UNIVERSITY OF MARYLAND

National Transportation Center

Project ID: NTC2016-??-?-?

STOCHASTIC DEPLOYMENT OF EMERGENCY VEHICLES CONSIDERING SEQUENCE OF INCIDENTS

Final Report

by

Principal Investigator Name: Ali Haghani
haghani@umd.edu

Hyoshin Park
University of Maryland

for

National Transportation Center at Maryland (NTC@Maryland)
1124 Glenn Martin Hall
University of Maryland
College Park, MD 20742

May, 2016

ACKNOWLEDGEMENTS

This project was funded by the National Transportation Center @ Maryland (NTC@Maryland), one of the five National Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT).

DISCLAIMER

The contents of this report reflect the views of the authors, who are solely responsible for the facts and the accuracy of the material and information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation University Transportation Centers Program in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The contents do not necessarily reflect the official views of the U.S. Government. This report does not constitute a standard, specification, or regulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

EXCUTIVE SUMMARY

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 that events over a time interval are independent. This study follows an assumption 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 study is a pioneer in relaxing the structural assumptions of independency 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 study 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 rule to original depot, 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.

1.0 INTRODUCTION

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 (Schrank et al. 2012). 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 (National Traffic Incident Management Coalition 2007). 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 (Koutsopoulos and Yablonski 1991). One way to enhance performance is applying a mobile facility concept (Halper and Raghavan 2011), 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.

1.1 PROBLEM STATEMENT

This research 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%.

$$\boldsymbol{\alpha} = \begin{Bmatrix} (\mathbf{2},\mathbf{3}) & \mathbf{1} & \mathbf{2} \\ (\mathbf{2},\mathbf{3}) & \mathbf{2} & \mathbf{1} \end{Bmatrix} \tag{1}$$

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 (Park et al. 2013a; Park and Haghani 2015a). 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 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 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 (Park and Haghani. 2014). Therefore, potential effects of secondary incidents on emergency response system have been overlooked.

1.2 RESEARCH OBJECTIVES

In this paper, the statistical properties of future sequence of incidents are considered in generating scenarios. We lookahead interdependent location-allocation of ERUs by taking stochastic information of future incidents explicitly into account. Stochastic programming hedges well against a wide range of scenarios in which probabilities of a sequence of incidents are assigned. Importantly in stochastic programming, location decisions should be made before the occurrence of a next incident. Even in the case of similar primary incidents, different candidate secondary incident locations are expected to have different delay times. This problem fits well into the framework of stochastic programming, which includes uncertainty in primary and secondary incident occurrences. Stochastic optimization will provide a prompt response to incidents and will play a crucial role in reducing delay, fuel consumption, and pollutant emission rates.

1.3 REPORT OUTLINE

The rest of this report is organized as follows. The rest of the paper is organized as follows. We review relevant previous studies about the emergency facility location problem in the next section. In Section 3, the structure of incident process is introduced to generate scenarios. Section 4 shows the proposed formulation and linearization. Section 5 presents numerical examples and sensitivity analysis. Conclusion and future researches are discussed in Section 6.

2.0 LITERATURE REVIEW

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.

2.1 DETERMINISTIC MODELS

The earlier versions of deterministic model are covering theories, such as location set-covering problems (Toregas et al. 1971). It provides 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 (Church and Revelle 1974). This model was extended to account for the chance when a demand arrives at the system that is engaged to serve other demands (Daskin 1983). 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 (Suzuki and Drezner 1996).

2.2 PROBABILISTIC FORMULATIONS

On the other hand, 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 (Hakimi 1964). One can use the maximum availability location problem (Revelle and Hogan 1989). An upper bound was imposed on the probability that a call on demand point does not receive immediate service (Ball and Lin 1993). To incorporate the busy probability, queuing-based models consider customers waiting for service in congested systems (Larson 1974). A spatial queuing model considers spatial and temporal demand characteristics such as the probability that a server is not available when required (Geroliminis et al. 2009).

2.3 LOCATION MODELS

2.3.1 Static Methods

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 assumes that the probability of incident occurrences follows a Poisson distribution (Daneshgar et al. 2013). A single incident rate, assuming independency between two incidents, has been considered (Kim et al. 2014; Skabardonis et al. 1999). 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 (Ng et al. 2013). Crash risk is higher in the presence of an earlier crash (Park and Haghani 2013, 2015a). Incidents frequently cause unexpected delay due to larger traffic demand than reduced capacity (Park et al. 2016). After a primary incident occurs, bottleneck quickly forms 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, 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.

2.3.2 Dynamic Methods

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 (Haghani et al. 2004). Solutions are computed one-by-one in an online fashion, while minimizing the response time of emergency vehicles (Yang et al. 2005). Dynamic double standard models incorporate practical dimensions addressing the dynamic nature of the problem (Gendreau et al. 1997). The real-time relocation models take service coverage concern when ERUs are dispatched (Nair and Miller-Hooks 2009). Dynamic relocation models pre-compute solutions in anticipation of events in the future stages (Gendreau et al. 2001). Recently, an interesting problem of determining stochastic emergency vehicle redeployment for an effective response to traffic incidents was introduced (Lei et al. 2015). 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 study, we estimate the number of available servers by comparing remaining time to clear the current incident and time to next incident occurrence.

