the incident sequence. If next incident is not covered yet, then BALANCE moves any vehicle that would have the smallest cumulative cost after moving. As indicated by its name, BALANCE tends to use all of its servers *equally*.

8.2.4 Evaluation Method

Let $C_{\text{ALG}}(\sigma)$ be the total cost incurred by ALG on σ, and $C_{\text{OPT}}(\sigma)$ be the minimum total cost on σ. We design an online algorithm that never does much worse than the optimal offline solution. An online algorithm ALG is c-competitive if its performance is estimated to be only a bounded number of times worse than that of OPT on any input with another constant a such that on every σ it holds:

$$C_{\text{ALG}}(\sigma) \leq c \times C_{\text{OPT}}(\sigma) + a \tag{8.4}$$

Suppose that the adversary generates a total of n requests. We can apply this concept to Figure 8.1: $\text{GREEDY}(\sigma) \geq \mu(y,z) + (n-1) \times \mu(x,y)$ and $\text{OPT}(\sigma) \leq \mu(x,y) + 2 \times \mu(y,z)$. As n can be made arbitrarily large, $\text{GREEDY}(\sigma)$ is unbounded. Hence, there are no constants c and a such that $\text{GREEDY}(I) \leq c \times C_{\text{OPT}}(I) + a$ on a sequence I, and so GREEDY is not competitive.

8.3 Numerical Examples

8.3.1 Application to a Real Network

Figure 8.3 presents an emergency operation during morning peak-hour on June 3^{th}, 2013. Eleven incidents were non-uniformly distributed, therefore different sequence of incidents would present different response times. Line after each incident symbol indicates clearance times and red-shaded cell indicates traffic congestion. Five CHART units were patrolling between minor incidents and emergencies.

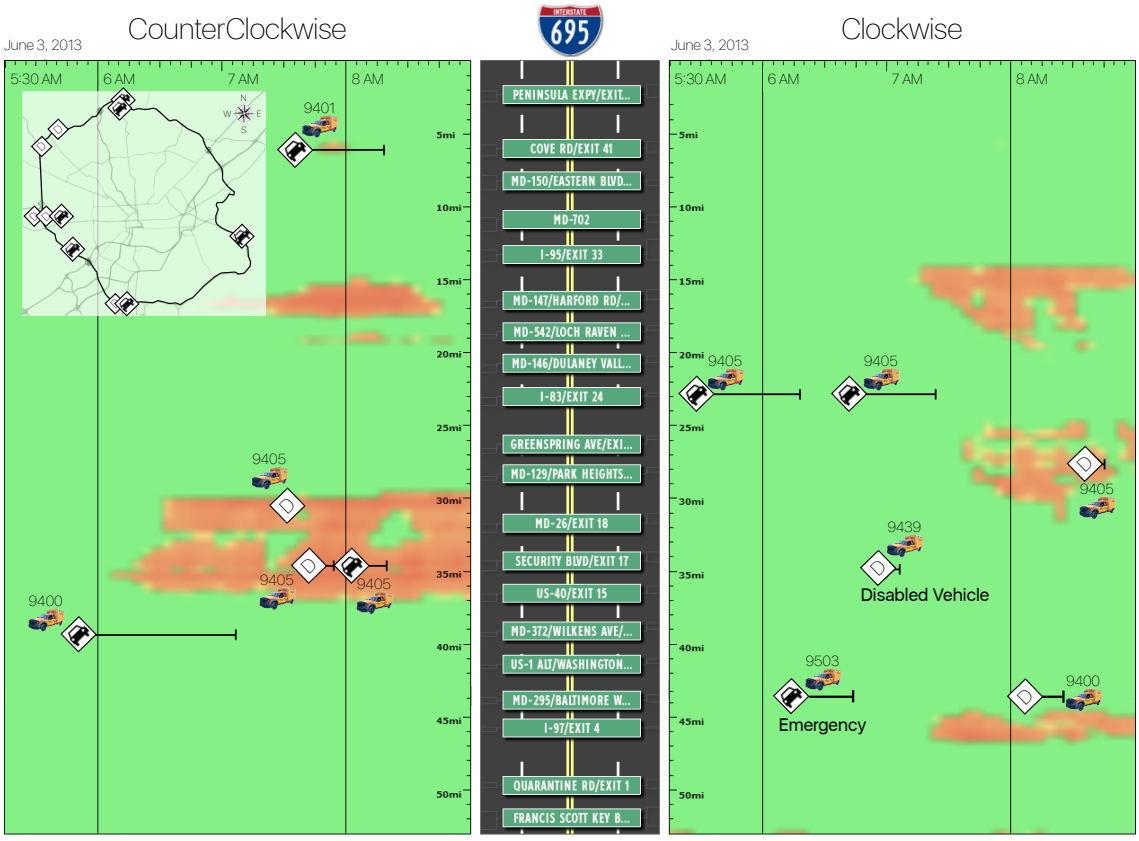


Figure 8.3: One-day example of emergency operation in the real-world (June 3^{rd}, 2013).

Table 8.1 compares the current strategy and the result of the proposed decision support system. It is challenging for the current operation to make a decision beforehand without all future information. Even though there were four units without any concurrent emergency nearby, the response time ranged between 1 min and 18 min for emergencies (total 45 min). After clearing request 6, CHART unit # 9405 was relocated in response to the patrolling plan or anticipation of future events. However, that anticipated incident was far away from emergency request 9, which takes 18 min for unit # 9405 to come back and serve. Regardless of an improvement of the current operation, if unit # 9405 stayed at exit 18 after the previous work, the response could be much quicker.

In this real-world network, WFA identifies the best unit to respond the same as OPT that makes the choice of each assignment based on the entire sequence of requests. It turns out that the solution of the previous step can provide a very good approximation for the next step. Events that require an emergency response can range from a minor incident that does not have a direct impact on the travel lane to a very major emergency that involves fatalities or hazardous material spills. The latter type of event may require faster response because they have a profound impact on the operation of the surrounding transportation network. In our proposed model, three units serve emergencies, and two units serve minor incidents, respectively.

Table 8.1: Performance of Current Strategy and Proposed Model (June 3^{rd}, 2013)

Requests Inputs	Current operation		Proposed model OPT = WFA	
	Vehicle	Response time (min)	Vehicle	Response time (min)
Emergency				
Request ID (Detect)	4 Units	-	3 Units	-
1 (5:30AM)	# 9405*	1	# 9405	1
2 (5:53AM)	# 9400*	5	# 9400	5
3 (6:19AM)	# 9503	5	# 9503	5
4 (6:45AM)	# 9405*	3	# 9405	3
7 (7:38AM)	# 9401	13	# 9403	**8**
9 (8:00AM)	# 9405*	18	# 9503	**12**
Total	-	45	-	**34**
Minor Incident				
Request ID (Detect)	3 Units	-	2 Units	-
5 (6:59AM)	# 9439	1	# 9439	1
6 (7:31AM)	# 9405*	1	# 9439	1
8 (7:45AM)	# 9405*	1	# 9439	1
10 (8:18AM)	# 9400*	1	# 9401	**2**
11 (8:39AM)	# 9405*	1	# 9439	**5**
Total		5		**10**

1. Asterisks * indicate vehicles that served both emergencies and minor incidents.

2. Bold font indicates decrease and increase in response time with the proposed model.

With fewer available units, WFA provides prompt response with 33% reduction from current operation. If emergency 2 occurred at 5:30AM before unit # 9405 arrives at request 1, we swap the order to consider assignment of two vehicles (unit # 9405 and # 9400). This example presents the importance of incident sequence on assignment decisions.

Table 8.2 presents the competitiveness of the algorithms as the ratio between

the cost incurred by the corresponding algorithm and the optimal cost incurred by OPT. The effectiveness of an online algorithm is measured by its competitive ratio that defines the worst-case ratio between its cost and that of a hypothetical off-line algorithm.

Table 8.2: Performance of Current Strategy and Proposed Model (June 3th, 2013)

Competitive Ratio	Number of available emergency vehicles		
	2 Vehicles	3 Vehicles	4 Vehicles
$C_{\text{WFA}}/C_{\text{OPT}}$	2.13	1.98	1.77
$C_{\text{BALANCE}}/C_{\text{OPT}}$	2.89	2.53	2.16
$C_{\text{GREEDY}}/C_{\text{OPT}}$	2.98	2.79	2.37

With four vehicles, performance of WFA (1.77) is better than BALANCE (2.16), and much better than GREEDY (2.37). As fewer vehicles are available, WFA outperforms compared to other reference algorithms. On typical request sequence, WFA performs well with a small competitive ratio and its behavior can never be too catastrophic.

8.3.2 A Visualization of the Algorithms

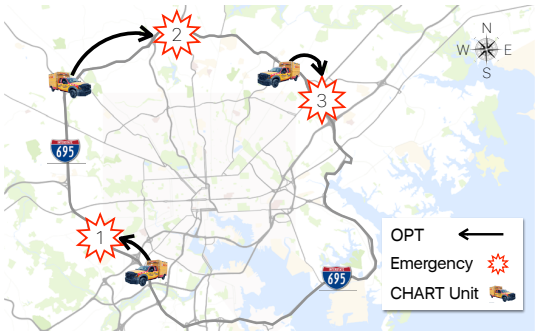
On the same network, Figure 8.4 presents a visualization of the algorithms. Table 8.3 shows a potential scenario about how other decisions would influence the performance, at each time a request arrives. The goal is to explain the experimental results intuitively in terms of the actual decisions taken by the algorithms and to provide insights on their differences in behavior and solution quality. Suppose that the initial location of emergency vehicles is $s_0 = (11, 20, 31)$, and sequence of emergency requests is $\sigma = (13, 25, 32, 27, 34, 31)$. The solutions (i.e., response times) obtained by four algorithms are 30 min (OPT), 60 min (GREEDY), 55 min (BALANCE), and 53 min (WFA) respectively. After the second request, algorithms began to have different performance. First, OPT serves the second request at exit 25 by vehicle 2, while other three algorithms serve the same request by vehicle 3. Consequently, vehicle 3 takes 14 min to serve the third request from exit 25 (Figure 8.4(b)), instead of taking just 3 min to serve the same request from exit 31 (Figure 8.4(a)). Another consequence of decisions from the past is that OPT serves the fourth request at exit 27 by vehicle 2 that was at exit 25 after serving the second request, which was significantly quicker (Figure 8.4(c)). In Figure 8.4(d), GREEDY serves the fourth request at exit 27 by vehicle 3 that was the nearest neighbor. However, Figure 8.4(f) shows that GREEDY was myopic because BALANCE and WFA can serve the fifth request with much shorter response time (3 min). Figure 8.4(h) shows BALANCE serves the last request at exit 31 by vehicle 2 that is a little farther away than vehicle 3, to equally use responses. Note that WFA made the same decisions as

OPT (Figure 8.4(c), 8.4(e), 8.4(g)), after fourth request $s_4 = (13, 27, 32, 27)$.