2.3.3 Markov Decision Processes

Alternatively, Markov Decision Processes (MDPs) were used on dynamic relocation of service units in early works (Berman 1981a, 1981b). A tree-search heuristic was applied for approaching optimal relocations to the Stockholm region in Sweden (Andersson and Varbrand 2006). A MDP approximates distribution of the response time and the number of busy ambulances to identify near-optimal compliance tables (Alanis et al. 2013). Recently, a look-ahead scheme was applied in ambulance locating models to approximate the temporally accrued rewards and discounted probabilities (Zhang 2010). 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 research, 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.

2.4 SUMMARY

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, Prodhon and Prins 2014). This research 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 taking significant portion of traffic delays, emergency responders' strategic concerns have been growing. Fortunately, scientific breakthroughs enabled us to develop thresholds as a consistent definition of secondary incidents (Park and Haghani 2015a, 2015b). This research 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 research, 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.

3.0 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 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 $E[\tau(i,r)]$ for each site (i = k) and stage (r = u) is a sum of $Pr^p{}_{(i,r)}$ and $Pr^s{}_{(k,u)}$.

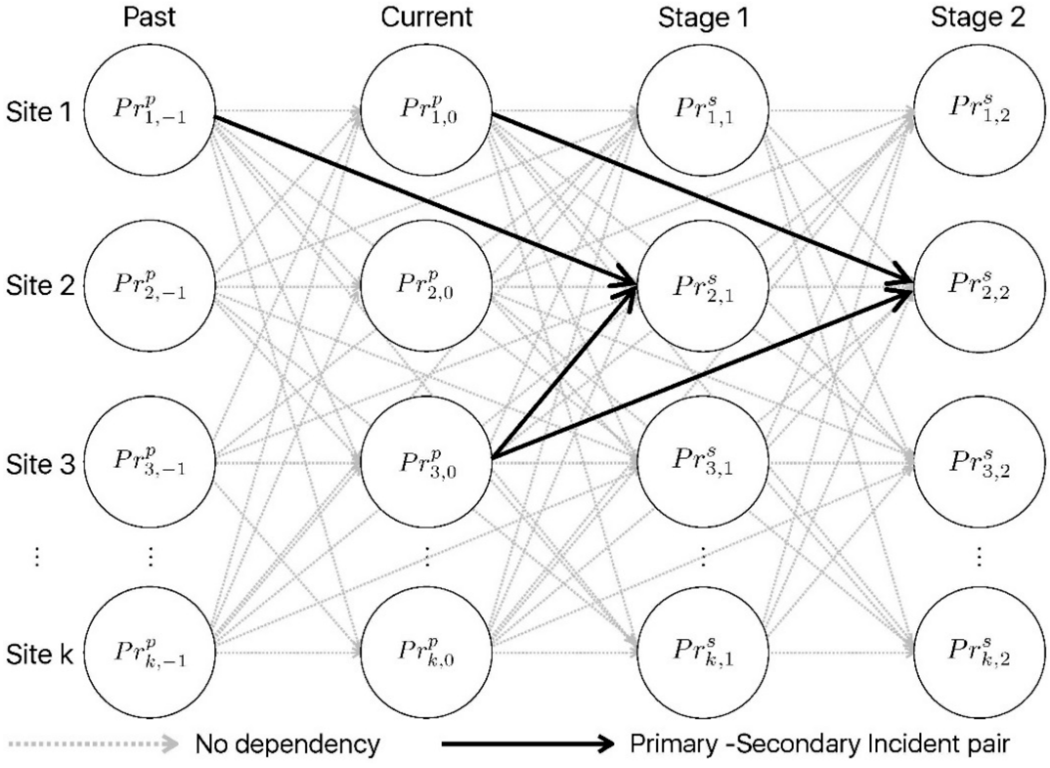


Figure 1: Stochastic process of incident occurrences (Two-steps ahead).

$Pr^p{}_{(i,r)}$ denotes corresponding probability of primary and independent incidents at site □ during stage r, and $Pr^s{}_{(k,u)}$ denotes corresponding probability of secondary incident occurrences at site k during stage u.

$$E[\tau(i,r)] = Pr^p{}_{(i,r)} + Pr^s{}_{(k,u)} \qquad for\ i = k, r = u \tag{2}$$

First, we use the Poisson process (Koutsopoulos and Yablonski 1991; Daneshgar et al. 2013) 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 (Kim et al. 2014) 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 is used in this study. The average of subintervals is $T\lambda^{-1}$(with variance $T\lambda^{-2}$). The discrete random incidents are assumed to be Poisson distributed with incident rate λ^r indicated by $\boldsymbol{X}$ ~*Poisson* (λ^r). Using probability mass function where the count of incidents is one, normalized probability of incident occurring at location i for each interval r is

$$Pr_{(i,r)} = \lambda_i^r e^{-\lambda_i^r}\left(\sum_i \lambda_i^r e^{-\lambda_i^r}\right)^{-1} \qquad \forall i, r \tag{3}$$

Second, the probability of secondary incidents $Pr^s{}_{(k,u)}$ is a function of $Pr^p{}_{(i,r)}$ 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 (Park and Haghani 2014, 2016a). Each primary incident at site *i* during stage *r* has different impact on future secondary incident occurrences. We introduce an indicator function, I(**Ω**, Δ)$_{(i,r)(k,u)}$, 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 $Pr^p{}_{(3,0)}$ to $Pr^s{}_{(2,1)}$ is $\delta(\Omega,\Delta)_{(3,0)(2,1)}$=1). With introduced parameters and variables, we propose the probability of secondary incidents $Pr^s{}_{(k,u)}$ in an explicit form:

$$Pr^s{}_{(k,u)} = \sum_i \delta(\Omega,\Delta)_{(i,r-1)(k,u)} Pr^p{}_{(i,r-1)} + \sum_i \delta(\Omega,\Delta)_{(i,r-2)(k,u)} Pr^p{}_{(i,r-2)} \tag{4}$$