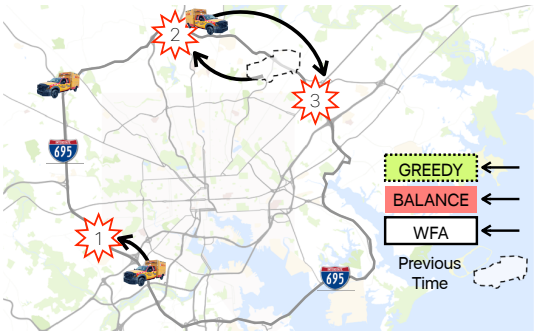
Table 8.3: Optimal Strategies for Six Sequence of Emergencies (I-695 Sample Scenario)

Emergency response	Original location	Sequence of emergencies						Total cost
		1	2	3	4	5	6	
	Exit No.	Exit 13	Exit 25	Exit 32	Exit 27	Exit 34	Exit 31	
OPT								
Vehicle 1	11	**13**	13	13	13	13	13	4 min
Vehicle 2	20	20	**25**	25	**27**	27	27	14 min
Vehicle 3	31	31	31	**32**	32	**34**	**31**	12 min
Cost	-	4 min	12 min	3 min	2 min	3 min	6 min	30 min
GREEDY								
Vehicle 1	11	**13**	13	13	13	13	13	4 min
Vehicle 2	20	20	20	20	20	20	20	0 min
Vehicle 3	31	31	**25**	**32**	**27**	**34**	**31**	56 min
Cost	-	4 min	11 min	14 min	11 min	14 min	6 min	60 min
BALANCE								
Vehicle 1	11	**13**	13	13	13	13	13	4 min
Vehicle 2	20	20	20	20	**27**	27	**31**	23 min
Vehicle 3	31	31	**25**	**32**	32	**34**	34	28 min
Cost	-	4 min	11 min	14 min	15 min	3 min	8 min	55 min
WFA								
Vehicle 1	11	**13**	13	13	13	13	13	4 min
Vehicle 2	20	20	20	20	**27**	27	27	15 min
Vehicle 3	31	31	**25**	**32**	32	**34**	**31**	34 min
Cost	-	4 min	11 min	14 min	15 min	3 min	6 min	53 min

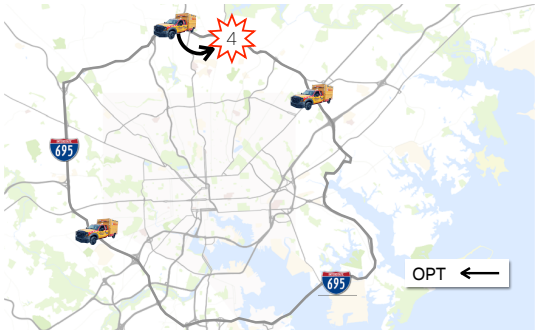
Bold texts indicate the decision of which vehicle was assigned to serve current emergency request.



(a) The 3^{rd} request with OPT

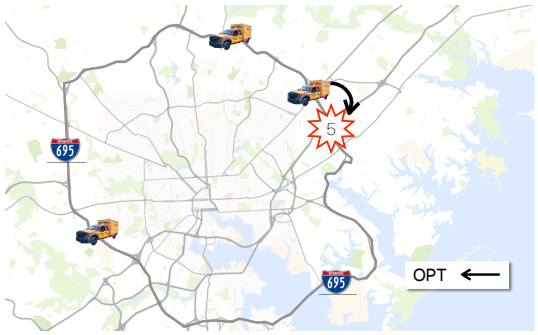


(b) The 3^{rd} request with other models

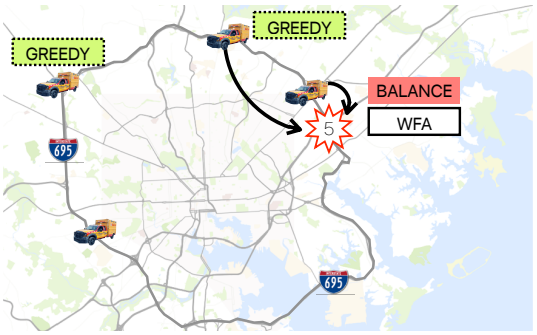


(c) The 4^{th} request with OPT

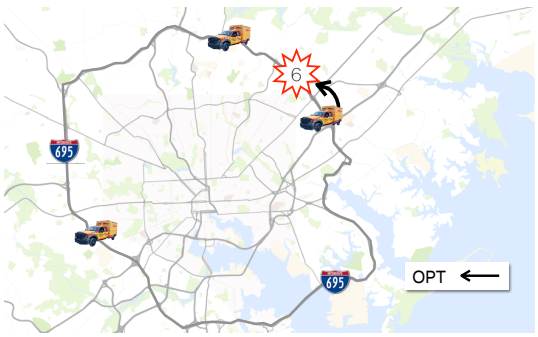
(d) The 4^{th} request with other models



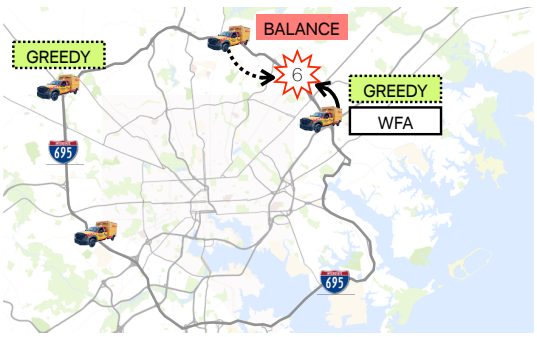
(e) The 5^{th} request with OPT



(f) The 5^{th} request with other models



(g) The 6^{th} request with OPT



(h) The 6^{th} request with other models

Figure 8.4: An illustration of response behavior of online dispatching strategies.

8.3.3 Performance Enhancement with Look-ahead

Partial information of future enhances performance of the model. A different behavior is observed when OPT serves the second request without knowing the next sequence of emergencies (Table 8.4). A blind OPT serves the second request by vehicle 3 (11 min) faster than serving by vehicle 2 (12 min). As discussed in Table 8.3, this decision turned out to be poor when the third request arrives. The consequence of this decision would become even worse if the third request arrives before clearance of the second request. This delayed service by vehicle 3 causes the third emergency request to block the traffic flow longer without a proper incident management tool.

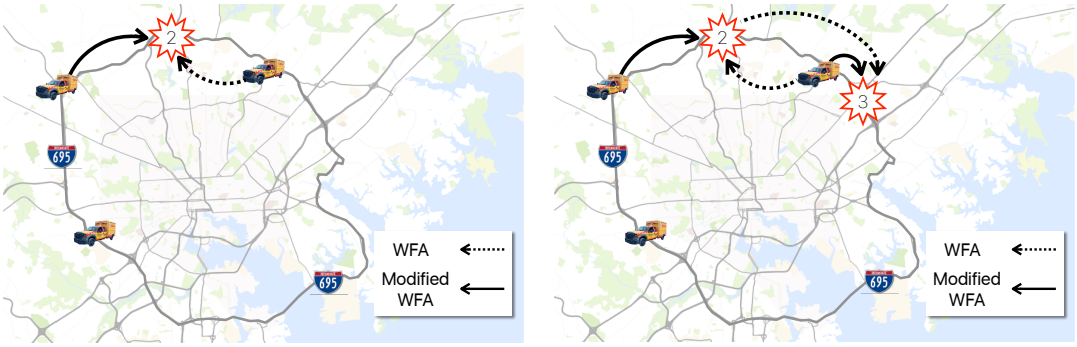
Table 8.4: Optimal Strategies without Knowing the Future Sequence of Emergencies

Emergency response	Original location Exit No.	Sequence of emergencies 1 Exit 13	2 Exit 25	Total cost
OPT				
Vehicle 1	11	**13**	13	4 min
Vehicle 2	20	20	20	-
Vehicle 3	31	31	**25**	11 min
Cost	-	4 min	11 min	15 min

Bold texts indicate the decision of which vehicle was assigned to serve current emergency request.

In the dynamic emergency nature, there are abrupt changes in the pattern of the request sequence. Conventional approaches [5, 6] that assume independent arrival times cannot justify the order of emergency requests and may result in an extremely poor performance without adapting to these changes. Instead of the independency assumption, an emergency may have an impact on the next one, a

secondary incident [7, 76]. We partially look-ahead the potential location of next requests. Note that we can predict the future only when certain sequence of emergencies has already observed. For example, Figure 8.5 presents the modification of the proposed WFA with look-ahead.



(a) Modified model for the 2^{nd} request (b) Modified model for the 3^{rd} request

Figure 8.5: An illustration of look-ahead for modification of the model.

With some belief, we can modify the original decision of WFA to serve the second request by vehicle 2 (Figure 8.5(a)). After this improvement, modified WFA works the same as OPT (Figure 8.5(b)). Even though we face the third request before clearance of the second request, the third request is served without any external delay.

We implement the diffuse adversary [148] with the sequence of emergency requests: $\sigma = (13, 25, x, , ,)$. We make a decision to serve a request under the best decisions between $V(\Delta_\alpha)$ and $V(\Delta_\beta)$: with the information of the next emergency location $x(\sigma =$ exit 25 $)$. Let $Pr_{D,\pi,\omega}(x = \tau|\sigma)$ be the probability distribution of a secondary incident occurring at x after an emergency at exit 25 (i.e., π: collision with injuries, three involved vehicles, and two lanes blocked, ω: near an interchange area with severe traffic congestions). Suppose the average of $Pr_{D,\pi,\omega}(x = \tau|\sigma)$

with τ = exit 32 is estimated to be 0.3, then there is a 30% chance of secondary incident at exit 32 under the circumstance. In other words, we have a 30% belief about the next emergency at exit 32. Let $\alpha = \sigma_{x=32} = (13, 25, 32, , ,)$ and $\beta = \sigma_{x\neq 32} = (13, 25, x, , ,)$. We estimate $V(\Delta_\alpha) = 2.1$ at 31.1% chance. Accordingly, $V(\Delta_\beta) = 2.8$ at 68.9% chance when exit 32 is not requested. We choose $V(\Delta_\alpha)$ that has 31.1% chance with less catastrophic results (2.1). It could greatly improve the performance of incident management especially when there is a significant likelihood of secondary incidents on the transportation networks.

8.4 Conclusions

In this dissertation, an emergency dispatch decision for the current incident has to made before the next incident occurs. Due to the belief that incidents on a transportation network may occur at unpredictable locations at unpredictable times, deciding which emergency vehicle to dispatch is an inherently online problem. Response requests arrive bit by bit and a sequence of dispatch decisions has to be made without perfect assumptions on the future incidents. The proposed algorithm based on dynamic programming presents better performance than current operation. It identifies the best unit to respond in the real-world operation, and its performance is close to optimal offline solution. We enhance the solution with a look-ahead to the next stage emergency.

Chapter 9: Stochastic ERU Location Problem

Determining where to locate response vehicles and how to serve incidents are important decisions that arise in developing ERU plans. While significant progress has been made in formulating and solving location and allocation problems, a number of challenging theoretical and practical issues remain to be addressed. In this section, we present limitations of previous studies and highlight the main contribution of our work. The non-linear formulation is lineralized and heuristics are introduced for a large scale problem.

9.1 Formulation

In incident management systems, the planning decision for locating ERUs needs to be made before the uncertainty is revealed. These decisions, mainly to deal with primary incidents, can be adjusted depending on the actual realization of uncertain parameters. If an incident in the past stage has not been cleared yet (depending on response and clearance), response to incidents in the present and future stages will be delayed. By considering the response delay, serious underestimation of incident duration that commonly appears in traditional models is prevented. We construct a stochastic programming model to distinguish different natures of pri-

mary and secondary incidents and to allow recourse for allocation decisions to deal with secondary incidents.

Under standard two-stage stochastic programming paradigm, the first-stage decision has to be made before realization of system uncertainties. The second-stage decisions are allowed to have recourse after a random incident occurs and affects the outcome of the first-stage decision. A recourse decision made in the second-stage is typically interpreted as corrective. Since the recourse decision is scenario-dependent, the second-stage is also a random variable.

Random events are represented by a finite, discrete set of realizations of scenarios. We consider two major sources of uncertainties, occurrence of the incidents and the locations of the incidents. In this study, ERUs are distributed to their designated locations before detection of an incident. After clearance of that incident, the ERU will remain at that location until the next incident happens. This assumption is justified because of the probability of a secondary incident happening in the vicinity of the incident. We want the response units to be as close as possible to the incidents to minimize the travel time of going to the next incident.

Our objective is to make a location decision to minimize the expected delay of all scenarios with constraints categorized as assignment, starting time of clearance, serving time, and variables. For convenience, Table 9.1 summarizes all notations used in the model formulations.

Table 9.1: Formulation Notation Table

Indexes	
n	index n, set for incident response-units (vehicles)
i	index i, set of candidate locations of origins for response units (vehicles)
j	index j, set of jobs for each incident-response unit, n
o	index o, set for defining requested incidents
ω	index ω, set of scenarios
parameters	
TT_{ij}	Travel time of response-unit going from location i to location j
CD_i	Service time required for incident at node i, also called as clearance duration (CD)
$L_{o\omega}$	Location of incident o under scenario ω
P_ω	probability of scenario ω
$H_{o\omega}$	Time that incident o happens under scenario ω
M	Big-M used for modelling
ϵ	A very small number used for modeling
Decision variables	
x_{in}	Binary decision variable which equals to one if candidate location i is selected as the starting point for vehicle n and 0 otherwise.
$a_{onj\omega}$	Binary decision variable equals one if incident o is assigned as the j^{th} job in scenario ω that vehicle n covers and 0 otherwise.
$sv_{on\omega}$	Service start time for incident o if which vehicle n is going to serve under scenario ω
$cv_{onj\omega}$	Time of clearance of incident o if done as the j^{th} job by vehicle n under scenario ω
$d_{o\omega}$	Delay of incident o under scenario ω
$s_{o\omega}$	Time at which incident o starts getting served under scenario ω and the vehicle is at the location of the incident
$c_{o\omega}$	Time at which incident o is cleared under scenario ω
$d1_{onj\omega}$	Dummy variable used for linearization
$d2_{onj\omega}$	Dummy variable used for linearization
$d3_{onj\omega}$	Dummy variable used for linearization
$f_{onj\omega}$	Binary variable indicating whether incident o is served as the j^{th} job of vehicle n under scenario ω $(= 1)$ or not $(= 0)$. The serving vehicle, n, has to be at the location of the incident for at least CD.

We formulate the ERU location-allocation problem as follows. The main goal of the objective function (9.1) is optimally locate ERUs by focusing on total delay as a function of waiting time until an ERU becomes available, travel time of the responding units from assigned location to incident site, and the clearance time of that incident.

$$minimize \;\; z = \sum_{\omega} \sum_{o} P_{\omega} d_{o\omega} \tag{9.1}$$

The first group of constraints presents rules for assignment of ERUs. Constraints (9.2) ensure that for each scenario ω and vehicle n, no incident o can be assigned as the j^{th} job unless a previous incident $p(< o)$ is assigned as the $(j-1)^{th}$ job.

$$a_{onj\omega} \leq \sum_{p<o} a_{pn(j-1)\omega} \qquad \forall\, \omega, n, o, j \neq 1 \tag{9.2}$$

Constraints (9.3) are in charge of ensuring that in each scenario, ω, at most one incident can be assigned as the j^{th} job for each vehicle, n.

$$\sum_{o} a_{onj\omega} \leq 1 \qquad \forall\, \omega, n, j \tag{9.3}$$

Constraints (9.4) make sure that each incident is assigned to one job of a vehicle.

$$\sum_{n} \sum_{j} a_{onj\omega} = 1 \qquad \forall\, \omega, o \tag{9.4}$$

Constraints (9.5) are added so that multiple similar solutions would not occur.

$$a_{111\omega} = 1 \qquad \forall\, \omega \tag{9.5}$$

Constraints (9.6) are enforcing that each vehicle has exactly one origin (starting).

$$\sum_{i} x_{in} = 1 \qquad \forall\, n \tag{9.6}$$

The second group of constraints shows starting time of each incident. Constraints (9.7) ensure that the starting time for the first job of each vehicle, under each scenario, is at least equal to the travel time of going from the vehicles origin to the location of the first assigned incident.

$$\boldsymbol{sv_{on\omega}} \times \boldsymbol{a_{on1\omega}} \geq \sum_{i} TT_{iL_{o\omega}} \times \boldsymbol{x_{in}} \times \boldsymbol{a_{on1\omega}} + H_{o\omega} \times a_{on1\omega} \qquad \forall\, \omega, o, n \tag{9.7}$$

Constraints (9.8) ensure that for each scenario, ω, the starting times for the next jobs ($j > 1$) should be at least greater or equal to the travel time of going from the previous job to this job plus the clearance duration of the previous job.

$$\begin{aligned} &\boldsymbol{sv_{on\omega}} \times \boldsymbol{a_{onj\omega}} \geq \sum_{p<o} TT_{L_pL_o} \times a_{pn(j-1)\omega} \\ &+ \sum_{p<o} \boldsymbol{cv_{pn(j-1)\omega}} \times \boldsymbol{a_{pn(j-1)\omega}} - M^{16}_{o,\omega} \times (1 - a_{onj\omega}) \quad \forall\, \omega, o, n, j \neq 1 \end{aligned} \tag{9.8}$$

The third group of constraints ensures serving time of each incidents. Constraints (9.9) and(9.10) define the starting and clearance times for each incident under each scenario, regardless of the vehicle covering it.

$$s_{o\omega} = \sum_{\boldsymbol{n}} \sum_{\boldsymbol{j}} \boldsymbol{sv_{onj\omega}} \times \boldsymbol{a_{onj\omega}} \qquad \forall\, \omega, o \tag{9.9}$$

$$c_{o\omega} = \sum_{\boldsymbol{n}} \sum_{\boldsymbol{j}} \boldsymbol{cv_{onj\omega}} \times \boldsymbol{a_{onj\omega}} \qquad \forall\, \omega, o \tag{9.10}$$

Constraints (9.11) ensure that each incident is not served any sooner than when it happens.

$$\begin{aligned} \boldsymbol{sv_{on\omega}} \times \boldsymbol{a_{onj\omega}} \geq H_{o\omega} \times a_{onj\omega} + \sum_{p<o} TT_{L_pL_o} \times a_{pn(j-1)\omega} \\ -M^{19}_{o\omega} \times (1 - a_{onj\omega}) \qquad \forall\, \omega, o, n, j \neq 1 \end{aligned} \tag{9.11}$$

Constraints (9.12) and (9.13) ensure that the serving time of an incident cannot start unless the vehicle which is in charge of serving that incident has finished its previous job.

$$\boldsymbol{sv_{onj\omega}} \times \boldsymbol{a_{onj\omega}} \leq M^{20}_{o,\omega} \times \sum_{p<o} f_{pn(j-1)\omega} \quad \forall\, \omega, o \neq 1, n, j \neq 1 \tag{9.12}$$

$$\begin{aligned} &\boldsymbol{cv_{onj\omega}} \times \boldsymbol{a_{onj\omega}} - \boldsymbol{sv_{on\omega}} \times \boldsymbol{a_{onj\omega}} - CD_{L_o} \times a_{onj\omega} \\ &+\varepsilon \times a_{onj\omega} \leq M^{21}_{o,\omega} \times f_{onj\omega} \quad \forall\, \omega, o, n, j \end{aligned} \tag{9.13}$$

Constraints (9.14) are for finding the soonest time an incident can be cleared.

$$c_{o\omega} \geq s_{o\omega} + CD_{L_{o\omega}} \qquad \forall\, \omega, o \tag{9.14}$$

The last group of constraints presents delay calculation based on above constraints and condition of each variable. Constraints (9.15) define the delay for an incident.

$$c_{o\omega} - H_{o\omega} = d_{o\omega} \qquad \forall\, \omega, o \tag{9.15}$$

Constraints (9.16) define non-negative and binary variables.

$$\begin{aligned} &f_{onj\omega},\, a_{onj\omega} \in \{0, 1\} \qquad \forall\, \omega, o, n, j \\ &x_{in} \in \{0, 1\} \qquad \forall\, i, n \end{aligned} \tag{9.16}$$

In the presented formulation, constraints (9.7), (9.8), (9.9), (9.10), (9.11), (9.12), (9.13) have non-linear terms (***bolded***). The solution procedure used for solving this problem is branch and bound. In branch and bound, at each node, we solve a linear programming relaxation of the problem by relaxing the integrality constraint for the integer variables. For this relaxation, if the program is not a linear program, it cannot be solved in polynomial time using algorithms that find the optimal solution. We transform the ERU location-allocation problem (a non-linear problem) into an equivalent linear programming problem in the next section.

9.2 Linearization

We find the optimal solution for the important linearization that is proven not to cut off the optimal solution. In this section, we address the problem of selecting an appropriate big-M. To prevent numerical issues and improve the solution time, it is the best practice to select the big-M as small as possible. Looking at the structure and inputs to the model, we have stated the value each M should assume for each constraint.

This approach enhances problem solvability by providing an equivalent linear representation. We introduce new variables and constrain these variables such that the new linear problem is a tight estimation of the original problem and contains those regions which the global minimum exists [149].

For linearizing $sv_{on\omega} \times a_{onj\omega}$ we have introduced a dummy variable $d1_{onj\omega}$ and added two constraints (9.17) and (9.18):

$$d1_{onj\omega} \leq sv_{on\omega} \qquad \forall\, \omega, o, n, j \tag{9.17}$$

$$d1_{onj\omega} \leq M \times a_{onj\omega} \qquad \forall\, \omega, o, n, j \tag{9.18}$$

The objective of adding constraints (9.17) is to enforce $d1_{onjw}$ to at most equal to sv_{onw}. Therefore $d1_{onjw}$ will be capped by sv_{onw}, which was the initial objective of the linearization. By adding constraints (9.18), we ensure that $d1_{onjw}$ will equal zero if a_{onjw} equals zero. The correctness of this type of linearization can be found in [149].

For linearizing the term, $x_{in} \times a_{on1\omega}$ we have introduced a dummy binary variable, $d2_{onj\omega}$ to equate that nonlinear term. Constraints (9.19) are added as a result:

$$d2_{onj\omega} \geq x_{in} + a_{on1\omega} - 1 \qquad \forall\, \omega, o, n, j \tag{9.19}$$

The purpose of constraints (9.19) is to bound $d2_{onjw}$ from assuming the value of zero when both of the other two binary variables (x_{in} and a_{on1w}) assume the value of 1. In that case we will have $d2_{onjw} \geq 1 + 1 - 1$ ($d2_{onjw} \geq 1$). Since $d2_{onjw}$ is binary it will assume the value of one.

Selecting good values for the big-M parameters in constraints (9.8), (9.11), (9.12), and (9.13) can be a challenge. To prevent such unwanted events, we present a range for the big - Ms based on the input parameters of the model (Table 9.2). It is advised to pick the smallest number within that domain.

Table 9.2: The Ranges for the Big-Ms

Constraints	Value of M based on inputs
17	$M^{17}_{o,\omega} \geq \sum_{p<o} TT_{L_pL_o} + \sum_{p<o} (CD_o + TT_{L_pL_o}) \quad \forall\, \omega, o$
20	$M^{20}_{o,\omega} \geq H_{o\omega} + \sum_{p<o} TT_{L_pL_o} \quad \forall\, \omega, o$
21	$M^{21}_{o,\omega} \geq \sum_{p<o} (CD_o + TT_{L_pL_o}) \quad \forall\, \omega, o$
22	$M^{22}_{o,\omega} \geq \sum_{p<o} (CD_o + TT_{L_pL_o}) + o \times \varepsilon \quad \forall\, \omega, o$

The objective is to minimize a function of delay whenever we start serving the incident the fastest based on constraints (9.7). The nonlinear term $x_{in} \times a_{on1\omega}$ would always try to assume the value of zero. By adding constraints (9.19), we prevent it from assuming the value of zero whenever both x_{in} and $a_{on1\omega}$ equal one.

To linearize $cv_{pn(j-1)\omega} \times a_{pn(j-1)\omega}$, we add a dummy variable $d3_{onj\omega}$ that is equal to nonlinear term through constraints (9.20) and (9.21):

$$d3_{onj\omega} \leq M \times a_{onj\omega} \qquad \forall\, \omega, o, n, j \tag{9.20}$$

$$d3_{onj\omega} \leq cv_{onj\omega} \qquad \forall\, \omega, o, n, j \tag{9.21}$$

To linearize the nonlinear constraints we replace the nonlinear terms with their linear equivalents. The linearized constraints (9.22), (9.23), (9.24), (9.25), (9.26), (9.27), and (9.28) are presented below:

$$d1_{on1\omega} \geq \sum_{i} TT_{iL_{o\omega}} \times d2_{onj\omega} \qquad \forall\, o, n, \omega \tag{9.22}$$

$$d1_{on1\omega} \geq \sum_{p<o} TT_{L_pL_o} \times a_{on(j-1)\omega} + \sum_{p<o} d3_{on(j-1)\omega} - M \times (1 - a_{onj\omega}) \qquad \forall\, \omega, o, n, j \neq 1 \tag{9.23}$$

$$s_{o\omega} = \sum_{n} \sum_{j} d1_{onj\omega} \qquad \forall\, \omega, o \tag{9.24}$$

$$_{o\omega} = \sum_{n} \sum_{j} d3_{onj\omega} \qquad \forall\, \omega, o \tag{9.25}$$

$$d1_{on1\omega} \geq H_{o,\omega} \times a_{onj\omega} + \sum_{p<o} TT_{L_p,L_o} \times a_{pn(j-1)\omega} - M \times (1 - a_{onj\omega}) \qquad \forall\, \omega, o, n, j \neq 1 \tag{9.26}$$

$$d1_{onj\omega} \leq M \times \sum_{p<o} f_{on(j-1)\omega} \;\; \forall\, \omega, o \neq 1, n, j \neq 1 \tag{9.27}$$

$$d3_{onj\omega} - d1_{onj\omega} - CD_{L_{o\omega}} \times a_{onj\omega} + \epsilon \times a_{onj\omega} \leq M \times f_{onj\omega} \qquad \forall\, \omega, o, n, j \tag{9.28}$$

9.3 Heuristics for a Large Scale Problem

As we look-ahead more future stages on a larger network, the problem size increases. The computational effort for solving scenario-based method depends on the scenario size. This dissertation is dealing with a complex stochastic problem with large number of constraints and variables. For example, suppose 3 stages on the freeway network with 2 ERUs on 17 nodes. Even though we linearize the non-linear terms, we have a matrix with columns more than $10\times17^3\times2\times3\times3$ (variables×scenarios×ERUs×order×job), and rows at least $17^3\times3\times16$ (scenarios× order×constraints). There may be some efficient heuristics, but this dissertation focuses on a fast scenario reduction method to meet the real-time requirements when we run the model.