Now, we insert the $Pr^s{}_{(k,u)}$ from Equation (4) to Equation (2). Suppose we are interested in incidents at site 2 in the first future-stage. The expected probability of incidents is:

$$E[\tau(2,1)] = Pr^p{}_{(2,1)} + \sum_i \delta(\Omega,\Delta)_{(i,0)(2,1)} Pr^p{}_{(i,0)}$$

$$+\sum_{i} \delta(\Omega, \Delta)_{(i,-1)(2,1)} Pr^{p}{}_{(i,-1)} \qquad (5)$$

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 $p(i, j)$.

$$\begin{array}{cccc}
r-1 & r & & r+2 \\
\begin{array}{c} 1 \\ (2) \\ 3 \\ \cdot \\ \cdot \\ \cdot \\ m \end{array} & \begin{array}{c} 1 \\ 2 \\ (3) \\ \cdot \\ \cdot \\ \cdot \\ m \end{array} & r+1 & \begin{bmatrix} p(1,1) & p(1,2) & \cdot\ \cdot\ \cdot & p(1,j) \\ p(2,1) & p(2,2) & \cdot\ \cdot\ \cdot & p(1,j) \\ p(3,1) & p(3,2) & \cdot\ \cdot\ \cdot & p(1,j) \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ p(i,1) & p(i,2) & \cdot\ \cdot\ \cdot & p(i,j) \end{bmatrix}
\end{array} \qquad (6)$$

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:

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

$$+ p(\forall i = j) \times \begin{bmatrix} p(1,1) = \mathbb{E}[\tau(1,1)] \times \mathbb{E}[\tau(1,2)] \\ p(2,2) = \mathbb{E}[\tau(2,1)] \times \mathbb{E}[\tau(2,2)] \\ p(3,3) = \mathbb{E}[\tau(3,1)] \times \mathbb{E}[\tau(3,2)] \\ . \\ . \\ . \\ p(i,j) = \mathbb{E}[\tau(i,1)] \times \mathbb{E}[\tau(i,2)] \end{bmatrix} \tag{7}$$

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 ($E[\tau\ (1, 1)]\times \mathrm{E}[\tau\ (2, 2)] \neq \mathrm{E}[\tau\ (2, 1)] \times \mathrm{E}[\tau\ (1, 2)]$).

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 (Park et al. 2016). For example, total delay, D_i, for each incident location *i* can be estimated using variables considered (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^{t}; 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:

$$D_i = (R_i + NC_i)\frac{(q_i - s_i')}{2q_i} \tag{8}$$

Uncertainty of incident clearance duration is another major challenge in quantifying the impact of incidents (Park et al. 2013b, 2015). 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 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.

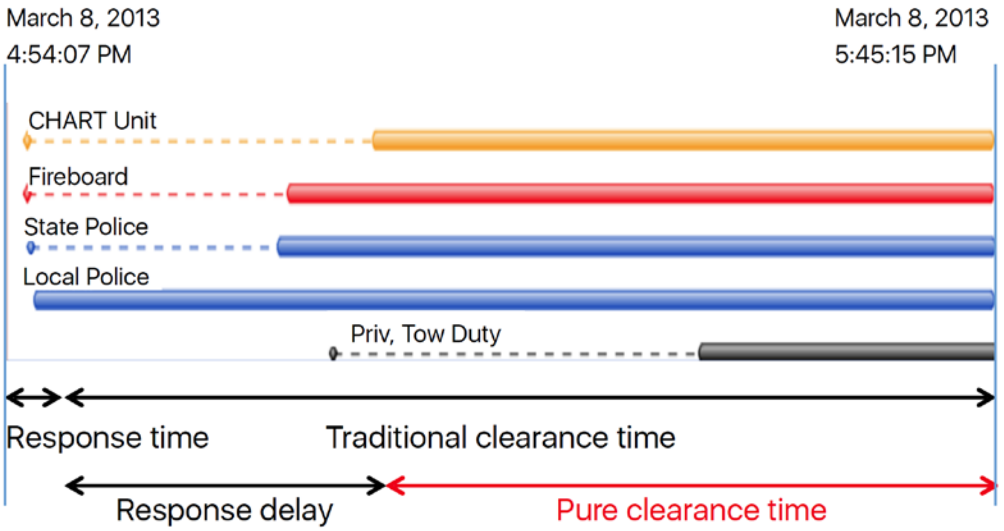


Figure 2: The concept of pure clearance time (Park and Haghani 2015a).

Potentially delayed clearance can be modeled by integrating delay-type with normal clearance time. A test (Park and Haghani 2016a) 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^{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 η = 1) from traditional normal clearance time NC_i at each location i,

$$C_i \ = \ NC_i \, \beta^{\eta} \tag{9}$$

In our optimization problem, the clearance time without delay is used as an in- put to minimize the total delay. For example, when we have the delay type (β^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[1]. Our main goal is getting required ERUs to the incident site as quickly as possible to reduce total incident-induced delay. See more details in Park and Haghani (2016a).

4.0 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 a 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 linearized and heuristics are introduced for a large scale problem.

4.1 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 primary 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 1 summarizes all notations used in the model formulations.

Table 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
E	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 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 (10) 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.

$$minimize\ z = \sum_{w}\sum_{o} P_w d_{o,w} \qquad (10)$$

The first group of constraints presents rules for assignment of ERUs. Constraints (11) 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.

$$a_{onjw} \leq \sum_{p<o} a_{pn(j-1)w} \qquad \forall\, w, n, o, j \neq 1 \qquad (11)$$

Constraints (12) 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*.

$$\sum_{o} a_{onjw} \leq 1 \qquad \forall w, n, v \qquad (12)$$

Constraints (13) make sure that each incident is assigned to one job of a vehicle.

$$\sum_{n}\sum_{j} a_{onjw} = 1 \qquad \forall w, o \qquad (13)$$

Constraints (14) are added so that multiple similar solutions would not occur.

$$a_{111w} = 1 \qquad \forall w \qquad (14)$$

Constraints (15) are enforcing that each vehicle has exactly one origin (starting location).

$$\sum_{i} x_{in} = 1 \qquad \forall n \qquad (15)$$

The second group of constraints shows starting time of each incident. Constraints (16) 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.

$$\boldsymbol{sv_{onw}} \times \boldsymbol{a_{on1w}} \geq \sum_{i} TT_{iL_o} \times \boldsymbol{x_{in}} \times \boldsymbol{a_{on1w}}$$

$$+H_{o,w} \times a_{on1w} \qquad \forall w, o, n \qquad (16)$$

Constraints (17) 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.

$$\boldsymbol{sv_{onw}} \times \boldsymbol{a_{onjw}} \geq \sum_{p<o} TT_{L_pL_o} \times a_{pn(j-1)w} + \sum_{p<o} \boldsymbol{cv_{pn(j-1)w}} \times \boldsymbol{a_{pn(j-1)w}} - M^{17}_{o,w} \times \left(1 - a_{onjw}\right) \qquad \forall w,o,n,j \neq 1 \tag{17}$$

The third group of constraints ensures serving time of each incidents. Constraints (18) and (19) define the starting and clearance times for each incident under each scenario, regardless of the vehicle covering it.

$$s_{ow} = \sum_{\boldsymbol{n}} \sum_{\boldsymbol{j}} \boldsymbol{sv_{onjw}} \times \boldsymbol{a_{onjw}} \qquad \forall w,o \tag{18}$$

$$c_{ow} = \sum_{\boldsymbol{n}} \sum_{\boldsymbol{j}} \boldsymbol{cv_{onjw}} \times \boldsymbol{a_{onjw}} \qquad \forall w,o \tag{19}$$

Constraints (20) ensure that each incident is not served any sooner than when it happens.

$$\boldsymbol{sv_{onw}} \times \boldsymbol{a_{onjw}} \geq H_{o,w} \times a_{onjw} + \sum_{p\,<\,o} TT_{L_pL_o} \times a_{pn(j-1)w} - M^{20}_{o,w} \times \left(1 - a_{onjw}\right) \qquad \forall w,o,n,j \neq 1 \tag{20}$$

Constraints (21) and (22) 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.

$$\boldsymbol{sv_{onjw}} \times \boldsymbol{a_{onjw}} \leq M^{21}_{o,w} \times \sum_{p<o} f_{pn(j-1)w} \qquad \forall w,o \neq 1,n,v \neq 1 \tag{21}$$

$$\boldsymbol{cv_{onjw}} \times \boldsymbol{a_{onjw}} - \boldsymbol{sv_{onw}} \times \boldsymbol{a_{onjw}} - CD_{L_o} \times a_{onjw} + \varepsilon \times a_{onjw} \leq M^{22}_{ow} \times f_{onjw} \qquad \forall w,o,n,j \tag{22}$$

Constraints (23) are for finding the soonest time an incident can be cleared.

$$c_{ow} \geq s_{ow} + CD_{L_{ow}} \qquad \forall w,o \tag{23}$$

The last group of constraints presents delay calculation based on above constraints and condition of each variable. Constraints (24) define the delay for an incident.

$$c_{ow} - H_{ow} = d_{ow} \qquad \forall w,o \tag{24}$$

Constraints (25) define non-negative and binary variables.

$$f_{onjw}, a_{onjw} \in \{0,1\} \qquad \forall w, n, o, j$$

$$x_{in} \in \{0,1\} \qquad \forall i, n \tag{25}$$

In the presented formulation, constraints (16), (17), (18), (19), (20), (21), (22) have non-linear terms. 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.

4.2 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 (McCormick 1976).

For linearizing $sv_{on\omega} \times a_{onj\omega}$ we have introduced a dummy variable $d1_{onj\omega}$ and added two constraints (26) and (27):

$$d1_{onjw} \leq sv_{onw} \qquad \forall w, o, n, j \tag{26}$$

$$d1_{onjw} \leq M \times a_{onjw} \qquad \forall w, o, n, j \tag{27}$$

The objective of adding constraints (26) 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 (27), we ensure that $d1_{onjw}$ will equal zero if a_{onjw} equals zero. The correctness of this type of linearization can be found in (McCormick 1976).

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 (28) are added as a result:

$$d2_{oniw} \geq x_{in} + a_{on1w} - 1 \qquad \forall w, o, n, i \tag{28}$$

The purpose of constraints (28) 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 (17), (20), (21), and (22) 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 2). It is advised to pick the smallest number within that domain.

The objective is to minimize a function of delay whenever we start serving the incident the fastest based on constraints (17). The nonlinear term $x_{in} \times a_{on1\omega}$ would always try to assume the value of zero. By adding constraints (20), 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 (21) and (22):

Table 2: The ranges for the big – Ms.