A particularly efficient implementation of scenario-reduction algorithm is a fast forward selection [150]. Starting from original set of scenarios Γ and set of scenarios to be selected $|S|$ and deleted $|J|$, we select one scenario reclusively. The algorithm produces a reduced set of scenarios $\Gamma_S^{[0]}$, $\Gamma_S^{[1]}$, ..., $\Gamma_S^{[i]}$, ..., $\Gamma_S^{[*]}$, where the set $\Gamma_S^{[*]}$ is the target of the search. Note that one of the main contributions of this study is the different ordering of incident sequences. To make r stages of ordering numerically tractable, we multiply $r!$ cases of sequences (permutation) by required number of scenarios ω. To select total representative scenarios ($\omega \times r!$) out of N, we implement the following procedure:

- *Step 0* : Before starting the process, the initial step consists of computing the delay d_ω (For simplicity, we know which incident o causes delay $d_{o\omega}$).

We solve each scenario independently as a deterministic case (very fast) and calculate the severity of each scenario as the total delay for that particular scenario. Suppose we have a goal of reduced set of 50 scenario ($\times 6$ for full combinatorial in 3 stages) among N, the value of d_ω can be conveniently arranged into a systematic matrix,

$$d = \begin{bmatrix} 0 & 10 & \cdot & \cdot & \cdot & 1000 \\ 10 & 0 & \cdot & \cdot & \cdot & 990 \\ 25 & 15 & \cdot & \cdot & \cdot & 975 \\ . & . & & . & & \\ . & . & & . & & \\ . & . & & . & & \\ 1000 & 990 & \cdot & \cdot & \cdot & 0 \end{bmatrix} sec \tag{9.29}$$

- *Step 1* : Compute delay for each scenario ω, and select ω that minimizes distance D between the reduces sets Γ_S and original sets Γ. The starting scenario can be obtained from

$$D_\omega = arg\{min \sum_{w \in \Gamma} P_\omega d_{\omega\omega'}\} \tag{9.30}$$

If ω=3 is selected, then $\Gamma_S^{[1]} = \{3\}$ and $\Gamma_J^{[1]} = \{1, 2, ..., 289\}$.

- *Step i* : Update delay matrix as follows:

$d_{\omega\omega'}^{[i]} = min\{d_{\omega\omega'}^{[i-1]}, d_{\omega\omega_{i-1}}^{[i-1]}\}, \forall \omega, \omega' \in \Gamma_J^{[i-1]}$

Considering new delay matrix, we select $D_\omega \in arg \min_{\omega \in \Gamma_J^{[i-1]}} D_\omega^{[i]}$

- *Step* $i+1$: Optimally redistribute probabilities. The new probability of a preserved scenario is equal to the sum of its formal probability and of all probabilities of deleted scenarios that are closes to it. All deleted probabilities have probability zero.

The process is continued until given number of scenarios are selected. The interested reader is referred to [150] for further information about the algorithm.

9.4 Illustrative Case Study

The case study site is the Baltimore Beltway (I-695) extending around Baltimore, Maryland, USA. It is a 51-mile segment, with 40 exits and intersects with other major roads (e.g. I-97, I-70, I-83, etc.). Interested readers can vary the distance to test different sizes in any freeway network. Traffic operation center 4 (near Exit 34) covers selected routes including I-695 (Figure 9.1). There were 4 field operation patrol units available for AM peak hours on weekdays until 2014.

Potential locations for the ERUs are the exits (treated as nodes) where incidents occur. We control the potential locations of emergency requests by clustering historical frequency of incidents. Two different network sizes (i.e., 17 nodes, 34 nodes) are generated by grouping nearby incidents.

The case study presents a ring shape network where two route exists for each trip. The proposed model can be applied to a complex freeway network in which more than two routes exist for each allocation. In that case, interested readers can choose the fastest route using a shortest path algorithm and change the travel time

input of an ERU [3], [111].

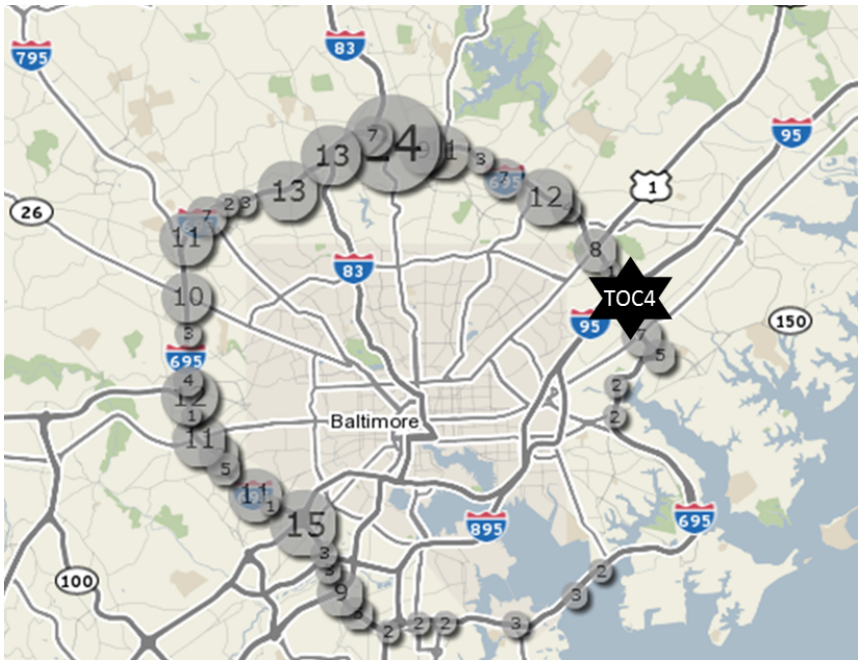


Figure 9.1: Spatial distribution of incidents on I-695 freeway

In total, 1,981 primary and independent incidents (e.g., disabled vehicles, collisions, vehicle on fire) during the morning peak hour (i.e. 6:30-9AM) for 1 year (i.e. from October 2012 to September 2013) are collected (i.e. 261 weekdays) along the I-695 corridor. As a result, an average of 7.7 (λ) incidents are occurring in each 150-minutes time period per day. Based on incident locations, the travel speeds of probe vehicles are represented on traffic message channels codes of each segment. The archived incident and probe vehicle database are provided by the Center for Advanced Transportation Technology Laboratory at the University of Maryland.

The proposed incident model (Section 3) is incorporated into the generation of scenarios. Generally, it takes an average 19.8 min for response units to clear an incident after the detection of the incident (i.e. incident duration). To respond to

another incident, it takes time for the response units to travel from previous incident location to another one after the notification. However, another incident has a high potential to be pending without appropriate response units, because the general tendency of the occurrence rate of incidents is one per every 18.5 min. Therefore, we break the morning peak hours into exponentially distributed intervals (mean 18.5 min). For an efficient emergency system, waiting time for the current request can be reduced with quick response in the previous request. Every time a request arrives, we look-ahead two future stages. Secondary incident probabilities majorly vary during the clearance or recovery of primary incidents. For the comparison of computational performance and efficiency, we also extended look-ahead setting from two to three future stages.

If next emergency occurs before previous emergency vehicle arrives at the destination, we can re-run the model with shifted sequences and choose a better solution. The new model considers updated probability of incident and real-time traffic information. However, as shown in the incident intervals, major incidents are less likely to occur concurrently over a short time period.

Clearance times are categorized with different delay types and locations. For example, exit 5($i = 1$) has average clearance duration (NC_1) of 19.6 min with following parameters: $\beta_1^1 = 0.68, \beta_1^2 = 0.94, \beta_1^3 = 1.05, \beta_1^4 = 1.35, \beta_1^5 = 0.98$. As an input to the optimization model, pure clearance time ($C_1 = 13.4$ min) is estimated for exit 5 without response delay. The same delay type (e.g., $\beta_i^2, \eta = 2$) varies for different location i with coefficient of variation (0.43) that is the ratio of the standard deviation (0.42) to the mean (0.96). This variation in delay presents more

non-uniformly distributed response delays on the network.

We test the model in two networks with different sizes ($i = 17, 34$). The main goal is to generate future stages of incident scenarios given information of past and current incidents. (Ω: number of blocked lanes, collision with injuries or property damage only) and traffic condition at upstream (Δ: difference in speed before and after incident occurrence) of primary incident [22].

We build a total of ω scenarios. For example, Table 9.3 presents 17×17 scenarios as a combinatorial of two future incidents (during stage 1 at site $i = 17$ and stage 2 at site $j = 17$). Suppose we estimate parameters based on the past incident which occurred at exit 11 ($\delta(\Omega, \Delta)_{(exit\,11,-1)(k,u)} = 0.207$, $\Omega = 2$ lanes blocked, collision with injuries; $\Delta = 30$mph speed difference), and the current incident occurred at exit 5 ($\delta(\Omega, \Delta)_{(exit\,5,0)(k,u)} = 0.098$, $\Omega = 1$ lanes blocked, collision with property damage; $\Delta = 10$mph speed difference). Based on the location of past and current incidents and the consequent traffic, we update the density in real-time. In the same logic, we estimate the expected clearance time [17].

9.4.1 Results

Our computational implementation of the formulation involves coding and solving Xpress on a computer with 2.6-GHz CPU and 32-GB RAM. Since our problems are formulated as Mixed Integer Programs (MIP) reaching the optimal solution is very time consuming. In most of the cases, running time was less than 30sec to get the near-optimal solution with a gap less than 1%. However, for 3-stages and

2-vehicle or 3-vehicle cases, we terminated most of the problems after 1400 seconds or 20% gap, since no significant improvements were observed after running the code more than that time. Starting from one available ERU vehicle, multiple ERU vehicles are tested to analyze the sensitivity of the optimal solutions and to find the number of vehicles after which increasing the vehicles will only improve the solution marginally.

Table 9.3 shows conditional probabilities that are calculated for each scenario in the example of 2 stages and 17 nodes. The expected probability of scenarios (P_ω) ranges from 0.001 to 0.041 (average probability of a scenario is 0.013). For example, the probability of the first scenario, P_1, $p(5,7) = \mathbb{E}[\tau(exit\,5,1)] \times \mathbb{E}[\tau(exit\,7,2)]$, is 0.009. Note that the probability of the scenario #17 is 0.011 which is 0.002 larger than first one. Since we have 289 scenarios, each assigned probability is small. However, the difference 0.002 takes 23% of the first scenario, and this difference may change the optimal solution of the problem. Note that the transition probabilities vary in real-time when next incident occurs, and we re-execute the optimization model.

Before an incident occurs, we pre-locate ERUs at the optimal locations with look-ahead. After an occurrence of an incident Q_i and an assignment of one of pre-located ERUs, a better relocation decision is made. At each point, the program updates current traffic condition, response and clearance status of the incident and ERU information such as the current location, the route to be taken, the destination, the time to the next incident. With new traffic condition (Δ) and incident severity (Ω), we update the probability of incident occurrences. These variables are used

Table 9.3: Probabilities of Scenarios

Scenario #	Stage 1	Stage 2	Probability
1	$\mathbb{E}[\tau(exit\,5,1)]$	$\mathbb{E}[\tau(exit\,7,2)]$	0.009
2	$\mathbb{E}[\tau(exit\,5,1)]$	$\mathbb{E}[\tau(exit\,11,2)]$	0.012
3	$\mathbb{E}[\tau(exit\,5,1)]$	$\mathbb{E}[\tau(exit\,13,2)]$	0.005
⋮	⋮	⋮	⋮
17	$\mathbb{E}[\tau(exit\,7,1)]$	$\mathbb{E}[\tau(exit\,5,2)]$	0.011
⋮	⋮	⋮	⋮
289	$\mathbb{E}[\tau(exit\,36,1)]$	$\mathbb{E}[\tau(exit\,36,2)]$	0.002
Sum			1.000

in estimating expected clearance of incidents $\hat{C}_i$ [17]. We relocate n ERUs if the expected clearance $\hat{C}_i$ of Q_i is earlier than next call $(Q+1)_i$, or $n-1$ ERUs if clearance is later than $(Q+1)_i$.

The illustrative example presents where to relocate ERUs after an occurrence of incidents. While previous literature has only considered travel time of ERUs, we calculate total delay time as the sum of travel time, response delay, and clearance time. Our model explicitly models the response delay when a server has not finished the clearing job yet. We test the performance of the emergency response model on two different sets of probabilities with maximum travel time. We obtain solutions for scenarios without considering secondary incident on freeways, and insert this solution into real-world scenarios with secondary incidents. When we have one or two available ERUs, the solution of two approaches are same. However, as more ERUs available, the benefit of considering probability of secondary incident becomes important. With 0% gap, the optimal objective function value (total delay time), was 58.69 min without consideration of secondary incidents (at 11, 18, 29). This

is worse than the solution if the locations were 11, 11, 27 (objective value= 57.13 min).

In the previous study [27] the travel time of ERUs were dependent on the traffic condition. The emergency medical service act of 1973 stipulates that 95% of service request be met within the required time [118]. However, in many cases, even though police units had been dispatched to the scene, the left lane can be blocked until available emergency units arrives. Maryland's "clear the road" policy provides ERUs (well-equipped vehicles) for the rapid removal of vehicles from the travel lanes rather than waiting for a private tow service. The proposed model repositions single type of ERUs to the best locations to serve future incidents. Most parts of United States and Canada enforce the "move over laws" that require motorists to move to the farthest roadside and stop, until the emergency vehicle has passed the vicinity. We consider freeway networks that have enough space on right lane/shoulder which are less likely to be influenced by severe traffic congestions. However, emergency vehicles still expect delays waiting for other traveling vehicles to become aware of their presence and yield. We explore both minimum (free-flow traffic) and maximum (congested traffic) response time as an input to the model (Table 9.4).

For cases with one ERU considering probability of secondary incidents, clearance of the second incident starts after waiting from previous service (9.84 min) and traveling to incident site (12.31 min). Including the actual clearance duration (17.51), total delay is 39.67 min. As we have more available ERUs, we have less waiting and travel times. It presents the importance of efficient response that has an influence on later stages of response delay. While the minimum expected total

delay with one vehicle case ranges from 27.68 min to 39.67 min, three vehicle case has a much lower value that ranges from 25.72 min to 27.68 min. For one available ERU, maximum expected delay is 1.31-1.36 times longer than minimum expected delay. As we have more available ERUs, the discrepancy between minimum and maximum delay becomes smaller (i.e., 1.26-1.28 times for 2 ERUs and 1.17-1.13 times for 3ERUs). This is due to the impact of traffic condition on the travel time of response vehicles. The real emergency response would be between somewhere in the free-flow and congested condition.

Table 9.4: The performance of the Proposed Model (Different Number of ERUs)

ERU #	Traffic	Expected time value (minutes)						
		Occur	Start	Clear	Wait	Travel	Duration	Delay
One ERU	Free	18	10.20	27.68	0.00	10.20	17.48	27.68
		36	40.16	57.67	9.84	12.31	17.51	39.67
	Real	18	14.31	31.79	0.00	14.31	17.48	31.79
		36	50.38	67.89	13.98	18.40	17.51	49.89
Two ERUs	Free	18	10.20	27.68	0.00	10.20	17.48	27.68
		36	27.45	44.96	0.58	8.86	17.51	26.96
	Real	18	14.31	31.79	0.00	14.31	17.48	31.79
		36	31.66	49.17	2.07	11.59	17.51	31.17
Three ERUs	Free	18	10.20	27.68	0.00	10.20	17.48	27.68
		36	26.20	43.71	0.73	7.47	17.51	25.71
	Real	18	14.31	31.79	0.00	14.31	17.48	31.79
		36	25.83	43.34	0.71	7.12	17.51	25.34

Figure 9.2 shows the optimal solutions for each scenario based on the travel time with real traffic condition (three ERU vehicles). We have considered response delay and clearance time compared to previous study. Response delay and clearance time take a larger portion (72.1%) of incident management process compared to

travel time only (27.9%). Our model further saves potential response delay because we have the assumption that ERUs stay as the current incident site instead of returning back to their originally assigned locations. If we add the return travel-time, the total delay time will increase with more response time to serve the next incident.

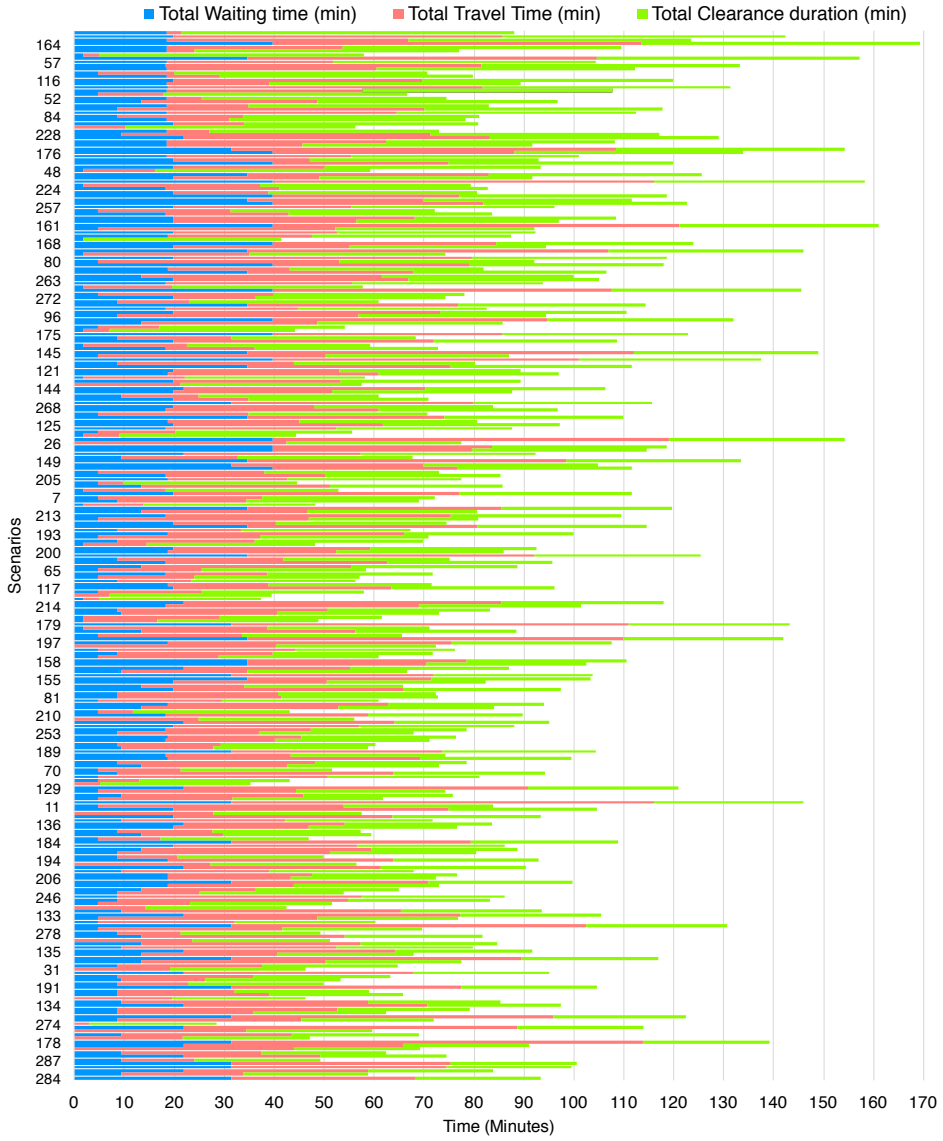


Figure 9.2: Optimal solutions for each scenario

The test problems are designed to evidence the significant effect of efficient allocation in the problems. Generally, the optimality gap drops as the number of response units is increased from two. If we have a deterministic solution based on expected value, the model will underestimate or overestimate the solution in different scenarios due to lack of flexibility. The scenario-based solution, on the other hand, generally provides a better estimate of the objective function. The quality of solutions is highly dependent on the scenarios, from worst-quality solutions to best solutions.

To gain further insight into the behavior of the model, we compared solutions with different the number of response units (Table 9.5). The response delay drops from 81.68 min to 57.13 min as the number of response units increases from one to three. This is because adding response units in the system becomes more effective in reducing response delay. If a given solution satisfies a threshold of response time for the overall system, we can save on operational cost under a budget limit.

Table 9.5: Assigned Locations and Performance

ERU #	Total Expected Time (mins)				Gap	Optimal locations
	Travel	Wait	Clear	Total		
1	32.71	13.98	34.99	81.68	0%	11
2	25.90	2.07	34.99	62.96	0.69%	11,11
3	21.42	0.71	34.99	57.13	13.06%	11,11,27

9.4.2 Discussions

We design a different experiment setup to compare the performance of the proposed model against the heuristic (scenario reduction). We have different combination of parameters such as nodes I, stages R, and number of ERUs U. Table 9.6 shows the result of computation time (s) and gap (%) for each case (No.). We reported the performance of the proposed model depending on the available time for execution of the model. We stop further execution after 1400s or less than 20% gap of the model and report the best found solution up to that point. The main reason is that the first feasible solution is usually found very fast (generally in less than 60 seconds). Most of the running time of the model is devoted to proving that a solution is optimal and only a small fraction of the running time is devoted to finding better feasible solutions that only marginally improve the previous best found solution.

As we have larger network size and more future stages, it is more time consuming. In this study, we use the heuristic method (fast forward selection) and the measure of the optimality gap to justify the quality of the solution. The optimality gap jumps as the network size increases from 17 to 34, and as we increase total stages from 2 to 3. Note that even the first case is very complex with 32042 variables. For larger scale cases (No. 5, 6, 8, 9), the heuristic method reaches a solution with less than 20% gap within 60s that can be used in real-time. These cases have fewer iterations as a result of the convergence. On the contrary, instead of quick solution, the proposed approach finds the solution with less gap compared to the

heuristic solution.

Table 9.6: Computational Performances for the Proposed Approach

No.	Parameters			Proposed approach		Fast forward selection	
	N	R	U	CPUtime	Gap	CPUtime	Gap
1	17	2	1	0.1s	0.00%	-	-
2	17	2	2	17.2s	0.92%	-	-
3	17	2	3	22.5s	19.19%	-	-
4	17	3	1	2.6s	0.00%	-	-
5	17	3	2	1400s	20.08%	8.3s	15.81 %
6	17	3	3	1400s	32.56%	54.9s	18.59 %
7	34	3	1	6.5s	0.00%	-	-
8	34	3	2	1400s	29.09%	54.2s	19.13%
9	34	3	3	1400s	35.89%	59.6s	19.91%

The presented mathematical model can be applied to real-time problems. The operator communicates with responders at each incident site by receiving messages or keeping track of ERU$'$ locations. Notifications can include available ERUs, travel time, probabilities of primary and secondary incidents at different node of the network.

The time to respond to an incident is relatively small compared to the time necessary to clear the incident. We dynamically incorporate position of ERUs in each stage, and this formulation causes a high complexity. In a planning stage, before an incident occurs, we can run the full model without a restriction of computational times. In an operational stage, after an incident happens at a node, a vehicle is dispatched to serve that incident based on the planning stage decision. After certain time intervals, number of available vehicles and the second stage scenarios are updated. We re-run our mathematical model to relocate the remaining vehicles

to be more prepared for future incidents. Upon the clearance of the incident, the ERU which was serving the incident is once more added to the pool of available fleet and therefore we need to re-run the model one more time based on the updated parameters.