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

$$d3_{onjw} \leq M \times a_{onjw} \qquad \forall w,o,n,j \qquad (29)$$

$$d3_{onjw} \leq cv_{onjw} \qquad \forall w,o,n,j \qquad (30)$$

To linearize the nonlinear constraints, we replace the nonlinear terms with their linear equivalents. The linearized constraints are presented below:

$$d1_{on1w} \geq \sum_{i} TT_{iL_o} \times d2_{oniw} \qquad \forall o,n \qquad (31)$$

$$d1_{onjw} \geq \sum_{p<o} TT_{L_pL_o} \times a_{pn(j-1)w} + \sum_{p<o} d3_{pn(j-1)w}$$

$$- M \times \left(1 - a_{onjw}\right) \qquad \forall w,o,n,j \neq 1 \qquad (32)$$

$$s_{ow} = \sum_{n} \sum_{j} d1_{onjw} \qquad \forall w,o \qquad (33)$$

$$c_{ow} = \sum_{n} \sum_{j} d3_{onjw} \qquad \forall w, o \tag{34}$$

$$d1_{onjw} \geq H_{o,w} \times a_{onjw} + \sum_{p < o} TT_{L_p L_o} \times a_{pn(j-1)w} - M \times (1 - a_{onjw}) \qquad \forall w, o, n, j \neq 1 \tag{35}$$

$$d1_{onjw} \leq M \times \sum_{p < o} f_{pn(j-1)w} \qquad \forall w, o \neq 1, n, v \neq 1 \tag{36}$$

$$d3_{onjw} - d1_{onjw} - CD_{L_{ow}} \times a_{onjw} + \varepsilon \times a_{onjw} \leq M \times f_{onjw} \quad \forall w, o, n, j \tag{37}$$

4.3 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 (Heitsch and R¨omisch 2003). 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 $\Gamma^{[0]}$, $\Gamma^{[1]}$, ..., $\Gamma^{[i]}$, ..., $\Gamma^{[*]}$, where the set $\Gamma^{[*]}$ 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 (ω × *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 scenarios (×6 for full combinatorial in 3 stages) among *N*, the value of d_ω can be conveniently arranged into a systematic matrix,

$$d = \begin{bmatrix} 0 & 10 & \cdots & 1000 \\ 10 & 0 & \cdots & 990 \\ 25 & 15 & \cdots & 975 \\ \vdots & \vdots & \vdots & \vdots \\ 1000 & 990 & \vdots & 0 \end{bmatrix} \tag{38}$$

- *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

$$D_\omega = arg\{\min \sum_{w\in\Gamma} P_\omega\, d_{\omega\omega'}\} \tag{39}$$

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

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

$$\mathrm{d}_{ww'}^{[i]} = \min\left\{\mathrm{d}_{ww'}^{[i]}, \mathrm{d}_{ww(i-1)}^{[i]}\right\}, \forall w, w' \in \Gamma_J^{[i-1]} \tag{40}$$

Considering new delay matrix, we select

$$D_{\omega} \in arg\{\min \sum_{w \in \Gamma_{J}^{[i-1]}} \mathrm{D}_{w}^{[i]}\} \tag{41}$$

- *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 (Heitsch and Romisch 2003) for further information about the algorithm.

5.0 NUMERICAL EXAMPLES

5.1 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 3). There were 4 field operation patrol units available for AM peak hours on weekdays until 2014.

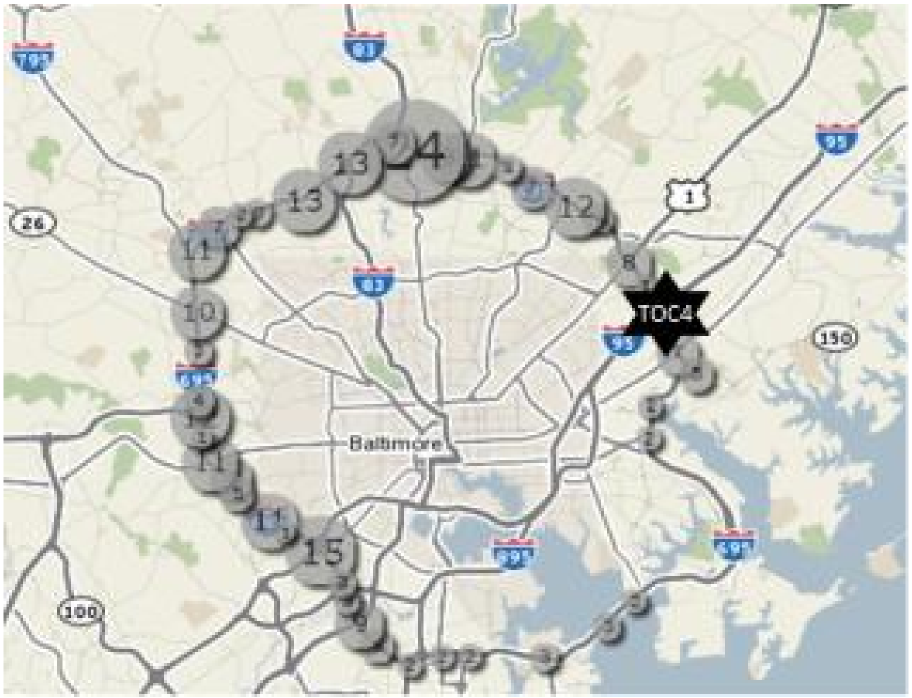


Figure 3: Spatial distribution of Incidents on I-695 freeway.