As we face later stages, the computational burdens are reduced. However, running the model iteratively is still more practical with reasonable solution times. One possible way to reduce the running time of the model for real time applications is decreasing the size of the problem. This can be done either by reducing the number of scenarios or analogously reducing the number of future stages being considered at each time we run the model. Another approach is to accept non-optimal good enough solutions by running the model as long as we are allowed. After the time limit is met, we can report the solution and relocate the ERU vehicles accordingly.

9.5 Conclusions

In this research, we present an analytical approach for ERUs location-allocation to protect the safety of victims, travelers, and emergency personnel. Generally, traffic operators have underestimated the impact of secondary incidents due to their low frequency. Our model represents two main phases. The first one is a location phase solved by a facility location problem that allocates ERUs to respond to primary incidents. The second phase is an allocation phase that deals with a series of stages based on secondary incidents scenarios.

After an incident occurs, clearance activities cause vehicles approaching from

upstream to reduce their speeds, and emergency units responding to a secondary incident site take longer to respond. Determination of the best solution without considering stochastic nature of incidents has limitation in coping with uncertainty, and it might produce practically infeasible solutions. This study proposed an advanced strategy for distributing incident response units by solving a stochastic programming problem. As we demonstrate in a case study, the proposed framework can be useful for reducing delay time caused by response to secondary incidents occurring under impact of primary incidents. We approach the problem from a long-term perspective that the flexible location of ERUs can be changed and is not fixed.

Chapter 10: Overall Conclusions and Directions for Future Research

This dissertation is innovative and important in mitigating non-recurring congestion on freeways. Instead of a limited assumption that we know nothing about the future, we take advantage of the prediction of future events (i.e., secondary crashes) in making dispatching and relocation decisions. The identification, prediction, interpretation of secondary crashes enables us to present scenarios in a matrix form with expected probabilities. The matrix works as an input for stochastic and dynamic optimization models. Comparative analyses present differences between conventional models and the proposed model, and justify the importance of a secondary crash study.

Our key results are summarized below. These findings raise a number of issues which merit further investigation. Many of these issues can be investigated through further in-depth analyses of emerging data sources, micro-simulation software, and other optimization strategies. Having summarized our key findings below, we discuss possible avenues for future research to enhance our understanding of emergency response processes and thereby promote improved system.

10.1 Summary of Key Findings

- Bayesian structure equation model recognizes congestion patterns using INRIX Data and an adjustment of the boxplot captures queued segments.
- A principle for stochastic incident occurrences is developed with advanced machine learning models. The likelihood of classified secondary incidents is sequentially predicted. The principled Bayesian learning approach to neural networks outperforms the logistic model.
- A pedagogical rule extraction approach improves the ability to understand secondary incidents.
- A deterministic and a stochastic capacity estimation model quantify the impact caused by non-recurring congestion.
- Proposed online dispatching model outperforms other models by considering past and future emergencies to respond current emergency.
- Proposed stochastic location model relaxes the structural assumptions of Poisson process and overcomes limitation of previous studies.

10.2 Future Research Directions

10.2.1 Identification of Secondary Crashes

This dissertation focused on comparing robust secondary incident identifications to traditional static method, rather than performance evaluation of secondary

incident detection to dynamic method. However, future research can focus on a comparison against ASDA model [151], deterministic queue estimations, and simulation models. While current methods are based on vehicle probe data, deterministic queuing methods will be based data on vehicle arrivals and departures from loop detector.

10.2.2 Application for Prediction of Secondary Crashes

An emergency system evolves from one time-stage to another in such a way that chance elements are involved in progressing from one state to the next. We are extending the first-order semi-Markov model to include higher order features. When we see the time after a primary incident, the semi-Markov model can estimate the time to secondary incidents. There is a close relationship between incident duration and secondary incident occurrences. A second-order semi-Markov model will be developed to capture the time to secondary incident considering incident duration based on vehicle arrivals.

As shown in Figure 10.1, the symbolic description represents a series of decisions to assist emergency response personnel in decision-making. A user can simply insert the values for different parameters into a tree and obtain the results. Smartphone application ($e.g. WAZE$) can help driversnavigate aroundroad closuresand get where they need to be. If the likelihood of secondary incidents is high, notifications like watch out could make driving safer. Moreover, a connection weight approach accurately quantifies the contributions each variable makes from the neural network.

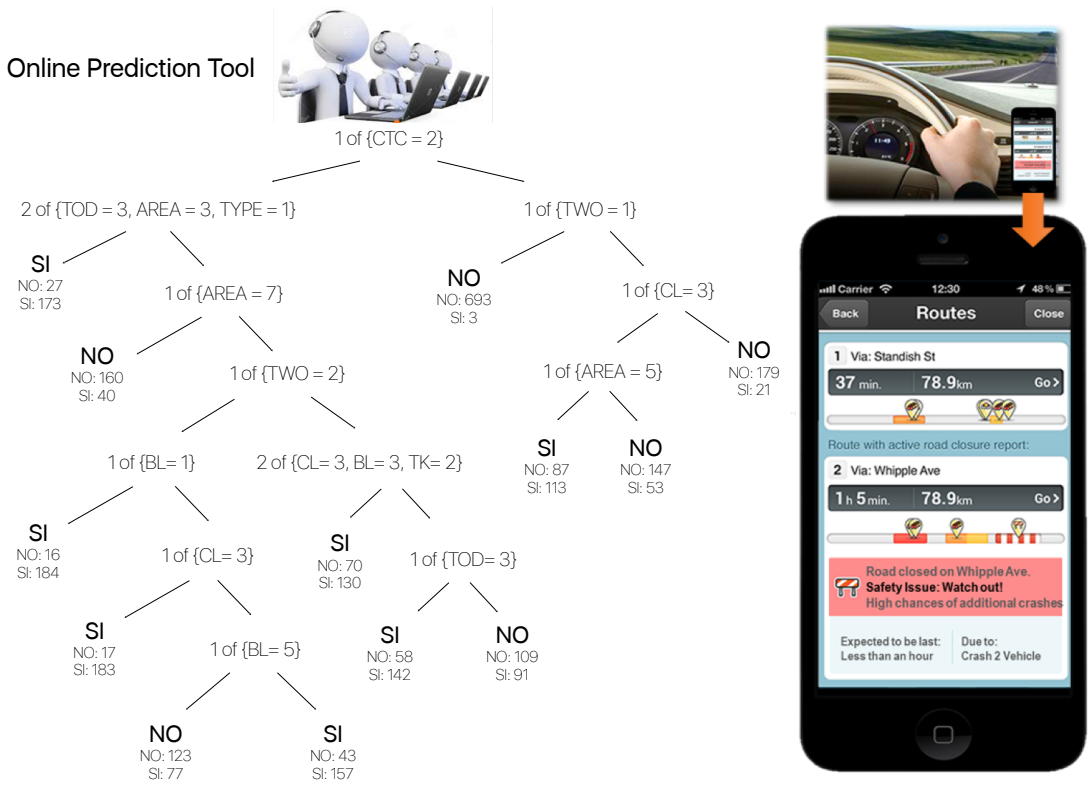


Figure 10.1: Application of incident online prediction tool

10.2.3 Capacity Adjustment for Estimation of Delay

Since the HCM capacity adjustment underestimates or overestimates impact of incidents, the proposed method can be implemented in incident management systems to calculate total incident delay. For traffic management of the network, traffic operators can control the system by regulating density. Occupancy is measured to detect congestion impact of incidents. However, it is a sample of traffic conditions that occur over a longer freeway segment. Its effectiveness of capacity drop depends on the locations of the loop detectors relative to the incident locations. Therefore, only secondary incidents occurring at the same segment of their primary incidents were considered. Advanced algorithms for measurement of densities over

extended-length freeway segment can be used to identify capacity drop regardless of the location of incident.

Furthermore, the current study focused on secondary incidents from large amount of field data at a single site. Since secondary incidents are rare, it was difficult to additionally find another site supporting capacity drop with enough observations. The proposed methodology can be applied to freeway segments where ample primary-secondary incident pairs are found. This will help interested readers to estimate incident-induced capacity.

10.2.4 Dispatching of Emergency Vehicles

In the future, we may decide to assign more than one vehicle to reduce expected clearance time. Reducing clearance time is also important, because the time to serve an incident is relatively large compared to the time to approach (response) the incident [108].

The introduced k-server problem has many applications in network modeling when we have a sequence of requests served by k-servers. For example, the k-server problem can be reduced to computing the minimal-cost maximal flow on a suitable constructed network [152]. Better competitive ratio can persuade dispatchers to use our algorithm. The proposed algorithm can be improved to accommodate asymmetry of emergency response service systems on arterial networks. However, complexity of the model will increase and the network will not have an advantage of using metric space. Game theoretical models such as Nash-like equilibrium [153]

can be a solution.

10.2.5 Relocation of Emergency Vehicles

Our results indicate that the expected waiting time omitted by previous studies can significantly impact the expected total delay compared to the relatively short travel time of response units. Allowing for flexibilities with secondary incidents decreases the expected total delay time compared to the solution without considering secondary incidents. As the number of available emergency response unit increases, shorter total delay is expected. Therefore, further assignment of ERUs that covers new locations occurs by using information about the most promising sites.

One of the challenges is generation of realistic incident scenarios. We can improve the model by allowing more than one vehicle routing for each stage. By investigating the structure of the transition probability of each stage, the scenario can be generalized and estimation method can be developed. The proposed model is executed in planning stage before occurrence of an incident. More efficient formulation can improve computation time and allow the use of the model in operation stage for dynamic scenarios.

We will use the capability for cars to communicate with one another for both travelers and emergency operators. This new data source improves the real-time traffic routing service as an input to the emergency vehicle location and dispatch model. The system will respond to transportation demand or emergencies in real-time by messaging and response between vehicles and dispatch.

References

[1] D. Schrank, B. Eisele, and T. Lomax. *Urban Mobility Report*. Texas A&M Transportation Institute, College Station, Texas., 2012.

[2] National Traffic Incident Management Coalition. Benefits of traffic incident management. 2007.

[3] H.N. Koutsopoulos and A. Yablonski. Design parameters of advanced information system: the case of incident congestion and small market penetration. In *Proceedings of the IEEE 2nd Vehicle Navigation Information Systems Conference, Dearborn, Michigan*, 1991.

[4] R. Halper and S. Raghavan. The mobile facility routing problem. *Transportation Science*, 45(3):413–434, 2011.

[5] H. Kim, W. Kim, G. L. Chang, and S. Rochon. Design of emergency response system to minimize incident impacts. *Transportation Research Record: Journal of the Transportation Research Board, No. 2470*, pages 65–77, 2014.

[6] F. Daneshgar, S. Mattingly, and A. Haghani. Evaluating beat structure and truck allocation for the tarrant county, texas, courtesy patrol. *Transportation Research Record: Journal of the Transportation Research Board, No. 2334*, pages 40–49, 2013.

[7] H. Park and A. Haghani. Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C: Emerging Technologies, Online Published*, 2015.

[8] H. Park, A. Shafahi, and A. Haghani. Stochastic emergency response units relocation considering secondary incidents. *Revision in the IEEE Transactions on Intelligent Transportation Systems*, 2015.

[9] M.G. Karlaftis, S.P. Latoski, N.J. Richards, and K.C. Sinha. Its impacts on safety and traffic management: an investigation of secondary crash causes. *Journal of Intelligent Transportation Systems: Technologies, Planning, and Operations. forthcoming*, 5(1):39–52, 1999.

[10] N. Owens, A. Armstrong, C. Mitchell, and R. Brewster. *Federal Highway Administration Focus States Initiative: Traffic Incident Management Performance Measures Final Report*. FHWA-HOP-10-010, Federal Highway Administration, U.S. Department of Transportation, Washington, DC, 2010.

[11] F.L. Hall and B.N. Persaud. Evaluation of speed estimates made with single-detector data from freeway traffic management systems. *Transportation Research Record: Journal of the Transportation Research Board, No. 1232*, pages 9–16, 1989.

[12] Y. Lao, G. Zhang, J. Corey, and Y. Wang. Gaussian mixture model-based speed estimation and vehicle classification using single-loop measurements. *Journal of Intelligent Transportation Systems: Technologies, Planning, and Operations. forthcoming*, 16(4):184–196, 2012.

[13] S. Washington, M.G Karlaftis, and F. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis*. Chapman and Hall/CRC Press., 2003.

[14] A. Haghani, M. Hamedi, K. Sadabadi, S. Young, and P. Tarnoff. Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record: Journal of the Transportation Research Board, No. 2160*, pages 60–68, 2010.

[15] M.L. Pack. Wide-area congestion performance monitoring using probe data. *The 92nd Annual Meeting of Transportation Research Board (CD-ROM), Washington, D.C.*, 2013.

[16] H. Park and A. Haghani. Optimal number and location of bluetooth sensors considering stochastic travel time prediction. *Transportation Research Part C: Emerging Technologies*, 55:203–216, 2015.

[17] H. Park, A. Haghani, and X. Zhang. Interpretation of bayesian neural network for predicting the duration of detected incidents. *Journal of Intelligent Transportation Systems: Technologies, Planning, and Operations. Online Published*, 2015.

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

[19] H. Park, A. Haghani, and Y. A. A pedagogical rule extraction from bayesian neural networks for prediction of secondary incidents. *The 1st International Conference on Engineering and Applied Sciences Optimization, Kos Island, Greece*, 2014.

[20] A. Haghani, D. Iliescu, M. Hamedi, and S. Yong. Methodology for quantifying the cost effectiveness of freeway service patrols programs, case study. *H.E.L.P.*

Program, I-95 corridor coalition, University of Maryland, College Park, MD., 2006.

[21] S. Tang and F. Y. Wang. A pci-based evaluation method for level of services for traffic operational systems. *IEEE Transactions on Intelligent Transportation Systems*, 4(7):494–499, 2006.

[22] H. Park and A. Haghani. Capacity adjustment considering the impact of secondary incidents. *Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C.*, 2015.

[23] HCM. *Highway Capacity Manual 2010*. Transportation Research Board, Washington, D.C., 2010.

[24] M. W. Ng, A. J. Khattak, and W. K. Talley. Modeling the time to the next primary and secondary incident: A semi-markov stochastic process approach. *Transportation Research Part B: Methodological*, 58:44–57, 2013.

[25] L. Fu and L.R. Rilett. Real-time estimation of incident delay in dynamic and stochastic networks. *Transportation Research Record: Journal of the Transportation Research Board, No. 1603*, pages 99–105, 1997.

[26] H. Park and A. Haghani. Stochastic capacity adjustment considering secondary incidents. *Revision in the IEEE Transactions on Intelligent Transportation Systems*, 2015.

[27] C. Lei, W.H. Lin, and L. Miao. A stochastic emergency vehicle redepolyment model for an effective response to traffic incidents. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):898–909, 2015.

[28] H. Park and A. Haghani. Online emergency vehicle dispatching with look-ahead on a transportation network. *Presentation at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C.*, 2016.

[29] A. Haghani, Q. Tian, and H. Hu. Simulation model for real-time emergency vehicle dispatching and routing. *Transportation Research Record: Journal of the Transportation Research Board, No.1882*, pages 176–183, 2004.

[30] S.L. Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, 12(3):450–459, 1964.

[31] H. Park and A. Haghani. An optimal fleet allocation of emergency response teams on freeway using a two stage stochastic programming. *The 20th Conference of the International Federation of Operational Research Societies, Barcelona, Spain*, 2014.

[32] H. Park and A. Haghani. Quantifying non-recurrent congestion impact on secondary incidents using probe vehicle data. *The 54th Annual Transportation Research Forum, Annapolis, MD*, 2013.

[33] H. Park, A. Haghani, and H. Masoud. Real-time filtering of vehicle probe data for secondary incident prediction. *The 8th Triennial Symposium on Transportation Analysis, San Pedro de Atacama, Chile*, 2013.

[34] R. Raub. Occurrence of secondary crashes on urban arterial roadways. *Transportation Research Record: Journal of the Transportation Research Board, No. 1581*, pages 53–58, 1997.

[35] J.E. Moore, G. Giuliano, and S. Cho. Secondary accident rates on los angeles freeways. *Journal of Transportation Engineering*, 130(3):280–285, 2004.

[36] W. Hirunyanitiwattana and S. Mattingly. Identifying secondary crash characteristics for the california highway system. *The 85th Annual Meeting of Transportation Research Board (CD-ROM), Washington, D.C.*, 2006.

[37] C. Zhan, L. Shen, M.A. Hadi, and A. Gan. Understanding the characteristics of secondary crashes on freeways. *The 87th Annual Meeting of Transportation Research Board (CD-ROM), Washington, D.C.*, 2008.

[38] A.J. Khattak, X. Wang, and H. Zhang. Incident management integration tool: Dynamically predicting incident durations, secondary incident occurrence and incident delays. *IET Intelligent Transportation Systems*, 6(2):204–214, 2012.

[39] E. Vlahogianni, M. Karlaftis, J. Golias, and B. Halkias. Freeway operations, spatiotemporal-incident characteristics, and secondary-crash occurrence. *Transportation Research Record: Journal of the Transportation Research Board, No. 2178*, pages 1–9, 2010.

[40] C. Chen, A. Skabardonis, and P. Varaiya. Systematic identification of freeway bottlenecks. *Transportation Research Record: Journal of the Transportation Research Board, No. 1867*, pages 46–52, 2004.

[41] H. Li and R.L. Bertini. Comparison of algorithms for systematic tracking of patterns of traffic congestion on freeways in portland, oregon. *Transportation Research Record: Journal of the Transportation Research Board, No. 2178*, pages 101–110, 2010.

[42] M. Imprialou, F. Orfanou, E. Vlahogianni, and M. Karlaftis. Methods for defining spatiotemporal influence areas and secondary incident detection in freeways. *Journal of Transportation Engineering*, 140(1):70–80, 2014.

[43] M. Schonhof and D. Helbing. Criticism of three-phase traffic theory. *Transportation Research Part B: Methodological*, 43:784–797, 2009.

[44] C. Antoniou, H. N. Koutsopoulos, and G. Yannis. Dynamic data-driven local traffic state estimation and prediction. *Transportation Research Part C: Emerging Technologies*, 34:89–107, 2013.

[45] C. Sun and V. Chilukuri. Dynamic incident progression curve for classifying secondary traffic crashes. *Journal of Transportation Engineering*, 136(12):1153–1158, 2010.

[46] S. Yang. On feature selection for traffic congestion prediction. *Transportation Research Part C: Emerging Technologies*, 26:160–169, 2013.

[47] A. Garib, A. Radwan, and H. Al-Deek. Estimating magnitude and duration of incident delays. *Journal of Transportation Engineering*, 123(6):459–466, 1997.

[48] K. Ozbay and P. Kachroo. *Incident management in Intelligent Transportation Systems*. Artech House, Boston, MA, 1999.

[49] K. Smith and B Smith. *Forecasting the Clearance Time of Freeway Accidents*. Publication STL-2001-012, Center for Transportation Studies, University of Virginia, 2002.

[50] B. Yang, X. Zhang, and L. Sun. Traffic incident duration prediction based on the bayesian decision tree method. *Transportation and Development Innovative Best Practices, American Society of Civil Engineers*, pages 338–343, 2008.

[51] Q. He, Y. Kamarianakis, K. Jintanakul, and L. Wynter. *Incident Duration Prediction with Hybrid Tree-based Quantile Regression*, volume 2. Advances in Dynamic Network Modeling in Complex Transportation Systems, Springer New York, 2013.

[52] W. Wu, S. Chen, and C. Zheng. *Traffic Incident Duration Prediction Based on Support Vector Regression*. American Society of Civil Engineers, 2011.

[53] T.F. Golob, W.W. Recker, and J.D. Leonard. An analysis of the severity and incident duration of truck-involved freeway accidents. *Accident Analysis and Prevention*, 19(5):375–395, 1987.

[54] G. Giuliano. Incident characteristics, frequency, and duration on a high volume urban freeway. *Transportation Research Part A: General*, 23(5):387–396, 1989.

[55] A. Skabardonis, K.F. Petty, and P.P. Varaiya. Los angeles i-10 field experiment: Incident patterns. *Transportation Research Record: Journal of the Transportation Research Board, No. 1683*, pages 22–30, 1999.

[56] Y. Chung and B. J. Yoon. Analytical method to estimate accident duration using archived speed profile and its statistical analysis. *KSCE Journal of Civil Engineering, Korean Society of Civil Engineers*, 16(6):1064–1070, 2012.

[57] K. Ozbay and N. Noyan. Estimation of incident clearance times using bayesian networks approach. *Accident Analysis and Prevention*, 38(3):542–555, 2006.

[58] S. Boyles, D. Fajardo, and S.T. Waller. A naïve bayesian classifier for incident duration prediction. *Proceedings of the 86rd TRB Annual Meeting (CD-ROM). Washington, DC*, 2007.

[59] W. Kim and G. L. Chang. Development of a hybrid prediction model for freeway incident duration: A case study in maryland. *International Journal of Intelligent Transportation Systems Research*, 10(1):22–33, 2012.

[60] Zou N. Lin, P.W. and G.L. Chang. Integration of a discrete choice model and a rule-based system for estimation of incident duration: A case study in maryland. *The 83rd Annual Meeting of the Transportation Research Board, Washington, D.C.*, 2004.

[61] B. Jones, L. Janssen, and F. Mannering. Analysis of the frequency and duration of freeway accidents in seattle. *Accident Analysis and Prevention*, 23(4):239–255, 1991.

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

[63] A. Stathopoulos and M. Karlaftis. Modeling duration of urban traffic congestion. *Journal of Transportation Engineering*, 128(6):587–590, 2002.

[64] Y. Qi and B. Smith. Identifying nearest neighbors in a large-scale incident data archive. *Transportation Research Record: Journal of the Transportation Research Board, No.1879*, pages 89–98, 2004.

[65] Y. Chung. Development of an accident duration prediction model on the korean freeway systems. *Accident Analysis and Prevention*, 42(1):282–289, 2010.

[66] A.T. Hojati, L. Ferreira, S. Washington, and P. Charles. Hazard based models for freeway traffic incident duration. *Accident Analysis and Prevention*, 52:171–181, 2013.

[67] H.J. Kim and H.K. Choi. A comparative analysis of incident service time on urban freeways. *International Association of Traffic and Safety*, 25(1):62–72, 2001.

[68] N. Gorjian, L. Ma, M. Mittinty, P. Yarlagadda, and Y. Sun. *A Review on Reliability Models with Covariates*. Engineering Asset Lifecycle Management, Springer London, 2010.

[69] W. Wang and J. Paliwal. Generalisation performance of artificial neural networks for near infrared spectral analysis. *Biosystems Engineering*, 94(1):7–18, 2006.

[70] L. Guan, W. Liu, X. Yin, and L. Zhang. Traffic incident duration prediction based on artificial neural network. *Intelligent Computation Technology and Automation*, 3:1076–1079, 2010.

[71] G. Valenti, M. Lelli, and D. Cucina. A comparative study of models for the incident duration prediction. *European Transport Research Review*, 2(2):103–111, 2010.

[72] A.J. Khattak, J.L. Schofer, and M.H. Wang. A simple time sequential procedure for predicting freeway incident duration. *I V H S Journal*, 2(2):113–138, 2015/08/06 1995.

[73] C.H. Wei and Y. Lee. Sequential forecast of incident duration using artificial neural network models. *Accident Analysis & Prevention*, 39(5):944–954, 2007.

[74] F.C. Pereira, F. Rodrigues, and M. Ben-Akiva. Text analysis in incident duration prediction. *Transportation Research Part C: Emerging Technologies*, 37:177–192, 12 2013.

[75] S.P. Miaou, A. Lu, and H. Lum. Pitfalls of using r2 to evaluate goodness of fit of accident prediction models. *Transportation Research Record: Journal of the Transportation Research Board, No. 1542*, pages 6–13, 1996.

[76] A. Khattak, X. Wang, and H. Zhang. Incident management integration tool: Dynamically predicting incident durations, secondary incident occurrence and incident delays. *Intelligent Transport Systems, IET*, 6(2):204–214, 2012.

[77] M.G. Karlaftis and E.I. Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399, 2011.

[78] W.L. Buntine and A.S. Weigend. Bayesian back-propagation. *Complex systems*, 5(6):603–643, 1991.