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 (Koutsopoulos and Yablonski 1991).

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 (A) 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 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: $\beta_1^1 = 0.68, \beta_1^2 = 0.94, \beta_1^3 = 1.05, \beta_1^4 = 1.358, \beta_1^5 = 0.98$. 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., β_i^2, η = 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 (Park and Haghani 2015b).

We build a total of u scenarios. For example, Table 3 presents 17 x 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 occurred at exit 11 ($□(\Omega, \Delta)_{(exit11,-1)(□,u)}$ = 0.207, Ω = 2 lanes blocked, collision with injuries; Δ = 30mph speed difference), and the current incident occurred at exit 5 ($\delta(\Omega, \Delta)_{(exit5,0)(k,u)}$= 0.098, Ω = 1 lanes blocked, collision with property damage; Δ = 10mph 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 (Park et al. 2015).

Table 3: Probabilities of scenarios.

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

5.2 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_w) ranges from 0.001 to 0.041 (average probability of a scenario is 0.013). For example, the probability of the first scenario, $P1$, $p(5,7) = \mathrm{E}[\tau(\text{exit}1, 2)] \times \mathrm{E}[\tau(\text{exit}7, 2)]$, 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 Ω 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, and 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 in estimating expected clearance of incidents C_1 (Park et al. 2015). We relocate n ERUs if the expected clearance C_i of Q 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 (Lei et al. 2015) 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 required time (Ball and Lin 1993). 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 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.

Figure 4 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 at 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.

Table 4: The performance of the proposed model (Different number of ERU vehicles).

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

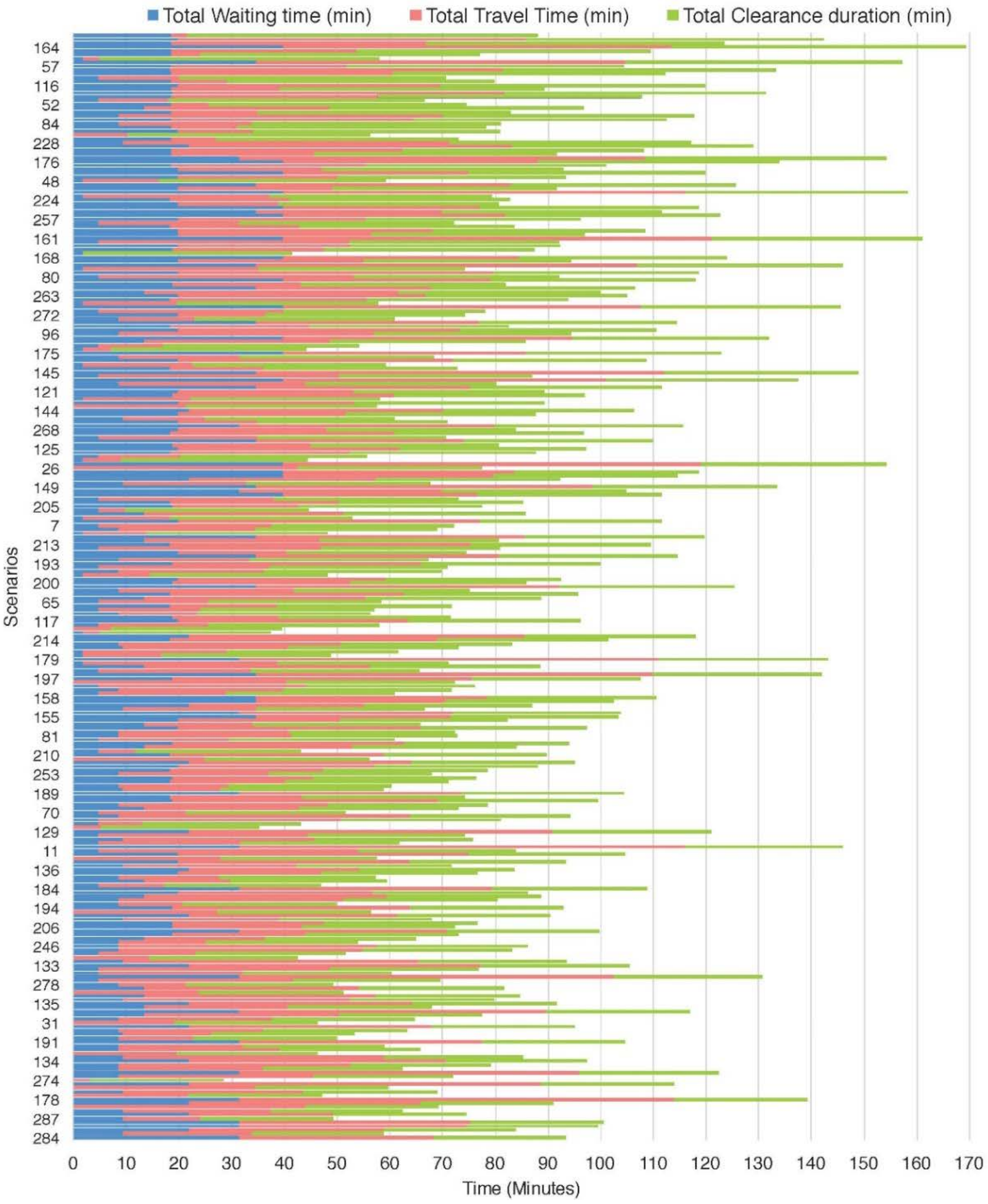


Figure 4: Optimal solutions for each scenario.

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 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 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

5.3 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 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 portion of the running time is devoted to finding better feasible solutions that only marginally improve the previous best found solution.

Table 6: Computational performances for the proposed approach.

No.	Parameters			Proposed approach		Fast forward selection	
	N	R	U	CPU time	Gap	CPU time	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%

As we have larger network size and more future stages, it is more time consuming. In this study, we used 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 less 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.

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.

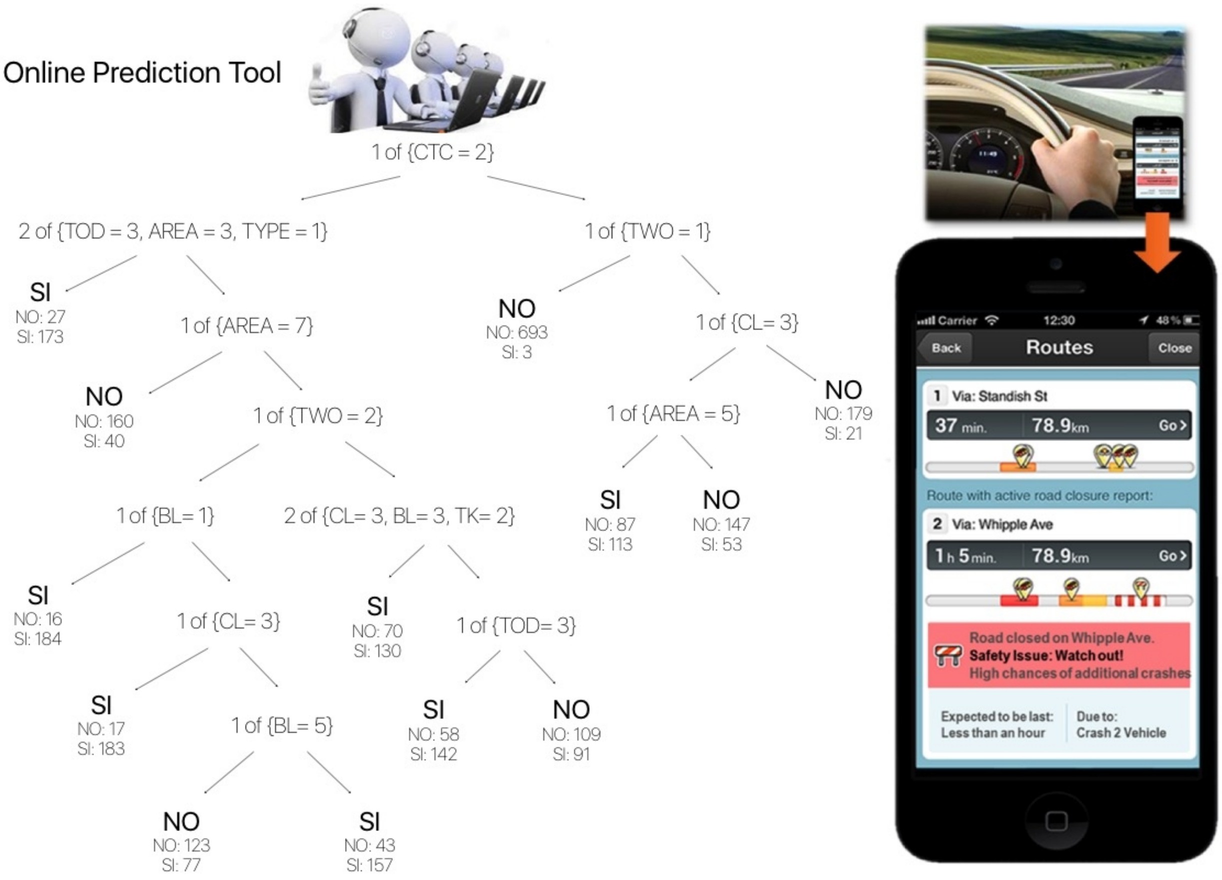
The time to respond to an incident is relatively small compared to the time necessary to clear the incident. We dynamically incorporate position of ERUs in each stage, and this formulation causes a high complexity. In a planning stage, before an incident occurs, we can run the full model without a restriction 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.

5.4 APPLICATION

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 can be developed to capture the time to secondary incident considering incident duration based on vehicle arrivals.

As shown in Figure 5, 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 (Park and Haghani 2015a). Smartphone application (e.g. WAZE) can help drivers navigate around road closures and get where they need to go. If the likelihood of secondary incidents is high, notifications like watch out could make driving safer.



Road closed on Whipple Ave.
Safety Issue: Watch out!
High chances of additional crashes

Expected to be last:
Less than an hour

Due to:
Crash 2 Vehicle

Figure 5: Application of incident online prediction tool.

Moreover, Smartphone application can have a mode for emergency services personnel to make a decision of relocating emergency vehicles. The conditional probability of a secondary incident at each location will be updated in real-time after incident sequence, incident severity, and environmental and traffic information. The updated information can be displayed in screen to provide an optimal route to emergency services personnel.

6.0 CONCLUSION AND RECOMMENDATIONS

6.1 CONCLUSION

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.

6.2 FUTURE RECOMMENDATION

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.

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. In the future study, we will 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 would flexibly re-computes the solution to react in real-time. Our practical online algorithm (Park and Haghani 2016b) has a look-ahead setting contingent on present requests in making future decisions.

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.

7.0 REFERENCES

References shall include only published works.

Andersson, T. and Varbrand, P. (2006) Decision support tools for ambulance dispatch and relocation. *Operation Research Society* 58(2), pp. 195-201.