[79] D.J.C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[80] D. Barber. *Bayesian methods for supervised neural networks, Handbook of Brain Theory and Neural Networks*. MIT Press, Cambridge, MA, 2002.

[81] J. Lampinen and A. Vethari. Bayesian approach for neural networks - review and case studies. *Neural Networks*, 14(3):7–24, 2001.

[82] H.K.H. Lee. *Bayesian Nonparametrics via Neural Networks*. ASA-SIAM Series on Statistics and Applied Mathematics, 2004.

[83] NETLAB Algorithms for Pattern Recognition. *Nabney, I.T.* Springer, New York, 2004.

[84] Y. Xie, D. Lord, and Y. Zhang. Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention*, 39(5):922–933, 2007.

[85] E. Castillo, J. Menéndez, and S. Sánchez-Cambronero. Traffic estimation and optimal counting location without path enumeration using bayesian networks. *Computer-Aided Civil and Infrastructure Engineering*, 23(3):189–207, 2008.

[86] H. Adeli and H. S. Park. Optimization of space structures by neural dynamics. *Neural Networks*, 8(5):769–781, 1995.

[87] S. Arangio and F. Bontempi. Soft computing based multilevel strategy for bridge integrity monitoring. *Computer-Aided Civil and Infrastructure Engineering*, 25(5):348–362, 2010.

[88] T. Dietterich, M. Kearns, and Y. Mansour. Applying the weak learning framework to understand and improve c4.5. *In Proceedings of the Thirteenth International Conference on Machine Learning*, pages 96–104, 1996.

[89] M.W. Craven and J.W. Shavlik. Understanding time-series networks: A case study in rule extraction. *International Journal of Neural Systems*, 4:373–384, 1997.

[90] J.W. Johnson. A heuristic method for estimating the relative weight of predictor variables in multiple regression. *Multivariate Behavioral Research*, 35(1):1–19, 2000.

[91] I. Dimopoulos, J. Chronopoulos, A. Chronopoulou-Sereli, and S. Lek. Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in athens city (greece). *Ecological Modelling*, 120(2–3):157–165, 1999.

[92] G.D. Garson. Interpreting neural-network connection weights. *International Journal of Artificial Intelligence and Expert Systems*, 6(4):46–51, 1991.

[93] A. T. C. Goh. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering*, 9(3):143–151, 1995.

[94] M. Scardi and L.W. Harding Jr. Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecological Modelling*, 120(2–3):213–223, 1999.

[95] S. Lek, A. Belaud, P. Baran, I. Dimopoulos, and M. Delacoste. Role of some environmental variables in trout abundance models using neural networks. *Aquatic Living Resources*, 9(1):23–29, 1996.

[96] H.R. Maier and G.C. Dandy. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environmental Modelling and Software*, 15(1):101–124, 2000.

[97] J.D. Olden, M.K. Joy, and R.G. Death. An accurate comparison of methods for quantifying variable importance in artificial neural networksu simulated data. *Ecological Modelling*, 178(3–4):389–397, 2004.

[98] V.L. Knoop, S.P. Hoogendoorn, and H.J. van Zuylen. Capacity reduction at incidents: Empirical data collected from a helicopter. *Transportation Research Record: Journal of the Transportation Research Board, No. 2071*, pages 19–25, 2008.

[99] V. Sisiopiku, X. Li, K. Mouskos, C. Kamga, C. Barrett, and A. Abro. Dynamic traffic assignment modeling for incident management. *Transportation Research Record: Journal of the Transportation Research Board, No. 1994*, (1):110–116, 2007.

[100] M. Menendez and C. Daganzo. Assessment of the impact of incidents near bottlenecks: Strategies to reduce delays. *Transportation Research Record: Journal of the Transportation Research Board, No. 1867*, pages 53–59, 2007.

[101] A. Khattak, X. Wang, and H. Zhang. Are incident durations and secondary incidents interdependent? *Transportation Research Record: Journal of the Transportation Research Board, No. 2099*, pages 39–49, 2009.

[102] C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The location of emergency service facilities. *Operations Research*, 19(6):1363–1373, 1971.

[103] M.S. Daskin. A maximum expected covering location model: Formulation, properties and heuristic solution. *Transportation Science*, 17(1):48–70, 1983.

[104] P.B. Mirchandani and A. R. Odoni. Locations of medians on stochastic networks. *Transportation Science*, 13(2):85–97, 1979.

[105] R. Nair and E. Miller-Hooks. Evaluation of relocation strategies for emergency medical service vehicles. *Transportation Research Record: Journal of the Transportation Research Board, No. 2137*, pages 63–73, 2009.

[106] M. Gendreau, G. Laporte, and F. Semet. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. *Parallel Computing*, 27(12):1641–1653, 2001.

[107] E.S. Savas. Simulation and cost-effectiveness analysis of new york's emergency ambulance service. *Management Science*, 15(12):608–627, 1969.

[108] R.C. Larson. A hypercube queuing model for facility location and redistricting in urban emergency services. *Computers and Operations Research*, 1(1):67–95, 1974.

[109] M.L. Brandeau. Extending and applying the hypercube queueing model to deploy ambulances in boston. *Delivery of urban services: with a view towards applications in management science and operations research*, pages 121–153, 1986.

[110] M.S. Daskin and A. Haghani. Multiple vehicle routing and dispatching to an emergency scene. *Environment and Planning A*, 16(10):1349–1359, 1984.

[111] S. Yang, M. Hamedi, and A. Haghani. Online dispatching and routing model for emergency vehicles with area coverage constraints. *Transportation Research Record: Journal of the Transportation Research Board, No. 1923*, pages 1–8, 2005.

[112] L.A. McLay and M.E. Mayorga. A model for optimally dispatching ambulances to emergency calls with classification errors in patient priorities. *IIE Transactions*, 45(1):1–24, 2012.

[113] M.E. Mayorga, D. Bandara, and L.A. McLay. Districting and dispatching policies for emergency medical service systems to improve patient survival. *IIE Transactions on Healthcare Systems Engineering*, 3(1):39–56, 2013.

[114] S. Albers. Online algorithms: A survey. *Mathematical Programming*, 97(1-2):3–26, 2003.

[115] R.L. Church and C.S Revelle. The maximal covering location problem. *Regional Science Association*, 32(1):101–118, 1974.

[116] A. Suzuki and Drezner Z. The p-center location problem in an area. *Location Science*, 4:69–82, 1996.

[117] C. Revelle and K. Hogan. The maximum reliability location problem and -reliablep-center problem: Derivatives of the probabilistic location set covering problem. *Annals of Operations Research*, 18(1):155–173, 1989.

[118] M.O. Ball and L.F. Lin. A reliability model applied to emergency service vehicle location. *Operations Research*, 41(1):18–36, 1993.

[119] N. Geroliminis, M.G. Karlaftis, and A. Skabardonis. A spatial queuing model for the emergency vehicle districting and location problem. *Transportation Research Part B: Methodological*, 43(7):798–811, 2009.

[120] M. Gendreau, G. Laporte, and F. Semet. Solving an ambulance location model by tabu search. *Location Science*, 5(2):75–88, 1997.

[121] O. Berman. Repositioning of distinguishable urban service units on networks. *Computers and Operations Research*, 8(2):105–118, 1981.

[122] O. Berman. Dynamic repositioning of indistinguishable service units on transportation networks. *Transportation Science*, 15(2):115–136, 05 1981.

[123] T. Andersson and P. Varbrand. Decision support tools for ambulance dispatch and relocation. *Operation Ressearch Society*, 58(2):195–201, 2006.

[124] R. Alanis, A. Ingolfsson, and B. Kolfal. A markov chain model for an ems system with repositioning. *Production and Operations Management*, 22(1):216–231, 2013.

[125] L. Zhang. Optimisation of small-scale ambulance move-up. In *Proceedings of the 45th Annual Conference of the New Zealand Operation Research Society*, 2010.

[126] C. Prodhon and C. Prins. A survey of recent research on location-routing problems. *European Journal of Operation Research*, 238:1–17, 2014.

[127] P. Marchesini and W. Weijermars. *The Relationship Between Road Safety and Congestion on Motorways. A Literature Review of Potential Effects, R-2010-12 Leidschendam.* R-2010-12 Leidschendam. SWOV Institute for Road Safety Research, The Netherlands., 2010.

[128] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

[129] J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

[130] D.C. Hoaglin, F. Mosteller, and J.W. Tukey. *Understanding Robust and Exploratory Data Analysis*. Wiley, New York, 1983.

[131] M. Hubert and E. Vandervieren. An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52:5186–5201, 2008.

[132] R. M. Neal. *MCMC using Hamiltonian dynamics*. Handbook of Markov Chain Monte Carlo, Chapman and Hall, CRC Press, 2011.

[133] *Portland Fire and Rescue: Emergency Response Time Goal not Met, Though Pf and R Strives for Excellence*. Office of the City Auditor, Portland, OR., 2010.

[134] Bureau of Labor Statistics. *Census of Fatal Occupational Injuries*. U.S. Department of Labor, Washington, D.C., 2003.

[135] L. Hou, Y. Lao, Y. Wang, Z. Zhang, Y. Zhang, and Z. Li. Modeling freeway incident response time: A mechanism-based approach. *Transportation Research Part C: Emerging Technologies*, 28:87–100, 2013.

[136] H.J. Payne and S.C. Tignor. Freeway incident-detection algorithms based on decision trees with states. *Transportation Research Record: Journal of the Transportation Research Board, No. 682*, pages 30–37, 1978.

[137] Armstrong A. Sullivan P. Mitchell C. Newton N. Brewster R. Owens, N. and T Trego. Traffic incident management handbook. *FHWA-HOP-10-013, Federal Highway Administration, Office of Transportation Operations, U.S. Department of Transportation, Washington, D.C.*, 2010.

[138] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[139] C.C. Chang and C.J. Lin. Libsvm: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[140] C. Marzban and A. Witt. A bayesian neural network for severe-hail size prediction. *Weather and Forecasting*, 16(5):600–610, 2001.

[141] E.I. Vlahogianni, M.G. Karlaftis, and F.P. Orfanou. Modeling the effects of weather and traffic on the risk of secondary incidents. *Journal of Intelligent Transportation Systems: Technologies, Planning, and Operations.*, 16(3):109–117, 2012.

[142] R.M. French. Catastrophic forgetting in connectionist networks: Causes, consequences and solutions. *Trends in Cognitive Science*, 3(4):128–35, 1999.

[143] J.H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis - Nonlinear methods and data mining*, 38(4):367–378, 2002.

[144] J. Li, C.J. Lan, and X. Gu. Estimation of incident delay and its uncertainty on freeway networks. *Transportation Research Record: Journal of the Transportation Research Board, No. 1959*, (1):37–45, 2006.

[145] Petrov V. *Sum of Independent Random Variables*. Springer, Berlin, 1975.

[146] M.S. Manasse, L.A. McGeoch, and D.D. Sleator. Competitive algorithms for server problems. *Journal of Algorithms*, 11(2):208–230, 1990.

[147] N. Bansal, N. Buchbinder, and J. Naor. Metrical task systems and the k-server problem on hsts. In *Proceedings of the 37th International Colloquium Conference on Automata, Languages and Programming*, ICALP'10, pages 287–298, Berlin, Heidelberg, 2010. Springer-Verlag.

[148] E. Koutsoupias and C.H. Papadimitriou. Beyond competitive analysis. *Society for Industrial and Applied Mathematics Journal on Computing*, 30(1):300–317, 2000.

[149] G.P. McCormick. Computability of global solutions to factorable nonconvex programs: Part i - convex underestimating problems. *Mathematical Programming*, 10(1):147–175, 1976.

[150] H. Heitsch and W. Römisch. Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications*, 24(2-3):187–206, 2003.

[151] B.S. Kerner, H. Rehborn, M. Aleksic, and A. Haug. Recognition and tracing of spatial-temporal congested traffic patterns on freeways. *Transportation Research Part C: Emerging Technologies*, 12(5):369–400, 2004.

[152] M. Chrobak and L. L. Larmore. An optimal on-line algorithm for k-servers on trees. *Society for Industrial and Applied Mathematics Journal on Computing*, 20(1):144–148, 1991.

[153] K. Han. *An Analytical Approach to Sustainable Transportation Network Design*. PhD dissertation, Pennsylvania State University, 2013.