Alanis, R. Ingolfsson, A. and Kolfal, B. (2013) A Markov chain model for an ems system with repositioning. *Production and Operations Management* 22(1), pp. 216-231.

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

Berman, O. (1981a) Repositioning of distinguishable urban service units on networks. *Computers and Operations Research* 8(2), pp. 105-118.

Berman, O. (1981b) Dynamic repositioning of indistinguishable service units on transportation networks. *Transportation Science* 15(2), pp. 115-136.

Brandeau, M.L. (1986) 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,* pp. 121-153.

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

Daneshgar, F. Mattingly, S. and Haghani, A. (2013) Evaluating beat structure and truck allocation for the Tarrant county, Texas, courtesy patrol. *Transportation Research Record: Journal of the Transportation Research Board,* No. 2334, pp. 40-49.

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

Daskin, M.S. and Haghani, A. (1984) Multiple vehicle routing and dispatching to an emergency scene. *Environment and Planning A,* 16(10), pp. 1349-1359.

Gendreau, M., Laporte, G., and Semet, F. (1997) Solving an ambulance location model by tabu search. *Location Science* 5(2), pp.75-88.

Gendreau, M., Laporte, G., and Semet, F. (2001) A dynamic model and parallel Tabu search heuristic for real-time ambulance relocation. *Parallel Computing* 27(12), pp. 1641-1653.

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

Haghani, A., Iliescu, D., Hamedi, M. and Yong. S. (2006) 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.*

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

Hakimi, S.L. (1964) Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research* 12(3), pp. 450-459.

Halper, R. and Raghavan, S. (2011) The mobile facility routing problem. *Transportation Science,* 45(3), pp. 413-434.

HCM. *Highway Capacity Manual 2010.* Transportation Research Board, Washington, DC, 2010.

Heitsch, H. and Romisch, W. (2003) Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications* 24(2-3), pp. 187-206.

Kim, H. Kim, W. Chang, G.L. and Rochon, S. (2014) Design of emergency response system to minimize incident impacts. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2470, pp. 65-77.

Koutsopoulos, H.N. and Yablonski, A. (1991) 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*.

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

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

Li, J. Lan, C.J., and Gu, X. (2006) Estimation of incident delay and its uncertainty on freeway networks. *Transportation Research Record: Journal of the Transportation Research Board* 1959(1), pp. 37-45.

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

Mirchandani, P.B. and Odoni, A.R. (1979) Locations of medians on stochastic networks. *Transportation Science* 13(2), pp. 85-97.

Nair, R. and Miller-Hooks, E. (2009) Evaluation of relocation strategies for emergency medical service vehicles. *Transportation Research Record: Journal of the Transportation Research Board*, No. 2137, pp. 63-73.

National Traffic Incident Management Coalition. (2007) Benefits of traffic incident management.

Ng, M.W., Khattak, A.J. and Talley, W.K. (2013) Modeling the time to the next primary and secondary incident: A semi-Markov stochastic process approach. *Transportation Research Part B: Methodological* 58, pp. 44-57.

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

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

Park, H., Haghani, A., and Zhang, X. (*2013b*) ATIS: Interpretation of Bayesian neural network for predicting the duration of detected incidents. *Presented at the 92nd Annual Meeting of Transportation Research Board (CD-ROM), Washington, DC.*

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

Park, H., Haghani, A., and Aliari, Y. (2014) 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*.

Park, H. and Haghani, A. (2015a) Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C: Emerging Technologies,* Article in advance.

Park, H. and Haghani, A. (2015b) Capacity adjustment considering the impact of secondary incidents. *Presented at 94th Annual Meeting of the Transportation Research Board, Washington, DC.*

Park, H., Haghani, A., and Zhang, X. (2015) Interpretation of Bayesian neural network for predicting the duration of detected incidents. *Journal of Intelligent Transportation Systems: Technologies, Planning, and Operations.* Article in advance.

Park, H. and Haghani, A. (2016a) Stochastic capacity adjustment considering secondary incidents. *IEEE Transactions on Intelligent Transportation Systems,* in press.

Park, H. and Haghani, A. (2016b) Online emergency vehicle dispatching with look-ahead on a transportation network. *submitted to the 95th Annual Meeting of the Transportation Research Board, Washington, DC.*

Park, H., Haghani, A., and Shafahi, A. (2016) Stochastic emergency response units allocation with secondary incident occurrences. *IEEE Transactions on Intelligent Transportation Systems,* In press.

Prodhon, C. and Prins, C. (2014) A survey of recent research on location-routing problems. *European Journal of Operation Research* 238, pp. 1-17.

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

Savas, E.S. (1969) Simulation and cost-effectiveness analysis of New York’s emergency ambulance service. *Management Science* 15(12), pp. 608-627.

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

Skabardonis, A., Petty, K.F., and Varaiya, P.P. (1999) Los Angeles I-10 field experiment: Incident patterns. *Transportation Research Record: Journal of the Transportation Research Board,* No. 1683, pp. 22-30.

Suzuki, A. and Drezner, Z (*1996)* The p-center location problem in an area. Location Science *4* pp. 69-82.

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

Yang, S. Hamedi, M., and Haghani, A. (2005) Online dispatching and routing model for emergency vehicles with area coverage constraints. *Transportation Research Record: Journal of the Transportation Research Board,* No. 1923, pp. 1-8.

Zhang, *L.* (2010) Optimization of small-scale ambulance move-up. In *Proceedings of the 45th Annual Conference of the New Zealand Operation Research Society.*